OXFORD

Post-selection inference in regression models for group testing data

Qinyan Shen, Karl Gregory [ORCID]*, Xianzheng Huang

Department of Statistics, University of South Carolina, Columbia, SC 29208, United States of America

*Corresponding author: Karl Gregory, Department of Statistics, University of South Carolina, 219 LeConte, 1523 Greene St., Columbia, SC 29208, United States of America (gregorkb@stat.sc.edu).

ABSTRACT

We develop a methodology for valid inference after variable selection in logistic regression when the responses are partially observed, that is, when one observes a set of error-prone testing outcomes instead of the true values of the responses. Aiming at selecting important covariates while accounting for missing information in the response data, we apply the expectation-maximization algorithm to compute maximum likelihood estimators subject to LASSO penalization. Subsequent to variable selection, we make inferences on the selected covariate effects by extending post-selection inference methodology based on the polyhedral lemma. Empirical evidence from our extensive simulation study suggests that our post-selection inference results are more reliable than those from naive inference methods that use the same data to perform variable selection and inference without adjusting for variable selection.

KEYWORDS: confidence intervals; EM algorithm; individual testing; LASSO; variable selection.

1 INTRODUCTION

When testing a large number of individuals for a disease or infection, it can be advantageous to test individuals in groups rather than one by one. This strategy is called group testing. In group testing, individuals are placed (usually at random) into groups (often small, of 2-10 individuals), and the specimens drawn from the individuals in each group are combined into a single pooled specimen. Then the pooled specimen is tested for the presence of the disease or infection, giving a positive or negative result for each group. If individual diagnoses are desired, individuals in groups that tested positive can be retested one by one; depending on the goals of the study, one may or may not retest individuals in groups that tested positive.

One advantage of group testing is that it can greatly reduce the cost of disease surveillance. Even if individuals in positive pools are retested, it is the case that when the disease is rare, the majority of the pools will test negative, resulting in a significant reduction in the total number of tests that need to be performed. A more hidden advantage to group testing is that, when positive pools are retested, one can, if the disease is rare, obtain more accurate estimators of disease prevalence, etc. than one would obtain if one tested individuals one by one. The reason for this is that each test is assumed to be liable to error; that is, it can give a false positive or a false negative result. The source of the hidden advantage can be put simply: The smaller the number of tests, the fewer the opportunities to make one of these errors. The sensitivity and specificity of the tests will therefore play an important role in any analysis of group testing data. The sensitivity of a test is the probability that it is positive when the infection is present, while its specificity is the probability that it is negative when the infection is absent. Throughout our work, we will assume that the sensitivity and specificity of the group or individual tests are known and that they are less than 1, so that each testing outcome is liable to error. Given the liability to error of each test, we will throughout this work make a distinction between testing outcomes associated with individuals, which we can observe, and individuals' true disease statuses, which we *cannot* observe.

Group testing has been implemented with success in many settings. Sinnott-Armstrong et al. (2020) demonstrated that the number of individuals it was possible to test was doubled with only a marginal increase in the number of testing kits needed when utilizing group testing for SARS-CoV-2 RNA detection; Verdun et al. (2021) further optimized group testing strategies tailored to the COVID-19 pandemic, leading to a 10-fold increase in testing efficiency. Pilcher et al. (2020) developed a model to assess the efficiency and accuracy of specimen pooling algorithms for SARS-CoV-2 and showed that group testing could enable rapid scale-up of testing and real-time surveillance of incidence. McMahan et al. (2017) report that the State Hygienic Laboratory at the University of Iowa, which tests thousands of Iowans annually for chlamydia and gonorrhea, has saved approximately $3.1 million from 2009 to 2014 by adopting a group testing approach.

Besides disease detection, researchers are interested in performing logistic regression with group testing data in order to identify risk factors or important covariates for the prediction of disease incidence at the individual level. The main challenge to

fitting regression models with group testing data is that instead of observing individuals' true disease statuses, we observe testing outcomes that are only guaranteed to be correct with certain probabilities. As we will see in the development of our methods, the likelihood functions based on the observed testing outcomes are much more complicated than the likelihood functions based on individuals' true disease statuses, and they depend intricately on the sensitivity and specificity of the tests. Xie (2001), Zhang et al. (2013), and Gregory et al. (2019) addressed these complications by employing the expectation-maximization (EM) algorithm to estimate parameters in generalized linear regression models for the individual response based on group testing data. McMahan et al. (2017) took a Bayesian approach for regression analysis of group testing data. Joyner et al. (2020) formulated a Bayesian generalized linear mixed model that can accommodate any group testing protocol, and used a spike-and-slab prior to facilitate variable selection. Focusing on multiple-infection group testing data, Lin et al. (2019) proposed a method exploiting the LASSO penalty for variable selection to identify risk factors for each infection.

Subsequent to variable selection, one may wish to make inferences on the effects of the selected covariates, such as testing hypotheses regarding covariate effects or constructing confidence intervals (CIs) for them. A naive approach to inference after variable selection is to fit the regression model using only the selected covariates and then to make inference as though the process of choosing the covariates had not used the same data. There is abundant evidence in the literature that this naive approach can lead to inflated Type I error rates and CIs that are too narrow (Berk et al., 2013; Lee et al., 2016; Tibshirani et al., 2016). As our own simulation studies demonstrate, naive post-selection inference will have the same undesired consequences in the case of regression with group testing data. We present a procedure for making valid post-selection inferences in this setting.

The rest of the paper is organized as follows: In Section 2, we provide a brief review of post-selection inference literature. In Section 3, we present the EM algorithm for LASSO-penalized estimation of regression parameters in a logistic regression model for individual testing data when the individual tests have some sensitivity Se and specificity Sp. We then derive a method for valid post-selection inference. Section 4 extends the method to group testing data. Section 5 presents simulation studies of the empirical performance of the proposed methodology and illustrates its implementation on a real data set. Section 6 outlines key takeaways and suggestions for future research.

2 OVERVIEW OF POST-SELECTION INFERENCE

Throughout this paper, we will use the terms *model selection* and *variable selection* interchangeably to mean the process of deciding from among a set of available predictors which ones to keep. Post-selection inference aims to account for the uncertainty introduced by the model selection process, in contrast to classical statistical inference, which assumes the model has been selected or is known before the data are observed. Kuchibhotla et al. (2022) provide a comprehensive review on 3 strategies for post-selection inference.

One strategy is sample splitting (Meinshausen et al., 2009; Wasserman and Roeder, 2009; Fithian et al., 2014; Rinaldo et al., 2019), whereby one splits the observed data into training and testing data sets, using the training set to perform model or variable selection and the testing set to perform inference in the selected model. This strategy avoids "double dipping" into the observed data and guarantees inferences unaffected by the selection process. A downside is the loss in power entailed by using part of the data solely for variable selection. Inference based on the data splitting approach may also be difficult to replicate, as the model selected as well as the inferences in the selected model may be very sensitive to which observations fall in the training versus the testing set (Rasines and Young, 2022).

The work on simultaneous inference by Berk et al. (2013), along with subsequent works by Zhang and Cheng (2017) and Bachoc et al. (2019, 2020), among many others, contributed to the second strategy of post-selection inference. A common goal of methods under this theme is to control the family-wise error rate for all parameters in the selected model. By striving for "universally valid post-selection inference for all model selection procedures..." (Berk et al., 2013, Section 4.4) and controlling the probability of making an error for *any* parameter in a selected model, this strategy can lead to overly conservative CIs.

The third strategy, which we adopt in this paper, attempts to make inferences in the selected model that are conditioned on the event that the selected model was selected. If one can properly condition on the selection event, one can, for example, construct CIs that maintain nominal coverage rates even when constructed with the same data used in model selection. The key step is to find the conditional sampling distributions of one's estimators given the model selection event. In linear regression, for example, one may begin by considering the sampling distribution of the least-squares estimators (LSEs) in a given, pre-specified model. Then one introduces a selection event, which is the event that the given model is selected by a variable selection procedure. If the selection event can be expressed as affine constraints on the responses, then analytic derivations of the conditional distributions of the LSEs given the selection event become more feasible. It has been shown that variable selection in linear regression via forward stepwise selection, least angle regression, and LASSO-penalization results in selection events that place the vector of responses in a polyhedral set (Lee et al., 2016; Tibshirani et al., 2016). Due to this property, one finds, in the case of Gaussian noise, that the sampling distributions of the LSEs in the selected model are truncated Gaussians. These truncated distributions properly account for variable selection and can thus be used to make valid post-selection inferences. Tibshirani et al. (2018) showed that this strategy works asymptotically in the case of non-Gaussian noise, and Taylor and Tibshirani (2018) outlined a generalization of these results to logistic regression and proportional hazards models. Hyun et al. (2018) developed inference conditioning on model selection events that are defined by the generalized LASSO regularization path. More recent developments in this direction aim to analytically derive exact or approximately valid post-selection inference under more complicated model settings with different model selection procedures (Kuchibhotla et al., 2020; Panigrahi et al., 2021; Neufeld et al., 2022; Rugamer et al., 2022; Zhao et al., 2022; Panigrahi and Taylor, 2023).

To make valid post-selection inferences in logistic regression with group testing data, we follow the third strategy. We first introduce our covariate effect estimators as well as a selection event. Then we describe how to make post-selection inferences based on the approximate conditional sampling distributions of our estimators given the selection event. We introduce our procedure in the context of logistic regression with individual testing data, in which each individual is tested with a test having some sensitivity Se and some specificity Sp. Then we extend our methods to the group testing setting in Section 4.

3 POST-SELECTION INFERENCE BASED ON INDIVIDUAL TESTING DATA

3.1 The post-selection statistic

Denote by $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ a vector of binary responses indicating the unobservable true disease status of n individuals, and by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ an $n \times p$ design matrix with values of p covariates for each of the n individuals. We assume a logistic regression model such that $Y_i \mid \mathbf{x}_i \sim \text{Bernoulli}(\pi_i(\boldsymbol{\theta}))$ for $i = 1, \ldots, n$, with $\pi_i(\boldsymbol{\theta}) = 1/\{1 + \exp(-\alpha - \mathbf{x}_i^\top \boldsymbol{\beta})\}$, where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\top)^\top$. Throughout we will regard $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as fixed, so we will no longer explicitly condition on the covariate values.

Suppose we observe individual testing outcomes $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ in place of $\mathbf{Y}$, where the tests have sensitivity Se and specificity Sp, so that $Z_i \mid Y_i \sim \text{Bernoulli}(\text{Se} \cdot Y_i + (1 - \text{Sp}) \cdot (1 - Y_i))$, for $i = 1, \ldots, n$. We assume $Z_i \mid Y_i$ does not depend on $\mathbf{x}_i$. Moreover, we assume Se and Sp to be known. Several works in group testing data make this same assumption, for example, Gregory et al. (2019), Xie (2001), and Zhang et al. (2013), whereas some other works aim to estimate Se and Sp in tandem with the regression coefficients $\boldsymbol{\theta}$, such as McMahan et al. (2017) and Joyner et al. (2020). If treating Se and Sp as unknown, one typically takes a Bayesian approach to estimating them (Puggioni et al., 2008). In some cases, point estimates of Se and Sp for tests that are in widespread use are available from studies like Haugland et al. (2010), which presented Se and Sp point estimates of 0.890 and 0.992, respectively, for a strand displacement assay on cervical swabs for *Chlamydia trachomatis* in women. Given that such information may be available, and so that we may focus on the area of post-selection inference, we will proceed under the assumption that Se and Sp are known.

Viewing $\mathbf{Y}$ as missing data, we obtain the regularized maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ using the EM algorithm subject to LASSO penalization, as outlined in Algorithm 1. With complete data $(\mathbf{Y}, \mathbf{Z})$, the complete data log-likelihood has the form

$$\ell_c(\boldsymbol{\theta}; \mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^n \left\{ Y_i \log(\pi_i(\boldsymbol{\theta})) + (1 - Y_i) \log(1 - \pi_i(\boldsymbol{\theta})) \right\} + C(\mathbf{Y}, \mathbf{Z}), \tag{1}$$

where $C(\mathbf{Y}, \mathbf{Z})$ is the log-likelihood of $\mathbf{Z}$ given $\mathbf{Y}$, which is assumed to be free of $\boldsymbol{\theta}$. In the E-step, we compute the expectation of Equation 1 given $\mathbf{Z}$ under the current estimate for $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(k)}$. This gives the objective function (into which we later incorpo-

rate the LASSO penalty)

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \mathbf{Z}) = \sum_{i=1}^n \left\{ \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Y_i | Z_i] \log(\pi_i(\boldsymbol{\theta})) + (1 - \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Y_i | Z_i]) \log(1 - \pi_i(\boldsymbol{\theta})) \right\}, \tag{2}$$

where we have removed the term involving $C(\mathbf{Y}, \mathbf{Z})$ since it does not depend on $\boldsymbol{\theta}$, and where, for $i = 1, \ldots, n$, we have

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Y_i | Z_i] = \frac{\text{Se}^{Z_i}(1 - \text{Se})^{1-Z_i} \pi_i(\boldsymbol{\theta}^{(k)})}{\{P_{\boldsymbol{\theta}^{(k)}}(Z_i = 1)\}^{Z_i} \{P_{\boldsymbol{\theta}^{(k)}}(Z_i = 0)\}^{1-Z_i}}, \tag{3}$$

in which $P_{\boldsymbol{\theta}^{(k)}}(Z_i = 0) = (1 - \text{Se})\pi_i(\boldsymbol{\theta}^{(k)}) + \text{Sp}\{1 - \pi_i(\boldsymbol{\theta}^{(k)})\}$, and $P_{\boldsymbol{\theta}^{(k)}}(Z_i = 1) = 1 - P_{\boldsymbol{\theta}^{(k)}}(Z_i = 0)$. In the M-step, one updates the estimate for $\boldsymbol{\theta}$ by maximizing $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \mathbf{Z})$ in Equation 2 penalized by LASSO with a penalty parameter $\lambda \geq 0$. The outputs at convergence of Algorithm 1 are the penalized MLEs for the regression coefficients in $\boldsymbol{\theta}$.

Algorithm 1 EM algorithm for LASSO-penalized logistic regression based on imprecise individual testing data

---

1: Initialize an estimate $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}$. Set the iteration counter at $k = 0$.
2: Compute $\hat{Y}_i^{(k)} = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Y_i | Z_i]$, for $i = 1, \ldots, n$, according to (3).
3: Obtain

$$\boldsymbol{\theta}^{(k+1)} \longleftarrow \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\text{argmax}} \sum_{i=1}^n \left\{ \hat{Y}_i^{(k)} \log(\pi_i(\boldsymbol{\theta})) + (1 - \hat{Y}_i^{(k)}) \log(1 - \pi_i(\boldsymbol{\theta})) \right\} - \lambda \|\boldsymbol{\beta}\|_1. \tag{4}$$

4: Repeat Steps 2 and 3 until $\frac{1}{\sqrt{n}} \|\hat{\mathbf{Y}}^{(k+1)} - \hat{\mathbf{Y}}^{(k)}\|_2 \langle 10^{-6}$.

Incorporating the LASSO penalty into each M-step allows one to perform variable selection. Let $M = \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}$ be the set containing the indices of the selected covariates, and let $|M|$ be the cardinality of M. Subsequent to variable selection, we consider making inferences on $\boldsymbol{\theta}_M = (\alpha, \boldsymbol{\beta}_M^\top)^\top$ based on $\mathbf{Z}$ and $\mathbf{X}_M$, where $\boldsymbol{\beta}_M$ contains the regression coefficients associated with the selected covariates, and $\mathbf{X}_M$ is the $n \times |M|$ design matrix with the columns of $\mathbf{X}$ with indices in M. Similarly, let $\hat{\boldsymbol{\theta}}_M = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_M^\top)^\top$. We will base our post-selection inferences on the estimator

$$\bar{\boldsymbol{\theta}}_M = (\bar{\alpha}, \bar{\boldsymbol{\beta}}_M^\top)^\top = \left([\mathbf{1}_n, \mathbf{X}_M]^\top \hat{\mathbf{W}}[\mathbf{1}_n, \mathbf{X}_M] \right)^{-1} \times [\mathbf{1}_n, \mathbf{X}_M]^\top \hat{\mathbf{W}} \dot{\mathbf{z}}, \tag{5}$$

where $\hat{\mathbf{W}} = \text{diag}(\hat{\pi}_1(1 - \hat{\pi}_1), \ldots, \hat{\pi}_n(1 - \hat{\pi}_n))$, and $\hat{\mathbf{z}} = \hat{\alpha} \mathbf{1}_n + \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{W}}^{-1}(\hat{\mathbf{Y}} - \hat{\boldsymbol{\pi}})$, in which $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$ and $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)^\top$, with $\hat{Y}_i = \mathbb{E}_{\hat{\boldsymbol{\theta}}}[Y_i | Z_i]$ and $\hat{\pi}_i = \mathbb{E}_{\hat{\boldsymbol{\theta}}}[Y_i | \mathbf{x}_i]$, for $i = 1, \ldots, n$. The post-selection estimator in Equation 5 is motivated by the iteratively reweighted least squares (IRLS) algorithm for fitting the logistic regression model for $\mathbf{Y}$ based on

the complete data log-likelihood given in Equation 1 evaluated at $\hat{\mathbf{Y}}$. In other words, we use the estimated individual responses in $\hat{\mathbf{Y}}$ at convergence from the EM algorithm as if they were the observed data, and, with $\hat{\boldsymbol{\theta}}_M$ as an initial estimate in the IRLS algorithm, we obtain a weighted LSE of $\boldsymbol{\theta}_M$ that minimizes $(\hat{\mathbf{z}} - \alpha - \mathbf{X}_M \boldsymbol{\beta}_M)^\top \hat{\mathbf{W}} (\hat{\mathbf{z}} - \alpha - \mathbf{X}_M \boldsymbol{\beta}_M)$. Without involving model selection, one may consider making inferences based on the MLE of $\boldsymbol{\theta}_M$. But this estimator in logistic regression is not available in closed form, which makes it a difficult statistic to use as the first ingredient for developing conditional selective inference. Our choice of the post-selection estimator $\bar{\boldsymbol{\theta}}_M$ in Equation 5 resembles an LSE in Gaussian linear regression, the setting in which we can establish post-selection inference results.

Next, we characterize the selection event and relate this event to the post-selection estimator to prepare for deriving its conditional distribution given the selection event.

3.2 The selection event

The event on which we will condition in order to make post-selection inference is the event that our LASSO-penalized estimation results in the selection of the covariate set M as well as in the vector of signs given by $\text{sign}(\hat{\boldsymbol{\beta}}_M)$ for the estimated coefficients of the selected covariates. Following the work of Lee et al. (2016), we show next that the selection event can be characterized as a set of affine constraints on the estimator $\bar{\boldsymbol{\beta}}_M$ introduced in Equation 5.

The selected model is revealed by $\hat{\boldsymbol{\theta}}$ at convergence of the EM algorithm. If one views the estimated individual probabilities of disease $\hat{\mathbf{Y}}$ at convergence as the actual response data, then the M-step in the last iteration of (4) is equivalent to the minimization of $-\ell_c(\boldsymbol{\theta}; \hat{\mathbf{Y}}, \mathbf{Z})$ plus the LASSO penalty. This can be written in weighted least squares form as

$$\underset{(\alpha, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}}{\text{argmin}} \ (\hat{\mathbf{z}} - \alpha - \mathbf{X}\boldsymbol{\beta})^\top \hat{\mathbf{W}} (\hat{\mathbf{z}} - \alpha - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1. \quad (6)$$

According to the Karush–Kuhn–Tucker (KKT) conditions, $\hat{\boldsymbol{\theta}}_M = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_M^\top)^\top$ satisfies

$$\mathbf{1}_n^\top \hat{\mathbf{W}} (\mathbf{1}_n \hat{\alpha} + \mathbf{X}_M \hat{\boldsymbol{\beta}}_M - \hat{\mathbf{z}}) = 0, \quad (7)$$

$$\mathbf{X}_M^\top \hat{\mathbf{W}} (\mathbf{1}_n \hat{\alpha} + \mathbf{X}_M \hat{\boldsymbol{\beta}}_M - \hat{\mathbf{z}}) + \lambda \mathbf{s}_M = \mathbf{0}, \quad (8)$$

$$\mathbf{X}_{-M}^\top \hat{\mathbf{W}} (\mathbf{1}_n \hat{\alpha} + \mathbf{X}_M \hat{\boldsymbol{\beta}}_M - \hat{\mathbf{z}}) + \lambda \mathbf{s}_{-M} = \mathbf{0}, \quad (9)$$

where $\mathbf{X}_{-M}$ is the $n \times (p - |M|)$ matrix containing the columns of $\mathbf{X}$ with indices not in M, $\mathbf{s}_M = \text{sign}(\hat{\boldsymbol{\beta}}_M)$, and $\mathbf{s}_{-M}$ is a $(p - |M|) \times 1$ vector with each entry in $[-1, 1]$. The Equations 7-9 are similar to the KKT conditions from Taylor and Tibshirani (2018). Setting $\mathcal{J}(\hat{\boldsymbol{\theta}}_M) = [\mathbf{1}_n, \mathbf{X}_M]^\top \hat{\mathbf{W}} [\mathbf{1}_n, \mathbf{X}_M]$, Equations 7 and 8 allow us to connect the penalized estimator $\hat{\boldsymbol{\beta}}_M$ and the post-selection estimator $\bar{\boldsymbol{\beta}}_M$ with the expression

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}_M^\top)^\top = (\bar{\alpha}, \bar{\boldsymbol{\beta}}_M^\top)^\top - \mathcal{J}^{-1}(\hat{\boldsymbol{\theta}}_M)[0, \lambda \mathbf{s}_M^\top]^\top. \quad (10)$$

From here we can translate the selection event to constraints on $\bar{\boldsymbol{\beta}}_M$.

Adapting the formulation of a selection event resulting from least squares estimation subject to the LASSO penalty in Lee et al. (2016, Theorem 4.3) to the penalized weighted least squares problem in Equation 6, the selection event can be expressed as affine constraints on $\bar{\boldsymbol{\beta}}_M$ implied by $\{\text{sign}(\hat{\boldsymbol{\beta}}_M) = \mathbf{s}_M\}$. Using Equation 10, one can show that these constraints are equivalent to $\{\mathbf{A}_1 \bar{\boldsymbol{\beta}}_M \leq \mathbf{b}_1\}$, where $\mathbf{A}_1 = -\text{diag}(\mathbf{s}_M)$, and

$$\mathbf{b}_1 = -\text{diag}(\mathbf{s}_M)[\mathbf{0} \ \mathbf{I}_{|M|}] \mathcal{J}^{-1}(\hat{\boldsymbol{\theta}}_M)[0, \lambda \mathbf{s}_M^\top]^\top. \quad (11)$$

We next adjust an approximate pre-selection sampling distribution of $\bar{\boldsymbol{\beta}}_M$ for these constraints, obtaining an approximate post-selection sampling distribution for a linear contrast $\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M$ conditional on the selection event.

3.3 The distribution of the post-selection estimator

Suppose the truth of $\boldsymbol{\theta}$ gives $\{j \in \{1, \ldots, p\} : \beta_j \neq 0\} \subseteq M$, that is, the selected model M is correct in the sense that it contains all the truly important covariates. Rearranging Equation 10 gives $\bar{\boldsymbol{\theta}}_M = \hat{\boldsymbol{\theta}}_M + \mathcal{J}^{-1}(\hat{\boldsymbol{\theta}}_M)[0, \lambda \mathbf{s}_M^\top]^\top$, where $\mathcal{J}(\hat{\boldsymbol{\theta}}_M)$ can be viewed as the observed Fisher information corresponding to the weighted least squares regression for fitting the selected model, and $[0, \lambda \mathbf{s}_M^\top]^\top$ as the corresponding score evaluated at $\hat{\boldsymbol{\theta}}_M$. Thus, $\bar{\boldsymbol{\theta}}_M$ can be viewed as a 1-step update from the initial value $\hat{\boldsymbol{\theta}}_M$ in the context of an unpenalized weighted least squares regression for a correct model. By Theorem 7.3.3 regarding a 1-step estimator in Lehmann (1999), we have $\sqrt{n}(\bar{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta}_M))$, provided $\hat{\boldsymbol{\theta}}_M$ is $\sqrt{n}$-consistent for $\boldsymbol{\theta}_M$, where $\mathcal{I}(\boldsymbol{\theta}_M)$ is the Fisher information corresponding to the weighted least squares regression. The $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_M$ as the solution to Equation 6 can be achieved by choosing the penalty parameter λ such that $\lim_{n \to \infty} \lambda / \sqrt{n} = 0$ (Theorem 2, Knight and Fu, 2000).

Hence, the asymptotic sampling distribution of $\bar{\boldsymbol{\theta}}_M$ without conditioning on the selection event is $N(\boldsymbol{\theta}_M, \mathcal{I}^{-1}(\boldsymbol{\theta}_M)/n)$ if the selected model M contains all truly important covariates and $\lambda / \sqrt{n} \to 0$, while viewing that $\hat{\boldsymbol{\theta}}_M$ results from Equation 6, where we ignore the dependence of $\hat{\mathbf{z}}$ and $\hat{\mathbf{W}}$ on estimates for $\boldsymbol{\theta}$. Under these assumptions, the asymptotic distribution of $\bar{\boldsymbol{\beta}}_M$ is $N(\boldsymbol{\beta}_M, \mathcal{I}_M^{-1}(\boldsymbol{\theta}_M)/n)$, where $\mathcal{I}_M^{-1}(\boldsymbol{\theta}_M)$ is the lower $|M| \times |M|$ block of $\mathcal{I}^{-1}(\boldsymbol{\theta}_M)$.

To see how the selection event $\{\mathbf{A}_1 \bar{\boldsymbol{\beta}}_M \leq \mathbf{b}_1\}$ affects the sampling distribution of $\bar{\boldsymbol{\beta}}_M$, we introduce 2 random quantities that are independent of each other. The first is a linear contrast $\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M$ in $\bar{\boldsymbol{\beta}}_M$, where $\boldsymbol{\xi} \in \mathbb{R}^{|M|}$ is a user-specified vector depending on one's inference target. For example, if one wishes to make inferences on β_j for $j \in M$, one sets $\boldsymbol{\xi}$ as the jth elementary basis vector in $\mathbb{R}^{|M|}$. The second quantity is $\mathbf{r} = (\mathbf{I}_{|M|} - \mathbf{c}\boldsymbol{\xi}^\top)\bar{\boldsymbol{\beta}}_M$, where $\mathbf{c} = \mathcal{I}_M^{-1}(\boldsymbol{\theta}_M)\boldsymbol{\xi}\{\boldsymbol{\xi}^\top \mathcal{I}_M^{-1}(\boldsymbol{\theta}_M)\boldsymbol{\xi}\}^{-1}$. By Lemma S.1 in Lee et al. (2016), we have $\{\mathbf{A}_1 \bar{\boldsymbol{\beta}}_M \leq \mathbf{b}_1\} = \{\nu^-(\mathbf{r}) \leq \boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \leq \nu^+(\mathbf{r}), \ \nu^0(\mathbf{r}) \geq 0\}$, where

$$\nu^-(\mathbf{r}) = \max_{\{j: (\mathbf{A}_1 \mathbf{c})_j < 0\}} \frac{\mathbf{b}_{1j} - (\mathbf{A}_1 \mathbf{r})_j}{(\mathbf{A}_1 \mathbf{c})_j},$$

$$\nu^+(\mathbf{r}) = \max_{\{j: (\mathbf{A}_1 \mathbf{c})_j > 0\}} \frac{\mathbf{b}_{1j} - (\mathbf{A}_1 \mathbf{r})_j}{(\mathbf{A}_1 \mathbf{c})_j}, \ \text{and}$$

$$\nu^0(\mathbf{r}) = \max_{\{j: (\mathbf{A}_1 \mathbf{c})_j = 0\}} \{\mathbf{b}_{1j} - (\mathbf{A}_1 \mathbf{r})_j\}.$$

Here, for a generic vector $\mathbf{t} \in \mathbb{R}^{|M|}$, $\mathbf{t}_j$ or $(\mathbf{t})_j$ denotes its jth entry. Now, using "$\overset{d}{=}$" to refer to "follows the same distribution as," we have

$$
\begin{aligned}
&\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \mid \{\mathbf{A}_1 \bar{\boldsymbol{\beta}}_M \leq \mathbf{b}_1\} \\
&\overset{d}{=} \boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \mid \{\nu^-(\mathbf{r}) \leq \boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \leq \nu^+(\mathbf{r}), \nu^0(\mathbf{r}) \geq 0\} \\
&\overset{d}{=} \boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \mid \{\nu^-(\mathbf{r}) \leq \boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \leq \nu^+(\mathbf{r})\},
\end{aligned}
$$

where the second equality owes to the independence between $\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M$ and $\mathbf{r}$. From here we see that the selection event corresponds to a truncation of the sampling distribution of our contrast $\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M$. Specifically, the conditional asymptotic distribution of $\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M$ given the selection event is the truncated Gaussian distribution $TN(\boldsymbol{\xi}^\top \boldsymbol{\beta}_M, \boldsymbol{\xi}^\top \mathcal{I}_M^{-1}(\boldsymbol{\theta}_M)\boldsymbol{\xi}/n, \nu^-(\mathbf{r}), \nu^+(\mathbf{r}))$, where $TN(\mu, \sigma^2, a, b)$ denotes the $N(\mu, \sigma^2)$ with support truncated to $[a, b]$.

3.4 Post-selection inference for linear contrasts of covariate effects

Denote by $F_{\mu,\sigma^2}^{[a,b]}$ the cumulative distribution function of $TN(\mu, \sigma^2, a, b)$. Then

$$
F_{\boldsymbol{\xi}^\top \boldsymbol{\beta}_M, \boldsymbol{\xi}^\top \mathcal{I}_M^{-1}(\boldsymbol{\theta}_M)\boldsymbol{\xi}/n}^{[\nu^-(\mathbf{r}), \nu^+(\mathbf{r})]} \left(\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \right) \Big| \{\mathbf{A}_1 \bar{\boldsymbol{\beta}}_M \leq \mathbf{b}_1\} \sim \mathrm{Unif}(0, 1)
$$

asymptotically, which reveals an asymptotic pivotal quantity for $\boldsymbol{\xi}^\top \boldsymbol{\beta}_M$. An asymptotic $(1 - a)100\%$ CI for $\boldsymbol{\xi}^\top \boldsymbol{\beta}_M$ is thus given by

$$
\left\{ \mu : a/2 \leq F_{\mu, \boldsymbol{\xi}^\top \hat{\mathcal{I}}^{-1}\boldsymbol{\xi}}^{[\nu^-(\hat{\mathbf{r}}), \nu^+(\hat{\mathbf{r}})]} \left(\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M \right) \leq 1 - a/2 \right\}, \quad (12)
$$

where $\hat{\mathbf{r}} = (\mathbf{I}_{|M|} - \hat{\mathbf{c}}\boldsymbol{\xi}^\top)\bar{\boldsymbol{\beta}}_M$, $\hat{\mathbf{c}} = \hat{\mathcal{I}}^{-1}\boldsymbol{\xi}(\boldsymbol{\xi}^\top \hat{\mathcal{I}}^{-1}\boldsymbol{\xi})^{-1}$, and $\hat{\mathcal{I}}$ is an estimate of $n\mathcal{I}_M(\boldsymbol{\theta}_M)$.

Following Louis' method (Louis, 1982), we construct $\hat{\mathcal{I}}$ based on the observed total information after the last iteration of the EM algorithm that is given by

$$
\begin{aligned}
\mathcal{I}_{\mathrm{obs}}(\boldsymbol{\theta}_M; \mathbf{Z}) = {}&\mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\mathcal{I}_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})|\mathbf{Z}] \\
&- \mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\mathcal{S}_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})\mathcal{S}_c^\top(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})|\mathbf{Z}],
\end{aligned}
$$
$$(13)$$

where $\mathcal{S}_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})$ and $\mathcal{I}_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})$ are the complete data score and observed information matrix, respectively, evaluated at $\boldsymbol{\theta}_M$. Recall that $\hat{\boldsymbol{\theta}}_M$ is the value of the estimator of $\boldsymbol{\theta}_M$ at convergence of the EM algorithm. From the complete data log-likelihood in Equation 1, we have

$$
\begin{aligned}
\mathcal{S}_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z}) &= \frac{\partial \ell_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})}{\partial \boldsymbol{\theta}_M^\top} \\
&= \sum_{i=1}^n \{Y_i - \pi_i(\boldsymbol{\theta}_M)\} \begin{bmatrix} 1 \\ \mathbf{x}_{M,i} \end{bmatrix}, \quad (14)
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z}) &= -\frac{\partial^2 \ell_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})}{\partial \boldsymbol{\theta}_M \partial \boldsymbol{\theta}_M^\top} \\
&= \sum_{i=1}^n \pi_i(\boldsymbol{\theta}_M)\{1 - \pi_i(\boldsymbol{\theta}_M)\} \begin{bmatrix} 1 & \mathbf{x}_{M,i}^\top \\ \mathbf{x}_{M,i} & \mathbf{x}_{M,i}\mathbf{x}_{M,i}^\top \end{bmatrix},
\end{aligned}
$$
$$(15)$$

where $\mathbf{x}_{M,i}^\top$ gives the ith row of $\mathbf{X}_M$. Since Equation 15 does not depend on $\mathbf{Y}$, the first expectation in Equation 13 is equal to Equation 15. To derive the second expectation in Equation 13, let $\mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y}) = \{\mathbf{Y} - \boldsymbol{\pi}(\boldsymbol{\theta}_M)\}\{\mathbf{Y} - \boldsymbol{\pi}(\boldsymbol{\theta}_M)\}^\top$, where $\boldsymbol{\pi}(\boldsymbol{\theta}_M) = (\pi_1(\boldsymbol{\theta}_M), \ldots, \pi_n(\boldsymbol{\theta}_M))^\top$. Then Equation 14 implies that

$$
\begin{aligned}
&\mathcal{S}_c(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z})\mathcal{S}_c^\top(\boldsymbol{\theta}_M; \mathbf{Y}, \mathbf{Z}) \\
&= \begin{bmatrix} \mathbf{1}_n^\top \mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y})\mathbf{1}_n & \mathbf{1}_n^\top \mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y})\mathbf{X}_M \\ \mathbf{X}_M^\top \mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y})\mathbf{1}_n & \mathbf{X}_M^\top \mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y})\mathbf{X}_M \end{bmatrix}.
\end{aligned}
$$

Hence, the second term in Equation 13 depends on $\mathbf{Z}$ only via $\mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y})|\mathbf{Z}]$, of which the $[i, k]$ entry is, for $i \neq k \in \{1, \ldots, n\}$,

$$
\begin{aligned}
&\mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\{Y_i - \pi_i(\boldsymbol{\theta}_M)\}\{Y_k - \pi_k(\boldsymbol{\theta}_M)\}|\mathbf{Z}] \\
&= \{\hat{Y}_i - \pi_i(\boldsymbol{\theta}_M)\}\{\hat{Y}_k - \pi_k(\boldsymbol{\theta}_M)\}, \quad (16)
\end{aligned}
$$

and, for $i = k \in \{1, \ldots, n\}$,

$$
\mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\{Y_i - \pi_i(\boldsymbol{\theta}_M)\}^2|\mathbf{Z}] = \{1 - 2\pi_i(\boldsymbol{\theta}_M)\}\hat{Y}_i + \pi_i^2(\boldsymbol{\theta}_M),
$$
$$(17)$$

where $\hat{Y}_i = \mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[Y_i|Z_i]$ is given by Equation 3 (with $\boldsymbol{\theta}^{(k)}$ there set at $\hat{\boldsymbol{\theta}}_M$), for $i = 1, \ldots, n$. We then use the lower $|M| \times |M|$ block of $\mathcal{I}_{\mathrm{obs}}(\hat{\boldsymbol{\theta}}_M; \mathbf{Z})$ as an estimate of $n\mathcal{I}_M(\boldsymbol{\theta}_M)$ in Equation 12, that is,

$$
\begin{aligned}
\hat{\mathcal{I}} = {}&\sum_{i=1}^n \pi_i(\hat{\boldsymbol{\theta}}_M)\{1 - \pi_i(\hat{\boldsymbol{\theta}}_M)\}\mathbf{x}_{M,i}\mathbf{x}_{M,i}^\top \\
&- \mathbf{X}_M^\top \mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\mathcal{D}(\hat{\boldsymbol{\theta}}_M; \mathbf{Y})|\mathbf{Z}]\mathbf{X}_M. \quad (18)
\end{aligned}
$$

4 ADAPTATION FOR GROUP TESTING DATA

We now adapt the strategy of Section 3 to group testing data. Suppose n individuals are randomly partitioned into J pools, and let $\mathcal{P}_1, \ldots, \mathcal{P}_J$ be the partition of $\{1, \ldots, n\}$ such that $\mathcal{P}_j$ is the set of indices of the individuals in pool j. Set $Y_j^* = \max_{i \in \mathcal{P}_j} Y_i$, which is the true disease status for pool j (equal to 1 if at least one individual is positive), and let Z_j be the observed testing result for the pool, where the test has sensitivity Se and specificity Sp.

In this setting, the complete data log-likelihood will still be of the form in Equation 1, so Algorithm 1 can still be used to compute the LASSO-penalized MLE: The only change is in how we compute the conditional expectations of the unobserved true disease statuses $Y_1, \ldots, Y_n$ given the observed testing outcomes, which are now the pooled testing outcomes $Z_1, \ldots, Z_J$. As long as we can compute these conditional expectations, we can compute $\bar{\boldsymbol{\theta}}_M$ and follow the arguments in Section 3 for making valid post-selection inferences on contrasts in $\bar{\boldsymbol{\beta}}_M$.

Under group testing, for $i \in \mathcal{P}_j$, we have

$$\mathbb{E}_{\boldsymbol{\theta}}[Y_i | Z_j] = \frac{\mathrm{Se}^{Z_j}(1 - \mathrm{Se})^{1-Z_j} \pi_i(\boldsymbol{\theta})}{\{P_{\boldsymbol{\theta}}(Z_j = 1)\}^{Z_j} \{P_{\boldsymbol{\theta}}(Z_j = 0)\}^{1-Z_j}}, \quad (19)$$

where $P_{\boldsymbol{\theta}}(Z_j = 0) = (1 - \mathrm{Se})[1 - \prod_{i \in \mathcal{P}_j} \{1 - \pi_i(\boldsymbol{\theta})\}] + \mathrm{Sp} \prod_{i \in \mathcal{P}_j} \{1 - \pi_i(\boldsymbol{\theta})\}$, and $P_{\boldsymbol{\theta}}(Z_j = 1) = 1 - P_{\boldsymbol{\theta}}(Z_j = 0)$. Now, the LASSO-penalized MLE is computed with Algorithm 1 as in the individual testing case, but with $\hat{Y}_i^{(k)}$ on line 2 computed according to the conditional expectation in Equation 19. The output of the EM algorithm gives the penalized MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, which reveals a selected model M, under which we obtain the estimated responses in $\hat{\mathbf{Y}}$, $\hat{Y}_i = \mathbb{E}_{\hat{\boldsymbol{\theta}}}[Y_i | Z_j]$, for $i \in \mathcal{P}_j$ and $j = 1, \ldots, J$, according to Equation 19. Using $\hat{\mathbf{Y}}$ and $\hat{\boldsymbol{\theta}}$, we define the post-selection estimator $\bar{\boldsymbol{\theta}}_M$ as in Equation 5.

Having obtained $\bar{\boldsymbol{\theta}}_M$, we follow the steps in Section 3 until we come to the estimation of the observed total information $n \mathcal{I}_M(\boldsymbol{\theta}_M)$. The identity in Equation 13 on which Louis' method is based still holds, only now $\mathbf{Z}$ represents the group testing outcomes. As a result, the estimated total information associated with $\bar{\boldsymbol{\beta}}_M$ is still given by Equation 18, but with the entries in $\mathbb{E}_{\bar{\boldsymbol{\theta}}_M}[\mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y}) | \mathbf{Z}]$ derived differently, as elaborated in Web Appendix A. The $[i, k]$ entry of $\mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\mathcal{D}(\boldsymbol{\theta}_M; \mathbf{Y}) | \mathbf{Z}]$ is now found as follows: For $i, k \in \mathcal{P}_j$ with $i \neq k$, we have

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[\{Y_i - \pi_i(\boldsymbol{\theta}_M)\}\{Y_k - \pi_k(\boldsymbol{\theta}_M)\} | \mathbf{Z}]$$
$$= \frac{\mathrm{Se}^{Z_j}(1 - \mathrm{Se})^{1-Z_j} \pi_i(\hat{\boldsymbol{\theta}}_M) \pi_k(\hat{\boldsymbol{\theta}}_M)}{\{P_{\hat{\boldsymbol{\theta}}_M}(Z_j = 1)\}^{Z_j} \{P_{\hat{\boldsymbol{\theta}}_M}(Z_j = 0)\}^{1-Z_j}}$$
$$- \pi_k(\boldsymbol{\theta}_M)\hat{Y}_i - \pi_i(\boldsymbol{\theta}_M)\hat{Y}_k + \pi_i(\boldsymbol{\theta}_M)\pi_k(\boldsymbol{\theta}_M),$$

where $\hat{Y}_i = \mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[Y_i | Z_j]$ and $\hat{Y}_k = \mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[Y_k | Z_j]$ are given by Equation 19 with $\hat{\boldsymbol{\theta}}_M$ replacing $\boldsymbol{\theta}$; for $i \in \mathcal{P}_j$ and $k \in \mathcal{P}_m$, $j \neq m$, the entry is still given by Equation 16, but with $\hat{Y}_i = \mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[Y_i | Z_j]$ and $\hat{Y}_k = \mathbb{E}_{\hat{\boldsymbol{\theta}}_M}[Y_k | Z_m]$; if $i = k \in \mathcal{P}_j$, the entry is also given by Equation 17, but with the estimated response $\hat{Y}_i$ obtained based on the group response Z_j according to Equation 19.

If M is an incorrect model, that is, M does not contain all truly active covariates, then we view $\boldsymbol{\theta}_M$ as the "truth" defined by an equation similar to Equation (1.2) in Lee et al. (2016). Then all discussions in Sections 3 and 4 still go through, but, in the presence of model misspecification, one would use the sandwich-form of variance–covariance matrix estimation in Elashoff and Ryan (2004) to estimate the variance–covariance instead of Louis' method. This estimator, which is derived in Web Appendix B, gives some protection against model misspecification in variance estimation.

Our method can be easily extended to accommodate more complicated group testing schemes, for example, when individuals in positive pools are retested. While more complicated testing schemes induce more complicated likelihood functions, one can still implement the EM algorithm described in Algorithm 1 and use Louis' method for evaluating the observed information matrix: One needs only to be able to compute conditional expectations of the true, unobserved disease statuses $Y_1, \ldots, Y_n$ given all the observed testing outcomes. Suppose $\mathcal{A}_i$ represents all the testing outcomes that play a role in the conditional probability

that $Y_i = 1$. For example, if groups of individuals are tested in pools and then individuals in positive pools are re-tested, $\mathcal{A}_i$ will consist of a single negative test if the pool in which individual i was placed tests negative; on the other hand, if the pool in which individual i was placed tests positive, $\mathcal{A}_i$ will consist of the positive pool test as well as the re-testing outcomes for all individuals in the pool. To implement our methods in this setting, we can replace the condition expectations given in Equation 3 or 19 by $\mathbb{E}_{\boldsymbol{\theta}}[Y_i | \mathcal{A}_i]$. See, for example, Gregory et al. (2019) for details on how to compute the conditional expectations $\mathbb{E}_{\boldsymbol{\theta}}[Y_i | \mathcal{A}_i]$ for some more complicated testing schemes.

5 EMPIRICAL EVIDENCE

We shall acknowledge that several approximations are made in developing the proposed post-selection inference procedure in Sections 3 and 4. First, in Equation 6, we view $\dot{\mathbf{z}}$ as response data and $\hat{\mathbf{W}}$ as the weight matrix even though they depend on an estimate of $\boldsymbol{\theta}$. The uncertainty in this estimate when formulating the selection event is ignored, for instance, by viewing $\mathbf{b}_1$ in Equation 11 as fixed given a selected model even though it depends on $\hat{\boldsymbol{\theta}}_M$. Second, we approximate the pre-selection sampling distribution of $\bar{\boldsymbol{\theta}}_M$ by a Gaussian distribution. Effects of these approximations permeate through the post-selection distribution of $\boldsymbol{\xi}^\top \bar{\boldsymbol{\beta}}_M$ and the CI in Equation 12. Despite being only approximately valid post-selection inference procedures, they are expected to improve over their naive counterparts, which assume the selection event to be known and fixed when making inferences on $\boldsymbol{\beta}_M$. This claim is supported by empirical evidence from our extensive simulation study.

5.1 Simulation design

In our simulation study, we generate true individual responses $\{Y_i\}_{i=1}^n$ from a logistic regression model with $p = 10$ covariates and regression coefficients $\boldsymbol{\theta} = (-5, 2, 1, 1, \mathbf{0}^\top)^\top$, that is, $\alpha = -5$ and only the first 3 covariates are truly important. Realizations of each of the 10 covariates are generated independently from the $N(0, 1)$ distribution. We then randomly form J pools such that $|\mathcal{P}_j| = m$ for all j for some pool size m, and compute true pool responses $\{Y_j^* = \max_{i \in \mathcal{P}_j} Y_i\}_{j=1}^J$. Lastly, we simulate the imprecise group testing response Z_j from Bernoulli(Se $\cdot Y_j^* + (1 - \mathrm{Sp})(1 - Y_j^*)$), for $j = 1, \ldots, J$, with Se = 0.95 and Sp = 0.97. We consider the sample sizes $n \in \{1000, 2000\}$ and the pool sizes $m \in \{1, 2, 4\}$, where $m = 1$ leads to $\{Z_j\}_{j=1}^J$ as imprecise individual testing data. A total of 5000 Monte Carlo replicates are generated at each combination of n and m.

We set the penalty parameter λ at a sequence of values ranging from 1 to 7 (when $n = 1000$) or 0.5 to 10 (when $n = 2000$) that are equally spaced on the logarithmic scale. Using each Monte Carlo replicate dataset, we fit the logistic regression model following Algorithm 1 at each pre-specified λ; then we carry out the proposed post-selection inference to construct 95% CIs for each entry of $\boldsymbol{\beta}_M$, where M is the selected model. We also obtain 95% CIs naively in the classical way assuming the selected model to be known a priori.

For each simulated dataset, we compute a realization of the Type I error rate in the following way: After model selection, we

build a CI for each entry of $\boldsymbol{\beta}_M$, where M is the selected model. If the selected model includes covariates for which the regression coefficient is equal to 0, then for each of these, we check whether the CI for the regression coefficient contains 0. The realized Type I error rate is the proportion of these CIs that did not contain 0. If the selected model M does not include any "extra" covariates (covariates with regression coefficient equal to 0), then the realized Type I error rate is recorded as 0 for that dataset. Specifically, we compute on each simulated dataset the quantity

$$\text{Type I error rate} = \frac{|\{j \in M : 0 \notin \text{CI}(\hat{\beta}_j) \text{ and } \beta_j = 0\}|}{|\{j \in M : \beta_j = 0\}|},$$

where $\text{CI}(\hat{\beta}_j)$ is a 95% CI for β_j according to the considered inference procedure, and we record 0 if $|\{j \in M : \beta_j = 0\}| = 0$. In addition, we record for each data set, at each preset value of λ, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) associated with the selected model, that is, based on the observed data log-likelihood evaluated at $\hat{\boldsymbol{\theta}}_M$.

5.2 Simulation results

Figure 1 depicts Monte Carlo averages of the realized Type I error rate associated with the naive and selection-adjusted inference procedures as λ increases. In all 6 simulation settings, our proposed method produces Type I error rates closely matching the nominal level 0.05 across all pre-set values of λ, whereas the naive method yields an increasingly inflated Type I error as λ increases. A larger penalty λ makes the regularization more aggressive and typically results in a smaller model size $|M|$. Ignoring the uncertainty in variable selection when inferring selected covariate effects does more harm when 1 more proactively penalizes large models, and having group testing data with a larger group size only exacerbates the problem. In contrast, our method adequately accounts for variable selection and provides more reliable inference on selected covariate effects based on either individual testing data or group testing data with arbitrary group sizes.

If one first uses AIC or BIC to choose a value of λ and then selects a model using the chosen λ, the plots in Figure 1 suggest that the naive inference will be liable to above-nominal Type I error rates; this problem could be worse under BIC selection than under AIC selection, as the BIC tends to select smaller models (larger values of λ). In contrast, the proposed post-selection inference procedure exhibits good control of the Type I error rate over the entire range of λ values typically chosen by the AIC or BIC.
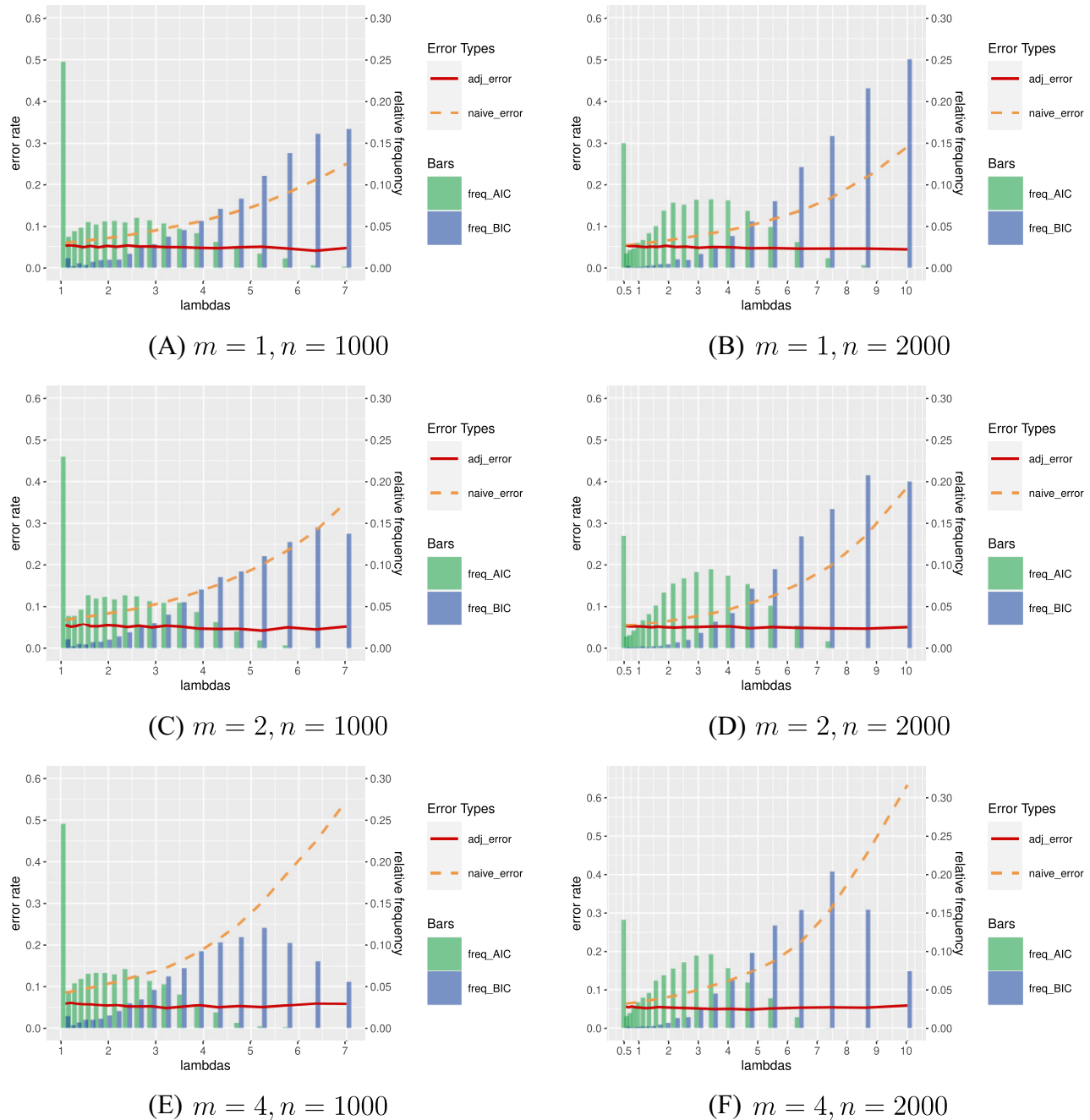
Table 1 summarizes the performance of 95% CIs for β_2, β_4, and β_6 resulting from the proposed method and from the naive method along with those for the corresponding odds ratios $\exp(\beta_j)$, $j = 2, 4, 6$, where the odds ratio CIs are constructed by exponentiating the endpoints of the coefficient CIs. Note that β_2 is nonzero, whereas β_4 and β_6 are both equal to 0 in the true model. We report for each coefficient the average of the upper as well as the lower CI bound, the average widths of the CI, and the proportions of the CIs that contain the true value of the coefficient—when the corresponding covariate is selected—from 5000 simulated datasets. Results are reported at a single, fixed value of λ chosen as follows: For each of the 5000 data

sets, we recorded the value of λ that minimized the AIC criterion across a grid of λ choices (where the same grid of choices was considered for each dataset). Then we set our fixed λ value equal to the median of these values. This was done at each combination of n and m. Table 1 shows that the naive CIs are on average much narrower than our selection-adjusted CIs constructed according to Equation 12. This explains the inflated Type I error of the naive method exhibited in Figure 1. For β_4 and β_6, which are equal to 0, the coverage probabilities of the naive CIs are considerably lower than the nominal level, while those of the adjusted CIs are close to the nominal level. For β_2, which is nonzero, we see that the coverage probability of the proposed method falls a little below the nominal level; this owes to the shrinkage toward 0 of the 1-step estimators of the covariate effects, which is not unexpected. For coefficients equal to 0, the shrinkage toward 0 of the 1-step estimator does not adversely affect the coverage of the selection-adjusted CIs, since the shrinkage makes the 1-step estimator more accurate for 0-valued coefficients.

A practical consideration is how our method for post-selected adjustment will perform under misspecification of the sensitivity and specificity of the tests. Table 2 and Figure 2 summarize the performance of naive and selection-adjusted inference when the true sensitivity and specificity of the tests are Se = 0.90 and Sp = 0.92, but when the higher values Se = 0.95 and Sp = 0.97 are assumed by the estimation procedure. This corresponds to the situation in which we assume the tests are more accurate than they really are. According to Figure 2, which focuses on Type I error rates, this kind of misspecification does not appear to degrade the performance of our selection-adjusted inference, as the patterns we see are similar to those in Figure 1. From Table 2, however, we see that for β_2, which is nonzero, the coverage probabilities of both the naive and the selection-adjusted CIs are nowhere close to the nominal level. This shows that misspecifying the sensitivity and specificity of the tests can severely distort both naive and selection-adjusted inferences for nonzero coefficients; this is not very surprising, as this kind of misspecification bases estimation on an incorrect likelihood function. Somewhat surprising, however, is that Type I error rates (inference on coefficients that are truly equal to 0) appear to be largely unaffected by the misspecification.

5.3 Application to chlamydia data

We now analyze data from the National Health and Nutrition Examination Survey from year 2015 to 2016 on $n = 1371$ participants aged 14-39 years. Besides individual testing results for chlamydia infection based on urine specimens, participants' demographic information was also collected, such as ethnicity, income level, and age. For illustration purposes, we consider a logistic regression model for the indicator of chlamydia infection, with covariates age, non-Hispanic White, non-Hispanic Black, Mexican American and Hispanic, family monthly poverty level index, new partner in the past 12 months, an indicator of poverty index > 5, gender, education level, and marriage status (married or never married). Variable selection was performed in order to identify the most significant risk factors, thereby improving the model performance and interpretability in assessing the likelihood of chlamydia infection. A sensitivity of 0.97 and a specificity of 0.99 are assumed for the test for the diagnosis of

(A) $m = 1, n = 1000$

(B) $m = 1, n = 2000$

(C) $m = 2, n = 1000$

(D) $m = 2, n = 2000$

(E) $m = 4, n = 1000$

(F) $m = 4, n = 2000$

FIGURE 1 Averages of Type I error rates across 5000 Monte Carlo replicates versus λ when the proposed post-selection inference procedure is used to construct 95% confidence intervals for selected covariate effects (solid lines), and when the naive inference procedure is used (dashed lines) as the sample size n and pool size m vary. Overlaying the plot of Type I error rate versus λ are the histogram of chosen λ according to Akaike information criterion and the counterpart histogram when Bayesian information criterion is used to choose λ.

chlamydia, which we take to be reasonable values according to Whellams (2021).

We first use the individual testing data to fit the logistic regression model subject to the LASSO penalty, with λ chosen by AIC, followed by constructing 95% CIs for covariate effects in the selected model. As in the simulation study, we compare CIs from the naive method to those from our proposed method. To introduce another competitor, we present results from the strategy of data splitting, using half the data for variable selection and half

for interval estimation. Web Appendix C contains results from this strategy implemented in the simulation setting described in Section 5.1. Table 3 (the upper half) presents these interval estimates for a selected model resulting from these 3 approaches. Figure 3 provides a pictorial comparison of these CIs.

Finally, we randomly partition the 1371 individuals into $J = 457$ groups of size $m = 3$ and artificially generate group testing responses. We repeat the exercise of model selection followed by interval estimation using the aforementioned 3 strategies. The

TABLE 1 Monte Carlo averages of empirical coverage probabilities (C.P.) of 95% confidence intervals (CIs) for β_2, β_4, and β_6, along with the corresponding average CI limits, average CI for odds ratio, average widths (A.W.), and average parameter estimates (P.E.).

n	Param	m	Proposed					Naive				
			CI	CI odds	A.W.	C.P.	P.E.	CI	CI odds	A.W.	C.P.	P.E.
		1	[0.28,1.82]	[1.32,6.17]	1.5	93.1	0.93	[0.53,1.68]	[1.70,5.37]	1.2	95.6	0.83
	β_2	2	[0.19,1.93]	[1.21,6.89]	1.7	92.4	0.93	[0.51,1.70]	[1.67,5.47]	1.2	94.7	0.83
		4	[0.02,2.16]	[1.02,8.67]	2.1	93.9	0.93	[0.47,1.81]	[1.60,6.11]	1.3	95.3	0.84
		1	[−0.89,1.57]	[0.41,4.81]	2.5	94.9	0.00	[−0.46,0.45]	[0.63,1.57]	0.9	92.8	0.00
1000	β_4	2	[−1.02,1.81]	[0.36,6.11]	2.8	94.1	0.00	[−0.50,0.49]	[0.61,1.63]	1.0	91.2	0.00
		4	[−1.19,2.33]	[0.30,10.29]	3.5	94.2	0.00	[−0.56,0.58]	[0.57,1.79]	1.1	90.1	0.00
		1	[−0.88,1.56]	[0.41,4.75]	2.4	95.0	−0.01	[−0.47,0.45]	[0.62,1.57]	0.9	92.5	0.00
	β_6	2	[−0.97,1.91]	[0.38,6.75]	2.9	94.2	−0.01	[−0.50,0.48]	[0.61,1.62]	1.0	91.6	0.00
		4	[−1.19,2.26]	[0.30,9.59]	3.5	94.0	0.00	[−0.57,0.57]	[0.57,1.77]	1.1	89.8	0.00
		1	[0.53,1.53]	[1.70,4.62]	1.0	92.9	0.96	[0.67,1.41]	[1.96,4.10]	0.7	94.9	0.90
	β_2	2	[0.48,1.57]	[1.62,4.81]	1.1	93.3	0.95	[0.66,1.43]	[1.93,4.18]	0.8	94.9	0.88
		4	[0.37,1.71]	[1.45,5.53]	1.3	93.9	0.95	[0.63,1.49]	[1.88,4.44]	0.9	94.6	0.89
		1	[−0.60,1.06]	[0.55,2.89]	1.7	94.8	0.00	[−0.30,0.31]	[0.74,1.36]	0.6	92.8	0.00
2000	β_4	2	[−0.67,1.21]	[0.51,3.36]	1.9	95.5	0.00	[−0.32,0.33]	[0.72,1.39]	0.7	92.4	0.00
		4	[−0.79,1.48]	[0.45,4.39]	2.3	94.3	0.00	[−0.37,0.38]	[0.69,1.46]	0.8	91.3	0.00
		1	[−0.60,1.08]	[0.55,2.94]	1.7	94.7	0.00	[−0.31,0.30]	[0.73,1.35]	0.6	92.6	0.00
	β_6	2	[−0.66,1.22]	[0.52,3.39]	1.9	95.2	0.00	[−0.33,0.32]	[0.72,1.38]	0.7	92.2	0.00
		4	[−0.80,1.42]	[0.45,4.13]	2.2	94.6	0.00	[−0.37,0.38]	[0.69,1.46]	0.8	91.1	0.00

The results are based on a single value of λ (the median Akaike information criterion choice of λ) across 5000 Monte Carlo simulations at each combination of n and m. Se = 0.95, Sp = 0.97.

TABLE 2 Monte Carlo averages of empirical coverage probabilities (C.P.) of 95% confidence intervals (CIs) for β_2, β_4, and β_6, along with the corresponding average CI limits, average CI for odds ratio, average widths (A.W.), and average parameter estimates (P.E.).

n	Param	m	Proposed					Naive				
			CI	CI odds	A.W.	C.P.	P.E.	CI	CI odds	A.W.	C.P.	P.E.
		1	[−0.01,0.86]	[0.99,2.36]	0.9	13.8	0.40	[0.15,0.75]	[1.16,2.12]	0.6	10.5	0.34
	β_2	2	[−0.06,1.36]	[0.94,3.90]	1.4	53.4	0.57	[0.25,1.09]	[1.28,2.97]	0.8	60.7	0.51
		4	[−0.22,1.90]	[0.80,6.69]	2.1	84.0	0.72	[0.31,1.44]	[1.36,4.22]	1.1	90.0	0.65
		1	[−0.57,1.01]	[0.57,2.75]	1.6	95.3	0.00	[−0.28,0.28]	[0.75,1.32]	0.6	92.8	0.00
1000	β_4	2	[−0.82,1.52]	[0.44,4.58]	2.3	94.9	0.00	[−0.39,0.38]	[0.68,1.46]	0.8	90.8	0.00
		4	[−1.15,2.05]	[0.32,7.77]	3.2	93.9	0.00	[−0.51,0.52]	[0.60,1.68]	1.0	89.3	0.00
		1	[−0.57,1.00]	[0.57,2.72]	1.6	95.7	0.00	[−0.28,0.27]	[0.75,1.31]	0.5	92.3	0.00
	β_6	2	[−0.83,1.50]	[0.44,4.48]	2.3	94.6	0.00	[−0.38,0.39]	[0.68,1.48]	0.8	90.6	0.00
		4	[−1.12,2.17]	[0.33,8.76]	3.3	93.6	−0.01	[−0.53,0.50]	[0.59,1.65]	1.0	89.0	−0.01

The results are based on a single value of λ (the median Akaike information criterion choice of λ) across 5000 Monte Carlo simulations at each combination of n and m. Misspecified Se = 0.90, Sp = 0.92 when generating data.

resulting 95% CIs are provided in Table 3 (the lower half), with a visual contrast given in Figure 4.

From Table 3, we see that the selection-adjusted method tends to be most conservative when making claims regarding statistical significance of a covariate effect, which is reflected in the count of CIs that exclude 0, as well as in the width of a CI clearly depicted in Figures 3 and 4. More specifically, when $m = 1$, the variable non-Hispanic Black is deemed significant by all 3 methods. The variable new sex partner is deemed significant by the naive method and data splitting method, but not according to the selected-adjusted method. Under $m = 3$, data splitting and the selection-adjusted method agree with the naive method on the significance of the variable never married, with the widest CI resulting from the selection-adjusted method. The naive method finds education level to be significant, but neither the proposed selection-adjusted method nor data splitting finds it significant. While data splitting is a sound method for obtaining CIs and

P-values that will maintain nominal coverage probabilities and Type I error rates, it is important to note that in the case of highly unbalanced binary response data (when the responses are mostly zeros with a small number of ones or vice versa), the set of selected covariates can vary wildly with the subset of the data chosen to perform variable selection. That is, different random splits of the data can result in selected models that are very different. In fact, in order to compare our method and the naive method to the data splitting method, we needed to try many different splits of the data to find a split under which the selected model was the same as the selected model from the full data set. These considerations pose considerable drawbacks to the data splitting method, in particular when one wants reproducible results. Lastly, Table 4 presents the odds ratio CIs and point estimates for the selected variables. This table is provided for readers who are interested in these additional details.

(A) $m = 1, n = 1000$ (B) $m = 2, n = 1000$
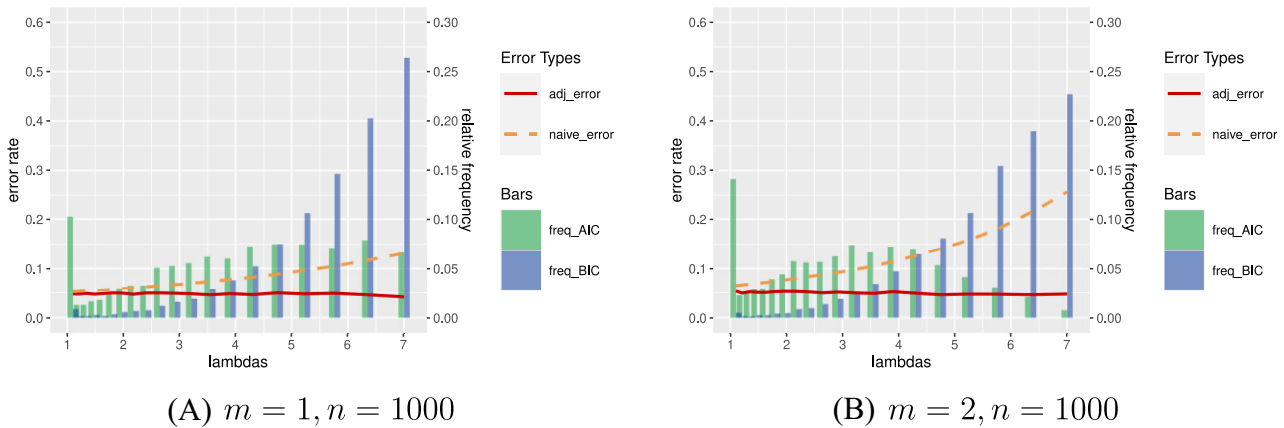
FIGURE 2 Averages of Type I error rates across 5000 Monte Carlo replicates versus λ when the proposed post-selection inference procedure is used to construct 95% confidence intervals for selected covariate effects (solid lines), and when the naive inference procedure is used (dashed lines) as the sample size n and pool size m vary, with the sensitivity and specificity of tests misspecified. Overlaying the plot of Type I error rate versus λ are the histogram of chosen λ according to Akaike information criterion and the counterpart histogram when Bayesian information criterion is used to choose λ.

TABLE 3 Confidence intervals (CIs) and point estimates for selected covariate effects from 3 considered methods based on the chlamydia data. CIs not containing zero are marked with '∗'.

m	Covariate	Proposed		Naive		Data splitting	
		CI	PE	CI	PE	CI	PE
1	Age	$[-2.62, 1.36]$	-1.22	$[-2.84, 0.17]$	-1.34	$[-2.91, 0.60]$	-1.16
	Non-Hispanic Black	$[0.31, 2.79]*$	1.72	$[0.61, 3.41]*$	2.01	$[0.41, 3.17]*$	1.79
	New sex partner	$[-0.16, 1.77]$	0.92	$[0.07, 1.82]*$	0.95	$[0.06, 2.07]*$	1.06
	Hispanic	$[-1.16, 2.23]$	1.07	$[-0.03, 2.84]$	1.41	$[-0.63, 2.38]$	0.87
	Poverty index > 5	$[-4.17, 2.53]$	1.06	$[0, 2.58]$	1.29	$[-0.89, 2.55]$	0.83
	Gender	$[-0.70, 1.60]$	0.72	$[-0.15, 1.75]$	0.80	$[-0.37, 1.77]$	0.70
	Education level	$[-1.68, 0.39]$	-0.82	$[-1.70, 0.09]$	-0.80	$[-1.90, 0.11]$	-0.90
3	Non-Hispanic White	$[-1.41, 9.91]$	-0.31	$[-2.55, 1.21]$	-0.67	$[-1.57, 0.24]$	1.05
	New sex partner	$[-5.06, 1.48]$	0.36	$[-0.77, 1.64]$	0.43	$[-1.04, 1.91]$	1.01
	Gender	$[-1.44, 4.03]$	-0.42	$[-1.79, 0.48]$	-0.65	$[-1.87, 0.56]$	-2.04
	Education level	$[-2.70, 0.05]$	-1.37	$[-2.73, -0.25]*$	-1.49	$[-3.03, 0.05]$	-0.37
	Never married	$[0.07, 4.16]*$	1.50	$[0.19, 2.79]*$	1.49	$[0.53, 2.45]*$	0.83

Abbreviation: PE, parameter estimate.
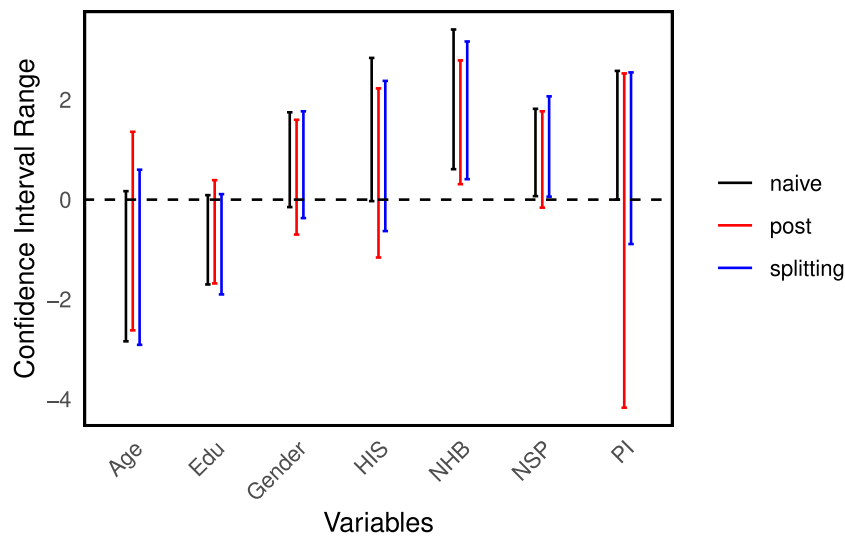
FIGURE 3 Naive, selection-adjusted, and data splitting confidence intervals for the regression coefficients of selected covariates under group size $m = 1$ based on the chlamydia data.
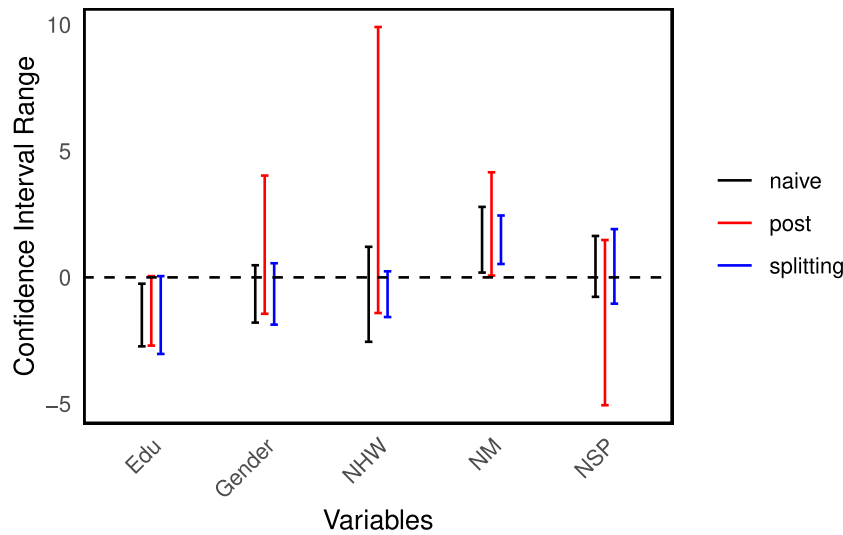
FIGURE 4 Naive, selection-adjusted, and data splitting confidence intervals for the regression coefficients of selected covariates under group size $m = 3$ based on the chlamydia data.

TABLE 4 Odds ratio confidence intervals (CIs) and point estimates for selected covariate effects from 3 considered methods based on the chlamydia data. CIs not containing 1 are marked with '*'.

m	Covariate	Proposed		Naive		Data splitting	
		CI	PE	CI	PE	CI	PE
1	Age	[0.07, 3.90]	0.30	[0.06, 1.18]	0.26	[0.06, 1.82]	0.31
	Non-Hispanic Black	[1.36, 16.33]*	5.59	[1.84, 30.31]*	7.46	[1.51, 23.82]*	5.99
	New sex partner	[0.85, 5.87]	2.51	[1.07, 6.17]*	2.59	[1.06, 7.93]*	2.89
	Hispanic	[0.31, 9.26]	2.92	[0.97, 17.13]	4.10	[0.54, 10.81]	2.39
	Poverty index > 5	[0.02, 12.56]	2.89	[1.00, 13.22]	3.63	[0.41, 12.83]	2.30
	Gender	[0.50, 4.95]	2.05	[0.86, 5.75]	2.23	[0.69, 5.87]	2.01
	Education level	[0.19, 1.48]	0.44	[0.18, 1.09]	0.45	[0.15, 1.12]	0.41
3	Non-Hispanic White	[0.24, 20135.96]	0.73	[0.08, 3.36]	0.51	[0.21, 1.27]	2.86
	New sex partner	[0.01, 4.39]	1.43	[0.46, 5.16]	1.54	[0.35, 6.77]	2.75
	Gender	[0.24, 56.36]	0.66	[0.17, 1.62]	0.52	[0.15, 1.75]	0.13
	Education level	[0.07, 1.05]	0.25	[0.07, 0.78]*	0.23	[0.05, 1.05]	0.69
	Never married	[1.07, 64.01]*	4.48	[1.21, 16.34]*	4.43	[1.70, 11.60]*	2.30

Abbreviation: PE, parameter estimate.

6 DISCUSSION

We develop in this study valid post-selection inference procedures based on binary response data that are partially observed due to testing error and/or pooling. The strategy of relating the penalized estimation and post-selection estimation of regression coefficients to a weighted least squares problem with Gaussian-like response data (ie, $\hat{z}$) was also used in Taylor and Tibshirani (2018) to develop post-selection inference for LASSO-penalized likelihood models. This strategy allows us to draw parallels between the current setting with a partially observed non-Gaussian response and the setting with a linear regression model for a Gaussian response, so that we can borrow established results in the latter setting (Lee et al., 2016; Tibshirani et al., 2016) to derive valid post-selection CIs for target parameters. Although we focus on logistic regression for a binary response here, we believe this strategy can be generalized to other non-Gaussian responses modeled by other generalized linear regression models. The fact that individuals' true response values are unobserv-

able, either due to testing error or pooling, may need to be addressed case-by-case. We use the EM algorithm to address this problem here; other imputation-based methods can be viable alternatives.

Despite the approximations (in formulating response data and sampling distributions) we invoke in the development of our post-selection inference procedure, their impact on the resulting inference is not noticeable in our simulation study, especially when the penalty parameter λ is pre-set at a level such that the LASSO regularization is not too aggressive. In the event of more aggressive penalization, say, by using a penalty chosen by BIC, the proposed method can have the tendency of over-adjusting for model selection, leading to overly conservative CIs that are too wide and produce Type I error much lower than the nominal level. An interesting follow-up research problem is the development of valid post(-double)-selection inference procedures that adequately account for both the selection of λ and model selection given the selected λ.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices referenced in Sections 4 and 5 as well as an R package that includes the data set analyzed in Section 5.3 are available with this paper at the Biometrics website on Oxford Academic. The R package is also available at https://github.com/kateshen28/InfGroupTest.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data that support the findings in this paper are included in the online supplementary materials. The data were drawn from the National Health and Nutrition Examination Survey at https://www.cdc.gov/nchs/nhanes/index.htm.

REFERENCES

Bachoc, F., Leeb, H. and Pötscher, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics*, 47, 1475–1504.

Bachoc, F., Preinerstorfer, D. and Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48, 440–463.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41, 802–837.

Elashoff, M. and Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, 13, 48–65.

Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection, arXiv, arXiv:1410.2597, preprint: not peer reviewed.

Gregory, K. B., Wang, D. and McMahan, C. S. (2019). Adaptive elastic net for group testing. *Biometrics*, 75, 13–23.

Haugland, S., Thune, T., Fosse, B., Wentzel-Larsen, T., Hjelmevoll, S. O. and Myrmel, H. (2010). Comparing urine samples and cervical swabs for chlamydia testing in a female population by means of Strand Displacement Assay (SDA). *BMC Women's Health*, 10, 1–6.

Hyun, S., G'sell, M. and Tibshirani, R. J. (2018). Exact post-selection inference for the generalized LASSO path. *Electronic Journal of Statistics*, 12, 1053–1097.

Joyner, C. N., McMahan, C. S., Tebbs, J. M. and Bilder, C. R. (2020). From mixed effects modeling to spike and slab variable selection: a Bayesian regression model for group testing data. *Biometrics*, 76, 913–923.

Knight, K. and Fu, W. (2000). Asymptotics for LASSO-type estimators. *The Annals of Statistics*, 28, 1356–1378.

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I. and Zhao, L. H. (2020). Valid post-selection inference in model-free linear regression. *The Annals of Statistics*, 48, 2953–2981.

Kuchibhotla, A. K., Kolassa, J. E. and Kuffner, T. A. (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, 9, 505–527.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E. et al. (2016). Exact post-selection inference, with application to the LASSO. *Annals of Statistics*, 44, 907–927.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. New York, NY: Springer.

Lin, J., Wang, D. and Zheng, Q. (2019). Regression analysis and variable selection for two-stage multiple-infection group testing data. *Statistics in Medicine*, 38, 4519–4533.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 226–233.

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

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). *P*-values for high-dimensional regression. *Journal of the American Statistical Association*, 104, 1671–1681.

Neufeld, A. C., Gao, L. L. and Witten, D. M. (2022). Tree-values: selective inference for regression trees. *The Journal of Machine Learning Research*, 23, 13759–13801.

Panigrahi, S. and Taylor, J. (2023). Approximate selective inference via maximum likelihood. *Journal of the American Statistical Association*, 118, 2810–2820.

Panigrahi, S., Taylor, J. and Weinstein, A. (2021). Integrative methods for post-selection inference under convex constraints. *The Annals of Statistics*, 49, 2803–2824.

Pilcher, C. D., Westreich, D. and Hudgens, M. G. (2020). Group testing for severe acute respiratory syndrome—coronavirus 2 to enable rapid scale-up of testing and real-time surveillance of incidence. *The Journal of Infectious Diseases*, 222, 903–909.

Puggioni, G., Gelfand, A. E. and Elmore, J. G. (2008). Joint modeling of sensitivity and specificity. *Statistics in Medicine*, 27, 1745–1761.

Rasines, D. G. and Young, G. A. (2022). Splitting strategies for post-selection inference. *Biometrika*, 110, 597–614.

Rinaldo, A., Wasserman, L. and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Annals of Statistics*, 47, 3438–3469.

Rugamer, D., Baumann, P. F. and Greven, S. (2022). Selective inference for additive and linear mixed models. *Computational Statistics & Data Analysis*, 167, 107350.

Sinnott-Armstrong, N., Klein, D. L. and Hickey, B. (2020). Evaluation of group testing for SARS-CoV-2 RNA. *MedRxiv*, 2020–03.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for penalized likelihood models. *Canadian Journal of Statistics*, 46, 41–61.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics*, 46, 1255–1287.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111, 600–620.

Verdun, C. M., Fuchs, T., Harar, P., Elbrächter, D., Fischer, D. S., Berner, J. et al. (2021). Group testing for SARS-CoV-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies. *Frontiers in Public Health*, 9, 583377.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37, 2178–2201.

Whellams, D. (2021). Screening for chlamydia and gonorrhea in primary care in populations with low prevalence. *Canadian Medical Association Journal*, 193, E1307–E1307.

Xie, M. (2001). Regression analysis of group testing samples. *Statistics in Medicine*, 20, 1957–1969.

Zhang, B., Bilder, C. R. and Tebbs, J. M. (2013). Group testing regression model estimation when case identification is a goal. *Biometrical Journal*, 55, 173–189.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112, 757–768.

Zhao, Q., Small, D. S. and Ertefaie, A. (2022). Selective inference for effect modification via the LASSO. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84, 382–413.