Probit time-to-event regression for misclassified group testing data

Lijun Fang[1], Tao Hu[2], Shuwei Li[1*], Lianming Wang [3], Christopher S. McMahan [4] and Joshua M. Tebbs[3]

[1] School of Economics and Statistics, Guangzhou University, Guangzhou, China

[2] School of Mathematical Sciences, Capital Normal University, Beijing, China

[3]Department of Statistics, University of South Carolina, Columbia, South Carolina, U.S.A.

[4] School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, U.S.A.

Abstract

Group testing has been used extensively to reduce the testing time and the screening costs in epidemiological studies involving low-prevalence diseases. This testing strategy works by first combining specimens (e.g., blood, urine, swabs, etc.) from several individuals to form a pool and then testing the pooled specimen for infection. When the endpoint of interest is a time-to-event outcome, for example, the time until infection or disease, and pools are tested only once, the resulting data are called group-tested current status data (Petito and Jewell, 2016). In this paper, we propose a new type of regression analysis for these data using a semiparametric probit model, an alternative to the proportional hazards model used in survival analysis. A sieve maximum likelihood estimation approach is developed that approximates the model's nonparametric nuisance function with logarithmic monotone splines. To facilitate sieve estimation, we develop a highly efficient expectation-maximization algorithm. The asymptotic properties of the resulting estimators are investigated by using empirical process techniques and sieve estimation theory. Numerical results from simulation studies suggest our proposed method performs nominally, even when pools are possibly misclassified due to assay error, and can outperform individual testing when the number of assays (tests) is fixed. We illustrate our work by estimating a time-to-event regression model for chlamydial infection using group testing data from a large public health laboratory in Iowa.

Keywords: Censored data; Expectation-maximization algorithm; Maximum likelihood estimation; Misclassification; Pooled testing; Sieve estimation.

*Address Correspondence to Shuwei Li, School of Economics and Statistics, Guangzhou University, Guangzhou, China 510006; E-mail: seslishuw@gzhu.edu.cn.

1 Introduction

Group testing was originally proposed by Dorfman (1943) to screen members of the United States military for syphilis during World War II. This testing strategy works by first collecting a biological specimen (e.g., blood, urine, swab, etc.) from a number of different individuals and then pooling these specimens together. The pooled specimen is then tested for infection or disease. If a pooled specimen tests negatively, then all individuals in the pool are declared to be negative at the expense of a single test. If a pooled specimen tests positively, individuals within it can be retested one at a time or in some other predetermined manner. When the disease of interest has low prevalence, group testing can save time and money when compared to testing each individual separately. It is not surprising this form of testing garnered widespread attention in the early stages of the recent covid-19 pandemic. Of course, group testing has also been adopted in a full panoply of application areas outside of testing for infectious diseases; some include DNA library screening (Berger et al., 2000), drug discovery (Xie et al., 2001; Remlinger et al., 2006), food contamination assessment (Fahey et al., 2006), environmental monitoring (Heffernan et al., 2014), and veterinary medicine (Baruch et al., 2020).

Since Dorfman's seminal work, research in group testing has flourished, and, more recently, a large number of regression methods have been developed for analyzing group testing data when covariate information is available. The first regression approach came from Farrington (1992), who estimated a specific generalized linear model under the restrictive assumption that individual covariates within pools were identical. Vansteelandt et al. (2000) and Xie (2001) separately extended this work to include any generalized linear model with pools having possibly different covariate values. Huang and Tebbs (2009) and Chen et al. (2009) examined group testing regression in the presence of covariate measurement error and random effects, respectively. Delaigle and Meister (2011) and Delaigle and Hall (2012) developed nonparametric approaches with a single continuous covariate and offered detailed asymptotic evaluations. Wang et al. (2014) proposed a general semiparametric framework that can incorporate multiple covariates and disease misclassification. McMahan et al. (2017) provided a Bayesian approach to estimate both a generalized linear model for disease status and the accuracy rates of the assays used.

All of the articles cited in the previous paragraph, and many others not cited, propose regression methods for group testing when the endpoint is binary, that is, an individual is diseased or not. However, in some applications, the endpoint of interest is not the disease status itself, but rather the *time* until the onset of the disease. Estimating time-to-event characteristics for individuals with group testing data is challenging, because individuals are tested in pools and the pools themselves are usually only tested at one time−at the time when screening occurs. An additional complication arises when pools are misclassified due to inherent assay error. Pools which are truly positive may test negatively if there are dilution effects; on the other hand, pools which are truly negative may test positively if there are synergistic or additive effects among the negative specimens (Xie et al., 2001). Therefore, the true individual disease onset times are not observed due to the current status data structure and the assessments of pools for disease status at the time of testing are potentially error-laden.

Despite these complex challenges, some progress has been made in combining time-to-event analysis with group testing. Petito and Jewell (2016) first studied the current status data problem with pools in the absence of covariates and proposed a constrained expectation-maximization algorithm to estimate the population-level survival function of the time until disease onset. These authors also performed an analysis for hepatitis C infection among American women of child-bearing age, showing that estimating time-to-disease characteristics with individual current status data can provide similar results and conclusions as those with current status data from group testing. More recently, when subject-specific covariates are available, Li et al. (2024) developed an expectation-maximization algorithm to estimate a proportional hazards regression model (Cox, 1972) for the time until disease onset with group testing data. These authors adopted a sieve estimation approach by first approximating the cumulative baseline hazard function with a piecewise constant function and then proceeded to derive asymptotic properties of the resulting maximum likelihood estimators. An interesting theoretical finding was that, under certain conditions, large-sample properties of estimators from group testing were identical to those from individual testing when the number of tests is fixed.

In this paper, we explore further the merger of time-to-event analysis with current

status data from studies which use group testing as a cost-saving strategy. We explore regression analysis of group-tested current status data with semiparametric probit model while accommodating misclassified testing results due to the use of imperfect tests. The semiparametric probit model provides an important alternative to the commonly used PH model. In contrast to the PH model, one primary advantage of the probit model is that the random error terms are assumed to follow the standard normal distribution, which can greatly facilitate developing an efficient, easy-to-implement and reliable inference procedure. Due to this desirable feature, many inference procedures have already been developed for the probit model under various types of survival data (Shiboski, 1998; Lin and Wang, 2010; Huang and Cai, 2016; Wu and Wang, 2019; Du et al., 2019; Fang et al., 2023). However, no estimation procedure has been reported for analyzing group-tested current status data using this model.

In this work, we propose a reliable and stable estimation approach based on sieve maximum likelihood for analyzing group-tested current status data with semiparametric probit model. Specifically, we first approximate the nonparametric nuisance function in the model with the logarithmic monotone splines and then propose an efficient EM algorithm to obtain the sieve estimators. The proposed algorithm has several enticing features. First, all of the conditional expectations involved in the E-step have closed-form expressions. Second, the objective function in the M-step is easy to optimize since it has a simple and tractable form as in the least squares estimation for complete data. In particular, the finite-dimensional spline coefficients can be readily updated with Newton-Raphson algorithm, and the regression parameters have closed-form solutions. By adopting the empirical process techniques and sieve estimation theory, the estimators of the regression parameters are shown to be consistent, asymptotically normally distributed and efficient. In addition, unlike Li et al. (2024), which adopted a time-consuming resampling procedure to estimate the covariance matrix of the regression parameter estimates, we obtain the variance estimates through a numerical profile likelihood method with a high estimation accuracy and easy implementation.

The rest of this paper is arranged as follows. Section 2 introduces the description of the data structure, considered model, observed likelihood, and assumptions needed for the

proposed approach as well as the adoption of the monotone spines for the nuisance function. Section 3 presents the proposed sieve maximum likelihood estimation method and some detail of the derivation of our proposed EM algorithm. In Section 4, we investigate the asymptotic properties of the proposed sieve estimators. Simulation studies are conducted to investigate the performance of the proposed methodology in Section 5, followed by an application to chlamydia data arising from the Infertility Prevention Project in Section 6. Section 7 provides some concluding remarks and discussions.

2 Model, Data and Likelihood

Consider a group testing study that involves screening N independent individuals for a disease of interest. We assume that these N individuals are randomly assigned to n groups, which are subsequently tested to identify the disease statuses. Denote by J_i the size of the ith group with $i = 1, \ldots, n$, that is, $N = \sum_{i=1}^{n} J_i$. Let T_{ij} be the disease onset time and $\boldsymbol{Z}_{ij}$ a $p \times 1$ vector of covariates for the jth individual in the ith group. To characterize the effects of $\boldsymbol{Z}_{ij}$ on T_{ij}, we consider a semiparametric probit model, which specifies the conditional cumulative distribution function (CDF) of T_{ij} given $\boldsymbol{Z}_{ij}$ as

$$F(t \mid \boldsymbol{Z}_{ij}) = \Phi\left\{\alpha(t) + \boldsymbol{\beta}^{\top}\boldsymbol{Z}_{ij}\right\}, \tag{1}$$

where $\Phi\{\cdot\}$ is the CDF of the standard normal random variable, $\boldsymbol{\beta}$ is a vector of covariate effects, and $\alpha(\cdot)$ is an unspecified and increasing nuisance function with $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$. The focus of the work herein is to fit model (1) based on group-tested current status data. Notably, the probit model (1) can be derived from the following model

$$\alpha(T_{ij}) = -\boldsymbol{\beta}^{\top}\boldsymbol{Z}_{ij} + \varepsilon_{ij},$$

where $\{\varepsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, J_i\}$ is a set of independent standard normal random variables. In other words, the probit model directly examines the covariate effects on the transformed failure time T_{ij}.

Define $\phi_{ij} = I(T_{ij} \leq X_{ij})$ as the disease occurrence status of the jth individual in the ith group at the testing time X_{ij}, where $I(\cdot)$ denotes the indicator function. Note that, under a group testing strategy, the disease onset current status ϕ_{ij} is unobserved, and the

testing time can be subject specific (e.g., age at testing). The true disease occurrence status of the ith group can be denoted as $\Delta_i = \max(\phi_{ij}; j = 1, \ldots, J_i)$ or $\Delta_i = I\left(\sum_{j=1}^{J_i} \phi_{ij} > 0\right)$. That is, $\Delta_i = 1$ means that ith group contains at least one infected individual and $\Delta_i = 0$ indicates that all individuals in the ith group are disease free. However, in many applications, the test used is not perfect, and consequently the true group statuses Δ_i's are not observable. Instead, one observes the testing outcome Y_i for ith group, with $Y_i = 1$ denoting the event that the ith group tests positively for disease and $Y_i = 0$ otherwise. Then the observed data include $\{Y_i, X_{ij}, \boldsymbol{Z}_{ij}; i = 1, \ldots, n,\ j = 1, \ldots, J_i\}$.

Define $\nu = P(Y_i = 1 \mid \Delta_i = 1)$ and $\omega = P(Y_i = 0 \mid \Delta_i = 0)$ as the sensitivity and specificity of the test, respectively. Throughout the paper, we make the following assumptions:

(I) ν and ω are known constants with $\nu + \omega > 1$.

(II) ν and ω are independent of the group size, the testing times, and the covariates.

(III) The contributed individuals in each group are independent of each other.

(IV) For each i and j, T_{ij} and X_{ij} are conditionally independent given the covariates.

Assumptions I – III are standard assumptions as given in the literature of group testing data (Wang et al., 2014) and group-tested current status data (Petito and Jewell, 2016). Assumption IV is usually referred to as the non-informative or conditional independent censoring in the traditional current status data analysis (Sun, 2006). Without losing generality, we temporarily assume that the sensitivity (specificity) are same for all groups for notational simplicity.

Under model (1) and the aforementioned assumptions, the observed data likelihood function can be derived as

$$L(\boldsymbol{\beta}, \alpha) = \prod_{i=1}^{n} \left\{ \nu - \gamma \prod_{j=1}^{J_i} \left(1 - \Phi\left\{\alpha(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right) \right\}^{Y_i} \times \left\{ 1 - \nu + \gamma \prod_{j=1}^{J_i} \left(1 - \Phi\left\{\alpha(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right) \right\}^{1-Y_i}, \tag{2}$$

where $\gamma = \nu + \omega - 1$ is a positive constant related to the test sensitivity and specificity. To derive likelihood $L(\boldsymbol{\beta}, \alpha)$ in (2), one needs to calculate both $P(Y_i = 1 \mid D_i)$ and $P(Y_i = 0 \mid D_i)$, where $D_i = \{X_{ij}, \boldsymbol{Z}_{ij}; j = 1, \ldots, J_i\}$ for $i = 1, \ldots, n$. By applying the law of total

probability and the conditional probability formula, $P(Y_i = 1 \mid D_i)$ is given by

$$\begin{aligned}
P(Y_i = 1 \mid D_i) &= P(Y_i = 1, \Delta_i = 1 \mid D_i) + P(Y_i = 1, \Delta_i = 0 \mid D_i) \\
&= P(Y_i = 1 \mid \Delta_i = 1,\, D_i)P(\Delta_i = 1 \mid D_i) + P(Y_i = 1 \mid \Delta_i = 0,\, D_i)P(\Delta_i = 0 \mid D_i) \\
&= \nu P(\Delta_i = 1 \mid D_i) + (1 - \omega)P(\Delta_i = 0 \mid D_i) \\
&= \nu(1 - P(\Delta_i = 0 \mid D_i)) + (1 - \omega)P(\Delta_i = 0 \mid D_i) \\
&= \nu - \gamma P(\Delta_i = 0 \mid D_i).
\end{aligned}$$

Furthermore, based on model (1) and assumption IV, we have

$$\begin{aligned}
P(\Delta_i = 0 \mid D_i) &= P(T_{i1} > X_{i1}, \ldots, T_{iJ_i} > X_{iJ_i} \mid D_i) \\
&= \prod_{j=1}^{J_i} S(X_{ij} \mid \boldsymbol{Z}_{ij}) = \prod_{j=1}^{J_i} \left(1 - \Phi\left\{\alpha(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right),
\end{aligned}$$

and $P(\Delta_i = 1 \mid D_i) = 1 - P(\Delta_i = 0 \mid D_i)$ for $i = 1, \ldots, n$.

Notably, the observed data likelihood (2) is a function of the regression vector $\boldsymbol{\beta}$ and nuisance function $\alpha(\cdot)$. Unlike the case for right censored data, no partial likelihood method is available for group-tested current status data, and instead, one needs to estimate $\boldsymbol{\beta}$ and $\alpha(\cdot)$ simultaneously. Since $\alpha(\cdot)$ is a infinite-dimensional function, approximating it with some smooth functions is a commonly adopted approach in the survival literature. In particular, we approximate $\alpha(t)$ with the logarithmic monotone splines as follows,

$$\alpha_n(t) = \log\left\{\sum_{l=1}^{L_n} \xi_l b_l(t)\right\},$$

where b_l's are integrated spline basis functions, each of which is nondecreasing from 0 to 1, and ξ_l's are nonnegative spline coefficients to guarantee the monotonicity (Ramsay, 1988). To construct these basis functions, it is necessary to specify a sequence of q_n increasing points as interior knots and choose an order denoted by k for the splines. In particular, one can obtain the linear, quadratic, and cubic functions by setting k to be 1, 2 and 3, respectively. The number of basis functions $L_n = q_n + k$ basis functions are fully determined if the order and interior knots have been specified for the monotone splines.

After approximating $\alpha(\cdot)$ with the logarithmic monotone splines, the observed likelihood

(2) becomes

$$L(\boldsymbol{\beta}, \boldsymbol{\xi}) = \prod_{i=1}^{n} \left\{ \nu - \gamma \prod_{j=1}^{J_i} \left(1 - \Phi\left\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right) \right\}^{Y_i} \times \left\{ 1 - \nu + \gamma \prod_{j=1}^{J_i} \left(1 - \Phi\left\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right) \right\}^{1-Y_i}, \tag{3}$$

where $\boldsymbol{\xi} = (\xi_l, \ldots, \xi_{L_n})^\top$ is the vector of spline coefficients. Since likelihood (3) exhibits an intractable form, identifying an estimator of $(\boldsymbol{\beta}, \boldsymbol{\xi})$ via direct maximization is quite challenging. To overcome this difficulty, we resort to latent variable techniques and seek to find the maximum likelihood estimators via an EM algorithm.

3 Estimation Procedure

To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ based on (3), we develop an EM algorithm based on a three-stage data augmentation. The first stage of our data augmentation procedure introduces the true disease status of the pools, Δ_i's, as latent random variables. This yields the following augmented data likelihood

$$\begin{aligned} L_1(\boldsymbol{\beta}, \boldsymbol{\xi}) &= \prod_{i=1}^{n} P(Y_i, \Delta_i \mid D_i) = \prod_{i=1}^{n} P(Y_i \mid \Delta_i) P(\Delta_i \mid D_i) \\ &= \prod_{i=1}^{n} \left\{ 1 - \prod_{j=1}^{J_i} \left(1 - \Phi\left\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right) \right\}^{\Delta_i} \\ &\quad \times \left\{ \prod_{j=1}^{J_i} \left(1 - \Phi\left\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right) \right\}^{1-\Delta_i} P(Y_i \mid \Delta_i), \end{aligned}$$

where

$$P(Y_i \mid \Delta_i) = \left\{\nu^{Y_i} (1-\nu)^{1-Y_i}\right\}^{\Delta_i} \left\{(1-\omega)^{Y_i} \omega^{1-Y_i}\right\}^{1-\Delta_i}, \tag{4}$$

for $i = 1, \ldots, n$.

The second stage introduces the disease statuses of the individuals ϕ_{ij}'s as latent random variables. This step leads to the following augmented data likelihood

$$L_2(\boldsymbol{\beta}, \boldsymbol{\xi}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J_i} \Phi\left\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}^{\phi_{ij}} \left(1 - \Phi\left\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\right\}\right)^{1-\phi_{ij}} \right\} P(Y_i \mid \Delta_i), \tag{5}$$

where $\Delta_i = I\left(\sum_{j=1}^{J_i} \phi_{ij} > 0\right)$.

To further simplify the form of (5), in the third stage, we introduce a set of independent latent variables $\{G_{ij}; j = 1, \ldots, J_i, i = 1, \ldots, n\}$, where $G_{ij} = \alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} + \varepsilon_{ij}$ and $\{\varepsilon_{ij}; j = 1, \ldots, J_i, i = 1, \ldots, n\}$ is a set of i.i.d. standard normal random variables. Then, one can easily find that

$$P(\phi_{ij} = 1 \mid D_i) = P(G_{ij} \geq 0 \mid D_i) = \Phi\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\}.$$

The above fact indicates that the likelihood function (5) can be equivalently expressed with G_{ij}'s as

$$L_3(\boldsymbol{\beta}, \boldsymbol{\xi}) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} P(G_{ij} \geq 0 \mid D_i)^{\phi_{ij}} P(G_{ij} < 0 \mid D_i)^{1-\phi_{ij}} P(Y_i \mid \Delta_i).$$

Treating G_{ij}'s as observable, our complete data likelihood function takes the form

$$L_c(\boldsymbol{\beta}, \boldsymbol{\xi}) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left\{G_{ij} - \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} - \alpha_n(X_{ij})\right\}^2\right) P(Y_i \mid \Delta_i)$$

with the following constraint: $G_{ij} \geq 0$ if $\phi_{ij} = 1$ and $G_{ij} < 0$ if $\phi_{ij} = 0$ for each i and j. After removing some constants that are irrelevant to the unknown parameters, the complete data log-likelihood function can be simplified as

$$\begin{aligned} l_c(\boldsymbol{\beta}, \boldsymbol{\xi}) &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \left\{G_{ij} - \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} - \alpha_n(X_{ij})\right\}^2 \left\{\phi_{ij} 1_{(G_{ij}>0)} + (1-\phi_{ij}) 1_{(G_{ij}<0)}\right\} \\ &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \left\{G_{ij} - \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} - \log\left(\sum_{l=1}^{L_n} \xi_l b_l(X_{ij})\right)\right\}^2 \left\{\phi_{ij} 1_{(G_{ij}>0)} + (1-\phi_{ij}) 1_{(G_{ij}<0)}\right\}. \end{aligned}$$

In the E-step of the algorithm, one takes the conditional expectation of $l_c(\boldsymbol{\beta}, \boldsymbol{\xi})$ with respect to all latent variables, which yields the objective function

$$\begin{aligned} Q_c(\boldsymbol{\beta}, \boldsymbol{\xi}; \boldsymbol{\beta}^{(m)}, \boldsymbol{\xi}^{(m)}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \Bigg\{ & E(\phi_{ij}) \left(\tau_{ij}^{+} + \left[\mu_{ij}^{+} - \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} - \log\left(\sum_{l=1}^{L_n} \xi_l b_l(X_{ij})\right)\right]^2\right) \\ & + \{1 - E(\phi_{ij})\} \left(\tau_{ij}^{-} + \left[\mu_{ij}^{-} - \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} - \log\left(\sum_{l=1}^{L_n} \xi_l b_l(X_{ij})\right)\right]^2\right) \Bigg\}. \end{aligned}$$

In the above, $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$ denote the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ at the mth iteration, respectively, μ_{ij}^{+} and μ_{ij}^{-} are the conditional expectations of G_{ij} under the constraints of

$G_{ij} \geq 0$ and $G_{ij} < 0$, respectively, and τ_{ij}^{+} and τ_{ij}^{-} are the conditional variances of G_{ij} under the constraints of $G_{ij} \geq 0$ and $G_{ij} < 0$, respectively. For notational simplicity, we ignore all the conditional arguments in the conditional expectations, including the observed data $\mathcal{O}_i = \{(X_{ij}, Y_i, \boldsymbol{Z}_{ij}); i = 1, \ldots, n,\ j = 1, \ldots, J_i\}$, $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$. After omitting some constants, the objective function $Q_c(\boldsymbol{\beta}, \boldsymbol{\xi}; \boldsymbol{\beta}^{(m)}, \boldsymbol{\xi}^{(m)})$ becomes

$$Q(\boldsymbol{\beta}, \boldsymbol{\xi}; \boldsymbol{\beta}^{(m)}, \boldsymbol{\xi}^{(m)}) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \left\{ E(\phi_{ij}) \left[\mu_{ij}^{+} - \boldsymbol{\beta}^{\top} \boldsymbol{Z}_{ij} - \log \left(\sum_{l=1}^{L_n} \xi_l b_l(X_{ij}) \right) \right]^2 \right.$$
$$\left. + \{1 - E(\phi_{ij})\} \left[\mu_{ij}^{-} - \boldsymbol{\beta}^{\top} \boldsymbol{Z}_{ij} - \log \left(\sum_{l=1}^{L_n} \xi_l b_l(X_{ij}) \right) \right]^2 \right\}. \tag{6}$$

By applying the properties of the truncated normal distribution and the Bayesian theorem, we are able to obtain closed-form expressions for all of the conditional expectations depicted in (6) and they are given by

$$\mu_{ij}^{+} = \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) + \frac{\phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\}}{\Phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\}},$$

$$\mu_{ij}^{-} = \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) - \frac{\phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\}}{1 - \Phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\}},$$

and

$$E(\phi_{ij}) = Y_i \frac{\nu \, \Phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\}}{\nu - \gamma \prod_{j=1}^{J_i} \left[1 - \Phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\} \right]}$$
$$+ (1 - Y_i) \frac{(1 - \nu) \, \Phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\}}{1 - \nu + \gamma \prod_{j=1}^{J_i} \left[1 - \Phi \left\{ \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m)} + \log \left(\sum_{l=1}^{L_n} \xi_l^{(m)} b_l(X_{ij}) \right) \right\} \right]},$$

where $\phi\{\cdot\}$ is the density function of the standard normal random variable. Details about these derivations are sketched in supplementary materials. Note, having closed-form expressions of the necessary conditional expectations contributes greatly to the computational efficiency of the proposed EM algorithm as the numerical integration techniques are not needed.

The M-step of the algorithm then updates $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$ by maximizing the Q function in (6) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. To this end, we first solve $\partial Q(\boldsymbol{\beta}, \boldsymbol{\xi}; \boldsymbol{\beta}^{(m)}, \boldsymbol{\xi}^{(m)})/\partial \boldsymbol{\beta} = 0$,

rendering a closed-form solution for $\boldsymbol{\beta}$ as a function of $\boldsymbol{\xi}$,

$$\boldsymbol{\beta}^{(m+1)}(\boldsymbol{\xi}) = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{J_i} \boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^{\top} \right\}^{-1} \times \left\{ \sum_{i=1}^{n} \sum_{j=1}^{J_i} \boldsymbol{Z}_{ij} \left[E(\phi_{ij}) \mu_{ij}^{+} + \{1 - E(\phi_{ij})\} \mu_{ij}^{-} - \log \left(\sum_{l=1}^{L_n} \xi_l b_l(X_{ij}) \right) \right] \right\}. \quad (7)$$

Note that the spline coefficients ξ_l's are nonnegative, to avoid the use of constraint optimization, we reparameterize ξ_l as $\exp(\xi_l^*)$, $l = 1, \ldots, L_n$. By plugging $\boldsymbol{\beta}^{(m+1)}$ into (6) and replacing each ξ_l with $\exp(\xi_l^*)$, the score equation for ξ_l^* can be derived as

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \left[E(\phi_{ij}) \mu_{ij}^{+} + \{1 - E(\phi_{ij})\} \mu_{ij}^{-} - \boldsymbol{Z}_{ij}^{\top} \boldsymbol{\beta}^{(m+1)} - \log \left\{ \sum_{l=1}^{L_n} \exp(\xi_l^*) b_l(X_{ij}) \right\} \right] \times \frac{\exp(\xi_l^*) b_l(X_{ij})}{\sum_{l=1}^{L_n} \exp(\xi_l^*) b_l(X_{ij})} = 0. \quad (8)$$

Since the estimating equation (8) has tractable form, one can readily obtain $\xi_l^{*(m+1)}$ with the simple Newton-Raphson algorithm and then obtain $\xi_l^{(m+1)} = \exp(\xi_l^{*(m+1)})$ for $l = 1, \ldots, L_n$.

In summary, the step-by-step implementation of our algorithm can be described as follows.

Step 1. Set $m = 0$ and choose initial values $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\xi}^{(0)}$.

Step 2. At the $(m+1)$th iteration, calculate the conditional expectations μ_{ij}^{+}, μ_{ij}^{-} and $E(\phi_{ij})$ at $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$.

Step 3. Compute $\boldsymbol{\beta}^{(m+1)}$ with the closed-form expression (7) by letting $\boldsymbol{\xi} = \boldsymbol{\xi}^{(m)}$.

Step 4. For each $l = 1, \ldots, L_n$, obtain $\xi_l^{*(m+1)}$ by solving (8), in which other components in $\boldsymbol{\xi}^* = (\xi_1^*, \ldots, \xi_{L_n}^*)^{\top}$ are fixed to their mth updates, and set $\xi_l^{(m+1)} = \exp(\xi_l^{*(m+1)})$.

Step 5. Increase m by 1 and repeat Steps 2–4 until the convergence is achieved.

The proposed EM algorithm above is found to be robust to initialization. In practice, one simply set the initial value of each component in $\boldsymbol{\beta}$ to be 0 and let the initial values of the spline coefficients to be L_n random values generated from the exponential distribution with mean 1/7. The algorithm is declared convergent if the sum of all absolute differences of estimates between two successive iterations is less than a small positive constant (e.g. 10^{-4}).

4 Asymptotic Properties

In this section, we investigate the asymptotic properties of the proposed sieve maximum likelihood estimator. Define $\boldsymbol{\Theta} = \{\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha) \in \mathcal{B} \otimes \mathcal{A}\}$, where $\mathcal{B}$ is a compact set in $\mathbb{R}^p$, and $\mathcal{A}$ contains all bounded and continuous nondecreasing functions over $[\tau_1, \tau_2]$ with $0 < \tau_1 < \tau_2 < \infty$. Define the sieve space as $\boldsymbol{\Theta}_n = \{\boldsymbol{\theta}_n = (\boldsymbol{\beta}, \alpha_n) \in \mathcal{B} \otimes \mathcal{A}_n\}$, where $\mathcal{A}_n = \{\alpha_n(t) = \log \Lambda_n(t) = \log\{\sum_{l=1}^{L_n} \xi_l b_l(t)\} : \xi_l \geq 0,\ 0 \leq b_l(t) \leq 1,\ t \in [\tau_1, \tau_2]\}$. Denote $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n, \hat{\alpha}_n)$ as the sieve maximum likelihood estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)$ obtained by maximizing $\log L(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}_n$, and let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \alpha_0)$ be the true value of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)$. Note that we approximate $\alpha(t)$ with the logarithmic monotone splines and thus simplify the estimation problem by reducing a semiparametric model to a weakly parametric one. However, since the number of spline basis functions increases with sample size, traditional parametric likelihood theory is no longer applicable. To establish the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$, we resort to the empirical process techniques and sieve estimation theory. Following Delaigle and Meister (2011), Wang et al. (2014) and others, we assume that the number of groups n tends to infinity as the sample size N goes to infinity while the group sizes J_i's remain finite.

Let $\|\boldsymbol{b}\|$ denote the Euclidean norm for a vector $\boldsymbol{b}$. Define the distance between any $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1, \alpha_1) \in \boldsymbol{\Theta}$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2, \alpha_2) \in \boldsymbol{\Theta}$ as

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \left(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2 + \|\alpha_1 - \alpha_2\|_2^2\right)^{1/2},$$

where $\|\alpha_1 - \alpha_2\|_2 = [\int_{\tau_1}^{\tau_2} \{\alpha_1(u) - \alpha_2(u)\}^2 \, \mathrm{d}Q(u)]^{1/2}$ and $Q(\cdot)$ is the cumulative distribution function of the observation time. Let $\mathcal{T}_n = \{t_i, i = 1, \ldots, q_n + 2k\}$, with

$$\tau_1 = t_1 = \cdots = t_k < t_{k+1} < \cdots < t_{q_n+k} < t_{q_n+k+1} = \cdots = t_{q_n+2k} = \tau_2,$$

denote a sequence of knots that partition $[\tau_1, \tau_2]$ into $q_n + 1$ subintervals, where $q_n = O(n^\kappa)$ for $0 < \kappa < 0.5$. To establish the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$, we require the following regularity conditions:

(A1) (a) The true value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_0$, lies in the known compact set $\mathcal{B}$. (b) The true value of α, denoted by α_0, is continuously differentiable with a positive first derivative and has a bounded rth derivative in $[\tau_1, \tau_2]$ for $r \geq 1$.

(A2) The covariate vector $\boldsymbol{Z}_j$ is bounded with probability one for $j = 1, \ldots, J$.

(A3) $E(\boldsymbol{Z}_j \boldsymbol{Z}_j^\top) > 0$ for $j = 1, \ldots, J$.

(A4) For $j = 1, \ldots, J$, if $p(x) + \boldsymbol{\beta}^\top \boldsymbol{Z}_j = 0$ for all $x \in [\tau_1, \tau_2]$ with probability one, then $p(x) = 0$ for $x \in [\tau_1, \tau_2]$ and $\boldsymbol{\beta} = 0$.

(A5) The maximum spacing of the knots satisfies $\tilde{\Delta}_{\max} = \max_{k+1 \le q \le q_n+k+1} |t_q - t_{q-1}| = O(n^{-\kappa})$ with $\kappa \in (0, 0.5)$, and $\tilde{\Delta}_{\max}/\tilde{\Delta}_{\min}$ is bounded, where $\tilde{\Delta}_{\min} = \min_{k+1 \le q \le q_n+k+1} |t_q - t_{q-1}|$ is the minimum spacing of the knots.

(A6) If $p(\boldsymbol{x}, \boldsymbol{Z}) + \sum_{j=1}^{J} \boldsymbol{\eta}_j^\top \boldsymbol{Z}_j = 0$ for all $\boldsymbol{x} \in [\tau_1, \tau_2]^J$ with probability one, where $\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_J^\top)^\top$, then $p(\boldsymbol{x}, \boldsymbol{Z}) = 0$ for $\boldsymbol{x} \in [\tau_1, \tau_2]^J$ and $\boldsymbol{\eta}_j = 0$ for $j = 1, \ldots, J$.

Conditions (A1)–(A3) are mild and commonly used in interval-censored data analysis (Huang and Rossini, 1997; Huang et al., 2008; Zhang et al., 2010). Condition (A4) holds if the matrix $E([1, \boldsymbol{Z}_j^\top]^\top [1, \boldsymbol{Z}_j^\top])$ is nonsingular for $j = 1, \ldots, J$, and is used to ensure the model identifiability (Zeng et al., 2016, 2017). Condition (A5) is required to derive the convergence rate and asymptotic normality and is the same as Condition 1 of Lu et al. (2007). Condition (A6) is used to prove the invertibility of the efficient Fisher information matrix and holds if the matrix $E([1, \boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_J^\top]^\top [1, \boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_J^\top])$ is nonsingular. Let $q_n = O(n^\kappa)$ for $1/2(1 + r) < \kappa < 1/2r$. We state the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$ with the following three theorems and present the detailed proofs of them in the supplementary materials.

Theorem 1: Under conditions (A1)–(A4), we have $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \to 0$ almost surely as $n \to \infty$.

Theorem 2: Under conditions (A1)–(A5), we have $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p\left(n^{-\min\{r\kappa, (1-\kappa)/2\}}\right)$.

Theorem 3: Under conditions (A1)–(A6), we have $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to N(0, I^{-1}(\boldsymbol{\beta}_0))$ in distribution as $n \to \infty$, where $I(\boldsymbol{\beta}_0)$ is the information matrix of $\boldsymbol{\beta}_0$ defined in the supplementary materials.

It is worth noting from Theorem 2 that the choice of $\kappa = 1/(1 + 2r)$ yields the optimal convergence rate, $n^{-r/(1+2r)}$, in the nonparametric regression. In particular, the convergence rate of the proposed estimator is $n^{-1/3}$ when $r = 1$ and can increase to $n^{-2/5}$ if $r = 2$. To make inference about $\boldsymbol{\beta}$, the finite-dimensional parameter of interest, one often needs to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_n$. Although Theorem 3 suggests that $\hat{\boldsymbol{\beta}}_n$ is asymptotically normally distributed, the covariance matrix of $\hat{\boldsymbol{\beta}}_n$ is not directly obtained

due to its intractable form as shown in the proof of Theorem 3. In what follows, we adopt a numerical profile likelihood method as in Zeng et al. (2017) and others and approximate the covariance matrix of $\hat{\boldsymbol{\beta}}_n$ by $(n\hat{V}_n)^{-1}$, where

$$\hat{V}_n = n^{-1} \sum_{i=1}^{n} \left[\left\{ \frac{\partial}{\partial \boldsymbol{\beta}} l_i(\boldsymbol{\beta}, \hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}}) \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \right\}^{\otimes 2} \right],$$

$l_i(\boldsymbol{\beta}, \hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}})$ is the log-likelihood function of the ith group, $\hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\xi}} \log L(\boldsymbol{\beta}, \boldsymbol{\xi})$, $L(\boldsymbol{\beta}, \boldsymbol{\xi})$ is the observed likelihood given by (3), and $\boldsymbol{b}^{\otimes 2} = \boldsymbol{b}\boldsymbol{b}^\top$ for a column vector $\boldsymbol{b}$. Note that $\hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}}$ can be easily obtained by using the proposed EM algorithm with the fixed $\boldsymbol{\beta}$, and the gradient $\frac{\partial}{\partial \boldsymbol{\beta}} l_i(\boldsymbol{\beta}, \hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}}) \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ can be approximated with the first-order numerical difference. The simulation results below show that this approach works reasonably well in practical situations.

5 Simulation Studies

In this section, we conducted simulation studies to evaluate the finite sample performance of the proposed method. In the first study, we considered a group testing strategy that randomly divided $N = 10000$ individuals into $n = 2000$ groups with each group size of 5. That is, $J_i = 5$, for $i = 1, \ldots, n = 2000$. To obtain group-tested current status data, we first generate the individuals' disease onset times T_{ij}'s from model (1), where $\alpha(t) = \log(t)$, $\boldsymbol{Z}_{ij} = (Z_{ij1}, Z_{ij2})^\top$, $Z_{ij1} \sim Bernoulli(0.5)$ and $Z_{ij2} \sim Uniform(0, 1)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (0.5, -0.5)^\top$. For the jth individual in the ith group, the true disease occurrence status at the time of testing was obtained by $\phi_{ij} = I(T_{ij} \leq X_{ij})$, where the observations times X_{ij}'s followed $Uniform(0, 0.5)$. On average, the right censoring rate was approximately 90%. Then we can obtain the true group-based disease statuses $\Delta_i = I(\sum_{j=1}^{J_i} \phi_{ij} > 0)$ for $i = 1, \ldots, n$. To generate the potentially misclassified group-tested current status data, we set the sensitivity and specificity to be $(\nu, \omega) = (1, 1)$, $(0.95, 0.95)$, $(0.90, 0.95)$, $(0.90, 0.90)$, or $(0.85, 0.85)$. In particular, $(\nu, \omega) = (1, 1)$ corresponds to the case of a perfect test. Given the values of ν and ω, the observed group testing results Y_i's were then generated from Bernoulli distributions based on equation (4). Five hundred data sets were generated for each simulation setup.

To apply the proposed method, we set the order of monotone splines to be $k = 3$ and the number of interior knots to be $q_n = 5$. By following McMahan et al. (2013b) and others, we specified the interior knots as the equally spaced points within the interval $[X_{min} - 10^{-5}, X_{max} + 10^{-5}]$, where X_{min} and X_{max} denote the minimum and maximum observation time, respectively. Table 1 presents the numerical results for the estimation of the regression parameters, β_1 and β_2, including the estimated bias (Bias) calculated by the average of 500 estimates minus the true value, the sample standard error (SSE) of 500 estimates, the average of 500 standard error estimates (SEE), and the 95% empirical coverage probability (CP). The results shown in Table 1 indicate that the proposed method performs reasonably well under the situations considered. Specifically, the proposed estimators seem to be unbiased, the SSEs align with the SEEs, and the coverage probabilities are close to the nominal value 95%.

We further focus on investigating the advantages of the proposed method over the individual-based method considering the individual level data. To this end, we first consider the situation in which both techniques testing $N = 10000$ individuals. We generated the potentially individual-based testing results from ϕ_{ij}'s with the pre-specified values of ν and ω. The individual-based method can be accomplished with the proposed algorithm by specifying the group sizes to be 1. The results are summarized in Table 1, form which one can find that the biases of the individual-based method are comparable to those of the proposed method based on the group-tested current status data. However, the SEEs based on group-tested data are larger than those based on the individual level data. These suggest that our method based on group-tested current status data loses some estimation efficiency compared with the method focusing on the individual level data, which is anticipated because the group-tested data contain much less information than the individual level data. On the other hand, a study design with group testing strategy can save 80% screening costs compared with the individual screening when the group size is 5. In other words, significant savings in the assaying efforts can be realized by employing the proposed method.

In the second comparison, we consider another practical situation with limited testing cost, where only n detection reagents are available and n individual-based tests are conducted. The analysis results are included in the rightmost panel of Table 1, from which one

can find that the proposed method based on the group-tested current status data yields more efficient estimators than this individual-based method. In other words, for the same cost we are able to obtain more precise estimators through the analysis of group-tested data. The findings clearly demonstrates the desirable advantages of conducting group testing in large-scale studies.

To further examine the performance of the proposed method for the estimation of the baseline survival function $S(t) = 1 - \Phi(\alpha(t))$, we plotted in Figure 1 the average estimate of $S(t)$ as well as the true curve with $(\nu, \omega) = (1, 1)$, $(0.90, 0.95)$ or $(0.85, 0.85)$. For comparison, the average estimates of $S(t)$ obtained by the two individual-based methods are also included in Figure 1. One can find from Figure 1 that the average estimate of $S(t)$ of the proposed method is extremely close to the true curve, and comparable to the average estimates of $S(t)$ of the two individual-based methods.

In the second study, we examined the setting with more covariates and larger group number n. In particular, we independently generated 5 covariates from $Bernoulli(0.5)$ and set the true values of the regression parameters to be $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top = (0.5, 0.5, -0.5, -0.5, -0.5)^\top$. The group sizes J_i's were specified to be 1, 2, 3 or 4 with equal probabilities. The sensitivity and specificity were set to be $(\nu, \omega) = (1, 1)$, $(0.90, 0.95)$ or $(0.85, 0.85)$, and other simulation configurations were kept to be the same as above. The numerical results are summarized in Table 2, from which one can draw similar conclusions as above. In addition, we plotted in Figure 2 the average estimates of $S(t)$ obtained by the proposed method and two individual-based methods, which again shows the satisfactory performances of the three methods.

Table 1: Simulation results from the proposed method based on group-tested current status data and the individual-based method based on misclassified individual data with two covariates, which include the estimated bias (Bias), the sample standard error (SSE) of the estimates, the average of the standard error estimates (SEE), and the 95% empirical coverage probability (CP).

		Proposed method				Individual-based method (N)				Individual-based method (n)			
(ν, ω)		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
(1.00,1.00)	$\hat{\beta}_1$	0.001	0.081	0.080	95.2	-0.001	0.040	0.038	94.6	0.005	0.087	0.087	95.0
	$\hat{\beta}_2$	-0.004	0.134	0.131	94.4	0.003	0.066	0.065	95.4	-0.001	0.139	0.147	95.8
(0.95,0.95)	$\hat{\beta}_1$	0.001	0.087	0.091	96.2	0.001	0.047	0.050	96.2	-0.001	0.111	0.112	95.4
	$\hat{\beta}_2$	-0.002	0.154	0.148	93.2	0.001	0.085	0.083	94.8	-0.015	0.187	0.187	95.0
(0.90,0.95)	$\hat{\beta}_1$	0.001	0.092	0.096	96.0	0.001	0.050	0.052	96.0	0.007	0.115	0.118	95.2
	$\hat{\beta}_2$	-0.002	0.158	0.155	92.8	0.000	0.088	0.086	95.4	-0.012	0.199	0.195	93.8
(0.90,0.90)	$\hat{\beta}_1$	0.000	0.097	0.104	96.0	0.002	0.061	0.060	94.6	0.013	0.136	0.137	95.2
	$\hat{\beta}_2$	-0.005	0.169	0.167	94.6	-0.006	0.099	0.099	94.8	0.002	0.218	0.223	96.0
(0.85,0.85)	$\hat{\beta}_1$	-0.004	0.117	0.120	95.4	0.002	0.076	0.073	94.2	0.007	0.161	0.165	95.6
	$\hat{\beta}_2$	-0.008	0.194	0.193	94.4	-0.011	0.123	0.119	96.0	0.008	0.263	0.268	97.2

Note: ν and ω are the sensitivity and specificity of a test, respectively. "Individual-based method (N)" and "Individual-based method (n)" denote the methods based on the misclassified individual current status data with sample sizes N and n, respectively.

Table 2: Simulation results from the proposed method based on group-tested current status data and the individual-based methods based on misclassified individual data with five binary covariates and varying group sizes, which include the estimated bias (Bias), the sample standard error (SSE) of the estimates, the average of the standard error estimates (SEE), and the 95% empirical coverage probability (CP).

		Proposed method				Individual-based method (N)				Individual-based method (n)			
(ν, ω)		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
(1.00,1.00)	$\hat{\beta}_1$	-0.001	0.043	0.045	96.0	0.000	0.035	0.032	94.6	-0.003	0.052	0.051	96.0
	$\hat{\beta}_2$	-0.002	0.047	0.045	94.4	-0.002	0.032	0.032	95.2	-0.001	0.050	0.051	94.8
	$\hat{\beta}_3$	-0.001	0.044	0.045	94.0	0.000	0.034	0.033	92.6	0.001	0.051	0.052	95.0
	$\hat{\beta}_4$	-0.001	0.047	0.045	93.0	0.000	0.034	0.033	94.6	0.000	0.051	0.052	95.2
	$\hat{\beta}_5$	-0.002	0.044	0.045	95.4	0.000	0.032	0.033	95.2	0.001	0.054	0.052	94.2
(0.90,0.95)	$\hat{\beta}_1$	0.000	0.058	0.057	95.8	0.000	0.047	0.044	94.2	-0.002	0.073	0.071	94.2
	$\hat{\beta}_2$	-0.003	0.059	0.057	93.4	-0.003	0.045	0.044	94.2	-0.003	0.071	0.071	95.4
	$\hat{\beta}_3$	-0.002	0.058	0.057	95.2	0.002	0.046	0.045	94.2	-0.002	0.071	0.071	96.0
	$\hat{\beta}_4$	-0.006	0.060	0.057	94.4	0.000	0.046	0.045	94.0	-0.004	0.075	0.072	94.6
	$\hat{\beta}_5$	-0.002	0.063	0.057	93.6	0.001	0.045	0.045	94.0	0.000	0.072	0.071	94.8
(0.85,0.85)	$\hat{\beta}_1$	0.001	0.077	0.075	94.4	-0.002	0.065	0.062	94.2	-0.001	0.099	0.098	94.4
	$\hat{\beta}_2$	-0.003	0.076	0.075	94.0	-0.004	0.063	0.062	93.8	-0.002	0.104	0.098	93.2
	$\hat{\beta}_3$	-0.006	0.078	0.076	94.6	-0.003	0.063	0.063	95.4	-0.011	0.098	0.100	96.4
	$\hat{\beta}_4$	-0.007	0.078	0.076	94.0	-0.007	0.066	0.064	93.6	-0.011	0.107	0.101	93.4
	$\hat{\beta}_5$	-0.007	0.084	0.076	93.9	-0.002	0.065	0.063	95.0	-0.007	0.101	0.101	95.0

Note: ν and ω are the sensitivity and specificity of a test, respectively. "Individual-based method (N)" and "Individual-based method (n)" denote the methods based on the misclassified individual current status data with sample sizes N and n, respectively.

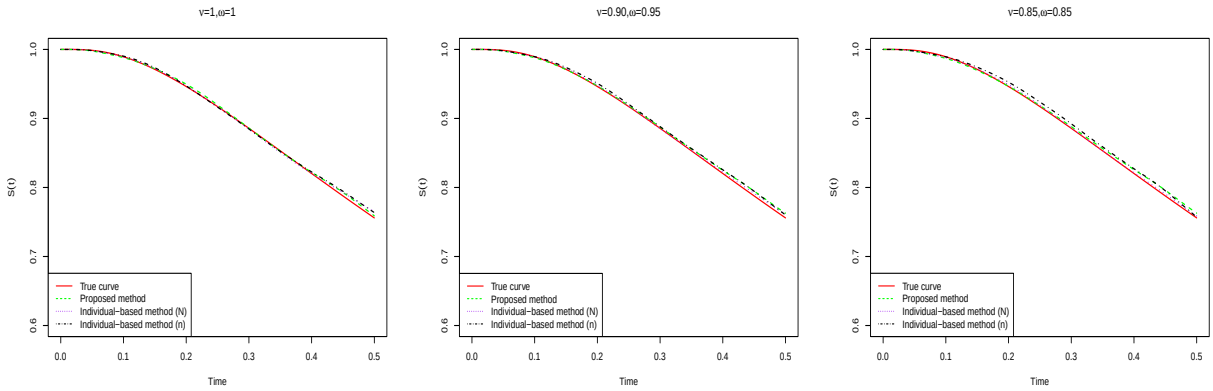

Fig. 1 Estimated baseline survival curves.

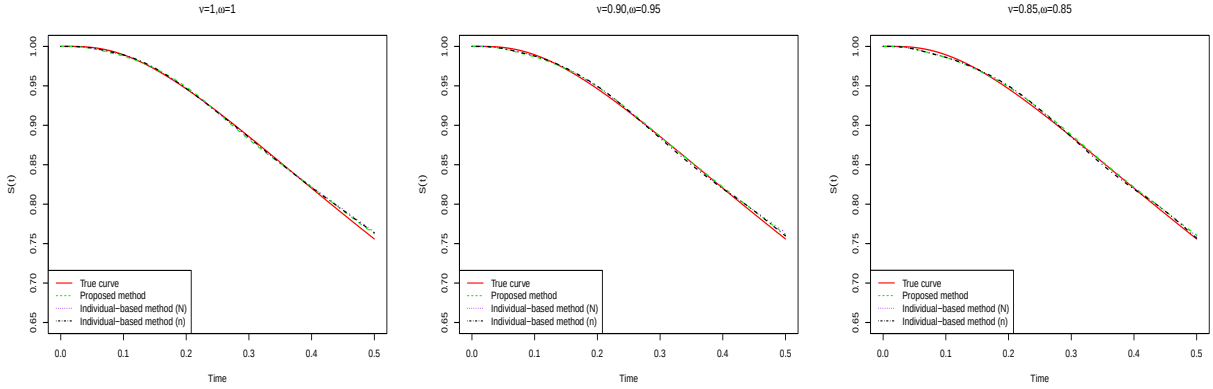

Fig. 2 Estimated baseline survival curves.

6 An Application

We apply the proposed method to a set of chlamydia data collected by the State Hygienic Laboratory (SHL) at the University of Iowa. Chlamydia is an asymptomatic, sexually transmitted disease that may cause various complications if left untreated. The SHL at the University of Iowa tests thousands of Iowa residents for monitoring chlamydia infection each year.

In the study, we focus on the chlamydia data involving $N = 13862$ female individuals collected during the 2014 calendar year. The data consist of test results taken on 2273 swab pools of size 4, 12 swab pools of size 3, 1 swab pool of size 2, 416 individual swab specimens, and 4316 individual urine specimens. The diagnostic test was performed using the Aptima

Combo 2 Assay (Hologic, San Diego) on the swab (urine) specimen of the individual, with a sensitivity and specificity of 0.942 (0.947) and 0.976 (0.989), respectively. The failure time of interest in our analysis is defined as the age when individual was infected with chlamydia, and the observation time is the age of the individual at testing. We consider the following three race categories: black, white, and other races. For our analysis, we take white race as the baseline group and adopt two dummy variables for black and other race, respectively.

To fit the proposed model, we specify the order of splines to be 2 or 3, and vary the number of equally spaced interior knots from 1 to 20 across the minimum and maximum of the observation times. We choose the optimal combination of the order and the number of interior knots by standard model selection criteria: Akaike's information criterion (AIC) and Bayesian information criterion (BIC), which are defined as

$$\text{AIC} = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) + 2(p + L_n) \text{ and } \text{BIC} = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) + (p + L_n)\log(n),$$

respectively. In the above, $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = \log\{L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})\}$, $L(\boldsymbol{\beta}, \boldsymbol{\xi})$ is given in (3), and n is the number of the groups. Through the analysis, it is turned out that the smallest AIC was given by the spline-based model with the order 3 and the number of interior knots 4, and the model with the order 3 and the number of interior knots 3 yielded the smallest BIC. Table 3 summarizes the obtained results under the optimal models selected by AIC and BIC, including the estimated regression coefficients (Est), the estimated standard errors (Std), and the corresponding p-values. It is clear from Table 3 that the two optimal models selected by AIC and BIC yield the same conclusions. In particular, the black people are more likely to get infected with chlamydia than the people belonging to the white race category. In contrast, there is no significant difference in the infection risk of contacting chlamydia between the white and other race categories.

Table 3: The analysis results of the chlamydia data

k	q_n	Race	Est	Std	p-value
3	4	`Black`	0.222	0.066	0.001
		`Other Race`	0.016	0.087	0.853
3	3	`Black`	0.224	0.066	0.001
		`Other Race`	0.013	0.087	0.883

Note: k and q_n denote the order and the number of interior knots in the spline, respectively. The white race category is the baseline for comparison.

7 Discussion and Concluding Remarks

Group-tested current status data are commonly encountered in large-scale infectious disease studies by employing group testing strategies. In this paper, we study regression analysis of such data with semiparametric probit model, a vital alternative to the popular proportional hazards model. We propose a sieve maximum likelihood estimation method that approximates the nuisance function with the logarithmic monotone splines. To facilitate the model fitting, we develop a novel EM algorithm based on three-stage data augmentation. Asymptotic properties of the resulting estimators are established by using some results of empirical process and sieve estimation theory. Numerical studies demonstrate clearly the satisfactory estimation performance and practical usefulness of the proposed method.

There are several interesting extensions that are worthwhile to investigate for future research. One direction is to extend the proposed method to accommodate group-tested current status data with retesting problems, in which if a pooled sample is tested positive, further individual decoding or retesting of the pooled sample is conducted to ascertain which individuals are positive. During the process of individual decoding, the covariate information of each individual in each positive group may be useful to help identify the individuals that are more likely to be positive (Bilder et al., 2010). Therefore, it would be important to develop informative retesting strategies to reduce the number of retesting and the cost. Also, as in Petito and Jewell (2016) and others, our method assumes that the

sensitivity and specificity of the test are constants and do not depend on the pool or group sizes of the pooled samples. However, in some applications, this assumption is unrealistic, and the sensitivity and specificity of the test may decrease when pool sizes become large, which is usually referred to as the dilution effect in the literature (McMahan et al., 2013a; Delaigle and Hall, 2015). It is worthwhile to extend the proposed method to accommodate the dilution effect, if it is present. Furthermore, the model checking techniques are useful and still lacking for the proposed probit model under the group-tested current status data. Future effort will be devoted to address this challenging problem.

Acknowledgments. This work was supported by the Nature Science Foundation of Guangdong Province of China (2022A1515011901), National Nature Science Foundation of China (12171328), Beijing Natural Science Foundation (Z210003), the National Institutes of Health (R01AI121351) and the Nature Science Foundation (OIA-1826715).

References

Baruch, J., Suanes, A., Piaggio, J., Gil, A., 2020. Analytic sensitivity of an elisa test on pooled sera samples for detection of bovine brucellosis in eradication stages in Uraguay. Frontiers in Veterinary Science, 7, 1–5.

Berger, T., Mandell, J.W., Subrahmanya, P., 2000. Maximally efficient two-stage screening. Biometrics 56, 833–840.

Bilder, C.R., Tebbs, J.M., Chen, P., 2010. Informative retesting. Journal of the American Statistical Association 105, 942–955.

Chen, P., Tebbs, J.M., Bilder, C.R., 2009. Group testing regression models with fixed and random effects. Biometrics 65, 1270–1278.

Cox, D., 1972. Regression models and life-tables. Journal of the Royal Statistical Society, Series B 34, 187–220.

Delaigle, A., Hall, P., 2012. Nonparametric regression with homogeneous group testing data. The Annals of Statistics 40, 131–158.

Delaigle, A., Hall, P., 2015. Nonparametric methods for group testing data, taking dilution into account. Biometrika 102, 871–887.

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

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

Du, M., Hu, T., Sun, J., 2019. Semiparametric probit model for informative current status data. Statistics in Medicine 38, 2219–2227.

Fahey, J.W., Ourisson, P.J., Degnan, F.H., 2006. Pathogen detection, testing, and control in fresh broccoli sprouts. Nutrition Journal 5, 1–6.

Fang, L., Li, S., Sun, L., Song, X., 2023. Semiparametric probit regression model with misclassified current status data. Statistics in Medicine 42, 4440–4457.

Farrington, C., 1992. Estimating prevalence by group testing using generalized linear models. Statistics in Medicine 11, 1591–1597.

Heffernan, A.L., Aylward, L.L., Toms, L.M.L., Sly, P.D., Macleod, M., Mueller, J.F., 2014. Pooled biological specimens for human biomonitoring of environmental chemicals: opportunities and limitations. Journal of Exposure Science and Environmental Epidemiology 24, 225–232.

Huang, J., Rossini, A., 1997. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. Journal of the American Statistical Association 92, 960–967.

Huang, J., Zhang, Y., Hua, L., 2008. A least-squares approach to consistent information estimation in semiparametric models. Technical Report. Department of Biostatistics, University of Iowa.

Huang, X., Tebbs, J.M., 2009. On latent-variable model misspecification in structural measurement error models for binary response. Biometrics 65, 710–718.

Huang, Y.T., Cai, T., 2016. Mediation analysis for survival data using semiparametric probit models. Biometrics 72, 563–574.

Li, S., Hu., T., Wang, L., McMahan, C.S., Tebbs, J.M., 2024. Regression analysis of group-tested current status data. Biometrika, doi:10.1093/biomet/asae006.

Lin, X., Wang, L., 2010. A semiparametric probit model for case 2 interval-censored failure time data. Statistics in Medicine 29, 972–981.

Lu, M., Zhang, Y., Huang, J., 2007. Estimation of the mean function with panel count data using monotone polynomial splines. Biometrika 94, 705–718.

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

McMahan, C.S., Tebbs, J.M., Bilder, C.R., 2013a. Regression models for group testing data with pool dilution effects. Biostatistics 14, 284–298.

McMahan, C.S., Wang, L., Tebbs, J.M., 2013b. Regression analysis for current status data using the EM algorithm. Statistics in Medicine 32, 4452–4466.

Petito, L.C., Jewell, N.P., 2016. Misclassified group-tested current status data. Biometrika 103, 801–815.

Ramsay, J.O., 1988. Monotone regression splines in action. Statistical Science 3, 425–441.

Remlinger, K.S., Hughes-Oliver, J.M., Young, S.S., Lam, R.L., 2006. Statistical design of pools using optimal coverage and minimal collision. Technometrics 48, 133–143.

Shiboski, S.C., 1998. Generalized additive models for current status data. Lifetime Data Analysis 4, 29–50.

Sun, J., 2006. *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer.

Vansteelandt, S., Goetghebeur, E., Verstraeten, T., 2000. Regression models for disease prevalence with diagnostic tests on pools of serum samples. Biometrics 56, 1126–1133.

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

Wu, H., Wang, L., 2019. Normal frailty probit model for clustered interval-censored failure time data. Biometrical Journal 61, 827–840.

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

Xie, M., Tatsuoka, K., Sacks, J., Young, S., 2001. Group testing with blockers and synergism. Journal of the American Statistical Association 96, 92–102.

Zeng, D., Gao, F., Lin, D., 2017. Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. Biometrika 104, 505–525.

Zeng, D., Mao, L., Lin, D., 2016. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. Biometrika 103, 253–271.

Zhang, Y., Hua, L., Huang, J., 2010. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. Scandinavian Journal of Statistics 37, 338–354.