Statistica Sinica Preprint No: SS-2024-0099	
Title	Probit Time-to-Event Regression for Misclassified Group Testing Data
Manuscript ID	SS-2024-0099
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0099
Complete List of Authors	Lijun Fang, Tao Hu, Shuwei Li, Lianming Wang, Christopher S. McMahan and Joshua M. Tebbs
Corresponding Authors	Shuwei Li
E-mails	seslishuw@gzhu.edu.cn
Notice: Accepted author version.	

PROBIT TIME-TO-EVENT REGRESSION FOR MISCLASSIFIED GROUP TESTING DATA

Lijun Fang[1], Tao Hu[1], Shuwei Li[*2], Lianming Wang[3],

Christopher S. McMahan[4], and Joshua M. Tebbs[3]

[1]*Capital Normal University*, [2]*Guangzhou University*,

[3]*University of South Carolina*, and [4]*Clemson University*

Abstract: Group testing has been used extensively to reduce screening costs in epidemiological studies involving low-prevalence diseases. This testing strategy involves 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 measured 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 in survival analysis. A sieve maximum likelihood estimation approach is developed that approximates the model's nonparametric nuisance function by using logarithmic monotone splines, and an efficient expectation-maximization algorithm

*Corresponding author.

is proposed. Asymptotic properties of the resulting estimators are investigated by using empirical process techniques and sieve estimation theory. Numerical results from simulation studies suggest our estimation methods perform 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.

Key words and phrases: Current status data, EM algorithm, Maximum likelihood estimation, Pooled testing, Sieve estimation.

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 strategy works by collecting a biological specimen (e.g., blood, urine, swab, etc.) from different individuals and pooling the 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. The sexually transmitted disease literature is replete with applications of pooling biological specimens for bacterial and viral infections (Westreich et al., 2008; Lewis et al., 2012), and, more recently, group testing garnered widespread attention in the early stages of the covid-19 pandemic (Abdalhamid et al., 2020; Pilcher et al., 2020). Other applications of group testing include DNA library screening (Berger et al., 2000), drug discovery (Xie et al., 2001), environmental monitoring (Heffernan et al., 2014), food pathogen testing (Mester et al., 2017), blood donor safety (Saá et al., 2018), and veterinary medicine (Baruch et al., 2020).

Since Dorfman's seminal work, statistical research in group testing has flourished, and a large number of regression methods have been developed for analyzing group testing data when individual covariate information is available. The first regression approach came from Farrington (1992), who estimated a specific generalized linear model under the 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) de-

veloped nonparametric approaches with a single continuous covariate and offered rigorous 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 accuracy rates of the assays used.

All of the articles cited in the previous paragraph, and many others not cited, propose regression techniques 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 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 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 (EM) algorithm to estimate the population-level survival function of the time until disease onset. These authors 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 results and conclusions similar to those with current status data from group testing. More recently, when subject-specific covariates are available, Li et al. (2024) developed an EM algorithm to estimate a proportional hazards (PH) 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 with the same number of tests.

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. Our work focuses on estimating a semiparametric probit regression model with current status responses from group testing, thereby extending previous work by authors who have considered this model with data from individual testing (Shiboski, 1998; Lin and Wang, 2010; Huang and Cai, 2016; Wu and Wang, 2019; Du et al., 2019; Fang et al., 2023). To estimate the model, we first approximate the nonparametric nuisance function with logarithmic monotone splines and propose an EM algorithm to obtain sieve maximum likelihood estimators of all model parameters. When compared to Li et al. (2024), a practical advantage of considering probit regression is its highly efficient implementation. All conditional expectations in the E-step are in closed form, and the objective function in the M-step has a tractable form making it easy to optimize. In particular, finite-dimensional spline coefficients can be quickly updated by using a Newton-Raphson algorithm, and regression parameter estimators have closed-form solutions. Adopting empirical process techniques and sieve estimation theory, maximum likelihood estimators of the regression parameters are shown to be consistent, asymptotically normal, and asymptotically efficient. Furthermore, unlike Li et al. (2024), which uses a time-consuming

resampling procedure to estimate the covariance matrix of the regression parameter estimators, we obtain variance estimates using a profile likelihood method which is straightforward to implement and computationally efficient.

Subsequent sections are organized as follows. In Section 2, we describe the data observed from a study which uses group testing, the proposed probit model for a time-to-event analysis, and the observed data likelihood. We also discuss the corresponding model assumptions and our use of monotone splines for the nuisance function in the model. In Section 3, we provide specific details on our EM algorithm, and, in Section 4, we summarize a rigorous asymptotic evaluation of the estimators. In Section 5, we present the results of a simulation study to characterize finite-sample performance. In Section 6, we illustrate our time-to-event methods using a chlamydia data set from the State Hygienic Laboratory at University of Iowa. In Section 7, we conclude with a summary discussion. Additional results and technical details are given in the Supplementary Material.

2. Model, Data, and Likelihood

Consider a study involving N individuals whose disease statuses (e.g., HIV, chlamydia, etc.) are mutually independent. We assume individual spec-

imens (e.g., blood, urine, swabs, etc.) are randomly assigned to n non-overlapping pools which are then tested for disease. Denote by J_i the size of the ith pool, for $i = 1, \ldots, n$, so that $N = \sum_{i=1}^{n} J_i$. Let T_{ij} denote the disease onset time and let $\boldsymbol{Z}_{ij}$ denote the $p \times 1$ vector of covariates for the jth individual in the ith pool. To relate T_{ij} to the covariates $\boldsymbol{Z}_{ij}$, we consider a semiparametric probit model, which specifies the conditional cumulative distribution function (cdf) of T_{ij} given $\boldsymbol{Z}_{ij}$ is

$$F(t \mid \boldsymbol{Z}_{ij}) = \Phi\{\alpha(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\}, \tag{2.1}$$

where $\Phi(\cdot)$ is the cdf of a standard normal random variable, $\boldsymbol{\beta}$ is a $p \times 1$ vector of covariate effects, and $\alpha(\cdot)$ is an increasing function with $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$. Note that the model in (2.1) arises from

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

where ε_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, J_i$ are mutually independent standard normal random variables. We note in passing that if one lets $\alpha(t) = \log(t)$ in (2.1) and allows the distribution of ε_{ij} to remain unspecified, the probit model above coincides with the popular accelerated failure time (AFT) model (Jin et al., 2003; Zeng and Lin, 2007; Chiou et al., 2015). That is, both the probit and AFT models directly relate a transformed disease onset time T_{ij} to the covariates. Our goal is to estimate (2.1) using possi-

bly misclassified current status responses from group testing. We assume throughout the covariates in $\boldsymbol{Z}_{ij}$ are not time dependent.

Let $\phi_{ij} = I(T_{ij} \leq X_{ij})$ denote the true disease status of the jth individual in the ith pool at testing time X_{ij}, where $I(\cdot)$ is the indicator function. Note that ϕ_{ij} is unobserved because individuals are pooled, and X_{ij} can be subject-specific, for example, an individual's age at testing. The true disease status of the ith pool is denoted by $\Delta_i = \max(\phi_{ij}; j = 1, \ldots, J_i)$, that is, $\Delta_i = 1$ if the ith pool contains at least one diseased individual and $\Delta_i = 0$ otherwise. To incorporate misclassification due to assay error, we let $Y_i = 1$ if the ith pool tests positively for disease, $Y_i = 0$ otherwise, and let $\nu = P(Y_i = 1 \mid \Delta_i = 1)$ and $\omega = P(Y_i = 0 \mid \Delta_i = 0)$ denote the sensitivity and the specificity, respectively, of the assay used to provide the test outcomes. We assume ν and ω are known constants (with $\nu + \omega > 1$) and do not depend on X_{ij} and $\boldsymbol{Z}_{ij}$. In practice, excellent estimates of ν and ω are usually available from assay validation experiments which are published in the infectious disease or product literature, a topic we discuss further in Section 6. We also demonstrate in Section 6 how different sets of ν and ω can be used for different pools.

The observed data for analysis consist of the group testing outcomes Y_i, the observation (testing) times X_{ij}, and the covariates $\boldsymbol{Z}_{ij}$. The observed

data likelihood is

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

where $\gamma=\nu+\omega-1$. The likelihood function $L(\boldsymbol{\beta},\alpha)$ disregards multiplicative constants which are not relevant and is derived by making a non-informative censoring assumption, that is, T_{ij} and X_{ij} are conditionally independent given $\boldsymbol{Z}_{ij}$. The expressions inside the largest parentheses are $P(Y_i=1\mid D_i)$ and $P(Y_i=0\mid D_i)$, respectively, where $D_i=\{X_{ij},\boldsymbol{Z}_{ij};\ j=1,\ldots,J_i\}$, for $i=1,\ldots,n$. Under our assumptions, it follows that

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

and $P(Y_i=1\mid D_i)=\nu-\gamma P(\Delta_i=0\mid D_i)$.

As noted in Li et al. (2024), there is no partial likelihood method available for current status data from group testing, so one must estimate $\boldsymbol{\beta}$ and $\alpha(\cdot)$ simultaneously. Because $\alpha(\cdot)$ is infinite-dimensional, we invoke an approximation for it as is common in the survival analysis literature; in particular, we use a logarithmic monotone splines approximation

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

where the b_l's are integrated spline basis functions, each non-decreasing over $(0, 1)$, and the ξ_l's are non-negative spline coefficients (Ramsay, 1988). To construct the basis functions, it is necessary to specify a sequence of q_n increasing points as interior knots and to choose the order k for the splines. One can use linear, quadratic, and cubic functions by setting $k = 1$, $k = 2$, and $k = 3$, respectively. The $L_n = q_n + k$ basis functions are determined when the interior knots and order have been specified. After approximation, the observed data likelihood can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\xi}) = \prod_{i=1}^{n} \left(\nu - \gamma \prod_{j=1}^{J_i} \left[1 - \Phi \left\{ \alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} \right\} \right] \right)^{Y_i} \times \left(1 - \nu + \gamma \prod_{j=1}^{J_i} \left[1 - \Phi \left\{ \alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} \right\} \right] \right)^{1-Y_i}, \quad (2.2)$$

where $\boldsymbol{\xi} = (\xi_l, \ldots, \xi_{L_n})^\top$ is regarded to be a vector of unknown spline coefficients. Maximizing $L(\boldsymbol{\beta}, \boldsymbol{\xi})$ or $\log L(\boldsymbol{\beta}, \boldsymbol{\xi})$ directly is terribly difficult due to its intractable form. We therefore develop an EM algorithm to determine maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$.

3. Estimation

Our estimation procedure uses three layers of data augmentation. In the first and second layer, we introduce the true pool and individual statuses, Δ_i and ϕ_{ij}, respectively, as latent random variables yielding the augmented

data likelihoods

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

and

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

respectively, where $P(Y_i \mid \Delta_i) = \{\nu^{Y_i}(1-\nu)^{1-Y_i}\}^{\Delta_i}\{(1-\omega)^{Y_i}\omega^{1-Y_i}\}^{1-\Delta_i}$ and $\Delta_i = I(\sum_{j=1}^{J_i}\phi_{ij} > 0)$, for $i = 1,\ldots,n$. In the third layer, we introduce the set of latent variables $\{G_{ij}; i = 1,\ldots,n, j = 1,\ldots,J_i\}$, where $G_{ij} = \alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} + \varepsilon_{ij}$ and ε_{ij} are mutually independent standard normal random variables, so that $P(\phi_{ij} = 1 \mid D_i) = P(G_{ij} \geq 0 \mid D_i) = \Phi\{\alpha_n(X_{ij}) + \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}\}$. Incorporating all three layers of augmentation, the complete data likelihood function is

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

with the constraints $G_{ij} \geq 0$ if $\phi_{ij} = 1$ and $G_{ij} < 0$ if $\phi_{ij} = 0$, for each i and j. Removing constants that are not relevant, the complete data log-likelihood function can be written as

$$l_c(\boldsymbol{\beta}, \boldsymbol{\xi}) =$$
$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\left[G_{ij} - \boldsymbol{\beta}^\top \boldsymbol{Z}_{ij} - \log\left\{\sum_{l=1}^{L_n}\xi_l b_l(X_{ij})\right\}\right]^2 \{\phi_{ij}1_{(G_{ij}\geq 0)} + (1-\phi_{ij})1_{(G_{ij}<0)}\},$$

where 1_A is an indicator function for the event A.

We now describe the expectation and maximization steps. In the E-step, one takes the expectation of $l_c(\boldsymbol{\beta}, \boldsymbol{\xi})$ with respect to all latent variables (Δ_i, ϕ_{ij}, and G_{ij}), conditional on the observed data $\mathcal{O}_i = \{(Y_i, X_{ij}, \boldsymbol{Z}_{ij}); i = 1, \ldots, n,\ j = 1, \ldots, J_i\}$ and current parameters $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$. Omitting unnecessary constants, this yields the objective function

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

where μ_{ij}^{+} and μ_{ij}^{-} are the expectations of G_{ij} under the constraints $G_{ij} \geq 0$ and $G_{ij} < 0$, respectively. Note that μ_{ij}^{+}, μ_{ij}^{-}, and $E(\phi_{ij})$ are really conditional expectations given $\mathcal{O}_i$, $\boldsymbol{\beta}^{(m)}$, and $\boldsymbol{\xi}^{(m)}$, but we do not emphasize this in the notation for ease of exposition. Using properties of the truncated normal distribution and Bayes' Theorem, we obtain closed-form expressions

for each expectation; these are

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

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

where $\varphi(\cdot)$ is the probability density function of a standard normal random variable and

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

Additional details on the derivation of these conditional expectations are given in the Supplementary Material. The M-step then updates $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\xi}^{(m)}$ by maximizing $Q(\boldsymbol{\beta},\boldsymbol{\xi};\boldsymbol{\beta}^{(m)},\boldsymbol{\xi}^{(m)})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. Solving $\partial Q(\boldsymbol{\beta},\boldsymbol{\xi};\boldsymbol{\beta}^{(m)},\boldsymbol{\xi}^{(m)})/\partial\boldsymbol{\beta} = \mathbf{0}$ renders a closed-form solution as a function of $\boldsymbol{\xi}$, that is,

$$\boldsymbol{\beta}^{(m+1)}(\boldsymbol{\xi}) = \left(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^{\top}\right)^{-1} \times \sum_{i=1}^{n}\sum_{j=1}^{J_i}\boldsymbol{Z}_{ij}\left[E(\phi_{ij})\mu_{ij}^{+} + \{1-E(\phi_{ij})\}\mu_{ij}^{-} - \log\left\{\sum_{l=1}^{L_n}\xi_l b_l(X_{ij})\right\}\right]. \quad (3.3)$$

Because the spline coefficients in $\boldsymbol{\xi}$ are non-negative, we reparameterize to avoid constrained optimization. Substituting $\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m+1)}(\boldsymbol{\xi})$ into

$Q(\boldsymbol{\beta}, \boldsymbol{\xi}; \boldsymbol{\beta}^{(m)}, \boldsymbol{\xi}^{(m)})$ and replacing each ξ_l with $\exp(\xi_l^*)$, the score equation for ξ_l^* is

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\Bigg[E(\phi_{ij})\mu_{ij}^{+} + \{1 - E(\phi_{ij})\}\mu_{ij}^{-} - \boldsymbol{Z}_{ij}^{\top}\boldsymbol{\beta}^{(m+1)} - \log\left\{\sum_{l=1}^{L_n}\exp(\xi_l^*)b_l(X_{ij})\right\}\Bigg]\frac{\exp(\xi_l^*)b_l(X_{ij})}{\sum_{l=1}^{L_n}\exp(\xi_l^*)b_l(X_{ij})} = 0. \quad (3.4)$$

This equation has a tractable form, so we can readily obtain $\xi_l^{*(m+1)}$ by using a simple Newton-Raphson algorithm and then calculate $\xi_l^{(m+1)} = \exp(\xi_l^{*(m+1)})$ for $l = 1, \ldots, L_n$.

Summarizing, a step-by-step description of our EM algorithm to determine maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\xi}}$ is provided below.

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

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

Step 3. Calculate $\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m+1)}(\boldsymbol{\xi}^{(m)})$ by using Equation (3.3).

Step 4. For each $l = 1, \ldots, L_n$, calculate $\xi_l^{*(m+1)}$ by solving Equation (3.4) where the other components in $\boldsymbol{\xi}^* = (\xi_1^*, \ldots, \xi_{L_n}^*)^{\top}$ are set at their mth updates. Set $\xi_l^{(m+1)} = \exp(\xi_l^{*(m+1)})$.

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

We have found that our algorithm's performance is robust to the choice of initialization. In practice, one can set the initial value of each component

in $\boldsymbol{\beta}$ to be 0 and initial values of the spline coefficients to be L_n randomly generated values from an exponential distribution with mean 1/7. Convergence is declared when the maximum of all absolute differences between two successive iterations is less than a small positive constant; e.g., $\epsilon = 10^{-4}$.

4. Asymptotic Properties

We now summarize the large-sample properties of the estimators in Section 3. 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 non-decreasing functions over $[\tau_1, \tau_2]$, with $0 < \tau_1 < \tau_2 < \infty$. Define $\boldsymbol{\Theta}_n = \{(\boldsymbol{\beta}, \alpha_n) \in \mathcal{B} \otimes \mathcal{A}_n\}$, where $\mathcal{A}_n = \{\alpha_n(t) = \log\{\sum_{l=1}^{L_n} \xi_l b_l(t)\} : \xi_l \geq 0,\ 0 \leq b_l(t) \leq 1,\ t \in [\tau_1, \tau_2]\}$, and let $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\beta}}_n, \hat{\alpha}_n)$ denote the estimator from Section 3, where $\hat{\alpha}_n(t) = \log\{\sum_{l=1}^{L_n} \hat{\xi}_l b_l(t)\}$. Note that $\hat{\boldsymbol{\theta}}_n$ is a sieve maximum likelihood estimator of $\boldsymbol{\theta}$ because it arises from maximization over the sieve space $\boldsymbol{\Theta}_n$ (Shen and Wong, 1994). In this section, we use empirical process techniques and sieve estimation theory to establish asymptotic properties of $\hat{\boldsymbol{\theta}}_n$. Following Delaigle and Meister (2011), Wang et al. (2014), and Li et al. (2024), we assume the number of pools $n \to \infty$ as the number of individuals $N \to \infty$, but where the pool sizes J_i are regarded as fixed.

Let $\|\boldsymbol{b}\|$ denote the Euclidean norm for the vector $\boldsymbol{b}$, and define the

distance between $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1, \alpha_1) \in \boldsymbol{\Theta}$ 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 $\|\alpha_1 - \alpha_2\|_2 = [\int_{\tau_1}^{\tau_2} \{\alpha_1(u) - \alpha_2(u)\}^2 \, dQ(u)]^{1/2}$ and $Q(\cdot)$ is the cdf of the observation time. Let $\mathcal{T}_n = \{t_q, q = 1, \ldots, q_n + 2k\}$ 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 collection of knots that partitions $[\tau_1, \tau_2]$ into $q_n + 1$ subintervals, where $q_n = O(n^\kappa)$ for $0 < \kappa < 1/2$. To make our large-sample arguments, we state the following regularity conditions:

(A1) The true value $\boldsymbol{\beta}_0$ is an interior point of $\mathcal{B}$. The true value $\alpha_0 \in \mathcal{A}$ is continuously differentiable, has a positive first derivative, and has a bounded rth derivative over $[\tau_1, \tau_2]$ for $r \geq 1$.

(A2) The covariate vector $\boldsymbol{Z}_j$ is bounded with probability 1.

(A3) The matrix $E(\boldsymbol{Z}_j \boldsymbol{Z}_j^\top)$ is positive definite.

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

(A5) Let $\tilde{\Delta}_{\max} = \max_{k+1 \leq q \leq q_n+k+1} |t_q - t_{q-1}|$ denote the maximum spacing between two adjacent knots. Then $\tilde{\Delta}_{\max} = O(n^{-\kappa})$ for $0 < \kappa < 1/2$, and $\tilde{\Delta}_{\max}/\tilde{\Delta}_{\min}$ is bounded, where $\tilde{\Delta}_{\min} = \min_{k+1 \leq q \leq q_n+k+1} |t_q - t_{q-1}|$.

(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 1, 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 $\boldsymbol{\eta}_j = \mathbf{0}$.

Note that conditions (A1)–(A3) are commonly assumed in the survival analysis literature; see, for example, Huang and Rossini (1997) and Zhang et al. (2010). Condition (A4) is used to ensure model identifiability (Zeng et al., 2016, 2017) and holds if the matrix $E([1, \boldsymbol{Z}_j^\top]^\top [1, \boldsymbol{Z}_j^\top])$ is nonsingular. Condition (A5) is required to establish asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ and to derive the convergence rate (Lu et al., 2007). Condition (A6) is used to prove invertibility of the efficient Fisher information matrix and holds if $E([1, \boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_J^\top]^\top [1, \boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_J^\top])$ is nonsingular.

We now present three theorems which summarize the asymptotic behavior of the sieve estimator $\hat{\boldsymbol{\theta}}_n$. Proofs are given in the Supplementary Material. In what follows, $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \alpha_0)$ denotes the true value of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha)$.

Theorem 1: Under conditions (A1)–(A4), the sieve estimator is strongly consistent, that is, $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \to 0$ and $\sup_{t \in [\tau_1, \tau_2]} |\hat{\alpha}_n(t) - \alpha_0(t)| \to 0$ almost surely as $n \to \infty$.

Theorem 2: Under conditions (A1)–(A5),

$$d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p\left(n^{-\min\{r\kappa, (1-\kappa)/2\}}\right).$$

Theorem 3: Under conditions (A1)–(A6), if $1/2(1+r) < \kappa < 1/2r$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to N(0, I^{-1}(\boldsymbol{\beta}_0))$ in distribution as $n \to \infty$, where the information matrix $I(\boldsymbol{\beta}_0)$ is given in the Supplementary Material.

Readers familiar with Li et al. (2024) will recognize the homology between our theoretical results herein and those for sieve estimators under the PH model. 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 particular, the convergence rate of the sieve estimator $\hat{\boldsymbol{\theta}}_n$ is $n^{1/3}$ when $r = 1$ and increases to $n^{2/5}$ when $r = 2$. Although Theorem 3 ensures asymptotic normality of the regression estimator $\hat{\boldsymbol{\beta}}_n$, the intractable form of the asymptotic covariance matrix $I^{-1}(\boldsymbol{\beta})$ renders it unhelpful for practical purposes, for example, writing large-sample confidence intervals for the regression parameters. We therefore adopt a numerical profile method, which has been used by Zeng et al. (2017) and others, to approximate the covariance matrix by $(n\hat{V}_n)^{-1}$, where

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

where $l_i(\boldsymbol{\beta}, \boldsymbol{\xi})$ is the log-likelihood function for the ith pool only, $\hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\xi}} \log L(\boldsymbol{\beta}, \boldsymbol{\xi})$, $L(\boldsymbol{\beta}, \boldsymbol{\xi})$ is the observed data likelihood in (2.2), and $\boldsymbol{b}^{\otimes 2} = \boldsymbol{b}\boldsymbol{b}^\top$ for the column vector $\boldsymbol{b}$. Note that $\hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}}$ can be obtained by using

the EM algorithm in Section 3 with $\boldsymbol{\beta}$ fixed, and, for each i, the gradient $\partial/(\partial\boldsymbol{\beta})\, l_i(\boldsymbol{\beta}, \hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ can be approximated by using a first-order numerical difference; i.e., the sth component of $\partial/(\partial\boldsymbol{\beta})\, l_i(\boldsymbol{\beta}, \hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ is approximated by $\{p\ell_i(\hat{\boldsymbol{\beta}}+h_n\boldsymbol{e}_s) - p\ell_i(\hat{\boldsymbol{\beta}}-h_n\boldsymbol{e}_s)\}/(2h_n)$, where $\boldsymbol{e}_s$ is a p-dimensional vector with 1 as its sth entry and 0 elsewhere, h_n is a perturbation constant with the same order as $n^{-1/2}$, and $p\ell_i(\boldsymbol{\beta}) = l_i(\boldsymbol{\beta}, \hat{\boldsymbol{\xi}}_{\boldsymbol{\beta}})$. Our simulation results in Section 5 demonstrate this approximation works well in practice.

5. Simulation Evidence

We performed two simulation studies to assess the finite-sample performance of our time-to-event estimation methods with group testing data. Results from the first study are shown in this section, and those from the second are in the Supplementary Material.

The first study considers $N = 10000$ individuals randomly assigned to pools of size five, that is, $J_i = 5$ for each $i = 1, ..., 2000$. Individual disease onset times T_{ij} are generated from the model in (2.1), where $\alpha(t) = \log(t)$, $\boldsymbol{Z}_{ij} = (Z_{ij1}, Z_{ij2})^\top$, $Z_{ij1} \sim$ Bernoulli(0.5), $Z_{ij2} \sim$ Unif$(0, 1)$, and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (0.5, -0.5)^\top$. Individual observation (testing) times X_{ij} are generated from a Unif$(0, 0.5)$ distribution, and the true individual disease statuses at the time of testing are recorded as $\phi_{ij} = I(T_{ij} \leq X_{ij})$.

These configurations provide an average right censoring rate of approximately 90%, which is consistent with our application in Section 6. The true disease status of the ith pool is $\Delta_i = I(\sum_{j=1}^{5} \phi_{ij} > 0)$, and the testing outcome Y_i is simulated as $Y_i \sim \text{Bernoulli}\{\nu - \gamma(1 - \Delta_i)\}$, where $\gamma = \nu + \omega - 1$. We consider five configurations of the assay sensitivity and specificity to allow for potentially misclassified testing outcomes, $(\nu, \omega) = (1, 1)$, $(0.95, 0.95)$, $(0.90, 0.95)$, $(0.90, 0.90)$, and $(0.85, 0.85)$, and generate 500 independent data sets for each configuration.

[Table 1 about here.]

Table 1 shows the results when using order $k = 3$ for the splines and $q_n = 5$ interior knots equally spaced within $[X_{min} - 10^{-5}, X_{max} + 10^{-5}]$, where X_{min} and X_{max} are the minimum and maximum observation times, respectively. We include the empirical bias and the sample standard deviation (SSD) of the 500 sieve maximum likelihood estimates of β_1 and β_2, along with the averaged estimated standard error (ESE) and the empirical coverage probability of nominal 95% Wald confidence intervals. For each data set, we chose initial values for the EM algorithm as described in Section 3, and we estimated standard errors using the profile likelihood method in Section 4 with $h_n = 5n^{-1/2}$, the perturbation constant used in Zeng et al. (2017). Estimating both the model in (2.1) and the large-sample

covariance matrix $I^{-1}(\boldsymbol{\beta})$ took approximately 2 minutes for each data set. This average time is for a computer that has an Intel Xeon Platinum 8375C CPU @ 2.90GHz and 128GB of RAM.

The results in Table 1 show our estimation methods work well for group testing. At all configurations of ν and ω, the empirical bias in the regression estimators is close to 0, and the ratio of SSD to ESE is consistently close to 1, suggesting that our profile likelihood approach to estimate $I^{-1}(\boldsymbol{\beta})$ performs adequately. For large-sample inference, estimated coverage probabilities of confidence intervals are all within the margin of Monte Carlo error; at the 99% confidence level, the margin of error is approximately ± 0.03.

Table 1 also includes results for individual testing, allowing one to compare group testing and individual testing when the number of individuals is fixed at $N = 10000$ (middle) and the number of tests is fixed at $n = 2000$ (right). In the former comparison, it is not surprising that individual testing is more efficient. However, the efficiency gain is small and would come at the cost of having to perform 8,000 more tests. For the latter, group testing is slightly more efficient. This comparison would be sensible if the cost of testing was the primary factor in choosing between group testing with $n = 2000$ tests and individual testing with $n = 2000$ individuals.

[Figure 1 about here.]

Finally, we show in Figure 1 estimates of the baseline survival function $S(t) = 1-\Phi\{\alpha(t)\}$ for group testing and individual testing. These estimates are averaged over 500 data sets and are shown for no misclassification ($\nu = \omega = 1$) and at the highest level of misclassification ($\nu = \omega = 0.85$). In both cases, averaged estimates of $S(t)$ under group testing are close to the true baseline survival function and are practically indistinguishable from the corresponding estimates under individual testing.

6. Application

The State Hygienic Laboratory (SHL) at the University of Iowa is the largest public health lab in Iowa. Each year, the SHL tests thousands of residents for chlamydia as part of national screening and surveillance efforts. Screening to detect positive cases is a public health imperative given the asymptomatic nature of the disease (Low, 2007) and the possible complications that could arise if the disease is left untreated (Land et al., 2010). At the same time, surveillance is also critical to understand the epidemiology of chlamydia, for example, understanding which risk factors are associated with time to disease onset. This helps state and federal organizations establish screening recommendations for the general population and for those at higher risk (LeFevre, 2014).

The SHL receives specimens every day for chlamydia testing, and, as a state-funded laboratory, it is important to be mindful of costs. The lab has implemented group testing as a cost-savings strategy to test swab specimens while urine specimens are tested individually. To illustrate our methods, we analyze a data set with $N = 13862$ female subjects whose specimens were collected during the 2014 calendar year. This data set consists of testing outcomes for 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. Thus, there are $n = 7018$ pools in total, where ostensibly we view an individual specimen to be a "pool" of size one. Our goal is to estimate a probit model for T_{ij}, the time to disease onset; i.e.,

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

where $\boldsymbol{Z}_{ij} = (Z_{ij1}, Z_{ij2})^\top$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$. The covariates Z_{ij1} and Z_{ij2} are indicator variables for race; specifically, $Z_{ij1} = 1$, if the jth subject in the ith pool is African American ($Z_{ij1} = 0$, otherwise) and $Z_{ij2} = 1$, if the jth subject in the ith pool is a race other than Caucasian or African American ($Z_{ij2} = 0$, otherwise). In other words, race in our analysis is regarded as a categorical variable with three levels, and Caucasian subjects form the baseline group for comparison. Our data set also includes the age at testing for each subject which serves as the observation time X_{ij}.

Along with most other public health labs in the United States, the SHL uses the Aptima Combo 2 Assay (AC2A, Hologic, San Diego) to test specimens for chlamydia. The AC2A product insert and Gaydos et al. (2003) summarize the results of a validation experiment and report $\nu = 0.942$ (0.947) and $\omega = 0.976$ (0.989) as the sensitivity and specificity, respectively, for swab (urine) specimens. In our analysis, we ignore sampling error associated with the validation experiment and treat these probabilities as true values for the AC2A. Our estimation framework in Section 3 can be easily adapted to use different sets of ν and ω for different specimen types.

To estimate the model, we consider $k = 2$ and $k = 3$ for the order of the splines and vary the number of interior knots q_n from 1 to 20 across the minimum and maximum of the observation times. We then select the combination of k and q_n that minimizes Akaike's information criterion (Akaike, 1974) and also the combination that minimizes the Bayesian information criterion; see, e.g., Li et al. (2017). These criteria are given by

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

where $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = \log\{L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})\}$, $L(\boldsymbol{\beta}, \boldsymbol{\xi})$ is the likelihood in (2.2), and $L_n = q_n + k$. The optimal selections of k and q_n identified under both criteria are shown in Table 2.

[Table 2 about here.]

Table 2 gives the regression estimates of β_1 and β_2, their estimated asymptotic standard errors, and Wald probability values for testing $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$, respectively. We used the numerical profile method in Section 4 to estimate standard errors with both $h_n = n^{-1/2}$ and $h_n = 5n^{-1/2}$ as perturbation constants in approximating the gradient. Estimates shown in Table 2 are nearly identical for both constants, suggesting our method is not overly sensitive to its selection. Furthermore, overall conclusions from the estimated models using $(k, q_n) = (3, 4)$ and $(k, q_n) = (3, 3)$ are the same. That is, the time to chlamydial disease onset is stochastically smaller for African American subjects when compared to Caucasian subjects. When making the same comparison with subjects of other races (e.g., Asian, American Indian, Pacific-Islander, etc.), the difference is not statistically significant.

Finally, for comparison purposes, we also estimated a PH regression model with the same covariates using the approach in Li et al. (2024). This analysis revealed the same findings and is shown in the Supplementary Material. Note that it took approximately two hours to estimate the PH model and the large-sample covariance matrix of the regression estimators. Doing the same for our probit analysis took about three minutes.

7. Discussion

Because the merger of time-to-event analysis with group testing is relatively new, it is easy to envision future research in this area—research that is both methodologically challenging and motivated by real biostatistical practice. First, although one can specify different values of the assay sensitivity ν for different pools in Section 2, one could extend this notion by positing a formal submodel which characterizes how the sensitivity depends on the pool size J and the (unknown) number of positive specimens in a pool. This idea, which acknowledges a possible dilution effect in group testing, has been implemented successfully in binary regression (McMahan et al., 2013; Delaigle and Hall, 2015). Estimating time-to-event characteristics could benefit from this extension too if dilution is suspected.

Second, to establish asymptotic properties of the sieve estimators in Section 4, we have assumed pool sizes are best regarded as fixed. This assumption is congruous with Li et al. (2024) and existing work in binary regression for group testing (Delaigle and Meister, 2011; Delaigle and Hall, 2012; Delaigle et al., 2014; Wang et al., 2014; Delaigle and Hall, 2015; Chatterjee and Bandyopadhyay, 2020). However, it might be of interest to relax this assumption if pool sizes are allowed to change throughout the course of data collection, similar to what Hughes-Oliver and Swallow (1994)

proposed when estimating a population prevalence. We are unsure how this generalization would affect arguments needed to establish our asymptotic properties in Section 4; however, they could become markedly more difficult. For example, Taylor series expansions under the fixed pool size assumption, shown in the Supplementary Material, might have non-ignorable remainder terms if pool sizes were viewed to be random.

Finally, many laboratories are now using multiplex assays to test for multiple diseases at once. In fact, such assays are currently available for chlamydia and gonorrhea (Hou et al., 2017), SARS-CoV-2 and influenza A/B (Neopane et al., 2022), HIV, HBV, and HCV (Stramer et al., 2013), and other disease combinations. Multiplex technology naturally motivates the development of multivariate current status regression methods which allow for group testing. In terms of estimation efficiency, one would expect these methods to be preferred to marginal modeling when unobserved disease onset times on the same individual are correlated.

Supplementary Material

The online Supplementary Material contains conditional expectation derivations, detailed proofs for Theorems 1-3, a second simulation study, and an analysis of the Iowa SHL data under the PH model. R code for data analysis

is available at `https://github.com/lishuwstat/GTEMProbit`.

Acknowledgements

We are grateful to two anonymous referees who provided helpful comments on an earlier version of this article. This research is supported by Grants 12471251 and 12171328 from the National Nature Science Foundation of China and and Grant Z210003 from the Beijing Natural Science Foundation. Authors at institutions in the United States are supported by the National Institutes of Health and the National Science Foundation.

References

Abdalhamid, B., Bilder, C., McCutchen, E., Hinrichs, S., Koepsell, S., and Iwen, P. (2020). Assessment of specimen pooling to conserve SARS CoV-2 testing resources. *American Journal of Clinical Pathology*, **6**:715–718.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**:716–723.

Baruch, J., Suanes, A., Piaggio, J., and 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., and Subrahmanya, P. (2000). Maximally efficient two-stage screening.

Biometrics, **56**:833–840.

Chatterjee, A. and Bandyopadhyay, T. (2020). Regression models for group testing: Identifiability and asymptotics. *Journal of Statistical Planning and Inference*, **204**:141–152.

Chen, P., Tebbs, J., and Bilder, C. (2009). Group testing regression models with fixed and random effects. *Biometrics*, **65**:1270–1278.

Chiou, S., Kang, S., and Yan, J. (2015). Rank-based estimating equations with general weight for accelerated failure time models: An induced smoothing approach. *Statistics in Medicine*, **34**:1495–1510.

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

Delaigle, A. and Hall, P. (2012). Nonparametric regression with homogeneous group testing data. *Annals of Statistics*, **40**:131–158.

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

Delaigle, A., Hall, P., and Wishart, J. (2014). New approaches to non- and semi-parametric regression for univariate and multivariate group testing data. *Biometrika*, **101**:567–585.

Delaigle, A. and 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. *Annals of*

Mathematical Statistics, **14**:436–440.

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

Fang, L., Li, S., Sun, L., and 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.

Gaydos, C., Quinn, T., Willis, D., Weissfeld, A., Hook, E., Martin, D., Ferrero, D., and Schachter, J. (2003). Performance of the APTIMA Combo 2 Assay for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in female urine and endocervical swab specimens. *Journal of Clinical Microbiology*, **41**:304–309.

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

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

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

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

Hughes-Oliver, J. and Swallow, W. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association*, **89**:982–993.

Jin, Z., Lin, D., Wei, L., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, **90**:341–353.

Land, J., Van Bergen, J., Morre, S., and Postma, M. (2010). Epidemiology of *Chlamydia trachomatis* infection in women and the cost-effectiveness of screening. *Human Reproduction Update*, **16**:189–204.

LeFevre, M. (2014). Screening for chlamydia and gonorrhea: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, **161**:902–910.

Lewis, J., Lockary, V., and Kobic, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for *Chlamydia trachomatis* and *Neisseria gonorrhoeae*. *Sexually Transmitted Diseases*, **39**:46–48.

Li, S., Hu, T., Wang, L., McMahan, C., and Tebbs, J. (2024). Regression analysis of group-tested current status data. *Biometrika*, **111**:1047–1061.

Li, S., Hu, T., Wang, P., and Sun, J. (2017). Regression analysis of current status data in the

presence of dependent censoring with applications to tumorigenicity experiments. *Computational Statistics and Data Analysis*, **110**:75–86.

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

Low, N. (2007). Screening programmes for chlamydial infection: When will we ever learn? *BMJ*, **334**:725–728.

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

McMahan, C., Tebbs, J., and Bilder, C. (2013). Regression models for group testing data with pool dilution effects. *Biostatistics*, **14**:284–298.

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

Mester, P., Witte, A., Robben, C., Streit, E., Fister, S., Schoder, D., and Rossmanith, P. (2017). Optimization and evaluation of the qPCR-based pooling strategy DEP-pooling in dairy production for the detection of Listeria monocytogenes. *Food Control*, **82**:298–304.

Neopane, P., Nypaver, J., Shrestha, R., and Beqaj, S. (2022). Performance evaluation of TaqMan SARS-CoV-2, Flu A/B, RSV RT-PCR multiplex assay for the detection of respiratory viruses. *Infection and Drug Resistance*, **15**:5411–5423.

Petito, L. and Jewell, N. (2016). Misclassified group-tested current status data. *Biometrika*,

103:801–815.

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

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, **3**:425–441.

Saá, P., Proctor, M., Foster, G., Krysztof, D., Winton, C., Linnen, J., Gao, K., Brodsky, J., Limberger, R., Dodd, R., and Stramer, S. (2018). Investigational testing for Zika virus among US blood donors. *New England Journal of Medicine*, **378**:1778–1788.

Shen, X. and Wong, W. (1994). Convergence rate of sieve estimates. *Annals of Statistics*, **22**:580–615.

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

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

Vansteelandt, S., Goetghebeur, E., and 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., and Kulasekera, K. (2014). Semiparametric group testing regression models. *Biometrika*, **101**:587–598.

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

Wu, H. and 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., and Young, S. (2001). Group testing with blockers and synergism. *Journal of the American Statistical Association*, **96**:92–102.

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

Zeng, D. and Lin, D. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, **102**:1387–1396.

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

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

Table 1: Simulation study. Empirical bias (Bias) and sample standard deviation (SSD) of 500 sieve maximum likelihood estimates. The averaged estimated standard error (ESE) and empirical coverage probabilities (CP) are also included. The second and third columns show results for individual testing (IT) when fixing the number of individuals ($n = 10000$) and the number of tests ($n = 2000$), respectively.

		Group testing				IT, $n = 10000$				IT, $n = 2000$			
(ν, ω)		Bias	SSD	ESE	CP	Bias	SSD	ESE	CP	Bias	SSD	ESE	CP
$(1, 1)$	$\hat{\beta}_1$	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

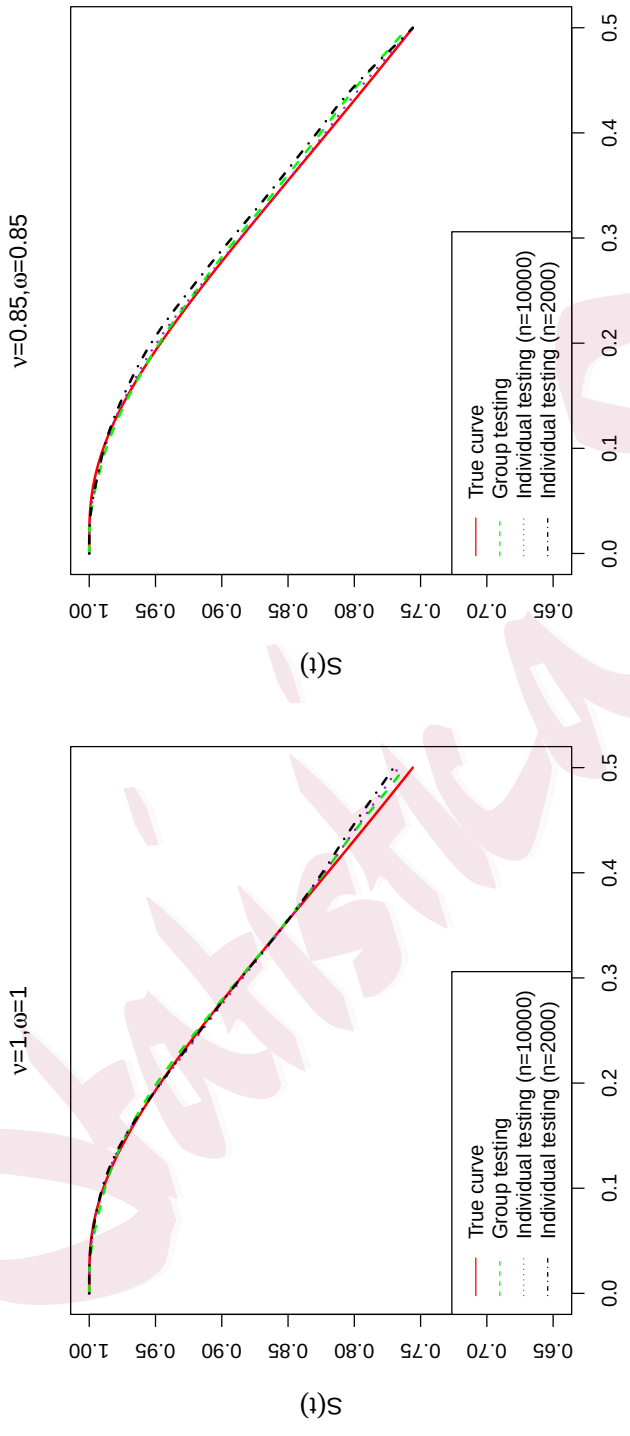

Figure 1: Simulation study. Estimated baseline survival curves for group testing and individual testing. Left: No misclassification. Right: Misclassification with sensitivity $\nu = 0.85$ and specificity $\omega = 0.85$.

Table 2: Iowa data analysis. Regression parameter estimates, estimated standard errors (SE) and probability values. Optimal combinations of k and q_n are shown when using AIC and BIC as model selection criteria. Different perturbation constants h_n are used to estimate standard errors.

					$h_n = n^{-1/2}$		$h_n = 5n^{-1/2}$	
Criterion	k	q_n	Parameter	Estimate	SE	p-value	SE	p-value
AIC	3	4	β_1	0.222	0.061	<0.001	0.066	0.001
			β_2	0.016	0.083	0.850	0.087	0.853
BIC	3	3	β_1	0.224	0.061	<0.001	0.066	0.001
			β_2	0.013	0.083	0.880	0.087	0.883