Fall November 2014

Data Analysis And Study Design In The Presence Of Error-prone Diagnostic Tests

Xiangdong Gu
University of Massachusetts Amherst

DATA ANALYSIS AND STUDY DESIGN IN THE PRESENCE OF ERROR-PRONE DIAGNOSTIC TESTS

A Dissertation Presented

by

XIANGDONG GU

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2014

Department of Public Health

DATA ANALYSIS AND STUDY DESIGN IN THE PRESENCE OF ERROR-PRONE DIAGNOSTIC TESTS

A Dissertation Presented

by

XIANGDONG GU

Approved as to style and content by:

Raji Balasubramanian, Chair

Andrea S. Foulkes, Member

Mahlet G. Tadesse, Member

Edward J. Stanek III, Department Chair
Department of Public Health

To Zefeng and Junyu

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my advisor and committee chair, Dr. Raji Balasubramanian, for her continuous support, guidance, and patience on my dissertation. I would like to thank my committee members, Dr. Andrea S. Foulkes and Dr. Mahlet G. Tadesse, for their constructive feedbacks and thoughtful comments that have greatly improved my research work. I would also like to thank Dr. David Shapiro, Dr. Michael D. Hughes, Dr. Yunsheng Ma, Christine Foley, and Hui Xu, with whom I have enjoyed working together in different projects.

I would like to thank Dr. Edward J. Stanek III for his help throughout my time in the department of Public Health and his great seminar classes full of interesting statistical topics and insightful discussions. I would also like to thank Gloria Seaman and Diane Wolf for their timely helps from my admission to graduation, which make my life in the department of Public Health much easier.

Lastly, I want to thank the Women's Health Initiative for providing me their data access, which is used throughout my dissertation work.

ABSTRACT

DATA ANALYSIS AND STUDY DESIGN IN THE PRESENCE OF ERROR-PRONE DIAGNOSTIC TESTS

SEPTEMBER 2014

XIANGDONG GU, B.Sc., UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Raji Balasubramanian

Interval censored time to event outcomes arise when a silent event of interest is known to have occurred within a specific time period, determined by the times of the last negative and first positive diagnostic tests. The four chapters comprising this thesis are tied together by a common theme in that the outcome of interest is an interval censored time to event random variable.

In Chapter 1, we describe a stratified Weibull model appropriate for interval censored outcomes and implement a new R package *straweib*. We compare the proposed approach with the log-linear form of the Weibull regression model that is currently implemented in the existing R package *survival*, and illustrate its use by analyzing data from a longitudinal oral health study on the timing of the emergence of permanent teeth in 4430 children.

In Chapter 2, we present methods to estimate the association of one or more covariates with an error-prone, self reported time to event outcome. We present simulation studies to assess the effect of error in self reported outcomes with regard to bias in the estimation of the regression parameter of interest. We apply the proposed methods to the data from Women's Health Initiative (WHI) to evaluate the effect of statin use with respect to incident diabetes risk.

In Chapter 3, we develop tools to calculate power and sample size for studies in which data from sequentially administered, error-prone, laboratory-based diagnostic tests or self-reported questionnaires are collected to determine the occurrence of a silent event. We evaluate the effects of the characteristics of the imperfect diagnostic test on resulting power and sample size calculations. We compare the relative efficiency of various study designs in the context of error-prone outcomes.

In Chapter 4, we propose a lasso and a Bayesian variable selection approach in the context of error-prone self reported outcomes to address the problem of variable selection in high dimensional data settings. We perform simulation studies to compare prediction performance of proposed methods and naive methods that ignore measurement error. We apply our proposed methods to the genome-wide association study data from the WHI to select biomarkers associated with diabetes.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

STRATIFIED WEIBULL REGRESSION MODEL FOR INTERVAL-CENSORED DATA

1.1 Introduction

In many clinical studies, the time to a silent event is known only up to an interval defined by the times of the last negative and first positive diagnostic test. Event times arising from such studies are referred to as 'interval-censored' data. For example, in pediatric HIV clinical studies, the timing of HIV infection is known only up to the interval from the last negative to the first positive HIV diagnostic test [10]. Examples of interval-censored outcomes can also be found in many other medical studies [19].

A rich literature exists on the analysis of interval-censored outcomes. Non-parametric approaches include the self-consistency algorithm for the estimation of the survival function [53]. A semi-parametric approach based on the proportional hazards model has been developed for interval-censored data [13, 18]. A variety of parametric models can also be used to estimate the distribution of the time to the event of interest, in the presence of interval-censoring [33]. An often used parametric approach for the analysis of interval-censored data is based on the assumption of a Weibull distribution for the event times [33]. The Weibull distribution is appropriate for modeling event times when the hazard function can be reliably assumed to be monotone. Covariate effects can be modeled through the assumption of proportional hazards (PH), which assumes that the ratio of hazard functions when comparing individuals in different strata defined by explanatory variables is time-invariant. The article by [19] presents a comprehensive review of the state-of-the-art techniques available for the analysis of interval-censored data.

In this chapter, we implement a parametric approach for modeling covariates applicable to interval-censored outcomes, but where the assumption of proportional hazards may be questionable for a certain subset of explanatory variables. For this setting, we implement a stratified Weibull model by relaxing the PH assumption across levels of a subset of explanatory variables. We compare the proposed model to an alternative stratified Weibull regression model that is currently implemented in the R package *survival* [50]. We illustrate the difference between these two models analytically and through simulation.

The chapter is organized as follows: In Section 1.2, we present and compare two models for relaxing the PH assumption, based on the assumption of a Weibull distribution for the time to event of interest. In this section, we discuss estimation of the unknown parameters of interest, hazard ratios comparing different groups of subjects based on specific values of explanatory covariates and tests of the PH assumption. These methods are implemented in a new R package, *straweib* [21]. In Section 1.3, we perform simulation studies to compare two stratified Weibull models implemented in R packages *straweib* and *survival*. In Section 1.4, we illustrate the use of the R package *straweib* by analyzing data from a longitudinal oral health study on the timing of the emergence of permanent teeth in 4430 children in Belgium [31, 19]. In Section 1.5, we discuss the models implemented in this chapter and present concluding remarks.

1.2 Weibull regression models

Let T denote the continuous, non-negative random variable corresponding to the time to event of interest, with corresponding probability distribution function (pdf) and cumulative distribution function (cdf), denoted by $f(t)$ and $F(t)$, respectively. We let $S(t) = 1 - F(t)$ to denote the corresponding survival function and $h(t) = lim_{\delta t \to 0} \frac{P(t \leq T < t+\delta t | T \geq t)}{\delta t}$ to denote the hazard function. We let $\mathbf{Z}$ denote the $p \times 1$ vector of explanatory variables or covariates.

We assume that the random variable $T \mid \boldsymbol{Z} = \mathbf{0}$ is distributed according to a Weibull distribution, with scale and shape parameters denoted by λ and γ, respectively. The well known PH model to accommodate the effect of covariates on T is expressed as:

$$h(t \mid \boldsymbol{Z}) = h(t \mid \boldsymbol{Z} = \mathbf{0}) \times \exp(\boldsymbol{\beta}'\boldsymbol{Z}),$$

where $\boldsymbol{\beta}$ denotes the $p \times 1$ vector of regression coefficients corresponding to the vector of explanatory variables, $\mathbf{Z}$.

Thus, under the Weibull PH model, the survival and hazard functions corresponding to T can be expressed as

$$S(t \mid \boldsymbol{Z}) = \exp\left(-\lambda \exp\left(\boldsymbol{\beta}'\boldsymbol{Z}\right) t^{\gamma}\right) \tag{1.1}$$

$$h(t \mid \boldsymbol{Z}) = \lambda \exp(\boldsymbol{\beta}'\boldsymbol{Z})\gamma t^{\gamma-1} \tag{1.2}$$

where, $\lambda > 0$ and $\gamma > 0$ correspond to the scale and shape parameters corresponding to T when $\boldsymbol{Z} = \mathbf{0}$. The hazard ratio comparing two individuals with covariate vectors $\mathbf{Z}$ and $\boldsymbol{Z}^*$ is equal to $\exp(\boldsymbol{\beta}'(\boldsymbol{Z} - \boldsymbol{Z}^*))$.

1.2.1 Stratified Weibull regression model implemented in the R package survival

In this section, we describe the stratified Weibull PH regression model implemented in the the R package *survival* ([50]).

Consider the following log-linear model for the random variable T:

$$log(T \mid \boldsymbol{Z}) = \mu + \alpha_1 Z_1 + \cdots \alpha_p Z_p + \sigma\epsilon$$

where, $\alpha_1, \cdots, \alpha_p$ denote unknown regression coefficients corresponding to the p dimensional vector of explanatory variables, μ denotes the intercept and σ denotes the scale parameter. The random variable ϵ captures the random deviation of event times

on the natural logarithm scale (i. e. $\log(T)$) from the linear model as a function of the covariate vector $\mathbf{Z}$. In general, the log-linear form of the model for T can be shown to be equivalent to the accelerated failure time (AFT) model ([5]).

The assumption of a standard Gumbel distribution with location and scale parameters equal to 0 and 1, respectively, implies that the random variable T follows a Weibull distribution. Moreover, in this case, both the PH and AFT assumptions (or equivalently, the log-linear model) lead to identical models with different parameterizations ([5]). The survival and hazard functions can be expressed as:

$$S(t \mid \boldsymbol{Z}) = \exp\left[-\exp\left(\frac{\log(t) - \mu - \boldsymbol{\alpha}^{'}\boldsymbol{Z}}{\sigma}\right)\right] \tag{1.3}$$

$$h(t \mid \boldsymbol{Z}) = \exp\left[-\frac{\mu + \boldsymbol{\alpha}^{'}\boldsymbol{Z}}{\sigma}\right]\frac{1}{\sigma}t^{\frac{1}{\sigma}-1} \tag{1.4}$$

The coefficients for the explanatory variables ($\boldsymbol{\beta}$) in the hazard function ($h(t \mid \boldsymbol{Z})$) are equal to $-\frac{\boldsymbol{\alpha}}{\sigma}$. Moreover, there is a one-to-one correspondence between the parameters $\lambda, \gamma, \boldsymbol{\beta}$ in equations (1.1)-(1.2) and the parameters $\mu, \sigma, \boldsymbol{\alpha}$ in equations (1.3)-(1.4), where $\lambda = \exp(-\frac{\mu}{\sigma}), \gamma = \sigma^{-1}$ and $\beta_j = -\frac{\alpha_j}{\sigma}$ ([5]).

The log-linear form of the Weibull model can be generalized to allow arbitrary baseline hazard functions within subgroups defined by a stratum indicator $S = 1, \cdots, s$. Thus, the stratified Weibull regression model for an individual in the j^{th} stratum is expressed as:

$$\log\left(T \mid \boldsymbol{Z}, S = j\right) \;\; = \;\; \mu_j + \alpha_1 Z_1 + \cdots \alpha_p Z_p + \sigma_j \epsilon$$

where μ_j and σ_j denote stratum specific intercept and scale parameters. This model is implemented in the R package *survival* ([50]). In this model, the regression coefficients $\boldsymbol{\alpha}$ on the AFT scale are assumed to be stratum independent.

However, the hazard ratio comparing two individuals with covariate vectors and stratum indicators denoted by $(\mathbf{Z},\, S=j)$ and $(\boldsymbol{Z}^{*}, S=k)$ is stratum specific and is given by:

$$\frac{h\left(t \mid S=j, \boldsymbol{Z}\right)}{h\left(t \mid S=k, \boldsymbol{Z}^{*}\right)} = t^{1/\sigma_j - 1/\sigma_k} \frac{\sigma_k}{\sigma_j} \exp\left(\frac{\mu_k}{\sigma_k} - \frac{\mu_j}{\sigma_j}\right) \exp\left(\boldsymbol{\alpha}^{'}\left(\boldsymbol{Z}^{*}/\sigma_k - \boldsymbol{Z}/\sigma_j\right)\right)$$

For $j \neq k$, the hazard ratio varies with time t. However, when $j = k$, the hazard ratio comparing two individuals within the same stratum $S = j$ is invariant with respect to time t but is stratum-dependent and reduces to:

$$\frac{h\left(t \mid S=j, \boldsymbol{Z}\right)}{h\left(t \mid S=j, \boldsymbol{Z}^{*}\right)} = \exp\left(\frac{\boldsymbol{\alpha}^{'}}{\sigma_j}\left(\boldsymbol{Z}^{*} - \boldsymbol{Z}\right)\right) \tag{1.5}$$

1.2.2 Stratified Weibull regression model implemented in R package straweib

In this section, we describe the stratified Weibull regression model that is implemented in the new R package, *straweib* ([21]).

To relax the proportional hazards assumption in the Weibull regression model, we propose the following model for an individual in the stratum $S = j$:

$$h(t \mid \boldsymbol{Z}, S=j) = \lambda_j \exp\left(\boldsymbol{\beta}'\boldsymbol{Z}\right) \gamma_j t^{\gamma_j - 1} \tag{1.6}$$

Equivalently, the model can be stated in terms of the survival function as:

$$S(t \mid \boldsymbol{Z}, S=j) = \exp\left(-\lambda_j \exp\left(\boldsymbol{\beta}'\boldsymbol{Z}\right) t^{\gamma_j}\right)$$

Here, we assume that the scale and shape parameters (λ, γ) are stratum specific - however, the regression coefficients $\boldsymbol{\beta}$ are assumed to be constant across strata

(S). The hazard ratio comparing two individuals with covariate vectors and stratum indicators denoted by $(\mathbf{Z},\, S = j)$ and $(\boldsymbol{Z^*}, S = k)$ is given by:

$$\frac{h(t \mid S = j, \boldsymbol{Z})}{h(t \mid S = k, \boldsymbol{Z}^*)} = t^{\gamma_j - \gamma_k} \exp\left(\boldsymbol{\beta}' \left(\boldsymbol{Z} - \boldsymbol{Z}^*\right)\right) \frac{\lambda_j \gamma_j}{\lambda_k \gamma_k}$$

For $j \neq k$, the hazard ratio varies with time t and thus relaxes the PH assumption. However, for $j = k$, the hazard ratio comparing two individuals within the same stratum $S = j$ reduces to:

$$\frac{h(t \mid S = j, \boldsymbol{Z})}{h(t \mid S = j, \boldsymbol{Z}^*)} = \exp\left(\boldsymbol{\beta}' \left(\boldsymbol{Z} - \boldsymbol{Z}^*\right)\right) \tag{1.7}$$

This hazard ratio is invariant with respect to time t and stratum S, as in the stratified Cox model [5].

Estimation:

Let $u_j = \log(\lambda_j)$ and $v_j = \log(\gamma_j)$. Let n_j denote the number of subjects in stratum $S = j$. For the k^{th} subject in stratum j, let $\boldsymbol{Z}_{jk}$ denote the p dimensional vector of covariates and let a_{jk} and b_{jk} denote the left and right endpoints of the censoring interval. That is, a_{jk} denotes the time of the last negative test and b_{jk} denotes the time of the first positive test for the event of interest. Then the log-likelihood function can be expressed as:

$$\begin{aligned} l(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\beta}) = \; & \textstyle\sum_{j=1}^{s} \sum_{k=1}^{n_j} \log\{\exp[-\exp[u_j + \boldsymbol{\beta}' \boldsymbol{Z}_{jk} + \exp(v_j)\log(a_{jk})]] \\ & - \exp[-\exp[u_j + \boldsymbol{\beta}' \boldsymbol{Z}_{jk} + \exp(v_j)\log(b_{jk})]]\} \end{aligned}$$

The unknown parameters to be estimated are $\boldsymbol{v}$, $\boldsymbol{u}$, and $\boldsymbol{\beta}$. The log-likelihood function can be optimized using the **optim** function in R. The shape and scale parameters can be estimated from the estimates of $\boldsymbol{v}$ and $\boldsymbol{u}$. The covariance matrix of the estimates of these unknown parameters can be obtained by inverting the negative Hessian matrix that is output from the optimization routine ([8]).

Test of the PH assumption:

One can test whether or not the baseline hazard functions of each strata are proportional to each other, by testing the equality of shape parameters across strata $S = 1, \cdots, s$. That is,

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_s$$

or equivalently,

$$H_0 : v_1 = v_2 = \cdots = v_s.$$

The null hypothesis H_0 can be tested using a likelihood ratio test, by comparing a reduced model that assumes that $\gamma_1 = \gamma_2 == \cdots = \gamma_s$ to the full model in (1.6) assuming stratum specific shape parameters. We note that the reduced model is equivalent to the Weibull PH model that includes the stratum indicator S as an explanatory variable. Thus the reduced model has $s - 1$ fewer parameters than the stratified model, or the full model. Let l_F and l_R denote the log-likelihoods of the full and reduced models evaluated at their MLE. Then the test statistic $T = -2(l_R - l_F)$ follows a χ^2_{s-1} distribution under H_0. In addition to the likelihood ratio test, one can also use a Wald test to test the null hypothesis H_0. The R package *strawieb* outputs both the Wald and Likelihood Ratio test statistics.

Estimating hazard ratios:

The log hazard ratio comparing two individuals with covariate vectors and stratum indicators denoted by ($\mathbf{Z}$, $S = j$) and ($\boldsymbol{Z}^*, S = j^*$) at time t can be expressed as:

$$r_{tjj^*} = \log\left(R_{tjj^*}\right) = u_j + v_j + \log\left(t\right)\exp\left(v_j\right) - u_{j^*} - v_{j^*} - \log\left(t\right)\exp\left(v_{j^*}\right) + \boldsymbol{\beta}^{'}\left(\boldsymbol{Z} - \boldsymbol{Z}^*\right)$$

Let $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimates for $\boldsymbol{v}$, $\boldsymbol{u}$ and $\boldsymbol{\beta}$, then r_{tjj^*} can be estimated by

$$\hat{r}_{tjj^*} = \hat{u}_j + \hat{v}_j + \log\left(t\right)\exp\left(\hat{v}_j\right) - \hat{u}_{j^*} - \hat{v}_{j^*} - \log\left(t\right)\exp\left(\hat{v}_{j^*}\right) + \hat{\boldsymbol{\beta}}^{'}\left(\boldsymbol{Z} - \boldsymbol{Z}^*\right)$$

Let $\boldsymbol{w} = (\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\beta}) = (v_1, v_2, \cdots, v_s, u_1, u_2, \cdots, u_s, \beta_1, \cdots, \beta_p)$. Let $\widehat{\Sigma}$ denote the estimate of the covariance matrix of $\hat{\boldsymbol{w}}$. Let $\boldsymbol{J}_{tjj^*}$ denote the Jacobian vector, $\boldsymbol{J}_{tjj^*} = \frac{\partial r_{tjj^*}}{\partial \boldsymbol{w}}|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$. Thus, the estimate of the variance of $\hat{r}_{tjj^*}$ is obtained by:

$$\widehat{Var}(\hat{r}_{tjj^*}) = \boldsymbol{J}_{tjj^*}^T \widehat{\Sigma} \boldsymbol{J}_{tjj^*}$$

We obtain a 95% confidence interval for r_{tjj^*} as $\left(\hat{r}_{tjj^*} - 1.96\sqrt{\widehat{Var}(\hat{r}_{tjj^*})}, \hat{r}_{tjj^*} + 1.96\sqrt{\widehat{Var}(\hat{r}_{tjj^*})}\right)$. We exponentiate $\hat{r}_{tjj^*}$ and its corresponding 95% confidence interval to obtain the estimate and the 95% confidence interval for the hazard ratio, R_{tjj^*}. We illustrate the use of the *straweib* R package for obtaining hazard ratios and corresponding confidence intervals in Section 1.4.

1.3 Comparison of models implemented in the R packages survival and straweib

In this section, we compare the stratified Weibull regression model implemented in the *survival* package to that implemented in our package, *straweib*.

In the absence of stratification, both models are identical and reduce to the Weibull PH model. However, in the presence of a stratification factor, the models implemented by *survival* and *straweib* correspond to different models, resulting in different likelihood functions and inference. As we discussed in Section 1.2, the hazard ratio between two subjects with different covariate values within the same stratum depends on their stratum in the model implemented in the R package *survival* (Equation (1.5)), whereas the hazard ratio comparing two individuals within the same stratum is invariant to stratum in the model implemented in the R package *straweib* (Equation (1.7)). In particular, the Weibull model implemented in the *straweib* shares similarities with the semi-parametric, stratified Cox model for right censored data.

To illustrate the difference between the models implemented in the R packages *survival* and *straweib*, we conducted a simulation study in which 1000 datasets were simulated under the model assumed in the *straweib* package (Equation (1.6)). For each simulated dataset, since both models have the same number of unknown parameters, we compare the values of the log-likelihood evaluated at the MLEs. Datasets were simulated based on the assumptions that there are 3 strata, each with a 100 subjects; the shape parameters ($\boldsymbol{\gamma}$) in the three strata were set to 1.5, 2, and 1, respectively; the baseline scale parameters in the three strata ($\boldsymbol{\lambda}$) were set to 0.01, 0.015, and 0.02, respectively. We assumed that there are two independent explanatory variables available for each subject, randomly drawn from $N(0, 1)$ random variables. The coefficients corresponding to each of the two covariates were set to 0.5 and 1, respectively. To simulate interval censored outcomes, we first simulated the true event time for each subject by sampling from a Weibull distribution with the appropriate parameters. We assumed that each subject has 20 equally spaced diagnostic tests, at which the true event status is observed. Each test has a 70% probability being missing. To obtain the maximum likelihood estimates under each model, we used the **survreg** function in the R package *survival* and the **icweib** function in the *straweib* package.

Figure 1.1 compares the maximized value of the log-likelihoods under both models, when the data are generated using a simulation mechanism that corresponds to the model implemented in the R package *straweib*. The maximized value of the log-likelihood from the R package *survival* is lower than that from the R package *straweib* for 93.1% of simulated datasets. This is expected as in this simulation study the data generating mechanism is identical to the model implemented in the R package *straweib*. In applications where the proportional hazards assumption is questionable, we recommend fitting both models and comparing the resulting maximized values of the log likelihood. Whether one model is better than another depends on the data.

1.4 Example

We illustrate the R package *straweib* with data from a study on the timing of emergence of permanent teeth in Flemish children in Belgium [31]. The data analyzed were from the Signal-Tandmobiel project [54], a longitudinal oral health study in a sample of 4430 children conducted between 1996 and 2001. Dental examinations were conducted annually for a period of 6 years and tooth emergence was recorded based on visual inspection. As in [19], we will illustrate our R package by analyzing the timing of emergence of the permanent upper left first premolars. As dental exams were conducted annually, for each child, the timing of tooth emergence is known up to the interval from the last negative to the first positive dental examination.

```
data(tooth24)
head(tooth24)
```

```
  id left right sex dmf
1  1  2.7   3.5   1   1
2  2  2.4   3.4   0   1
3  3  4.5   5.5   1   0
4  4  5.9   Inf   1   0
5  5  4.1   5.0   1   1
6  6  3.7   4.5   0   1
```

The dataset is formatted to include 1 row per child. The variable denoted **id** corresponds to the ID of the child, **left** and **right** correspond to the left and right endpoints of the censoring interval in years, **sex** denotes the gender of the child (0 = boy, and 1 = girl), and **dmf** denotes the status of primary predecessor of the tooth (0 = sound, and 1 = decayed or missing due to caries or filled). Right censored observations are denoted by setting the variable **right** to "Inf".

In our analysis below, we use the function **icweib** in the package *straweib* to fit a stratified Weibull regression model, where the variable **dmf** is the stratum indicator (S) and the variable **sex** is an explanatory variable (Z).

```
fit <- icweib(L = left, R = right, data = tooth24, strata = dmf,
  covariates = ~sex)
```

```
fit

Total observations used: 4386. Model Convergence: TRUE

Coefficients:
    coefficient     SE    z p.value
sex       0.331 0.0387 8.55       0

Weibull parameters - gamma(shape), lambda(scale):
 straname strata gamma   lambda
      dmf      0  5.99 1.63e-05
      dmf      1  4.85 1.76e-04

Test of proportional hazards for strata
  (H0: all strata's shape parameters are equal):
             test TestStat df  p.value
             Wald     44.2  1 2.96e-11
 Likelihood Ratio     44.2  1 3.00e-11

Loglik(model)=  -5501.781   Loglik(reduced)=  -5523.87
Loglik(null)=  -5538.309  Chisq= 73.05611   df= 1  p.value= 0
```

The likelihood ratio test of the PH assumption results in a p value of 3.00e-11, indicating that the PH model is not appropriate for this dataset. Or in other words, the data suggest that the hazard functions corresponding to the strata defined by $dmf = 0$ and $dmf = 1$ are not proportional. From the stratified Weibull regression model, the estimated regression coefficient for **sex** is 0.331, corresponding to a hazard ratio of 1.39 (95% CI: 1.29 - 1.50). In the output above, the maximized value of the log likelihood of the null model corresponds to the model stratified by covariate **dmf** but excluding the explanatory variable **sex**.

The p value from the Wald test of the null hypothesis of no effect of gender results in a p value of approximately 0 ($p < 10^{-16}$), which indicates that the timing of emergence of teeth is significantly different between girls and boys.

To test the global null hypothesis that both covariates **sex** and **dmf** are not associated with the outcome (time to teeth emergence), we obtain the log-likelihood for global null model, as shown below.

```
fit0 <- icweib(L = left, R = right, data = tooth24)
fit0

Total observations used: 4386. Model Convergence: TRUE

Weibull parameters - gamma(shape), lambda(scale):
 straname strata gamma   lambda
   strata    ALL   5.3 7.78e-05

Loglik(model)=  -5596.986
Loglik(null)=  -5596.986
```

The likelihood ratio test testing the global null hypothesis results in a test statistic $T = -2(l_R - l_F) = -2(-5596.986 + 5501.781) = 190.41$, which follows a χ^2_3 distribution under H_0, resulting in a p value of approximately 0 ($p < 10^{-16}$).

We illustrate the **HRatio** function in the *straweib* package to estimate the hazard ratio and corresponding 95% confidence intervals for comparing boys without tooth decay ($dmf = 0$) to boys with evidence of tooth decay ($dmf = 1$), where the hazard ratio is evaluated at various time points from 1 through 7 years.

```
HRatio(fit, times = 1:7, NumStra = 0, NumZ = 0, DemStra = 1, DemZ = 0)

  time NumStra DemStra beta*(Z1-Z2)        HR      low95    high95
1    1       0       1            0 0.1143698 0.06596383 0.1982972
2    2       0       1            0 0.2520248 0.18308361 0.3469262
3    3       0       1            0 0.4000946 0.33112219 0.4834339
4    4       0       1            0 0.5553610 0.49863912 0.6185351
5    5       0       1            0 0.7162080 0.66319999 0.7734529
6    6       0       1            0 0.8816470 0.79879884 0.9730878
7    7       0       1            0 1.0510048 0.91593721 1.2059899
```

The output indicates that the hazard ratio for boys comparing the stratum $dmf = 0$ to stratum $dmf = 1$ is small initially (e.g. 0.11 at 1 year) but tends to 1 in later

years (e.g. 0.88 at 6 years and 1.05 at 7 years). Prior to 6 years, the hazard ratio is significantly less than 1, indicating that the timing of teeth emergence is delayed in children with tooth decay ($dmf = 1$) when compared to children without tooth decay ($dmf = 0$).

We illustrate estimation of the survival function in Figure 1.2 by plotting the survival functions and corresponding 95% point wise confidence intervals for girls ($Z = 1$), with and without tooth decay.

```
plot(fit, Z = 1, tRange = c(1, 7), xlab = "Time (years)",
   ylab = "Survival Function",
   main = "Estimated survival function for girls")
```

We compare our results from the *straweib* package to that obtained from the *survival* package.

```
library(survival)
tooth24.survreg <- tooth24
tooth24.survreg$right <- with(tooth24, ifelse(is.finite(right), right, NA))
fit1 <- survreg(Surv(left, right, type="interval2") ~ sex + strata(dmf) +
      factor(dmf), data = tooth24.survreg)
fit1

Call:
survreg(formula = Surv(left, right, type = "interval2") ~ sex +
    strata(dmf) + factor(dmf), data = tooth24.survreg)

Coefficients:
 (Intercept)          sex factor(dmf)1
  1.84389938  -0.06254599  -0.06491729

Scale:
dmf=Sound1 dmf=Sound2
 0.1659477  0.2072465

Loglik(model)= -5499.3   Loglik(intercept only)= -5576.2
        Chisq= 153.8 on 2 degrees of freedom, p= 0
n= 4386
```

The maximized value of the log-likelihood from the R package *survival* is -5499.3 (shown below), as compared to the maximized value of the log-likelihood of -5501.8 from the R package *straweib*.

To clarify the specific assumptions made by the models implemented in the *survival* and *straweib* packages, we carried out subgroup analyses in which we fit a Weibull PH model separately to each of the strata $dmf = 0$ and $dmf = 1$. The results from the Weibull PH model fit to the subgroup of children in the $dmf = 0$ stratum is shown below:

```
fit20 <- icweib(L= left, R=right, data=tooth24[tooth24$dmf==0, ],
           covariates = ~sex)
fit20 ### Partial results shown below
Coefficients:
    coefficient     SE    z  p.value
sex       0.448 0.0543 8.25 2.22e-16
```

The results from the Weibull PH model fit to the subgroup $dmf = 1$ is shown below:

```
fit21 <- icweib(L= left, R=right, data=tooth24[tooth24$dmf==1, ],
          covariates = ~sex)
fit21 ### Partial results shown below
Coefficients:
    coefficient     SE    z  p.value
sex       0.208 0.0554 3.76 0.000169
```

The model using the PH scale (implemented by *straweib* package) replaces the stratum specific hazard ratios for sex of $e^{0.448} = 1.57$ for the subgroup $dmf = 0$ and $e^{0.208} = 1.23$ for the subgroup $dmf = 1$ with a common value, $e^{0.331} = 1.39$.

Since the Weibull distribution has both the PH and accelerated failure time (AFT) property ([5]), the identical set of subgroup analyses can be fit using the *survival* package. Results from the fit using the *survival* package for the subgroup $dmf = 0$ are shown below:

```
fit20.survreg <- survreg(Surv(left, right, type="interval2") ~ sex,
                 data = tooth24.survreg[tooth24.survreg$dmf==0, ])
fit20.survreg ### Partial results shown below
Coefficients:
(Intercept)         sex
 1.85029150 -0.07453785
```

Similar results using the *survival* package for the subgroup $dmf = 1$ are shown below:

```
fit21.survreg <- survreg(Surv(left, right, type="interval2") ~ sex,
                 data = tooth24.survreg[tooth24.survreg$dmf==1, ])
fit21.survreg ### Partial results shown below
Coefficients:
(Intercept)         sex
 1.76931556 -0.04303767
```

In particular, the model assuming a common sex coefficient in the AFT scale (implemented by *survival* package) replaces the value of sex coefficient -0.075 for the subgroup with $dmf = 0$ and sex coefficient of -0.043 for the subgroup $dmf = 1$ with a shared common value, -0.063.

To assess the goodness of fit of the stratified Weibull model implemented by *straweib*, we created a multiple probability plot, as described in chapter 19 of [39]. This diagnostic plot was created by splitting the dataset into 4 subgroups based on the values of **sex** and **dmf**. Within each group, we estimated the cumulative incidence at each visit time using a non-parametric procedure for interval censored data ([53]). The non-parametric estimates of cumulative incidence within each subgroup were compared to that obtained from the stratified Weibull model implemented by *straweib* package. We use the R package *interval* ([12]) to obtain Turnbull's NPMLE estimates and the R package *straweib* for the estimates from the stratified Weibull model (code available upon request). Figure 1.3 shows the diagnostic plot.

Table 1.1 presents the estimates of hazard ratio for **sex**, within each of the strata defined by $dmf = 0$ and $dmf = 1$, comparing three different analyses - (1) Using the

survival package to stratify on the variable **dmf** and including **sex** as an explanatory variable; (2) Using the *straweib* package to stratify on the variable **dmf** and including **sex** as an explanatory variable; (3) Fitting a Weibull PH model with **sex** as an explanatory variable, separately within each of the two subgroups defined by $dmf = 0$ and $dmf = 1$.

```
HR.straweib <- exp(fit$coef[1, 1])
HR.survreg <- exp(-fit1$coefficients['sex']/fit1$scale)
HR.subgroup <- exp(c(fit20$coef[1, 1], fit21$coef[1, 1]))
```

Table 1.1. Hazard ratio estimates for gender, comparing the models implemented in the R packages *survival*, *straweib* and subgroup analyses

	R package		
stratum	*survival*	*straweib*	Stratum specific subgroup analyses
dmf = 0	1.46	1.39	1.56
dmf = 1	1.35	1.39	1.23

1.5 Concluding remarks

We have developed and illustrated an R package *straweib* for the analysis of interval-censored outcomes, based on a stratified Weibull regression model. The proposed model shares similarities with the semi-parametric stratified Cox model. We illustrated the R package *straweib* using data from a prospective study on the timing of emergence of permanent teeth in Flemish children in Belgium [31].

Although the models and R package are illustrated for the analysis of interval-censored time-to-event outcomes, the methods proposed here are equally applicable for the analysis of right-censored outcomes. The syntax for the analysis of right-censored observations is explained in the manual accompanying the *straweib* package available on CRAN ([21]).

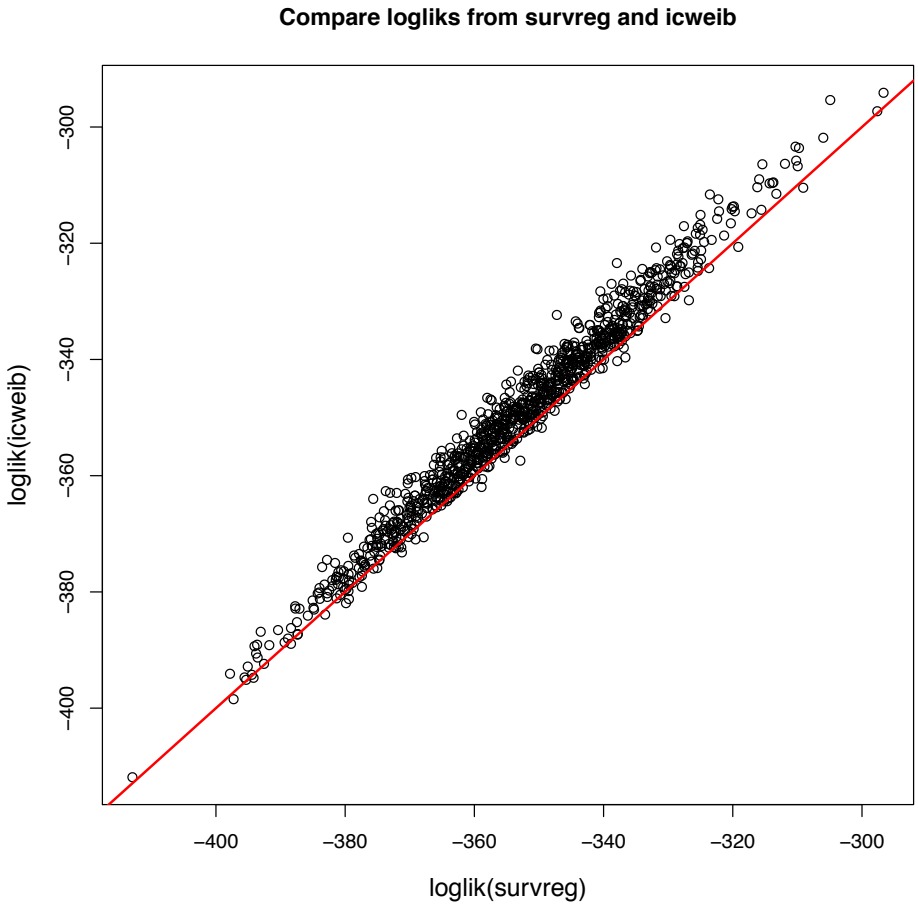

Figure 1.1. Comparing the maximized values of the log-likelihood obtained from the models implemented in the R package *survival* (X axis) to that from the R package *straweib* (Y axis), when the data is simulated under the model implemented in the R package *straweib*

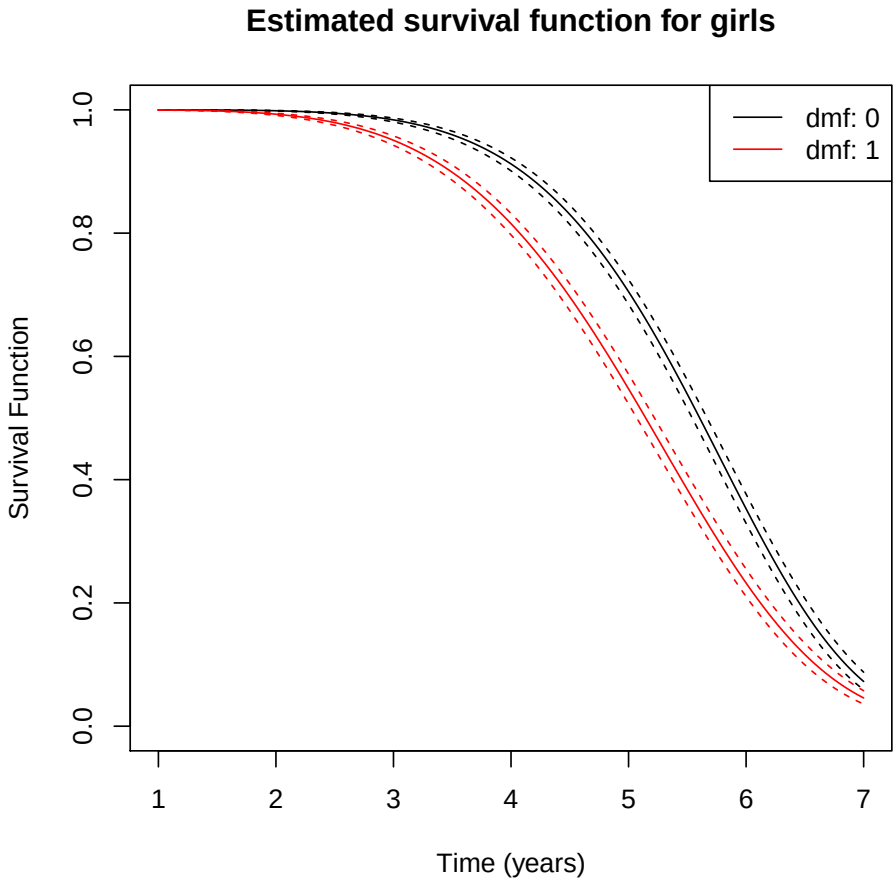

Figure 1.2. Estimated survival functions for girls, comparing the subgroup with sound primary predecessor of the tooth (dmf = 0) to the subgroup with unsound primary predecessor of the tooth (dmf = 1).

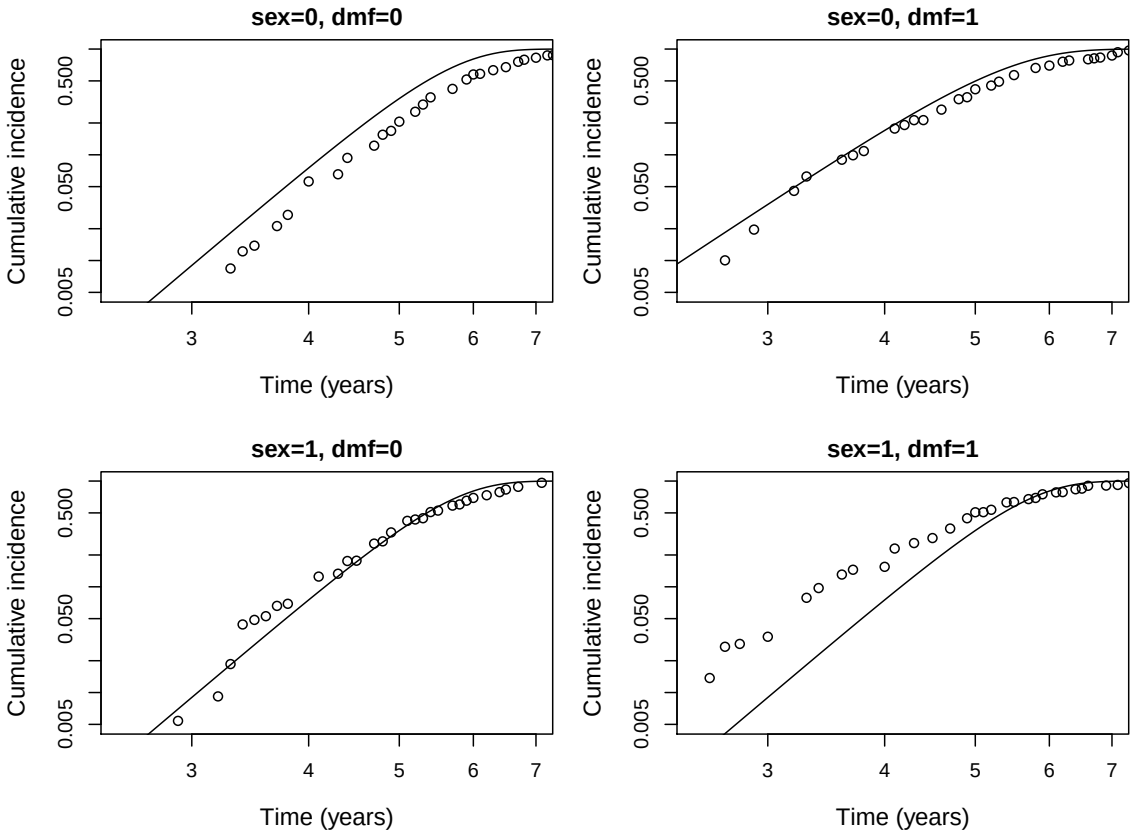

Figure 1.3. Comparing non-parametric (points) and Weibull model (lines) based estimates of cumulative incidence within each group based on covariates **sex** and **dmf**.

CHAPTER 2

SEMI-PARAMETRIC REGRESSION MODELS IN THE PRESENCE OF ERROR-PRONE, SELF-REPORTED OUTCOMES

2.1 Introduction

The onset of several chronic diseases such as diabetes are asymptomatic and can be detected only through diagnostic tests. For example, diabetes can be detected by measuring levels of fasting blood glucose or glycosylated hemoglobin levels (HbA1c). However, the costs of such gold standard diagnostic tests can be prohibitive in large-scale epidemiological studies such as the Women's Health Initiative (WHI) that enroll and follow over a hundred thousand subjects. Disease prevalence and incidence in large observational cohorts are often ascertained through error-prone, self-reported questionnaires. In this chapter, we propose a semi-parametric regression model to assess the association of specific covariates of interest with a silent time to event outcome that is assessed through periodic self-reports subject to error.

The motivating application in this chapter is the evaluation of the hypothesis that the use of cholesterol lowering medications (statins) can result in an increased risk of diabetes, using data from postmenopausal women enrolled in the WHI. The WHI recruited women (N=161,808) aged 50-79 at 40 clinical centers across the U.S. from 1993-1998 with ongoing follow-up ([1]). Prevalent and incident diabetes during the course of follow up was ascertained by self-report obtained at each annual visit. In a recent paper, [9] presented an analysis of the effects of statin use on the risk of incident diabetes in the WHI using Cox proportional hazards models. The analyses were conducted based on the assumption that self-reported outcomes of prevalent and

incident diabetes are error-free. The validity of self-reports of incident and prevalent diabetes have been evaluated by [34] using data from a substudy nested within the WHI - when compared to fasting glucose levels (treated as the gold standard), diabetes self-reports had a positive predictive value of 74% and negative predictive value of 97%. Other studies such as the Nurses' Health Study and Physicians' Health Study also commonly use self-reported outcomes ([24, 25]).

When a perfect diagnostic test is given sequentially at different points in time to the same individual, the time until the event of interest can be determined to lie in the interval between the last negative test and the first positive test - that is, the time until the event is interval censored. In this context, methods for estimating the survival distribution and assessing the effect of covariates have been developed ([53, 13]). However, when error-prone diagnostic procedures such as self-reports are used, standard methods for interval censored outcomes are rendered invalid. Previous work in this area includes methods for error-prone outcomes with application to studies in HIV, HPV and STD ([2, 3, 36, 40]). [3] developed a formal likelihood framework to estimate the distribution of the time to event of interest in the presence of error-prone laboratory-based diagnostic tests, in the context of data obtained from pediatric HIV clinical trials. [40] extended the discrete proportional hazard model to incorporate mismeasured outcomes and also covariates. In related work, [47, 6, 7] considered generalized Cox models in settings involving time to event outcomes with incomplete event adjudication. Other related work includes that proposed by [36] in the context of HPV studies, where the authors accommodate misclassification by incorporating ideas of binary generalized linear models with outcomes subject to misclassification ([42]). The problem of error-prone time to event outcome can also be handled through Hidden Markov Model (HMM) framework. Previous applications of HMM based methods include in the areas of breast cancer ([4]), HIV ([43, 23]), lung transplantation ([26]) and cervical smear tests ([29]). [27] present a general

framework for staged Markov models to handle misclassification due to error prone screening tests.

In this chapter, we present a likelihood based approach to incorporate time-varying covariate effects specific to the setting in which the prevalence and incidence of a chronic condition such as diabetes is ascertained through error-prone self-reports. We incorporate the situation where an unknown proportion of subjects who have already experienced the event of interest at baseline are mistakenly included into the study, due to the use of error-prone self-reports at study entry. We also provide a freely available R software package ([20]) and illustrate its use. In Section 2.2, we present notation, form of the likelihood function, address issues related to estimation and extensions to incorporate misclassification of subjects at study entry. In Section 2.3, we perform simulation studies to evaluate the effects of various degrees of error in self-reports. We investigate the effects of erroneous inclusion of subjects who have already experienced the event of interest due to less than perfect negative predictive values associated with self-reports. In Section 2.4, we evaluate the association between statin use with the risk of incident diabetes in a subset of 152,830 women enrolled in the WHI. Lastly, in Section 2.5 we discuss the findings of this study and highlight future directions.

2.2 Methods

In this section, we first present the notation, likelihood, and estimation then we extend our model to incorporate the possibility of misclassification at study entry.

2.2.1 Notation, likelihood, estimation

Let X refer to the random variable denoting the unobserved time to event for an individual, with associated survival, density and hazard functions denoted by $S(x), f(x)$ and $\lambda(x)$, for $x \geq 0$ respectively. The time origin is set to 0, corresponding

to the baseline visit at which all subjects enrolled in the study are event-free. In other words, $Pr(X > 0) = 1$. Without loss of generality, we set $X = \infty$ when the event of interest does not occur. Let N denote the number of subjects and n_i denote the number of pre-scheduled visits for the i^{th} subject. At each visit, we assume that each subject would self report their disease status. For example, in the Women's Health Initiative, information on incident diabetes was collected at periodicallly scheduled visits using self-reported questionnaires. For the i^{th} subject, we let $\mathbf{R_i}$ and $\mathbf{t_i}$ denote the $1 \times n_i$ vectors of binary self-reported, binary outcomes and corresponding visit times, respectively. In particular, R_{ij} is equal to 1 if the j^{th} self-report for the i^{th} subject is positive (indicating occurrence of the event of interest such as diabetes) and 0 otherwise. Let $\tau_1, \cdots, \tau_J$ denote the distinct, ordered visit times in the dataset among N subjects, where $0 = \tau_0 < \tau_1 < ... < \tau_J < \tau_{J+1} = \infty$ - thus, the time axis can be divided into $J + 1$ disjoint intervals, $[\tau_0, \tau_1), [\tau_1, \tau_2), \cdots, [\tau_J, \infty)$.

In general, the likelihood contribution for the ith subject can be expressed as:

$$\begin{aligned} L(\mathbf{R_i} \mid \mathbf{t_i}) &= \sum_{j=1}^{J+1} Pr(\tau_{j-1} < X \leq \tau_j) Pr(\mathbf{R_i} \mid \tau_{j-1} < X \leq \tau_j, \mathbf{t_i}) \\ &= \sum_{j=1}^{J+1} \theta_j Pr(\mathbf{R_i} \mid \tau_{j-1} < X \leq \tau_j, \mathbf{t_i}) \end{aligned}$$

where $\theta_j = Pr(\tau_{j-1} < X \leq \tau_j)$, $\tau_0 = 0$ and $\tau_{J+1} = \infty$.

To simplify the form of the likelihood above, we make the assumption that the probability of a positive self-report at the kth visit at t_k **conditional** on all previous self reported outcomes and the true time of the event can be simplified as:

$$Pr(r_{ik} = 1 \mid r_{i1}, \cdots r_{i,k-1}, \tau_{j-1} < X \leq \tau_j, t_k) = Pr(r_{ik} = 1 \mid r_{i,k-1}, \tau_{j-1} < X \leq \tau_j, t_k)$$

Thus, the likelihood for the ith subject can be simplified as:

$$
\begin{aligned}
L(\mathbf{R_i} \mid \mathbf{t_i}) &= \sum_{j=1}^{J+1} \theta_j \left[Pr(r_{i1} \mid \tau_{j-1} < X \leq \tau_j, t_1) \prod_{k=2}^{n_i} Pr(r_{ik} \mid r_{i,k-1}, \tau_{j-1} < X \leq \tau_j, t_k) \right] \\
&= \sum_{j=1}^{J+1} \theta_j C_{ij}
\end{aligned}
\tag{2.1}
$$

where $C_{ij} = [Pr(r_{i1} \mid \tau_{j-1} < X \leq \tau_j, t_1) \prod_{k=2}^{n_i} Pr(r_{ik} \mid r_{i,k-1}, \tau_{j-1} < X \leq \tau_j, t_k)]$. We first assume that once a subject has self-reported the occurrence of the event of interest, then all subsequent self-reports will be positive for the event of interest. In other words, $Pr(r_{ik} = 1 \mid r_{i,k-1} = 1) = 1$ for all j. Moreover, we assume that for a subject whose last self-report was negative for the event of interest (i.e. $r_{i,k-1} = 0$), the probability of a positive self report at the next visit ($r_{ik} = 1$) conditional on the interval containing the true time of the event of interest is independent of visit time and can be expressed as:

$$
Pr(r_{ik} = 1 | r_{i,k-1} = 0, \tau_{j-1} \leq X < \tau_j, t_k) = \begin{cases} \varphi_1, & t_k \geq \tau_j \\ 1 - \varphi_0, & t_k < \tau_{j-1} \end{cases}
$$

Thus the terms C_{ij}, for $j = 1, \cdots, J+1$ in equation (1) can be expressed as a product involving the constants φ_1, $1-\varphi_1$, φ_0, or $1-\varphi_0$. We note that an equivalent expression for the form of the likelihood can be obtained by assuming independence of self-reported results conditional on true event time ([3]). Thus, in the absence of covariates, the likelihood for a random sample of N subjects can be expressed as:

$$
l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta})) = \sum_{i=1}^{N} \log(\sum_{j=1}^{J+1} C_{ij} \theta_j) \tag{2.2}
$$

In most settings, including the WHI, it is of interest to evaluate the association of a vector of covariates with respect to the time to the event of interst. Let $\boldsymbol{Z}$ denote the $P \times 1$ vector of explanatory variables with corresponding $P \times 1$ vector

of regression coefficients denoted by $\boldsymbol{\beta}$. To incorporate the effect of covariates, we assume the proportional hazards model, $\lambda(t|\boldsymbol{Z} = \boldsymbol{z}) = \lambda_0(t)e^{\boldsymbol{z}'\boldsymbol{\beta}}$, or equivalently, $S(t|\boldsymbol{Z} = \boldsymbol{z}) = S_0(t)^{e^{\boldsymbol{z}'\boldsymbol{\beta}}}$.

To derive the form of the log-likelihood based on the assumption of the proportional hazards model, we first re-parameterize the log likelihood in (2) in terms of the of the survival function, $\boldsymbol{S} = (1 = S_1, S_2, \cdots, S_{J+1})^T$, where $S_j = Pr(X > \tau_{j-1})$. Since $S_j = \sum_{l=j}^{J+1} \theta_l$, the vector of interval probabilities can be expressed as $\boldsymbol{\theta} = T_r\boldsymbol{S}$, where T_r is the $(J+1) \times (J+1)$ transformation matrix. Let $C = [C_{ij}]$ denote the $N \times (J+1)$ matrix of the coefficients, C_{ij}, and let the $N \times (J+1)$ matrix D be defined as $D_{N\times(J+1)} = C \times T_r$. Then, the log-likelihood function for the one-sample setting in (2) can be expressed as

$$l(\mathbf{S}) = \sum_{i=1}^{N} \log(\sum_{j=1}^{J+1} D_{ij}S_j), \tag{2.3}$$

where $S_1 = 1$ and $S_2, S_3, \cdots, S_{J+1}$ are the unknown parameters of interest.

Let $1 = S_1 > S_2 > ... > S_{J+1}$ denote the baseline survival functions (i.e. corresponding to $\boldsymbol{Z} = \mathbf{0}$), evaluated at the left boundaries of the intervals $[0, \tau_1), [\tau_1, \tau_2)$, $\cdots, [\tau_J, +\infty)$. Then, for subject i with corresponding covariate vector $\boldsymbol{z}_i$, $S_j^{(i)} = (S_j)^{e^{\boldsymbol{z}_i'\boldsymbol{\beta}}}$. Thus, the log-likehood function for a random sample of N subjects can be expressed as

$$l(\boldsymbol{S}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \log(\sum_{j=1}^{J+1} D_{ij}(S_j)^{e^{\boldsymbol{z}_i'\boldsymbol{\beta}}}). \tag{2.4}$$

The elements of the D matrix are functions of the observed data including the visit times and corresponding self-reported results, as well as the constants φ_0, φ_1. Assuming that φ_0, φ_1 are known constants, the maximum likelihood estimates of the unknown parameters $\beta_1, \cdots, \beta_P, S_2, \cdots, S_{J+1}$ can be obtained by numerical maximization of the log-likelihood function, subject to the constraints that $1 > S_2 > S_3 > \cdots > S_{J+1} > 0$. Statistical inference regarding the parameters of interest

$(\beta_1, \cdots, \beta_P, S_2, \cdots, S_{J+1})$ can be made by using asymptotic properties of the maximum likelihood estimators ([8]). The estimated covariance matrix of the maximum likelihood estimates can be obtained by inverting the Hessian matrix. Hypothesis tests regarding the unknown parameters can be carried out using the likelihood ratio or Wald test.

2.2.2 Misclassification at study entry

In this section, we incorporate the setting in which a self-report of being event (disease) free at baseline or study entry is used as the inclusion criterion. The evaluation of the association between statin use and risk of incident diabetes in the WHI was based on all women who self-reported to be diabetes free at baseline ([9]). However, diabetes self reports at study entry in the WHI have been found to be less than perfect - the study by [34] found that the negative predictive value of prevalent diabetes at baseline was approximately 97% - that is, 3% of women who self-reported as being diabetes free were in fact diabetic. In this situation, the assumption in the model developed in Section 2.2.1 that $S(0) = 1$ is invalid.

For the i^{th} subject, let G_i denote the baseline binary self-report, where $G_i = 1$ denotes a self report indicating that the event of interest has already occurred and $G_i = 0$ denotes otherwise. Similarly, let B_i denote the true event status at baseline. In other words, $B_i = 1 \stackrel{def}{=} X_i \leq 0$ and $B_i = 0 \stackrel{def}{=} X_i > 0$. Consider a subject who has a negative self-report result at baseline (i.e. $G_i = 0$) and is thus, included in the dataset. As before, let the observed self-report results for the i^{th} subject be denoted by $\boldsymbol{R}_i$. Let the negative predictive value of the self-report at baseline $Pr(B_i = 0|G_i = 0)$ be denoted by η, which we assume to be constant for all subjects. Then the likelihood function for the i^{th} subject can be expressed as:

$$
\begin{aligned}
L_i &= Pr(\boldsymbol{R}_i|G_i = 0, \mathbf{t_i}) \\
&= \eta Pr(\boldsymbol{R}_i|B_i = 0, G_i = 0, \mathbf{t_i}) + (1-\eta) Pr(\boldsymbol{R}_i|B_i = 1, G_i = 0, \mathbf{t_i})
\end{aligned} \tag{2.5}
$$

We assume that for those who are true negative at baseline, the self report result at baseline is non-informative. That is, those who self report negative and are also truly negative for at baseline are a random sample from all subjects who are true negative at baseline. Then we have $Pr(\boldsymbol{R}_i|B_i = 0, G_i = 0, \mathbf{t_i}) = Pr(\boldsymbol{R}_i|B_i = 0, \mathbf{t_i})$, which corresponds to the likelihood function derived in Section 2.2.1, based on the assumption that all subjects included in the analysis have not experienced the event of interest at baseline (i.e. $X > 0$). Thus, $Pr(\boldsymbol{R}_i|B_i = 0, G_i = 0, \mathbf{t_i}) = \sum_{j=1}^{J+1} D_{ij}(S_j)^{e^{\boldsymbol{z}_i'\boldsymbol{\beta}}}$. Since the probability of self report result conditional on $X \leq 0$ is equal to the probability of self report result conditional on $\tau_0 < X \leq \tau_1$, $Pr(\boldsymbol{R}_i|B_i = 1, G_i = 0, \mathbf{t_i}) = C_{i1} = D_{i1} = D_{i1}(S_1)^{e^{\boldsymbol{z}_i'\boldsymbol{\beta}}}$.

The likelihood function for the i^{th} subject has the form,

$$
\begin{aligned}
L_i(\boldsymbol{\beta}, \boldsymbol{S}) &= \eta \sum_{j=1}^{J+1} D_{ij}(S_j)^{e^{\boldsymbol{z}_i'\boldsymbol{\beta}}} + (1-\eta) D_{i1}(S_1)^{e^{\boldsymbol{z}_i'\boldsymbol{\beta}}} \\
&= \sum_{j=1}^{J+1} D'_{ij}(S_j)^{e^{\boldsymbol{z}_i'\boldsymbol{\beta}}}
\end{aligned} \tag{2.6}
$$

where $D'_{i1} = D_{i1}$ and $D'_{ij} = \eta D_{ij}$ for $j > 1$. Thus, the likelihood function incorporating baseline misclassification has the same general form as in equation (2.4). The likelihood function in equation (2.4) can be obtained as a special case when $\eta = 1$ in equation (2.6).

2.2.3 Time varying covariates

We consider the situation where covariate values can change with time and are collected at each visit. Let $\boldsymbol{z_{ij}}$ denote the $p \times 1$ vector of covariate values for subject i at

time τ_j. In extending the likelihood function (Equation (2.4)) to handle time-varying covariates, we make the additional assumption that the values of the covariates $\boldsymbol{z_{ij}}$ remains constant during the interval $[\tau_j, \tau_{j+1})$. Let Λ_j denote the cumulative hazard function during the period of $[\tau_j, \tau_{j+1})$ for the subjects in the reference group (i.e. $\boldsymbol{Z} = 0$). Under the model $\lambda_{\boldsymbol{z_i}}(t) = \lambda_{\boldsymbol{0}}(t)e^{\boldsymbol{\beta z_i}}$, the corresponding cumulative hazard function during the period $[\tau_j, \tau_{j+1})$ for subject i is equal to $\Lambda_j \exp(\boldsymbol{z'_{ij}\beta})$. The survival function at τ_{j-1} can then be expressed as,

$$S_j^{(i)} = \exp\left(-\sum_{j'=0}^{j-2} \Lambda_{j'} \exp\left(\boldsymbol{z'_{ij'}\beta}\right)\right)$$

where $j = 2, \cdots, J+1$, where $S_1^{(i)} = 1$. The likelihood function can be expressed as function of the derived $S_j^{(i)}$,

$$l(\boldsymbol{S}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{J+1} D_{ij} S_j^{(i)}\right)$$

The log-likelihood function can be optimized with respect to the parameters $\Lambda_0, \cdots, \Lambda_{J-1}$ and $\beta_1, \cdots, \beta_P$ subject to constrains $\Lambda_j \geq 0$. In practice, if a subject has missing visits or missing covariate values at some visits, one can carry forward the last observation as one approach to impute missing covariate values.

2.3 Simulation

In this section, we present results from simulation studies to illustrate the effects of (1) error-prone self reported outcomes; and (2) misclassification at study entry (baseline). We present the effects of the factors noted above with regard to the bias associated with the estimated regression parameter of interest.

2.3.1 Effects of error-prone self-reported outcomes

The simulation studies are based on 1000 subjects randomly assigned to two exposure groups with equal proportion, where all subjects are event-free at baseline (i.e. $X_i > 0$ for all i). We assume that there is a single binary covariate of interest Z_i, corresponding to the exposure status of the ith subject. The associated regression parameter in the likelihood (Equation 2.4) is set to $\beta = 1$. For each subject, self reported questionnaires were collected at 8 scheduled visits over a duration of 8 years, each with random missing probability of 30%. All self reports following the first positive report were assumed to be positive with probability 1. The simulation mechanism assumed that the time to the event of interest X followed an exponential distribution. The hazard rate λ governing the time to the event of interest in the reference group ($Z_i = 0$) is set to equal 0.0132 or 0.0866, corresponding to cumulative incidence by study end ($1 - S_{J+1}$) of 0.10 or 0.50, respectively. As shown in Table 2.1, we compare results across several sets of values for the parameters (φ_1, φ_0) governing the characteristics of self-reports.

Table 2.1 presents the results from the simulation study, averaged over 1000 datasets. For each parameter setting, we present estimates of bias, associated standard error, root mean square error (RMSE), and coverage probability associated with the estimation of the regression parameter of interest, β. Coverage probability is calculated as the proportion of datasets in which the 95% confidence interval for β contains its true value. We compare results from two sets of analyses for estimating β - (a) maximizing the likelihood presented in Equation (2.4), assuming that the true values of φ_1, φ_0 are known; and (b) maximizing the likelihood presented in Equation (2.4) assuming that self-reports are perfect (that is, $\varphi_1 = \varphi_0 = 1$). In general, when the true values of φ_0, φ_1 are incorporated into the analysis, the estimates of β are nearly unbiased. Similarly, the true coverage probability corresponding to a 95% confidence interval is close to its nominal value. On the other hand, when self-reports

are incorrectly assumed to be perfect, the estimates of β may be significantly biased, especially in settings where φ_0 is low. When $\varphi_0 << 1$, early false positive results result in significant loss of information due to premature cessation of data collection. In this case, coverage probabilities deviated significantly from 95% especially in settings where $\varphi_0 << 1$. Lastly, incorporating the uncertainty in error-prone self-reports increases the standard error of the maximum likelihood estimates of β.

2.3.2 Effects of misclassification at study entry

In this simulation, we incorporate the setting in which an error-prone, self-report of being event (disease) free at study entry is used as the inclusion criterion. As before, let η denote the negative predictive value of the baseline self-report. That is, each subject included in the study has a probability of $1 - \eta$ of having already experienced the event of interest prior to study entry. We assumed that 1000 subjects are enrolled in the study, of whom $1000 \times (1 - \eta)$ have already experienced the event of interest prior to entry into the study (i.e. $X < 0$). The data are simulated as described in Section 2.3.1 , where $\varphi_1 = 0.61$ and $\varphi_0 = 0.995$. We compare results for various settings by varying the cumulative incidence of the event of interest $(1 - S_{J+1})$ to equal 0.10 or 0.50, and by varying the value of η to equal 0.99, 0.96 or 0.93.

Table 2.2 presents the simulation results, averaged over 1000 datasets. We present results from an "adjusted" model that properly accounts for misclassification at baseline based on the likelihood presented in Equation (2.6) compared to the model that incorrectly assumes that $\eta = 1$ (denoted "Unadjusted"). As expected, the adjusted model is nearly unbiased and has uniformly lower bias when compared to the unadjusted model. The bias of the unadjusted model increases with decreasing values of negative predictive value (η), and it is more pronounced when the cumulative incidence is low $(1 - S_{J+1} = 0.10)$. In general, the inclusion of subjects who have already experienced the event of interest at study entry results in the exposure groups becom-

ing less distinguishable. Thus, ignoring this issue in data analysis results in estimates of exposure effects (β) that are biased towards the null. In contrast, incorporating the effect of baseline misclassification increases the standard error of $\hat{\beta}$. The effects on the bias and the standard error of $\hat{\beta}$ are reflected in the RMSE values - the adjusted model has smaller RMSE than the unadjusted model in all settings except when $S_{J+1} = 0.9$ and $\eta = 0.99$. The coverage probability of the adjusted model is approximately 95% in all settings considered in this study. However, the coverage probability of the unadjusted model decreases with decreasing negative predictive value (η) due to increased bias.

2.4 Application: Risk of diabetes mellitus with statin use in the Women's Health Initiative

Background: We analyze data collected on 152,830 women from the Women's Health Initiative (WHI) to evaluate the effects of statin use on the risk of incident diabetes mellitus (DM). [9] reported an increased risk of incident DM with baseline statin use (multivariate-adjusted HR, 1.48; 95% CI, 1.38-1.59). These results were based on Cox PH models where the time to event variable was calculated as the interval between enrollment date and the earliest of the following: 1) date of annual medical history update when new diabetes is self-reported (positive outcome); 2) date of last annual medical update during which diabetes status can be ascertained (censorship); or 3) date of death (censorship). The methods used in [9] were based on the assumptions that: (1) all subjects who self-reported as being diabetes free at baseline were truly not diabetic (that is, $\eta = 1$); and (2) the self-reports of incident diabetes at each follow up visit were error-free (i.e. $\varphi_1 = \varphi_0 = 1$). We compare the results from [9] to results based on application of the likelihood based methods described in this chapter.

Diabetes self-reports: Prevalent diabetes at baseline and incident diabetes were assessed through self reported questionnaires in the WHI. At baseline and at each annual visit, participants were asked whether she has ever received a physician diagnosis of and/or treatment for diabetes when not pregnant since the time of the last self-report/visit. Using data from a WHI substudy ([34]), estimates of sensitivity, specificity, and baseline negative predictive value of self reported diabetes outcomes were obtained by comparing self reported outcomes to fasting glucose levels and medication data. A woman was considered to be truly diabetic if she had either taken anti-diabetic medication and/or had a fasting glucose level $\geq$ 126mg/dL. By using a subset of 5485 women, with information at baseline on diabetes self reports, fasting glucose levels and medication inventory, we estimated that self reports have a sensitivity of 0.61, the specificity of 0.995, and a negative predictive value of 0.96 at baseline. These estimated parameter values are used in our analysis.

Methods: The analysis dataset included 152,830 women out of a total of 161,808 women enrolled in the WHI. Women with self reported diabetes at baseline or missing diabetes status or medication inventory at baseline were excluded. In addition, women who ever took cerivastatin were excluded from our analysis ([9]). The results presented here are based on follow up until 2010. The median follow up time was 12.1 years, including 1,688,967 person-years of total follow up. During the course of follow up, 10.4% of women self reported being diagnosed with diabetes. Information on statin use was obtained from medical inventory information, which was available for selected follow-up years. Information on statin use was available for 59,505, 128,507, 55,043 and 12,039 subjects at years 1, 3, 6, and 9. Models included either baseline statin use or statin use as a time varying covariate - in the latter case, missing medication information data was imputed by carrying forward the last observation. In multivariable models, other covariates included race, smoking status, alcohol intake, age, education, WHI study, BMI, recreational physical activity, dietary energy intake,

family history of diabetes, and hormone therapy use. We assumed that self reports following the first report of incident diabetes are non-informative. Annual visit times were rounded to the nearest year in order to limit the number of parameters estimated to describe the baseline survival function $(S_2, \cdots, S_{J+1})$.

Results: Table 2.3 presents the estimated hazard ratio (95% confidence interval) for statin use by modeling statin use at baseline or as a time-varying covariate. For each, we present results from univariable models as well as multivariable models incorporating potential confounders. In each setting, the results from the methods proposed in this chapter are compared to results from Cox models. In all models, by incorporating self-report measurement error and potential misclassification at study entry, the hazard ratio of statin use is consistently increased when comparing to the corresponding Cox models. Using the proposed methods, the hazard ratio for baseline statin use from univariate analysis was 2.33(95% CI: 2.12-2.56). In the multivariable model, the hazard ratio of baseline statin use was 1.81(95% CI: 1.65-1.99) , suggesting a relatively strong confounding effect. When statin use was modeled as a time-varying covariate, the hazard ratios of statin use from univariate and multivariate models were 2.49(95% CI: 2.31 -2.68) and 1.88 (95% CI: 1.75-2.02), respectively.

The goodness of fit of the multivariable model incorporating statin use as a time-varying covariate was assessed in an augmented model that included 2 additional terms corresponding to the interactions of time periods (in years) (3,6] and (6,16] with statin use. This model allows the effect of statin use to vary between the time periods (0,3], (3,6] and (6,16] years. The Wald test p values corresponding to the interactions of statin use with the time periods (3,6] and (6,16] were 0.89 and 0.11, respectively - these results indicate that there are no significant time-varying effects of statin use on incident diabetes risk.

To evaluate how the results depend on the choice of parameters such as sensitivity, specificity and baseline negative predictive value of self-reported diabetes,

we performed a sensitivity analysis by varying each of these parameters. Table 2.4 presents how the estimated hazard ratio of statin use changes with different combinations of the parameters. Statin use was modeled as a time-varying covariate while simultaneously adjusting for potential confounders. We observed that the estimated hazard ratio of statin use is most sensitive to change in specificity. This is largely due to the fact that the cumulative incidence of diabetes was low (10.4%), and thus false positive test results due to imperfect specificity have a big influence on estimated parameters. In general, the hazard ratio of statin use decreases as specificity increases. Changes in sensitivity and negative predictive value at baseline have modest effects on the resulting model fit.

The models presented here can be implemented using our freely available R software package *icensmis* ([20]).

2.5 Discussion

Due to cost considerations, the use of self reported outcomes is common to diagnose prevalent and incident disease in large scale epidemiologic investigations such as the Women's Health Initiative and the Nurses Health Study. In this chapter, we present a likelihood based framework to model the association of a time varying covariate with a time to event outcome, that is observed through periodically collected, error-prone, self reported data. We incorporate the possibility of erroneous inclusion of subjects who have already experienced the event of interest prior to study entry as a result of the use of self reported outcomes at baseline in determining the study population.

We presented results from simulation studies to assess the impact of ignoring error in self reported outcomes - in all cases considered, the use of statistical models that correctly accomodate the error inherent in self reports resulted in nearly unbiased estimates of regression parameter of interest. The largest bias as a result of ignoring

error in self reported outcomes was found in settings where the cumulative incidence was low and specificity was less than perfect. Models that correctly accomodate error in self reports also resulted in increased variance of the estimated regression parameters. However, in most settings, the RMSE values that combine the impact of bias and variance of the estimated regression parameter favored the use of methods that appropriately account for error in self reported outcomes.

The methods proposed in this chapter were applied to prospective data from 152,830 women enrolled in the WHI to evaluate the effect of statin use and risk of incident diabetes. By accounting for the imperfect sensitivity, specificity and negative predictive value at baseline for diabetes self reports, we observed that the hazard ratio for statin use was significantly larger than that estimated in naive analysis that ignored the error in self reported outcomes. In particular, the hazard ratio of statin use in a multivariable model adjusted for potential confounders was 1.88 (95% CI: 1.75-2.02) as compared to the multivariable hazard ratio estimate from Cox model 1.48(95% CI: 1.42-1.54).

In the methods developed here, we assumed that the sensitivity and specificity of self reported outcomes are invariant with respect to time and independent of covariates. In many real world settings, this assumption may result in over-simplified models. In addition, the methods developed here assumed that the parameters governing the characteristics of self reported outcomes are known. However, in many cases these are estimated values - in this context, it would be useful to extend the methods proposed here to allow joint estimation of the sensitivity and specificity of self reported outcomes together with the parameters governing the distribution of the time to event of interest and associated regression parameters. Lastly, our models assumed that the probability of a positive self report conditional on a preceding negative self report is independent of the time duration between the self reports - while this assumption may be plausible for laboratory based diagnostic tests, it may be un-

duly strong for self reported outcomes in settings where the the visits are unequally spaced. However, in the WHI data, visits were equally spaced and we observed a low rate of missingness.

Table 2.1. Comparing estimates of the regression parameter β from an "adjusted" analysis that accounts for the error in self-reported outcomes to an "unadjusted' analysis that incorrectly assumes that self-reports are perfect.

φ_1	φ_0	S_{J+1}	Analysis type	Bias(%)	Std Err	RMSE	Coverage(%)
0.75	1.00	0.90	Adjusted	0.3%	0.17	0.17	96.8%
0.75	1.00	0.90	Unadjusted	0.1%	0.17	0.17	97.0%
1.00	0.75	0.90	Adjusted	-6.7%	0.82	0.82	93.8%
1.00	0.75	0.90	Unadjusted	-90.2%	0.07	0.90	0.0%
0.61	0.995	0.90	Adjusted	1.4%	0.21	0.22	94.9%
0.61	0.995	0.90	Unadjusted	-16.4%	0.17	0.23	82.9%
0.75	1.00	0.50	Adjusted	0.1%	0.09	0.09	95.1%
0.75	1.00	0.50	Unadjusted	-1.9%	0.09	0.09	93.5%
1.00	0.75	0.50	Adjusted	0.2%	0.19	0.19	94.4%
1.00	0.75	0.50	Unadjusted	-59.2%	0.07	0.60	0.0%
0.61	0.995	0.50	Adjusted	0.5%	0.09	0.09	94.2%
0.61	0.995	0.50	Unadjusted	-6.9%	0.08	0.11	86.7%

Table 2.2. Comparing estimates of the regression parameter β from an "adjusted" analysis that incorporates the possibility of misclassification at baseline to an "unadjusted' analysis that incorrectly assumes that all subjects are event-free at study entry or that $\eta = 1$. We assume that $\varphi_1 = 0.61$ and $\varphi_0 = 0.995$.

S_{J+1}	η	Analysis type	Bias(%)	Std Err	RMSE	Coverage(%)
0.90	0.99	Adjusted	2.6%	0.22	0.23	95.0%
0.90	0.99	Unadjusted	-4.5%	0.20	0.21	94.1%
0.90	0.96	Adjusted	1.2%	0.24	0.24	95.8%
0.90	0.96	Unadjusted	-22.9%	0.17	0.29	72.7%
0.90	0.93	Adjusted	0.1%	0.25	0.25	95.2%
0.90	0.93	Unadjusted	-36.4%	0.15	0.40	36.3%
0.50	0.99	Adjusted	0.0%	0.09	0.09	95.2%
0.50	0.99	Unadjusted	-1.5%	0.09	0.09	94.1%
0.50	0.96	Adjusted	0.1%	0.10	0.10	94.2%
0.50	0.96	Unadjusted	-5.7%	0.09	0.11	89.2%
0.50	0.93	Adjusted	0.6%	0.10	0.10	94.1%
0.50	0.93	Unadjusted	-9.4%	0.09	0.13	80.9%

Table 2.3. Analysis of the effects of statin use on incident diabetes mellitus risk in the WHI.

Statin variable type	Type of Analysis	Univariable/ Multivariable*	N	Hazard ratio (95% CI)
Baseline statin	Proposed model	Univariable	152830	2.33(2.12, 2.56)
Baseline statin	Proposed model	Multivariable	138338	1.81(1.65, 1.99)
Baseline statin	Cox model	Univariable	152830	1.69(1.60, 1.78)
Baseline statin	Cox model	Multivariable	138338	1.54(1.46, 1.63)
Time varying statin	Proposed model	Univariable	152830	2.49(2.31, 2.68)
Time varying statin	Proposed model	Multivariable	138338	1.88(1.75, 2.02)
Time varying statin	Cox model	Univariable	152830	1.65(1.59, 1.72)
Time varying statin	Cox model	Multivariable	138338	1.48(1.42, 1.54)

* Covariates adjusted include race, smoking status, alcohol intake, age, education, WHI study, BMI, recreational physical activity, dietary energy intake, family history of diabetes, and hormone therapy use

Table 2.4. Statin use versus risk of incident diabetes mellitus in the WHI - sensitivity analysis for varying sensitivity, specificity and baseline negative predictive value (η) associated with diabetes self reports. All models incorporate statin use as a time varying covariate and adjust for potential confounders.

sensitivity	specificity	η	Hazard Ratio (95% CI)
0.50	0.993	0.96	2.11(1.92, 2.31)
0.50	0.993	0.98	2.10(1.92, 2.30)
0.50	0.995	0.96	1.93(1.79, 2.08)
0.50	0.995	0.98	1.93(1.79, 2.07)
0.50	0.997	0.96	1.76(1.65, 1.88)
0.50	0.997	0.98	1.77(1.66, 1.88)
0.61	0.993	0.96	2.05(1.88, 2.24)
0.61	0.993	0.98	2.06(1.89, 2.24)
0.61	0.995	0.96	1.88(1.75, 2.02)
0.61	0.995	0.98	1.89(1.76, 2.03)
0.61	0.997	0.96	1.73(1.63, 1.84)
0.61	0.997	0.98	1.74(1.64, 1.84)
0.70	0.993	0.96	2.02(1.85, 2.20)
0.70	0.993	0.98	2.03(1.86, 2.21)
0.70	0.995	0.96	1.86(1.73, 2.00)
0.70	0.995	0.98	1.87(1.74, 2.00)
0.70	0.997	0.96	1.71(1.61, 1.82)
0.70	0.997	0.98	1.72(1.62, 1.82)

CHAPTER 3

STUDY DESIGN IN THE PRESENCE OF ERROR-PRONE DIAGNOSTIC TESTS AND SELF-REPORTED OUTCOMES

3.1 Introduction

In Chapter 2, we discussed the effects of error-prone diagnostic tests such as self-report on the estimation of covariate effects. In this chapter, we discuss the study design issues arising when error-prone diagnostic tests are used to ascertain the occurrence of a silent event. When error-prone measurement is used, study design issues have been considered in other experimental settings. [37, 38] developed methods to find optimal designs given a fixed budget when both error-prone and error-free measurements are used in two-stage studies for the estimation of sensitivity, specificity and positive predictive value associated with a diagnostic test. In this chapter, we consider studies aimed at estimating a treatment effect in the presence of time to event outcomes measured with repeatedly administered, error-prone diagnostic procedures - to our knowledge, no previous studies have considered issues related to study design for this setting. The goal of this chapter is to describe effects of various factors influencing the sample size and statistical power in a regression context, when a silent event such as diabetes is detected via sequentially administered diagnostic procedures subject to imperfect sensitivity and/or specificity. We provide a freely available R software package *icensmis* ([20]). In Section 3.2, we describe the methods to calculate power and sample size, incorporate the effects of missing tests and censoring, and present the trade-off in power when comparing perfect and imperfect tests. In Section 3.3, we illustrate the effect of different factors influencing power and

sample size. In Section 3.4, we extend our methods to incorporate settings in which the sensitivity and specificity of the diagnostic test are unknown. Lastly, in Section 3.5, we discuss the findings of this study and highlight future directions.

3.2 Method

In this section, we first present the notation, likelihood, and estimation and describe the approach for deriving analytical expressions for power and sample size calculations. The derivation on likelihood function is very similar to that presented in Chapter 2, but it is more general and applicable to not only self-report but also other diagnostic tests. In Section 3.2.5, we illustrate the effect on power by comparing the use of perfect versus imperfect diagnostic tests.

3.2.1 Notation, likelihood, estimation

Let T refer to the random variable denoting the unobserved time to event for an individual, with associated survival, density and hazard functions denoted by $S(t), f(t)$ and $\lambda(t)$, for $t \geq 0$ respectively. Without loss of generality, we set $T = \infty$ when the event of interest does not occur. Let N denote the number of subjects and n_i denote the number of tests (visits) for the i^{th} subject. Then, for the i^{th} subject, we let $\mathbf{R_i}$ and $\mathbf{t_i}$ denote the $1 \times n_i$ vectors of binary test results and corresponding test times. In particular, R_{ij} is equal to 1 if the j^{th} test result for the i^{th} subject is positive (indicating occurrence of the event of interest) and 0 otherwise. We note that the term *test result* is used to denote both self-reported outcomes and results of laboratory based diagnostic assays. For ease of notation, we consider the special case where all tests are of the same type, with fixed sensitivity and specificity. Let $\tau_1 ... \tau_J$ denote the distinct test (visit) times in the data, where $0 = \tau_0 < \tau_1 < ... < \tau_J < \tau_{J+1} = \infty$ - thus, the time axis can be divided into $J + 1$ disjoint intervals, $[\tau_0, \tau_1), [\tau_1, \tau_2), \cdots, [\tau_J, \infty)$. Let $\boldsymbol{S} = (1 = S_1, S_2, \cdots, S_{J+1})^T$, where $S_j = Pr(T > \tau_{j-1})$.

Assuming that an individual's test results are independent conditional on the true event time T and that the tests are scheduled at pre-determined times, [3] showed that the log-likelihood function for a random sample of N subjects can be written as:

$$l(\boldsymbol{\theta}) = log(L(\boldsymbol{\theta})) = \sum_{i=1}^{N} \log(\sum_{j=1}^{J+1} C_{ij}\theta_j) \tag{3.1}$$

where $C_{ij} = p(\boldsymbol{R_i}|\tau_{j-1} \leq T_i < \tau_j, \boldsymbol{t_i})$ or, in other words, C_{ij} is the conditional probability of observing the vector of test results $\boldsymbol{R_i}$ given that the event is known to have occurred in the j^{th} interval. For example, consider a study where all subjects are tested at two times τ_1 and τ_2 with interval probabilities $\theta_1, \theta_2, \theta_3$ corresponding to intervals $[0, \tau_1), [\tau_1, \tau_2), [\tau_2, \infty)$, respectively. Let the constant sensitivity and specificity of the diagnostic test (or self report) be denoted by φ_1 and φ_0, respectively. For the i^{th} subject, if we observe a negative test result at τ_1 and a positive test result at τ_2, then the coefficients of the C matrix can be expressed as $C_{i1} = (1 - \varphi_1)\varphi_1$, $C_{i2} = \varphi_0\varphi_1$, and $C_{i3} = \varphi_0(1 - \varphi_0)$. The likelihood function for this subject is equal to $L_i = (1 - \varphi_1)\varphi_1\theta_1 + \varphi_0\varphi_1\theta_2 + \varphi_0(1 - \varphi_0)\theta_3$.

For the general setting, let $C = [C_{ij}]$ denote the $N \times (J + 1)$ matrix of the coefficients, C_{ij}. For a given dataset, we details with regard to computing the C matrix are given below.

Computing the C matrix: To calculate each element in the C matrix, we first calculate the occurrence matrix O, where $O_{is,j}$ denotes whether or not event has occurred at s^{th} test time for subject i given his event time is in j^{th} interval. O has same number of rows as the the number of subjects and $J + 1$ columns. $O_{is,j} = I(t_{is} \geq \tau_j)$, where $I(.)$ is an indicator function. We obtain the matrix L with same dimensions as O, where $L_{is,j}$ is the likelihood of each single test result R_{is} at each single test time t_{is} given T_i in j^{th} interval,

$$L_{is,j} = p(R_{is}|\tau_{j-1} \leq T_i < \tau_j) = \begin{cases} sensitivity, & R_{is} = 1 \text{ and } O_{is,j} = 1; \\ 1 - sensitivity, & R_{is} = 0 \text{ and } O_{is,j} = 1; \\ 1 - specificity, & R_{is} = 1 \text{ and } O_{is,j} = 0; \\ specificity, & R_{is} = 0 \text{ and } O_{is,j} = 0; \end{cases}$$

Then we have $C_{ij} = \prod_{s=1}^{n_i} L_{is,j}$.

Assuming that the sensitivity and specificity of the diagnostic test are known, the maximum likelihood estimates of $\boldsymbol{\theta}$ can be obtained by numerical optimization of the log likelihood, subject to the constraints $0 \leq \theta_1, \cdots, \theta_{J+1} < 1$ and $\sum_{j=1}^{J+1} \theta_j = 1$.

We re-parameterize the likelihood function in equation (3.1) in terms of the survival function, $\boldsymbol{S} = (1 = S_1, S_2, \cdots, S_{J+1})^T$, where $S_j = Pr(T > \tau_{j-1})$. Since $S_j = \sum_{l=j}^{J+1} \theta_l$, the vector of interval probabilities can be expressed as $\boldsymbol{\theta} = T_r \boldsymbol{S}$, where T_r is the $(J+1) \times (J+1)$ transformation matrix.

$$T_r = \begin{pmatrix} 1 & -1 & 0 & . & . & 0 & 0 \\ 0 & 1 & -1 & . & . & 0 & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 1 & -1 \\ 0 & 0 & 0 & . & . & 0 & 1 \end{pmatrix}_{(J+1)\times(J+1)}$$

Let the $N \times (J+1)$ matrix D be defined as $D_{N\times(J+1)} = C_{N\times(J+1)} T_{r(J+1)\times(J+1)}$. Then, the log-likelihood function (3.1) can be expressed as

$$l(\mathbf{S}) = \sum_{i=1}^{N} \log(\sum_{j=1}^{J+1} D_{ij} S_j), \tag{3.2}$$

where $S_1 = 1$ and $S_2, S_3, \cdots, S_{J+1}$ are the unknown parameters of interest.

Incorporating covariates: Let $\boldsymbol{X}$ denote the $P \times 1$ vector of explanatory variables with corresponding $P \times 1$ vector of regression coefficients denoted by $\boldsymbol{\beta}$. As-

suming the proportional hazards model, we obtain $\lambda(t|\boldsymbol{X}=\boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$, or equivalently, $S(t|\boldsymbol{X}=\boldsymbol{x}) = S_0(t)^{e^{\boldsymbol{x}'\boldsymbol{\beta}}}$. Let $1 = S_1 > S_2 > ... > S_{J+1}$ denote the baseline survival functions (i.e. corresponding to $\boldsymbol{X}=\boldsymbol{0}$), evaluated at the left boundaries of the intervals $[0,\tau_1), [\tau_1,\tau_2), \cdots, [\tau_J, +\infty)$. Then, for subject i, $S_j^{(i)} = (S_j)^{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}$. Thus, the log-likehood function for a random sample of N subjects can be expressed as

$$l(\boldsymbol{S},\boldsymbol{\beta}) = \sum_{i=1}^{N} \log(\sum_{j=1}^{J+1} D_{ij}(S_j)^{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}). \tag{3.3}$$

Statistical inference regarding the parameters of interest $(\beta_1, \cdots, \beta_P, S_2, \cdots, S_{J+1})$ can be made by using asymptotic properties of the maximum likelihood estimator. The estimated covariance matrix of the MLEs can be obtained by inverting the Hessian matrix. Hypothesis tests regarding the unknown parameters can be carried out using the likelihood ratio or Wald test. [40] and our simulation results (not shown here) both show that ignoring the error-prone nature of test results leads to bias in estimation and using likelihood based approaches incorporating measurement error with known sensitivity and specificity can correct the bias.

Note on conditional independence assumption: We note that the conditional independence assumption with regard to the vector of test results for each individual may be more plausible for laboratory-based diagnostic tests. Here we show that under different set of assumptions that are more plausible for self-reported outcomes, we derive a likelihood that is equivalent to that obtained assuming conditional independence. The probability of an individual's vector of test results $\boldsymbol{R_i}$ conditional on the true time of the event of interest can be expressed as:

$$p(\boldsymbol{R_i}|\tau_{j-1} \leq T_i < \tau_j) = \prod_l p(R_l|R_1, \cdots, R_{l-1}, \tau_{j-1} \leq T_i < \tau_j)$$

We assume that the sensitivity and specificity are constant for each self-report as long as there is no prior positive result reported. We further assume that no further

reports are collected following the first positive report of the event of interest (NTFP study design discussed in Section 3.2.4). That is,

$$p(R_l = 1 | R_1 = 0, \cdots, R_{l-1} = 0, \tau_{j-1} \leq T_i < \tau_j) = \begin{cases} sensitivity & t_l \geq \tau_j \\ 1 - specificity & t_l \leq \tau_{j-1} \end{cases}$$

Under these assumptions, the form of the likelihood is identical to that obtained by assuming conditional independence.

3.2.2 Power

Analytical solutions for power calculations can be obtained by deriving expressions for the variance of $\hat{\beta}$ under the alternative hypothesis.

First, for simplicity, we assume that each subject is tested at J times denoted $\tau_1, \tau_2 ... \tau_J$ and that there are no missing tests. This assumption is relaxed in Section 3.2.4, where we allow the possibility of missing tests. Assuming that each test result is binary (positive or negative), there are 2^J possible patterns of test results for each subject. Let r_{ki} denote the number of subjects in the k^{th} exposure group having the i^{th} pattern of test results, where $k = 1, 2$ and $i = 1, 2, \cdots, 2^J$. Let l_{ki} represent the corresponding log-likelihood function. Note that the index i here refers to the index of possible patterns of test results. Then the log-likelihood function for a random sample of N subjects can be expressed as

$$l = \sum_{k=1}^{2} \sum_{i=1}^{2^J} r_{ki} l_{ki}$$

, where

$$l_{ki} = \log(\sum_{j=1}^{J+1} D_{ij} S_{kj})$$

where $\boldsymbol{S}_k = (1 = S_{k1}, \cdots, S_{k(J+1)})$ represents the survival function in the k^{th} group $(k = 1, 2)$, evaluated at visit times $\tau_0, \cdots, \tau_J$. We note that the $2^J \times (J + 1)$ matrix

D can be obtained by enumerating the 2^J possible patterns of test results and then computing the corresponding coefficients as described in Section 3.2.1.

Under the proportional hazards model, the log-likelihood function specific to each exposure group is given by:

$$l_{1i} = \log(\sum_{j=1}^{J+1} D_{ij} S_j)$$

$$l_{2i} = \log(\sum_{j=1}^{J+1} D_{ij} (S_j)^{e^\beta})$$

, where S_j denotes the survival function at the left boundary of j^{th} interval in group 1.

Analytical expressions for the second derivatives of log-likelihood functions for the two-treatment group case (Group 1 Vs Group 2) are presented below: **Group 1:**

$$\frac{\partial^2 l_{1i}}{\partial S_j^2} = \frac{-D_{ij}^2}{(\sum_{l=1}^{J+1} D_{il} S_l)^2}; \frac{\partial^2 l_{1i}}{\partial \beta^2} = \frac{\partial^2 l_{1i}}{\partial S_j \partial \beta} = 0; \frac{\partial^2 l_{1i}}{\partial S_j \partial S_{j'}} = \frac{-D_{ij} D_{ij'}}{(\sum_{l=1}^{J+1} D_{il} S_l)^2}$$

Group 2:

$$\frac{\partial^2 l_{2i}}{\partial S_j^2} = \frac{D_{ij} e^\beta (e^\beta - 1)(S_j)^{e^\beta - 2}}{\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta}} - \frac{(D_{ij} (S_j)^{e^\beta - 1} e^\beta)^2}{(\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta})^2}$$

$$\frac{\partial^2 l_{2i}}{\partial \beta^2} = \frac{\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta} \log(S_l) e^\beta (1 + \log(S_l) e^\beta)}{\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta}} - \frac{(\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta} \log(S_l) e^\beta)^2}{(\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta})^2}$$

$$\frac{\partial^2 l_{2i}}{\partial S_j \partial S_{j'}} = \frac{-D_{ij} D_{ij'} (S_j S_{j'})^{e^\beta - 1} e^{2\beta}}{(\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta})^2}$$

$$\frac{\partial^2 l_{2i}}{\partial S_j \partial \beta} = \frac{D_{ij} (S_j)^{e^\beta - 1} e^\beta (1 + \log(S_j) e^\beta)}{\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta}} - \frac{(\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta} \log(S_l) e^\beta) D_{ij} (S_j)^{e^\beta - 1} e^\beta}{(\sum_{l=1}^{J+1} D_{il} (S_l)^{e^\beta})^2}$$

where $j, j' = 2, 3, \cdots, J+1$ and $j \neq j'$.

Let $\gamma_1, \gamma_2, \gamma_3, \cdots, \gamma_{J+1}$ denote the parameters $\beta, S_2, S_3, \cdots, S_{J+1}$. Then the elements of the expected Fisher information matrix $I_{(J+1)\times(J+1)}$ can be obtained by

$$I_{jj'} = -E(\frac{\partial^2 l}{\partial\gamma_j \partial\gamma_{j'}}) = -\sum_{k=1}^{2}\sum_{i=1}^{2^J} E(r_{ki})\frac{\partial^2 l_{ki}}{\partial\gamma_j \partial\gamma_{j'}} \tag{3.4}$$

Let N_1 and N_2 refer to the sample sizes in groups 1 and 2, respectively. Then the expected number of subjects with the i^{th} pattern of test results in the k^{th} exposure group is

$$E(r_{ki}) = N_k \exp(l_{ki}) = N_k \sum_{l=1}^{J+1} D_{il} S_{kl}$$

Thus,

$$I_{jj'} = -\sum_{k=1}^{2} N_k \sum_{i=1}^{2^J} (\frac{\partial^2 l_{ki}}{\partial\gamma_j \partial\gamma_{j'}} \sum_{l=1}^{J+1} D_{il} S_{kl}) \tag{3.5}$$

Thus, the expected Fisher information matrix can be expressed as a function of the sensitivity and the specificity of the diagnostic test, the survival function at each test time for the reference group, β or hazard ratio (HR) under alternative, and the sample sizes in each group. The covariance matrix corresponding to the MLEs of $\beta, S_2, \cdots, S_{J+1}$ can be obtained by inverting the expected Fisher information matrix, from which we estimate the variance of $\hat{\beta}$. The power can be calculated using the asymptotic normal property of the test statistic,

$$power = p\left(Z > z_{1-\frac{\alpha}{2}} - \frac{\beta}{\sqrt{Var(\hat{\beta})}}\right) + p\left(Z < -z_{1-\frac{\alpha}{2}} - \frac{\beta}{\sqrt{Var(\hat{\beta})}}\right)$$

where Z is standard normal random variable, α is type I error, and $Var(\hat{\beta})$ is the variance of $\hat{\beta}$.

3.2.3 Sample size

We derive expressions for sample size by assuming equal sample size allocation into the two exposure groups, that the the type I error is denoted by α corresponding

to a two-sided hypothesis test, and that the desired power is ρ. We set $N_1 = N_2 = \frac{1}{2}$ in (3.5) to obtain the expected Fisher information matrix for one unit. The expression for total sample size is given by:

$$N = \frac{(z_{1-\alpha/2} + z_\rho)^2 Var_1(\hat{\beta})}{\beta^2}$$

, where the $Var_1(\hat{\beta})$ is obtained as the inverse of the expected Fisher information for one unit.

3.2.4 Incorporating Missing Tests

In this section, we incorporate the possibility of missing test results due to missed visits in deriving expressions for power and sample size. Let J denote the number of scheduled test times, let $\boldsymbol{R}$ denote the $J \times 1$ vector of binary test results and let $\boldsymbol{I}$ denote the $J \times 1$ vector of binary indicators corresponding to the pattern of missingness. That is, $I_j = 0$ denotes that the j^{th} test is missing and $I_j = 1$ denotes that the j^{th} test result is observed. Note that $\boldsymbol{I}$ is fully observed. Let $\boldsymbol{R}_{obs}$ and $\boldsymbol{R}_{mis}$ denote the observed and missing components of the vector $\boldsymbol{R}$. Let Θ denote the parameters governing the likelihood function (i.e. $\Theta = (\boldsymbol{\theta}, \boldsymbol{\beta})$), and let Φ denote the parameters governing the missing mechanism. We assume the missing mechanism is missing at random (MAR), which implies that the missing mechanism is independent of the unobserved results $\boldsymbol{R}_{mis}$. Under this assumption, the joint distribution of $\boldsymbol{R}_{obs}$ and $\boldsymbol{I}$ can be expressed as ([16]),

$$p(\boldsymbol{R}_{obs}, \boldsymbol{I}|\Phi, \Theta) = p(\boldsymbol{I}|\boldsymbol{R}_{obs}, \Phi) p(\boldsymbol{R}_{obs}|\Theta)$$

In this application, $p(\boldsymbol{R}_{obs}|\Theta)$ corresponds to the previously described likelihood function based on the observed data. $\phi = p(\boldsymbol{I}|\boldsymbol{R}_{obs}, \Phi)$ denotes the distribution of the missing mechanism. Since $p(\boldsymbol{I}|\boldsymbol{R}_{obs}, \Phi)$ is independent of the parameters Θ governing the likelihood, this can be ignored when obtaining both the maximum likelihood

estimates of the parameters of interest $(\boldsymbol{\theta}, \boldsymbol{\beta})$, as well as the second derivatives of the likelihood function required for calculating the Fisher information matrix.

In the presence of missing data, the expected number of subjects in k^{th} group with the i^{th} pattern of test results (r_{ki}) is given by

$$\begin{aligned} E(r_{ki}) &= N_k p(\boldsymbol{I}_i \mid \boldsymbol{R}_i, \Phi) p(\boldsymbol{R}_i \mid \Theta_k) \\ &= N_k \phi_i L_{ki} \\ &= N_k \phi_i \sum_{l=1}^{J+1} D_{il} S_{kl} \end{aligned}$$

Thus, when incorporating the possibility of missed visits, the only difference in deriving expressions for sample size is due to the presence of the extra term ϕ_i, corresponding to the distribution of the missing data mechanism.

Below, we consider some common missing mechanisms.

Missing completely at random (MCAR): In this setting, we assume that each test can be independently missing with probability p_{miss}. Let p_{miss} denote the probability of a missing test. For the i^{th} pattern of test results

$$\phi_i = \prod_{j=1}^{J} (1 - p_{miss})^{I_{ij}} (p_{miss})^{1 - I_{ij}} = (1 - p_{miss})^{m_i} (p_{miss})^{J - m_i}$$

, where m_i is the number of non-missing tests.

Missing all tests following the first positive (NTFP): In several studies, no additional diagnostic tests are administered following the first positive test result, indicating the occurrence of the event of interest. This study design is especially useful when the specificity of the diagnostic test is perfect. This study design is referred to as NTFP or 'No Tests after First Positive'. To derive sample sizes for this setting, we assume that at each test time prior to the first positive test result, each test can

be missing with probability p_{miss}. On the other hand, following the first positive test result, all tests are missing with probability 1.

$$\phi_i = p(\boldsymbol{I}_i|\boldsymbol{R}_i, \Phi) = p(I_{i1}|\boldsymbol{R}_i, \Phi)p(I_{i2}|I_{i1}, \boldsymbol{R}_i, \Phi)\cdots p(I_{iJ}|I_{i1}\cdots I_{i,J-1}, \boldsymbol{R}_i, \Phi)$$

If the first positive is at test time J_i, then for $j > J_i$ we have $p(I_{ij} = 0|I_{i1}\cdots I_{i,j-1}, \boldsymbol{R}_i, \Phi) = 1$, and for $j \leq J_i$ we have $p(I_{ij}|I_{i1}\cdots I_{i,j-1}, \boldsymbol{R}_i, \Phi) = (1 - p_{miss})^{I_{ij}}(p_{miss})^{1-I_{ij}}$.

$$\phi_i = (1 - p_{miss})^{m_i}(p_{miss})^{J_i - m_i}$$

, where m_i is the number of non-missing tests. If no positive result is observed, then $J_i = J$.

Censoring: We present the derivation of power and sample size calculations in the presence of censoring or loss to follow-up. We assume non-informative censoring and that the distribution of censoring times is identical for all subjects - under these assumptions, the missing mechanism can be shown to be MAR. The presence of non-informative censoring can be incorporated into power and sample size calculations through the distribution of missingness, denoted by ϕ_i. Let T_c denote the random variable corresponding to censoring time and let the probability of being censored in j^{th} interval is $p(\tau_{j-1} \leq T_c < \tau_j) = c_j$. Thus, the distribution of missing pattern when censoring is present can be expressed as:

$$\phi_i = \sum_{j=J_i+1}^{J+1} c_j p(\boldsymbol{I}_i|\boldsymbol{R}_i, \Phi, \tau_{j-1} \leq T_c < \tau_j)$$

, where J_i represents the index of the last observed test time for i^{th} pattern of test results.

For example, assume that there are 4 test times, $\tau_1, \cdots, \tau_4$, dividing the time axis into 5 intervals. The probability of being censored in the j^{th} interval is denoted by c_j, for $j = 1, \cdots, 5$, where $\sum_{j=1}^{5} c_j = 1$. We assume that each test has a random missing probability of p_{miss} and that the random missing process is independent on the censoring process. Consider a pattern of test results that has an associated missingness pattern $\boldsymbol{I}_i = (1, 0, 1, 0)$, where $I_{ij} = 1$ denotes that a test result is available (observed) at visit τ_j and 0 otherwise. In this example, the last non-missing observation is at τ_3, thus $J_i = 3$. Therefore, in this example, censoring could have occurred in either the 4^{th} interval ($[\tau_3, \tau_4)$) or the 5^{th} interval ($[\tau_4, +\infty)$). Given that the censoring time is in the interval $[\tau_3, \tau_4)$, there are 1 missing and 2 non-missing tests before censoring, thus the probability of the missing pattern is $p(\boldsymbol{I}_i|\boldsymbol{R}_i, \Phi, \tau_3 \leq T_c < \tau_4) = (1 - p_{miss})^2 p_{miss}$. Similarly, given that censoring occurs in the interval $[\tau_4, +\infty)$, we have $p(\boldsymbol{I}_i|\boldsymbol{R}_i, \Phi, \tau_4 \leq T_c < +\infty) = (1 - p_{miss})^2 p_{miss}^2$. Thus, the probability of the missing pattern, is expressed as $\phi_i = c_4(1 - p_{miss})^2 p_{miss} + c_5(1 - p_{miss})^2 p_{miss}^2$.

3.2.5 Comparing perfect versus imperfect tests

We present an analysis of the trade-off between perfect and imperfect diagnostic tests, by considering parameter settings that reflect the characteristics of diabetes self-reports in the WHI study.

Motivating application: The WHI recruited postmenopausal women (N=161,808) aged 50-79 at 40 clinical centers across the U.S. from 1993-1998 with ongoing follow-up. Prevalent diabetes was ascertained by self-report at the baseline visit. Similarly, incident diabetes was also determined by self-reports obtained at each annual visit. At each visit, participants were asked whether she has ever received a physician diagnosis of and/or treatment for diabetes when not pregnant. The accuracy of self-report results have been shown in Chapter 2 to be 0.61 and 0.995 for sensitivity and specificity respectively.

Results: In this analysis, we present the trade-off in power when comparing a perfect test to tests with varying degrees of imperfect sensitivity and/or specificity. We assume two treatment groups of interest, equal sample size allocation to the two groups and that the type I error is fixed at 0.05 for a two-sided hypothesis test. Further, to specify the survival function at each test time, we assume that the time to the event of interest (T) within each group follows an exponential distribution, where the parameter of the exponential distribution with each exposure group is determined by the survival function at the last test time, S_{J+1} (i.e. 1 - cumulative incidence in group 1).

Figure 3.1 shows the power versus sample size curves corresponding to different values of (sensitivity, specificity), assuming that the hazard ratio (HR) between the two groups is equal to 1.25. We assumed that the duration of the study is 8 years, that tests (visits) are scheduled either annually or every 4 years, and that the cumulative incidence in the reference group is either $1 - S_{J+1} = 0.1$ or $1 - S_{J+1} = 0.5$ - this corresponds to a mean event time of 75.9 years or 11.5 years, respectively.

As the frequency of testing increases, the curves corresponding to imperfect tests tend towards the curve of the perfect test, indicating an increase in power with more frequent tests even in the presence of error-prone diagnostic tests. As expected, when the cumulative incidence rates are larger, power increases for all values of (sensitivity, specificity). Moreover, when the cumulative incidence rate is low ($S_{J+1} = 0.9$), reduction in specificity has a significant impact on power. However, for this setting, a corresponding reduction in sensitivity does not have a similarly big effect on power. On the other hand, when cumulative incidence rates are higher ($S_{J+1} = 0.5$), both sensitivity and specificity have similar impact on power. For example, we consider the power curves corresponding to diagnostic tests with (sensitivity, specificity) of (1.00, 0.75) versus (0.61, 0.995) - the power curve corresponding to the test with lower sensitivity but higher specificity (i.e. (0.61, 0.995)) has higher power when cumulative

incidence rates are low ($S_{J+1} = 0.9$) but has lower power when cumulative incidence rates are high ($S_{J+1} = 0.5$). In settings of low cumulative incidence rates, reduced specificity results in a large number of false positive results that have a deleterious effect on power. However, in settings of larger cumulative incidence rates (≈ 0.5), reduced sensitivity would result in significant number of false negative results, whereas reduced specificity would also result in a significant number of false positive results - thus, the effects of reduced sensitivity and specificity are more pronounced when cumulative incidence rates are larger. Diabetes self-reports in the WHI have sensitivity and specificity of approximately 0.61 and 0.995, respectively - however, since the cumulative incidence of diabetes over a 8-year follow up period until 2005 is relatively low within the WHI ($\leq 10\%$), the low sensitivity of self-reports is expected to not have a drastic impact on statistical power.

3.3 Study Design

In this section, we address the following questions that arise during design of biomedical investigations, namely (1) What is the minimum sample size to achieve a desired level of statistical power? (2) What is the optimal number tests per subject? (3) Should the study incorporate different testing schedules for different subjects in the study? and (4) Should further tests be administered after the first positive test result is observed?. In all the examples we assume there are no missing tests unless otherwise specified, that type I error and power are fixed at 0.05 and 0.90, respectively for a two-sided hypothesis test.

3.3.1 Sample size

The desired sample size typically depends on the specified statistical power and minimum clinically meaningful difference between groups (hazard ratio) as well as the characteristics of the diagnostic test (sensitivity,specificity), cumulative incidence in

each group, and the testing frequence and associated schedule. Figure 3.2 illustrates the effects of varying hazard ratio, cumulative incidence and frequency of tests (visits) on resulting sample size estimates for diagnostic tests with varying levels of sensitivity and specificity. For each setting, we assume that the desired power is 0.90, and compare the perfect diagnostic test with sensitivity and specificity of (1.00, 1.00), to imperfect tests with values of sensitivity and specificity given by (i) (0.75, 1.00); (ii) (1.00, 0.75); and (iii) (0.61, 0.995), which corresponds the characteristics of self-reports of incident diabetes in the WHI ([34]).

Figure 3.2(a) presents how sample size varies with hazard ratio, assuming statistical power of 0.90, equally spaced 4 test times (visits) during the study period and a cumulative incidence of 0.10 (or $S_{J+1} = 0.9$) in the reference group. As expected, the required sample size decreases with increasing values of hazard ratio. In addition, we observed that the diagnostic tests with perfect specificity have the lowest required sample size (all other factors assumed fixed), whereas the diagnostic tests with imperfect specificity require the largest sample sizes. In particular, sensitivity has little effect on sample size estimates for this setting - there is little difference between the perfect test and (0.75, 1.00) test. These observations are driven by the low cumulative incidence of 0.10 - in this case, decreasing specificity results in a corresponding large number of false positive test results which in turn results in a correspondingly larger sample size in order to detect statistically significant differences between the two groups of comparison.

Figure 3.2(b) shows how sample size depends on the cumulative incidence in the reference group, where we fix statistical power to equal 0.90, the hazard ratio to equal 1.25 and assume four equally spaced test times (visits) during the study period. As expected, regardless of the characteristics of the diagnostic tests, sample size decreases as cumulative incidence is increased. Since low specificity of a diagnostic test results in a high rate of false positive tests and low sensitivity results in high rate of false

negative test results, the impact of imperfect sensitivity and specificity on sample size depends on the true positive proportion, which is determined by the cumulative incidence. As discussed earlier, when cumulative incidence is low, specificity plays a dominant role with respect to sample size. When cumulative incidence is higher resulting in more true positives, the relative importance of false positives as a result of low specificity is diminished. In contrast, the rate of false negative test results as a result of low sensitivity has a larger impact on the sample size when there are fewer true negatives (i.e. high cumulative incidence). This discussion is supported by the observation that the curves for (1.00, 0.75) and (0.61, 0.995) cross at a cumulative incidence rate of approximately 0.4 - when cumulative incidence in lower than 0.4, the required sample size for the test with imperfect specificity [(1.00, 0.75)] is larger than the test with imperfect sensitivity [(0.61, 0.995)]. However, this trend is reversed when cumulative incidence is larger than 0.4, in which case the test with imperfect sensitivity [(0.61, 0.995)] requires a larger sample size than the test with imperfect specificity [(1.00, 0.75)]. We also observed that for cumulative incidence below 0.65, the required sample size for the test (0.75, 1.00) is smaller than that for the test (1.00, 0.75), illustrating that even when the cumulative incidence is high, at early test times the majority of subjects may still be event-free or true negative and thus, a test with high specificity will still result in lower sample sizes. In practice, when cumulative incidence over the study period is modest, it is more important to choose a test with high specificity than high sensitivity. Since a higher cumulative incidence can decrease the resulting sample size, one practical option may be to increase the follow-up time assuming that is feasible and cost-effective to do so.

Figure 3.2(c) presents how the required sample size changes with the number of tests, assuming equally spaced tests (visits) during the follow up period. The statistical power was fixed at 0.90, hazard ratio was fixed at 1.25 and the cumulative incidence in the reference group at 0.10 (or equivalently, $S_{J+1} = 0.9$). For the case of

perfect test [(1.00, 1.00)], the required sample size does not increase when the number of tests increases. As more tests are administered, the underlying survival function can be estimated more precisely but at an increased cost or loss in power resulting from an increase in the number of parameters estimated. For imperfect tests, the required sample size decreases with increased number of tests but reaches a limiting value when the number of tests becomes large. Thus the gain in statistical power reaches a limiting value, which is determined by how well the test can discriminate the survival distributions of the two groups being compared. In practice, when diagnostic tests are imperfect, increasing the frequency of testing can reduce the required sample size upto a limit.

3.3.2 Optimal number of tests per subject

As seen in Section 3.3.1, for an imperfect diagnostic test, as the number of tests administered over the duration of the study period increases, the required sample size decreases upto a limiting value. However, increasing the frequency of testing can also increase the total cost of a study as a result of the more frequent testing or visits. Here, we consider the tradeoff between increasing the number of subjects versus frequency of tests per subject, with regard to the total cost of the study.

Assume C_0 is the cost of recruiting a subject into the study including the cost of a diagnostic test at baseline and C_1 is the additional cost for each test (visit) per subject. Let N denote total sample size and J denote the number tests per subject, then the total cost C can be expressed as:

$$C = N(C_0 + JC_1)$$

Assume $C_0 = 1$ and let $C_1 = 1$ for perfect test and $C_1 = 0.1$ for self-report. That is, we assume that collecting a self-reported questionnaire costs 10% of the corresponding cost of a laboratory based diagnostic test result, which is assumed to be the gold

standard . Figure 3.3 presents how total cost C changes with increasing frequency of tests (visits), by comparing a perfect diagnostic test to self-reports with sensitivity and specificity of 0.61 and 0.995, respectively. For each type of test (self report Vs perfect test), we obtain the total cost C by calculating N assuming J equally spaced tests (visits) during the study period, that the hazard ratio is 1.25, the cumulative incidence in the reference group is 0.10 (i.e. $S_{J+1} = 0.9$) and the desired power is 0.90. We observe that self-report has much lower total cost than a perfect test and that the cost is minimized at $J = 3$ visits.

3.3.3 Varying schedule of testing

Our previous results in Section 3.3.2 indicate that performing too many tests per subject may not be cost-effective in comparing group effects. However, performing too few tests per subject would limit our ability to estimate the survival distributions within each group. One strategy to overcome this limitation is to assign different test schedules for different subjects in the study. For example, if study follow up is 4 years, one half of the subjects in each group can be tested at years 2,4 and the other half of the subjects tested at years 1, 3. In this case, the survival function within each group can be estimated at times 1,2,3,4; however, each subject is tested only at two time points during follow-up. This more complex design may come with increased administrative costs and may result in some loss in efficiency (or corresponding increase in sample size) when compared to the design where all subjects are tested twice according to the same schedule.

Assume we have $2J$ possible test time points equally spaced during the study period, we compare two schedules. In schedule 1, all subjects are tested at times $2, 4, \cdots, 2J$. In schedule 2, half of the subjects are tested at times $1, 3, \cdots, 2J-1$ and another half are tested at times $2, 4, \cdots, 2J$. In both schedules, each subject receives J tests. Figure 3.4(a) shows how the ratio of sample sizes of schedule 2 to

schedule 1 depends on the hazard ratio. Here, we assumed $J = 4$ or 8 equally spaced possible test time points during the study period, statistical power is fixed at 0.90 and the cumulative incidence in the reference group is fixed at 0.10 ($S_{J+1} = 0.9$). The relative increase in sample size for schedule 2 compared to schedule 1 is modest and remains approximately constant as a function of hazard ratio for this setting. Figure 3.4(b) shows how the relative sample size depends on the number of equally spaced test points, $2J$. These results were generated by setting power to equal 0.90, the hazard ratio to equal 1.25 and cumulative incidence in the reference group to equal 0.10 ($S_{J+1} = 0.9$). We observed that with increasing frequency of tests, the loss of efficiency for schedule 2 relative to schedule 1 becomes progressively smaller. However, if the frequency of testing is already large enough, using a more complex study design involving different schedules for different subjects may not provide additional value in terms of estimation of the survival distribution. Figure 3.4(c) shows how the relative sample size depends on cumulative incidence, where we assume that there are 8 test points and hazard ratio is 1.25. We observed that the relative increase in sample size for schedule 2 compared to schedule 1 is slightly decreased with increased cumulative incidence.

In summary, using different schedules for different subjects will result in a finer estimation of the survival function within each group, but this is accompanied with an increase in sample size in addition to the burden of implementing a more complex study design. In practice, we recommend that one first determines the optimal number of tests per subject assuming a fixed cost as illustrated in Section 3.3.2. If the optimal number of tests is small and better estimation of the underlying survival is of value, we recommend further investigation into more complex testing paradigms involving varying testing schedules as illustrated in this section.

3.3.4 Stop testing after first positive result is observed

When specificity is perfect, the first positive result indicates the occurrence of event with probability 1. Test results after the first positive are non-informative and thus unnecessary. However, when specificity is imperfect, tests after first positive can provide additional information and increase power.

Here, we compare two designs: In the first design (Design 1), each subject is tested at all pre-specified test times. In the second design, each subject receives all tests until the first positive is observed and there is no test performed afterwards. We refer to the second design as 'NTFP' (or "No Test following First Positive''). Figure 3.5 presents the relative efficiency of NTFP when compared to Design 1 (i.e. sample size ratio of NTFP to Design 1). We observe that the relative sample sizes increase rapidly with decreasing specificity, driven by the low cumulative incidence of 0.1 (i.e. $S_{J+1} = 0.9$). This is because when specificity is low for the NTFP design, early false positive test results prevent further tests, thus leading to a significant loss in information.

In summary, when specificity is perfect, NTFP design is optimal. When specificity is less than perfect and cumulative is low, NTFP can lead to substantial loss in power. In the case when specificity is only slightly less than perfect, sample sizes of MCAR and NTFP may be comparable - however, using NTFP requires fewer tests per subject and thus may lead to a more cost-effective study design. A cost analysis similar to Section 3.3.2 can be conducted to compare these two designs.

3.4 Unknown sensitivity and specificity

In this section, we consider studies that incorporate an imperfect diagnostic test with unknown sensitivity and specificity. We assume that such studies would also include a subset of subjects who would be enrolled in a validation study, for whom

both the imperfect and a perfect diagnostic test are administered at pre-determined time points throughout the duration of the study.

Notation: Let φ_1 and φ_0 denote unknown sensitivity and specificity of the imperfect diagnostic test, respectively. Let N denote the number of subjects in the study and let $N_1 < N$ denote the number of subjects enrolled in the validation study. For the i^{th} subject, we let $\boldsymbol{t}_i$ and $\boldsymbol{R}_i$ denote the $1 \times n_i$ vectors of pre-scheduled test times and corresponding binary test results from the imperfect diagnostic test, respectively. For the i^{th} subject in the subgroup of N_1 subjects in the validation study, we let $\boldsymbol{V}_i$ denote the $1 \times n_i$ vector of binary test results from the perfect diagnostic test, administered at the pre-scheduled test times $\boldsymbol{t}_i$. For simplicity, we present this derivation assuming that there are no missed visits and no censoring. We let $\boldsymbol{X}_i$ denote the $P \times 1$ vector of explanatory variables measured on the i^{th} subject, with corresponding $P \times 1$ vector of regression coefficients denoted by $\boldsymbol{\beta}$.

Likelihood: Let p_v denote likelihood function corresponding to the N_1 subjects who are given both perfect and imperfect tests and let p_u denote likelihood function corresponding to the $N - N_1$ subjects who are scheduled to receive only the imperfect diagnostic test. Then the log-likelihood function for the N subjects can be expressed as:

$$l(\boldsymbol{S}, \boldsymbol{\beta}, \varphi_1, \varphi_0) = \sum_{i=1}^{N_1} \log(p_v(\boldsymbol{V}_i, \boldsymbol{R}_i \mid \boldsymbol{S}, \boldsymbol{\beta}, \varphi_1, \varphi_0, \boldsymbol{x}_i)) + \sum_{j=1}^{N-N_1} \log(p_u(\boldsymbol{R}_j \mid \boldsymbol{S}, \boldsymbol{\beta}, \varphi_1, \varphi_0, \boldsymbol{x}_j))$$

, where $\boldsymbol{S}$ is as defined in equation (3.3). The likelihood function p_u has same form as in equation (3.3), with the exception that the sensitivity and specificity parameters φ_1, φ_0 are unknown. The likelihood function for the validation set can be further expressed as,

$$p_v(\boldsymbol{V}_i, \boldsymbol{R}_i \mid \boldsymbol{S}, \boldsymbol{\beta}, \varphi_1, \varphi_0, \boldsymbol{x}_i) = p(\boldsymbol{V}_i \mid \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{x}_i) p(\boldsymbol{R}_i \mid \boldsymbol{V}_i, \varphi_1, \varphi_0)$$

$p(\boldsymbol{V}_i \mid \boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{x}_i)$ represents likelihood function for the perfect test results only, which is special case of previous derived likelihood function (3.3) when both sensitivity and specificity are 1. $p(\boldsymbol{R}_i \mid \boldsymbol{V}_i, \varphi_1, \varphi_0)$ represents the conditional likelihood of the observed test results from the imperfect diagnostic test conditional on the test results from the perfect diagnostic test - this is product of terms φ_1, $1-\varphi_1$, φ_0, $1-\varphi_0$. Thus, this conditional probability allows the estimation of the sensitivity and specificity of the imperfect diagnostic test (i.e. φ_1, φ_0).

Power and sample size calculations: To derive power and sample size, we obtain the expected Fisher Information matrix by enumerating all possible patterns of test results in both the validation study as well as the subjects receiving only the imperfect diagnostic test. We present the key steps for calculating the expected Fisher information matrix by considering separately the subgroup of N_1 subjects in the validation study and the subgroup of $N-N_1$ subjects receiving only the imperfect diagnostic test. The mathematical derivations are straightforward but algebraically tedious, so we do not show them here. Instead, we present the key steps in the approach leading to the derivation.

First, we consider the subgroup of $N-N_1$ subjects who receive only the imperfect diagnostic test. This setting is identical to that discussed in previous sections, with the exception that the sensitivity and specificity of the imperfect diagnostic tests denoted by φ_1, φ_0 are unknown parameters. We refer to equation (3.5) for the calculation of the expected Fisher information matrix, where the dimension of the expected Fisher information matrix increases from $(J+1) \times (J+1)$ to $(J+3) \times (J+3)$ due to the inclusion of additional parameters φ_1, φ_0. Note that the coefficient matrices C and D depend on φ_1, φ_0, and thus are not constant in this setting. As shown in Section 3.2.1, each element in matrix C is a product of powers of $\varphi_1, 1-\varphi_1, \varphi_0$ and $1-\varphi_0$. Thus, we can calculate the first and second derivatives on sensitivity and specificity for each element in the C matrix and hence the D matrix by transformation. We basically

need to derive 5 additional derivative matrices for D matrix: two first derivative matrices and three second derivative matrices on sensitivity and specificity. With the derivative matrices, we can easily calculate additional elements involving sensitivity and specificity for the expected Fisher Information matrix.

Second, we consider the subgroup of N_1 subjects in the validation study. To enumerate all possible patterns of test results $(\boldsymbol{V}, \boldsymbol{R})$, we first enumerate all possible patterns of test results from the perfect diagnostic tests (i.e. $\boldsymbol{V}$) - in this case, we can refer to the discussion in previous sections by setting the sensitivity and specificity to equal 1.0 corresponding to a perfect test. Conditional on each pattern of perfect test results $\boldsymbol{V}$, we enumerate all possible patterns of test results from the imperfect diagnostic test ($\boldsymbol{R}$). Let $\boldsymbol{V}_i$ denote i^{th} possible pattern of test results from the perfect test and $\boldsymbol{R}_{il}$ denote l^{th} possible pattern of test results from the imperfect test *conditional* on $\boldsymbol{V}_i$. As in equation(3.5), the Fisher information matrix can be expressed as:

$$I_{jj'} = -\sum_{k=1}^{2} N_k \sum_i \sum_l \frac{\partial^2 \log(p_{vk}(\boldsymbol{V}_i, \boldsymbol{R}_{il}))}{\partial \gamma_j \partial \gamma_{j'}} p_{vk}(\boldsymbol{V}_i, \boldsymbol{R}_{il})$$

where p_{vk} denotes the is likelihood function corresponding to the k^{th} treatment group $(k = 1, 2)$ in the validation study. The expected Fisher information is obtained as the sum of the expected Fisher information matrices corresponding to the two subgroups, namely those in the validation study and those in receiving only the imperfect diagnostic test. Missed visits can be incorporated into sample size calculations as described previously by incorporating the probability of each pattern of missingness (ϕ_i).

Results: We illustrate the effect of unknown sensitivity and specificity by comparing the following two scenarios: all subjects are tested using an imperfect diagnostic test with known sensitivity and specificity Versus a study in which all subjects receive an imperfect diagnostic test with unknown sensitivity and specificity (assuming no

validation set in this case). In Figure 3.6(a), we present the relative sample size as a function of the number of tests (visits) scheduled during the study, where the relative sample size (or sample size ratio) defined as the ratio of the sample size when sensitivity and specificity are unknown to the sample size when sensitivity and specificity are known. We assume no missing tests, that the true sensitivity and specificity of the imperfect diagnostic test are 0.61 and 0.995, respectively, the hazard ratio is 1.25, power is set at 0.90 and type I error is assumed to be 0.05 corresponding to a two-sided hypothesis test. For both high ($S_{J+1} = 0.5$) and low ($S_{J+1} = 0.9$) rates of cumulative incidence, we observed that when the number of tests is greater than 2, the relative sample size is close to 1, indicating that the loss of power due to unknown sensitivity and specificity is negligible when there are sufficient number of tests scheduled per subject. A similar result was observed when incorporating the effects of missingness under the MCAR mechanism. However, when the NTFP study design is assumed in settings of low cumulative incidence rates, the sample size estimates when sensitivity and specificity are unknown are dramatically larger when compared to the setting when sensitivity and specificity is known.

In Figure 3.6(b), we evaluate the effect of incorporating a validation study of sample size N_1, in which subjects receive both a perfect and imperfect diagnostic test at all scheduled visits. We assume no missed visits and consider total sample sizes for two study designs - a study in which all subjects are tested at all scheduled test (visit) times (‘ Design 1’) Versus a study in which testing ceases following the first positive test result (‘ NTFP”). Fig 3.6(b) shows the total sample size (N) for each setting as function of the proportion of subjects included in a validation study ($\frac{N_1}{N}$). In general, incorporating a validation study can reduce sample size. As shown previously, the loss in power due to unknown sensitivity and specificity under Design 1 is negligible - thus, in this case, the reduction of sample size is driven by more precise estimation of the treatment effect resulting from the inclusion of a validation study. For the NTFP

design, the use of validation set can greatly reduce the sample size even when the proportion of subjects included in the validation study is modest.

3.5 Discussion

In this chapter, we have presented methods to estimate sample size and power applicable to studies in which an error-prone diagnostic procedure is administered sequentially to ascertain disease status. The methods developed in this chapter are motivated by self-reported outcomes of diabetes in the WHI. In our development, we illustrated the trade-off with regard to sample size and power when comparing perfect and imperfect tests. We observed that in studies with low rates of cumulative incidence (in the order of 10%), using a diagnostic test with low specificity results in dramatic increase in required sample size. However, when cumulative incidence rates are larger, both imperfect sensitivity and specificity result in decreased power for a given sample size. We have also illustrated the effects of various factors related to study design that influence the power and samples size, including (1) sensitivity and specificity of test, (2) hazard ratio, (3) testing frequency, (4) cumulative incidence during the study period, (5) total cost, (6) varying testing schedules for different subgroups within the study, and (6) whether or not to stop testing after observing the first positive test result. Lastly, we extended our methods to incorporate settings in which a diagnostic test with unknown sensitivity and specificity is used. The methods illustrated in this chapter can be readily implemented using our freely available R software package *icensmis* ([20]), which can be downloaded from the Comprehensive R Archive Network (CRAN).

Our proposed methods can be generalized in several ways. In some studies, the diagnostic procedure used at baseline or study entry is subject to imperfect sensitivity and specificity. For example, in the WHI, diabetes status at baseline was assessed through self reports. A study by [34] found that the negative predictive value of

prevalent diabetes at baseline in the WHI was approximately 97% - in other words, 3% of women who self-reported as being diabetes free were in fact diabetic. In this case, some subjects whose events have already occurred at baseline are included in the study. This specific setting can be readily incorporated into power and sample size calculations as shown in Appendix A.

In other studies, several diagnostic tests with varying sensitivity and specificity may be used, where the sensitivity and specificity values can depend on test times and/or subject-specific attributes. These more general settings can be accommodated through appropriate modifications to the C matrix.

Missing visits are common in practical applications, and our methods allow the specification of the missing probabilities at each test time to estimate the power and sample size. Our methods can be easily extended to allow the missing probabilities to vary with treatment group. Finally, our methods can also be extended to k-sample settings. Lastly, we note that the computation complexity of our proposed method grows rapidly with the number of test times. In our analysis, we found that analysis for settings up to 15 test times are computationally tractable using currently available computing power on a standard desktop.

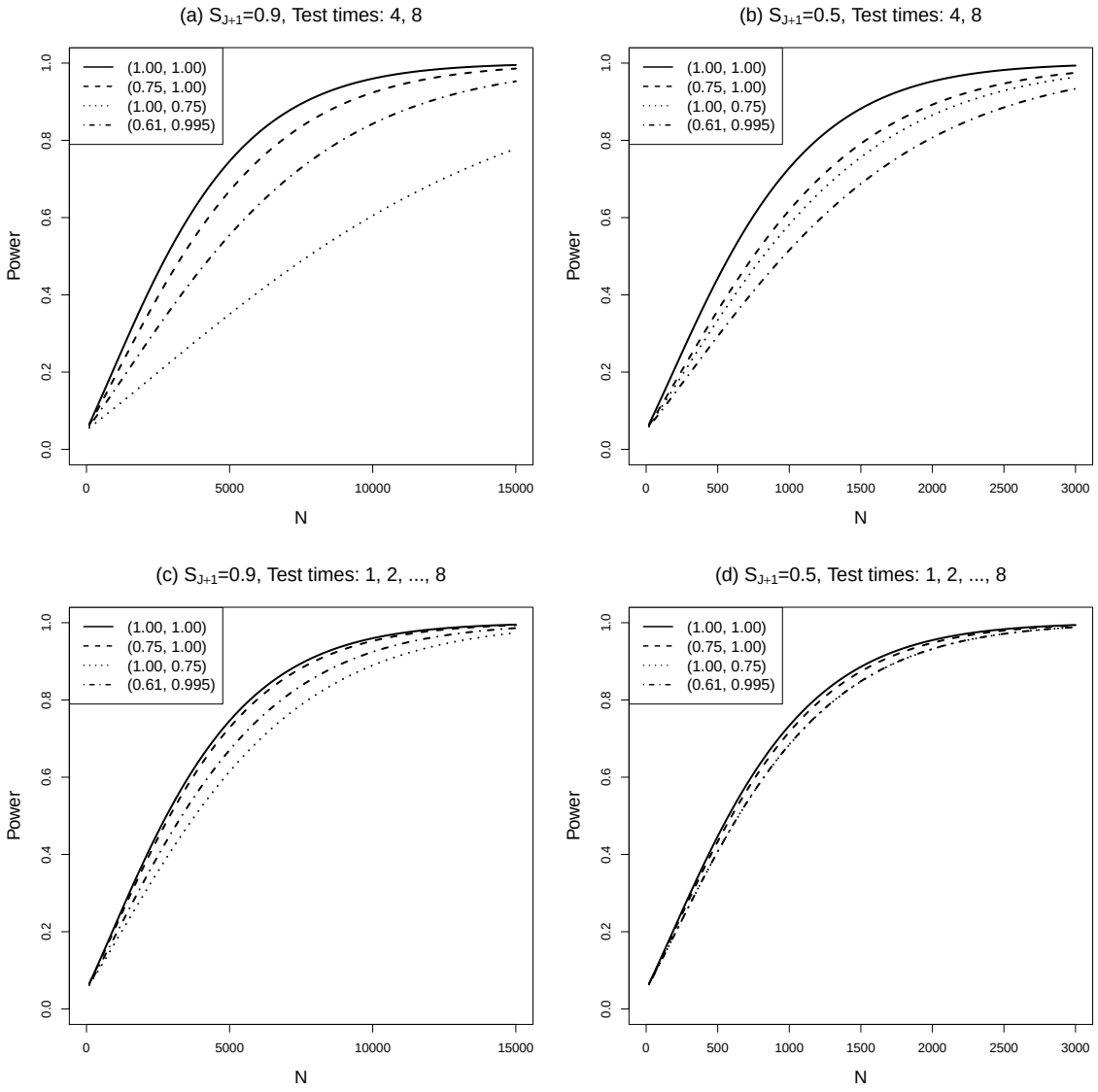

Figure 3.1. Comparison of power versus total sample size (N) for different values of the (sensitivity, specificity) of the diagnostic test, with varying cumulative incidence and testing schedules. The results are based on assuming that there are no missed visits, HR=1.25 and type I error is fixed at 0.05, corresponding to a two-sided hypothesis test.

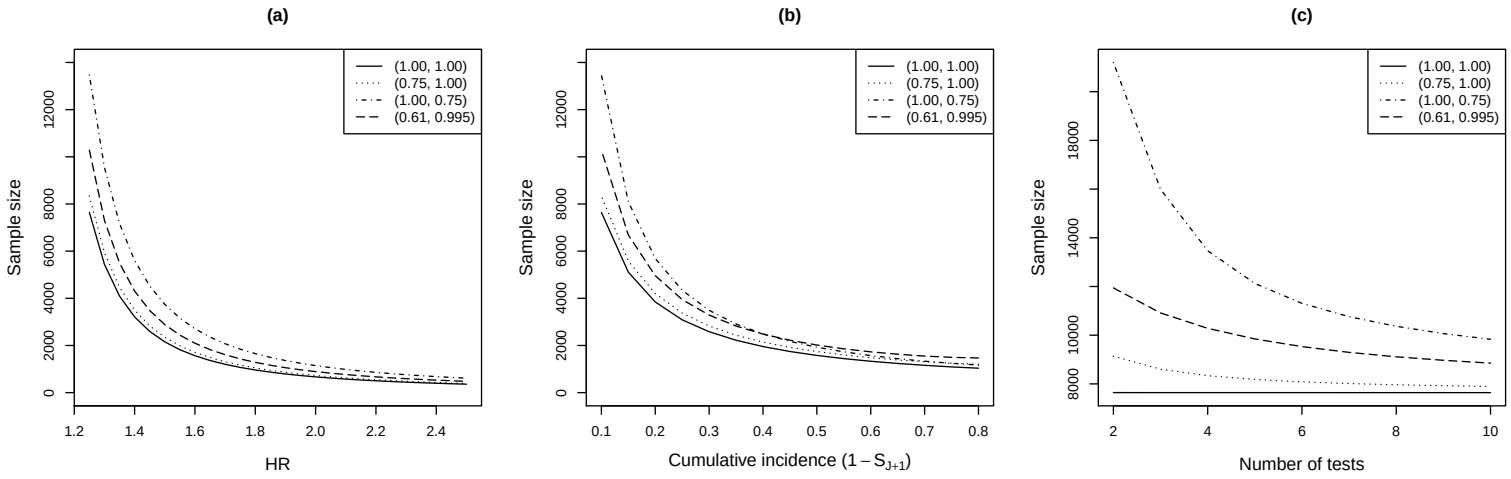

Figure 3.2. Effects of hazard ratio, cumulative incidence, and number of tests with respect to sample size for different values of (sensitivity, specificity). The results are based on assuming no missing tests, that type I error and power are fixed at 0.05 and 0.90, respectively, corresponding to a two-sided hypothesis test. (a) Sample size as a function of hazard ratio, assuming $S_{J+1} = 0.9$ and 4 equally spaced tests during the study period. (b) Sample size as a function of cumulative incidence of baseline group (1 - S_{J+1}), assuming HR = 1.25 and 4 equally spaced tests during the study period. (c) Sample size as a function of the number of equally spaced tests during study period, assuming that HR=1.25 and $S_{J+1} = 0.9$

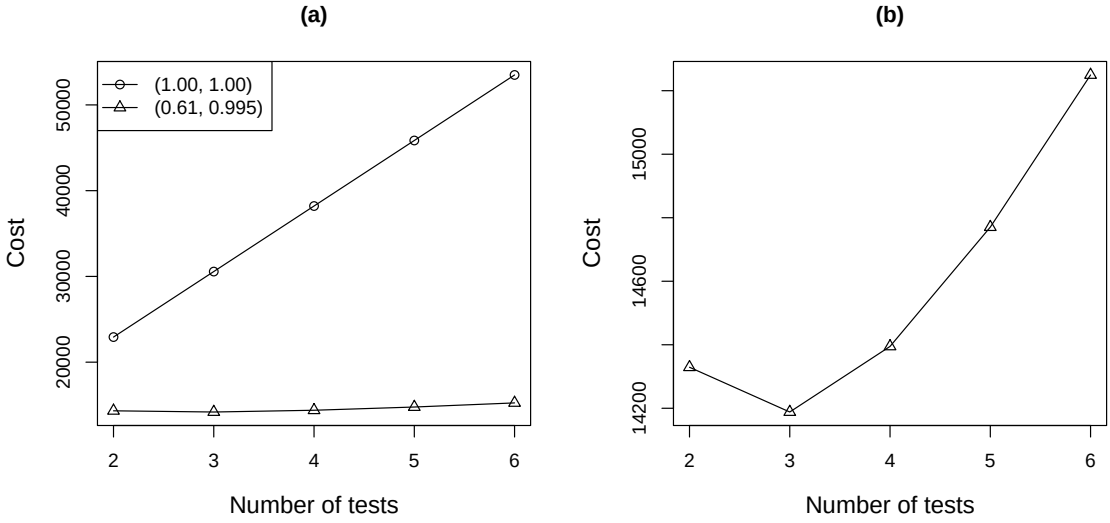

Figure 3.3. Total cost as function of the number of tests. The results are based on assuming no missing tests, that type I error and power are fixed at 0.05 and 0.9, respectively, corresponding to a two-sided hypothesis test. (a) Assume the recruitment and administration cost for each subject is 1. Assume that the cost of a single perfect test is 1.00, with (sensitivity, specificity) given by (1.00, 1.00), and the cost for a single self-report is 0.1, with corresponding (sensitivity, specificity) equal to (0.61, 0.995). The results are based on assuming that HR is fixed at 1.25 and $S_{J+1} = 0.9$. (b) The plot corresponding to self-reports shown in (a) is displayed using a narrower Y-axis range.

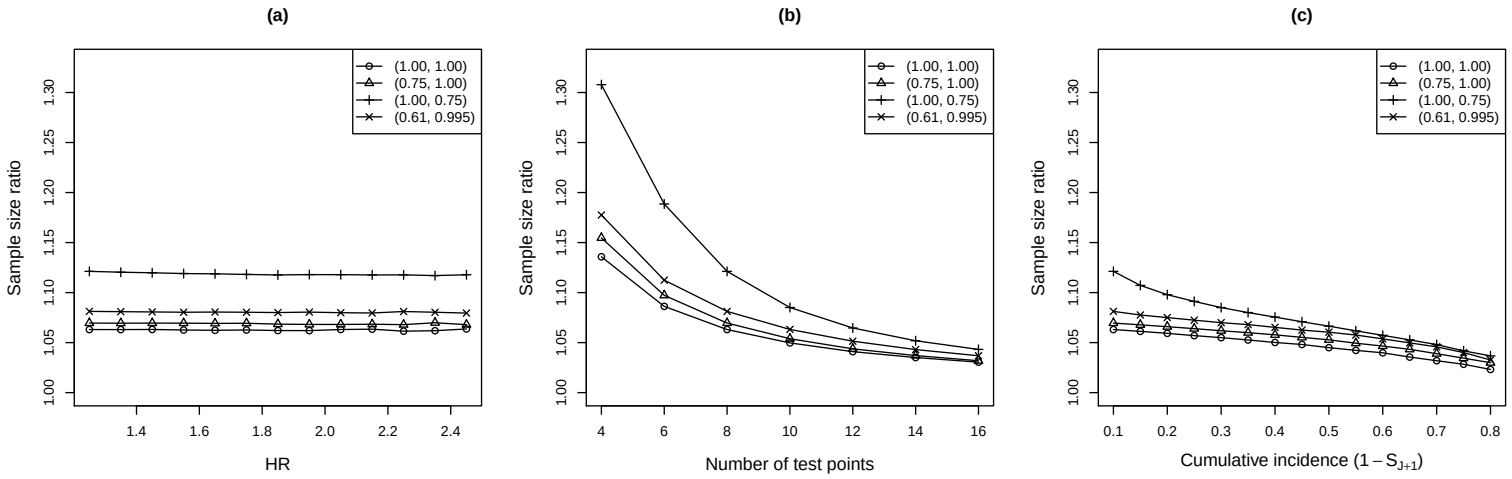

Figure 3.4. Relative efficiency of different testing schedules. Sample size ratio is defined as the ratio of sample size corresponding to schedule 2 relative to the sample size for schedule 1. The results are based on assuming no missing tests and that type I error and power are fixed at 0.05 and 0.90, respectively, corresponding to a two-sided hypothesis test. (a) Sample size ratio as a function of hazard ratio, assuming there are 8 equally spaced tests and $S_{J+1} = 0.9$. (b) Sample size ratio as function of number of equally spaced tests, assuming HR=1.25 and $S_{J+1} = 0.9$. (c) Sample size ratio as function of cumulative incidence for the study duration, assuming that there are 8 equally spaced tests and HR=1.25.

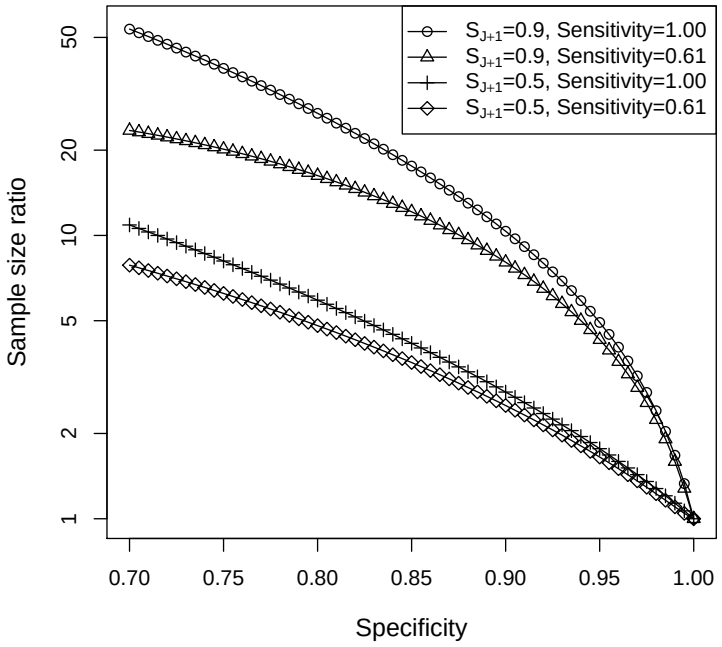

Figure 3.5. Relative efficiency of NTFP when compared to the Design 1. Relative efficiency or sample size ratio is calculated as the ratio of sample size for NTFP relative to Design 1 and is shown as a function of specificity of imperfect diagnostic test. The results are based on assuming no missing tests, that type I error and power are fixed at 0.05 and 0.90, respectively, corresponding to a two-sided hypothesis test, HR=1.25 and that there are 8 equally spaced test times over the study period.

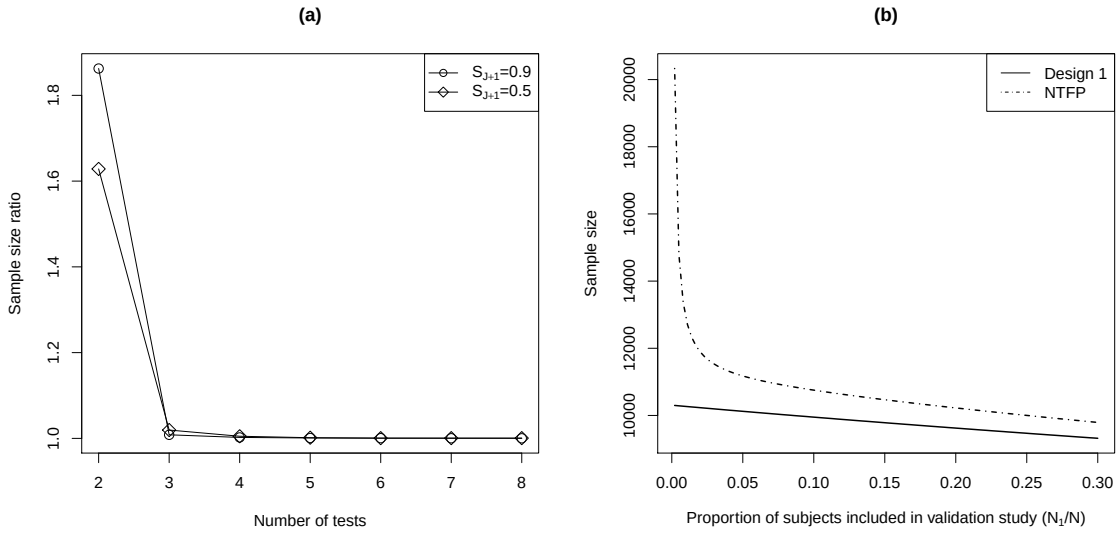

Figure 3.6. Sample sizes when sensitivity and specificity are unknown.
Panel (a): Sample size ratio (or relative efficiency) as a function of number of tests (visits). Sample size ratio (Y-axis) is defined as the ratio of sample size for studies including diagnostic tests with unknown sensitivity and specificity relative to sample size for studies including diagnostic tests with known sensitivity and specificity. Results are based on the following assumptions: type I error and power are fixed at 0.90 and 0.05, respectively, corresponding to a two-sided hypothesis test, sensitivity is 0.61, specificity is 0.995, HR=1.25, and that there are no missing tests.
Panel (b): Sample size as function of proportion of subjects (N_1/N) included in a validation study. Results are based on the following assumptions: type I error and power are fixed at 0.90 and 0.05, respectively, corresponding to a two-sided hypothesis test, sensitivity is 0.61, specificity is 0.995, HR=1.25, $S_{J+1} = 0.9$, 4 tests scheduled during the study period and no missed visits.

CHAPTER 4

VARIABLE SELECTION IN HIGH DIMENSIONAL DATASETS IN THE PRESENCE OF ERROR-PRONE DIAGNOSTIC TESTS

4.1 Introduction

In previous chapters, we have addressed issues of covariate effect estimation and study design in settings of error-prone time to event outcomes. In this chapter, we consider settings characterized by high dimensional data. The motivating application in this chapter is biomarker discovery for diabetes using data from the WHI clinical and observational study SHARe, which includes extensive genotypic ($>$ 900,000 SNPs) and phenotypic data on 12,008 African American and Hispanic women. We propose and apply two approaches for variable selection in high dimensional datasets, while accounting for error-prone, self-reported outcomes.

While a rich literature exists to handle estimation and hypothesis testing in the presence of error-prone survival outcomes, none of the previous studies have considered the setting of high-dimensional data, in which the number of features (p) far exceeds the number of subjects (n). In this setting, likelihood based estimation approaches are intractable. The lasso ([51])and Bayesian variable selection (BVS) ([41], [17]) approaches are two methods that are appropriate for variable selection in high-dimensional datasets.

The lasso algorithm proceeds by adding the absolute value of the coefficients as a penalty term to the objective function. This results in a sparse solution with some coefficients set to zero, thus performing automatic variable selection. The idea of lasso is general and has been applied to various statistical models including linear

and generalized linear models, and the Cox proportional hazards model ([51, 52]). In a recent study, [55] demonstrated the use of the lasso to select relevant biomarkers from a large number of SNPs in a case-control, gene-disease mapping study. The lasso problem can be solved using very computationally efficient algorithms such as the pathwise coordinate descent algorithm ([14, 15, 46]), which is implemented in the R package *glmnet*.

The BVS approach proceeds by assigning the prior distribution of the regression coefficients ($\boldsymbol{\beta}$) to be a mixture distribution - [41] proposed a mixture of a point mass at 0 and a uniform distribution, whereas [17] propose a mixture of two normal distributions centered at 0 but with distinct variances. The estimated posterior distribution of the probability of being included in the model can be used for variable selection. Several papers have applied this approach for identifying important features in high-dimensional microarray data for various settings such as binary outcomes ([30]), multi-category responses ([45]), and censored outcomes ([44]). In a recent study, [22] demonstrated the use of the BVS method in large-scale settings such as genome-wide association studies (GWAS). In addition to variable selection, the BVS approach has also been shown to be able to simultaneously cluster variables to identify group structures ([49, 28, 11]). One advantage of the BVS approach when compared to other variable selection methods is that it can be naturally extended to incorporate external information such as biological pathway membership ([32, 48]).

In this chapter, we extend the lasso and BVS algorithms to accommodate error-prone, self-reported outcomes. Through simulation studies, we compare our proposed algorithms to naive approaches that ignore the error in self-reported outcomes. We apply our proposed approaches to GWAS data on 12,008 African American and Hispanic women enrolled in the WHI to discover biomarkers of self-reported incident diabetes. This chapter is organized as follows: In Section 4.2, we present notation and the form of the likelihood function that accommodates error in self-reported out-

comes. We incorporate the lasso and BVS algorithms into this loglikelihood, to handle high-dimensional datasets. In Section 4.3, we perform simulation studies to compare the variable selection performance of different approaches. In Section 4.4, we apply our proposed methods to the GWAS data from the WHI. Lastly, in Section 4.5 we discuss the findings of this study and highlight future directions.

4.2 Method

In this section, we introduce notation and present the form of the likelihood function to accommodate error-prone, self-reported outcomes. We propose two variable selection methods by adapting the Lasso and the Bayesian variable selection strategies.

4.2.1 Notation, likelihood function

Consider a dataset of N subjects each with P observed features. Let x_{ip} denote the p^{th} covariate for subject i. We assume that the visits for each subject are prescheduled, and at each visit the diagnostic test result for occurrence of event is either positive or negative. We treat self-report as a type of diagnostic test. Let T refer to the random variable denoting the unobserved time to event for an individual, with associated survival, density and hazard functions denoted by $S(t), f(t)$ and $\lambda(t)$, for $t \geq 0$ respectively. Without loss of generality, we set $T = \infty$ when the event of interest does not occur. Let $\tau_1, \cdots, \tau_J$ denote the distinct, ordered visit times in the dataset, where $0 = \tau_0 < \tau_1 < ... < \tau_J < \tau_{J+1} = \infty$ - thus, the time axis can be divided into $J + 1$ disjoint intervals, $[\tau_0, \tau_1), [\tau_1, \tau_2), \cdots, [\tau_J, \infty)$. Let S_j denote the survival function at time τ_{j-1} for the reference group (i.e. $\boldsymbol{X} = 0$), where $j = 1, 2, \cdots, J + 1$. Therefore $S_1 = 1$ and $1-S_{J+1}$ refers to the cumulative incidence in the reference group at the end of follow up. We assume the sensitivity and specificity of the diagnostic test are known and constant and are denoted by φ_1 and φ_0. Assuming a proportional

hazards model, $\lambda(t|\boldsymbol{X} = \boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{x}'\boldsymbol{\beta}}$, it can be shown that the log-likelihood function can be written as:

$$l\left(\boldsymbol{S}, \boldsymbol{\beta}\right) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{J+1} D_{ij}\left(S_j\right)^{\exp\left(\sum_{p=1}^{P}\beta_p x_{ip}\right)}\right). \tag{4.1}$$

where $\boldsymbol{\beta}$ are regression coefficients of the covariates. The elements of the D matrix are functions of the observed data including the visit times and corresponding self-reported results, as well as the constants φ_0, φ_1 denoting the specificity and sensitivity of the diagnostic test (self-report), respectively. We define sensitivity and specificity as $\varphi_1 = p(\text{Positive test result} \mid \text{Disease})$ and $\varphi_0 = p(\text{Negative test result} \mid \text{Disease free})$. $\beta_1, \cdots, \beta_P, S_2, \cdots, S_{J+1}$ denote the unknown parameters of interest and can be estimated by numerical maximization of the log-likelihood function, subject to the constraints that $1 > S_2 > S_3 > \cdots > S_{J+1} > 0$.

4.2.2 Lasso

In this section, we adapt the Lasso to handle error-prone self-reported outcomes. For computational considerations, we reparameterize the log-likelihood function by setting $\alpha_j = \log(-\log(S_{j+1}))$. The objective function including penalty term can be expressed as:

$$\begin{aligned} f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \;& \tfrac{2}{n} l(\boldsymbol{S}, \boldsymbol{\beta}) - \lambda \textstyle\sum_{p=1}^{P} |\beta_p| \\ = \;& \tfrac{2}{n} \textstyle\sum_{i=1}^{n} \log\left(D_{i1} + \sum_{j=1}^{J} D_{i,j+1} \exp\left(-\exp\left(\alpha_j + \sum_{p=1}^{P} X_{ip}\beta_p\right)\right)\right) \\ & -\lambda \textstyle\sum_{p=1}^{P} |\beta_p|, \end{aligned} \tag{4.2}$$

where λ is the tuning parameter controlling the magnitude of shrinkage. When λ is large, all coefficients $\boldsymbol{\beta}$ approach 0.

The Lasso objective function can be optimized using the pathwise coordinate descent algorithm. Our algorithm is based on the procedure developed by [46] for the

penalized Cox proportional hazards model. First, assume that $\boldsymbol{\alpha}$ is fixed and use a Taylor series expansion to approximate the log-likelihood function around some constant $\tilde{\boldsymbol{\beta}}$. $\tilde{\boldsymbol{\beta}}$ can be set to equal a set of starting values or the values of $\boldsymbol{\beta}$ from a previous iteration. Then the log-likelihood function in Equation (4.1) can be approximated as:

$$l(\boldsymbol{\beta}) = l(\boldsymbol{\beta} \mid \boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}) = \frac{1}{2} \left(z\left(\tilde{\boldsymbol{\eta}}\right) - X\boldsymbol{\beta}\right)^T l^{''}\left(\tilde{\boldsymbol{\eta}}\right) \left(z\left(\tilde{\boldsymbol{\eta}}\right) - X\boldsymbol{\beta}\right) + C$$

where C is independent of $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\eta}} = X\tilde{\boldsymbol{\beta}}$, $l^{''}(.)$ is the Hessian of the log-likelihood function with respect to $\boldsymbol{\eta} = X\boldsymbol{\beta}$, and $z(\tilde{\boldsymbol{\eta}}) = \tilde{\boldsymbol{\eta}} - l^{''}(\tilde{\boldsymbol{\eta}})^{-1} l^{'}(\tilde{\boldsymbol{\eta}})$. , For ease of computation, $l^{''}(\tilde{\boldsymbol{\eta}})$ can be replaced with a diagonal matrix - this follows since the off diagonal elements of the Hessian matrix can be shown to be of much smaller magnitude when compared to the diagonal elements ([46]). Let $w(\tilde{\boldsymbol{\eta}})_i = -l^{''}(\tilde{\boldsymbol{\eta}})_{ii}$. Then with fixed $\boldsymbol{\alpha}$, maximizing the objective function (4.2) is equivalent to minimizing the following function,

$$M(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} w\left(\tilde{\boldsymbol{\eta}}\right)_i \left(z\left(\tilde{\boldsymbol{\eta}}\right)_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 + \lambda \sum_{p=1}^{P} |\beta_p|, \tag{4.3}$$

which can be solved using pathwise coordinate descent algorithm.

The Lasso algorithm using pathwise coordinate descent:

1. Standardize the design matrix X.

2. Set initial values of the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ to equal $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$, respectively.

3. Optimize the log-likelihood function $l(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha}$ while fixing $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$. The optimization can be achieved using the Newton Raphson algorithm. Update $\tilde{\boldsymbol{\alpha}}$ to equal the optimized value of $\boldsymbol{\alpha}$.

4. Compute the values of $\tilde{\boldsymbol{\eta}} = X\tilde{\boldsymbol{\beta}}$, $z(\tilde{\boldsymbol{\eta}})$ and $w(\tilde{\boldsymbol{\eta}})$ by fixing $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$ as follows:

$$y_{ij} = \exp\left(-\exp\left(\tilde{\alpha}_j + \tilde{\eta}_i\right)\right)$$

$$l_i = D_{i1} + \sum_{j=1}^{J} D_{i,j+1} y_{ij}$$

$$l1_i = \frac{\sum_{j=1}^{J} D_{i,j+1} y_{ij} \log(y_{ij})}{l_i}$$

$$l2_i = \frac{\sum_{j=1}^{J} D_{i,j+1} y_{ij} \log(y_{ij})(1 + \log(y_{ij}))}{l_i} - l1_i^2$$

$z(\tilde{\boldsymbol{\eta}})$ and $w(\tilde{\boldsymbol{\eta}})$ are both $N \times 1$ vectors, with elements obtained as:

$$w(\tilde{\boldsymbol{\eta}})_i = -l2_i$$

$$z(\tilde{\boldsymbol{\eta}})_i = \tilde{\eta}_i - \frac{l1_i}{l2_i}$$

5. Iterate and update each element of $\boldsymbol{\beta}$ while fixing other elements. The kth element of $\boldsymbol{\beta}$ (i.e. β_k) is updated as:

$$\hat{\beta}_k = \frac{S\left(\frac{1}{n}\sum_{i=1}^{n} w\left(\tilde{\boldsymbol{\eta}}\right)_i x_{i,k}\left(z\left(\tilde{\boldsymbol{\eta}}\right)_i - \sum_{j \neq k} x_{ij}\beta_j\right), \lambda\right)}{\frac{1}{n}\sum_{i=1}^{n} w\left(\tilde{\boldsymbol{\eta}}\right)_i x_{ik}^2}$$

where $S(x, \lambda) = \text{sign}(x)(|x| - \lambda)_+$.

6. Repeat step (5) until convergence of the objective function in equation (4.3).

7. Set $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

8. Repeat steps (3)-(7) until convergence of the objective function in equation (4.2).

An important aspect with implementing the Lasso algorithm is the selection of the shrinkage parameter λ. In linear regression models, the optimal value for λ is usually chosen using a cross validation approach, in which the optimal value of λ is that value which minimizes the cross-validated sum of squared error. This approach can also be applied to our setting where the cross-validated sum of squared error is replaced by

the cross-validated log-likelihood as the criterion for selection of the optimal value of λ.

k-fold cross validation approach for selecting the optimal value of λ:

1. Define a grid of values for λ: $\lambda_1, \cdots, \lambda_m$, from which the optimal value of λ will be chosen.

2. Split data equally into k groups.

3. Compute the cross validated log-likelihood l_{ij} as follows: Set $\lambda = \lambda_i$ and fit the Lasso model to estimate the parameters $\boldsymbol{\beta}$ on the subset of the data by excluding the j^{th} partition. Use the fitted parameters to compute the log-likelihood function on the j^{th} partition of the dataset using the fitted values of the parameters $\boldsymbol{\beta}$.

4. Compute the cross validated log-likelihood for λ_i by summing over the j partitions, $CV_i = \sum_j l_{ij}$.

5. Select the value of λ that maximizes the cross validated log-likelihood function.

When the value of λ is large, all Lasso solutions for βs approach 0. To determine the candidate list of λ values, we first obtain the maximum value of λ that gives at least one non-zero solution. As described in [46], we first set $\boldsymbol{\beta} = \mathbf{0}$ and obtain $\hat{\boldsymbol{\alpha}}$ as the value that optimizes the log-likelihood function. We obtain values of $w(\mathbf{0})$ and $z(\mathbf{0})$. The maximum value of λ can then be obtained as $\lambda_{max} = \max_j \frac{1}{n} \sum_{i=1}^{n} w_i(\mathbf{0}) z_i(\mathbf{0})$.

4.2.3 Bayesian Variable Selection

In this section, we adapt the Bayesian variable selection approach to incorporate error-prone self reported outcomes. We introduce a latent vector $\boldsymbol{\gamma} = (\gamma_i, 1 \leq i \leq P)$, where each γ_i is an indicator variable denoting whether the ith covariate is included or not in the model. The Bayesian variable selection analysis proceeds via MCMC

methods to estimate the posterior distribution $\boldsymbol{\gamma}$. With this latent variable formulation for variable selection, the log likelihood function in Equation (4.1) is a function of the parameters $S_2, \cdots, S_{J+1}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and is denoted $l(\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. We assume the following hierarchical structure of the prior distributions corresponding to the unknown parameters in the model:

$$\begin{aligned} \beta_p \mid \gamma_p &\sim \gamma_p N(0, b^2) + (1 - \gamma_p)\delta_0 \\ \gamma_p \mid \omega &\sim \text{Bernoulli}(\omega) \\ \omega &\sim \text{Beta}(w_1, w_2) \end{aligned}$$

where δ_0 is delta function corresponding to a point mass at 0 and b, w_1, w_2 are treated as known hyper-parameters.

By treating the survival parameters $\boldsymbol{S}$ as nuisance parameters, we propose the following Metropolis-Hasting algorithm:

1. Initialization: For fixed values of w_1, w_2, set $\omega^{(0)}$ to equal a randomly generated value from Beta(w_1, w_2) distribution. Set $\boldsymbol{\gamma}^{(0)} = \mathbf{0}$ and $\boldsymbol{\beta}^{(0)} = \mathbf{0}$. Optimize the log-likelihood function in Equation (4.1) with respect to $\boldsymbol{S}$ by fixing $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$, and then set $\boldsymbol{S}^{(0)}$ to equal the optimized value for $\boldsymbol{S}$.

2. At iteration t, we let the indices $t-1$ and $*$ denote the current and proposed values of the parameters, respectively.

3. Update main effects: Select a covariate $p \in (1 \cdots P)$ at random and let the proposed value $\gamma_p^* = 1 - \gamma_p^{t-1}$. If $\gamma_p^* = 0$, the corresponding regression coefficient β_p^* is set to 0. If $\gamma_p^* = 1$, the proposed regression coefficient β_p^* is sampled from $N(0, b^2)$ distribution. Compute:

$$\Delta = l\left(\boldsymbol{S}^{(t-1)}, \boldsymbol{\beta}^*\right) - l\left(\boldsymbol{S}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}\right) + \left(2\gamma_p^* - 1\right) \log\left(\frac{\omega^{(t-1)}}{1 - \omega^{(t-1)}}\right)$$

Generate a uniformly distributed random sample U. If $\log(U) \leq \Delta$ then accept the proposed update for β_p and γ_p.

4. Update regression coefficients: For each included main effect, update the coefficient with a user defined probability p_{main}, for example 30%. If a main effect β_p is chosen for update, let the proposed value, β_p^*, be a random sample from the distribution $N(\beta_p^{(t-1)}, b^2)$. Compute:

$$\Delta = l\left(\boldsymbol{S}^{(t-1)}, \boldsymbol{\beta}^*\right) - l\left(\boldsymbol{S}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}\right) + \frac{\left(\beta_p^{(t-1)}\right)^2 - \left(\beta_p^*\right)^2}{2b^2}$$

Generate a uniformly distributed random sample U. If $\log(U) \leq \Delta$ then accept the proposed update for β_p.

5. Update $\boldsymbol{S}$: Optimize the log-likelihood function with respect to $\boldsymbol{S}$ by fixing $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$. Set $\boldsymbol{S}^{(t)}$ to equal the optimized value of $\boldsymbol{S}$.

6. Update ω: Using Gibbs sampling, update ω by generating a sample from Beta($w_1 + K_\gamma, w_2 + P - K_\gamma$), where K_γ is the number of main effects selected.

After the burn-in period, the features that have highest inclusion probabilities based on the posterior distribution of γ_p, are selected.

4.3 Simulation studies

In this section, we report results from simulation studies to evaluate the performance of the proposed adaptions of the Lasso and Bayesian variable selection algorithms in the presence of error-prone self-reported outcomes, under various parameter settings. To closely mimic real data settings, the design matrix for the simulation study was selected as a randomly selected subset of the existing GWAS data from the WHI Clinical Trial and Observational Study SHARe that includes extensive genotypic (> 900K SNPs) and phenotypic data 12,008 African American and Hispanic

women. We randomly selected 1000 SNPs on 300 subjects from the existing WHI dataset. All missing genotypes were imputed to be genotype "AA". Corresponding to each SNP, we created binary variables by coding genotype "AA" as 0 (reference), and genotype "Aa" and "aa" as 1.

To simulate results from error-prone self-reports (diagnostic tests) for each subject, we assumed that there are 8 pre-scheduled test times with no missed visits. The distribution of event times for the reference group (i.e. $\boldsymbol{X} = \mathbf{0}$) was assumed to be exponential with baseline hazard rate denoted by λ_0. λ_0 was determined by the corresponding cumulative incidence rate (CIR) during the study, which was varied between 5%, 10%, 25% and 50%. Event times for each subject in the study were simulated from an exponential distribution, where the hazard function (λ) was determined through the proportional hazards model ($\lambda = \lambda_0 e^{\boldsymbol{\beta X}}$). A total of 5 SNPs were randomly sampled from the 1000 SNPs as true biomarkers, each with a coefficient $\beta = 0.7$ in the proportional hazards model, corresponding to a hazard ratio of $e^{\beta} = 2$. The regression coefficients for the remaining 995 SNPs in the proportional hazards model were set to 0. To simulate the results from the error-prone self-reports or diagnostic tests for each subject, we first simulated the true event time from an exponential distribution with hazard rate $\lambda = \lambda_0 e^{\beta_1 SNP_1 + \cdots + \beta_5 SNP_5}$, where $SNP_1, \cdots, SNP_5$ are the selected 5 SNPs with regression coefficients $\beta = 0.7$. The true disease status at each visit was obtained by comparing the true event time and visit time. Lastly, for each subject, the results from the error-prone diagnostic test result at each visit can be simulated from an appropriate Bernoulli distribution determined by the true disease status at that time and the sensitivity (φ_1), specificity (φ_0) parameters governing the behavior of the diagnostic test (self-report). The values of (sensitivity, specificity) were varied between $[(1,1),(0.9,1),(1,0.9),(0.55,0.99),(0.75,0.98)]$. The sensitivity and specificity values $(0.55, 0.99)$ correspond to the properties of diabetes self-repo rts in the WHI ([34]).

We consider two study design settings: (1) Diagnostic (self-reported) test results are collected at all predetermined times/visits (denoted ‘ No missed visits” in Table 4.1); (2) Results following the first positive test are discarded or considered to be non-informative (denoted “NTFP” in Table 4.1). The latter scenario commonly applies to self-reported outcomes.

For each parameter setting, we compared the variable selection performance of the following three strategies: (1) Lasso algorithm implemented for the Cox proportional hazards model (*glmnet* R package); (2) proposed Lasso algorithm; and (3) proposed Bayesian variable selection algorithm.

(1) Lasso for the Cox proportional hazards model: To apply the algorithm implemented in the *glmnet* R package, the event time was defined as the time of the first positive test result (observed event) or the time of last observation (censored observation). We note that the two study design settings (“No missed visits”, “NTFP”) will yield identical results under the Cox proportional hazards Lasso model. 10-fold cross validation was used to select the optimal tuning parameter λ. Features were ranked based on the absolute values of the estimated regression coefficients $\boldsymbol{\beta}$.

(2) Proposed Lasso algorithm: 10-fold cross validation was used to obtain the optimal tuning parameter, λ.

(3) Proposed Bayesian variable selection algorithm: The MCMC algorithm was run up to 100,000 iterations and the first 20,000 iterations were discarded as the burn-in period. We assumed the following parameter values corresponding to the hyper-parameters determining the prior distributions: $b = 0.7, w_1 = 5, w_2 = 1000$. The 1000 SNPs were ranked based on posterior inclusion probabilities determined by the posterior distribution of $\boldsymbol{\gamma}$.

For each simulated dataset, the performance of each of the three approaches with regard to identifying the true five biomarkers was quantified through the area under curve (AUC) metric. For each simulated dataset and analysis approach, the ROC

curve was generated by varying the threshold with regard to the magnitude of the estimated regression coefficients in (methods (1) and (2)) or the magnitude of the posterior inclusion probabilities (method (3)). The AUC for each parameter setting and approach was obtained as the average of AUCs obtained over 500 simulated datasets. The confidence interval for the AUC corresponding to each approach was obtained by $\pm 1.96\sigma$, where σ corresponds to the estimated standard error of the AUC estimates across the 500 simulations.

Table 4.1 presents the simulation results. When the study design is NTFP, the AUC of proposed Lasso and Bayesian variable selection algorithms are not significantly different from that of Cox Lasso which ignores the error in self-reported outcomes. On the other hand, when there are no missed visits, the AUC estimates corresponding to the proposed Lasso and Bayesian variable selection algorithms are significantly larger than that of Cox Lasso, especially in settings where the specificity is less than perfect and the cumulative incidence is low. In all settings considered, the results from the proposed Lasso were comparable to that from the proposed Bayesian variable selection algorithm.

For all parameter settings, the AUC increases with increased incidence rate - this result is expected as a higher incidence rate results in higher statistical power to distinguish true biomarkers from non-markers, regardless of the approach. In the parameter settings considered, the AUC was more sensitive to the change in specificity. This effect is especially dramatic when cumulative incidence rate is low because the false positive events dominate all observed positive events. On the other hand, for all approaches, the AUC was not affected by changes in sensitivity - the AUCs when $\varphi_1 = 1.0, \varphi_0 = 1.0$ was always close to that when $\varphi_1 = 0.9, \varphi_0 = 1.0$. As expected, for both the proposed Lasso and Bayesian variable selection algorithms, the AUC when there are no missed visits was generally higher than that under the NTFP study design - this was pronounced when specificity was less than perfect and

cumulative incidence was low. When specificity is perfect, there was no difference in AUC when comparing the study design NTFP to the setting of no missed visits - this is expected as when specificity is perfect, the first positive event indicates the occurrence of event with probabilty 1 and hence further diagnostic tests are non-informative.

4.4 Application: Genetic biomarkers of incident diabetes mellitus in the Women's Health Initiative

Background: The proposed methods were applied to data from the WHI Clinical Trial and Observational Study SHARe, to identify biomarkers of incident diabetes mellitus. The dataset includes extensive genotypic (909, 622 SNPs) and phenotypic information on 12,008 African American and Hispanic women. The biomarkers identified in this analysis were compared to the set of 40 genes (and associated SNPs) in the review by [35] that have been previously identified in candidate gene studies and genome-wide investigations of Type 2 Diabetes. The WHI SHARe dataset provides a unique opportunity for the discovery of genetic biomarkers of diabetes in a population of African American and Hispanic women, who are at higher risk for developing diabetes when compared to Caucasians (http://www.cdc.gov/diabetes/consumer/groups.htm).

Diabetes self-reports: Prevalent diabetes at baseline and incident diabetes were assessed through self reported questionnaires in the WHI. At baseline and at each annual visit, participants were asked whether they had ever received a physician diagnosis of and/or treatment for diabetes when not pregnant since the time of the last self-report/visit. Using data from a WHI substudy ([34]), estimates of sensitivity, specificity, and baseline negative predictive value of self reported diabetes outcomes were obtained by comparing self reported outcomes to fasting glucose levels and medication data. A woman was considered to be truly diabetic if she had either taken anti-diabetic medication and/or had a fasting glucose level $\geq$ 126mg/dL. By

using a subset of 5485 women, with information at baseline on diabetes self reports, fasting glucose levels and medication inventory, we estimated that self reports have a sensitivity of 0.61, the specificity of 0.995, and a negative predictive value of 0.96 at baseline. These estimated parameter values are used in our analysis.

Methods: Subjects who reported diabetes at baseline were excluded from the analysis, resulting in a dataset of 8,293 women. The results presented here are based on follow up until 2010. The average follow up from baseline was 11 years, with a maximum follow up of 16 years. During the period of follow up, 10.1% of the women self-reported a new diagnosis of diabetes.

To reduce the dimension of the dataset of 909,622 SNPs, the following two step procedure was followed: First, SNPs that satisfied one or more of the following three criteria were excluded from the analysis: (i) more than 1% missing values; or (ii) less than 5% minor allele frequency; or (iii) a Hardy-Weinberg equilibrium test p-value less than 0.05. Second, the univariate association of each SNP with incident self-reported diabetes was estimated using a two sided likelihood ratio test from fitting the model described in (4.1). SNPs with a univariate p value greater than 0.4 were excluded from the analysis. This two-step filtering procedure resulted in a dataset of 133,781 SNPs available for analysis. All missing genotypes were imputed to be genotype "AA" (wildtype). SNPs were incorporated into the analysis by creating binary variables where genotype "AA" was coded as 0 (reference), and genotypes "Aa" or "aa" as 1.

The dataset of 133,781 SNPs on 8,293 subjects was analyzed using three different methods:

(1) Univariate Cox proportional hazards (PH) model was fit to each SNP to evaluate the association with incident self-reported diabetes. Here, the time to event variable was calculated as the interval between enrollment date and the earliest of the following: (i) date of annual medical history update when new diabetes is self-

reported (positive outcome); (ii) date of last annual medical update during which diabetes status can be ascertained (censorship); or (iii) date of death (censorship). SNPs were ranked based on the maximized value of the log-likelihood.

(2) Proposed Lasso algorithm: We found for this particular dataset, the algorithm does not converge when the value of λ is small. Thus, we use the smallest possible value of λ based on computational feasibility. SNPs were ranked based on the absolute value of the estimated regression coefficients $\boldsymbol{\beta}$.

(3) Proposed Bayesian variable selection algorithm: The values of the hyper-parameters governing the prior distributions were assumed to be $b = 0.5, w_1 = 5, w_2 = 50000$. The MCMC algorithm was run for 5,000,000 iterations and the first 500,000 iterations were discarded as the burn-in period.

Results: The final model from the proposed Lasso algorithm resulted in 36 SNPs with nonzero coefficients (Table 4.2). The Bayesian variable selection algorithm resulted in 174 SNPs with non-zero posterior inclusion probability (Table 4.3). To compare the results from each of the three approaches, we present the top 10 SNPs identified by each method in Table 4.4. In Tables 4.2 - 4.4, each SNP is annotated with associated genes (intron, left and right) using a publicly available database (http://www.scandb.org). A total of 24 SNPs were identified in the top 10 list by at least one analysis approach. Two SNPs (rs2575507, rs17028352) were ranked in the top 10 by all three approaches, suggesting that these two SNPs are very likely associated with risk of diabetes. One of the genes associated with SNP rs2575507 (ATP8A1) has previously been implicated with risk of type 2 diabetes (http://www.disgenet.org). However, none of these identified genes is among the list of the genes associated with type 2 diabetes that is presented in [35]

We observed that the results from the Lasso and the univariate Cox PH model aligned well with each other and resulted in a similar ranking. However, the ranks

from the Bayesian variable selection approach were generally different from other methods.

4.5 Discussion

The use of of error-prone diagnostic tests or self-report is common in large scale epidemiology studies. While likelihood based statistical approaches exist to evaluate the association of a small, targeted set of covariates with an error-prone outcome, these methods are not appropriate for the analysis of high dimensional datasets. In this chapter, we extend the likelihood based approach for error-prone self-reported outcomes to handle high dimensional datasets by adapting the Lasso and the Bayesian variable selection algorithms for this setting. We proposed a pathwise coordinate descent algorithm to solve the extension of the Lasso approach for error-prone outcomes. For the extension of the Bayesian variable selection method, we proposed a Metropolis-Hasting algorithm to stochastically search for important variables associated with the error-prone outcome.

We performed simulation studies to compare variable selection performance of the proposed approaches under various parameter settings corresponding to the sensitivity and specificity of the error-prone diagnostic tests. Our simulation results suggest that the Bayesian variable selection algorithm generally has better variable selection performance when compared to the Lasso based approaches. While the naive Cox Lasso method ignores measurement error, its variable selection performance is not significantly different from our proposed Lasso method when the study design specifies that no tests are administered following the first positive result (NTFP). However, in the absence of missed visits, our proposed Lasso and Bayesian variable selection algorithms outperform the naive Cox Lasso, in settings where the cumulative incidence rates are modest and specificity is imperfect.

The proposed methods were applied GWAS data from the WHI Clinical Trial and Observational Study SHARe to identify biomarkers of incident self-reported diabetes mellitus. Two SNPs rs2575507 and rs17028352 were consistently ranked among the top 10 SNPs by both proposed approaches, suggesting a true association with incident diabetes.

In the methods developed in this chapter, we assumed that the variables are independent and that their effects are additive. In real world settings, this assumption is likely to be an over simplification. For example, it is known that there exists complex network relationships between different SNPs based on genetic inter-relationships and membership in co-acting biological pathways. In this context, the proposed Bayesian variable selection approach can be extended to incorporate external biological information into the prior distributions for $\boldsymbol{\gamma}$.

Table 4.1. Comparing variable selection performance of different methods as quantified by the area under curve (AUC), for varying sensitivity φ_1, φ_0 and cumulative incidence rate (CIR). NMISS denotes the setting in which no visits are missed. NTFP denotes the study design in which no further testing is carried out following the first positive test result.

			Proposed Lasso		Cox Lasso	Proposed BVS	
φ_1	φ_0	CIR	NMISS	NTFP	Lasso	NMISS	NTFP
1.00	1.00	0.05	0.78(±.014)	0.77(±.014)	0.76(±.014)	0.81(±.011)	0.83(±.010)
0.90	1.00	0.05	0.76(±.014)	0.77(±.014)	0.75(±.014)	0.81(±.011)	0.82(±.010)
1.00	0.90	0.05	0.73(±.014)	0.51(±.009)	0.52(±.004)	0.79(±.011)	0.54(±.010)
0.55	0.99	0.05	0.72(±.015)	0.66(±.016)	0.63(±.012)	0.78(±.011)	0.73(±.012)
0.75	0.98	0.05	0.74(±.015)	0.63(±.018)	0.61(±.011)	0.79(±.011)	0.70(±.012)
1.00	1.00	0.10	0.91(±.011)	0.90(±.011)	0.90(±.011)	0.91(±.009)	0.91(±.008)
0.90	1.00	0.10	0.91(±.010)	0.92(±.009)	0.89(±.011)	0.91(±.008)	0.92(±.008)
1.00	0.90	0.10	0.90(±.010)	0.55(±.016)	0.57(±.009)	0.90(±.009)	0.65(±.012)
0.55	0.99	0.10	0.85(±.015)	0.82(±.013)	0.80(±.014)	0.88(±.009)	0.85(±.010)
0.75	0.98	0.10	0.89(±.011)	0.78(±.015)	0.77(±.014)	0.89(±.009)	0.84(±.010)
1.00	1.00	0.25	0.99(±.004)	0.98(±.004)	0.98(±.005)	0.98(±.004)	0.98(±.004)
0.90	1.00	0.25	0.98(±.004)	0.98(±.005)	0.98(±.005)	0.98(±.004)	0.98(±.004)
1.00	0.90	0.25	0.98(±.004)	0.77(±.020)	0.77(±.014)	0.98(±.004)	0.84(±.010)
0.55	0.99	0.25	0.95(±.007)	0.95(±.007)	0.94(±.008)	0.96(±.006)	0.95(±.007)
0.75	0.98	0.25	0.97(±.005)	0.95(±.007)	0.93(±.009)	0.97(±.005)	0.95(±.006)
1.00	1.00	0.50	0.99(±.003)	0.99(±.003)	0.99(±.003)	0.99(±.003)	0.99(±.002)
0.90	1.00	0.50	0.99(±.003)	0.99(±.003)	0.99(±.003)	0.99(±.003)	0.99(±.003)
1.00	0.90	0.50	0.99(±.003)	0.91(±.011)	0.91(±.010)	0.99(±.002)	0.93(±.008)
0.55	0.99	0.50	0.95(±.007)	0.94(±.008)	0.93(±.008)	0.95(±.006)	0.95(±.007)
0.75	0.98	0.50	0.98(±.004)	0.97(±.005)	0.96(±.006)	0.98(±.004)	0.97(±.005)

Table 4.2. Biomarkers of incident diabetes in the WHI based on the proposed Lasso algorithm (top 30 SNPs presented).

SNP	Intron gene	Left gene	Right gene	Rank
rs2575507	.	ATP8A1	GRXCR1	1
rs2191331	MAGI2	MGC34774	LOC100124402	2
rs11655073	ACCN1	LOC100129255	TLK2P1	3
rs4945056	.	WNT11	PRKRIR	4
rs11628414	KCNK13	TDP1	GLRXP2	5
rs565503	PLA2G5	PLA2G2A	PLA2G2D	6
rs17028352	.	TMEM182	LOC728815	7
rs8097803	.	hCG_1776047	TXNL1	8
rs11988314	KCNK9	COL22A1	TRAPPC9	9
rs8082986	SETBP1	KRT8P5	LOC100131669	10
rs13419210	.	LOC100130841	KLHL29	11
rs2505140	.	LOC646348	ANKRD30A	12
rs500090	FLI1	ETS1	KCNJ1	13
rs1546031	.	KCNH7	FIGN	14
rs756930	DYNC1I1	PDK4	SLC25A13	15
rs17239028	SLC12A8	HEG1	ZNF148	16
rs6944339	MAGI2	MGC34774	LOC100124402	17
rs2698723	SNX10	LOC442659	LOC100129036	18
rs9925238	.	LOC100131080	TMEM114	19
rs16889988	ZMAT4	C8orf4	SFRP1	20
rs136622	.	TBC1D22A	RP11-191L9.1	21
rs6930750	C6orf195	LOC100128372	MYLK4	22
rs4677987	SEMA5B	LOC100129550	PDIA5	23
rs638234	ALKBH8	CWF19L2	LOC100132695	24
rs10830770	.	CTSC	GAPDHL15	25
rs1200160	NME7	ATP1B1	BLZF1	26
rs41485749	.	.	.	27
rs9814339	.	LOC644662	KLF15	28
rs9848926	PPP2R3A	EPHB1	MSL2L1	29
rs2130806	CUBN	RSU1	TRDMT1	30

Table 4.3. Biomarkers of incident diabetes in the WHI based on the proposed Bayesian variable selection (top 30 SNPs presented).

SNP	Intron gene	Left gene	Right gene	Rank
rs2575507	.	ATP8A1	GRXCR1	1
rs644818	FLI1	ETS1	KCNJ1	2
rs10431977	.	LOC100128497	MON1B	3
rs4074769	.	LOC100131080	TMEM114	4
rs11857642	LCTL	ZWILCH	SMAD6	5
rs263173	.	GPR126	HIVEP2	6
rs17028352	.	TMEM182	LOC728815	7
rs4506998	.	LOC388474	KC6	8
rs989975	.	CYCSP14	LOC100132483	9
rs6658894	EPHB2	C1QB	LOC646262	10
rs4395106	.	CBX4	TBC1D16	11
rs2698723	SNX10	LOC442659	LOC100129036	12
rs11589612	SLC35F3	KCNK1	LOC100130965	13
rs16889988	ZMAT4	C8orf4	SFRP1	14
rs2741757	OR10A4	OR10A2	OR10A4	15
rs11192261	SORCS3	CCDC147	YWHAZP5	16
rs12306145	.	LOC100131418	SOX5	17
rs13117180	.	LOC729902	NPY2R	18
rs2226798	.	C21orf34	C21orf37	19
rs11867749	.	NUFIP2	TAOK1	20
rs756930	DYNC1I1	PDK4	SLC25A13	21
rs11668998	C19orf48	ACPT	C19orf48	22
rs7485690	.	LOC100129937	BCAT1	23
rs2683173	.	LOC728073	RPL38	24
rs10494795	.	LOC647202	NR5A2	25
rs11209403	.	AF357533	LOC100133218	26
rs6967983	MAGI2	MGC34774	LOC100124402	27
rs638234	ALKBH8	CWF19L2	LOC100132695	28
rs6953785	.	LOC100129606	LOC401316	29
rs446695	KIAA1576	NUDT7	CLEC3A	30

Table 4.4. Biomarkers of incident diabetes in the WHI based on: (1) univariate Cox proportional hazards model (2) proposed Lasso and (3) proposed Bayesian variable selection algorithm. SNPs ranking among the top 10 by at least one method are presented.

				Rank		
SNP	Intron gene	Left gene	Right gene	Univariate	Lasso	BVS
rs2575507	.	ATP8A1	GRXCR1	8	1	1
rs644818	FLI1	ETS1	KCNJ1	.	.	2
rs10431977	.	LOC100128497	MON1B	.	.	3
rs4074769	.	LOC100131080	TMEM114	.	.	4
rs11857642	LCTL	ZWILCH	SMAD6	.	.	5
rs263173	.	GPR126	HIVEP2	.	.	6
rs17028352	.	TMEM182	LOC728815	1	7	7
rs4506998	.	LOC388474	KC6	.	.	8
rs989975	.	CYCSP14	LOC100132483	.	.	9
rs6658894	EPHB2	C1QB	LOC646262	.	.	10
rs2698723	SNX10	LOC442659	LOC100129036	7	18	12
rs756930	DYNC1I1	PDK4	SLC25A13	6	15	21
rs2191331	MAGI2	MGC34774	LOC100124402	4	2	51
rs1546031	.	KCNH7	FIGN	2	14	83
rs9918753	HMBOX1	INTS9	KIF13B	9	32	107
rs8097803	.	hCG_1776047	TXNL1	3	8	113
rs11628414	KCNK13	TDP1	GLRXP2	13	5	153
rs8082986	SETBP1	KRT8P5	LOC100131669	21	10	278
rs11655073	ACCN1	LOC100129255	TLK2P1	18	3	.
rs4945056	.	WNT11	PRKRIR	45	4	.
rs565503	PLA2G5	PLA2G2A	PLA2G2D	32	6	.
rs11988314	KCNK9	COL22A1	TRAPPC9	40	9	.
rs9925238	.	LOC100131080	TMEM114	10	19	.
rs4389218	.	LOC388458	PPIAP14	5	.	.

APPENDIX

MISCLASSIFICATION AT STUDY ENTRY

In this appendix, we consider the setting in which the diagnostic procedure used at baseline or study entry is subject to imperfect sensitivity and specificity and see how it affects the sample size calculation. For example, in the WHI, women's diabetes status at baseline is also assessed through self reports. However, the study by [34] found that the negative predictive value of prevalent diabetes at baseline was approximately 97% - in other words, 3% of women who self-reported as being diabetes free were in fact diabetic. As discussed in Chapter 2, the incorporation of baseline misclassification can be achieved by simply modifying the D matrix using the baseline negative predictive value.

Figure A.1 illustrates the effect of misclassification at baseline with respect to sample size calculations. We consider parameters that mimic the properties of diabetes self-reports in the WHI. In particular, we assume a study duration of 8 years with annual visits, with an event rate of 10% during the course of the study (i.e. $S_{J+1} = 0.9$), assume no missing tests and that the sensitivity and specificity are equal to 0.61 and 0.995, respectively. Total sample sizes are calculated assuming that HR=2, the desired power is 0.9 and the Type I error is fixed at 0.05. As expected, the required sample size increases with decreasing negative predictive value (η). In the context of the WHI, using self-reports to ascertain diabetes status at baseline results in a sample size of 954(or a 20% increase) when compared to a sample size of 792 when using a perfect test at baseline.

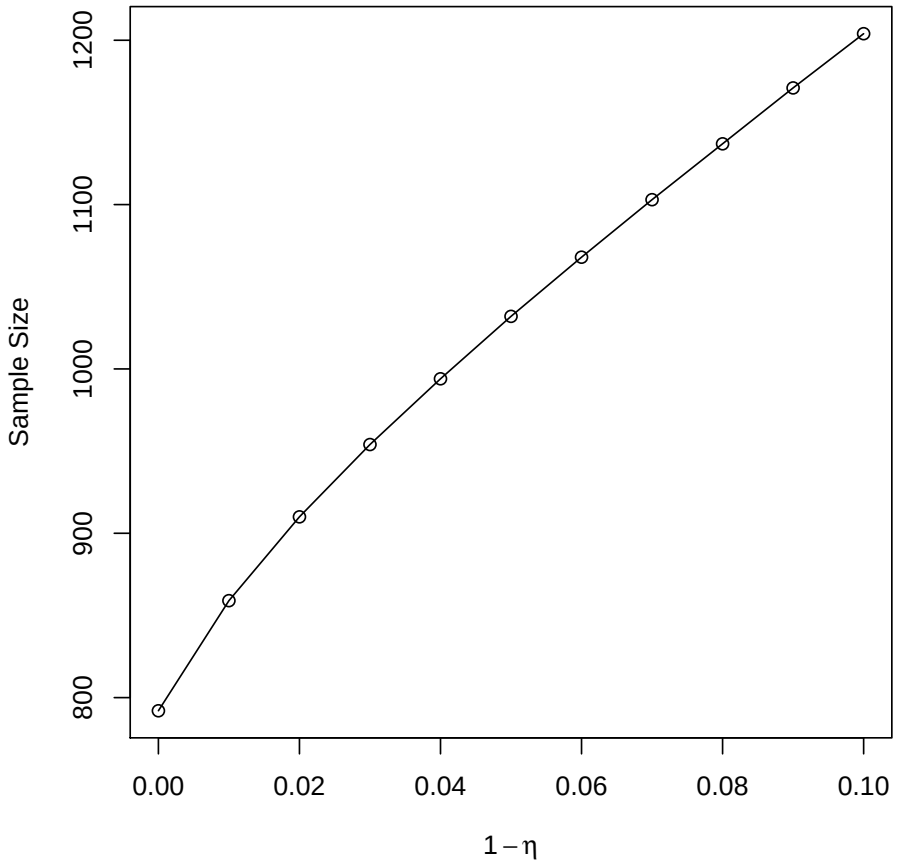

Figure A.1. Effect of using an imperfect diagnostic procedure at study entry. Results are based on the assumptions of annual visits over a study duration of 8 years, sensitivity=0.61, specificity=0.995, HR=2, $S_{J+1} = 0.9$, and that there are no missing tests, where type I error is fixed at 0.05 and power is fixed at 0.9 corresponding to a two-sided hopothesis test. η denotes the negative predictive value of the diagnostic test at baseline.

BIBLIOGRAPHY

[1] Anderson, G., Cummings, S., Freedman, L. S., Furberg, C., Henderson, M., Johnson, S. R., Kuller, L., Manson, J., Oberman, A., Prentice, R. L., Rossouw, J. E., and Grp, Women's Hlth Initiative Study. Design of the women's health initiative clinical trial and observational study. *Controlled Clinical Trials 19*, 1 (1998), 61–109.

[2] Balasubramanian, R., and Lagakos, S. W. Estimation of the timing of perinatal transmission of hiv. *Biometrics 57* (2001), 1048–1058.

[3] Balasubramanian, R., and Lagakos, S. W. Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika 90* (2003), 171–182.

[4] Chen, H. H., Duffy, S. W., and Tabar, L. A markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *Statistician 45* (1996), 307–317.

[5] Collett, D. *Modelling Survival Data in Medical Research, Second Edition*. Texts in statistical science. Taylor & Francis, 2003.

[6] Cook, T. D. Adjusting survival analysis for the presence of unadjudicated study events. *Controlled Clinical Trials 21* (2000), 208–222.

[7] Cook, T. D., and Kosorok, M. R. Analysis of time-to-event data with incomplete event adjudication. *Journal of the American Statistical Association 99* (2004), 1140–1152.

[8] Cox, D. R., and Hinkley, D. V. *Theoretical Statistics*. Chapman and Hall, 1979.

[9] Culver, A. L., Ockene, I. S., Balasubramanian, R., Olendzki, B. C., Sepavich, D. M., Wactawski-Wende, J., Manson, J. E., Qiao, Y. X., Liu, S. M., Merriam, P. A., Rahilly-Tierny, C., Thomas, F., Berger, J. S., Ockene, J. K., Curb, J. D., and Ma, Y. S. Statin use and risk of diabetes mellitus in postmenopausal women in the women's health initiative. *Archives of Internal Medicine 172*, 2 (2012), 144–152.

[10] Dunn, D. T., Simonds, R. J., Bulterys, M., Kalish, L. A., Moye, J., de Maria, A., Kind, C., Rudin, C., Denamur, E., Krivine, A., Loveday, C., and Newell, M. L. Interventions to prevent vertical transmission of HIV-1: effect on viral detection rate in early infant samples. *AIDS 14*, 10 (2000), 1421–1428.

[11] Dunson, D. B., Herring, A. H., and Engel, S. M. Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association 103*, 482 (2008), 534–546.

[12] Fay, Michael P., and Shaw, Pamela A. Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software 36*, 2 (2010), 1–34.

[13] Finkelstein, D. M. A proportional hazards model for interval-censored failure time data. *Biometrics 42* (1986), 845–854.

[14] Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. Pathwise coordinate optimization. *Annals of Applied Statistics 1*, 2 (2007), 302–332.

[15] Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1 (2010), 1–22.

[16] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003.

[17] George, E. I., and Mcculloch, R. E. Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*, 423 (1993), 881–889.

[18] Goetghebeur, E., and Ryan, L. Semiparametric regression analysis of interval-censored data. *Biometrics 56*, 4 (2000), 1139–1144.

[19] Gomez, G., Calle, M. L., Oller, R., and Langohr, K. Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling 9*, 4 (2009), 259–297.

[20] Gu, Xiangdong, and Balasubramanian, Raji. *icensmis: Study Design and Data Analysis in the presence of error-prone diagnostic tests and self-reported outcomes*, 2012. R package version 1.0.

[21] Gu, Xiangdong, and Balasubramanian, Raji. *straweib: Stratified Weibull Regression Model*, 2013. R package version 1.0.

[22] Guan, Y. T., and Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics 5*, 3 (2011), 1780–1815.

[23] Guihenneuc-Jouyaux, C., Richardson, S., and Longini, I. M. Modeling markers of disease progression by a hidden markov process: Application to characterizing cd4 cell decline. *Biometrics 56* (2000), 733–741.

[24] He, C. Y., Zhang, C. L., Hunter, D. J., Hankinson, S. E., Louis, G. M. B., Hediger, M. L., and Hu, F. B. Age at menarche and risk of type 2 diabetes: Results from 2 large prospective cohort studies. *American Journal of Epidemiology 171* (2010), 334–344.

[25] Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., and Willett, W. C. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine 345* (2001), 790–797.

[26] Jackson, C. H., and Sharples, L. D. Hidden markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine 21* (2002), 113–128.

[27] Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., and Couto, E. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society Series D-the Statistician 52* (2003), 193–209.

[28] Kim, S., Tadesse, M. G., and Vannucci, M. Variable selection in clustering via dirichlet process mixture models. *Biometrika 93*, 4 (2006), 877–893.

[29] Kirby, A.J., and Spiegelhalter, D.J. *Statistical modelling for the precursors of cervical cancer*. Wiley, New York, 1994.

[30] Lee, K. E., Sha, N. J., Dougherty, E. R., Vannucci, M., and Mallick, B. K. Gene selection: a bayesian variable selection approach. *Bioinformatics 19*, 1 (2003), 90–97.

[31] Leroy, R., Bogaerts, K., Lesaffre, E., and Declerck, D. The emergence of permanent teeth in flemish children. *Community Dentistry and Oral Epidemiology 31*, 1 (2003), 30–39.

[32] Li, F., and Zhang, N. R. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association 105*, 491 (2010), 1202–1214.

[33] Lindsey, J. C., and Ryan, L. M. Tutorial in biostatistics - methods for interval-censored data. *Statistics in Medicine 17*, 2 (1998), 219–238.

[34] Margolis, K. L., Qi, L. H., Brzyski, R., Bonds, D. E., Howard, B. V., Kempoinen, S., Liu, S. M., Robinson, J. G., Safford, M. M., Tinker, L. T., Phillips, L. S., and Womens Hlth, Initiative. Validity of diabetes self-reports in the women's health initiative: comparison with medication inventories and fasting glucose measurements. *Clinical Trials 5* (2008), 240–247.

[35] McCarthy, M. I. Genomic medicine genomics, type 2 diabetes, and obesity. *New England Journal of Medicine 363*, 24 (2010), 2339–2350.

[36] McKeown, Karen, and Jewell, Nicholas P. Misclassification of current status data. *Lifetime Data Analysis 16* (2010), 215–230.

[37] McNamee, R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Statistics in Medicine 21*, 23 (2002), 3609–3625.

[38] McNamee, R. Optimal design and efficiency of two-phase case-control control studies with error-prone and error-free exposure measures. *Biostatistics 6*, 4 (2005), 590–603.

[39] Meeker, W.Q., and Escobar, L.A. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Wiley, 1998.

[40] Meier, A. S., Richardson, B. A., and Hughes, J. P. Discrete proportional hazards models for mismeasured outcomes. *Biometrics 59* (2003), 947–954.

[41] Mitchell, T. J., and Beauchamp, J. J. Bayesian variable selection in linear-regression. *Journal of the American Statistical Association 83*, 404 (1988), 1023–1032.

[42] Neuhaus, J. M. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika 86* (1999), 843–855.

[43] Satten, G. A., and Longini, I. M. Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistics-Journal of the Royal Statistical Society Series C 45* (1996), 275–295.

[44] Sha, N., Tadesse, M. G., and Vannucci, M. Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics 22*, 18 (2006), 2262–8.

[45] Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics 60*, 3 (2004), 812–9.

[46] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software 39*, 5 (2011), 1–13.

[47] Snapinn, S. M. Survival analysis with uncertain endpoints. *Biometrics 54* (1998), 209–218.

[48] Stingo, F. C., Chen, Y. A. A., Tadesse, M. G., and Vannucci, M. Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics 5*, 3 (2011), 1978–2002.

[49] Tadesse, M. G., Sha, N., and Vannucci, M. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association 100*, 470 (2005), 602–617.

[50] Therneau, Terry. *A Package for Survival Analysis in S*, 2012. R package version 2.36-14.

[51] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B-Methodological 58*, 1 (1996), 267–288.

[52] Tibshirani, R. The lasso method for variable selection in the cox model. *Statistics in Medicine 16*, 4 (1997), 385–395.

[53] Turnbull, B. W. Empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B-Methodological 38* (1976), 290–295.

[54] Vanobbergen, J., Martens, L., Lesaffre, E., and Declerck, D. The Signal-Tandmobiel project a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results. *Eur J Paediatr Dent 2* (2000), 87–96.

[55] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics 25*, 6 (2009), 714–721.