Published in final edited form as:

Stat Med. 2016 September 30; 35(22): 3961–3975. doi:10.1002/sim.6962.

Study design for non-recurring, time to event outcomes in the presence of error-prone diagnostic tests or self-reports

XIANGDONG GU^a and RAJI BALASUBRAMANIAN*

^aDepartment of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, USA

Abstract

Sequentially administered, laboratory-based diagnostic tests or self-reported questionnaires are often used to determine the occurrence of a silent event. In this paper, we consider issues relevant in design of studies aimed at estimating the association of one or more covariates with a non-recurring, time to event outcome that is observed using a repeatedly administered, error-prone diagnostic procedure. The problem is motivated by the Women's Health Initiative, in which diabetes incidence among the approximately 160,000 women is obtained from annually collected self-reported data. For settings of imperfect diagnostic tests or self-reports with known sensitivity and specificity, we evaluate the effects of various factors on resulting power and sample size calculations and compare the relative efficiency of different study designs. The methods illustrated in this paper are readily implemented using our freely available R software package *icensmis* [1], which is available at the Comprehensive R Archive Network (CRAN) website. An important special case is that when diagnostic procedures are perfect, resulting in interval-censored, time to event outcomes. The proposed methods are applicable for the design of studies in which a time to event outcome is interval-censored.

Keywords

imperfect diagnostic tests; interval censoring; self-reports; study design; time to event outcomes

1. Introduction

The onset of several chronic diseases such as diabetes and cardiovascular disease (CVD) are asymptomatic and can only be detected by diagnostic tests. For example, clinical diabetes is diagnosed by measuring fasting blood glucose or glycosylated hemoglobin (HbA1c) levels. While gold standard diagnostic tests exist for several conditions, in many large-scale epidemiological studies such as the Women's Health Initiative (WHI), cost considerations dictate that disease status ascertainment be based on self-reports. Self reported outcomes of prevalent and incident disease have been found to be less than perfect. In a WHI substudy, diabetes self-reports were characterized by sensitivity and specificity estimates of approximately 55% and 99%, respectively [2]. Other studies such as the Nurses Health Study and Physician's Health Study also commonly use self-reported outcomes [3, 4]. A

*Correspondence to: Raji Balasubramanian, Division of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, USA. rbalasub@schoolph.umass.edu.

study in a prospective Finnish cohort showed that the sensitivity of self-reports for incident disease (diabetes, hypertension, asthma, coronary heart disease and rheumatoid arthritis) was between 55% and 63% [5]. Moreover, several laboratory based diagnostic tests such as DNA PCR tests used to detect HIV infection in infants are also subject to imperfect sensitivity and specificity [6]. Lastly, there is much interest in the scientific community in utilizing electronic medical records (EMRs) for clinical research [7, 8, 9]. These databases are especially attractive in studying rare outcomes. A fundamental challenge in utilizing data from EMRs is the possibility of misclassified outcomes and/or covariates. The inherent error in outcome ascertainment in all these settings renders standard regression methods for interval-censored, time-to-event outcomes invalid.

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. For example, consider a subject who is event free at baseline (time 0) and who is given a perfect diagnostic test at follow-up times 1 and 2 years, with test results that are negative and positive, respectively - the true time to event can be inferred to lie in the interval (1,2]. In this context, methods for estimating the survival distribution and assessing the effect of covariates have been developed [10, 11]. However, if the diagnostic tests are imperfect, the true time to event is no longer constrained to lie in the interval (1,2], due to the possibility of false positive and/or false negative results. In this setting of error-prone diagnostic procedures such as self-reports, standard regression methods for interval censored outcomes are rendered invalid. In recent work, Gu, X. *et al.* (2015) [12] present a likelihood-based approach to estimate the association of specific covariates with a time-to-event outcome that is observed through periodic, imperfect self-reports. In this work, the authors show through simulations that analysis procedures that ignore the error inherent in imperfect diagnostic procedures can produce results that are substantially biased. Consider the setting in which the cumulative incidence of events in the study is 10% and the diagnostic procedure has perfect specificity, but imperfect sensitivity of 0.75. When analysis proceeds by ignoring the error in the diagnostic procedure, the resulting estimates of the regression coefficients of interest can be biased as significantly as 90% towards the null. However, when the regression analysis is based on a likelihood procedure that incorporates the error-prone nature of the diagnostic test, this bias is eliminated [12].

Previous work in a regression context include likelihood-based methods for error-prone outcomes that incorporate specific characteristics of diagnostic tests used in studies in HIV, HPV and STD research [13, 14, 15, 16, 17]. Regression models in the presence of error-prone time to event outcomes can also be developed through a Hidden Markov model (HMM) framework. Applications of HMM based methods have been published in the areas of breast cancer [18], HIV [19, 20], lung transplantation [21] and cervical smear tests [22]. Other related work include methods for incomplete event adjudication [23, 24, 25] and misclassification due to error-prone screening tests [26]. However, none of these previously published papers consider the problem of power and sample size calculations, relevant in the design of studies that incorporate imperfect, self-reports or laboratory-based diagnostic procedures. In the context of study design, McNamee, R. (2002, 2006) [27, 28] developed methods to find optimal designs when both error-prone and error-free measurements are

used in two-stage studies that are used to estimate the sensitivity, specificity and positive predictive value of a diagnostic test. However, these methods are not appropriate for settings in which only imperfect diagnostic tests or self-reported outcomes are collected at repeated visits, in lieu of expensive, perfect diagnostic tests.

In this paper, we extend the semi-parametric likelihood based approach described in Gu, X. *et al.* (2015) [12] to obtain power and sample size calculations. We consider studies aimed at estimating the association of one or more covariates with a time to event outcome that is observed using a repeatedly administered, error-prone diagnostic procedure such as a self-reported questionnaire. We describe the effects of various factors influencing the sample size and statistical power and consider the relative efficiency of different study designs. An important special case of the proposed methods is that when a perfect diagnostic procedure is administered repeatedly, to detect a silent event. In this setting, the outcomes are interval-censored and the proposed methods for study design equally apply to this setting. We provide a freely available R software package *icensmis* [1], the use of which is illustrated in the Tutorial included in the Supplemental Material. In Section 2, we describe the methods to calculate power and sample size including extensions to incorporate the effects of missing tests and censoring. In Section 3, we illustrate the impact of error in self-reported outcomes on power and sample size calculations, motivated by the WHI. In Section 4, we illustrate the effects of different factors influencing power and sample size. Lastly, in Section 5, we discuss the findings of this study and highlight future directions.

2. Method

We present the notation, likelihood, and discuss procedures for power and sample size calculations in Section 2.1. In Section 2.2, we present expressions for power and sample size in the special case of a single follow-up visit. We present extensions to account for missing data and censoring in Section 2.3.

2.1. Likelihood, power and sample size calculations

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. We denote the probability of event A by $p(A)$. Thus, $p(X > 0) = 1$. Let N denote the number of subjects and n_i denote the number of visits (tests) for the i^{th} subject. At each visit, we assume that each subject would self report their disease status (yes/no) or obtain a laboratory based diagnostic test result (positive/negative). For example, in the WHI, information on incident diabetes was collected at periodically 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 self-reported or laboratory-based, binary outcomes and corresponding visit times, respectively. In the following, we use the term *test result* to denote both self-reported outcomes or results from laboratory-based diagnostic tests. 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 such as diabetes) and 0 otherwise. We assume that test results are collected at pre-scheduled visits or according to an adaptive paradigm. For example, several studies

are designed such that self-reports are collected only up to the time of the first positive self-report - thus, the vectors of test results ($\mathbf{R}_i$), visit times ($\mathbf{t}_i$) and the number of self-reports collected per subject (n_i) are random. Let $\tau_1, \dots, \tau_J$ denote the distinct, ordered visit times in the dataset among N subjects, where $0 = \tau_0 < \tau_1 < \dots < \tau_J < \tau_{J+1} = \infty$ - thus, the time axis can be divided into $J+1$ disjoint intervals, $[0, \tau_1), [\tau_1, \tau_2), \dots, [\tau_J, \infty)$.

The joint probability of the observed data for the i th subject can be expressed as:

$$\begin{aligned} g(\mathbf{R}_i, \mathbf{t}_i, n_i) &= \sum_{j=1}^{J+1} p(\tau_{j-1} < X_i \leq \tau_j) p(\mathbf{R}_i, \mathbf{t}_i, n_i | \tau_{j-1} < X_i \leq \tau_j) \\ &= \sum_{j=1}^{J+1} \theta_j p(\mathbf{R}_i, \mathbf{t}_i, n_i | \tau_{j-1} < X_i \leq \tau_j) \end{aligned}$$

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

To simplify the form of the expression above, we make the assumption that given the true time of event X_i , an individual's n_i test results are independent. That is,

$$p(\mathbf{R}_i | X_i, \mathbf{t}_i) = \prod_{k=1}^{n_i} p(r_{ik} | X_i, t_{ik})$$

This assumption implies that the observed values of other test results do not provide additional information about the distribution of a particular test result from that provided by the actual time of the event.

Based on the derivation in [14], it can be shown that the joint probability of the observed data for the i th subject can be simplified as:

$$\begin{aligned} g(\mathbf{R}_i, \mathbf{t}_i, n_i) &= \sum_{j=1}^{J+1} \theta_j \left[\prod_{k=1}^{n_i} p(r_{ik} | \tau_{j-1} < X_i \leq \tau_j, t_{ik}) \right] \\ &= \sum_{j=1}^{J+1} \theta_j C_{ij} \end{aligned} \quad (1)$$

where $C_{ij} = \left[\prod_{k=1}^{n_i} p(r_{ik} | \tau_{j-1} < X_i \leq \tau_j, t_{ik}) \right]$. We assume that the probability of a positive test result at the k th visit ($r_{ik} = 1$) conditional on the interval containing the true event time and visit time can be expressed as:

$$p(r_{ik} = 1 | \tau_{j-1} < X_i \leq \tau_j, t_{ik}) = \begin{cases} \varphi_1, & t_{ik} \geq \tau_j \\ 1 - \varphi_0, & t_{ik} \leq \tau_{j-1} \end{cases}$$

Here, φ_1 and φ_0 denote the sensitivity and specificity of self-reports, respectively. Thus the terms C_{ij} for $j = 1, \dots, J+1$ in equation (1) can be expressed as a product involving the constants φ_1 and φ_0 . A specific data example and the corresponding form of the matrix C is

shown in Section 1 of the Supplemental Material. Thus, in the absence of covariates, the log-likelihood for a random sample of N subjects can be expressed as:

$$l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta})) = \sum_{i=1}^N \log\left(\sum_{j=1}^{J+1} C_{ij} \theta_j\right) \quad (2)$$

For the special case where self-reports are perfect ($\varphi_1 = \varphi_0 = 1$), equation (2) reduces to the non-parametric likelihood for interval censored observations given in [10].

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 event of interest. Let $\mathbf{Z}$ 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, $\lambda(x|\mathbf{Z} = \mathbf{z}) = \lambda_0(x)e^{\mathbf{z}'\boldsymbol{\beta}}$, or equivalently, $S(x|\mathbf{Z} = \mathbf{z}) = S_0(x)e^{-\mathbf{z}'\boldsymbol{\beta}}$.

To derive the form of the log-likelihood incorporating the proportional hazards model, we first re-parameterize the log-likelihood in (2) in terms of the survival function, $\mathbf{S} = (1$

$= S_1, S_2, \dots, S_{J+1})^T$, where $S_j = p(X > \tau_{j-1})$. Since $S_j = \sum_{l=j}^{J+1} \theta_l$, the vector of interval probabilities can be expressed as $\boldsymbol{\theta} = T_r \mathbf{S}$, where T_r is the $(J+1) \times (J+1)$ transformation matrix. Let $\mathbf{C} = [C_{ij}]$ denote the $N \times (J+1)$ matrix of the coefficients, C_{ij} , and let the $N \times (J+1)$ matrix $\mathbf{D}$ be defined as $\mathbf{D}_{N \times (J+1)} = \mathbf{C} \times T_r$. The form of the transformation matrix T_r is given in Section 1 of the Supplemental Material. Then, the log-likelihood function for the one-sample setting in (2) can be expressed as

$$l(\mathbf{S}) = \sum_{i=1}^N \log\left(\sum_{j=1}^{J+1} D_{ij} S_j\right), \quad (3)$$

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

Let $1 = S_1 > S_2 > \dots > S_{J+1}$ denote the baseline survival functions (i.e. corresponding to $\mathbf{Z} = \mathbf{0}$), evaluated at the left boundaries of the intervals $[0, \tau_1), [\tau_1, \tau_2), \dots, [\tau_J, \infty)$. Then, for

subject i , with corresponding covariate vector $\mathbf{z}_i$, $S_j^{(i)} = (S_j)^{e^{\mathbf{z}_i' \boldsymbol{\beta}}}$. Thus, the log-likelihood function for a random sample of N subjects in equation (3) can be extended to incorporate covariates as

$$l(\mathbf{S}, \boldsymbol{\beta}) = \sum_{i=1}^N \log\left(\sum_{j=1}^{J+1} D_{ij} (S_j)^{e^{\mathbf{z}_i' \boldsymbol{\beta}}}\right). \quad (4)$$

The elements of the $\mathbf{D}$ matrix are functions of the observed data including the visit times and corresponding test 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 $\beta_1, \dots, \beta_P, S_2, \dots, S_{J+1}$ can be obtained by numerical maximization of the log-likelihood function in equation (4), subject to the constraints that $1 > S_2 > S_3 > \dots > S_{J+1} > 0$. Statistical inference regarding the parameters of interest ($\beta_1, \dots, \beta_P, S_2, \dots, S_{J+1}$) can be made by using asymptotic properties of the maximum likelihood estimators [29]. 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.

To derive expressions for power and sample size, we consider the setting of two groups, where the hazard and survival functions in each group are denoted by $\lambda_k(x)$ and $S_k(x)$, for $k = 1, 2$. Assuming the proportional hazards model, $\lambda_2(x) = e^{\beta} \lambda_1(x)$, the null hypothesis of equality of survival or hazard functions is equivalent to the null hypothesis, $H_0 : \beta = 0$. Hypothesis tests regarding β can be carried out based on the test statistic

$$Y = \frac{\hat{\beta}}{\sqrt{\text{Var}(\hat{\beta})}},$$

which under the null hypothesis $H_0 : \beta = 0$, is asymptotically normal. Power (denoted by ρ) can be calculated using the asymptotic normal property of the test statistic,

$$\rho = p \left(Z_G > z_{1-\frac{\alpha}{2}} - \frac{\beta}{\sqrt{\text{Var}(\hat{\beta})}} \right) + p \left(Z_G < -z_{1-\frac{\alpha}{2}} - \frac{\beta}{\sqrt{\text{Var}(\hat{\beta})}} \right), \quad (5)$$

where $Z_G \sim N(0, 1)$, α denotes type I error, $p(Z_G > z_{1-\frac{\alpha}{2}}) = \alpha/2$ and $\text{Var}(\hat{\beta})$ is the variance of $\hat{\beta}$ obtained from the inverse of the expected Fisher information matrix.

Alternatively, the expression for total sample size obtained by assuming equal allocation to the two exposure groups is given by:

$$N = \frac{(z_{1-\alpha/2} + z_{\rho})^2 \text{Var}_1(\hat{\beta})}{\beta^2}, \quad (6)$$

where $\text{Var}_1(\hat{\beta})$ is obtained as the inverse of the expected Fisher information for one unit (i.e. $\text{Var}(\hat{\beta}) = \frac{\text{Var}_1(\hat{\beta})}{N}$). For the general setting, details of the derivation of the analytical expressions for $\text{Var}(\hat{\beta})$ are presented in Section 2 of the Supplemental Material. While not explicitly stated, the expressions for power (ρ) in equation (5) and sample size N in equation (6) are evaluated at the value of β under the alternate hypothesis, $H_a : \beta = \beta_a$.

2.2. Special case: Single visit

We present expressions for power and sample size for the special case in which all subjects come back for a single follow-up visit at time, $\tau_1 > 0$. At τ_1 , all subjects are tested for the event of interest. Let $R = 0$ and $R = 1$ denote a negative and positive test result, respectively. We note that all subjects are assumed to be event-free at baseline, $\tau = 0$. As before, let φ_1 and φ_0 denote the sensitivity and specificity of the diagnostic test, respectively. Assume a total sample size of N subjects, with equal allocation of subjects to the two groups indicated by $Z = 0$ and $Z = 1$. Let $p = P(X > \tau_1 | Z = 0) = S(\tau_1 | Z = 0)$. We derive expressions for the likelihood contribution for all possible (Z, R) - let L_{ik} ($L_{00}, L_{01}, L_{10}, L_{11}$) denote the likelihood function for group i ($i = 0, 1$) with k^{th} test result ($k = 0, 1$).

Thus,

$$\begin{aligned} L_{00} &= 1 - \varphi_1 + (\varphi_0 + \varphi_1 - 1)p \\ L_{01} &= \varphi_1 + (1 - \varphi_0 - \varphi_1)p \\ L_{10} &= 1 - \varphi_1 + (\varphi_0 + \varphi_1 - 1)p^{e^\beta} \\ L_{11} &= \varphi_1 + (1 - \varphi_0 - \varphi_1)p^{e^\beta} \end{aligned}$$

The elements of the expected Fisher Information matrix $I(p, \beta)$ are given by:

$$\begin{aligned} I_{11} &= \frac{N}{2} (1 - \varphi_1 - \varphi_0)^2 e^{2\beta} p^{2e^\beta} \log^2(p) \frac{1}{L_{10}L_{11}} \\ I_{12} &= \frac{N}{2} (1 - \varphi_1 - \varphi_0)^2 e^{2\beta} p^{2e^\beta - 1} \log(p) \frac{1}{L_{10}L_{11}} \\ I_{21} &= I_{12} \\ I_{22} &= \frac{N}{2} (1 - \varphi_1 - \varphi_0)^2 \left(\frac{1}{L_{00}L_{01}} + \frac{e^{2\beta} p^{2e^\beta - 2}}{L_{10}L_{11}} \right) \end{aligned}$$

Thus, the expression for $Var_1(\hat{\beta})$ is:

$$Var_1(\hat{\beta}) = \frac{2}{(1 - \varphi_1 - \varphi_0)^2 p^2 \log^2(p)} \left(L_{00}L_{01} + \frac{L_{10}L_{11}}{e^{2\beta} p^{2e^\beta - 2}} \right)$$

Power and sample size formulae can be obtained by substituting the expression for $Var_1(\hat{\beta})$ into equations (5) and (6).

For the special case of a perfect diagnostic test, resulting in interval-censored time to event outcomes, the expression for $Var_1(\hat{\beta})$ reduces to:

$$Var_1(\hat{\beta}) = \frac{2}{p \log^2(p)} (1 - p + p^{1-e^\beta} e^{-2\beta} (1 - p^{e^\beta}))$$

2.3. Missing tests and censoring

Sample size and power calculations can be extended to incorporate the possibility of missing tests and/or censoring. Let J denote the number of scheduled test times, let $\mathbf{R}$ denote the $J \times 1$ vector of binary test results and let $\mathbf{I}$ denote the $J \times 1$ vector of binary indicators

corresponding to the pattern of missing-ness. 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 $\mathbf{I}$ is fully observed. Let $\mathbf{R}_{obs}$ and $\mathbf{R}_{mis}$ denote the observed and missing components of the vector $\mathbf{R}$. Let Θ denote the parameters governing the likelihood function (i.e. $\Theta = (S_2, \dots, S_{J+1}, \beta)$), and let Φ denote the parameters governing the missing mechanism.

We apply Rubin's taxonomy with regard to missing data mechanisms, namely Missing Completely At Random (MCAR) versus Missing At Random (MAR) [30, 31, 32]. Specifically, to handle missing data and censoring, we invoke the concepts defined below:

- **MCAR:** Under this mechanism, the probability of non-response $\mathbf{I}$ depends neither on an individual's test results, $\mathbf{R}$ nor on their covariates, $\mathbf{Z}$. In other words,

$$p(\mathbf{R}_{obs}, \mathbf{I} | \Phi, \Theta, \mathbf{Z}) = p(\mathbf{I} | \Phi) p(\mathbf{R}_{obs} | \Theta, \mathbf{Z}) \quad (7)$$

As described in [32], an example of the MCAR mechanism is when the number of follow-up visits differs by individual due to staggered entry and administrative censoring at a fixed calendar time.

- **MAR:** Under this mechanism, the probability of non-response $\mathbf{I}$ can depend on observable test results $\mathbf{R}_{obs}$ and covariates $\mathbf{Z}$, but conditional on these components, is independent of $\mathbf{R}_{mis}$.

$$p(\mathbf{R}_{obs}, \mathbf{I} | \Phi, \Theta, \mathbf{Z}) = p(\mathbf{I} | \Phi, \mathbf{R}_{obs}, \mathbf{Z}) p(\mathbf{R}_{obs} | \Theta, \mathbf{Z}) \quad (8)$$

For example, in a longitudinal study, among participants with the same covariate profile ($\mathbf{Z}$), those who are observed to be sicker through their values of Y_{obs} are more likely to have missing values, so long as their missingness probability does not depend on the missing responses (Y_{mis}) [32].

- **Non-informative censoring:** is the condition under which standard counting process statistics (e.g. logrank test statistic, Cox partial likelihood score statistic, etc.) retain their martingale properties. Non-informative censoring holds if the hazard of an uncensored failure time X is equivalent to the hazard under censoring. That is,

$$p(x \leq X \leq x + \delta x | X \geq x, \mathbf{Z}) = p(x \leq X \leq x + \delta x | X \geq x, C \geq x, \mathbf{Z}),$$

where C denotes censoring time. Non-informative censoring can be shown to be equivalent to Sequential MAR (S-MAR), under which the probability of drop-out at time τ_j may depend on both the covariate history up to and including τ_j (i.e. $\{X_1, \dots, X_j\}$) and observable test results up to but not including τ_j ($\{R_1, \dots, R_{j-1}\}$) [32]. We note that the S-MAR

condition is more restrictive than MAR - under MAR, the probability of missingness can depend on past, current and future values of R_{obs} , Z [32].

In equations (7) and (8) under MCAR and MAR, the term $p(\mathbf{R}_{obs}|\Theta, \mathbf{Z})$ corresponds to the previously described likelihood function based on the observed data [equation (4)]. We let ϕ denote the likelihood corresponding to the missing mechanism - under MCAR [equation (7)], $\phi = p(\mathbf{I} | \Phi)$; under MAR [equation (8)], $\phi = p(\mathbf{I} | \mathbf{R}_{obs}, \mathbf{Z}, \Phi)$. Since ϕ 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 ($\Theta = (S_2, \dots, S_{J+1}, \beta)$), as well as the second derivatives of the likelihood function that are required for calculating the Fisher information matrix.

Let $\mathbf{R}_{i,obs}$ and I_i denote the i^{th} pattern of test results and the corresponding vector of indicator variables denoting observed versus missing data, respectively. Let q_{ki} denote the number of subjects in the k^{th} exposure group having the i^{th} pattern of test results, where $k = 0, 1$ and $i = 1, 2, \dots, 2^J$. To derive expressions for power and sample size, we obtain expressions for the elements of the expected Fisher information matrix, $I_{(J+1) \times (J+1)}$ (See Section 2.1 of the Supplement), whose elements are given by:

$$I_{jj'} = -E\left(\frac{\partial^2 l}{\partial \gamma_j \partial \gamma_{j'}}\right) = -\sum_{k=0}^1 \sum_{i=1}^{2^J} E(q_{ki}) \frac{\partial^2 l_{ki}}{\partial \gamma_j \partial \gamma_{j'}}$$

In the presence of missing data under MCAR or MAR assumptions, the expected number of subjects in k^{th} group with the i^{th} observed pattern of test results (q_{ki}) is given by

$$\begin{aligned} E(q_{ki}) &= N_k \phi_i p(\mathbf{R}_{i,obs} | \Theta_k, Z=k) \\ &= N_k \phi_i \sum_{l=1}^{J+1} D_{il} S_{kl}, \end{aligned}$$

where ϕ_i denotes the probability of the i^{th} pattern of missingness and N_k (for $k = 0, 1$) denotes the number of subjects in the k^{th} group defined according to the values of the covariate Z . 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 present the form of ϕ for specific settings:

MCAR—We assume that each test can be independently missing with probability p_{miss} . 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.

Adaptive testing schedule—Studies that collect self-reported outcomes will often cease to collect further information following the first positive self-report. This study design is especially useful when the specificity of the diagnostic test (or self-report) is perfect. We refer to this design as NTFP (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 independently with probability p_{miss} . On the other hand, following the first positive test result, all tests are missing with probability 1. Thus,

$$\phi_i = p(I_i | \Phi, \mathbf{R}_{i,obs}, \mathbf{Z}) = p(I_{i1} | \Phi, \mathbf{R}_{i,obs}, \mathbf{Z}) p(I_{i2} | I_{i1}, \Phi, \mathbf{R}_{i,obs}, \mathbf{Z}) \cdots p(I_{iJ} | I_{i1}, \dots, I_{i,J-1}, \Phi, \mathbf{R}_{i,obs}, \mathbf{Z})$$

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

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

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

Censoring—The presence of non-informative censoring can also be incorporated into power and sample size calculations through the distribution of the missing data mechanism, ϕ . Let X_c denote the random variable corresponding to censoring time and the probability of being censored in j^{th} interval, $p(\tau_{j-1} < X_c < \tau_j) = c_j$. Thus, the distribution of missing pattern when censoring is present can be expressed as:

$$\phi_i = \sum_{j=J_i+1}^{J+1} c_j p(I_i | \Phi, \mathbf{R}_{i,obs}, \mathbf{Z}, \tau_{j-1} \leq X_c < \tau_j),$$

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, $\tau_1, \dots, \tau_4$, dividing the time axis into 5 intervals. The probability of being censored in the j^{th} interval is denoted by c_j for $j = 1, \dots, 5$,

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 of the censoring process. Consider a pattern of test results that has an associated missing-ness pattern $\mathbf{I}_i = (1, 0, 1, 0)$, 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 4th interval ($[\tau_3, \tau_4)$) or the 5th interval ($[\tau_4, +\infty)$).

Conditional on $X_c \in [\tau_3, \tau_4)$, there are 1 missing and 2 non-missing tests before censoring, thus the probability of the missing pattern is $p(\mathbf{I}_i | \Phi, \mathbf{R}_{i,obs}, \mathbf{Z}, \tau_3 < X_c < \tau_4) = (1 - p_{miss})^2 \times p_{miss}$. Similarly, conditional on $X_c \in [\tau_4, +\infty)$, we have

$$p(\mathbf{I}_i | \Phi, \mathbf{R}_{i,obs}, \mathbf{Z}, \tau_4 \leq X_c < +\infty) = (1 - p_{miss})^2 \times p_{miss}^2. \text{ Thus,}$$

$$\phi_i = c_4 \times (1 - p_{miss})^2 \times p_{miss} + c_5 \times (1 - p_{miss})^2 \times p_{miss}^2.$$

3. Impact of error-prone self-reports on power and sample size

We present an analysis of the trade-off between perfect and imperfect diagnostic tests with respect to statistical power and sample size. We incorporate parameter settings that reflect the characteristics of diabetes self-reports in the WHI.

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-reports have been evaluated by [2] using data obtained in a sub-study 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%. This corresponds to a sensitivity and specificity of 0.55 and 0.99, respectively based on the fact that approximately 5% of women self-reported diabetes at baseline [2]. Based on fasting glucose levels, the true prevalence of diabetes at baseline in the WHI substudy was approximately 6.6% [2].

Results—We present the trade-off in power versus sample size, comparing a perfect test to tests with varying degrees of imperfect sensitivity and/or specificity. We assume that the primary goal of the study is to use a two-sided test, to test $H_0 : \beta = 0$ corresponding to a single binary exposure variable, Z . We assume equal allocation to both groups and that the type I error is fixed at 0.05. Further, to specify the survival function at each test time ($S_2, \dots, S_{J+1}$), we assume that the time to the event of interest (X) within each group follows an exponential distribution, with density function $f_k(x) = \lambda_k e^{-\lambda_k x}$, for $k = 0, 1$. The parameter governing the exponential distribution in the reference group ($k = 0$) is determined by the cumulative incidence by the end of study ($1 - S_{J+1}$, assuming a study duration of eight years). The parameter governing the exponential distribution in the non-reference group ($k = 1$) is determined by assuming a proportional hazards model ($\lambda_1 = \lambda_0 e^\beta$). We note that the methods proposed in the paper do not require the exponential distribution assumption for X nor specifying the duration of the study - in this illustration, these assumptions are only used to set the values of $S_2, \dots, S_{J+1}$ at each visit time $\tau_1, \dots, \tau_J$ respectively.

Figure 1 shows the power versus sample size curves corresponding to different values of (sensitivity, specificity), assuming that the hazard ratio ($HR = e^\beta$) between the two groups is equal to 1.25. We assumed that the tests (visits) are scheduled either annually ($J = 8$) or every 4 years ($J = 2$), and that the cumulative incidence in the reference group is either $1 - S_{J+1} = 0.1$ or $1 - S_{J+1} = 0.5$ - corresponding to a mean time to event 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 (ϕ_1, ϕ_0) . The effects of imperfect sensitivity and specificity on sample size and power vary considerably with cumulative

incidence rate. When the cumulative incidence rate is low ($S_{J+1} = 0.9$), reduction in specificity (φ_0) has a significant impact on power. However, for this setting, a corresponding reduction in sensitivity (φ_1) 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, when compared to the test with (φ_1, φ_0) of $(1.00, 0.75)$, the test with (φ_1, φ_0) of $(0.55, 0.99)$ has higher power when cumulative incidence rates are low ($S_{J+1} = 0.9$) but has lower power when cumulative incidence rates are high ($S_{J+1} = 0.5$). Specifically, to achieve a power of 0.90 when cumulative incidence is low ($S_{J+1} = 0.9$) and assuming 8 visits, a test with (φ_1, φ_0) of $(0.75, 1.00)$ would require a sample size of approximately 6200 subjects, when compared to a sample size of approximately 8000 subjects for a test with (φ_1, φ_0) of $(1.00, 0.75)$ [Figure 1 (c)]. 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 result in a significant number of false positive results - thus, the combined 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 0.55 and 0.99, respectively - however, since the cumulative incidence of diabetes over a 8-year follow-up period until 2005 [33] is relatively low within the WHI ($\approx 10\%$), the low sensitivity of self-reports is not expected to have a drastic impact on statistical power.

4. Study Design

In this section, we address several issues that arise during design of biomedical investigations, namely: (1) What is the minimum sample size required to achieve a desired level of statistical power?; (2) What are the effects of missing test results and/or loss to follow-up? and (3) How to incorporate adaptive testing paradigms into study design? In all the examples, we assume that the primary goal of the study is to use a two-sided test, to test $H_0: \beta = 0$ corresponding to a single binary exposure variable, Z . We assume equal allocation to both groups and that the type I error and power are fixed at 0.05 and 0.90, respectively. As described in Section 3, we determine values of $(S_2, \dots, S_{J+1})$ by assuming an exponential distribution for the time to event within each group ($k = 0, 1$). The parameter λ_0 governing the distribution of X in the reference group ($Z = 0$) is determined by the cumulative incidence at the end of the study (at 8 years), $1 - S_{J+1}$. As noted earlier, the exponential distribution is not an assumption underlying the proposed methods - it is used here to set values of the parameters $(S_2, \dots, S_{J+1})$. In Sections 3.1, 3.2 and 3.3 of the Supplemental Material, we discuss: (1) The effects of misclassification at study entry; (2) The trade-off in costs considering test characteristics (φ_1, φ_0) versus number of subjects enrolled; and (3) The effects of designs that incorporate different testing schedules for different subsets of subjects.

4.1. Factors important in study design: sensitivity, specificity, cumulative incidence and frequency of tests (visits)

The desired sample size typically depends on the specified statistical power and minimum clinically meaningful difference between groups (HR) as well as the characteristics of the diagnostic test (sensitivity, specificity), cumulative incidence in each group, the testing frequency and associated schedule. Figure 2 illustrates the effects of varying HR, S_{J+1} and frequency of tests (visits) on resulting sample size estimates for diagnostic tests with varying levels of sensitivity and specificity. For each setting, we compare the perfect diagnostic test with (ϕ_1, ϕ_0) 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) (0.55, 0.99), which corresponds the characteristics of self-reports of incident diabetes in the WHI [2].

Figure 2(a) presents how sample size varies with HR, by assuming $S_{J+1} = 0.90$ and $J = 4$ - as expected, the required sample size decreases with increasing values of HR. 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 a small effect on sample size estimates for this setting - there is only a modest difference between the perfect test and the (0.75, 1.00) test. These observations are driven by the low cumulative incidence in the reference group - in this case, decreasing specificity results in a corresponding large number of false positive test results which in turn results in a larger sample size in order to detect differences in hazards between the two exposure groups.

Figure 2(b) presents the dependence of sample size on cumulative incidence in the reference group, $1 - S_{J+1}$, assuming $HR = 1.25$ and $J = 4$. 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 positives and low sensitivity results in high rate of false negatives, the impact of imperfect sensitivity and specificity on sample size depends on the true positive proportion (i.e. cumulative incidence). As discussed earlier, when cumulative incidence is low, specificity plays a dominant role. 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 (0.55, 0.99) cross at a cumulative incidence rate of approximately 0.25 - when cumulative incidence is lower than 0.25, the required sample size for the test with imperfect specificity [(1.00, 0.75)] is larger than the test with imperfect sensitivity [(0.55, 0.99)]. However, this trend is reversed when cumulative incidence is larger than 0.25, in which case the test with imperfect sensitivity [(0.55, 0.99)] 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 is modest, it may be optimal to

select a test with high specificity. Since a higher cumulative incidence can decrease sample size, one practical option may be to increase the follow-up assuming that it is feasible and cost-effective to do so.

Figure 2(c) presents how the required sample size changes with the number of tests, J , assuming $HR = 1.25$ and $S_{J+1} = 0.90$. 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 functions of the two groups being compared. In practice, when diagnostic tests are imperfect, increasing the frequency of testing can reduce the required sample size up to a limit.

4.2. Censoring and missed visits

To illustrate the impact of censoring and missed visits on power and sample size, we fixed values of sensitivity and specificity of 0.55 and 0.99, respectively, $HR = 1.25$, $S_{J+1} = 0.50$ and assumed that there are eight equally spaced test times (visits) during the study period. We illustrate the effects of missed visits and censoring (loss to follow-up) by considering designs implementing two distinct testing paradigms: (1) Tests or visits are carried out according to a ‘predetermined’ schedule; and (2) Tests or visits are conducted according to pre-determined schedule up to the time of the first positive test result - after which, further testing ceases (‘NTFP’).

Missed visits—To incorporate the occurrence of missed visits, we assume the following specific to each study design: (1) ‘Pre-determined’ schedule: Each test has an independent, constant probability of being missing, denoted by p_{miss} . In other words, the missing

mechanism is MCAR and thus, $p(I_i|\Phi) = \prod_{j=1}^J p(I_{ij}|\Phi) = (1-p_{miss})^{m_i} (p_{miss})^{J-m_i}$, where m_i is the number of non-missing tests; (2) Adaptive (NTFP) schedule: we assume that tests can be missing independently with probability, p_{miss} - however, after the first positive test result, all subsequent tests are missing. Thus, $p(I_i | \mathbf{R}_{i,obs}, \mathbf{Z}_i, \Phi) = (1 - p_{miss})^{m_i} (p_{miss})^{J-m_i}$, where the first positive test is observed at the J_i^{th} visit. If no positive result is observed, then $J_i = J$. In this setting, we assume that there is no censoring or loss to follow-up.

In Figure 3(a), we present the relative sample size (sample size ratio) to achieve at least a power of 0.90, as a function of p_{miss} . Relative sample size or sample size ratio is defined as the ratio of the sample size under a specific non-zero value of p_{miss} to the sample size when $p_{miss} = 0$. As expected, the sample size ratio increases with p_{miss} for both the pre-determined and adaptive (NTFP) testing paradigms, indicating that increasing probabilities of missing test results results in decreased power. In addition, while the absolute sample sizes for the pre-determined testing schedule are lower when compared to the adaptive (NTFP) schedule, the sample size ratio for a specific non-zero value of p_{miss} under the pre-determined schedule is larger than that for NTFP - indicating that missing visits result in a larger

decrease in power when the schedule is pre-determined when compared to NTFP. For example, when $p_{miss} = 0.5$, the sample size ratio is 1.38 for the pre-determined schedule, when compared to 1.28 under NTFP.

Censoring or loss to follow-up—Let $p_{censor} = P(X_c \in (\tau_j, \tau_{j+1}])$ - in other words, p_{censor} denotes the probability of being censored between any two consecutive visits. For example, if $p_{censor} = 0.1$, then over the course of eight scheduled visits, each subject has a probability of 0.80 (0.1×8) of being censored during the study (or equivalently, $P(X_c = \tau_j) = 0.8$). As described above, we evaluate the effect of censoring in the context of the two testing paradigms, namely the pre-determined testing schedule versus the adaptive (NTFP) testing schedule. In Figure 3(b), we present the effects of varying values of the probability being censored [$P(X_c = \tau_j)$] on the sample size ratio to achieve a power of 0.90, for the pre-determined and adaptive (NTFP) testing paradigms. Relative sample size or sample size ratio is defined as the ratio of the sample size under a specific non-zero value of $P(X_c = \tau_j)$ to the sample size when $P(X_c = \tau_j) = 0$. Under the pre-determined testing paradigm, all scheduled tests are carried out until either the subject is lost to follow-up or the study is terminated. Under the adaptive (NTFP) setting, all pre-scheduled tests are carried out until the earlier of the following - observation of the first positive result, loss to follow-up or termination of study. In this setting, we assume that there is no random probability of missed visits (i.e. $p_{miss} = 0$). For values of $P(X_c = \tau_j) \leq 0.8$, the effect of $P(X_c = \tau_j) > 0$ on resulting sample size does not differ significantly by testing paradigm, pre-determined or NTFP. For example, when $P(X_c = \tau_j) = 0.4$, the sample size to achieve power of 0.90 is 30% larger when compared to the setting of no loss to follow-up, for both the pre-determined and adaptive (NTFP) testing paradigms.

4.3. Adaptive versus pre-determined testing schedules

To compare the sample sizes required between the predetermined and adaptive (NTFP) testing paradigms, we set $HR = 1.25$, $J = 8$, $S_{J+1} = 0.9$ or 0.5 and $\phi_1 = 0.55$ or 1.0 . We further assumed no missed-visits or censoring. In Figure 4, we present the relative sample size (or sample size ratio), comparing the NTFP design to the pre-determined testing schedule as a function of varying specificity (ϕ_0), for fixed values of sensitivity (ϕ_1) and S_{J+1} . The sample size ratio is calculated as the ratio of the sample size required to achieve power of at least 0.90 for the NTFP testing paradigm to the corresponding sample size required for the pre-determined testing paradigm.

As expected, when $\phi_0 = 1$, the relative sample size is 1, indicating that the sample size required for NTFP is equal to that for the pre-determined testing schedule - in this case, NTFP is optimal as further testing after the observation of the first positive test result provides no additional information. When cumulative incidence is low ($S_{J+1} = 0.9$), relative sample sizes increase rapidly even with modest decreases in specificity - for example, when $S_{J+1} = 0.9$, $\phi_1 = 1$ and $\phi_0 = 0.95$, the sample size for NTFP is 6 times that for the pre-determined testing paradigm. When specificity is low for the NTFP design, early false positive test results prevent subsequent testing, thus leading to a significant loss in information. When cumulative incidence is high ($S_{J+1} = 0.5$), decreasing specificity results in more modest increases in relative sample size. In addition, the difference in relative

sample sizes comparing tests with perfect to imperfect sensitivity is most pronounced when specificity is also low - for example, when $S_{J+1} = 0.9$ and $\varphi_0 = 0.70$, the relative sample size comparing NTFP to the pre-determined testing paradigm is 20 when $\varphi_1 = 1$; however, when $\varphi_1 = 0.55$, the relative sample size increases to approximately 50. This trend is also observed for settings of high cumulative incidence ($S_{J+1} = 0.5$) and low specificity (φ_0).

In summary, when specificity is less than perfect and cumulative incidence is low, NTFP can lead to substantial loss in power. For settings of high cumulative incidence, and when specificity is slightly less than perfect, sample sizes for MCAR and NTFP may be comparable - however, since using NTFP requires fewer tests per subject, it may lead to a more cost-effective study design. The effects of decreased sensitivity is negligible when specificity is high (> 0.90) and cumulative incidence is high.

4.4. Interplay of various factors affecting study design

As in all experiments, power and sample size calculations are influenced by the magnitude of the clinical difference of interest (HR) and type I error of the hypothesis test. However, statistical power (and associated sample size) in studies incorporating error-prone diagnostic tests or self-reports is also affected by a variety of other factors including the sensitivity and specificity of the diagnostic test (self-report), cumulative incidence of the event during the course of the study, frequency and timing of test times (visits), the testing paradigm (pre-determined versus adaptive) and the possibility of misclassification at study entry.

As expected, power increases when HR and/or cumulative incidence of the event increases. The effects of imperfect sensitivity and specificity are dependent on the cumulative incidence of the event of interest during the period of follow-up. In settings of high cumulative incidence, a decrease in sensitivity and/or specificity can result in a significant increase in sample size (for a given power). In many biomedical studies, the event of interest is a rare outcome - in these settings, a decrease in sensitivity can have a negligible effect on the resulting sample size; however, low specificity can have a significant, adverse impact on power. The paradigm used to determine the testing schedule can also impact power and sample size calculations. Here we compared two testing paradigms - namely, pre-determined versus adaptive (NTFP). The adaptive (NTFP) paradigm is optimal in settings in which the specificity of the diagnostic test and cumulative incidence are high. The use of the adaptive (NTFP) testing paradigm in settings of low cumulative incidence combined with a test with reduced specificity can have a significant, adverse effect on power. As expected, missing visits (at random), censoring and the use of an imperfect diagnostic test at entry can each adversely effect power and sample size. Lastly, an increase in the number of visits (tests) can offset a smaller sample size up to a threshold. In settings of low cumulative incidence and low specificity, an increase in the number of tests can reduce the required sample size - however, this effect reaches a plateau after 8 or more tests during the study period.

5. Discussion

In this paper, we have presented methods to estimate sample size and power applicable in studies evaluating the association of one or more covariates with a non-recurrent, time to event outcome that is observed using an error-prone diagnostic test or self-report. The

methods developed in this paper are motivated by self-reported outcomes of Type II diabetes in the WHI. An important special case of the proposed methods is when the diagnostic test is perfect - resulting in interval-censored time to event outcomes.

In our development, we address several issues that arise frequently during design of biomedical investigations, namely: (1) What is the minimum sample size required to achieve a desired level of statistical power?; (2) What are the effects of missing test results and/or loss to follow-up? and (3) How to incorporate adaptive testing paradigms into study design? In the Supplemental Material, we also discuss: (1) The effects of misclassification at study entry; (2) The trade-off in costs considering test characteristics (ϕ_1 , ϕ_0) versus number of subjects enrolled; and (3) The effects of designs that incorporate different testing schedules for different subsets of subjects. The methods illustrated in this paper can be readily implemented using our freely available R software package *icensmis* [1], which can be downloaded from the Comprehensive R Archive Network (CRAN). The corresponding R code to recreate the plots presented in this paper is included in the Tutorial section of the Supplemental Material.

Our proposed methods can be generalized in several ways. We derived expressions for power and sample size based on the assumption that the sensitivity and specificity of the possibly imperfect diagnostic test is known. In many settings, either one or both these parameters may be unknown. Methods for the estimation of power and sample size to account for estimation of the characteristics of diagnostic tests would be an important extension of the methods presented here. In other studies, a battery of diagnostic tests with varying sensitivity and specificity may be used, including settings 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Contract/grant sponsor: This work was supported by the National Institutes of Health 1R01HL122241-01A1 to R.B.

References

1. Gu, X.; Balasubramanian, R. *icensmis*: Study Design and Data Analysis in the Presence of Error-Prone Diagnostic Tests and Self-Reported Outcomes. 2016. <http://CRAN.R-project.org/package=icensmis>, r package version 1.3.1

2. Margolis KL, Qi LH, Brzyski R, Bonds DE, Howard BV, Kempainen S, Liu SM, Robinson JG, Safford MM, Tinker LT, et al. Validity of diabetes self-reports in the women's health initiative: comparison with medication inventories and fasting glucose measurements. *Clinical Trials*. 2008; 5:240–247. [PubMed: 18559413]
3. He CY, Zhang CL, Hunter DJ, Hankinson SE, Louis GMB, Hediger ML, Hu FB. Age at menarche and risk of type 2 diabetes: Results from 2 large prospective cohort studies. *American Journal of Epidemiology*. 2010; 171:334–344. [PubMed: 20026580]
4. Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, Willett WC. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine*. 2001; 345:790–797. [PubMed: 11556298]
5. Oksanen T, Kivimäki M, Pentti J, Virtanen M, Klaukka T, Vahtera J. Self-report as an indicator of incident disease. *Annals of Epidemiology*. 2010; 20:547–554. [PubMed: 20538198]
6. Dunn D, Brandt CD, Krivine A, Cassol PSA, Roques Borkowsky W, De Rossi A, Denamur E, Ehrnst A, Loveday C. The sensitivity of HIV-1 DNA polymerase chain reaction in the neonatal period and the relative contributions of intra-uterine and intra-partum transmission. *AIDS*. 1995; 9:F7–11. [PubMed: 8527070]
7. Kohane I. Using electronic medical records to drive discovery in disease genomics. *Nature Reviews*. 2011; 12:417–428.
8. Sinnot J, Dai W, Liao K, Shaw S, Ananthakrishnan AN, Gainer G, Karlson EW, Churchill S, Szolovits P, Murphy S, et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics*. 2014; 133:1369–82. [PubMed: 25062868]
9. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Cai T. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*. 2015; 22:993–1000. [PubMed: 25929596]
10. Turnbull BW. Empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B-Methodological*. 1976; 38:290–295.
11. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics*. 1986; 42:845–854. [PubMed: 3814726]
12. Gu X, Ma Y, Balasubramanian R. Semi-parametric time to event models in the presence of error-prone, self-reported outcomes - with application to the women's health initiative. *Annals of Applied Statistics*. 2015; 9(2):714–730. [PubMed: 26834908]
13. Balasubramanian R, Lagakos SW. Estimation of the timing of perinatal transmission of HIV. *Biometrics*. 2001; 57:1048–1058. [PubMed: 11764243]
14. Balasubramanian R, Lagakos SW. Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika*. 2003; 90:171–182.
15. McKeown K, Jewell NP. Misclassification of current status data. *Lifetime Data Analysis*. 2010; 16:215–230. [PubMed: 20157848]
16. Meier AS, Richardson BA, Hughes JP. Discrete proportional hazards models for mismeasured outcomes. *Biometrics*. 2003; 59:947–954. [PubMed: 14969473]
17. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*. 1999; 86:843–855.
18. Chen HH, Duffy SW, 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*. 1996; 45:307–317.
19. Satten GA, Longini IM. 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*. 1996; 45:275–295.
20. Guihenneuc-Jouyaux C, Richardson S, Longini IM. Modeling markers of disease progression by a hidden markov process: Application to characterizing CD4 cell decline. *Biometrics*. 2000; 56:733–741. [PubMed: 10985209]

21. Jackson CH, Sharples LD. Hidden markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*. 2002; 21:113–128. [PubMed: 11782054]
22. Kirby, A.; Spiegelhalter, D. Statistical modelling for the precursors of cervical cancer. Wiley; New York: 1994.
23. Snapinn SM. Survival analysis with uncertain endpoints. *Biometrics*. 1998; 54:209–218. [PubMed: 9574966]
24. Cook TD. Adjusting survival analysis for the presence of unadjudicated study events. *Controlled Clinical Trials*. 2000; 21:208–222. [PubMed: 10822119]
25. Cook TD, Kosorok MR. Analysis of time-to-event data with incomplete event adjudication. *Journal of the American Statistical Association*. 2004; 99:1140–1152.
26. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society Series D-the Statistician*. 2003; 52:193–209.
27. McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Statistics in Medicine*. 2002; 21(23):3609–3625. [PubMed: 12436459]
28. McNamee R. Optimal design and efficiency of two-phase case-control control studies with error-prone and error-free exposure measures. *Biostatistics*. 2005; 6(4):590–603. [PubMed: 15860543]
29. Cox, DR.; Hinkley, DV. Theoretical Statistics. Chapman and Hall; 1979.
30. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
31. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*. 1995; 90:1112–1121.
32. Hogan JW, Roy J, Korkontzelou C. Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Statistics in Medicine*. 2004; 23:1455–1497. [PubMed: 15116353]
33. Culver AL, Ockene IS, Balasubramanian R, Olendzki BC, Sepavich DM, Wactawski-Wende J, Manson JE, Qiao YX, Liu SM, Merriam PA, et al. Statin use and risk of diabetes mellitus in postmenopausal women in the women's health initiative. *Archives of Internal Medicine*. 2012; 172(2):144–152. [PubMed: 22231607]

Figure 1. Comparison of power versus total sample size (N) for different values of (sensitivity, specificity)

We set $HR = 1.25$, $S_{J+1} = 0.90$ or 0.50 , $J = 2$ or 8 and assume no missed visits and/or censoring.

Figure 2. Effects of hazard ratio (HR), cumulative incidence ($1 - S_{J+1}$), and number of tests (J) with respect to sample size (N), for different values of (sensitivity, specificity)

a) Sample size as a function of HR, assuming $S_{J+1} = 0.9$ and $J = 4$ equally spaced test times (visits). (b) Sample size as a function of cumulative incidence in the reference group ($1 - S_{J+1}$), assuming $HR = 1.25$ and $J = 4$ equally spaced test times (visits). (c) Sample size as a function of the number of equally spaced test times, J , assuming $HR = 1.25$ and $S_{J+1} = 0.9$.

Figure 3. Effects of censoring and missed visits on sample size

The results are based on assuming sensitivity and specificity of 0.55 and 0.99, respectively, $HR = 1.25$, $S_{J+1} = 0.50$ and $J = 8$ equally spaced visits during the study period. (a) Sample size ratio comparing $p_{miss} > 0$ to $p_{miss} = 0$ is shown as a function of p_{miss} , under the pre-determined and adaptive (NTFP) study designs. (b) Sample size ratio comparing $p(X_c = \tau_j) > 0$ to $p(X_c = \tau_j) = 0$ is shown as a function of the censoring probability $[p(X_c = \tau_j)]$, under the pre-determined and adaptive (NTFP) study designs.

Figure 4. Relative efficiency of adaptive (NTFP) versus the pre-determined testing paradigms
 Relative efficiency or sample size ratio is calculated as the ratio of sample size for NTFP to that for the pre-determined testing paradigm and is shown as a function of specificity of the imperfect diagnostic test. The results are based on assuming $HR = 1.25$, $S_{J+1} = 0.9$ or 0.5 , sensitivity of 1.00 or 0.55 and $J = 8$ equally spaced test times (visits).