UW Biostatistics Working Paper Series

1-24-2003

Estimating Disease Prevalence in Two-Phase Studies

Todd A. Alonzo
University of Southern California, talonzo@childrensoncologygroup.org

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

1 Introduction

Estimation of disease prevalence is often important for public policy reasons. When ascertainment of disease status is costly, for example in-depth interviews for determining depression, or invasive, such as biopsy for diagnosing cancer, a two-phase or double sampling design (Neyman, 1938; Tenenbein, 1970) is often employed. In the first phase, a non-invasive and inexpensive screening test is given to all subjects in the study. Disease status is then verified for a subset of subjects in the second phase of the study, with the selection of subjects for phase 2 dependent on the measurements made in phase 1. An example of such a study is the Great Smoky Mountain Study (Costelo et al., 1996) which sought to estimate the prevalence of adolescent depression in the southeastern United States. All subjects in this study were given an inexpensive screening questionnaire, and then screen positive subjects were oversampled to receive an in-depth interview that determined depression. An analysis of data from this study is provided in Section 5.

This problem also arises in contexts other than survey sampling. Consider a clinical trial to compare treatment arms in regards to the occurrence of a dichotomous outcome or event. An easily measured surrogate outcome measure may be available soon after treatment. The surrogate could be used to select subjects on which to measure the true outcome if the latter is difficult to ascertain. Alternatively, if the true outcome requires a follow-up period, it may simply be unavailable for a subset of patients at the time of analysis. In either case, estimation of the probability of a favorable outcome event (prevalence) is of interest and the study can be conjured as a two-phase study with the surrogate or auxiliary outcome assuming the role of the "screening test" (Pepe et al., 1994).

In studies of new diagnostic tests, estimation of true and false positive rates relative to a gold standard are sought. When the gold standard is difficult to measure, selection of subjects for ascertainment of the gold standard diagnosis often depends on the result of the new test (Begg & Greenes, 1983). Estimation of disease prevalence, as measured by the gold standard, is required in order to quantify the error rates associated with the new test. Such estimation is necessarily based on data from a two-phase study, with the new test playing the role of the screen.

Naive estimates of prevalence can be biased in two-phase studies. In the context of medical diagnostic testing this is known as verification bias (Begg & Greenes, 1983) or work-up bias (Ransohoff & Feinstein, 1978). Formulated as a missing data problem, two main approaches used to obtain valid prevalence estimators are re-weighting (Horvitz & Thompson, 1951) and imputation (Clayton et al., 1998, Roberts et al., 1987). Imputation estimators require a model for the probability of disease conditional on the screening test results. In this paper, we show that model misspecification can lead to biased prevalence estimates. Conversely, the classic re-weighting approach requires estimates of verification probabilities. Since verification probabilities are often under the control of the investigator, the disadvantage of the re-weighting estimator is not lack of robustness to model misspecification, but rather poor efficiency. In this paper, we derive a semiparametric efficient estimator that is consistent if either the disease or verification model is correct and is generally more efficient than the classic re-weighting estimator.

In Section 2, we describe and contrast a variety of commonly used prevalence estimators for two-phase studies, including imputation and re-weighting estimators. By formulating estimators in a common context, a unified approach to the development of asymptotic theory is facilitated (Section 3). In Section 4, we describe extensive simulation studies. Results pertaining to relative efficiencies and small sample inference are presented. The methods are applied to data from the Great Smoky Mountain Study in Section 5. We end with some recommendations and a discussion.

2 Prevalence Estimators

Consider a two-phase study to estimate disease prevalence, $\theta = P(D = 1)$. Let T denote variables collected in phase 1 that are used to select subjects for phase 2. The notation A is used for additional variables that may be predictive for D but that are not used to select subjects for verification of D. Therefore, the n_V subjects with disease verification, which we refer to as the verification sample, have data $(D, T, A, V = 1)$. The other $n - n_V$ subjects have data $(T, A, V = 0)$. Although T and A are assumed to be continuous, the methods that follow are applicable to discrete data as well.

We make the standard missing at random (MAR) assumptions (Little & Rubin, 1987) $P(D|V, T, A) = P(D|T, A)$ or equivalently $P(V|D, T, A) = P(V|T, A)$. That is, although the disease process affects T and A, it only affects selection for disease verification through its influence on T and A. By definition, A can be dropped from the conditioning arguments in the latter equality. We next consider prevalence estimators that are based on two-phase study data. For ease of presentation, they are referenced throughout using acronyms summarized in Table 1.

2.1 Imputation Estimators

In this section three imputation estimators are considered. The first estimator, the naive complete case (CC) estimator, only includes data from subjects with complete data, i.e.

$$\widehat{\theta}_{\mathrm{CC}} = \frac{1}{n_V} \sum_{i=1}^{n} V_i D_i, \qquad (1)$$

where i indexes the $1, \cdots, n$ subjects. The CC estimator imputes observed disease probabilities for those verified. It is well known that $\widehat{\theta}_{\mathrm{CC}}$ is guaranteed to be unbiased only when the verification sample is a simple random sample of subjects in the study. If, for example, those at higher risk of disease are more likely to be selected for verification, then $\widehat{\theta}_{\mathrm{CC}}$ will be biased upwards.

The second estimator imputes disease values for all subjects (Roberts et al., 1987). Specifically, the full imputation (FI) estimator is

$$\widehat{\theta}_{\mathrm{FI}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\rho}_i \tag{2}$$

where $\widehat{\rho}_i$ is an estimator of $\rho_i = P(D_i = 1|T_i, A_i)$ which can be obtained from a parametric model, e.g. logistic regression, applied to the verification sample. The estimator is not consistent if the parametric model is misspecified.

The third imputation estimator uses the score contributions for those in the verification sample and imputes mean scores for those not in the verification sample over the distribution $P(D|T, A)$. Pepe et al. (1994) and Reilly & Pepe (1995) proposed estimating θ by solving $\sum_{i=1}^{n}\{V_i\ S_\theta(D_i) + (1 - V_i)\ \widehat{E}[S_\theta(D)|T_i, A_i]\} = 0$, where $S_\theta(D)$ is the binomial score, $\widehat{E}[S_\theta(D)|T_i, A_i] = \int S_\theta(D)\ \widehat{P}(D|T_i, A_i)dD$ and $\widehat{P}(D|T, A)$ is estimated non-parametrically using the verification sample. Clayton et al. (1998) proposed using a parametric model for $P(D|T, A)$ to accommodate settings where T and A are continuous. For the binomial score $S_\theta(D_i) = D_i - \theta$, solving the above estimating equation yields the mean score imputation (MSI) estimator of disease prevalence:

$$\widehat{\theta}_{\mathrm{MSI}} = \frac{1}{n} \sum_{i=1}^{n} \{V_i\ D_i\ +\ (1 - V_i)\ \widehat{\rho}_i\}. \tag{3}$$

Similar to FI, this estimator will yield asymptotically biased results if $\widehat{\rho}_i$ is biased.

All three imputation estimators belong to the class of imputation estimators. Estimators in this class aim to estimate θ by imputing ρ_i for various subsets of subjects. Suppose that complete data were available and consider the estimating function $S^C(k) = \sum_{i=1}^n \{kD_i + (1-k)\rho_i - \theta\}$ where k is a constant giving weight to the observed versus predicted true disease status. Clearly, setting $S^C(k) = 0$ yields a consistent prevalence estimator, $n^{-1}\sum_{i=1}^n \{kD_i + (1-k)\rho_i\}$, when ρ_i is estimated consistently. For a two-phase study, we take the expectation of S^C conditional on the observed data, which we denote by W. The ith contribution of $E[S^C(k)|W]$ is $kD_i + (1-k)\rho_i - \theta$ when $V_i = 1$ and $k\rho_i + (1-k)\rho_i - \theta$ when $V_i = 0$ and we solve the equation $E[S^C(k)|W] = 0$ for θ where

$$E[S^C(k)|W] = \sum_{i=1}^n \{V_i[kD_i + (1-k)\rho_i - \theta] + (1-V_i)[\rho_i - \theta]\}. \tag{4}$$

Values of k result in familiar estimating functions. If $k = 0$, then (4) equals the estimating function for the FI estimator. Conversely, if $k = 1$, then (4) equals the estimating function for the MSI estimator. In addition, if $k = 1$ and ρ_i is estimated using the observed disease probability for those verified, then (4) corresponds to the estimating function for the CC estimator.

2.2 Re-weighting Estimators

The classic re-weighting estimator (Horvitz & Thompson, 1951) weights each observation in the verification sample by the inverse of the sampling fraction, $\pi_i = P(V_i = 1|T_i, A_i)$. It is easy to show that $E\left[V_i\pi_i^{-1}S_\theta(D_i)\right] = 0$ under MAR. This suggests estimating θ by solving

$$\sum_{i=1}^n V_i\pi_i^{-1}S_\theta(D_i) = 0. \tag{5}$$

For $S_\theta(D_i) = D_i - \theta$, this yields the well-known Horvitz-Thompson inverse probability weighting

(IPW) estimator of prevalence for sample surveys

$$\widehat{\theta}_{\text{IPW}} = \left(\sum_{i=1}^{n} \pi_i^{-1} V_i \right)^{-1} \sum_{i=1}^{n} V_i \pi_i^{-1} D_i. \tag{6}$$

Sampling fractions may be known or may need to be estimated depending on the study design. Estimating π_i, even when known, may increase efficiency (Pepe et al., 1994).

Classes of semiparametric estimators based on augmentations to the IPW estimating equations have been proposed for conditional mean models, i.e. the conditional expectation of the response variable given covariates is known up to a finite dimensional parameter, with missing outcomes (Robins et al., 1995; Rotnitzky & Robins, 1995; Robins & Rotnitzky, 1995; Rotnitzky et al., 1998). Next we derive a class of semiparametric estimators of disease prevalence as a special case of the classes proposed by Robins, Rotnitzky, and colleagues.

In the setting when there are covariates of interest, the re-weighted score contribution for the IPW estimator, $\pi_i^{-1} V_i S_\theta(D_i)$, can be rewritten as $\frac{V_i}{\pi_i} X_i \epsilon_i(\theta)$, where $\epsilon_i = (D_i - \theta)$ is the residual and X is a design matrix containing the covariates of interest. Robins et al. (1995) modify this by replacing X_i with $h(X_i)$ and subtracting the term $\frac{V_i - \pi_i}{\pi_i} \phi(X_i, T_i, A_i)$, where $h(\cdot)$ and $\phi(\cdot)$ are unspecified functions. Thus, the following augmented estimating equation is used to define their class of semiparametric estimators

$$\sum_{i=1}^{n} \left\{ \frac{V_i}{\pi_i} h(X_i) \epsilon_i(\theta) - \frac{V_i - \pi_i}{\pi_i} \phi(X_i, T_i, A_i) \right\} = 0. \tag{7}$$

This class of estimators is semiparametric because it requires parametric conditional mean models to be specified for $P(D|T, A)$ and $P(V|T, A)$ but is non-parametric with respect to the joint distribution of the data $P(D, T, A)$.

To derive the most efficient estimator in this class of estimators, Robins et al. (1995) note that for a fixed $h(\cdot)$ the asymptotic variance of $\widehat{\theta}$ is minimized at $\phi(X_i, T_i, A_i) = E[h(X_i)\epsilon_i(\theta)|X_i, T_i, A_i]$. Even for fixed $h(\cdot)$ this expectation may be intractable, and the optimal choice of $h(\cdot)$ typically requires a complicated adaptive selection process. However, in our setting, only estimating the overall disease prevalence is of interest. Thus, since there are no covariates of interest, the design matrix, X, only contains a vector of 1's and $h(X)$ is equal to a constant which, without loss of generality, will be assumed to equal 1. It can then be shown that the optimal choice of $\phi(X_i, T_i, A_i)$ is $\rho_i - \theta$. Therefore, solving $\sum_{i=1}^{n} \left\{ \frac{V_i}{\pi_i}(D_i - \theta) - \frac{V_i - \pi_i}{\pi_i}(\rho_i - \theta) \right\} = 0$ yields the most efficient estimator in this class

$$\widehat{\theta}_{\text{SPE}} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \left\{ V_i \frac{D_i}{\pi_i} - \left(\frac{V_i - \pi_i}{\pi_i} \right) \widehat{\rho}_i \right\}. \tag{8}$$

We refer to this estimator as the semiparametric efficient estimator (SPE).

Since the IPW estimator is a member of this class of estimators for which SPE is the most efficient estimator, SPE is more efficient than IPW. Not only is SPE the most efficient estimator in the class defined by (7), it attains the semiparametric variance bound (Robins et al., 1995; Begun et al., 1983). That is, any more efficient estimator must be inconsistent under some misspecification of $E[D|T, A]$. This implies that FI and MSI may be more efficient than SPE because, as we have already noted, they are inconsistent under misspecification of $E[D|T, A]$, and hence are outside of the class of semiparametric estimators.

If π_i is estimated consistently, then the first and second terms in (8) converge to the true prevalence and zero, respectively. Moreover, by writing (8) as $\frac{1}{n}\sum_{i=1}^{n}\{V_i(D_i - \widehat{\rho}_i)/\pi_i + \widehat{\rho}_i\}$ it is clear that the SPE estimator converges to the true prevalence if ρ_i is estimated consistently. Thus, the SPE estimator has the attractive attribute that it is "doubly robust" in the sense that it is consistent if either π_i or ρ_i is estimated consistently. That is, misspecification of either π_i or ρ_i (but not both) permits consistency of $\widehat{\theta}_{\mathrm{SPE}}$.

2.3 Qualitative comparison of estimators

Observation-specific contributions to each of the disease prevalence estimators are given in Table 1. The FI estimator (2) imputes disease status for all subjects in the study. In contrast, the MSI estimator (3) imputes disease status only for those subjects not in the verification sample and uses the observed disease status for those in the verification sample. The IPW estimator (6) is similar to the CC estimator (1) in that it uses the observed disease status for the verification sample. Unlike the CC estimator, however, it corrects for the biased sampling by weighting the observed values by the probability of selection for verification. Similar to FI and MSI, the SPE estimator (8) estimates the probability of disease for those not verified. For those subjects in the verification sample, it weights the observed disease status by the probability of being selected for disease verification just like the IPW estimator, but also subtracts the product of the probability of being diseased and the odds of not being verified. The term subtracted compensates for systematic or random differences between V_i and its predicted value π_i.

The CC, MSI, SPE, and IPW estimators are identical when all subjects are verified, i.e., $\pi_i = 1$ for all i. Since all the estimators only require a regression model to be fit for a binary response (D or V), all the approaches are easy to implement, even as the dimension of A increases. CC yields consistent estimates only if the verification sample is a simple random sample. FI and MSI are consistent only if $\widehat{\rho}_i$ consistently estimates ρ_i. IPW is unbiased if $\widehat{\pi}_i$ is consistent. Conversely, SPE is consistent if either $\widehat{\rho}_i$ or $\widehat{\pi}_i$ is consistent. Since π_i is often under the control of the investigators designing the study, there are situations when it is known that $\widehat{\pi}_i$ is consistent; thus, resulting in consistent IPW and SPE estimators.

3 Asymptotic Distribution Theory

All the prevalence estimators discussed can be derived as solutions to estimating equations of the same form. Therefore, we develop asymptotic distribution theory in a unified framework.

3.1 Notation

Let $\beta = (\theta, \alpha)^T$ where α is a vector of nuisance parameters. The parameters α pertain to the models for $P(D = 1|T, A)$ and/or $P(V = 1|T, A)$, as appropriate for the prevalence estimator. For example, the FI and MSI estimators only involve $P(D = 1|T, A)$ as the nuisance function and hence α pertains to parameters in that function only. We assume that $\widehat{\alpha}$ is the solution to a classic estimating equation of the form $U_n^\alpha(\beta) = \sum_{i=1}^n U_i^\alpha(\beta) = 0$ where $U_i^\alpha(\beta)$ satisfy some conditions (below). For example, for the FI and MSI estimators these are standard binary regression estimating equations for the model $P(D = 1|T, A)$ based on the verification sample. The simultaneous estimation of θ and α implies that $\widehat{\beta}$ is the solution of an estimating equation of the form $U_n(\beta) = \sum_{i=1}^n U_i(\beta) = 0$ where $U_i(\beta) = \begin{pmatrix} U_i^\theta(\beta) \\ U_i^\alpha(\beta) \end{pmatrix}$ and the first component $U_i^\theta(\beta)$ is the estimating function component corresponding to $\widehat{\theta}$, as defined in Section 2. Denote $\beta_0^\star = (\theta_0^\star, \alpha_0^\star)$ to be a value of the parameter vector β and let $N_\delta(\beta_0^\star)$ be a bounded δ-neighborhood of $\beta_0^\star$. Further denote $\beta_0 = (\theta_0, \alpha_0)$ to be the true value of the parameter vector β.

3.2 Asymptotic results for solutions to estimating equations

Assuming that the data (D_i, T_i, A_i, V_i) and, hence, $U_i(\beta_0^\star)$ are independent and identically distributed (iid), $E[U_i(\beta_0^\star)] = 0$, $E[\frac{\partial}{\partial\beta}U_i(\beta_0^\star)]$ is negative definite, and the conditions

C1: $\beta_0^\star$ exists

C2: Elements of $U_n(\beta)$, $\frac{\partial}{\partial\beta}U_n(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_n(\beta)$ exist in $N_\delta(\beta_0^\star)$

C3: $U_i(\beta)$, $\frac{\partial}{\partial\beta}U_i(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_i(\beta)$ are uniformly bounded in $N_\delta(\beta_0^\star)$

we obtain the following consistency and asymptotic distribution theory results

Theorem 1 *(consistency) A unique solution $\widehat{\beta}$ to $U_n(\beta) = 0$ exists with probability converging to 1 as n converges to ∞ and $\widehat{\beta}$ converges in probability to $\beta_0^\star$.*

Theorem 2 *(asymptotic normality) $n^{\frac{1}{2}}(\widehat{\beta} - \beta_0^\star)$ converges in distribution to a mean 0 normally distributed random variable with variance-covariance matrix $\not\sum = E\left[\frac{\partial}{\partial\beta}U_i(\beta_0^\star)\right]^{-1} Cov(U_i(\beta_0^\star))E\left[\frac{\partial}{\partial\beta}U_i(\beta_0^\star)\right]^{-1}$.*

Theorem 3 *(consistent variance estimator) A consistent estimator of $\not\sum$ is*

$$\widehat{\not\sum} \;=\; n\left[\sum_{i=1}^{n}\frac{\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1}\left[\sum_{i=1}^{n}U_i(\widehat{\beta})U_i(\widehat{\beta})^T\right]\left[\sum_{i=1}^{n}\frac{\partial}{\partial\beta}U_i(\widehat{\beta})\right]^{-1}.$$

If generalized linear models, for example logistic or probit regression, with bounded covariates are used to estimate the nuisance parameters, then the existence of $\beta_0^\star$ (C1) is satisfied (Wedderburn, 1976) and the corresponding components, $U_i^\alpha(\beta)$, of the estimating functions will satisfy the conditions C2 and C3. In the Appendix, we show that under minimal assumptions they also hold for the estimating function components $U_i^\theta(\beta)$ corresponding to the class of semiparametric prevalence estimators. It can be shown in a similar manner that the estimating function components $U_i^\theta(\beta)$ corresponding to the class of imputation estimators satisfy the conditions.

Our main interest is in $\widehat{\theta}$, the disease prevalence estimator. Theorems 1 and 2 along with the observation that, in our setting, θ is a scalar, $E\left[\frac{\partial}{\partial\theta}U_i^\alpha(\beta)\right] = 0$, and $E\left[\frac{\partial}{\partial\theta}U_i^\theta(\beta)\right] = -1$ imply that

Corollary 1 $n^{\frac{1}{2}}(\widehat{\theta} - \theta_0^\star) \rightarrow_d N\left(0, \sigma_\theta^2\right)$, *where* $\sigma_\theta^2 = Var(U_i^\theta(\beta)) \; + \; \vartheta_{\theta\alpha}\, Cov(\widehat{\alpha})\, \vartheta_{\theta\alpha}^T \; -$ $2\, Cov(U_i^\theta(\beta), U_i^\alpha(\beta))\, \vartheta_\alpha^{-1}\, \vartheta_{\theta\alpha}^T$, $\vartheta_{\theta\alpha} = -E\left[\frac{\partial}{\partial\alpha}U_i^\theta(\beta)\right]$, $\vartheta_\alpha = -E\left[\frac{\partial}{\partial\alpha}U_i^\alpha(\beta)\right]$, *and* $Cov(\widehat{\alpha}) = \vartheta_\alpha^{-1} Cov(U_i^\alpha(\alpha))(\vartheta_\alpha^{-1})^T$.

Provided the nuisance estimating functions $U_i^\alpha(\beta)$ are unbiased, $\theta_0^\star = \theta_0$ so that Corollary 1 implies that $\widehat{\theta}$ for FI, MSI, and IPW are consistent for the true disease prevalence, θ_0. For the SPE prevalence estimator, $\theta_0^\star = \theta_0$ and thus the estimator is consistent provided either the nuisance estimating function for the disease model or the verification model is correctly specified.

3.3 Insights into relative performance

We observe from Corollary 1 that σ_θ^2, the asymptotic variance of the prevalence estimator, is comprised of three components. The first component is the variability of $\widehat{\theta}$ if α were known while the last two components take into account the sampling variability in $\widehat{\alpha}$. When α are known, the variance of an estimator in the imputation class of estimators is $Var(\rho_i) \; + \; k^2 E[\pi_i\rho_i(1-\rho_i)]$. Thus,

$$Var\left(\widehat{\theta}_{\text{FI}}\right) \quad = \quad Var(\rho_i) \tag{9}$$

$$Var\left(\widehat{\theta}_{\text{MSI}}\right) \quad = \quad Var(\rho_i) \; + \; E[\pi_i\rho_i(1-\rho_i)]. \tag{10}$$

Since FI is the MLE in a parametric model, it is not surprising that FI is the most efficient estimator in this class of estimators when α are known. That is, if the disease probabilities are known, then it is more efficient to use them instead of the observed D_i.

Furthermore, it can be shown that

$$Var\left(\widehat{\theta}_{\text{SPE}}\right) = Var(\rho_i) + E[\rho_i(1-\rho_i)/\pi_i] \tag{11}$$

$$Var\left(\widehat{\theta}_{\text{IPW}}\right) = Var(\rho_i) + E[\rho_i(1-\rho_i\pi_i)/\pi_i] \tag{12}$$

Since the second terms in equations (11) and (12) are always non-negative, FI is also more efficient than SPE and IPW when α are known. Comparing equations (10)-(12) also suggests that MSI is more efficient than SPE which is more efficient than IPW. The FI and MSI estimators do not belong to the class of semiparametric estimators for which SPE is most efficient because they make stronger assumptions about the disease model. Thus, it is not surprising that FI and MSI are more efficient.

Next, we consider the setting when α are not known. Since $\vartheta_{\theta\alpha} = 0$ for SPE, Corollary 1 yields that the variance of SPE is also (11) when α are not known. Variance expressions for the other prevalence estimators depend on the model used to estimate ρ_i or π_i. Thus, the relative efficiencies of the estimators when α are not known is investigated in the next section using simulation studies.

The above insights into relative performance are valid even when the disease and verification models are misspecified. However, $\widehat{\theta}$ is no longer unbiased when π_i and ρ_i are not estimated consistently so the bias in the estimators also should be considered. The magnitude of the bias relative to the variance will affect the relative performance of estimators as quantified, e.g., by the mean squared error (MSE).

4 Simulation Study

A disease can arise from an underlying continuous disease process which remains subclinical until it reaches some threshold, at which point the disease becomes apparent. We simulate an underlying continuous variable Z with a $N(0, 1)$ distribution and disease status, D, as the indicator variable $D = I[Z > h]$, where h is the threshold. By varying h, we can vary the prevalence of disease, θ.

Frequently there are multiple components to a disease process. Thus, we considered Z to be the sum of Z_1 and Z_2, where each was distributed $N(0, 0.5)$. Since screening tests or other variables associated with disease can relate differently to the different components, we simulated T and A as linear combinations of Z_1 and Z_2 plus random normal error, ϵ_1 and ϵ_2. In particular, $T = \nu_1 Z_1 + \tau_1 Z_2 + \epsilon_1$ and $A = \nu_2 Z_1 + \tau_2 Z_2 + \epsilon_2$, where $\epsilon_1 \sim N(0, 0.25)$ and $\epsilon_2 \sim N(0, 0.25)$ are independent.

By varying ν and τ, different aspects of the disease are weighted differently and the correlation between T and A is varied. The inherent accuracy of T and A for D also can be altered by changing the values of ν and τ. Table 2 summarizes the area under the receiver operating characteristic (ROC) curve for T corresponding to ν_1 and τ_1 between 0 and 1. Values of the area under the ROC curve (AUC) range from 0.5 at $\nu_1 = \tau_1 = 0$ where T does not discriminate between $D = 1$ and $D = 0$, to 0.96 at $\nu_1 = \tau_1 = 1$ indicating almost perfect discrimination.

To induce verification bias, verification status, V, was generated with $P(V = 1)$ as a function of T. Specifically, we consider $P(V = 1|T > t^{(q)}) = 1$, and $P(V = 1|T \leq t^{(q)}) = \delta$, where $t^{(q)}$ denotes the qth quantile of the distribution of T. That is, similar to the protocol in the Great Smoky Mountain Study, all subjects with a test result greater than the threshold were selected for disease verification while only a random fraction below the threshold were selected. Results presented are for

$q = 0.8$ and $\delta = 0.2$. These values resulted in an average of 36% of study subjects receiving disease verification, with 20% and 16% having a value of T above and below the threshold, respectively.

FI, MSI, and SPE require a parametric model for the probability of disease. We chose a probit model linear in T and A because this form is induced under our simulation model. Estimators that included both T and A are indexed by 'B' for 'both'; whereas, models that only included T are indexed by 'T'. IPW and SPE estimators used either known or estimated verification probabilities, indexed by 'K' and 'E' respectively. The empirical estimates used were $\widehat{\pi}_i = 1.0$ if $T_i > t^{(0.8)}$ and $\widehat{\pi}_i = \frac{\sum_{i=1}^{n} V_i \, I[T_i \leq t^{(0.8)}]}{\sum_{i=1}^{n} I[T_i \leq t^{(0.8)}]}$, if $T \leq t^{(0.8)}$.

Results are shown for 1000 simulations of data for n=1000 subjects and $\theta = 0.1$. Since verification is a function of T, the extent of biased sampling was varied by changing the parameters that determine T, ν_1 and τ_1. For larger values of ν_1 and τ_1, the discrepancy in the percentages of diseased and non-diseased subjects verified is larger, and, consequently, the larger the bias in CC, the naive estimator of prevalence (Table 2). FI, MSI, SPE, and IPW were essentially unbiased (not shown).

4.1 Relative efficiency

Mean prevalence variance estimates were similar to the simulation variance and the variance estimator had good coverage probabilities, suggesting that large sample results can be used to perform small sample inference. Both asymptotic relative efficiency (ARE) and small sample relative efficiency (SSRE) were calculated, with the latter calculated as the ratio of the variances for two estimators across the 1000 simulation realizations. Asymptotic variance expressions (Corollary 1) were calculated using Monte Carlo integration. Since ARE and SSRE were similar, only SSRE results are presented.

SSRE relative to FI-B is presented in Table 3 for $A = Z_1 + Z_2 + \epsilon_2$. SSRE less than 1 implies the estimator is less efficient than FI-B. Estimators are listed in the rows while the informativeness of T and potential for verification bias decreases moving across the columns. MSI-B and FI-B are the most efficient estimators and have similar performance to each other. Conversely, the IPW estimator is clearly the most inefficient estimator, 33-44% less efficient than the FI estimator. As anticipated, estimating π_i increased the efficiency of the IPW estimator. IPW with empirically estimated π_i is up to 14% more efficient than IPW using known π_i. Conversely, using estimated π_i does not appear to affect the efficiency of SPE. These observations are consistent with previous literature (Pepe et al., 1994; Robins et al., 1995). Table 3 also suggests that SPE estimators are 8-17% less efficient than FI and MSI estimators that use the same disease probability model. This is not surprising because FI and MSI do not belong to the class of semiparametric estimators (containing SPE) that makes fewer assumptions. Observe, however, that the SPE estimator is up to 37% more efficient than the IPW estimators that belong to the semiparametric class of estimators.

To assess gains in efficiency that can be achieved by incorporating auxiliary data, the informativeness of A was varied from highly informative (i.e., $\nu_2 = \tau_2 = 1$) to non-informative (i.e., $\nu_2 = \tau_2 = 0$). Small sample efficiency of FI-T relative to FI-B that includes A in the disease model is given in Table 4. Moving down a column decreases the informativeness of A and, not surprisingly, increases the relative efficiencies. It is reassuring to observe that estimators that included non-informative A were not less efficient than estimators that did not incorporate these data. We also observe that relative efficiency generally decreases moving across the columns, as the informativeness of T for D decreases. That is, the less informative T is, the more efficiency that can be gained by incorporating

A. Similar results were observed for MSI and SPE.

Qualitative results seemed to generalize for various δ and θ (not shown). As δ increases to 1, the difference in efficiency between FI and MSI estimators and other estimators decreased. This is not surprising since CC, MSI, IPW, and SPE estimators are equivalent when all subjects are verified. Moreover, there was a trend for differences in efficiencies between FI (MSI) and IPW to be augmented by increasing θ. Conversely, the difference in efficiency between FI (MSI) and SPE was smaller for larger θ.

4.2 Robustness to model misspecification

The estimators discussed in this paper require a binary regression model for disease and/or verification probabilities. These models can be misspecified if either the form or components of the model are incorrect. For a generalized linear model, these correspond to an incorrect link function or form of the linear predictor. The latter can occur if the model does not include pertinent explanatory variables or if the form of the explanatory variables included is not correct. Next, we discuss the robustness of the prevalence estimators to these types of model misspecification.

Not surprisingly, simulation results were similar when probit and logit links were used (results not shown). Thus, estimators appear to be robust to these link functions. Next, we investigate the robustness of the estimators when the form of the linear predictor is incorrect. FI, MSI, and SPE require a model for the probability of disease. Since verification is a function of T, the disease model is misspecified if A is included in the model instead of T. In this case, the MAR assumption is violated, $P(D|A, V = 1) \neq P(D|A)$. As expected, SPE yielded unbiased estimates of prevalence in all scenarios considered (Table 5). Conversely, FI and MSI overestimated prevalence except when a simple random sample of subjects was verified. The bias in FI and MSI was smaller for more informative A and considerably less than the bias in CC when A was somewhat informative. The variance of FI and MSI were larger than SPE except when A was highly informative or when a simple random sample was verified. However, the mean squared error (MSE) of FI and MSI was only smaller than the MSE of SPE when a simple random sample was verified. Similar qualitative results were obtained for other forms of linear predictor misspecification.

To investigate the robustness of IPW and SPE to linear predictor misspecification, a model linear in T was fit instead of the correct model containing the threshold $I[T > t^{(0.8)}]$. IPW estimates of prevalence clearly underestimated prevalence when the former model was used (not shown). Conversely, the SPE estimator yielded unbiased estimates with slightly smaller variance than the correctly specified estimator. However, when both disease and verification probability models were misspecified, SPE (denoted SPE-L) had similar bias and variance to FI and MSI estimators (Table 5).

5 Great Smoky Mountain Study

One of the main goals of the Great Smoky Mountain Study (GSMS) was to assess the prevalence of depression in adolescents residing in Western North Carolina (Costelo et al., 1996). The GSMS was a two-phase study where in phase 1 an inexpensive 57-question screening questionnaire, designed to uncover general behavioral and emotional disorders, was given to the parents of all study subjects. All subjects scoring above a pre-determined threshold on the questionnaire were selected for phase 2 while subjects scoring below were selected with probability 0.1. The subjects selected for phase 2 as well as their parents were administered in-depth Child and Adolescent Psychiatric Assessment (CAPA) interviews (Angold et al., 1995). Computer algorithms applied to the results of the

CAPA interviews were used to diagnose depression, as defined by DSM-III-R taxonomy (American Psychiatric Association, 1987).

5.1 Data

Screening questionnaire scores ranged from 0 to 79 with 25th, 50th, and 75th percentiles of 6, 12, and 20, respectively. Of the 3878 subjects for which we had data, 1005 (25.9%) had a questionnaire score higher than the threshold of 20. Due to refusal, loss to follow-up, and other reasons, only 758 (75.4%) of the screen positives and 257 (8.9%) of the screen negatives received depression status ascertainment in phase 2. Overall 1015 (26.2%) received disease verification, of which 36 (3.5%) were diagnosed with depression.

5.2 Estimators

In the notation used in previous sections, T are the continuous results of the screening questionnaire, V is the binary variable indicating whether subjects received disease verification, and D is the binary depression status. No auxiliary information, A, were available.

FI, MSI, and SPE estimators require a model for disease probabilities conditional on T. Generalized Additive Models (GAM) were used as an exploratory tool to suggest transformations of T, if necessary, to be included in the disease model (Hastie & Tibshirani, 1990). SPE and IPW estimators were fit using both the known (theoretical) weights of 1.0 and 0.1 and the empirically estimated weights of 0.754 and 0.089 for the screen positive and screen negative subjects, respectively.

5.3 Results

Table 6 presents estimates of the prevalence of depression for the various estimators. Based on GAM results, T was included as a linear term in the model for $P(D|T)$. Logistic regression and probit regression yielded similar results so only probit regression results are reported. Clearly, the CC estimate of prevalence (3.5%) is substantially larger than the other estimates. In fact, the 95% confidence interval for CC does not contain any of the other prevalence estimates. FI and MSI estimates are the least variable, but are roughly 0.02% higher than the SPE and IPW estimates. SPE and IPW yield similar estimates whether theoretical or estimated weights are used. SPE estimates are marginally less variable than the IPW estimates. Although the SPE, IPW, FI, and MSI prevalence estimates are similar for these data, we recommend reporting results of the SPE estimator because simulations reported in the previous section indicate that SPE is more robust to model misspecification and is only marginally less efficient than FI and MSI.

6 Discussion

A variety of approaches have been taken to deal with missing data in two-phase study designs. We have considered the simplest statistical task in this paper, namely estimation of disease prevalence, and considered the corresponding variety of prevalence estimators. We conclude that the SPE method is the preferred approach. We observed that the SPE estimator is only marginally less efficient than the FI and MSI estimators when the disease model is correct and has the important advantage that it is consistent if either the disease or verification model is correct. Moreover, it is substantially more efficient than the most commonly applied IPW estimator.

The SPE estimator derived in Section 2.2 is similar to the regression estimator proposed by Gao et al. (2000) for two-phase sample surveys. Gao et al. (2000) did not recognize this estimator to be the most efficient estimator in the class of semiparametric estimators and did not develop distribution theory for their estimator, that we have now provided here. An alternative approach to estimating prevalence is an efficient fully parametric analysis that parameterizes the joint distribution of T, A, and D as $P_{\theta,\gamma}(T, A, D) = P_\theta(D)P_\gamma(T, A|D)$ and estimates θ and γ simultaneously using maximum likelihood estimation methods. However, not only is the nuisance component $P_\gamma(T, A|D)$ not of interest, but misspecification of this model can lead to inconsistent estimation of θ (Pepe, 1992). Another important disadvantage to a fully parametric approach is that it is usually difficult to specify a model for $P_\gamma(T, A|D)$, especially as the dimension of A increases (Clayton et al., 1998).

Distribution theory has been developed for the SPE estimator in a more general setting (Robins et al., 1995). Clayton et al. (1998) obtain a variance expression for the MSI prevalence estimator by considering a vector of estimating equations for the simultaneous estimation of the parameter of interest, here disease prevalence, and nuisance parameters. In this paper, we formalized and generalized this approach to obtain theory that is applicable to all the estimators considered in this paper, and that can easily be extended to the clustered data setting.

Although our results pertain only to the estimation of prevalence, they may be informative about more complex analysis such as regression analysis. Complete case, imputation, mean score, and semiparametric methods can also be applied when fitting regression models. Application of semiparametric efficient methods in those settings is much more challenging, however, and they may not be practical. Further work to compare the approaches would be of interest.

The probability subjects are selected for phase 2 of a two-phase study is often under the control of the investigators designing the study. The question arises as how to "optimally" design such a study. Pepe et al. (1994) propose optimal sampling fractions for the mean score estimator with discrete phase 1 data. Identifying optimal two-phase sampling strategies for other estimators of prevalence discussed in this paper and for continuous phase 1 data also warrants further attention.

ACKNOWLEDGMENTS

This research is partially funded by NIH/NCI 5U10 CA13539 and NIH R01 GM54438. We thank Alaatin Erkanli for providing data from the Great Smoky Mountain Study.

APPENDIX

Asymptotic theory stated in Theorems 1-3 is proven in an unpublished technical report that can be accessed at *http://www-rcf.usc.edu/∼talonzo/techrep.html*. These results can be applied to the class of semiparametric estimators provided $\beta_0^\star$ exists, $U_i(\beta_0^\star)$ are iid, $E[U_i(\beta_0^\star)] = 0$, and the estimating functions satisfy conditions C2 and C3 (Section 3.2). Clearly, these are satisfied for $U_n^\alpha(\beta)$ if logistic or probit regression with bounded covariates are used to estimate nuisance parameters. The partial derivatives of $U_n^\alpha(\beta)$ with respect to θ equal 0 because $U_n^\alpha(\beta)$ are not a function of θ. Therefore, C2 and C3 are true for $\frac{\partial}{\partial\theta}U_n^\alpha(\beta)$ and all that remains to be shown is that $U_i^\theta(\beta_0^\star)$ are iid, $E[U_i^\theta(\beta_0^\star)] = 0$, and $U_i^\theta(\beta)$ and $\frac{\partial}{\partial\alpha}U_n^\theta(\beta)$ have properties C2 and C3. Next, we show that these hold under the following assumptions: (A1) D is MAR; (A2) (D_i, T_i, A_i, V_i) are iid; (A3) (T, A) is bounded; (A4) $E[\frac{\partial}{\partial\beta}U_i(\beta_0)]$ is negative definite; (A5) ρ_i and π_i are bounded away from 0.

$U_i^\theta(\beta) = \pi_i^{-1}V_i(D_i - \theta) - k\pi_i^{-1}(V_i - \pi_i)(\rho_i - \theta)$ are iid since each $U_i^\theta(\beta)$ is only a function of the data which are assumed to be iid (A2). Using conditional expectations and (A1), $E[U_i^\theta(\beta_0^\star)]$ equals

$$E\left[E\left[U_i^\theta(\beta_0^\star)\right]|T_i, A_i\right] = E\left[\pi_i^{-1}\left(E[D_i|T_i, A_i]\,\pi_i - \pi\,\theta_0^\star\right)\right] - kE\left[\pi_i^{-1}\,E\left[(V_i - \pi_i)(\rho_i - \theta_0^\star)|T_i, A_i\right]\right] = 0.$$

Expressions for $U_n^\theta(\beta)$, $\frac{\partial}{\partial\beta}U_n^\theta(\beta)$, and $\frac{\partial^2}{\partial\beta\partial\beta^T}U_n^\theta(\beta)$ can be easily derived. Hence existence (C2) is satisfied. Furthermore, these expressions are bounded (C3) by assumptions A3-A5.

References

American Psychiatric Association Committee on Nomenclature and Statistics. (1987). *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition* Washington, DC.

Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E., & Rutter, M. (1995). The child and adolescent psychiatric assessment (CAPA). *Psychological Medicine* **25**, 739–753.

Begg, C. B. & Greenes, R. A. (1983). Assessment of diagnostic tests when disease is subject to selection bias. *Biometrics* **39**, 207–216.

Begun, J. M., Hall, W. J., Huang, W. M., & Wellner, J. A. (1983). Information and asymptotic effiency in parametric-nonparametric models. *Annals of Statistics* **11**, 432–452.

Clayton, D., Dunn, G., Pickles, A., & Spiegelhalter, D. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B* **60**, 71–87.

Costelo, E. J., Angold, A., Burns, B. J., Stangl, D., Tweed, D., Erkanli, A., & Worthman, C. (1996). The Great Smoky Mountains Study of Youth: Prevalence and correlates of DSM-III-R disorders. *Archives of General Psychiatry* **53**, 1129–1136.

Gao, S., Hui, S. L., Hall, K. S., & Hendrie, H. C. (2000). Estimating disease prevalence from two-phase surveys with non-response at the second phase. *Statistics in Medicine* **19**, 2101–2114.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Horvitz, D. G. & Thompson, D. J. (1951). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Little, R. J. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley, New York.

Neyman, J. (1938). Contributions to the theory of sampling of human populations. *Journal of the American Statistical Association* **33**, 101–116.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355–365.

Pepe, M. S., Reilly, M., & Fleming, T. R. (1994). Auxiliary outcome data and the mean-score method. *Journal of Statistical Planning and Inference* **42**, 137–160.

Ransohoff, D. F. & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* **299**, 926–930.

Reilly, M. & Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.

Roberts, G., Rao, J. N. K., & Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**, 1–12.

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rotnitzky, A. & Robins, J. M. (1995). Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics* **22**, 323–333.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association* **65**, 1350–1361.

Wedderburn, R. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27–32.

Table 1: Contributions to prevalence estimation. $\rho_i = P(D_i = 1|T_i, A_i)$ and $\pi_i = P(V_i = 1|T_i, A_i)$.

Estimator (Acronym)	$V_i = 1$	$V_i = 0$
Complete case (CC)	D_i	0
Full imputation (FI)	ρ_i	ρ_i
Mean score imputation (MSI)	D_i	ρ_i
Inverse probability weighting (IPW)	$D_i\pi_i^{-1}$	0
Semiparametric efficient (SPE)	$\pi_i^{-1}[D_i - \rho_i(1 - \pi_i)]$	ρ_i

Table 2: Empirical AUC, $P(V = 1|D = 1)$, $P(V = 1|D = 0)$, and bias in the CC estimator when verification depends on the value of $T = \nu_1 Z_1 + \tau_1 Z_2 + \epsilon_1$.

$\nu_1,\ \tau_1$	AUC	$P(V = 1\|D = 1)$	$P(V = 1\|D = 0)$	Bias of CC
1, 1	0.96	0.92	0.30	160%
1, 0.5	0.92	0.83	0.31	130%
0.5, 0.5	0.88	0.76	0.32	110%
1, 0	0.81	0.66	0.33	80%
0.5, 0	0.72	0.56	0.34	60%
0, 0	0.50	0.36	0.36	0%

Table 3: Small sample efficiency relative to FI-B for estimating disease prevalence when verification depends on $T = \nu_1 Z_1 + \tau_1 Z_2 + \epsilon_1$. A is fixed to be $Z_1 + Z_2 + \epsilon_2$. Estimators indexed with "B" and "T" use the disease probability models $P(D|T, A)$ and $P(D|T)$, respectively. SPE and IPW use known or estimated verification probabilities, indexed by 'K' and 'E' respectively.

	$\nu_1,\ \tau_1$			
Estimator	1, 1	0.5, 0.5	1, 0	0, 0
FI-T	0.97	0.84	0.81	0.63
MSI-B	0.99	0.99	1.00	1.00
MSI-T	0.96	0.83	0.81	0.63
IPW-K	0.67	0.59	0.64	0.56
IPW-E	0.81	0.64	0.67	0.56
SPE-K-B	0.87	0.83	0.88	0.92
SPE-K-T	0.83	0.67	0.69	0.55
SPE-E-B	0.87	0.83	0.87	0.91
SPE-E-T	0.83	0.67	0.69	0.55

Table 4: Small sample efficiency for FI-T relative to FI-B for estimating disease prevalence when verification depends on $T = \nu_1 Z_1 + \tau_1 Z_2 + \epsilon_1$.

	$\nu_1,\ \tau_1$			
$\nu_2,\ \tau_2$	1, 1	0.5, 0.5	1, 0	0, 0
1, 1	0.968	0.837	0.815	0.625
0.5, 1	0.984	0.901	0.833	0.741
0.5, 0.5	0.990	0.932	0.924	0.810
0, 1	0.995	0.967	0.866	0.892
0, 0	0.998	0.999	1.002	0.999

Table 5: Mean (variance $\times 10^{-4}$) estimated prevalence when verification depends on $T = Z_1 + Z_2 + \epsilon_1$. True prevalence is 0.1. Different values of ν_2 and τ_2 are considered for $A = \nu_2 Z_1 + \tau_2 Z_2 + \epsilon_2$. FI, MSI, and SPE incorrectly use the disease probability model $P(D|A)$. SPE-L incorrectly fits the verification probability model $P(V|T)$ linear in T.

	$\nu_2,\ \tau_2$			
Estimator	1, 1	0.5, 0.5	1, 0	0, 0
FI	0.13 (1.48)	0.18 (2.82)	0.20 (3.68)	0.26 (5.31)
MSI	0.13 (1.49)	0.18 (2.83)	0.20 (3.69)	0.26 (5.31)
SPE-E	0.10 (1.34)	0.10 (1.69)	0.10 (1.74)	0.10 (1.28)
SPE-L	0.13 (1.49)	0.18 (2.80)	0.20 (3.67)	0.26 (5.31)
CC	0.26 (5.28)	0.26 (5.28)	0.26 (5.28)	0.26 (5.28)

Table 6: Estimates of depression prevalence in the Great Smoky Mountain Study.

Estimator	$\widehat{\text{Prevalence}}$	Standard Deviation	95% CI
CC	0.036	0.362	(0.024, 0.047)
FI	0.019	0.253	(0.011, 0.026)
MSI	0.019	0.251	(0.011, 0.026)
SPE-E	0.017	0.280	(0.008, 0.026)
SPE-K	0.017	0.272	(0.008, 0.025)
IPW-E	0.017	0.286	(0.007, 0.025)
IPW-K	0.016	0.281	(0.007, 0.025)