NIH Public Access
Author Manuscript
Stat Med. Author manuscript; available in PMC 2010 March 12.

Published in final edited form as:
Stat Med. 2000 August 30; 19(16): 2101–2114.

Estimating disease prevalence from two-phase surveys with non-response at the second phase

Sujuan Gao[1,*,†], **Siu L. Hui**[1,3], **Kathleen S. Hall**[2], and **Hugh C. Hendrie**[2]

[1] Division of Biostatistics, Department of Medicine, Indiana University School of Medicine, Indianapolis, U.S.A

[2] Department of Psychiatry, Indiana University School of Medicine, Indianapolis, U.S.A

[3] Regenstrief Institute for Health Care, Indianapolis, U.S.A

SUMMARY

In this paper we compare several methods for estimating population disease prevalence from data collected by two-phase sampling when there is non-response at the second phase. The traditional weighting type estimator requires the missing completely at random assumption and may yield biased estimates if the assumption does not hold. We review two approaches and propose one new approach to adjust for non-response assuming that the non-response depends on a set of covariates collected at the first phase: an adjusted weighting type estimator using estimated response probability from a response model; a modelling type estimator using predicted disease probability from a disease model; and a regression type estimator combining the adjusted weighting type estimator and the modelling type estimator. These estimators are illustrated using data from an Alzheimer's disease study in two populations. Simulation results are presented to investigate the performances of the proposed estimators under various situations.

1. INTRODUCTION

Two-phase sampling designs are often used in epidemiological studies where a disease is rare and diagnosis of the disease is expensive or difficult [1]. In the first phase of the study a large random sample from the targeted population is screened with less intensive and expensive screening test for the disease. Based on the results of the screening tests, subjects are stratified and randomly selected within each stratum for extensive clinical evaluations at the second phase to determine disease status. The sampling plans are usually designed to identify as many diseased subjects as possible for risk factor studies and at the same time allow efficient estimation of disease prevalence for the population. The two-phase sampling design has been used to estimate the prevalence rates of dementia and Alzheimer's disease [2], heart disease [3] and sexually transmitted disease [4].

Two types of estimator have been used in estimating population percentages from complex survey data. The first is the so-called weighting type estimator, or the standardization estimator. This estimator is widely applicable to a variety of sampling designs [5] and has been used in various medical studies for prevalence estimation [4]. The second estimator is the modelling type estimator applicable when there is auxiliary information in addition to the main outcome variable. For binary outcome variables, smoothed means from logistic regression models are

*Correspondence to: Sujuan Gao, Division of Biostatistics, Department of Medicine, Indiana University School of Medicine, 1050 Wishard Blvd, RG4101, Indianapolis, IN 46202-2872, U.S.A.
†sgao@iupui.edu

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

often used to estimate population percentages. The modelling type estimator utilizes additional information and is especially advantageous over the weighting type estimator when the event of interest is rare and estimation for small strata is desired.

Most large scale epidemiological studies with the two-phase sampling design encounter non-response at both phases. Since we do not have sufficient information on the non-response occurring at the first phase, correction for biases from this type of non-response is not possible without additional information on the non-respondents. The non-response that occurs at the second phase of the study is sometimes perceived as more problematic because these subjects are screened at the first phase and they constitute part of the sampling frame. Correcting for biases due to the second phase non-response may also be possible because there is information on the non-respondents from the screening phase. The reasons for non-response are varied, with refusal being the most common. Death and severe sickness are also major causes for non-response in studies involving elderly subjects if there is considerable time lapse between the two phases. The non-response from the second phase complicates statistical analysis of complex survey data because most of the conventional analysis procedures assume random sampling within strata. When there is non-response, randomization assumptions for stratified random sampling may be violated. A ‘naive’ approach ignoring the non-response may lead to biased and inconsistent estimates [6].

In this paper we first review the two types of estimator of population percentages without non-response. We then review and propose methods to obtain adequate disease prevalence estimates in the presence of non-response assuming that non-response depends on a set of covariates collected at the first phase of the study (the covariate-dependent missing data mechanism as defined by Little and Rubin [6]). The estimators are illustrated using data from the Indianapolis–Ibadan Dementia Study. A simulation study is presented in Section 5 to investigate the effects of adjustment and model misspecifications. We conclude the paper in Section 6.

2. ESTIMATING DISEASE PREVALENCE FROM TWO-PHASE SAMPLING WITHOUT NON-RESPONSE

Suppose that in the first phase N subjects are sampled by simple random sampling from the target population and information is collected from all N subjects on a set of characteristics X. X can be a vector containing several predictors, such as age, gender etc. that relate to the disease of interest. The N subjects are then stratified into S strata, labelled as $\mathrm{I}_1,\dots,\mathrm{I}_S$, based on values of X. The total numbers of subjects in the respective strata are denoted by $N_1,\dots,N_S$. In the second phase n_s subjects are sampled from the N_s subjects in stratum s using stratified random sampling. Disease status y_{si} is ascertained on the ith subject from the sth stratum, with $y_{si}=1$ denoting disease and $y_{si}=0$ for non-disease. We are interested in estimating the prevalence of disease in the population from which the N subjects are sampled.

2.1. The weighting type estimator

The weighting type estimator, also referred to as the direct standardization approach, assumes that subjects within each stratum are homogeneous and random sampling is used within stratum. The weighting type estimator for a stratified random sampling is

$$\widehat{p}_{\mathrm{wt}}=\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{n_s}\frac{N_s}{n_s}y_{si} \quad (1)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

This estimator is design unbiased under repeated samplings. The estimator can be inefficient when the event (disease) is rare and it does not utilize any additional information available from the first phase in a two-phase survey.

2.2. The modelling type estimator

An alternative method of estimating disease prevalence from two-phase surveys is the modelling type estimator where a model is assumed for the superpopulation from which the finite population is sampled and smoothed estimates from the model are used to estimate disease prevalence. The modelling type estimator for binary data was first proposed by Roberts *et al.* [8] and used by Beckett *et al.* [2] to estimate the prevalence of Alzheimer's disease from two-phase surveys. The modelling type estimator is preferred in situations where the disease is rare and estimates from strata containing few or zero events are desired.

Let X_{si} be a set of covariates collected at the first phase. Therefore, X_{si} is available for all N subjects. Let Prob(y_{si} = 1) = p_{si}. A logistic regression model is often assumed for the disease model:

$$\log \frac{p_{si}}{1-p_{si}} = X_{si}\beta \tag{2}$$

where β is a $p\times 1$ vector of parameter. If β is known, then the average of the predicted probability of disease from the model is an unbiased estimator of disease prevalence. However, in practice one has to estimate β from the sample. A maximum likelihood estimate $\hat{\beta}$ is obtained and estimate of disease prevalence is then obtained using the average predicted probabilities of disease:

$$\widehat{p}_{\text{model}} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N_s}\frac{1}{1+\mathrm{e}^{-X_{si}\widehat{\beta}}} \tag{3}$$

Let p be the true disease prevalence rate. Using a second-order Taylor series expansion:

$$\begin{aligned}\mathrm{E}(\widehat{p}_{\text{model}}) = {} & p + \frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N_s}\frac{1}{(1+\mathrm{e}^{-X_{si}\beta})^2}X'_{si}E(\widehat{\beta}-\beta)X_{si}\\ & +\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N_s}\frac{X'_{si}X_{si}}{(1+\mathrm{e}^{-X_{si}\beta})^3}X'_{si}\mathrm{var}(\widehat{\beta})X_{si}\end{aligned}$$

where the expectation and variance are under the superpopulation. It can be seen that the modelling type estimator using the estimated $\hat{\beta}$ is generally biased.

3. ESTIMATING PREVALENCE WITH NON-RESPONSE

Owing to non-response, the numbers clinically evaluated from the S strata are $r_1,\ldots, r_S$. The layout is summarized in Table I. Here we assume that $r_s<n_s$, for some stratum s, that is, there is non-response in the study.

Following Särdal and Swensson [9], we propose that two-phase samplings with non-response be conceptualized as three-phase samplings where the third phase consists of Bernoulli samplings from an unknown response mechanism. Therefore, the sampling scheme can be conceptualized as:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Phase I: Simple random sampling of N subjects from the target population and stratifications.

Phase II: Stratified random sampling of $n_1,\ldots, n_S$ subjects from strata $1,\ldots, S$, respectively.

Phase III: Bernoulli sampling, the ith sampled subject in the sth stratum has probability θ_{si} of responding.

With the establishment of the above sampling framework, estimators of disease prevalence can be derived similarly as in the two-phase sampling case taking into account one more level of randomization. Two different modelling approaches can be taken, one to model the response probability and the other to model the disease.

Let R_{si} be a binary random variable of response defined for the n sampled subjects, with R_{si} = 1 for responses and R_{si} = 0 for non-response, $s = 1,\ldots, S$, and $i = 1,\ldots, n_s$.

3.1. The adjusted weighting type estimator

An unadjusted weighting approach, often used in practice, is to ignore the column of sampled numbers in Table I. Prevalence estimates are then obtained using

$$\widehat{p}_{\text{unadjusted}}=\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{r_s}\frac{N_s}{r_s}y_{si}$$

This approach may result in biased prevalence estimates if the non-response within stratum is not missing completely at random (MCAR) as defined by Little and Rubin [6]. Furthermore, the variance of prevalence estimates may also be inaccurate using this approach.

Let $\Pr(R_{si} = 1) = \theta_{si}$. A Horvitz–Thompson type estimator adjusting for non-response is

$$\widehat{p}_{\text{wtadj}}=\frac{1}{N}\sum_{s=1}^{S}\frac{N_s}{n_s}\sum_{i=1}^{r_s}\frac{y_{si}}{\theta_{si}} \tag{4}$$

Under the assumption that R_{si} and y_{si} are independent, we have

$$\begin{aligned}&E(\widehat{p}_{\text{wtadj}})\\ &\quad=E_I E_{II} E_{III}\left(\sum_{s=1}^{S}\sum_{i=1}^{n_s}\frac{1}{N}\frac{N_s}{n_s}\frac{y_{si}R_{si}}{\theta_{si}}\right)\\ &\quad=E_I\left(\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N_s}y_{si}\right)=p\end{aligned} \tag{5}$$

where E_I, E_{II} and E_{III} are expectations taken over the three corresponding sampling phases and p is the population disease prevalence. Therefore, the adjusted estimator of disease prevalence is unbiased under a known missing data mechanism. In practice, the response mechanism is unknown. We can substitute an estimated probability of non-response $\hat{\theta}_{si}$ for θ_{si} in $\hat{p}_{\text{wtadj}}$ and the properties of the adjusted estimator will then depend on the modelling of the non-response mechanism.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

To obtain the estimated probability of response, we adopt the notion of a response model. Following the ‘model-based’ approach by Rosenbaum [7], we model the probability of responding to the clinical evaluation from the data and then incorporate the estimated probabilities into the prevalence estimator in (5). This approach has been advocated by Robins *et al.* [10] to obtain efficient parameter estimates from a wide range of models with missing data.

To build a response model, we can use a set of covariates obtained at the first phase of the study, say X, to model the probability of responding to the second phase. The most commonly used model for the response mechanism is the logistic regression model where responses are dichotomized into two categories of respondents and non-respondents. The logistic model obtains parameter estimate α from the following model:

$$\log \frac{\text{Prob}(R_{si}=1)}{\text{Prob}(R_{si}=0)} = X_{si}\alpha \tag{6}$$

The estimated response probability

$$\widehat{\theta}_{si} = \frac{1}{1+\mathrm{e}^{-X_{si}\widehat{\alpha}}}$$

can then be used in the place of θ_{si} in $\hat{p}_{\text{wtadj}}$ of (5).

Most survey literature on weight adjustment have mainly focused on classifying the samples into ‘homogeneous response groups’ according to demographic factors collected in the surveys. The above weighting adjustment is the most general and is applicable in many community based surveys where sample sizes are not large enough to afford further stratification.

3.2. The modelling type estimator

Estimation approach using maximum likelihood principle in the case of non-response has to take into account the joint likelihood of three random variables for disease, sampling and response, instead of just the disease and sampling variables in the case of complete data. We demonstrate that under certain assumptions maximum likelihood estimates from the disease model using data from respondents only are appropriate. Let y_{si} and R_{si} be defined as in the previous section. In addition let I_{si} be the sampling indicator variable for the ith subject in the sth stratum. Let $X_{si}^{(y)}, X_{si}^{(I)}$ and $X_{si}^{(R)}$ be the covariates for the conditional distributions of $f(y_{si}|X_{si}^{(y)};\beta)$, $f(I_{si}|y_{si}, X_{si}^{(I)};\varphi)$ and $f(R_{si}|I_{si}, y_{si}, X_{si}^{(R)};\theta)$, respectively, where β, φ and θ are parameters in the corresponding distributions. Let X_{si} be the union of $X_{si}^{(y)}, X_{si}^{(I)}$ and $X_{si}^{(R)}$. The joint conditional likelihood function based on all data can be written as

$$\begin{aligned} f(y_{si}, I_{si}, R_{si}|X_{si};\beta \\ ,\varphi,\theta) = f(y_{si}|X_{si};\beta) f(I_{si}|y_{si} \\ ,X_{si};\varphi) f(R_{si}|I_{si} \\ ,y_{si}, X_{si};\theta) \end{aligned} \tag{7}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Let y_{obs} and y_{miss} denote the observed and the missing y_{si}. Note here y_{miss} includes those non-sampled subjects in addition to the non-respondents. If the following assumptions hold:

- **i.** the second-phase sampling scheme is completely determined by X_{si};
- **ii.** the response probability is independent of sampling and disease status conditional on X_{si};
- **iii.** the parameters β, φ and θ are distinct,

following Little and Rubin [6], the joint conditional likelihood based on the respondent data can be rewritten as

$$\begin{aligned} f(y_{\text{obs}}, I, R|X;\beta,\varphi,\theta) &= \int f(y_{\text{obs}}, y_{\text{miss}}|X;\beta) f(I|y,X;\varphi) f(R|I,y,X;\theta)\,\mathrm{d}y_{\text{miss}} \\ &= f(I|X;\varphi) f(R|X;\theta) \int f(y_{\text{obs}}, y_{\text{miss}}|X;\beta)\,\mathrm{d}y_{\text{miss}} \\ &= f(I|X;\varphi) f(R|X;\theta) f(y_{\text{obs}}|X;\beta) \\ &= f(I|X^{(I)};\varphi) f(R|X^{(R)};\theta) f(y_{\text{obs}}|X^{(y)};\beta) \end{aligned} \tag{8}$$

The above equation implies that under the above assumptions maximum likelihood estimates of β obtained using just the respondents' data are the same as those obtained from the full joint likelihood. Hence prevalence estimates using the modelling type estimator are not affected by this type of non-response.

Note that the first assumption is typically true in two-phase studies where stratifications and sampling are based on the information collected at the first screening phase. The second assumption on missing data is a special type of the missing at random assumption of Little and Rubin [6] where missingness depends on covariates but not on outcome variables. Note also that the above conclusion is reached assuming that the correct conditional distributional assumption is made on $f(y|X;\beta)$. When the assumptions are violated, biased and/or inconsistent estimates may be expected.

Note also that the validity of (8) under the assumptions holds for the general situations where a distribution is assumed for the variables and where estimation of β is desired. Hence the conclusion can be extended from logistic models to other models such as linear regression models for continuous outcome variables.

3.3. A regression type estimator

The adjusted weighting type estimator and modelling type estimator differ in motivation and approach. The adjusted weighting type estimator attempts to model $R|X$ without any restriction on the outcome variable. The sampled units contribute to the estimation of prevalence by the modelling of response. The respondents contribute directly to prevalence estimates and the rest of the subjects only contribute to the weighting. The modelling type estimator, on the other hand, uses the respondents to derive the disease model, but every subject contributes to the prevalence estimates through the set of covariates in the prediction model. Therefore, there may be efficiency differences in the two types of estimator. Furthermore, both types of estimator depend on the correct specification of either a disease model or a response model. Although we have demonstrated in Section 3.2 that the derivation of maximum likelihood estimate of β can be done ignoring the sampling scheme under the conditions outlined there, tests of model fitting will nevertheless have to take into account of the sampling scheme [8], a task not handled by many standard statistical software.

To guard against model misspecification, we propose a third estimator analogous to the regression estimator in survey sampling [5]. Suppose that the disease model (2) holds in the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

study population and response probability is θ_{si} for the ith subject in the sth stratum, then a regression type estimator of disease prevalence is

$$\begin{aligned}\widehat{p}_{\text{reg}}=&\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N_s}\widehat{p}_{si}\\&+\frac{1}{N}\sum_{s=1}^{S}\frac{N_s}{n_s}\sum_{i=1}^{r_s}\frac{y_{si}-\widehat{p}_{si}}{\widehat{\theta}_{si}}\\&=\widehat{p}_{\text{model}}+\widehat{p}_{\text{wtadj}}\\&-\frac{1}{N}\sum_{s=1}^{S}\frac{N_s}{n_s}\sum_{i=1}^{r_s}\frac{\widehat{p}_{si}}{\widehat{\theta}_{si}}\end{aligned}\tag{9}$$

When both disease model and response model are known, the above estimator is unbiased. When the disease model is true, but the response model may not be, the above estimator can be rewritten as

$$\widehat{p}_{\text{reg}}=\widehat{p}_{\text{wtadj}}+\left(\frac{1}{N}\sum_{s=1}^{S}\sum_{i=1}^{N_s}p_{si}-\frac{1}{N}\sum_{s=1}^{S}\frac{N_s}{n_s}\sum_{i=1}^{r_s}\frac{p_{si}}{\widehat{\theta}_{si}}\right)$$

The last two terms on the right hand of the above equation is a correction measure of bias due to the use of response model when the average of p_{si} is estimated from the respondents.

On the other hand, when the response model is thought to be true, but not the disease model, the regression type estimator can be rewritten as

$$\widehat{p}_{\text{reg}}=\widehat{p}_{\text{model}}+\left(\frac{1}{N}\sum_{s=1}^{S}\frac{N_s}{n_s}\sum_{i=1}^{r_s}\frac{y_{si}}{\theta_{si}}-\frac{1}{N}\sum_{s=1}^{S}\frac{N_s}{n_s}\sum_{i=1}^{r_s}\frac{\widehat{p}_{si}}{\theta_{si}}\right)$$

The last two terms in the right hand of the above equation is a correction measure of bias due to the use of the disease model in the respondents.

The regression type estimator may be robust against model misspecification for the response model or the disease model, although when both are misspecified, it is less clear whether the regression type estimator will still be robust.

The estimators discussed above all assumed non-response depends on a set of covariates, which is a weaker assumption than the missing completely at random (MCAR) assumption required for the unadjusted weighting estimator. Although estimation approach under the non-ignorable missing data mechanism can be taken by using the joint likelihood of the disease, sampling and response variables, it will be complicated and requires additional assumptions on the joint distribution. We will not pursue the topic of estimation under the non-ignorable missing assumption in this paper.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

3.4. Variance estimation

Under the assumptions in Section 3.2, variance estimator for the modelling type prevalence estimator can be derived for logistic models as in Roberts *et al* [8]. Let

$$\widehat{p}_{si}=\frac{1}{1+\mathrm{e}^{-X_{si}\widehat{\beta}}}$$

be the predicted probability of disease from the disease model (2), then the estimated variance of the estimated prevalence from (3) is

$$\mathrm{var}(\widehat{p})=W'QXVX'Q'W$$

where W is an $N \times 1$ vector with elements equal to $\frac{1}{N}$, Q is an $N \times N$ diagonal matrix with elements $\hat{p}_{si}(1-\hat{p}_{si})$, X is the covariate matrix and V is the estimated variance covariance matrix of the logistic regression parameter β. Note that the above variance estimator assumes that X is fixed. Variance estimators accounting for the variability in X for survey data is given by Graubard and Korn [11].

Variance estimators for the adjusted weighting type estimators can be derived using variance formulae for multi-stage sampling [5]. However, since estimated response probabilities are used for the estimator, the variance estimates of $\hat{p}$ which involves variance estimation from the three phases would be very complicated to derive. A non-parametric approach for variance estimation, the jack-knife estimator of variances, can be used. The jack-knife estimator is useful in situations where estimators are derived through several modelling procedures in complex surveys [12,13]. Suppose that $\hat{p}_{(si)}$ is the estimate obtained by omitting the *i*th clinically diagnosed subject in the *s*th stratum from the sample, and $\hat{p}_{(s.)}$ is the mean of $\hat{p}_{(si)}$ in the *s*th stratum, the jack-knife estimate of the variance of $\hat{p}$ is

$$\mathrm{var}(\widehat{p})=\sum_{s=1}^{S}\frac{r_s-1}{r_s}\sum_{i=1}^{r_s}(\widehat{p}_{(si)}-\widehat{p}_{(s.)})^2 \qquad (10)$$

An advantage of using the jack-knife variance estimator is that no specialized software is required for the variance derivation for complex survey data. An SAS macro is written by the authors to implement the approaches which gives prevalence estimates and standard errors of the prevalence estimates using the jack-knife approach.

4. A TWO-PHASE STUDY FOR ALZHEIMER'S DISEASE IN TWO POPULATIONS

The Indianapolis and Ibadan Study of Health and Aging is an on-going longitudinal study of dementia and Alzheimer's disease in the elderly. The study populations include African Americans and Africans age 65 and older living in Indianapolis, U.S.A. and in Ibadan, Nigeria, respectively. One of the study goals is to estimate prevalence rates of dementia and Alzheimer's disease in the two populations to test if the rates differ. Details of the study have been published elsewhere [14,15]. Population-based two-phase surveys were conducted to estimate the prevalence of dementia and Alzheimer's disease in both populations. At the first phase, 2212 subjects from Indianapolis and 2494 subjects from Ibadan were randomly selected from the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

two communities and administered screening tests aimed at measuring the subject's cognitive functions. Each subject received a cognitive score which ranged from 0 to 33. Based on the screening scores, the subjects were grouped into four strata: good performance and age ≤74; good performance and age ≥75; intermediate performance, and poor performance. The initial sampling plan was to sample 100 per cent from the poor performance group, 50 per cent from the intermediate group, 5 per cent from the good performance group of which 75 per cent should come from those older than 75. However, due to refusal, death, severe sickness and other reasons, the study had to sample more than the prespecified percentages in all strata except the poor performance stratum to achieve the targeted number of total clinical diagnosis. Non-response rates in each of the strata for the two study sites are summarized in Table II.

In Table III we present some of the characteristics of the two samples by their response groups. Prevalence estimates ignoring the non-response indicates that prevalence of Alzheimer's disease is higher in the African American population than in the African population. However, since the non-respondents in Indianapolis have higher cognitive scores than the respondents and higher cognitive scores is correlated with lower prevalence of Alzheimer's disease, prevalence estimates for the two sites without accounting for non-response may be subject to the question that differences in prevalence estimates are perhaps due to differential response rates. Although in this data set the prevalence difference is so large that we do not anticipate that adjusting for the non-response will change the original conclusion, we hope to illustrate the approaches outlined in the previous section and to contrast the difference in estimates.

We derive prevalence estimates of dementia using the approaches of Section 3. Two types of weighting estimators are used: the unadjusted estimator, which ignores the non-response; the adjusted estimator, which models the probability of response by a logistic regression model using age, sex and cognitive scores as covariates. The modelling estimator is obtained using a logistic model for dementia with age, sex and cognitive scores as covariates. The regression type estimator is used with the same response model as in the adjusted weighting estimator and the same disease model as in the modelling estimator. Standard error estimates from the jack-knife variance estimators are also included. Although the standard error estimates for the modelling type estimator can be obtained from the model-based variance estimator, we choose to use the jack-knife variance estimators for all estimators so that differences in standard error estimates could not be attributed to differences in estimation approaches. The prevalence estimates are presented in Table IV.

The prevalence estimates from the adjusted weighting type estimator, the modelling estimator and the regression estimator still indicate significant difference between the two populations. However, it can be noted that the differences in prevalence rates for the two populations estimated by the three methods are smaller than those given by the unadjusted weighting estimates. There seems to be differences between the estimates, especially in the age-specific rates, given by the three types of estimators. The small difference in the overall rates is perhaps due to the fact that in this data set stratification leading to sampling weight captured essentially the same information as in the covariates used in the modelling estimator. Greater discrepancies may be anticipated when this is not the case. A simulation study was conducted to investigate the properties of the two estimators under various missing data assumptions and disease model specifications.

5. A SIMULATION STUDY

Data from the Indianapolis population of the dementia example are used to provide covariate information in the simulation. Specifically, age, sex and cognitive scores from the Indianapolis population are used as covariates. Disease and response are generated as random Bernoulli variables from the prespecified models. With each simulated data set the unadjusted weighting

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

estimator, the adjusted weighting estimator, the modelling type estimator and the regression type estimators are used to derive prevalence estimates. The parameter configuration for the simulation are classified into four basic design groups: the first group includes simulations where both disease and response models are correctly specified; the second group uses correct disease model but misspecified response models; the third group uses correct response model but misspecified disease models; the fourth group uses misspecified disease and response models. Details of the model and design specifications for the simulations are summarized in Table V.

In Table V disease models A, B and C represent increasing true prevalence rates. The disease model A represents prevalence similar to the dementia prevalence rate in our example data. Disease model D includes a second-order age effect in the model. Two true response models are used to generate the response data; response model *a* assumes a binary logistic model, and response model *b* assumes a multinomial logit model with three response categories (respondents, died or too sick and refused).

A total of 2000 simulations were generated for each configuration. True disease prevalence was derived from the underlying model used to generate the disease variable. The mean prevalence estimates and standard deviations from the simulation were used as the estimates derived from the corresponding approaches. Since the computational time on simulation studies evaluating the performances of the jack-knife variance estimator was beyond our resources at this time, simulations were only conducted to evaluate the performances of the prevalence estimators using the approaches discussed in this paper. Results of the simulation are presented in Table VI.

We first comment on the performances of the adjusted weighting type estimator. From the simulation results we can see that the unadjusted weighting estimators overestimate disease prevalence, owing to the fact that in our data the non-respondents have higher cognitive scores. From Table VI we can also see that when the response models are correctly specified (designs 1 and 3), the adjusted weighting type estimators yield prevalence estimates very close to the true prevalence, effectively correcting the biases of the weighted estimator. However, from designs 2 and 4 we can also see that the adjusted weighting type estimator is very sensitive to misspecifications of the response models. The biases due to misspecifications of the response models are small in the case of fitting a binary logistic response model for a true multinomial logit response model. However, the biases are substantial in the cases of fitting models without all the covariates in the true models. Therefore, careful examination on model fittings for the response models is needed before one proceeds with the adjusted weighting type estimator.

The modelling type estimator performs very well in situations where the assumptions in Section 3.2 hold (designs 1 and 2). The modelling type estimator seems fairly robust in omitting a second-order age effect in the disease model, which is in agreement with the observation made in Scott and Wild [16]. However, the modelling type estimator yields large bias when a covariate is omitted from the fitted model (simulation numbers 7 and 8). For both response and disease models the impact of omitting cognitive scores from the models seems to be larger than omitting age, probably due to the fact that cognitive scores are better predictor of dementia than age.

The regression type estimator performs well when either the response model or the disease model is correctly specified, making it the most robust estimator among the three estimators. However, in the cases where both disease and response models are misspecified, the regression type estimator does not seem to perform any better than the two other estimators.

These simulation results are expected properties of the various estimators. The simulations suggest that the modelling type estimator is robust when a reasonable model can be fitted to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the disease data and the non-response are ignorable in the likelihood based inference, because no additional adjustment is needed. The unadjusted weighting type estimator yields large bias even under the MAR mechanism. The biases can be corrected by the method proposed in Section 3.1, provided that the fitted response model is correctly specified. The regression type estimator is robust when either the disease model or the response model is correctly specified.

6. CONCLUSIONS

In this paper we considered methods for estimating disease prevalence from two-phase survey studies with non-response. The unadjusted weighting type estimator may result in large bias if the non-response are not that of missing completely at random. When the non-response depends on covariates collected at the first phase of the study, an adjusted weighting approach may be used. The adjusted weighting type estimator can be especially useful in a multi-purpose survey where many outcomes are of interest so that the adjusted weights can be used in a straightforward way. The modelling type estimator, on the other hand, is dependent on specific outcomes. We demonstrate that under the missing at random assumptions and certain other conditions the modelling type estimator obtained using maximum likelihood estimates from correctly specified disease models on respondents only is unaffected by this type of non-response. The regression type estimator can be more robust under a misspecified response model or a misspecified disease model. The approaches are illustrated using data from the Indianapolis–Ibadan dementia study. Our simulation results indicate that the adjusted weighting type estimator can effectively correct for biases of the unadjusted weighting estimator, but it may be sensitive to misspecification of the response model. The modelling type estimator is fairly robust under the stated assumptions. However, it may also be sensitive to misspecification of the disease model. The regression type estimator may be robust against misspecifications in the disease model or the response model. Nevertheless, it may not be robust for misspecifications in both disease and response models.

Acknowledgments

We thank Professor T. M. F. Smith (University of Southampton, U.K.) for helpful comments on an earlier version of this manuscript. This work was supported in part by NIA grants R01 09956, AG15813, AG04518 and P30 AG10133.

Contract/grant sponsor: NIA; contract/grant numbers: R01 09956, AG15813, AG04518, P30 AG10133

References

1. Pickles A, Dunn G, Vazquez-Barquero JL. Screening for stratification in two-phase ('two-stage') epidemiological surveys. Statistical Methods in Medical Research 1995;4:73–89. [PubMed: 7613639]
2. Beckett LA, Scherr PA, Evans DA. Population prevalence estimates from complex samples. Journal of Clinical Epidemiology 1992;45:393–402. [PubMed: 1569435]
3. Chambless LE, Boyle KE. Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazard models. Communications in Statistics — Theory and Methods 1985;14:1377–1392.
4. Warszawski J, Messiah A, Lellouch J, Meyer L, Deville JC. Estimating means and percentages in a complex sampling surveys: application to a French national survey on sexual behaviour. Statistics in Medicine 1997;16:397–423. [PubMed: 9044529]
5. Cochran, WG. Sampling Techniques. 3. Wiley; New York: 1977.
6. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. Wiley; New York: 1987.
7. Rosenbaum PR. Model based direct adjustment. Journal of the American Statistical Association 1987;82:387–394.
8. Roberts G, Rao JNK, Kumar S. Logistic regression analysis of sample survey data. Biometrika 1987;74:1–12.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

9. Särndal CE, Swensson B. A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. International Statistical Review 1987;55:279–294.

10. Robins JM, Rotintzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 1995;90:106–121.

11. Graubard BI, Korn EL. Predictive margins with survey data. Biometrics 1999;55:652–659. [PubMed: 11318229]

12. Rao JNK, Wu CFJ. Inference from stratified samples: second-order analysis of three methods for non-linear statistics. Journal of the American Statistical Association 1985;80:620–630.

13. Skinner, CJ.; Holt, D.; Smith, TMF., editors. Analysis of Complex Surveys. Wiley; New York: 1989.

14. Hendrie HC, Osuntokun BO, Hall KS, Ogunniyi AO, Hui SL, Unverzagt FW, Gureje O, Rodenberg CA, Baiyewu O, Musick BS, Adeyinka A, Farlow MR, Oluwole SO, Class CA, Komolafe O, Brashear A, Burdine V. Prevalence of Alzheimer's disease in two communities: Migerian Africans and African Americans. American Journal of Psychiatry 1995;152:1485–1492. [PubMed: 7573588]

15. Hall KS, Ogunniyi AO, Hendrie HC, Osuntokun BO, Hui SL, Musick BS, Rodenberg CA, Unverzagt FW, Guerje O, Baiyewu O. A cross-cultural community based study of dementias: methods and performance of the survey instrument Indianapolis, U.S.A. and Ibadan, Nigeria. International Journal of Methods in Psychiatry Research 1996;6:129–142.

16. Scott, AJ.; Wild, CJ. Selection based on the response variable in logistic regression. In: Skinner, CJ.; Holt, D.; Smith, TMF., editors. Analysis of Complex Surveys. Wiley; New York: 1989. p. 191-205.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table I

Sampling scheme and achieved samples for two-phase survey studies with non-response.

Strata	Total number in strata	Number sampled	Number evaluated
I_1	N_1	n_1	r_1
I_2	N_2	n_2	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_s	N_s	n_s	r_s
Total	N	n	r

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table II

Sampling scheme and non-response rates for the Indianapolis–Ibadan study of elderly African Americans and Africans.

Study site	Strata	Total numbers	Sampled	Clinically diagnosed	Demented	Non-response rate
Indianapolis	Total	2212	614	351	65	42.8%
	Good, age 65–74	1133	58	27	0	53.4%
	Good, age 75+	650	138	72	1	47.8%
	Intermediate	167	156	78	5	50.0%
	Poor	262	262	174	59	33.6%
Ibadan	Total	2494	615	423	28	31.2%
	Good, age 65–74	1593	65	35	0	46.2%
	Good, age 75+	427	115	77	2	33.0%
	Intermediate	195	156	97	2	37.8%
	Poor	279	279	214	24	23.3%

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table III

Characteristic of the response groups for the two study populations. Data from the Indianapolis–Ibadan study of health and ageing.

Study site	Variables	Clinically diagnosed	Died or too sick	Refused	*p*-value
Indianapolis		*n*= 351	*n*= 67	*n*= 196	
	Mean age	77.5	78.6	76.6	0.1500
	(SD)	(7.4)	(6.8)	(7.2)	
	Male %	40.5	26.9	30.6	0.019
	Mean cognitive scores (SD)	27.5 (4.9)	28.3 (4.0)	29.1 (2.9)	0.0002
Ibadan		*n*= 423	*n*= 65	*n*= 127	
	Mean age	77.7	82.8	75.4	0.0001
	(SD)	(9.2)	(9.5)	(7.7)	
	Male %	33.1	29.2	26.0	0.296
	Mean cognitive scores (SD)	24.2 (5.2)	21.6 (6.0)	26.2 (4.2)	0.0001

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table IV

Age specific prevalence estimates of dementia for Indianapolis and Ibadan, standard error estimates using the jack-knife variance estimator included in parenthesis.

Study site	Age group	Unadjusted weighting (%)	Adjusted weighting (%)	Modelling type (%)	Regression type (%)
Indianapolis	65–74	1.152 (0.355)	1.101 (0.339)	2.030 (0.415)	1.314 (0.372)
	75–84	9.433 (1.789)	9.413 (1.969)	6.308 (1.000)	9.210 (1.985)
	85+	13.484 (2.852)	13.099 (2.784)	14.605 (2.660)	11.965 (2.708)
	overall	4.908 (0.641)	4.837 (0.687)	4.534 (0.652)	4.795 (0.715)
Ibadan	65–74	0.509 (0.176)	0.548 (0.188)	0.856 (0.229)	0.667 (0.218)
	75–84	4.494 (1.510)	4.737 (1.591)	2.679 (0.567)	4.751 (1.581)
	85+	7.038 (2.836)	7.642 (3.077)	9.787 (2.914)	8.533 (3.258)
	overall	1.861 (0.403)	1.990 (0.430)	2.016 (0.464)	2.159 (0.484)

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table V

Model and design specifications for the simulations.

Design	Simulation number	Disease model		Response model	
		True	Fitted	True	Fitted
1	1	A	correct	*a*	correct
	2	B	correct	*a*	correct
	3	C	correct	*a*	correct
2	4	A	correct	*b*	binary logit
	5	A	correct	*a*	omit cogscore
3	6	D	omit age^2	*a*	correct
	7	A	omit cogscore	*a*	correct
4	8	A	omit cogscore	*a*	omit cogscore
	9	A	omit age	*a*	omit male

True disease models:

A: $\mathrm{logit}(y = 1) = 3.6804 + 0.0685\mathrm{age} - 0.4183\mathrm{cogscore}$

B: $\mathrm{logit}(y = 1) = 5.4 + 0.0685\mathrm{age} - 0.4183\mathrm{cogscore}$

C: $\mathrm{logit}(y = 1) = 6.5 + 0.0685\mathrm{age} - 0.4183\mathrm{cogscore}$

D: $\mathrm{logit}(y = 1) = -41.5199 + 1.1548\mathrm{age} - 0.3248\mathrm{cogscore} - 0.00679\mathrm{age}^2$

True response models:

a: $\mathrm{logit}(R = 1) = 12.0 + 0.5076\mathrm{male} - 0.3870\mathrm{cogscore}$

b: $\Pr(R = 1) = e^{U1}/(1 + e^{U1} + e^{U2})$, where

$U1 = 2.8155 + 0.00928\mathrm{age} - 0.2428\mathrm{male} - 0.1012\mathrm{cogscore}$, and $U2 = -2.1255 + 0.0312\mathrm{age} + 0.0527\mathrm{male} - 0.0485\mathrm{cogscore}$.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table VI

Estimated disease prevalence from 2000 simulations. Sampling standard deviations of the prevalence estimates are included in parenthesis. See Table V for parameter and model specifications.

Design	Simulation number	True prevalence (%)	Unadjusted weighting (%)	Adjusted weighting (%)	Modelling type (%)	Regression type (%)
1	1	4.989	5.790 (1.391)	4.952 (1.145)	4.959 (1.219)	4.946 (1.125)
	2	14.571	16.537 (2.808)	14.557 (2.493)	14.529 (2.567)	14.550 (2.463)
	3	27.263	30.154 (4.105)	27.296 (3.853)	27.257 (3.726)	27.290 (3.623)
2	4	4.989	5.543 (1.343)	4.991 (1.263)	4.979 (1.262)	4.978 (1.212)
	5	4.989	5.790 (1.391)	7.461 (1.344)	4.959 (1.219)	4.965 (1.070)
3	6	6.851	7.778 (1.720)	6.854 (1.524)	6.853 (1.588)	6.854 (1.506)
	7	4.989	5.790 (1.391)	4.952 (1.145)	5.665 (1.396)	4.946 (1.162)
4	8	4.989	5.790 (1.391)	7.461 (1.344)	5.665 (1.396)	6.219 (1.414)
	9	4.989	5.790 (1.391)	4.953 (1.152)	4.957 (1.218)	4.950 (1.238)