STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:4569–4580
Published online 9 July 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3304

Evaluating assumptions of weighting class methods for partial response using a selection model

Philip J. Smith[1,*,†] and Lawrence C. Marsh[2]

[1]*Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases, 1600 Clifton Road, NE, Mail Stop E-32, Atlanta, GA 30333, U.S.A.*
[2]*Department of Economics and Econometrics, University of Notre Dame, IN, U.S.A.*

SUMMARY

In survey sampling, information about the prevalence of a health outcome Y for a defined target population is frequently obtained using a two-stage data collection process. In the first stage, households that have members of the target population are identified and socio-demographic data that are believed to be associated with Y are collected. At the end of the first stage of data collection, permission is requested to contact the member's health providers so that accurate information about Y can be obtained. When permission is obtained, a second phase of data collection is conducted in which those health providers are contacted and Y is obtained. A 'complete response' results when data are obtained from both the first and the second phases of the survey. A 'partial response' results when data are collected from the first phase, but Y is not obtained in the second phase. To adjust for selection bias in estimating the prevalence of Y caused by partial responders' missing Y values, potential differences between complete and partial responders are typically taken into account by using weighting class methods. These methods assume that missing Y values are missing at random (MAR). This paper describes statistical tests for evaluating whether missing data are missing completely at random or MAR. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: missing values; partial response; response propensity; selection bias; weighting class

1. INTRODUCTION

Information about the prevalence $\bar{Y}$ of a health outcome Y in a defined target population is often obtained in two phases of data collection. In the first phase, sampled households are screened to determine whether they have a member belonging to the survey's target population. Sampled

*Correspondence to: Philip J. Smith, Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases, 1600 Clifton Road, NE, Mail Stop E-32, Atlanta, GA 30333, U.S.A.
†E-mail: pzs6@cdc.hhs.gov

Received 9 March 2007
Accepted 12 March 2008

households that have a member of the target population are interviewed to obtain socio-demographic data. First-phase sampling weights are constructed to account for the probability of sampling the member. At the end of the interview, consent is requested to contact the member's healthcare providers. Provided consent is obtained, the second phase of data collection is conducted and the member's healthcare providers are contacted to obtain data on Y. We let $Y=1$ if the member has the health outcome of interest, and $Y=0$, otherwise. The purpose of the survey is to estimate the prevalence $\bar{Y}$ of a health outcome Y in the defined target population.

In surveys with two phases of data collection, 'complete response' refers to the response pattern in which data are collected from the first data collection of the survey, consent is provided to contact healthcare providers, and data are collected from health providers in the second phase of data collection. 'Partial response' refers to the response pattern in which data are obtained from the first phase of data collection, but provider data from the second phase are missing either because consent to contact providers was not obtained or because data were not returned from healthcare providers in spite of obtaining consent. If partial responders' missing Y values are missing completely at random (MCAR), estimates of $\bar{Y}$ based only on data from complete responders will be unbiased. However, when partial responders are different from complete responders, estimates of $\bar{Y}$ based only on data from complete responders may be biased because partial respondents may be either more or less likely to have the health outcome of interest. This is a type of 'selection bias.'

The most common approach for adjusting for selection bias arising from partial response in sample surveys is the 'weighting class' methodology [1]. This method assigns each sampled child to a stratum-specific weighting class. Within each weighting class, the first-phase sample weights of the partial responders are redistributed among the complete responders, so that (with their adjusted first-phase sampling weights) complete responders represent the target population.

Within each weighting class, partial responders' missing Y values are assumed to be missing at random (MAR) [2, 3] [4, pp. 14–15]. The extent to which missing second-phase data are MAR determines how well the weighting class method reduces selection bias. Little and Rubin [4] suggest that when there is partial response, partial responders' missing second-phase data are usually not MAR. Thus, whenever weighting class adjustments are used, a relevant question is 'Is the MAR assumption correct?' In Section 2, we present a selection model and likelihood ratio tests for evaluating whether partial responder's missing Y values are MCAR or MAR. In Section 3, we use data from the U.S. National Immunization Survey (NIS) to illustrate our methods. Finally, we conclude our paper with a discussion of other relevant literature.

2. THE SELECTION MODEL

Let $R_t=1$ if the tth responder is a complete responder and $R_t=0$ if the tth responder is a partial responder. Also, let $Y_t=1$ if the tth responder has the health outcome of interest and $Y_t=0$ if the tth responder does not have the health outcome of interest. The joint probability distribution of Y_t and R_t is given in Table I.

Within the context of the joint distribution of Y_t and R_t, P_{t1} is the marginal probability that $Y_t=1$, and $P_{t2}=1-P_{t1}$ is the marginal probability that $Y_t=0$. Among respondents who have the health outcome of interest $Y_t=1$, and $\alpha_{t1}=P_{t1p}/P_{t1c}$ is the relative risk of being a partial responder. Among respondents who do not have the health outcome of interest $Y_t=0$, and $\alpha_{t2}=P_{t2p}/P_{t2c}$ is the relative risk of being a partial responder. Table II describes the joint probability distribution of Y_t and R_t in terms of relative risks $\{\alpha_{tj}\}$ and marginal probabilities $\{P_{tj}\}$.

DOI: 10.1002/sim

Table I. Probabilities for the joint distribution of the health outcome Y_t and respondent type R_t.

		Binary outcome, Y_t		
		$Y_t=1$	$Y_t=0$	Marginal probability
Respondent type, R_t	Complete, $R_t=1$	P_{t1c}	P_{t2c}	P_{tc}
	Partial, $R_t=0$	P_{t1p}	P_{t2p}	P_{tp}
	Marginal probability	P_{t1}	P_{t2}	1

Table II. Probabilities for the joint distribution of the health outcome Y_t and respondent type R_t in terms of relative risks α_{t1} and α_{t2} and marginal probabilities for the binary outcome.

		Binary outcome, Y_t		
		$Y_t=1$	$Y_t=0$	Marginal probability
Respondent type, R_t	Complete, $R_t=1$	$(1+\alpha_{t1})^{-1}P_{t1}$	$(1+\alpha_{t2})^{-1}P_{t2}$	P_{tc}
	Partial, $R_t=0$	$\alpha_{t1}(1+\alpha_{t1})^{-1}P_{t1}$	$\alpha_{t2}(1+\alpha_{t2})^{-1}P_{t2}$	P_{tp}
	Marginal probability	P_{t1}	P_{t2}	1

Models for $\{P_{tj}\}$ *and* $\{\alpha_{tj}\}$. Assuming that $x_t=(x_{t1},\ldots,x_{td_1})$ denotes the tth respondent's vector of covariates measured in the first phase of data collection and β denotes a column vector of unknown regression parameters, we express the relationship between x_t and the marginal probability P_{tj} $(j=1,2)$ using the logistic model

$$P_{tj}=\begin{cases}e^{x_t\beta}/(1+e^{x_t\beta}) & \text{if } j=1 \text{ and}\\ 1/(1+e^{x_t\beta}) & \text{if } j=2\end{cases} \tag{1}$$

Also, assuming that $z_t=(z_{t1},\ldots,z_{td_2})$ denotes another vector of covariates measured in the first phase of data collection for the tth respondent and $\gamma_j=(\gamma_{j1},\ldots,\gamma_{jd_2})'$ $(j=1,2)$ denotes two column vectors of unknown regression parameters, we express the relationship between z_t and the relative risks α_{tj} $(j=1,2)$ using the log-linear models

$$\alpha_{tj}=e^{z_{tj}\gamma_j}, \quad j=1,2 \tag{2}$$

The log-linear models (2) provide an extension of the selection model described by Lee and Marsh [5], and allows a more general way of allowing the relative risk of being a partial respondent to be different depending on whether $Y_t=1$ or $Y_t=0$. Within the context of the models for missing data described by Little [6], the model described in this section is a 'selection model' because it accounts for respondents' differential self-selection (i.e. differential risk) of being a partial respondent that can depend on differences in their health outcome Y_t and differences in their risk factors z_{tj}. When the risk of being a partial responder $(R_t=1)$ depends on z_t and Y_t, then partial and complete responders will be different with respect to the distributions of z_t and Y_t. These 'selective' differences will cause the estimate of the prevalence of Y to be biased when it is based on data from the complete responders only.

DOI: 10.1002/sim

The likelihood function for $\{P_{tj}\}$ *and* $\{\alpha_{tj}\}$. Let $\mathbb{S}_{R=1}$ denote the set of complete responders and $\mathbb{S}_{R=0}$ denote the set of partial responders. Also, among complete respondents let

$$y_{t1}=\begin{cases}1 & \text{if the } t\text{th complete respondent has the health outcome } (Y_t=1) \text{ and}\\ 0 & \text{otherwise}\end{cases}$$

Then, the log-likelihood function of $\{\alpha_{tj}\}$ and $\{P_{tj}\}$ is

$$l(\{\alpha_{tj}\},\{P_{tj}\})=\sum_{t\in\mathbb{S}_{R=1}}\{y_{t1}\ln((1+\alpha_{t1})^{-1}P_{t1})+(1-y_{t1})\ln((1+\alpha_{t2})^{-1}P_{t2})\}$$

$$+\sum_{t\in\mathbb{S}_{R=0}}\ln\left(\sum_{j=1}^{2}\alpha_{tj}(1+\alpha_{tj})^{-1}P_{tj}\right) \tag{3}$$

Substituting expressions (1) and (2) into (3), the log-likelihood may be expressed as a function of the unknown regression vector parameters β, γ_1, and γ_2. We denote that log-likelihood by $l(\beta,\gamma_1,\gamma_2)$. Assuming that $\hat{\beta}$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$ denote the values of β, γ_1, and γ_2 that jointly maximize the log-likelihood, $\max_{\beta,\gamma_1,\gamma_2} l(\beta,\gamma_1,\gamma_2)=\max l(\hat{\beta},\hat{\gamma}_1,\hat{\gamma}_2)$. Then, for the tth responder, the maximum likelihood estimate of the probability of having the health outcome $(Y_t=1)$ that corrects for selective differences between complete and partial responders is

$$\hat{P}_{t1}=e^{x_t\hat{\beta}}/(1+e^{x_t\hat{\beta}}) \tag{4}$$

We refer to (4) as the selection-bias-corrected probability of having the health outcome $(Y_t=1)$. This estimate corrects for selective differences between partial and complete responders that depend on the correlation between Y_t and R_t as modeled by the joint distribution of Y_t and R_t, and taken into account in the log-likelihood by $l(\beta,\gamma_1,\gamma_2)$.

A likelihood ratio test for evaluating whether missing data are MCAR. If the risk of being a partial respondent is the same for all complete and partial responders, then, regardless of their values of Y_t or z_t, the estimate of the prevalence $\bar{Y}$ will unbiased if it is based on data from the complete responders only. In this case complete responders' observed Y_t values are a random subsample from all responders' Y_t values and partial responders' missing Y_t values are 'ignorable' and are said to be MCAR [4].

With respect to the selection model, partial responders' missing Y_t values are MCAR if the intercepts of the vector parameters γ_j of the relative risk models (2) are identical for complete and partial responders and the other coefficients in the relative risk models are zero. That is, the null hypothesis for the MCAR assumption is $\gamma_{11}=\gamma_{21}$ and $\gamma_{1k}=\gamma_{2k}=0,\ k=2,\ldots,d_2$. To test the MCAR assumption, assuming that $\gamma^{(\mathrm{MCAR})}$ denotes γ_j under these conditions, the MCAR assumption may be evaluated using the likelihood ratio test statistic

$$\chi^2=2\times\left\{\max_{\beta,\gamma_1,\gamma_2} l(\beta,\gamma_1,\gamma_2)-\max_{\beta,\gamma^{(\mathrm{MCAR})}} l(\beta,\gamma^{(\mathrm{MCAR})},\gamma^{(\mathrm{MCAR})})\right\} \tag{5}$$

which has a chi-squared distribution with $2d_2-1$ degrees of freedom, asymptotically, when the MCAR assumption is true and the models are correctly specified. Severini [7] provides a simple proof demonstrating that, under certain regularity conditions, the likelihood ratio statistics have an asymptotic chi-squared distribution when the null hypothesis is true.

DOI: 10.1002/sim

A likelihood ratio test for evaluating whether missing data are MAR. If the risk of being a partial responder does not depend on Y_t but depends on z_t, complete responders' observed Y_t values are not necessarily a random subsample from all responders' Y_t values, but they are a random subsample of the sampled values within weighting classes defined by z_t. In this case, partial responders' missing Y_t values are said to be MAR [4]. With respect to the selection model, partial responders' missing Y_t values are MAR if the intercepts of vector parameters γ_j of the relative risk models (2) are identical for complete and partial responders, and each of the other coefficients is the same for complete and partial responders, also, but not necessarily zero. That is, the null hypothesis for the MAR model is $\gamma_{11}=\gamma_{21}$ and $\gamma_{1k}=\gamma_{2k},\ k=2,\ldots,d_2$. Assuming that $\gamma^{(\mathrm{MAR})}$ denotes γ_1 and γ_2 under these conditions, the MAR hypothesis may be evaluated using the likelihood ratio test statistic

$$\chi^2=2\times\left\{\max_{\beta,\gamma_1,\gamma_2} l(\beta,\gamma_1,\gamma_2)-\max_{\beta,\gamma^{(\mathrm{MAR})}} l(\beta,\gamma^{(\mathrm{MAR})},\gamma^{(\mathrm{MAR})})\right\} \tag{6}$$

which has a chi-squared distribution with d_2 degrees of freedom, asymptotically, when the MAR assumption is true and the models are correctly specified.

3. ILLUSTRATION USING DATA FROM THE NIS

The NIS. The NIS is a survey of U.S. children 19–35 months of age conducted by the Centers for Disease Control and Prevention (CDC) for the purposes of monitoring vaccination coverage rates in the United States. Data collection in the NIS occurs in two phases: a list-assisted random-digit-dialing (RDD) survey of households with landline telephones to identify households with children 19–35 months of age, followed by a mail survey to the age-eligible children's vaccination providers to obtain their vaccination histories.

When a household with an age-eligible child is identified in the first phase of the survey, the RDD interview is conducted, which collects demographic information about each age-eligible child in the household, demographic information about the age-eligible child's mother, and socio-demographic information about the household. First-phase sampling weights are constructed to account for the probability of sampling the age-eligible child. At the end of the first phase of data collection, consent is asked to contact the age-eligible children's vaccination providers. If consent is given, the second phase of data collection is conducted in the NIS. In the second phase, all vaccination providers named by the RDD respondent are contacted by mail to obtain the age-eligible child's provider-reported vaccination history. Provider-reported vaccination histories obtained from the mail survey of providers were used to evaluate the vaccination status of children sampled in the NIS. The NIS was reviewed and approved by the institutional review board at the CDC in 2001 and 2006. Smith *et al*. [8–10] have described the sampling design and statistical methods that have been used by the NIS to produce official estimates of vaccination coverage.

To illustrate the methods described in this paper, we used data sampled from California in the 2002 NIS. The health outcome Y is whether a sampled child is 'up-to-date' (UTD) as a result of having been administered 4 or more doses of diphtheria, tetanus toxoids, and acellular pertussis vaccine, 3 or more doses of polio vaccine, 1 or more doses of measles–mumps–rubella vaccine, and 3 or more doses of *Haemophilus influenzae* type b vaccine, as determined by the vaccination histories reported by their healthcare providers.

DOI: 10.1002/sim

Complete and partial response in California. Among sampled households in California that completed the first phase of the survey, 1101 were complete responders for whom consent was obtained to contact age-eligible children's healthcare providers and for whom vaccination histories were returned by those children's providers in the mail survey. Therefore, the UTD indicator Y was observed for all complete responders.

Also, among sampled households in California that completed the first phase of the survey, 213 were partial responders for whom consent was not obtained to contact healthcare providers and for whom a survey was not mailed to providers to obtain their provider-reported vaccination histories. Finally, among sampled households in California that completed the first phase of the survey, 416 were partial responders for whom consent was obtained to contact healthcare providers and for whom an adequate vaccination history was not returned in the mail survey of providers. The UTD indicator Y was missing for all partial responders.

A two-step weighting class method for adjusting for partial response. Because there are two different ways in which the UTD indicator Y can be missing for partial responders, we describe a weighting class approach that uses two independent steps to account for those two patterns of nonresponse.

In the first step, weighting classes are tailored to account for the differences between complete responders and partial responders for whom consent was not obtained. For the first step, let G denote the number of weighting classes, $\mathbb{S}^{g,1}_{R=0,\text{no consent}}$ denote the set of partial responders ($R=0$) for whom consent was not obtained that belong to weighting class $g=1,\ldots,G$ formed in the first step, $\mathbb{S}^{g,1}_{R=1}$ denote the set of complete responders ($R=1$) that belong to weighting class $g=1,\ldots,G$ formed in the first step, and W_j denote the first-phase sampling weight of the jth sampled unit. Then, let g_j denote the weighting class to which the jth sampled unit belongs in the first step, $\pi_j^{(1)}=W_j/\sum_{i\in\mathbb{S}^{g_j,1}_{R=1}} W_i$ denote the proportion of the first-phase sampling weights that the jth complete responder accounts for among complete responders in the g_jth weighting class, and $I_j^{(1)}=\pi_j^{(1)}\times\sum_{i\in\mathbb{S}^{g_j,1}_{R=0,\text{no consent}}} W_i$ denote the 'first step adjustment increment' for the jth complete responder.

In the second step, another 'new' specification of weighting classes is tailored to account for differences between complete responders and partial responders for whom consent was obtained and for whom an adequate vaccination history was not returned in the mail survey of providers. For the second step, let H denote the number of weighting classes, $\mathbb{S}^{h,2}_{R=0,\text{consent}}$ denote the set of partial responders ($R=0$) for whom consent was obtained that belong to weighting class $h=1,\ldots,H$ formed in the second step, $\mathbb{S}^{h,2}_{R=1}$ denote the set of complete responders ($R=1$) that belong to weighting class $h=1,\ldots,H$ formed in the second step, and W_j denote the first-phase sampling weight of the jth sampled unit. Then, let h_j denote the weighting class that the jth unit belongs to in the second step, $\pi_j^{(2)}=W_j/\sum_{i\in\mathbb{S}^{h_j,2}_{R=1}} W_i$ denote the proportion of the first-phase sampling weights that the jth complete responder accounts for among complete responders in the h_jth weighting class, and $I_j^{(2)}=\pi_j^{(2)}\times\sum_{i\in\mathbb{S}^{h_j,2}_{R=0,\text{consent}}} W_i$ denote the 'second step adjustment increment' for the jth complete responder. The jth complete responder's revised sampling weight W_j^* that is adjusted for the two types of partial response is obtained by adding its first-phase sampling weight W_j to its adjustment increments obtained in the first and second steps, $W_j^*=W_j+I_j^{(1)}+I_j^{(2)}$. Note that $\sum_{j\in\mathbb{S}_{R=1}} W_j^*$ is equal to the size of the target population.

DOI: 10.1002/sim

In practice, complete respondent's revised survey weights W_j^* are then raked on other variables associated with Y. Raking ensures that the marginal distribution determined from complete respondents revised survey weights matches the marginal distribution determined from complete and partial responder's first-phase survey weights. Smith *et al*. [10] describes raking in more detail. The next subsection describes how we tailored weighting classes in the two-step weighting class method to evaluate whether partial respondent's missing Y values were MCAR or MAR.

Formation of weighting classes using response propensities. To form weighting classes for both steps of the two-step weighting class method, a separate response propensity model was developed using logistic regression. For the first step, complete responders and partial responders for whom consent was not obtained were included in the modeling process, and the dependent variable was R_t, the binary indicator of whether a case was a complete or partial respondent. Predictor variables for the logistic model were selected among demographic variables (Table III) collected in the RDD phase of the NIS using a forward-stepwise variable selection procedure based on Akaike's [11] information criterion. Table IV shows the order in which selected demographic variables entered the stepwise selection procedure. Estimated predicted probabilities of being a complete responder (i.e. $\Pr[R_t = 1]$) were obtained from the response propensity model for each complete responder and partial responder for whom consent was not obtained. Because there were only 213 partial responders for whom consent was not obtained, two weighting classes were defined using the median of the distribution of the estimated predicted probabilities. With respect to the log-linear model for relative risks (2), the vector of covariates z_t consists of an intercept term that is always 1 and a binary indicator that codes the membership the tth respondent as belonging to one of the two weighting classes.

For the second step of the two-step weighting class method, complete responders and partial responders for whom consent was obtained were included in the logistic modeling process, and the dependent variable was R_t, the binary indicator of whether a case was a complete or partial respondent. Similar to the process described for the first-step, predictor variables for the logistic were selected among the demographic variables using Akaike's criterion. Table IV shows the order in which selected demographic variables entered the stepwise selection procedure for the second step of the two-step weighting class method. Estimated predicted probabilities of being a complete

Table III. Demographic variables used as candidates for inclusion in the forward-stepwise logistic analyses.

Variable name	Description of predictor variable
shotcard	Household (HH) used a shot card in reporting immunization status
itrueiap	Stratum identifier
educ1	Educational status of the mother
marital	Marital status of the mother
magegrp	Maternal age group
racekid	Race of the child
agegrp	Age group of the child
gender	Gender of the child
frstbrn	First-born status of the child
childnm	Number of children under 18 years of age living in the HH
incpov1	Poverty status
mobil	Mobility status
msa	HH within central city of MSA, suburban, or nonmetro area
c5	Relationship of the HH respondent to the child (mother, father, or others)

DOI: 10.1002/sim

Table IV. Predictors x_t of being UTD and predictors z_t of the risk of being a partial responder selected in the forward-stepwise logistic regression analyses.

Variable name	Order of selection for the logistic model of being UTD, x_t	Order of selection for the response propensity models for the two-step weighting class method	
		Step 1 weighting classes, z_t	Step 2 weighting classes, z_t
mobil	1		4*
shotcard	2		2
agegrp	3		
incpov1	4		1
educ1	5		3
racekid	6	1*	
shotcard: agegrp	7*		

Numbers listed in the table identify the order of entry in the forward-stepwise procedure. Variables separated by a colon denote a second-order interaction. Variables listed with an asterisk were the last variables among those listed in Table III that were selected by the forward-stepwise variable selection procedure.

responder (i.e. $\Pr[R_t = 1]$) were obtained from the response propensity model for each complete responder and partial responder for whom consent was obtained. Because there were 416 partial responders for whom consent was obtained, four weighting classes were defined using the quartiles of the distribution of the estimated predicted probabilities. With respect to the log-linear model for relative risks (2), the vector of covariates z_t consists of an intercept term that is always 1 and three binary indicators that code the membership the tth respondent as belonging to one of the four weighting classes.

To specify the predictors x_t for the logistic model of the marginal probability P_{tj} of being UTD (1), logistic regression was used. In this model the dependent variable was the binary indicator of vaccination status, Y_t, and predictor variables were selected from the demographic variables listed in Table III using a forward-stepwise variable selection procedure based on Akaike's information criterion [11]. Table IV shows the order in which selected demographic variables entered the stepwise selection procedure.

Evaluating whether partial respondents missing Y values are MCAR or MAR for the first step weighting classes of the two-step weighting class method. To evaluate whether the missing Y values of partial responders for whom consent was obtained were MCAR, we used likelihood ratio statistic (5). For this statistical test, $\chi^2 = 20.1$ with 3 degrees of freedom, and the associated p-value is <0.01, providing very little support for the MCAR hypothesis.

To evaluate whether the missing Y values of partial responders for whom consent was obtained were MAR, we used likelihood ratio statistic (6). For this statistical test, $\chi^2 = 3.0$ with 2 degrees of freedom, and the associated p-value is 0.22. This suggests that the MAR assumption underlying the first step weighting class strategy is supported by data used by the selection model.

Evaluating whether partial respondents missing Y values are MCAR or MAR for the second step weighting classes of the two-step weighting class method. To evaluate whether the missing Y values of partial responders for whom consent was not obtained were MCAR, we used likelihood ratio statistic (5). For this statistical test, $\chi^2 = 118.2$ with 7 degrees of freedom, and the associated p-value is <0.01, providing very little support for the MCAR hypothesis.

To evaluate whether the missing Y values of partial responders for whom consent was not obtained were MAR, we used likelihood ratio statistic (6). For this statistical test, $\chi^2 = 1.6$ with

DOI: 10.1002/sim

4 degrees of freedom, and the associated p-value is 0.81. This suggests that the MAR assumption underlying the second step weighting class strategy is supported by data used by the selection model.

In our study, we used the Nelder–Mead simplex method [12] to maximize the likelihood (3). All computations were conducted in *R* [13].

4. DISCUSSION

This paper describes a selection model for testing whether partial responders' missing Y values are MCAR or MAR. The model is useful for evaluating MAR assumption that is made when the weighting class method is used to adjust for partial nonresponse. For any weighting class approach used for adjusting for partial nonresponse, if the MAR hypothesis is rejected, the selection-bias-corrected probabilities of having the health outcome ($Y_t = 1$) given by equation (4) may be used to form weighting classes, as the basis for providing model-based multiple imputations or as the basis for designating hot-deck donors.

In other research, Little [14] proposed methods based on Wald tests for evaluating the MCAR assumption, which are tailored primarily for outcomes that are nondiscrete, continuous variables. Little's methods are based on the assumption that the missing values are MAR and therefore cannot be used to evaluate the validity of the MAR assumption used by the weighting class method.

Other literature [15] suggests that if weighting classes are formed using estimated response propensities or estimated predicted means, selection bias will be controlled. Rosenbaum and Rubin [16, 17] have provided a discussion of the theory of propensity scores, and Little [15] has provided a discussion of their use in the context of survey nonresponse. Little and Rubin [4] have shown that if weighting classes can be formed so that, within each weighting class, partial respondents' missing Y values are MAR and complete responders' observed Y are observed at random, the mechanism causing partial response is independent of Y and the sampling plan. In this case the biasing effect of nonresponse on the weighting class estimator will be zero, provided the response propensity model is correctly specified in the sense described by White [18] and Domowitz and White [19].

The importance of adjusting statistical methods so that inferences are not effected by selection bias has become a major theme in the econometric and statistical literature [5, 20, 21]. Greenlees *et al*. [22] describe an imputation model for missing values when the probability of response depends on the variable being imputed. Allison [23] shows that imputation based on subclassification on propensity scores with the approximate Bayesian bootstrap produces biased estimates. Fay [24] and Baker and Liard [25] discuss models for categorical data when there is nonignorable nonresponse. Jinn and Sedransk [26] and Wang *et al.* [27] report on the effect on secondary data analysis of the use of imputed values in the case where missing data are not MAR. Diggle and Kenward [28] and Rotnitsky *et al.* [29] describe models for longitudinal data when there is informative drop-out. Raghunathan and Grizzle [30] describe imputation methods when blocks of data are missing.

There are important weaknesses and challenges of selection models that should be noted. In an important work conducted by Kenward [31] and Molenberghs *et al*. [32, 33], it has been shown that different models that account for missing Y values as not being MCAR (non-MCAR models) may produce the same fit to the observed data but yield different predictions for the missing values. These authors suggest that, in this case, choosing between these different models cannot be based on examining the data alone. Rather, these authors assert that this choice must be based on the plausibility of each model's underlying assumptions in conjunction with other scientific

DOI: 10.1002/sim

knowledge that provides the context within which the data were collected. In the case of childhood vaccination coverage rates, because of societal expectations of a mother to protect their child, one might expect a mother who knows that their child is not UTD to be reluctant to give consent for the NIS to contact their child's vaccination providers, out of fear of revealing that their child is not UTD. Also, a vaccination provider who is selective about which vaccination to administer may be reluctant to respond to the second phase of data collection of the NIS out of fear of revealing that they are not following a nationally recommended standard of care. In both cases, the child is likely to be a partial responder. Further, not being UTD is positively correlated with being a partial nonresponder. Because of this it seems reasonable to entertain the possibility that this child's missing Y values are not MAR.

A further challenge of selection models pertain to model specification. Specifically, the assessments of MCAR and MAR are subject to a correct specification of the predictors $\{x_t\}$ for the marginal probability model (1) and a correct specification of the predictors $\{z_{tj}\}$ for the log-linear models (2). Numerous texts describe methods for selecting independent variables in logistic and log-linear models [34, 35]. In our study, we selected predictors for the logistic model (1) using a forward-stepwise procedure that was separate from the forward-stepwise selection procedure that was used to select predictors for both log-linear models (2). In this regard, because our selection model has three separate linear predictors, it shares similar challenges in specifying those predictors with other models that have more than one linear predictor, such as hierarchical models. A more careful selection of predictors for those linear predictors could be based on manually inserting and/or deleting predictors for either the logistic or the log-linear models in our selection model, and keeping or dropping predictors, depending on the significance of the decrease and/or increase in the selection model's log-likelihood.

Kenward and Molenberghs *et al*. rightly claim that it is impossible to prove that the missing Y values are MAR or MCAR, since they are not observed. Does this mean that the weighting class method should be used without trying to examine its assumptions? Such an inclination seems unwise. In an earlier study, Thomsen [36] showed that estimates based on the weighting-class method may not necessarily be unbiased. Our method provides one way of evaluating the assumptions of the weighting class method. An advantage of the selection model we present is that both MAR and MCAR models are special cases of this general model. Because of this, the model we describe allows analysts to test the MAR and MCAR assumptions and to conduct more extensive sensitivity analyses of the bias of weighting class estimates, assuming the selection model is correct. In other settings, Vach and Blettner [37] proposed the use of several different plausible non-MCAR models to conduct further sensitivity analyses.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the comments made by Sadeq Chowdury during the development of this paper.

REFERENCES

1. Brick JM, Kalton G. Handling missing data in survey research. *Statistical Methods in Medical Research* 1996; **5**:215–238.
2. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.
3. Rubin DB. *Multiple Imputation*. Wiley: New York, 1987.
4. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.

DOI: 10.1002/sim

5. Lee B-J, Marsh LC. Sample selection bias correction for missing observations. *Oxford Bulletin of Economics and Statistics* 2000; **62**(2):305–322.
6. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**:1112–1121.
7. Severini TA. *Likelihood Methods in Statistics*. Oxford University Press: New York, 2000.
8. Smith PJ, Rao JNK, Battaglia MP, Ezzati-Rice TM, Daniels D, Khare M. Compensating for provider nonresponse using response propensities to form weighting classes: the National Immunization Survey. National Center for Health Statistics. *Vital Statistics* 2001; **2**(133):1–17.
9. Smith PJ, Battaglia MP, Huggins VJ, Hoaglin DC, Rodén A-S, Khare M, Ezzati-Rice TM, Wright RA. Overview of the sampling design and statistical methods used in the National Immunization Survey. *American Journal of Preventive Medicine* 2001; **20**(Suppl. 4):17–24.
10. Smith PJ, Hoaglin DC, Battaglia MP, Khare M, Barker LE. Statistical methodology of the National Immunization Survey: 1994–2002. National Center for Health Statistics. *Vital Statistics* 2005; **2**(138):1–56.
11. Akaike H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Petrov BN, Csáki F (eds). Akademia Kiadó: Budapest, 1973; 267–281.
12. Nelder JA, Mead R. A simplex method for function minimization. *Computer Journal* 1965; **7**:308–315.
13. Venables WN, Smith DM, The R Development Core Team. *An Introduction to R*. Network Theory Ltd.: Bristol, U.K., 2002. (Available from: http://www.cran.r-project.org/doc/manuals/R-intro.pdf.)
14. Little RJA. A test for completely missing at random for multivariate data with missing values. *Journal of the American Statistical Association* 1988; **83**(404):1198–1202.
15. Little RJA. Survey nonresponse adjustments for estimates of means. *International Statistical Review* 1986; **54**:139–157.
16. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies. *Biometrika* 1983; **70**:41–55.
17. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling incorporating the propensity score. *The American Statistician* 1985; **39**:33–38.
18. White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; **50**:1–25.
19. Domowitz I, White H. Misspecified models with dependent observations. *Journal of Econometrics* 1982; **20**(1): 35–58.
20. Heckman JJ. Sample selection bias as a specification error. *Econometrica* 1979; **47**:153–161.
21. Groves RM, Dillman DA, Eltinge JL, Little RJA. *Survey Nonresponse*. Wiley-Interscience: New York, 2002.
22. Greenlees JS, Reece WS, Zieschang KD. Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* 1982; **77**:251–261.
23. Allison PD. Multiple imputation for missing data: a cautionary tale. *Sociological Methods and Research* 2000; **28**(3):301–309.
24. Fay RE. Causal models for nonresponse. *Journal of the American Statistical Association* 1986; **81**(394):354–365.
25. Baker SG, Laird NM. Regression analysis of categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* 1988; **83**(401):62–69.
26. Jinn JH, Sedransk J. Effect on secondary data analysis of the use of imputed values in the case where missing data are not missing at random. *Proceedings of the Section on Survey Research Methods*. American Statistical Association: Alexandria, VA, 1989; 51–61.
27. Wang R, Sedransk J, Jinn JH. Secondary data analysis when there are missing observations. *Journal of the American Statistical Association* 1992; **87**(420):952–961.
28. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. *Applied Statistics* 1994; **43**(1):49–93.
29. Rotnitsky A, Robins JM, Scharfstein DO. Semi-parametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* 1998; **93**:1321–1339.
30. Raghunathan TE, Grizzle JE. A split questionnaire survey design. *Journal of the American Statistical Association* 1995; **90**:54–63.
31. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity analysis. *Statistics in Medicine* 1998; **17**:2723–2732.
32. Molenberghs G, Goetghesbeur Lipsitz SR, Kenward MG. Nonrandom missingness in categorical data: strengths and limitations. *The American Statistician* 1999; **53**(2):110–117.
33. Molenberghs G, Kenward MG, Goetghesbeur E. Sensitivity analysis for incomplete contingency tables: the Solvenian plebisite case. *Applied Statistics* 2001; **50**(1):15–29.
34. Christianson R. *Log-linear Models and Logistic Regression* (2nd edn). Springer: New York, 1997.

DOI: 10.1002/sim

35. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*: *A Practical Information-theoretic Approach*. Springer: New York, 2002.
36. Thomsen I. A note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift* 1973; **4**:278–293.
37. Vach W, Blettner M. Logistic regression with incompletely observed categorical covariates—investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine* 1995; **14**:1315–1329.

DOI: 10.1002/sim