A large health survey was combined with a simulation study to contrast the reduction in bias achieved by double sampling versus two weighting methods based on propensity scores. The survey used a census of one New York county and double sampling in six others. Propensity scores were modeled as a logistic function of demographic variables and were used in conjunction with a random uniform variate to simulate response in the census. These data were used to estimate the prevalence of chronic disease in a population whose parameters were defined as values from the census. Significant \( p < 0.0001 \) predictors in the logistic function included multiple (vs. single) occupancy \( (OR = 1.3) \), bank card ownership \( (OR = 2.1) \), gender \( (OR = 1.5) \), home ownership \( (OR = 1.3) \), head of household’s age \( (OR = 1.4) \), and income \( >$18,000 \) \( (OR = 0.8) \). The model likelihood ratio chi-square was significant \( p < 0.0001 \), with the area under the receiver operating characteristic curve = 0.59. Double-sampling estimates were marginally closer to population values than those from either weighting method. However, the variance was also greater \( p < 0.01 \). The reduction in bias for point estimation from double sampling may be more than offset by the increased variance associated with this method.

After receiving a less than 100 percent initial response to a survey, an investigator is faced with three choices. The first is to estimate population parameters with the existing data either with or without applying some weighting methodology. Second, a random sample of the nonresponders can be induced to complete the survey and estimation procedures applicable to stratified (double) sampling used. Third, all nonresponders can be recontacted in an attempt to increase the overall response rate.

The extent to which increasing the overall response rate of a survey will decrease estimation bias is a complex subject, with conflicting reports in the literature. Many studies report that the intuitive result that increased nonresponse follow-up resulted in decreased bias \( (1, 2) \). Others, however \( (3–5) \), report little or no improvement. Keeter et al. \( (3) \) found only small differences even when the overall response rate was increased from 36 percent to 61 percent.

Avoiding nonresponse follow-up completely and estimating parameters by weighting data from the first wave of respondents has the primary advantage of saving time and money. Adjustment methods differ, with this paper concentrating on those relating to propensity scores.

The use of propensity scores to adjust for unit nonresponse has recently received considerable attention in the literature. The propensity score is defined as the probability of response and is typically modeled as a logistic function of covariates available from both responding and nonresponding subjects. A good overview of this method can be found in Little \( (6) \) and in Rosenbaum and Rubin \( (7) \).

After propensity scores are assigned, subjects are grouped into propensity classes or quantiles. Previous research by Cochran \( (8) \), Rosenbaum and Rubin \( (9) \), and Eltinge and Yansaneh \( (10) \) suggests that five classes are adequate to achieve most of the possible bias reduction. A weight is then assigned to each propensity class. The two most common of these classes are propensity stratification, in which the weight is the inverse of the proportion of subjects with complete data within the stratum, and propensity weighting,
The 1999 Otsego County Health Census

A detailed description of the 1999 Otsego County census can be found in Jenkins et al. (16). Briefly, every nonseasonal, noncommercial residence in the county was enumerated via 911 emergency address listings, tax maps, and visual confirmation. Surveys were mailed in June 1999, with door-to-door collection 3 weeks later. Subsequent telephone contacts for nonresponders increased the final response rate to 78.6 percent.

For both this study and the simulation, the sampling unit was the household. The person filling out the survey, either by phone or mail, was designated the “index” person, regardless of household role. This person provided the health information used in the current study for all individuals in the household older than age 18 years.

Double sampling of the six counties adjacent to Otsego in 1999

The six counties surrounding Otsego were double sampled by using a sampling frame provided by Genesys Corporation of Fort Washington, Pennsylvania. Initially, a survey was mailed to a random sample of 5,000 permanent residences in Herkimer, Madison, Schoharie, Montgomery, Delaware, and Chenango counties (30,000 total residences). After 2 weeks, a reminder letter was sent to nonresponders. A random sample of 200 households from each county that did not respond to this reminder letter within 3 weeks was telephoned and was offered $25 to participate. Details behind the theory of double sampling can be found in Neyman (14, 17) and in Hanson et al. (18, 19).

Response rates in these double-sampled counties averaged 37.5 percent and 60.3 percent for the first and second waves of sampling, respectively. Point estimates and their variances were computed by using composite formulas found in such texts as Levy and Lemeshow (20), Cochran (21), and Thompson (22). Further details of the results of the double sampling can be found in Jenkins et al. (23).

Methods for the simulation study

For simulation purposes, the population was defined as all individuals older than age 18 years living in households that responded to the 1999 Otsego County Health Census and could be matched to the sampling frame provided by Genesys Corporation.

Identification of predictors of household response and the four chronic diseases

To identify predictors of response, demographic variables were compared between homes that responded to the 1999 survey mailing and those that did not in the six counties surrounding Otsego. These demographic data were provided by Genesys Corporation in the file used as the sampling frame for the 1999 study.

This process involved a separate analysis of response versus nonresponse for each stage of the double sampling. Household variables compared were single versus multiple occupancy, home value ($25,000–<75,000, $75,000–<150,000, ≥$150,000), number of vehicles (1–2, 3, 4–7, ≥8), aggregate vehicle market value, and highest vehicle market value. Head-of-household variables compared were male versus female, married versus single, high versus low (≤$18,000) income, owner versus nonowner, bank card ownership, and less than age 45 years versus age 45 years or older.

MATERIALS AND METHODS

The 1999 Otsego County Health Census

A detailed description of the 1999 Otsego County census can be found in Jenkins et al. (16). Briefly, every nonseasonal, noncommercial residence in the county was enumerated via 911 emergency address listings, tax maps, and visual confirmation. Surveys were mailed in June 1999, with door-to-door collection 3 weeks later. Subsequent telephone contacts for nonresponders increased the final response rate to 78.6 percent.

For both this study and the simulation, the sampling unit was the household. The person filling out the survey, either by phone or mail, was designated the “index” person, regardless of household role. This person provided the health information used in the current study for all individuals in the household older than age 18 years.

Double sampling of the six counties adjacent to Otsego in 1999

The six counties surrounding Otsego were double sampled by using a sampling frame provided by Genesys Corporation of Fort Washington, Pennsylvania. Initially, a survey was mailed to a random sample of 5,000 permanent residences in Herkimer, Madison, Schoharie, Montgomery, Delaware, and Chenango counties (30,000 total residences). After 2 weeks, a reminder letter was sent to nonresponders. A random sample of 200 households from each county that did not respond to this reminder letter within 3 weeks was telephoned and was offered $25 to participate. Details behind the theory of double sampling can be found in Neyman (14, 17) and in Hanson et al. (18, 19).

Response rates in these double-sampled counties averaged 37.5 percent and 60.3 percent for the first and second waves of sampling, respectively. Point estimates and their variances were computed by using composite formulas found in such texts as Levy and Lemeshow (20), Cochran (21), and Thompson (22). Further details of the results of the double sampling can be found in Jenkins et al. (23).

Methods for the simulation study

For simulation purposes, the population was defined as all individuals older than age 18 years living in households that responded to the 1999 Otsego County Health Census and could be matched to the sampling frame provided by Genesys Corporation.

Identification of predictors of household response and the four chronic diseases

To identify predictors of response, demographic variables were compared between homes that responded to the 1999 survey mailing and those that did not in the six counties surrounding Otsego. These demographic data were provided by Genesys Corporation in the file used as the sampling frame for the 1999 study.

This process involved a separate analysis of response versus nonresponse for each stage of the double sampling. Household variables compared were single versus multiple occupancy, home value ($25,000–<75,000, $75,000–<150,000, ≥$150,000), number of vehicles (1–2, 3, 4–7, ≥8), aggregate vehicle market value, and highest vehicle market value. Head-of-household variables compared were male versus female, married versus single, high versus low (≤$18,000) income, owner versus nonowner, bank card ownership, and less than age 45 years versus age 45 years or older.
Variables significantly related to response were then compared between subjects with versus without asthma, heart disease, diabetes, and cancer. All comparisons described, except those for aggregate and highest vehicle value, which were analyzed via t-test, were made using chi-square tests.

Creation of propensity scores

A binary logistic model was used to create household propensity scores for both the first and second responses. These scores were defined as the probability of responding at each of these two stages and were referred to as IPROPEN and SPROPEN, respectively.

The requirement for a variable to be included in these equations was that it be related to at least one of the four chronic diseases and also to the probability of response for whichever of the two stages of sampling was being modeled.

Overview of sampling simulation

The simulation was designed to model the sampling that occurred in the six counties surrounding Otsego in 1999 and was repeated for 500 iterations. These simulations and all data analyses were performed by using SAS software (SAS Institute, Inc., Cary, North Carolina).

Simulation of initial sampling. A random sample of 5,000 homes in Otsego County was selected by using the SAS SURVEYSELECT procedure with the SIZE = IPROPEN option. This procedure weighted the likelihood of response by the propensity score for the first stage of sampling such that the probability of being selected was proportional to it.

Simulation of the second stage of sampling. A random sample of 200 homes was selected from the 3,500 homes that did not “respond” to the initial mailing. This process was meant to simulate telephone follow-up of the nonresponders. From these 200 homes, 120 (60 percent as per the 1999 study) would be selected to “respond” according to the SPROPEN propensity score. This step was conducted in a manner analogous to that described above for IPROPEN.

Endpoint estimation. The population parameters for the four chronic diseases were defined as the prevalences observed among all subjects in the combined group of homes from which the samples were generated. In this context, a case is defined as an individual reporting ever having received the diagnosis from a health professional. These parameters were estimated in the same manner as for the double-sampled counties in 1999, as follows:

\[ \text{prev}_{\text{overall}} = \left( \frac{\text{prev}_r \times p_r}{C_0} \right) + \left( \frac{\text{prev}_n \times p_n}{C_0} \right), \]

where

- \( \text{prev}_{\text{overall}} \) is the estimate of the overall population prevalence;
- \( \text{prev}_r \) is the prevalence of the attribute in the responders;
- \( p_r \) is the proportion of responders in the initial sample;
- \( \text{prev}_n \) is the prevalence of the attribute in the nonresponders; and
- \( p_n \) is the proportion of the nonresponders in the initial sample; more explicitly, \( p_n \) is estimated as \( 1 - p_r \).

Estimating \( p_n \) as \( 1 - p_r \) is equivalent to assuming that the nonresponse follow-up obtained a 100 percent response rate and thus constitutes a random sample of all nonresponders. Since the 1999 study had a 60.3 percent follow-up response rate, this rate was modeled in the simulation. Thus, the assumption of a 100 percent response rate is known to not be valid but was used because it was considered the alternative that would produce the least bias and was also used in 1999. The variance over the 500 replications was estimated as for any continuous random variable.

Prevalences of the four chronic diseases were also estimated from the initial wave of sampling by using both propensity stratification and propensity weighting. For propensity stratification, the formula was

\[ \text{prev}_{\text{overall}} = \frac{\sum_{i=1}^{5} (\text{prev}_i \times w_i^{-1})}{\sum_{i=1}^{5} w_i^{-1}}, \]

where

- \( \text{prev}_{\text{overall}} \) is the estimate of the overall population prevalence;
- \( \text{prev}_i \) is the prevalence in the \( i \)th stratum; and
- \( w_i \) is the proportion of subjects in the \( i \)th stratum who “responded” to the initial wave of sampling.

For propensity weighting, the same formula was used, with \( w_i \) being the mean propensity score in the \( i \)th stratum for subjects who “responded” to the initial wave of sampling.

The estimates were compared between the four methods separately for each of the four diseases by using one-way analysis of variance models with pairwise comparisons via Scheffe’s test. Differences in the variances of these estimates were tested with the ratio variance F test.

RESULTS

Demographic data were available from the Genesys file on 27,210 subjects from the six counties. Significant differences were seen between responders and nonresponders for single versus multiple occupancy, bank card possession, head of household age greater than 45 years, head of household male, homeowner versus renter, and income greater than $18,000. Significant responder versus nonresponder differences regarding the second wave of sampling were seen for owner versus renter only. All of these variables, with the exception of head of household gender, which was marginally related to asthma only, were significantly related to at least one of the four chronic diseases. Regression coefficients for the logistic propensity function to predict response to the initial mailing (IPROPEN) are shown in table 1. All six variables significantly related to response at the univariate level were found to be independently predictive when combined in this single model. The model likelihood ratio chi-square of 687.34 was significant at \( p < 0.0001 \). The area under the receiver operating characteristic curve of 0.59 indicated better than chance prediction of response (24).

The model for response to the second phase of sampling (table 1) contained the owner versus renter \( (p = 0.0016) \) and head of household male \( (p = 0.1100) \) variables. Although the head of household male variable did not reach significance at \( p = 0.05 \), it was retained as a means of producing adequate variance in the second propensity score (SPROPEN). The
model likelihood ratio chi-square was 11.00 \( (p = 0.0041) \), and the area under the receiver operating characteristic curve was 0.55.

Table 2 shows disease prevalence in the population from which the data were simulated and also from each of the two phases of the simulated sampling. For three of the four chronic diseases, prevalence was slightly lower in the second phase of sampling and was closer to the population value. The exception to this trend was for asthma, which had slightly lower prevalence in the first phase that was closer to the population value. The analysis-of-variance models showed a significant main effect of at least \( p < 0.0025 \) for estimating method for all four of the diseases. Pair-wise comparisons showed that the unweighted estimates differed significantly from those of the other three methods for cancer, heart disease, and diabetes. For asthma, the unweighted estimate differed significantly from only the propensity weighted estimate.

Estimates from propensity weighting and propensity stratification did not differ significantly from each other for any disease. Propensity stratification estimates differed significantly from double-sampling estimates for heart disease \( (p = 0.0006) \) and diabetes \( (p = 0.0311) \). Estimates for propensity weighting differed significantly from those for double sampling for heart disease only \( (p = 0.0186) \).

Double-sampling variances (table 4) were significantly greater \( (p < 0.01) \) than those obtained from the other three methods. Propensity stratification variances were identical to unweighted variances and were the smallest for all four diseases. Propensity weighting variances were significantly greater \( (p < 0.01) \) than unweighted or propensity stratification variances.

Jenkins et al. (23) also reported a tendency toward 2–4 percent higher rates of heart disease, diabetes, and cancer in responders than in nonresponders in the double-sampled counties in 1999. The simulation produced similar, statistically significant effects in Otsego County for diabetes,
cancer, and heart disease, although the magnitude of the differences was not as great, ranging from 0.6 percent for heart disease to 0.3 percent for both cancer and diabetes. Overall, the simulation seems to have successfully recreated the essential elements of the double-sampling outcomes that occurred in the six counties in 1999. If so, it can be inferred that the lower levels of disease observed in the nonresponders in the actual 1999 study were probably closer to the unknown population values and that their incorporation into the estimation process reduced bias.

**DISCUSSION**

Double sampling was the only method that produced prevalence estimates identical to the population parameters for all four diseases. However, the difference between these estimates and those obtained from the weighting methods that utilized only the first wave of sampling—although in some instances statistically significant—was, for practical purposes, trivial. This fact, in conjunction with the significantly greater variances associated with the double-sampling estimates, as well as the significant cost of performing a second wave of sampling, seriously calls into question whether the double sampling was justified.

It would be advantageous if investigators were able to decide before conducting a second wave of sampling whether it would be justified by the reduction in nonresponse bias. The literature does contain helpful examples (25) of the dependence of bias on both the response rate and differences between responders and nonresponders. Some investigators (26), however, would contend that setting cutoffs for what constitutes an acceptable response rate is meaningless. Others (5, 27, 28) have also shown that a low initial response rate, such as the 38.0 percent rate from 1999 that was simulated, does not necessarily guarantee that bias exists.

Demographic differences between responders and nonresponders may not be an indicator of nonresponse bias either. For example, in a study with an overall response rate of 67.0 percent, Shahar et al. (29) found bias in disease prevalence estimates to be less than 5 percent despite seeing significant differences between responders and nonresponders in smoking, marital status, and education.

These results lend credence to a position summarized by Mayer and Pratt, who stated we wish to suggest that a mechanistic approach to controlling nonresponse bias (e.g., insistence on an arbitrarily high response rate) could result in the utilization of expensive research tools to avoid the possibility of unknown bias in some characteristic, when, in fact, known data on other characteristics indicate that extensive biases are not likely to be present (30, p. 637).

Rather than condition the probability of the subject’s responses directly on the disease endpoints, the study conditioned this probability on demographic variables related to the diseases.

The resulting bias in the prevalence estimates that was observed remains attributable to outcome-dependent nonresponse.


Groves and Couper (31) emphasized the importance of differentiating between active refusal and noncontact when modeling and adjusting for nonresponse. The finding of different covariates in the propensity models for the first stage (a mailed survey in which 90 percent of the missing data was due to active refusal) versus the second stage (a phone survey with 25 percent of the missing data due to active refusal) underscores this point.

**Weaknesses**

Developing the propensity model for the Otsego County simulation by using the 1999 data from the six surrounding counties required the assumption that a common response pattern would be seen in both areas. Although this assumption seems reasonable, it is not directly testable. A second issue arises from the fact that, in both the 1999 and simulated study, “response” was determined at the household level whereas disease was determined at the individual level. As such, the decision by a single individual in the home regarding whether to respond determined whether the data from all individuals in the household were included. Because this constraint existed for all of the estimation methods being compared, it is unlikely that it affected the conclusions drawn.

We have assumed that nonresponse depends upon measured covariates and not on disease status conditional on those covariates. If, instead, nonresponse is a function of disease status even after accounting for covariates, then bias can exist in estimating disease prevalence, even with the adjustments we have described.

It is important to note that the conclusions of this study are limited to the case in which a survey is being used to estimate disease prevalence. Understanding the extent to which the results may also generalize to other areas, such as the associations between disease and risk factors in the population, requires further study.

**Conclusions**

These results indicate that, even when significant differences exist between responders and nonresponders regarding demographic variables known to be related to study endpoints, undertaking a second wave of sampling may only marginally reduce estimation bias. In this study, nearly comparable bias reduction was demonstrated with either of two propensity weighting methods that did not require double sampling. These alternatives are particularly attractive when one considers that they completely avoid the time and expense involved in undertaking a second wave of sampling. In addition, they do not produce the variance inflation that accompanies double sampling.

**ACKNOWLEDGMENTS**

Conflict of interest: none declared.
REFERENCES