American Journal of Epidemiology

Vol. 171, No. 5
DOI: 10.1093/aje/kwp440
Advance Access publication:
February 4, 2010

Practice of Epidemiology

Estimating Model-Adjusted Risks, Risk Differences, and Risk Ratios From Complex Survey Data

Gayle S. Bieler*, G. Gordon Brown, Rick L. Williams, and Donna J. Brogan

* Correspondence to Gayle S. Bieler, RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194 (e-mail: gbmac@rti.org).

Initially submitted March 31, 2009; accepted for publication December 9, 2009.

There is increasing interest in estimating and drawing inferences about risk or prevalence ratios and differences instead of odds ratios in the regression setting. Recent publications have shown how the GENMOD procedure in SAS (SAS Institute Inc., Cary, North Carolina) can be used to estimate these parameters in non-population-based studies. In this paper, the authors show how model-adjusted risks, risk differences, and risk ratio estimates can be obtained directly from logistic regression models in the complex sample survey setting to yield population-based inferences. Complex sample survey designs typically involve some combination of weighting, stratification, multistage sampling, clustering, and perhaps finite population adjustments. Point estimates of model-adjusted risks, risk differences, and risk ratios are obtained from average marginal predictions in the fitted logistic regression model. The model can contain both continuous and categorical covariates, as well as interaction terms. The authors use the SUDAAN software package (Research Triangle Institute, Research Triangle Park, North Carolina) to obtain point estimates, standard errors (via linearization or a replication method), confidence intervals, and *P* values for the parameters and contrasts of interest. Data from the 2006 National Health Interview Survey are used to illustrate these concepts.

health surveys; logistic regression; logistic risk; odds ratio; prevalence; risk; risk ratio; survey analysis

Abbreviations: CRN, cost-related nonadherence; NHIS, National Health Interview Survey.

There is increasing interest in the public health community in estimating and drawing inferences about risk ratios and risk differences instead of odds ratios in the binary-response regression setting (e.g., see Greenland (1, 2)). Spiegelman and Hertzmark (3) have recently shown how the GENMOD procedure in SAS (SAS Institute Inc., Cary, North Carolina) can be used to estimate these parameters, using the log-binomial regression model for the risk ratio and the binomial regression model for the risk difference. Zou (4) has recommended the modified Poisson model with a Huber (5) robust variance estimate when maximum likelihood estimation of the log-binomial or binomial model fails to converge. Cheung (6), noting that there are situations when even the modified Poisson model can fail to converge, has shown that risk differences can instead be reliably estimated via an ordinary least-squares linear regression model with a binary response variable and a robust variance estimate. Although there is no risk of nonconvergence with Cheung's modified least-squares regression method, it has the drawback that estimates of the model-adjusted risk, when risk is itself a parameter of interest, are not bounded by 0 and 1.

All of the previous work was proposed in the context of non-population-based studies. Complex sample surveys are designed to yield population-based estimates and inferences, and they typically involve some combination of sample weighting, stratification, multistage sampling, clustering, and perhaps finite population adjustments. Many public-use data files contain information from national public health surveys employing complex sampling schemes, including the National Health and Nutrition Examination Survey and the National Health Interview Survey (NHIS). Special statistical methods are needed to account for these complex sample designs in order to obtain unbiased estimates of population parameters, appropriate standard errors and confidence intervals, and valid

population inferences. Brogan (7, 8) has discussed the impact of sample survey design on data analysis and has illustrated the possible consequences of ignoring the survey design in analysis of national health survey data. In the logistic regression setting, accounting for the sample design via design-based methods typically implies weighted maximum likelihood estimation of model parameters and either Taylor linearization or a replication method for variance estimation (9).

In this paper, we show how model-adjusted risks, risk differences, and risk ratio estimates can all be obtained directly from logistic regression models in the complex sample survey setting. Point estimates of model-adjusted risks, risk differences, and risk ratios are obtained as functions of average marginal predictions, as originally defined by Graubard and Korn (10), from the fitted logistic regression model. Average marginal predictions allow comparisons of predicted outcomes (risk) between groups of people in the population, after controlling for differences in covariate distributions between the groups. Graubard and Korn (10, 11) show how variances of average marginal predictions in complex sample surveys can be estimated via Taylor linearization or a replication method—the most popular being jackknife and balanced repeated replication. The regression model can contain both continuous and categorical covariates, as well as interaction terms.

We demonstrate the approach using data from the 2006 NHIS (12). We use SUDAAN (Research Triangle Institute, Research Triangle Park, North Carolina) (13), a commercial software package designed specifically for analysis of data from complex sample surveys and other cluster-correlated study designs, to fit the logistic regression model and obtain parameter estimates, standard errors, confidence intervals, and *P* values for the contrasts of interest.

ESTIMATION

Model-adjusted risks

In the regression setting, we are interested in making comparisons of predicted outcomes (risk) between groups of individuals in the population, after controlling for differences in covariate distributions between the groups. Below we illustrate the estimation of model-adjusted risk discussed by Korn and Graubard (11)—the *average marginal prediction*—in the context of the logistic regression model. We then derive the risk ratio and risk difference as simple functions of the risk. Note that Korn and Graubard's method for estimating model-adjusted risk can be applied to both categorical and continuous covariates; for the sake of simplicity, however, we present estimation of average marginal predictions in the context of a categorical group variable with R categories.

Suppose a sample of size n is obtained from a population-based sample survey. Let Y_i $(i = 1, \ldots, n)$ be the binary response variable of interest. The logistic model with R groups, p-dimensional vectors of covariates x_i, and regression coefficients β is as follows:

$$E\left[\log\frac{\pi_i}{1-\pi_i}\right] = \alpha_1 I_{i1} + \alpha_2 I_{i2} + \ldots + \alpha_R I_{iR} + \beta' x_i, \quad (1)$$

where $I_{ir} = 1$ if the ith observation is in group r and 0 otherwise and $\pi_i = \text{prob}(Y_i = 1)$. The weighted maximum likelihood estimators of the regression coefficients are $\hat{\alpha}_r$ and $\hat{\beta}$.

Korn and Graubard (11) define the average marginal prediction, M_r, for the probability that $Y_i = 1$ for an observation in group r with covariate values x_i as follows:

$$M_r = \frac{\sum_{i=1}^{n} w_i\left\{\exp\left(\hat{\alpha}_r + \hat{\beta}' x_i\right)/\left[1 + \exp\left(\hat{\alpha}_r + \hat{\beta}' x_i\right)\right]\right\}}{\sum_{i=1}^{n} w_i}, \quad (2)$$

where w_i is the sample weight of the ith observation. In linear regression models, the average marginal predictions are referred to as adjusted treatment means. In addition, the average marginal prediction in equation 2 is the standardized (population-averaged) risk from the logistic model described by Greenland (1), adapted to complex sample surveys. In the sample survey setting, standard errors for the average marginal predictions are obtained using linearization or a replication method, such as jackknife or balanced repeated replication (9–11).

Let the vector $\boldsymbol{M}$ contain the estimates of the average marginal prediction for each level of a categorical study group variable, written as $\boldsymbol{M}' = [M_1, M_2, \ldots, M_R]$. Alternatively, a continuous variable could be substituted for the categorical group variable, and the R values in $\boldsymbol{M}$ would correspond to specified values of the continuous variable—for example, 20, 30, 40, 50, and 60 years of continuous age.

Model-adjusted risk differences

Contrasts among average marginal predictions, an important special case being risk differences, are calculated by defining a contrast D as the product of a user-defined contrast coefficient vector $\boldsymbol{C}$ and the average marginal prediction vector $\boldsymbol{M}$, as follows: $D = \boldsymbol{C}'\boldsymbol{M}$. The estimated variance of the contrasted average marginal predictions is then $\text{Var}(D) = \text{Var}(\boldsymbol{C}'\boldsymbol{M}) = \boldsymbol{C}'\text{Var}(\boldsymbol{M})\boldsymbol{C}$, where $\text{Var}(\boldsymbol{M})$ is the estimated variance-covariance matrix of $\boldsymbol{M}$ based upon one of the survey design-based methods.

The test statistic for evaluating the null hypothesis that the contrast equals zero is $t = D/\sqrt{\text{Var}(D)}$. For computing the significance level, the observed value of t is compared with a t distribution with degrees of freedom specific to the variance estimation method. Note that only single-df contrasts of average marginal predictions are demonstrated here (in particular, risk differences); however, contrasts among average marginal predictions could easily be extended to obtain multiple-df tests.

Model-adjusted risk ratios

The model-adjusted risk ratio based on average marginal predictions is a function of 2 elements of the $\boldsymbol{M}$ vector, $RR_{s/t} = M_s/M_t$, where the subscripts s and t represent 2 different values of the study variable. Confidence intervals for the model-adjusted risk ratios can be obtained using $\text{Var}(\boldsymbol{M})$, a survey design-based estimate of the variance-covariance matrix of $\boldsymbol{M}$, along with a Taylor series

transformation (also known as the delta method) to approximate the variance of $\log(RR_{s/t})$. Confidence intervals are first calculated on the log scale, and the endpoints are then exponentiated to generate a confidence interval for the risk ratio.

ILLUSTRATIVE EXAMPLE

In this example, we analyze data from the 2006 NHIS (12) to determine for white adults the age-specific prevalence of not being able to afford prescription medications in the past year, controlling for sex, region of the country, education, and marital status. For the sake of generality, the terms *prevalence* and *risk* will be used interchangeably. The analysis contained in this example is not intended to be a thorough epidemiologic investigation of a specific research question; rather, it is intended as an illustration of the basic techniques described above.

The NHIS is an annual multipurpose health survey conducted by the National Center for Health Statistics. It uses area probability sampling of housing units with face-to-face interviews in the home. Information is obtained about the health and other characteristics of each member of the household. In addition, 1 sample adult is selected via probability sampling from each family to answer several supplemental questions, including the affordability of prescription medications. The 2006 NHIS collected data on 24,275 sample adults aged 18 years or older.

Each sample adult was asked, ''During the past 12 months, was there any time when you needed prescription medicine but didn't get it because you couldn't afford it?'' Possible answers are yes, no, don't know, refused, and not ascertained. Only 0.96% of sample adults were coded as responding with something other than yes or no. The variable CANTAFMEDS is coded as 1 = yes (could not afford needed prescription medication at least once in the past 12 months) or 0 = no (event did not occur).

This example uses the LOGISTIC procedure in SUDAAN to model the probability that the dependent variable CANTAFMEDS is equal to 1 as a function of the set of independent variables. For variance estimation purposes, the complex sampling plan is described as 300 pseudostrata with 2 pseudo-primary sampling units per stratum. Each unit of analysis, a sample adult, is clustered within his/her primary sampling unit. The denominator df is 300 (i.e., 600 primary sampling units minus 300 strata).

The stratification and primary sampling unit variables on this file are named STRAT_P and PSU_P, respectively. The weight variable for the sample adult file is WTFA_SA. The survey design-based variance estimation method is Taylor linearization.

There are 5 independent variables in the model (all modeled as categorical): sex (1 = male, 2 = female); age (1 = 25–44 years, 2 = 45–64 years, 3 = ≥65 years); education (1 = high school or less, 2 = some college, 3 = college graduate); region of the United States (1 = Northeast, 2 = Midwest, 3 = South, 4 = West); and marital status (1 = married, 2 = widowed, 3 = unmarried), where unmarried includes never married, divorced, separated, and living as married.

The subpopulation of interest is defined as white and at least 25 years old (defined using the variables MRACRPI2 for race and AGE_P for age). Adults aged 18–24 years were eliminated from the analysis because they often have health insurance coverage under their parents' policies. The number of subjects defined by the subpopulation is 16,469. Of these, there were 427 observations, or 2.6% of the subpopulation observations, with missing values on 1 or more variables that did not get used in the regression analysis. Assuming that this latter group of observations is missing completely at random, the results of the logistic regression analysis can be generalized to the population of white adults aged 25 years or more in the civilian, noninstitutionalized household population.

Shown below are the basic SUDAAN statements used to fit the logistic regression model on the subpopulation of interest and obtain the model-adjusted risks, risk ratios, and risk differences:

```
PROC LOGISTIC;
NEST STRAT_P PSU_P;
WEIGHT WTFA_SA;
SUBPOPN AGE_P>24 AND MRACRPI2=1 / NAME=
  "WHITES AGE 25+";

CLASS SEX AGE EDUC REGION MARRY;
MODEL CANTAFMEDS = SEX AGE EDUC REGION
  MARRY;

PREDMARG AGE / ADJRR;
PRED_EFF AGE=(1 0 -1) / NAME="25-44 vs.
  65+";

PRED_EFF AGE=(0 1 -1) / NAME="45-64 vs.
  65+";
```

The PREDMARG statement requests the average marginal prediction (model-adjusted risk) for each level of age. For a given level of the age variable, the fitted model is used to predict the probability of the event for each person as if all subjects were from the same age group, while the person's actual covariate values (except for age) are used in the fitted prediction model. Then the weighted mean (using WTFA_SA) of the predicted probabilities yields the average marginal prediction for the given level of age. Thus, the model-adjusted risk for each age group is predicted as if the covariates other than age were distributed in that age group as they are distributed in the total population. The ADJRR option on the PREDMARG statement computes the model-adjusted risk ratio for each age group in comparison with the reference cell (age ≥65 years). The PRED_EFF statement performs pairwise comparisons (model-adjusted risk differences) among the 3 levels of age, based on the average marginal predictions.

Results from this analysis show that the model fits well ($P = 0.9248$, Hosmer-Lemeshow goodness-of-fit chi-squared test in SUDAAN LOGISTIC). The main effect of age group, conditional on all other variables in the model, is statistically significant ($P < 0.0001$ via a Wald F test with 2 numerator df and 300 denominator df). The effects of sex, education, region, and marital status are also statistically

Table 1. Estimated Percentage of White Adults Over Age 25 Years Who Were Not Able to Afford Their Prescription Medication in the Past Year, by Age Group, National Health Interview Survey, 2006[a]

Age Group, years	Unadjusted %[b] (SE)	Average Marginal Prediction[c] (SE)
25–44	9.27 (0.50)	9.23 (0.52)
45–64	8.35 (0.40)	8.63 (0.41)
≥65	3.05 (0.34)	2.89 (0.36)

Abbreviation: SE, standard error.
[a] Data were obtained from the National Center for Health Statistics.
[b] Percentage estimates were unadjusted for model covariates but accounted for the complex survey design (weighting, clustering, and stratification) (DESCRIPT procedure of SUDAAN).
[c] Percentage estimates were adjusted for the survey design as well as model covariates (LOGISTIC procedure of SUDAAN).

significant ($P < 0.0001$ for sex, education, and marital status and $P < 0.0015$ for region, all obtained via Wald F tests). Only the effects due to age group are further discussed here. Each of the 2 younger groups of adults has higher odds of incurring the event than those aged 65 years or older. Since the event occurs with a somewhat low probability in the specified population, estimated as 7.69% (standard error, 0.27), the odds ratio could be considered an approximate estimate of the prevalence ratio. The ''prevalence'' is the percentage of white adults in 2006 who were not able to afford needed prescription medicine at least once in the previous 12 months.

The average marginal predictions (model-adjusted risk) for each level of age are given in Table 1, with their estimated standard errors. Controlling on all other variables in the model, the estimated probability of incurring the event (being unable to afford needed prescription drugs at least once during the past 12 months) decreases with age for white adults aged 25 years or older. The estimated unadjusted prevalence percentages in Table 1 are unadjusted for model covariates; however, they still account for the complex sample design. While the average marginal predictions are close to the unadjusted prevalences, there are striking differences among the 3 age groups with regard to the percentage who incur the event.

Table 2 shows estimation of prevalence ratios (also referred to here as risk ratios) by age group, using persons aged 65 years or older as the reference group. The ratio of the marginal predictions yields a slightly smaller estimate but one that is fairly comparable to the adjusted odds ratio, based on the logistic regression analysis.

Table 2. Two Estimates of the Age-Specific Prevalence Ratio for Not Having Been Able to Afford One's Prescription Medication in the Past Year Among White Adults Over Age 25 Years, National Health Interview Survey, 2006[a]

Age Group Comparison, years	Model-Adjusted Odds Ratio[b]		Ratio of Average Marginal Predictions[b]	
	Point Estimate	95% CI	Point Estimate	95% CI
25–44 vs. ≥65	3.50	2.59, 4.73	3.19	2.40, 4.23
45–64 vs. ≥65	3.24	2.45, 4.30	2.98	2.28, 3.89

Abbreviation: CI, confidence interval.
[a] Data were obtained from the National Center for Health Statistics.
[b] Results were adjusted for the complex survey design as well as model covariates (LOGISTIC procedure of SUDAAN).

Table 3. Estimated Age-Specific Prevalence Differences for Not Having Been Able to Afford One's Prescription Medication in the Past Year Among White Adults Over Age 25 Years, National Health Interview Survey, 2006[a]

Age Group Comparison, years	Difference in Unadjusted Prevalence[b] (SE)	Difference in Average Marginal Predictions[c] (SE)
25–44 vs. ≥65	6.22 (0.62)*	6.34 (0.68)*
45–64 vs. ≥65	5.30 (0.52)*	5.73 (0.57)*

Abbreviation: SE, standard error.
* $P < 0.0001$ (2-sided P value from a t test with 300 df).
[a] Data were obtained from the National Center for Health Statistics.
[b] Results were unadjusted for model covariates but accounted for the complex survey design (weighting, clustering, and stratification) (DESCRIPT procedure of SUDAAN).
[c] Results were adjusted for the complex survey design as well as model covariates (LOGISTIC procedure of SUDAAN).

Table 3 contains the estimated unadjusted and model-adjusted prevalence differences (also referred to as risk differences) corresponding to pairwise comparisons of the 3 age groups, with persons aged 65 years or over constituting the reference group. Prevalence is first estimated from percentages unadjusted for any other covariates, then from average marginal predictions. The contrasts in Table 3 show that the younger 2 age groups differ significantly (2-sided $P < 0.0001$) from the older group (≥65 years) on both unadjusted and model-adjusted risk. The bigger difference appears to be in the youngest age category versus the oldest. Furthermore, the 2 younger age groups do not differ significantly from one another in terms of unadjusted prevalence or average marginal predictions (results not shown).

In summary, younger persons were more likely than older persons to report not being able to afford needed prescription drugs at least once in the past year, adjusted for sex, education, region, and marital status. In terms of risk ratios, persons aged 25–44 years were well over 3 times as likely to report the event as those aged 65 years or older, and persons aged 45–64 years were approximately 3 times as likely as those aged 65 years or older. Without the sample adult's health insurance status in the model as a covariate, the observed age group differences may be explained totally or partly by access of the senior population (persons aged ≥65 years) to Medicare health insurance, including Medicare's Part D prescription drug benefit program, which was first implemented in 2006.

DISCUSSION

There is increasing interest in the public health community in estimating and drawing inferences about model-adjusted risks, risk ratios, and risk differences instead of odds ratios in the binary-response regression setting (1, 2). This is no less true in the analysis of health survey data. With the public release of data sets from many national and

state health surveys over the past 25 years in the United States, many health researchers, including epidemiologists, conduct statistical analyses of these data sets, where the analytical results may have implications for identifying correlates of disease or risk behavior or for supporting new governmental health policy. Logistic regression is commonly used because the dependent variable often is dichotomous, and further, researchers desire to control for several covariates while investigating the association of a given independent variable with the dependent variable. A complicating factor is that many current public-use data files, including those from the National Health and Nutrition Examination Survey and the NHIS, are associated with national public health surveys which employ complex sampling schemes designed to yield population-based estimates and inferences and which are known to require special statistical analysis techniques (7, 8).

Our purpose in this paper was to make an available approach for estimating model-adjusted risks, risk ratios, and risk differences in the context of complex sample surveys more widely known, understood, and accessible. All parameter estimates were obtained directly from logistic regression models, as functions of average marginal predictions defined by Graubard and Korn (10). We used SUDAAN, a commercial software package (13) designed specifically for analysis of data from complex surveys, to fit the logistic model and to obtain parameter estimates, standard errors, confidence intervals, and significance levels for the contrasts of interest.

The use of the logistic regression model to estimate model-adjusted risks, risk differences, and risk ratios has many benefits. First, it avoids the problems of nonconvergence of the log-binomial, binomial, and modified Poisson models, regardless of the type of covariate. Also, the method is based on published and well-accepted procedures for computing predictive margins for logistic models estimated from survey data (10, 11). Further, it is currently implemented in at least 1 commercial software package. Another important point is that estimates of model-adjusted risks from the logistic model are bounded by 0 and 1, unlike the modified Poisson and modified least-squares linear regression models. This is important for analysis of complex survey data, where the goal of many analyses is to produce descriptive statistics for the population, which means that estimating the model-adjusted risk or prevalence is often just as important as subsequent estimation of risk ratios and differences. There are no restrictions on the types of covariates or terms in the model—it can contain both continuous (e.g., linear age) and/or categorical covariates, main effects as well as interaction terms. Risk ratios and differences can be defined for categorical as well as continuous and ordinal independent variables, as long as there are at least 2 specific values of interest from which to form comparison groups. Finally, focusing exclusively on the logistic model and simple functions of estimated predictive margins helps researchers avoid the need to change distribution and link function parameters when estimating risk ratios versus risk differences, as well as situations where the log-binomial and binomial models fail to converge.

The method is easily extended to multinomial logistic models with either cumulative logit (proportional odds model) or generalized logit link functions. In the multinomial logistic model, predictive margins can be used to estimate population risk for each category of the response. The multinomial logistic models (cumulative logit and generalized logit link functions) are also available in SUDAAN.

Limitations and assumptions behind the estimation of average marginal predictions from logistic models in sample surveys are fully described by Graubard and Korn (10). Briefly, the linearization and replication variance methods perform well when the denominator degrees of freedom (number of primary sampling units minus the number of survey strata) are at least 20; further, the model is assumed to be correctly specified (no missing covariates and covariates in the correct functional form) and does not include covariates on the causal pathway. Finally, it is assumed that any missing data are missing completely at random.

This paper was intended to be methodological in content. Although the example uses data on cost-related nonadherence (CRN) to medication protocols from the 2006 NHIS to illustrate the statistical technique, a more substantive examination of the important issue of CRN and its determinants can be found in the growing body of literature on this subject. Much of the CRN literature is focused on the elderly and disabled Medicare beneficiary populations. Briesacher et al. (14) recently reviewed the literature on patients at risk for CRN and concluded that in addition to the presence of prescription drug coverage (including enrollment in Medicare's Part D prescription drug plan), other factors such as the presence of chronic disease, multiple morbidity, mood disorders (e.g., depression), doctor-patient relationships, and patient demographic factors may either contribute to or moderate the risk of CRN. These concepts are further elaborated elsewhere (15–19). For example, Madden et al.'s (15) population-based study of elderly and disabled Medicare recipients, carried out between 2005 and 2006 (before and after the first year of implementation of Medicare Part D), showed that while the odds of CRN were decreased following implementation of Medicare Part D in the study population as a whole, there was no corresponding decrease in CRN among the sickest beneficiaries. Further, Piette et al. (16) determined that patient-physician trust and other noncost factors can moderate the impact of financial pressures on patients' adherence behavior.

ACKNOWLEDGMENTS

Author affiliations: Statistics and Epidemiology Unit, RTI International, Research Triangle Park, North Carolina (Gayle S. Bieler, G. Gordon Brown, Rick L. Williams); and Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, Georgia (Donna J. Brogan).

No external funding was obtained for this research. RTI International is the developer of SUDAAN software.

The authors thank Dr. Barry Graubard for consultation during the preparation of the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am J Epidemiol*. 2004;160(4):301–305.
2. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761–768.
3. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *Am J Epidemiol*. 2005;162(3):199–200.
4. Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702–706.
5. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Vol 1. Berkeley, CA: University of California Press; 1967:221–233.
6. Cheung YB. A modified least-squares regression approach to the estimation of risk difference. *Am J Epidemiol*. 2007;166(11):1337–1344.
7. Brogan D. Software for sample survey data: misuse of standard packages. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. 2nd ed. Chichester, United Kingdom: John Wiley & Sons Ltd; 2005:5057–5064.
8. Brogan D. Sampling error estimation for survey data. (Chapter XXI and annex). In: Yansaneh IS, Kalton G, eds. *Household Sample Surveys in Developing and Transition Countries.* (Studies in methods, series F, no. 96). New York, NY: United Nations; 2005: 447–490. (http://unstats.un.org/unsd/HHsurveys/pdf/Chapter_21.pdf, http://unstats.un.org/unsd/HHsurveys/pdf/Annex_CD-Rom.pdf). (Accessed March 1, 2009).
9. Wolter KM. *Introduction to Variance Estimation*. 2nd ed. New York, NY: Springer-Verlag; 2007.
10. Graubard BI, Korn EL. Predictive margins with survey data. *Biometrics*. 1999;55(2):652–659.
11. Korn E, Graubard B. *Analysis of Health Surveys*. New York, NY: John Wiley and Sons, Inc; 1999.
12. National Center for Health Statistics, Centers for Disease Control and Prevention. *National Health Interview Survey (NHIS). 2006 Data Release* [data file and documentation]. Hyattsville, MD: National Center for Health Statistics; 2007. (http://www.cdc.gov/nchs/nhis/nhis_2006_data_release.htm). (Accessed June 11, 2008).
13. Research Triangle Institute. *SUDAAN Language Manual, Release 10.0*. Research Triangle Park, NC: Research Triangle Institute; 2008.
14. Briesacher BA, Gurwitz JH, Soumerai SB. Patients at-risk for cost-related medication nonadherence: a review of the literature. *J Gen Intern Med*. 2007;22(6):864–871.
15. Madden JM, Graves AJ, Zhang F, et al. Cost-related medication nonadherence and spending on basic needs following implementation of Medicare Part D. *JAMA*. 2008;299(16):1922–1928.
16. Piette JD, Heisler M, Krein S, et al. The role of patient-physician trust in moderating medication nonadherence due to cost pressures. *Arch Intern Med*. 2005;165(15):1749–1755.
17. Bambauer KZ, Safran DG, Ross-Degnan D, et al. Depression and cost-related medication nonadherence in Medicare beneficiaries. *Arch Gen Psychiatry.* 2007;64(5):602–608.
18. Piette JD, Heisler M, Wagner TH. Cost-related medication underuse among chronically ill adults: the treatments people forgo, how often, and who is at risk. *Am J Public Health.* 2004;94(10):1782–1787.
19. Soumerai SB, Pierre-Jacques M, Zhang F, et al. Cost-related medication nonadherence among elderly and disabled Medicare beneficiaries: a national survey 1 year before the Medicare drug benefit. *Arch Intern Med*. 2006;166(17):1829–1835.