Doubly Robust Testing And Estimation Of Model-Adjusted Effect-Measure Modification With Complex Survey Data

By: Hao W. Zheng, Babette A. Brumback, Xiaomin Lu, **Erin D. Bouldin**,
Michael B. Cannell, and Elena M. Andresen

Abstract

Model-based standardization enables adjustment for confounding of a population-averaged exposure effect on an outcome. It requires either a model for the probability of the exposure conditional on the confounders (an exposure model) or a model for the expectation of the outcome conditional on the exposure and the confounders (an outcome model). The methodology can also be applied to estimate averaged exposure effects within categories of an effect modifier and to test whether these effects differ or not. Recently, we extended that methodology for use with complex survey data, to estimate the effects of disability status on cost barriers to health care within three age categories and to test for differences. We applied the methodology to data from the 2007 Florida Behavioral Risk Factor Surveillance System Survey (BRFSS). The exposure modeling and outcome modeling approaches yielded two contrasting sets of results. In the present paper, we develop and apply to the BRFSS example two doubly robust approaches to testing and estimating effect modification with complex survey data; these approaches require that only one of these two models be correctly specified. Furthermore, assuming that at least one of the models is correctly specified, we can use the doubly robust approaches to develop and apply goodness-of-fit tests for the exposure and outcome models. We compare the exposure modeling, outcome modeling, and doubly robust approaches in terms of a simulation study and the BRFSS example.

Zheng, H. W., Brumback, B. A., Lu, X., **Bouldin, E. D.**, Cannell, M. B. and Andresen, E. M. (2013), Doubly robust testing and estimation of model-adjusted effect-measure modification with complex survey data. Statist. Med., 32: 673-684. doi:10.1002/sim.5532. Publisher version of record available at: https://onlinelibrary.wiley.com/doi/10.1002/sim.5532

Doubly Robust Testing And Estimation Of Model-Adjusted Effect-Measure Modification With Complex Survey Data

Hao W. Zheng
Babette A. Brumback
Xiaomin Lu
Erin D. Bouldin
Michael B. Cannell
Elena M. Andresen

First published: 26 July 2012

https://doi.org/10.1002/sim.5532

Zheng, H. W., Brumback, B. A., Lu, X. , **Bouldin, E. D.**, Cannell, M. B. and Andresen, E. M. (2013), Doubly robust testing and estimation of model-adjusted effect-measure modification with complex survey data. Statist. Med., 32: 673-684. doi:10.1002/sim.5532

Abstract

Model-based standardization enables adjustment for confounding of a population-averaged exposure effect on an outcome. It requires either a model for the probability of the exposure conditional on the confounders (an exposure model) or a model for the expectation of the outcome conditional on the exposure and the confounders (an outcome model). The methodology can also be applied to estimate averaged exposure effects within categories of an effect modifier and to test whether these effects differ or not. Recently, we extended that methodology for use with complex survey data, to estimate the effects of disability status on cost barriers to health care within three age categories and to test for differences. We applied the methodology to data from the 2007 Florida Behavioral Risk Factor Surveillance System Survey (BRFSS). The exposure modeling and outcome modeling approaches yielded two contrasting sets of results. In the present paper, we develop and apply to the BRFSS example two doubly robust approaches to testing and estimating effect modification with complex survey data; these approaches require that only one of these two models be correctly specified. Furthermore, assuming that at least one of the models is correctly specified, we can use the doubly robust approaches to develop and apply goodness-of-fit tests for the exposure and outcome models. We

compare the exposure modeling, outcome modeling, and doubly robust approaches in terms of a simulation study and the BRFSS example.

1 Introduction

Recently, Brumback *et al.* [1] presented a statistical methodology to address the question of whether the difference in risk of a cost barrier to health care between persons with and without disability differed by age category. Our colleagues at the Florida Office on Disability and Health were interested in using a population-based sample, such as the Florida Behavioral Risk Factor Surveillance System Survey (BRFSS) [2], to answer this question to demonstrate for policy makers that interventions targeting younger persons with disability are much needed and cost-effective [3]. To argue for cost-effectiveness, it is helpful to use the risk difference. For example, even if the relative risk were constant across age groups, a greater risk difference in younger persons might imply that intervening to help that group could benefit more people overall than an intervention in another equal-sized group. Brumback *et al.* [1] presented the crude (unadjusted) risk differences together with two contrasting sets of adjusted risk differences, based on two different methods for model-based standardization [4]. The first method requires a correct model for the probability of the exposure conditional on the confounders (an exposure model), and the second requires a correct model for the expectation of the outcome conditional on the exposure and the confounders (an outcome model). Referring to Table 1, the results based on the outcome modeling approach indicated that the youngest adults (aged 18–29 years) have the highest adjusted risk difference and that it is statistically significant. Thus, these results demonstrate what our colleagues hypothesized. However, the results based on the exposure modeling approach suggest that the youngest adults do not have the highest risk difference and moreover that it is not statistically different from 0. The aim of the present paper is to revisit the two sets of adjusted risk differences, this time developing and applying 'doubly robust' methods [5], [6] to determine which of the exposure or outcome models better fits our data. In the process, we aim to address gaps in the methodology and implementation of doubly robust methods for complex survey data and for testing and estimating effect modification.

Table 1. Risk difference and 95% CI for the effect of disability on cost barriers to health care by age using three approaches, with complex survey data, Florida BRFSS Survey, 2007.

Age group (years)	Crude	Exposure model	Outcome model
18–29	0.234 (0.108,0.360)	0.107 (− 0.043,0.256)	0.158 (0.023,0.293)
30–64	0.180 (0.149,0.212)	0.118 (0.080,0.156)	0.097 (0.067,0.127)
65+	0.038 (0.018,0.058)	0.021 (− 0.001,0.044)	0.032 (0.010,0.054)
χ^2_2test of effect-measure modification	30.87 ($P < 0.001$)	9.55 ($P < 0.001$)	14.38 ($P < 0.001$)

The goal of standardization [4] is to compare outcomes across two groups, after the groups have been standardized to have the same distribution of confounders. When the confounders can be represented as a single categorical variable without too many categories (e.g., representing different categories of age-by-gender), standardization of each group typically proceeds by estimating the average outcome within each category and then taking a weighted average across categories according to a 'standard' distribution. When the confounders are inherently high-dimensional, a modeling approach is needed. The 'exposure modeling' enlists a model for the probability of belonging to the group conditional on the confounders [7], [8], whereas 'outcome modeling' instead uses a model for the expectation of the outcome conditional on the group and the confounders [9]. 'Model-based standardization' [4] is thus standardization using one of these two modeling approaches; the first approach requires that the exposure model be correct, whereas the second requires that the outcome model be correct. Bieler *et al.* [10] explained how to implement the outcome modeling approach with complex survey data and SUDAAN (Research Triangle Institute, Research Triangle Park, NC, USA), whereas Brumback *et al.* [1] explained how to implement both approaches with complex survey data and SAS (SAS Institute, Inc., Cary, NC, USA).

The standard distribution we shall consider in this paper is the distribution of the confounders in the two groups combined (within a given level of the effect modifier). Using the combined distribution as the standard distribution is equivalent to comparing differences in outcomes assuming every participant in the study population belonged to one exposure group versus the other. In addition to yielding an intuitive interpretation, this choice of standard distribution is the most natural to implement. With some modification, the methods we present can also be applied to other standard distributions, such as the distribution of confounders in one exposure group.

The approaches of interest in this paper combine the exposure model and the outcome model such that they have the property of being 'doubly robust', so termed because they give consistent estimators if either the exposure model or the outcome model is correct. Scharfstein, Rotnitzky, and Robins [11] first discovered doubly robust estimators and showed how to construct them. Bang and Robins [5] provided an overview of doubly robust estimation and gave details for several types of problems. More recently, Kang and Schafer [6] discussed a variety of ways to construct doubly robust estimators in the simple context of estimating a population mean from incomplete data. In this paper, we will extend the Bang and Robins [5] approach and also the Kang and Schafer [6] ,section 3.2 approach to test and estimate effect modification with complex survey data. There have been other efforts in evaluating and improving the performance of doubly robust estimators for a wide range of data settings. Austin [12] compared doubly robust methods with alternative strategies of estimating differences in proportions. Seaman and Copas [13] investigated doubly robust estimators in the longitudinal data setting with missing response. Tchetgen Tchetgen [14] described a doubly robust method for standard logistic regression with covariates missing at random and also extended it for complex survey data. Cao *et al.* [15] proposed methods for improving the efficiency of doubly robust estimators of a population mean. Tchetgen Tchetgen and Rotnitzky [16] examined an alternative doubly robust estimator in the logistic regression context. However, we are not aware of any literature other than [14] on extending doubly robust methods for use with complex survey data, and we are not aware of literature pertaining to testing and estimating effect modification using model-based standardization with doubly robust methods.

Complex survey data are typically obtained using a multistage sampling design enlisting stratification and clustering at each stage, which typically results in unequal probability sampling. The population is first divided into primary strata and then into primary sampling units (PSUs, clusters) within each of the primary strata. PSUs are sampled from each of the primary strata at the first stage, then the process is repeated with secondary sampling units being sampled within the substrata, and so on. Sampled observations are assigned survey weights representing the inverse probability of selection. Estimators for complex survey data incorporate these survey weights, and the sampling distribution of the estimators is typically approximated as though the PSUs (first-stage clusters) are resampled with replacement from each of the primary strata. This approach to estimating the sampling distribution has good properties, provided that a large number of PSUs have been selected. Otherwise, estimates of the sampling distribution need to account for subsequent stages of the sampling design. The BRFSS survey enlists stratification, and then individuals, designated as the PSUs, are sampled with unequal probabilities. Thus, there is no clustering. Technically, the design of the BRFSS is slightly more complicated, with household as the true PSU, followed by one individual within each household being selected at random. However, the difference is negligible between the two designs because of the small likelihood that individuals within the same household would be selected if this were permitted.

The paper is organized as follows. In Section 2, we review the analyses and results of Brumback *et al.* 1. In Section 3, we develop a modified version of the doubly robust estimator due to Bang and Robins 5, as well as a modified version of the doubly robust estimator due to Kang and [6, section 3.2]. We also develop goodness-of-fit tests for the exposure and outcome models, considered separately, assuming that one of the two is correctly specified. In Section 4, we conduct simulation studies to compare the performance of the doubly robust approaches under various model specifications; we also evaluate the goodness-of-fit tests. In Section 5, we apply the new methodology to the Florida BRFSS data and compare it with the analyses of Brumback *et al.* 1; we also use the new methodology to determine which of the previous results based on either exposure modeling or outcome modeling is more plausible. Section 6 concludes with a discussion.

2 Exposure modeling versus outcome modeling using the Florida Behavioral Risk Factor Surveillance System Survey

We first define our targets of estimation, namely standardized risks within levels of an effect modifier and functions of these risks, such as the risk differences and contrasts of the risk differences. Let Y_i be a binary outcome, X_i be a binary categorical exposure of interest, M_i be an effect modifier with K levels, Z_i be a vector of confounders, and W_i be the complex survey weights. Our basic targets of estimation are the set of standardized risks $\theta(x,m)=E_{Z|M=m}E(Y|X=x,M=m,Z)$ for varying levels of x and m. We are also interested in contrasts of these risks, that is, risk differences and differences of risk differences (to measure effect modification).

Brumback *et al.* 1 described an exposure modeling approach that relies on correctness of a parametric model $\pi(x,M,Z;\alpha)$ for $P(X=x|M,Z)$. It is common to use the additive logistic regression model $\text{logit}(\pi(1,M,Z))=M\alpha_0+Z\alpha_1$. Note that for polytomous X, one could use ordinal or multinomial logistic regression models. The parameter α is estimated using weighted logistic regression with the complex survey weights W_i (e.g., one could use SAS PROC SURVEYLOGISTIC), and then estimates $\hat{\theta}_e(x,m)$ of $\theta(x,m)$ are obtained using weighted linear regression with weights $W_i/\pi(X_i, M_i, Z_i; \hat{\alpha})$ and covariates X, M, and D, where D is a vector of dummy variables representing the interaction between X and M. The covariance matrix of the standardized risks can be computed using the sandwich estimator implemented in, for example, SAS PROC SURVEYREG, although these ignore estimation of α. Alternatively, one can use a nonparametric bootstrap, resampling PSUs with replacement within primary strata. Then, either that bootstrap or the delta method can be used to estimate covariances of contrasts of functions of the standardized risks. As discussed in the previous section, in the BRFSS, the individual is the PSU, and the number of PSUs per stratum is very large. With some other surveys, for example the National Health Interview Survey, there are only two PSUs per stratum. In those cases, simple modifications to the bootstrap [17], [18] are effective.

The outcome modeling approach described by Brumback *et al.* [1] relies on correctness of a parametric model $f(X,M,Z;\beta)$ for $E[Y\,|\,X,M,Z]$. In general, we recommend using a generalized linear model with canonical link, including all main effects for X and M, all interaction terms between X and M (via D), and additional covariates that involve Z. Brumback *et al.* [1] used the model $\operatorname{logit}(f(X,M,Z)) = X\beta_1 + M\beta_2 + D\beta_3 + Z\beta_4$. The parameter β is estimated using weighted regression with weights W_i, and then estimates $\hat{\theta}_o(x,m)$ of $\theta(x,m)$ are constructed as

$$\frac{\sum_{i:M_i=m} W_i \hat{f}(x,m,Z_i)}{\sum_{i:M_i=m} W_i}, \quad (1)$$

where $\hat{f}(x,m,Z_i)$ is $f(x,m,Z_i)$ with $\hat{\beta}$ in place of β. As described for the exposure modeling approach, the nonparametric bootstrap can be used to estimate the covariance of the standardized risks and covariances of their contrasts.

In the Florida BRFSS analysis, Y_i is the presence or absence of a cost barrier to health care, X_i is disability status, M_i is age, and Z_i includes race/ethnicity, income, and education. Further, M has three levels and X two; thus, we are interested in six standardized risks. We are also interested in the three risk differences $\theta(1,m) - \theta(0,m)$, $m = 1,2,3$, and in testing whether the measures of effect-measure modification $\tau_1 = (\theta(1,3) - \theta(0,3)) - (\theta(1,1) - \theta(0,1))$ and $\tau_2 = (\theta(1,2) - \theta(0,2)) - (\theta(1,1) - \theta(0,1))$ are both equal to 0. Note that we could instead focus on relative risks such as $\theta(1,m) / \theta(0,m)$ or on odds ratios, but risk differences are of primary interest to our collaborators given their focus on cost-effectiveness in public health practice.

Table [1] shows the estimated risk differences, as well as results of a chi-squared test of effect modification, that is, $\tau_1 = \tau_2 = 0$. It presents the results of implementing three approaches: the crude stratification (unadjusted), the exposure modeling approach, and the outcome modeling approach. With the exposure model, persons between the ages of 18 and 29 years do not have a greater risk difference when compared with those between 30 and 64 years. Using the outcome model, however, suggests that the risk difference is indeed higher for those between 18 and 29 years. To conclude whether our colleagues' goal of promoting interventions targeting younger persons with disability is justified, it is of interest in the present paper to determine which model fits the data better. We do this using the doubly robust approaches described in the next section.

3 Two doubly robust approaches and goodness-of-fit tests

Our modified Bang and Robins 5 doubly robust approach resembles the outcome modeling approach described in the previous section, but with the addition of a set of special covariates, one for each joint setting of (x,m). We substitute an 'augmented outcome model' $f_{\mathrm{dr}}(X,M,Z;\beta,\phi)$ in place of $f(X,M,Z;\beta)$ and then follow the outcome modeling approach. For this approach to work, one must specify $f(X,M,Z;\beta)$ as a generalized linear model with canonical link. We recommend including the covariates X, M, D, and Z, and possibly also interactions between Z and D. Then, the original model is augmented with a set of 'special covariates', one for each joint level of x and m, which we denote by $C(x,X,m,M)=I(M=m)I(X=x)/\hat{\pi}(x,m,Z)$, with $I(A=a)$ the indicator function for $A=a$, and $\hat{\pi}(x,m,Z)=\pi(x,m,Z;\hat{\alpha})$. For the BRFSS data, we use the logistic regression model $\operatorname{logit}(f_{\mathrm{dr}}(X,M,Z))=X\beta_1+M\beta_2+D\beta_3+Z\beta_4+\sum_{x,m}\phi(x,m)C(x,X,m,M)$. Let ϕ be a vector with components $\phi(x,m)$. The parameters (β,ϕ) are estimated using weighted logistic regression with the complex survey weights W_i, and then estimates $\hat{\theta}_{\mathrm{dr,\ br}}(x,m)$ of $\theta(x,m)$ are constructed as

$$\frac{\sum_{i:M_i=m}W_i\hat{f}_{\mathrm{dr}}(x,m,Z_i)}{\sum_{i:M_i=m}W_i}, \qquad (2)$$

where $\hat{f}_{\mathrm{dr}}(x,m,Z_i)$ is $f_{\mathrm{dr}}(x,m,Z_i)$ with $(\hat{\beta},\hat{\phi},\hat{\alpha})$ in place of (β,ϕ,α), where α has been estimated as described earlier for the exposure modeling approach. One can use the nonparametric bootstrap, as described in Section 2, to estimate covariances.

Why does this doubly robust approach work? It can be shown that when at least one of the two models is correct, the estimating equation that $\hat{\theta}_{\mathrm{dr,\ br}}(x,m)$ solves has expectation 0 and therefore is a consistent estimator of $\theta(x,m)$. When the augmented outcome model $f_{\mathrm{dr}}(X,M,Z)$ is correct, the doubly robust approach works for the same reason that the outcome modeling approach works. When the original outcome model $f(X,M,Z)$ is correct, then the augmented outcome model is automatically correct (let $\phi(x,m)=0$ for all (x,m)). When the original outcome model is incorrect but the exposure model $\pi(x,M,Z)$ is correct, two mathematical 'tricks' can be used to show that the estimating equation has expectation 0. The first trick is used to show that the estimator $\hat{\theta}_{\mathrm{dr,\ br}}(x,m)$ solves the estimating equation

$$\sum_i W_iI(M_i=m)\left\{\left(\hat{f}_{\mathrm{dr}}(x,m,Z_i)-\theta(x,m)\right)+\frac{I(X_i=x)}{\hat{\pi}(x,m,Z_i)}\left(Y_i-\hat{f}_{\mathrm{dr}}(x,m,Z_i)\right)\right\}=0, \qquad (3)$$

where $\hat{\pi}(x, m, Z_i)$ is $\pi(x,m,Z_i)$ with $\hat{\alpha}$ in place of α. The trick is to realize that

$$\sum_i W_i I(M_i = m) \frac{I(X_i = x)}{\hat{\pi}(x, m, Z_i)} \left(Y_i - \hat{f}_{\mathrm{dr}}(x, m, Z_i)\right) = 0 \tag{4}$$

because this is one component of the weighted estimating equations that are solved to produce $(\hat{\beta}, \hat{\phi})$, assuming that an ordinary weighted generalized linear model regression algorithm is used, as in SAS PROC GENMOD or, for a binary outcome, SAS PROC SURVEYLOGISTIC.

The second trick is to rewrite Equation 3 by adding and subtracting $W_i I(M_i = m)(I(X_i = x)/\hat{\pi}(x, m, Z_i))\theta(x, m)$ and then rearranging terms, so that the left-hand side of 3 equals

$$\sum_i W_i I(M_i = m) \left\{ \frac{I(X_i = x)}{\hat{\pi}(x, m, Z_i)}(Y_i - \theta(x, m)) - \left(\frac{I(X_i = x)}{\hat{\pi}(x, m, Z_i)} - 1\right) \left(\hat{f}_{\mathrm{dr}}(x, m, Z_i) - \theta(x, m\right.\right.$$

(5)

Note that $\hat{\alpha}$ is a consistent estimator of α if the exposure model $\pi(x,m,Z_i)$ is correct. The first term of 5 is the exposure modeling estimating equation for $\vartheta(x,m)$ and hence has mean 0 if the exposure model is correct. The second term is itself a product of two terms, the second of which is a function of Z_i only and the first of which has mean 0 conditional on Z_i and $M_i = m$ if the exposure model is correct. Therefore, Equation 5 and hence Equation 3 both have 0 mean, which implies that, subject to regularity conditions that are easily satisfied in our relatively simple setting, $\hat{\theta}_{\mathrm{dr,\ br}}(x, m)$ consistently estimates $\vartheta(x,m)$.

Instead of the outcome model being augmented with special covariates, our modified Kang and Schafer 6 approach achieves double robustness using the original outcome model but estimated with augmented weights. Specifically, we estimate β of the original outcome model $f(X,M,Z;\beta)$ using the product of the complex survey weights and an additional weight $1/\hat{\pi}(X, M, Z)$ constructed from the estimated exposure model. We assume that $f(X,M,Z;\beta)$ is a generalized linear model with canonical link. For this approach to work, one must include the covariates X, M, and D. We recommend also including Z, and possibly also interactions between Z and D. For our analysis of the BRFSS data, we did not include interactions between Z and D for the sake of parsimony and also to match the outcome model used by Brumback *et al.* 1. We estimate the parameter β by $\hat{\beta}$ using weighted generalized linear model regression with weights $W_i/\hat{\pi}(X_i, M_i, Z_i)$, where $\hat{\pi}(X_i, M_i, Z_i)$ is the estimated exposure model. Then, estimates $\hat{\theta}_{\mathrm{dr,\ ks}}(x, m)$ of $\theta(x,m)$ are constructed as

$$\frac{\sum_{i:M_i=m} W_i \hat{f}(x, m, Z_i)}{\sum_{i:M_i=m} W_i}, \tag{6}$$

where $\hat{f}(x, m, Z_i)$ is $f(x,m,Z_i)$ with $\hat{\beta}$ in place of β and $\hat{\beta}$ solves the estimating equation

$$\sum_i \frac{W_i}{\hat{\pi}(X_i, M_i, Z_i)} S_i^{\mathrm{T}}(Y_i - f(X_i, M_i, Z_i)) = 0, \tag{7}$$

where S_i is a row vector from the design matrix (e.g., for the BRFSS data, $S_i = \{1, X_i, M_i, D_i, Z_i\}$). Note that the estimating equation is a weighted version of the estimating equation for a generalized linear model with a canonical link,

$$\sum_i S_i^{\mathrm{T}}(Y_i - f(X_i, M_i, Z_i)) = 0. \tag{8}$$

When the outcome model is correct, that is, $f(X,M,Z) = E[Y \mid X,M,Z]$, then taking the iterated expectation of 7 with respect to (X_i,M_i,Z_i) shows that 7 has expectation 0 regardless of the correctness of the exposure model $\pi(X,M,Z)$. Therefore, $\hat{\beta}$ is a consistent estimator of β, and $\hat{\theta}_{\mathrm{dr,\ ks}}(x, m)$ is in turn a consistent estimator of $\theta(x,m)$, for all values of x and m. To see why the estimator remains valid when the outcome model is incorrect but the exposure model is correct, consider first a hypothetical large finite population in which a complete set of potential outcomes is observed for each individual at each level of X_i; let N denote the number of observations in that population. Then, the estimating Equation 8 implies that

$$\sum_{i=1}^{N} S_i^{\mathrm{T}} Y_i = \sum_{i=1}^{N} S_i^{\mathrm{T}} f(X_i, M_i, Z_i), \tag{9}$$

and

$$\frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(X_i, M_i, Z_i) \tag{10}$$

is a consistent estimator of the overall $E(Y)$, where $\hat{f}(X_i, M_i, Z_i)$ is $f(X_i,M_i,Z_i)$ with estimates $\hat{\beta}$ (from 9) in place of β. This estimator is consistent regardless of the correctness of the outcome model. Equation 9 also leads to

$$\frac{1}{N_{x,m}} \sum_{i:X_i=x,M_i=m} Y_i = \frac{1}{N_{x,m}} \sum_{i:X_i=x,M_i=m} \hat{f}(X_i, M_i, Z_i), \tag{11}$$

where $N_{x,m}$ is the total number of participants with $X = x$ and $M = m$ in the hypothetical population. Therefore, $(1/N_{x,m}) \sum_{i:X_i=x,M_i=m} \hat{f}(X_i, M_i, Z_i)$ is a consistent estimator of $\vartheta(x,m)$. For a simple random sample, only one outcome is observed for any individual, corresponding to the individual's actual level of X_i. By correctly modeling and consistently estimating the exposure probability, we can use

$$\sum_{i=1}^{N} \frac{1}{\hat{\pi}(X_i, M_i, Z_i)} S_i^{\mathrm{T}}(Y_i - f(X_i, M_i, Z_i)) = 0 \quad (12)$$

to construct a consistent estimator of $\vartheta(x,m)$. The weighting term $1/\hat{\pi}(X_i, M_i, Z_i)$reweights the observed population to resemble the hypothetical population with a complete set of potential outcomes for each individual at each level of X_i. In complex survey settings, we must further augment 12 with complex survey weights W_i to account for varying sampling probabilities of individuals; this leads to estimating Equation 7. Therefore, estimating Equation 7 also results in a consistent estimator 6 of $\vartheta(x,m)$ when the exposure model has the correct form but the outcome model is specified incorrectly.

We now describe the test for the presence of effect modification. Let τ be a column vector of linearly independent measures of effect-measure modification. For assessment of the differences of risk differences, τ has entries of the form $(\theta(1,m_s) - \theta(0,m_s)) - (\theta(1,m_t) - \theta(0,m_t))$, for m_s and m_t, two different levels of the effect modifier. Let $\hat{\tau}$be the estimated version of τ using a doubly robust model. To test whether $\tau = 0$, we use the test statistic $T_{\mathrm{em}} = \hat{\tau}^{\mathrm{T}} V_{\mathrm{em}}^{-1} \hat{\tau}$, where V_{em} is the estimated covariance matrix of $\hat{\tau}$obtained from the nonparametric bootstrap for complex survey data, described in Section 2. Specifically, doubly robust estimation is applied to the bootstrap samples, and bootstrap estimates $\{\hat{\tau}_b, b = 1, \ldots, N_{\mathrm{b}}\}$are obtained, where N_{b} is the number of bootstrap samples. V_{em} is then calculated as the sample covariance matrix of $\{\hat{\tau}_b\}$. Under the null hypothesis of no effect-measure modification, the test statistic T_{em} is distributed as χ^2 with degrees of freedom equal to the dimension of τ.

Next, we turn our attention to goodness-of-fit tests of the exposure and outcome models, assuming that at least one of the models is correct. Let $\hat{\psi}_{\mathrm{o}} = \{\hat{\theta}_{\mathrm{o}}(1, k) - \hat{\theta}_{\mathrm{o}}(0, k), k = 1, \ldots, K\}$be a column vector of estimated risk differences at K levels of M, using the outcome modeling approach. Let $\hat{\psi}_{\mathrm{e}}$and $\hat{\psi}_{\mathrm{dr}}$be similarly defined, for the exposure modeling approach and a doubly robust approach, respectively. To test the outcome and exposure models against a doubly robust model, we use $T_{\mathrm{o}} = (\hat{\psi}_{\mathrm{o}} - \hat{\psi}_{\mathrm{dr}})^{\mathrm{T}} V_{\mathrm{o}}^{-1} (\hat{\psi}_{\mathrm{o}} - \hat{\psi}_{\mathrm{dr}})$and $T_{\mathrm{e}} = (\hat{\psi}_{\mathrm{e}} - \hat{\psi}_{\mathrm{dr}})^{\mathrm{T}} V_{\mathrm{e}}^{-1} (\hat{\psi}_{\mathrm{e}} - \hat{\psi}_{\mathrm{dr}})$, respectively, where V_{o} and V_{e} are the sample covariance matrices of bootstrap estimates $\{(\hat{\psi}_{\mathrm{o},b} - \hat{\psi}_{\mathrm{dr},b}), b = 1, \ldots, N_{\mathrm{b}}\}$and $\{(\hat{\psi}_{\mathrm{e},b} - \hat{\psi}_{\mathrm{dr},b}), b = 1, \ldots, N_{\mathrm{b}}\}$. Under the null hypothesis of no lack of fit, the test statistics T_{o} and T_{e} are each distributed as χ^2 with degrees of freedom equal to K.

Table 2. Simulation results of estimating effect modification under four model specifications using modified Bang and Robins doubly robust estimation.

Scenario	**Risk difference for *M* = 1**	**Risk difference for *M* = 0**	**Effect modification by *M***	**95% CI coverage for *M* = 1 (%)**	**95% CI coverage for *M* = 0 (%)**	**95% CI coverage for effect mod. (%)**
True	0.392	0.232	0.160	N/A	N/A	N/A

Scenario	Risk difference for *M* = 1	Risk difference for *M* = 0	Effect modification by *M*	95% CI coverage for *M* = 1 (%)	95% CI coverage for *M* = 0 (%)	95% CI coverage for effect mod. (%)
Both outcome and exposure correct	0.390	0.231	0.158	94	94	94
Outcome correct, exposure incorrect	0.391	0.229	0.162	95	94	93
Exposure correct, outcome incorrect	0.385	0.230	0.155	93	97	94
Both outcome and exposure incorrect	0.419	0.238	0.180	88	93	92

We illustrate these tests in our simulation study, which follows, and also in a re-analysis of the BRFSS data.

4 Simulation study

We conducted a simulation study to investigate performance of the methods under four different model specifications: both models correct, outcome model correct but exposure model incorrect, exposure model correct but outcome model incorrect, and neither model correct. We assessed bias as well as coverage of 95% CI under each specification, and we also evaluated performance of our tests of model specification.

The correct population exposure model is logit $(P(X=1\mid M,Z))=0.5-1M+1Z+0.5MZ$, and the correct population outcome model is logit $(P(Y=1\mid X,M,Z))=-0.5+1X-0.5M+1XM+0.5Z+1XMZ$, where X is the exposure, M is a binary effect modifier distributed as Bernoulli(0.4), and Z is a confounder distributed as $N(0,1)$, independent of M. We used Monte Carlo integration to obtain true values of parameters of interest. For the outcome model, the incorrect specification is logit $(P(Y=1 \mid X,M,Z))=\beta_0+\beta_1 X+\beta_2 M+\beta_3 XM$, and for the exposure model, the incorrect specification is logit $(P(X=1 \mid M,Z))=\alpha_0+\alpha_1 Z$.

To simulate complex survey data, we first simulated an independent and identically distributed sample of size 3500 from the correct population models. Then, if $Y_i=X_i$, the observation was kept with probability 0.5 and was assigned a survey weight of $W_i=2$. Otherwise, the observation was kept with probability 1 and was assigned a survey weight of $W_i=1$. We estimated parameters and conducted hypothesis tests as described in the preceding section. For investigating estimation of the parameters, we used the nonparametric bootstrap with 100 samples for inference, and we simulated 500 datasets. For investigating the performance of our hypothesis tests, we used the nonparametric bootstrap with 500 samples for inference, and we simulated 200 datasets.

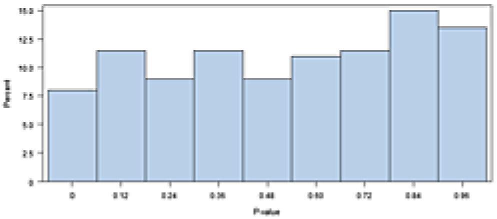
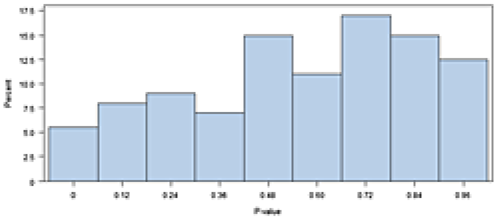
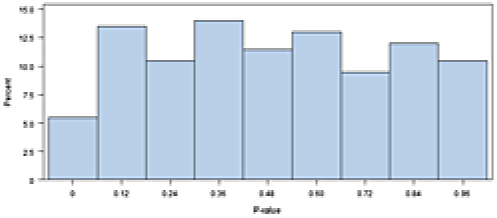
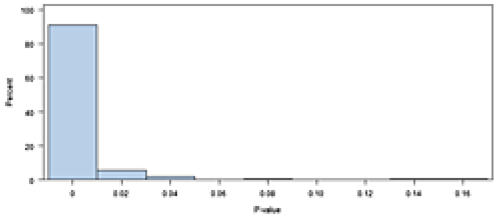
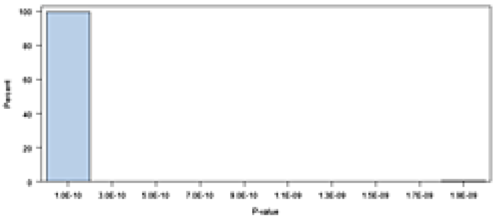
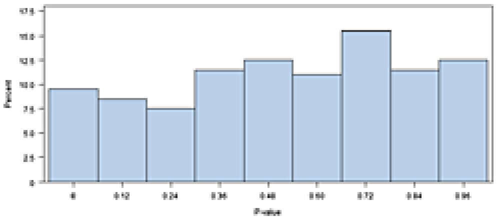
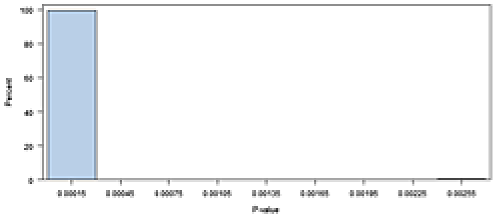
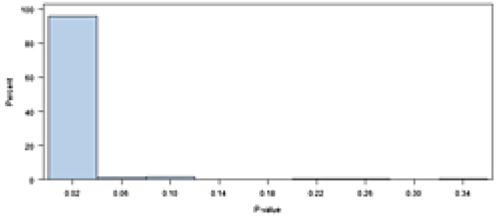
Table 2 presents the results of estimating the risk differences for $M=1$ and $M=0$ as well as estimating a measure of effect modification (the difference of the risk differences) using the modified Bang and Robins approach. The true values are presented in the first row. When at least one model is correctly specified, biases are slight, and coverages are close to 95%. When both models are incorrectly specified, the bias increases, and the coverage degrades, particularly for $M=1$. However, the coverage probability for the measure of effect modification (92%) does not deviate too much from 95%. Table 3 presents the results of estimation using the modified Kang and Schafer approach. Just as in Table 2, when at least one model is correctly specified, estimation results are very good. However, when both models are incorrectly specified, the modified Kang and Schafer method performs poorly, with 34% and 47% coverage probabilities for $M=1$ and the measure of effect modification, respectively. This is not surprising because when both models are incorrectly specified as before, the Kang and Schafer approach is identical to the exposure modeling approach with the incorrect exposure model. For comparison, we also applied the following: (i) the exposure modeling approach using the incorrect exposure model and (ii) the outcome modeling approach using the incorrect outcome model. For $M=1$, the estimated risk differences were 0.471 and 0.585, respectively, with 95% CI coverage of 32% and 0%, respectively. For $M=0$, the estimated risk differences were 0.215 and 0.323, respectively, with 95% CI coverage of 92% and 6%, respectively. For the differences of the two risk differences, the estimates were 0.257 and 0.261, respectively, with 95% CI coverage of 42% and 16%, respectively. We note that the estimated risk differences from both methods (i) and (ii) for $M=1$ and $M=0$ differ substantially from those obtained with the modified Bang and Robins approach when both models are incorrectly specified. However, the estimated risk differences from method (i) are essentially identical to, and from method (ii) differ substantially from, those obtained with the modified Kang and Schafer approach when both models are incorrectly specified.

Table 3. Simulation results of estimating effect modification under four model specifications using modified Kang and Schafer doubly robust estimation.

Scenario	Risk difference for *M* = 1	Risk difference for *M* = 0	Effect modification by *M*	95% CI coverage for *M* = 1 (%)	95% CI coverage for *M* = 0 (%)	95% CI coverage for effect mod. (%)
True	0.392	0.232	0.160	N/A	N/A	N/A
Both outcome and exposure correct	0.384	0.230	0.154	92	96	94
Outcome correct, exposure incorrect	0.393	0.230	0.163	93	96	93
Exposure correct, outcome incorrect	0.391	0.230	0.161	95	95	95
Both outcome and exposure incorrect	0.471	0.216	0.256	34	93	47

Table 4 presents the performance of our tests of model specification, in terms of the distribution of the 200 *P*-values and the mean of the χ^2_2 test statistic, using the modified Bang and Robins approach. Note that when the null hypothesis is true, the distribution should be approximately Uniform(1), and the mean of the test statistic should be approximately 2.0. The first column shows that the *P*-values for the test of the outcome model follow the Uniform(1) distribution, and the mean of the test statistic is approximately 2.0 when both models are correct or when the outcome model is correct but the exposure model is incorrect; when the outcome model is incorrect, the *P*-values are quite small, and the mean of the test statistic equals 129.5 when the exposure model is correct and 95.2 when the exposure model is incorrect. The second column shows similar results for the test of the exposure model. Particularly noteworthy are the results of the test when both outcome and exposure models are specified incorrectly. In this scenario, we would not necessarily expect the hypothesis testing methods to correctly reject the null hypotheses; however, both tests reject almost all of the time (percent rejected is 100% for the test of the outcome model and 96% for the test of the exposure model). For comparison, we also assessed the performance of our tests of model specification using the modified Kang and Schafer approach when both models are incorrectly specified as before. Under these conditions, the modified Kang and Schafer approach is identical to the exposure modeling approach. Therefore, for the test of the exposure model, the test statistic equals 0 each time, and the percent rejected is 0%. For the test of the outcome model, however, the mean of the test statistic equals 83.0, and the percent rejected is 100%; this reflects the substantial discrepancy between the estimates presented earlier from the modified Kang and Schafer approach when both models are incorrectly specified and those obtained from the outcome modeling approach when the outcome model is incorrectly specified.

Table 4. *P*-value distribution of goodness-of-fit test under four model specifications using modified Bang and Robins doubly robust approach.

Scenario	**Test of Outcome Model**	**Test of Exposure Model**
1.	 Mean of test statistic: 2.1 Percent rejected: 7.0%	 Mean of test statistic: 1.7 Percent rejected: 5.0%
2.	 Mean of test statistic: 2.0 Percent rejected: 4.5%	 Mean of test statistic: 20.6 Percent rejected: 98.5%
3.	 Mean of test statistic: 129.5 Percent rejected: 100%	 Mean of test statistic: 2.2 Percent rejected: 9%
4.	 Mean of test statistic: 95.2 Percent rejected: 100%	 Mean of test statistic: 21.2 Percent rejected: 96.0%

- Scenarios: 1, both outcome and exposure correct; 2, outcome correct, exposure incorrect; 3, exposure correct, outcome incorrect; 4, both outcome and exposure incorrect.

The simulation studies showed that when at least one of the two models is specified correctly, the modified Bang and Robins approach and the modified Kang and Schafer approach perform similarly well. The estimates have little bias, and the 95% CIs have close to 95% coverage. Interestingly, for our simulated data in which neither the exposure model nor the outcome model is correct, the modified Bang and Robins approach performs significantly better than the modified Kang and Schafer approach. This stands in contrast to the results presented by Kang and Schafer 6 for their simulation settings, albeit in a much simpler context. For our simulation settings, our results overcame a serious limitation of doubly robust methods. When neither the exposure model nor the outcome model is correct, the behavior of doubly robust models is unpredictable. Under these circumstances, a doubly robust model is not a reliable gauge for how well the exposure and outcome models fit the data. Nevertheless, for our simulation settings, our hypothesis tests based on the modified Bang and Robins approach correctly detected that neither model was correct when in fact neither model was correct. Those based on the modified Kang and Schafer approach could not reject the exposure model but do detect that the mean model is incorrect. As documented by Kang and Schafer 6, using a doubly robust model with two incorrect models does not necessarily provide improvement in point estimation over using either single incorrect model.

5 Application to the Florida Behavioral Risk Factor Surveillance System Survey data

Tables 5 and 6 present the results of applying our modified Bang and Robins 5 approach and our modified Kang and Schafer [6, section 3.2] approach, respectively, to the 2007 Florida BRFSS Survey data. For comparison, Table 1 presents the results from Brumback *et al.* 1. The outcome represents whether a survey participant could afford to visit a doctor in the past year. The exposure is disability status, determined for each respondent by the definition used by the Centers for Disease Control and Prevention. The modifier is age, categorized into three groups: 18–29, 30–64, and 65 years or older. The confounders are race/ethnicity categorized into four groups (non-Hispanic White; non-Hispanic Black; Hispanic of any race; others), annual household income categorized into five groups (less than 20,000;

20,000–24,999; 25,000–34,999;

35,000–49,000; $50,000 or more), and education categorized into four groups (less than high school; high school graduation or equivalent; some college; college degree or higher). Respondents with any missing data were excluded from the analysis. Total sample size was 31,590.

Table 5. Testing and estimating effect-measure modification by age of the adjusted risk difference for the effect of disability on cost barriers to health care, using modified Bang and Robins doubly robust estimation with complex survey data, Florida Behavioral Risk Factor Surveillance System Survey, 2007.

Age group (years)	Risk for PWD (95% CI)	Risk for PWOD (95% CI)	Risk difference (95% CI)
18–29	0.366 (0.238, 0.494)	0.224 (0.183, 0.265)	0.142 (0.008, 0.275)
30–64	0.249 (0.216, 0.281)	0.146 (0.133, 0.160)	0.102 (0.068, 0.137)
65+	0.071 (0.052, 0.091)	0.040 (0.028, 0.051)	0.032 (0.009, 0.054)

- χ^2_2test of effect-measure modification: 13.73 (*P* < 0.001); χ^2_3test of doubly robust model versus outcome model: 0.90 (*P* = 0.82); χ^2_3test of doubly robust model versus exposure model: 21.94 (*P* < 0.001).PWD, person with disability; PWOD, person without disability.

Table 6. Testing and estimating effect-measure modification by age of the adjusted risk difference for the effect of disability on cost barriers to health care, using modified Kang and Schafer doubly robust estimation with complex survey data, Florida Behavioral Risk Factor Surveillance System Survey, 2007.

Age group (years)	Risk for PWD (95% CI)	Risk for PWOD (95% CI)	Risk difference (95% CI)
18–29	0.361 (0.237, 0.486)	0.222 (0.181, 0.263)	0.140 (0.009, 0.270)
30–64	0.245 (0.213, 0.277)	0.145 (0.132, 0.159)	0.100 (0.066, 0.134)
65+	0.072 (0.052, 0.092)	0.039 (0.028, 0.051)	0.033 (0.009, 0.057)

- χ^2_2test of effect-measure modification: 11.18 (*P* = 0.004); χ^2_3test of doubly robust model versus outcome model: 0.50 (*P* = 0.92); χ^2_3test of doubly robust model versus exposure model: 23.94(*P* < 0.001).PWD, person with disability; PWOD, person without disability.

The second column of Table 1 presents the results from the exposure modeling approach. There is strong evidence of modification of the risk differences by age (*P* < 0.001). The youngest adults (18–29 years) have a slightly lower risk difference than the middle-aged adults (30–64 years), whereas the oldest adults (65+ years) have the lowest risk difference; only the risk difference for the middle-aged adults is statistically significantly different from 0. The third column of Table 1 presents the results from the outcome modeling approach. There is still strong evidence of modification by age (*P* < 0.001). However, in this case, all risk differences are statistically significantly different from 0, and the risk differences trend higher as the respondents trend younger. The results from the outcome model are qualitatively similar to the crude (unadjusted) results of column one of Table 1, and they better correspond to the hypothesis of our colleagues in the Florida Office on Disability and Health than do the results from the exposure model. Therefore, we were very curious to determine which results were more plausible.

Tables 5 and 6 show that the outcome model fits the data better, assuming that at least one of the exposure or outcome model is correct. Consequently, the adjusted risk differences for both doubly robust approaches are very similar to those of the outcome modeling approach, and the *P*-values for testing the outcome model versus the doubly robust model are 0.82 and 0.92, for the modified Bang and Robins and modified Kang and Schafer approaches, respectively. On the other hand, the *P*-values for testing the exposure model versus the doubly robust model are < 0.001 for both doubly robust approaches; that is, the exposure model and ensuing results are rejected.

6 Discussion

Motivated by the research question of our colleagues in the Florida Office on Disability and Health, we have developed and applied two doubly robust approaches for testing and estimating effect-measure modification with complex survey data, and furthermore, we have constructed two hypothesis tests for determining the plausibility of the results of the exposure modeling approach and the outcome modeling approach applied previously. The results of our simulation study confirmed that the approaches work well, and surprisingly, our hypothesis testing procedure was able to detect dual-model misspecification (i.e., misspecification of both the exposure and outcome models). The results of applying both doubly robust methods to our application perfectly documented the suspicions of our colleagues that in youngest (18–29 years) adults, the effect of disability on the presence of a cost barrier to health care is the strongest, when measured by the risk difference.

We were lucky that in our application, one model appeared plausible and the other implausible. In other applications, it could happen that neither model provides a satisfactory fit to the data. It would be of further interest in future research to investigate such an application and develop diagnostics that might be useful in improving either the exposure model or the outcome model.

It is worth noting that the research was motivated by our initially unsatisfactory analyses, in which the exposure modeling and outcome modeling approaches led to different conclusions. This paper presents a case study of our research, aimed at providing a convincing statistical argument for our colleagues to achieve their goal of promoting intervention efforts targeting younger persons with disability. If one were solely interested in correctly estimating functions of risks from the beginning of the study, then the analyst could deploy a doubly robust model without regards as to which of the component models is in fact correct. On the other hand, if one were additionally interested in which of the exposure or outcome model better fits the data, then the goodness-of-fit tests are also of use. Furthermore, our simulation studies have indicated that, although they cannot be expected to perform well all of the time, the goodness-of-fit tests do have some potential as a diagnostic tool to detect dual-model misspecification. In our simulation studies, we found that the modified Bang and Robins goodness-of-fit test detected dual misspecification but that the Kang and Schafer goodness-of-fit test only detected misspecification of the exposure model. The reliability of either method for detecting dual misspecification depends on the extent of disagreement between the doubly robust estimate on the one hand and the estimate based on the exposure model or that based on the outcome model on the other hand. When the doubly robust estimate happens to agree with either of the other two, we will fail to detect dual misspecification.

Acknowledgements

This research was supported by CDC grant number U59 DD000273. We would like to thank Janet Hess, MPH, CHES at the University of South Florida; Dr. David Wood, MD, MPH at the University of Florida in Jacksonville; Dr. Phyllis Sloyer, RN, PhD, FAHM, FAAP at Children's Medical Services; and Dr. John Reiss, PhD at the University of Florida, on behalf of the Health Care Transition Services Task Force for Youth and Young Adults with Disabilities and the FloridaHATS program for their collaboration and direction on the disability research. We would also like to thank Youjie Huang and Melissa Murray from the Florida BRFSS Office for their helpful support.

References

- 1 Brumback BA, Bouldin ED, Zheng HW, Cannell MB, Andresen EM. Testing and estimating model-adjusted effect-measure modification using marginal structural models and complex survey data. *American Journal of Epidemiology* 2010; 172:1085–1091.

- 2 Remington PL, Smith MY, Williamson DF, Anda RF, Gentry EM, Hogelin GC. Design, characteristics, and usefulness of state-based behavioral risk factor surveillance: 1981–1987. *Public Health Reports* 1988; 103:366–375.

- 3 Cannell MB, Brumback BA, Bouldin ED, Hess J, Wood DL, Sloyer PJ, Reiss JB, Andresen EM. Age group differences in health care access for people with disabilities: are young adults at increased risk? *Journal of Adolescent Health* 2011; 49:219–221.

- 4 Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Lippincott Williams & Wilkins: Philadelphia, 2008.

- 5 Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; 61:962–972.

- 6 Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; 22:523–539.

- 7 Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11:550–560.

- 8 Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology* 2003; 14:680–686.

- 9 Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics* 1982; 38:613–621.

- 10 Bieler GS, Brown GG, Williams RL, Brogan DJ. Estimating model-adjusted risks, risk differences, and risk ratios from complex survey data. *American Journal of Epidemiology* 2010; 171:618–623.

- 11 Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 1999; 94:1096–1120. (with rejoinder, 1135–1146).

- 12 Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 2010; 29:2137–2148.

- 13 Seaman S, Copas A. Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine* 2009; 28:937–955.

- 14 Tchetgen Tchetgen EJ. A simple implementation of doubly robust estimation in logistic regression with covariates missing at random. *Epidemiology* 2009; 20:391–394.

- 15 Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 2009; 96:723–734.

- 16 Tchetgen Tchetgen EJ, Rotnitzky A. Double-robust estimation of an exposure–outcome odds ratio adjusting for confounding in cohort and case–control studies. *Statistics in Medicine* 2011; 30:335–347.

- 17 McCarthy PJ, Snowden CB. The bootstrap and finite population sampling. In Vital and Health Statistics 2-95, Public Health Service Publication 85-1369. US Government Printing Office: Washington, DC.

- 18 Shao J. Impact of the bootstrap on sample surveys. *Statistical Science* 2003; 18(2):191–198.