Using official surveys to reduce bias of estimates from nonrandom samples collected by web surveys

Vladislav Beresovsky[*]

[*]National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782

Abstract

Recent trends in U.S. official statistics are characterized by the rising cost of data collection and increased nonresponse. At the same time, inexpensive data from commercial nonrandom web panels have become readily available. This paper discusses the possibility of capitalizing on both sources of information by calibrating data from web panels on estimates from conventional randomized surveys. We treat the probability of being included in a web panel as similar to a response probability and use propensity score adjustment (PSA) of sampling weights and generalized calibration to produce g-weights, thus potentially making nonrandom samples representative of the general population. The simulation study discussed in this paper demonstrates that propensity score model or calibration using covariates correlated with both web sample indicator and target variable can eliminate bias of estimates from web sample. Variances estimated using a two-phase sampling approach match Monte Carlo variances of point estimators. If the web panel inclusion probability depends on the target variable Y, bias can be removed by using Y as an instrumental variable and calibrating on a closely correlated covariate. However, this approach may lead to a significant increase in variances. Conclusions of the simulation study were validated by applying the methods used in that study to a real web sample representing a subset of the National Health Interview Survey (NHIS) questions. The NHIS public-use file was used as an auxiliary for calculating estimates from the web sample data and for their subsequent validation. The g-weight adjusted estimates from the web and random NHIS samples matched within the limits of statistical significance.

Key Words: web panel, propensity score adjustment, generalized calibration, instrumental variable, informative nonresponse, two-phase sampling

Introduction

As with the trend in statistics in general, the National Center for Health Statistics (NCHS) has recently experienced rising costs and growing nonresponse on questionnaires for the National Health Interview Survey (NHIS). One of the reasons is the long questionnaire and consequent burden falling on respondents. Simultaneously, rapid proliferation of commercially maintained web panels offers an expedient and relatively inexpensive way of collecting data through web surveys. The obvious disadvantage of these data sources is that they are in many cases collected from nonrandomly selected samples and therefore may not be representative of the general population. The difficulties of producing population estimates from web survey data were reported by Chmura et al. (2013), DiSogra et al. (2011) and Dever et al. (2008).

Recently, an NCHS team of methodologists decided to test the possibility of producing national estimates for some of the NHIS variables using data collected from a web survey. One of the commercial web panel vendors was selected to collect data for a subset of NHIS questions, while the same data were simultaneously collected using the regular NHIS random sample. This NHIS questionnaire was treated as consisting of two sets of

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

questions, providing data for both *core* and *detail* variables. Core questions of a more general type serve as a gateway to a group of detailed questions, which are usually focused on more specific aspects of the studied characteristics. Since survey participants are presented with detailed questions conditionally as a result of giving certain answers to core questions, these two sets of variables must be strongly correlated. The hope is that these correlations can be utilized to produce reliable estimates of population parameters for detail variables from web panel data. Because detail variables were collected from both web and regular NHIS surveys, it was possible to use the data to verify a proposed estimation methodology.
To recruit web panelists, the vendor utilized a protocol that can be characterized as a two-phase stochastic process. Potential respondents were initially contacted using random digit dialing. A subset of the contacted people, limited by access to the Internet and email systems and other unknown factors, might agree to join a web panel. The web panel is subject to attrition as well. As a result, the distribution of web panelists by demographic and other characteristics may differ from the general population. Dever et al. (2008) argued that using random digit dialing for initial contact may provide better coverage of various demographic groups than using a purely volunteer web panel.
The vendor fielded the NHIS web survey to a subset of the available panelists specially selected using stratified sampling, so the resulting web sample was more or less balanced with respect to the general population. Because the recruitment rate of the web panel varied mostly by age, race and ethnicity, and education level, these variables were used to define stratification cells. As with regular surveys, web surveys are subject to unit and item nonresponse. For the NHIS web survey, average unit response rate was below 25% and differed greatly by the same demographic strata. The vendor supplied post-stratification weights, calculated using a set of demographic variables. These weights may provide for unbiased estimation of some of the population parameters, but this cannot be stated unequivocally because of unknown mechanisms of the panel recruitment probability and survey nonresponse. This uncertainty leaves the possibility of biased estimates and requires additional study focused on specific target variables.
Since data collection using web panels depends on unexplained recruitment and response mechanisms, it is natural to consider estimation methods developed for regular nonresponse adjustment. Little (1986) conducted simulations to compare performance of the three methods: propensity-score adjustment (PSA) of sample selection weights, imputation of the outcome variable within adjustment cells defined by the response model, and post-stratification, which is a particular case of calibration. Imputation was found to be superior to other methods because it resulted in less bias and greater efficiency. Imputation of missing data requires a separate model for every outcome variable and is widely used for estimating basic population characteristics in case of item nonresponse and for various analytical purposes; see Little and Rubin (1987) and Schenker et al. (2010). PSA relies on a modeling response probability conditional on covariates available for both responding and non-responding units. If this model is correctly specified, PSA is proved by Kim and Kim (2007) to be asymptotically unbiased and consistent for any outcome variable. Deville and Sarndal (1992) proposed a generalized calibration theory for the unified treatment of post-stratification, raking, and generalized regression estimator (GREG); see also Sarndal (2007). Generalized calibration was initially proposed as a method to reduce the variance of the Horwitz-Thomson estimator in the case of complete data. It utilizes covariates available for sampled units and their finite population totals available from the external sources to produce modified sampling weights by calibrating weighted sample totals of auxiliary covariates to their population totals. Application of the same routine in the case of nonresponse was described in Sarndal and Lundstrom (2005) and Kott (2006). Extension of generalized calibration utilizing instrumental variables, discussed in Kott and Chang (2010),

Kott (2006), allows in principle handling the situation of informative nonresponse, when the response probability directly depends on the target variable. This is because instrumental variables need to be observed only on the respondents and their population totals are not used for calibration. Generalized calibration was shown by Kott (2006) to produce an unbiased and consistent estimator of population parameters if the calibration covariates and instrumental variables are good predictors of outcome variable and response indicator. Combining different nonresponse adjustment methods is also possible. Haziza and Lesage (2016) argue that the two-step procedure, using PSA on the first step to model the nonresponse mechanism and generalized calibration on the second step, is more effective at eliminating potential bias due to model misspecification, compared with one-step calibration. The two-step procedure allows for sophisticated modeling of the response probability and, therefore, is more flexible than one-step calibration, which is restricted to the generalized linear model with some form of link function.

Estimation methods for bias reduction in case of nonresponse assume correct models for response probability or outcome variables, but they are difficult to identify. The same is expected for estimates from web samples. Any improvement of robustness of these methods to model misspecification would greatly improve the chances of using web sample data for reliable estimation of population parameters. Bethlehem (1988) and Haziza and Lesage (2016) express bias of estimates of totals as a correlation between residuals of models for response indicator and target variables. This provides support for the accepted interpretation of "double robustness" property of estimation methods in case of nonresponse, which allows misspecification of one of these two models if another is correct. The same expression for bias suggests another understanding of double robustness: Each of these two models should only specify dependencies on covariates which are also relevant to the other model. This property of a response propensity model was noted in review by Brick (2013). Accounting for any other covariates, no matter how relevant they are to *just one* of the models, should not affect the bias of estimates from nonrandom samples for a specific outcome variable.

In Section 1 of this paper, the modified PSA of sampling weights was derived for the web samples. It involves modeling the web sample inclusion indicator on the dataset combining the web sample and the randomized NHIS reference sample utilizing the core variables available in both samples. It produces the set of weights projecting the web sample total to the weighted total over the reference NHIS sample. In this sense, the proposed PSA is similar to Sarndal's generalized calibration. This similarity suggests using the same expression in both cases for variance estimation from Chapter 11 of Sarndal and Lundstrom (2005). Section 2 presents the results of simulations demonstrating the similarity of PSA and generalized calibration estimators when used with web sample data. The variance estimation procedure is validated by comparing estimated standard errors with Monte Carlo variability of estimates over the simulations and by calculating the coverage of the finite population parameter by the estimated confidence intervals. This proposed interpretation of double robustness is clearly illustrated by showing the lack of bias of any estimator accounting for covariates essential for *both* target variables and web sample indicator. However, if the probability to belong to a web sample depends on a target variable, there is no set of auxiliary covariates that could provide for unbiased estimation. In this case, unbiased estimates are possible only when the outcome variable is used as the instrumental variable in the generalized calibration framework. In Section 3, generalized calibration estimates of population means for two detail variables from the real web sample data were validated by comparison with regular NHIS estimates. Estimates of totals of the selected demographic and core covariates from the reference NHIS sample were used for calibration. Model covariates were selected to minimize conditional correlation between the outcome variable and the web sample inclusion indicator. It was demonstrated that such criteria, following from the notion of double robustness, provides for a reliable inference procedure. In conclusion we discuss the feasibility and challenges of producing reliable estimates from web sample data.

1. PSA estimator in case of web samples and its similarity to generalized calibration

The PSA of sampling weights to reduce bias of estimates in case of nonresponse requires modeling the sample unit response indicator R_i on the combined sample of respondents and non-respondents utilizing covariates $\mathbf{X}_i$ available for all sampled units:

$$\Pr\left(R_i = 1 | \mathbf{X}_i = \mathbf{x}_i\right) = \phi\left(\mathbf{x}_i\right) = \phi_i \tag{1.1}$$

Estimates of the propensity score $\hat{\phi}\left(\mathbf{x}_i\right)$ may be obtained by parametric [Little (1986)] or non-parametric methods [Da Silva and Opsomer (2009); Phipps and Toth (2012)] and used in the PSA estimator of population total [Brick (2013); Haziza and Lesage (2016)]:

$$\hat{t}^{NR}_{PSA} = \sum_{i \in s_r} d_i \hat{\phi}_i^{-1} y_i \tag{1.2}$$

where d_i is sampling weights before nonresponse and summation goes over the sample of respondents s_r. Estimator $\hat{t}^{NR}_{PSA}$ is proved by Kim and Kim (2007) to be unbiased and consistent if the propensity score $\hat{\phi}\left(\mathbf{x}_i\right)$ is estimated from the correct model.

A similar expression can be derived for web samples. Every unit of the web sample s_W is characterized by the set of variables $\left(\mathbf{X}_i, d_i^W, y_i\right)$, where $\mathbf{X}_i$ is demographic and core covariates, d_i^W is possible web sample weights and y_i is the target detail variable. Due to the nonrandom nature of web sample selection, there is no guarantee that weights d_i^W can provide for unbiased estimation of any detail variable.

A reference sample s_H, selected following usual randomized sampling, has detailed questions omitted from the questionnaire to reduce the burden on respondents, so the target variable is missing $\left(\mathbf{X}_i, d_i^H\right)$. Sampling weights d_i^H provide for unbiased estimation of any population characteristic. The question becomes: how to estimate the finite population total of target detail variable y_i from web and reference samples.

This question can be answered by following the similarity between nonresponse in randomized surveys and nonrandom recruitment to web panels. In both cases, the final sample is drawn in two stages. In the case of nonresponse, a random sample s of size n is first drawn from a population U of size N, so each population unit may be drawn with probability π_i. Then it is assumed that a sample of respondents s_R of size n_R is drawn from the random sample according to a Poisson sampling, when each unit is drawn independently with the probability of selection depending on a set of covariates $\phi_R\left(\mathbf{X}_i\right)$. The described two-stage process can be presented as:

$$U\left(N\right) \xrightarrow{\pi_i} s\left(n\right) \xrightarrow{\phi_R(\mathbf{X}_i)} s_R\left(n_R\right) \tag{1.3}$$

The lack of bias of the nonresponse PSA estimator (1.2) is easily demonstrated:

$$\begin{gathered} E_\pi\left(E_R\left(\hat{t}^{NR}_{PSA}\right)\right) = E_\pi\left(E_R\left(\sum_{i \in U} \frac{I_i R_i y_i}{\pi_i \phi_R(\mathbf{X}_i)}\right)\right) = E_\pi\left(\sum_{i \in U} \frac{I_i E_R(R_i) y_i}{\pi_i \phi_R(\mathbf{X}_i)}\right) = \\ \sum_{i \in U} \frac{E_\pi(I_i) y_i}{\pi_i} = \sum_{i \in U} y_i \end{gathered} \tag{1.4}$$

The sample inclusion indicator I_i is defined for the population, and survey response indicator R_i is defined for the units of random sample before nonresponse. The propensity to respond to a survey is defined conditionally on covariates $\mathbf{X}_i$ as:

$$\phi_R\left(\mathbf{X}_i\right) = n_R\left(\mathbf{X}_i\right)/n\left(\mathbf{X}_i\right) \tag{1.5}$$

and can be estimated by modeling the response indicator R_i on the sample s data before nonresponse. Consequently, response propensity $\phi_R\left(\mathbf{X}_i\right)$ should not depend on sampling weights $d_i = \pi_i^{-1}$. This is consistent with the conclusions of Little and Vartivarian (2003), whose simulations have shown that "weighting the response adjustment rates by the sampling weights is either incorrect or unnecessary. It is incorrect, in the sense of yielding biased estimates of population quantities if the design variables are related to survey nonresponse; it is unnecessary if the design variables are unrelated to survey nonresponse."
Compared with the situation of nonresponse, the two stages of selecting a web sample occur in a reversed order. The population of the prospective web panelists U_W of size N_W can be imagined resulting from the Poisson sampling from the general population U of size N with probabilities $\phi_W\left(\mathbf{X}_i\right)$ depending on covariates. At the second stage, web sample s_W having n_W units is drawn at random from the population U_W according to a certain design, so each unit is drawn with probability π_i. Sampling with different probabilities at the second stage actually happens when a web panel vendor decides to field a survey to a certain subset of available web panelists to better satisfy the goal of a specific survey. The web sample selection can be presented as the following two-stage process:

$$U\left(N\right) \xrightarrow{\phi_W(\mathbf{X}_i)} U_W\left(N_W\right) \xrightarrow{\pi_i} s_W\left(n_W\right) \tag{1.6}$$

Lack of bias of the PSA estimator from web samples is proven similarly to (1.4), but expectations over the random sample selection and the Poisson response are taken in reverse order:

$$\begin{aligned} E_W\left(E_\pi\left(\hat{t}_{PSA}^W\right)\right) = E_W\left(E_\pi\left(\sum_{i\in U}\frac{I_iW_iy_i}{\pi_i\phi_W(\mathbf{X}_i)}\right)\right) = E_W\left(\sum_{i\in U}\frac{W_iE_\pi(I_i)y_i}{\pi_i\phi_W(\mathbf{X}_i)}\right) = \\ \sum_{i\in U}\frac{E_W(W_i)y_i}{\phi_W(\mathbf{X}_i)} = \sum_{i\in U} y_i \end{aligned} \tag{1.7}$$

Random variable I_i indicates selected web sample units of the population U_W of potential web panelists. Indicator W_i points to units of population U_W among the general population U. Estimation of conditional probability of being a web panelist:

$$\phi_W\left(\mathbf{X}_i\right) = N_W\left(\mathbf{X}_i\right)/N\left(\mathbf{X}_i\right) \tag{1.8}$$

requires modeling the indicator variable W_i on general population U with covariates $\mathbf{X}_i$. Alternatively, the propensity $\phi_W\left(\mathbf{X}_i\right)$ may be defined by considering the combined population $U_C = U_W \cup U$ and the new indicator variable Q_i on this population:

$$Q_i = \begin{cases} 1, i \in U_W \\ 0, i \in U \end{cases} \tag{1.9}$$

Some of the units of the original population U are duplicated in the population U_C, since they belong to both U and U_W. Index i running through the combined population U_C differentiates such units. One of each pair of duplicated units will have $Q_i = 1$, while another will have $Q_i = 0$. Nevertheless, the probability that indicator $Q_i = 1$ can still be defined as:

$$q_W(\mathbf{X}_i) = \frac{N_W(\mathbf{X}_i)}{N_W(\mathbf{X}_i) + N(\mathbf{X}_i)} = \frac{\phi_W(\mathbf{X}_i)}{1 + \phi_W(\mathbf{X}_i)} \tag{1.10}$$

Note the relation of this probability to the probability to be a prospective web panelist (1.8), which is the ultimate goal of the estimation effort. In principle, $q_W(\mathbf{X}_i)$ can be estimated by modeling indicator Q_i on the combined population U_C.

The fact that this population is unavailable may be circumvented by considering the randomly selected reference sample s_H from the general population U, which shares covariates $\mathbf{X}_i$ with the web sample s_W. Having populations $N_W(\mathbf{X}_i)$ and $N(\mathbf{X}_i)$ estimated from the web and reference samples s_W and s_H, the sample-based definition of propensity $q_W(\mathbf{X}_i)$ (1.10) becomes:

$$q_W^s(\mathbf{X}_i) = \frac{\sum\limits_{s_W, \mathbf{X}_j = \mathbf{X}_i} d_{W,j}}{\sum\limits_{s_W, \mathbf{X}_j = \mathbf{X}_i} d_{W,j} + \sum\limits_{s_H, \mathbf{X}_j = \mathbf{X}_i} d_{H,j}} \tag{1.11}$$

Here d_W and d_H are sampling weights associated with web and reference samples, randomly drawn from the corresponding populations. According to sampling theory [Sarndal et al. (1992)], $q_W^s(\mathbf{X}_i)$ is an unbiased and consistent estimator of $q_W(\mathbf{X}_i)$. In turn, sample probability $q_W^s(\mathbf{X}_i)$ may be estimated by modeling indicator variable Q_i (1.9) on the combined sample $s_C = s_H \cup s_U$. Estimated probability $\hat{q}_W^s(\mathbf{X}_i)$, substituted in (1.10), produces the unbiased estimator of probability of being a web panelist (1.8):

$$\hat{\phi}_W^{-1}(\mathbf{X}_i) = \frac{1 - \hat{q}_W^s(\mathbf{X}_i)}{\hat{q}_W^s(\mathbf{X}_i)} \tag{1.12}$$

Having this probability estimated, the unbiased estimator of the population total from web sample data (1.7) may be expressed similarly to the corresponding estimator in case of nonresponse (1.2):

$$\hat{t}_{PSA}^W = \sum_{i \in s_W} d_{W,i} \hat{\phi}_W^{-1}(\mathbf{X}_i) y_i = \sum_{i \in s_W} d_{W,i} \frac{1 - \hat{q}_W^s(\mathbf{X}_i)}{\hat{q}_W^s(\mathbf{X}_i)} y_i = \sum_{i \in s_W} d_{W,i} \hat{o}_i(\mathbf{X}_i) y_i \tag{1.13}$$

where $\hat{o}_i = (1 - \hat{q}_W^s(\mathbf{X}_i))/\hat{q}_W^s(\mathbf{X}_i)$ are the estimated odds of belonging to the web sample s_H for units of the combined sample s_C.

Expressions (1.10, 1.11) support *finding (a)* from a simulation study conducted by Valliant and Dever (2011) "that estimators of means based on estimates of propensity models that do not use the weights associated with the reference sample are biased even when the probability of volunteering is correctly modeled."

To summarize, PSA of sampling weights in case of nonresponse can be modeled from the available sample before nonresponse without accounting for sampling weights. In the case of web samples, as it explicitly follows from the expression (1.11), modeling corresponding adjustments requires an additional reference sample, and must account for the sampling weights in both the reference and web samples.

However, we make a conjecture that the estimate of total (1.13) does not ultimately depend on the web sample weights $d_{W,i}$, if these weights are simultaneously ignored in both modeling of q_W^s and in (1.13). The last fact can be easily shown for the saturated model $q_W^s(\mathbf{X}_i)$. In this case, index i designates adjustments cells, for which it is assumed that web and reference samples have $n_{W,i}$ and $n_{H,i}$ units with identical weights $d_{W,i}$ and $d_{H,i}$.

Expression (1.13) then becomes:

$$\hat{t}^{W}_{PSA,Sat} = \sum_{i \in s_W} \frac{n_{H,i}}{n_{W,i}} d_{H,i} y_i \tag{1.14}$$

Simulations described in Section 2 demonstrate independence of the estimates of total of the web sampling weights (supporting results are not presented here). If proved to be the case, this independence means that any method of a web sample data collection is suitable for estimation of population characteristics, as long as basic demographic groups are reasonably well represented, so estimated propensity scores $\hat{\phi}_W (\mathbf{X}_i)$ are not very small. In application to web sample data, the estimator $\hat{t}^{W}_{PSA}$ is very similar to the generalized calibration estimator of Deville and Sarndal (1992):

$$\hat{t}^{W}_{C} = \sum_{s_W} d_i^W F \left(\hat{\lambda}_w^T \mathbf{x}_i \right) y_i \tag{1.15}$$

Choice of calibration function $F \left(\hat{\lambda}_w^T \mathbf{x}_i \right)$ corresponds to different calibration methods, such as generalized regression estimator (GREG), raking, etc. The four most used functions are described by Haziza and Lesage (2016) and implemented in statistical software for calibration. Parameters $\hat{\lambda}_w^T$ are determined by minimizing deviation of the adjusted weights $w_i = d_i^W F \left(\hat{\lambda}_w^T \mathbf{x}_i \right)$ from the original sample weights d_i^W with additional calibration constraints on population totals for covariates $\mathbf{X}$:

$$\mathbf{t}_{\mathbf{x}} = \sum_{s_W} d_i^W F \left(\hat{\lambda}_w^T \mathbf{x}_i \right) \mathbf{x}_i \tag{1.16}$$

In the situation of web and reference samples, population totals $\mathbf{t}_{\mathbf{x}}$ in calibration equations (1.16) may be substituted with their estimates from the reference sample $\hat{\mathbf{t}}_{\mathbf{x}} = \sum_{s_H} d_i^H \mathbf{X}_i^H$. Then the PSA (1.13) and generalized calibration (1.15) estimators are very close in the sense that they both produce sampling weight adjustments for every unit of a web sample:

$$\hat{v}_i = \begin{cases} \hat{o}_i = (1 - \hat{q}_i)/\hat{q}_i \text{ , PSA} \\ F \left(\hat{\lambda}_w^T \mathbf{x}_i \right), \text{ Calibration} \end{cases} \tag{1.17}$$

Note, that expression (1.13) holds asymptotically for any variable y, while calibration (1.16) holds exactly, but only for covariates $\mathbf{X}$.

Despite these similarities, motivations for covariate selection for the PSA model and generalized calibration are different. Covariates in (1.13) must be explanatory of the indicator variable Q, while covariates for GREG and other calibration estimators must correlate with the target variable [Sarndal et al. (1992)]. This apparent dichotomy is explained by the dependence of nonresponse bias of PSA and calibration estimators on the correlation between the residuals of estimated response propensity and outcome variable models, first noted by Bethlehem (1988). Sarndal and Lundstrom (2005) and Haziza and Lesage (2016) give the following expression for the bias of the calibration estimator:

$$Bias \left(t_C^{NR} \right) \approx - \sum_{i \in U} \left(1 - \varphi_i \hat{\varphi}_i^{-1} \right) \left(y_i - \mathbf{x}_i^T \mathbf{B} \right) \tag{1.18}$$

where φ_i is the true response propensity, $\hat{\varphi}_i^{-1} = F \left(\hat{\lambda}_i \mathbf{x}_i^T \right)$, and $\mathbf{B}$ is the regression coefficient calculated for the population.

Bias is zero if either of the models for response propensity or target variable y is correct, because corresponding residuals are randomly distributed and therefore, uncorrelated with residuals from the other model. This phenomenon is known in the literature as "double robustness". It can be formulated using random variables $\Delta\varphi_i = \varphi_i - \hat{\varphi}_i$ and $\Delta y_i = y_i - \mathbf{x}_i^T \mathbf{B}$, representing residuals of both models. In these terms, expression for bias becomes Bias $\left(t_C^{NR}\right) \sim \mathrm{cov}\left(\Delta\varphi, \Delta y\right)$. Accepted interpretation of double robustness suggests that unbiased estimates are possible when either $E\left(\Delta\varphi|Z\right) = 0$ or $E\left(\Delta y|X\right) = 0$, if covariates Z and X are used by the propensity and target variable models.

It can be pointed out that expression (1.18) allows for more general interpretation of double robustness, which does not require correctness of either model. Suppose that covariates of the response propensity model $Z = (Z_0, U)$ and of the target variable model $X = (X_0, U)$ are correlated by a function of a covariate vector U, such that $\mathrm{cov}(Z, X) = f_{Z,X}\left(U\right)$ and $X_0 \perp Z_0$. Here and below, a function of a random variable is actually treated as a function of its realized value. For zero bias, it is sufficient if residuals of both models remain dependent on two uncorrelated sets of covariates: $E\left(\Delta\varphi|Z\right) = f_{\Delta\varphi}\left(Z_0\right)$ and $E\left(\Delta y|X\right) = f_{\Delta y}\left(X_0\right)$. In other words, only accounting for covariates U for both models is required for unbiased estimates. Such relaxed understanding of double robustness in case of nonresponse was noted by Brick (2013) and will be demonstrated for simulated web samples.

Suggested interpretation of the double robustness allows for more robust modeling and also may be used to guide the process of covariate selection for either sample indicator Q or the target variable model. These variables must be independent conditionally on model covariates $\mathbf{X}$:

$$f\left(y|\mathbf{X}, Q = 1\right) = f\left(y|\mathbf{X}\right) \tag{1.19}$$

They remain correlated if selected covariates are not adequate for bias reduction. This correlation can be estimated from the following regression:

$$\begin{aligned} &Y \sim \mathbf{X}\beta_{\mathbf{X}} + Q\beta_Q, \\ &\beta_Q = \mathrm{cov}\left(Y, Q|\mathbf{X}\right) \end{aligned} \tag{1.20}$$

Because the target variable is usually unavailable for the reference sample, this criteria cannot be applied in practice for selecting covariates. In the NCHS experiment, however, target variables are available for both web and reference NHIS samples. Expression (1.19) will be first checked in simulations and then used to justify covariate selection for the real data, just to demonstrate the possibility of using demographic and core covariates to reduce bias of estimates for selected detail variables.

Making inferences from web samples requires the ability to estimate variances and confidence intervals. Sarndal and Lundstrom (2005) generalized a well-known variance estimation formula for the GREG estimator to the case of nonresponse. Samples with nonresponse were considered resulting from a two-stage selection process: random sampling followed by nonresponse, which is treated as a Poisson sampling. The resulting estimator of full variance is a sum of the sampling variance of the GREG estimator and variance due to nonresponse. Estimates of these variances in cases of nonresponse, presented by expressions

(11.3-6) in Sarndal and Lundstrom (2005), can be applied to web samples:

$$\hat{V}\left(\hat{t}_C^W\right) = \hat{V}_{SAM} + \hat{V}_{NR} \tag{1.21a}$$

$$\begin{aligned} \hat{V}_{SAM} = \sum\sum\nolimits_{s^W}\left(d_k^W d_l^W - d_{kl}^W\right)\left(v_k \hat{e}_k\right)\left(v_l \hat{e}_l\right) - \\ \sum\nolimits_{s^W} d_k^W\left(d_k^W - 1\right) v_k\left(v_k - 1\right)\left(\hat{e}_k\right)^2 \end{aligned} \tag{1.21b}$$

$$\hat{V}_{NR} = \sum\nolimits_{s^W} v_k\left(v_k - 1\right)\left(d_k^W \hat{e}_k\right)^2 \tag{1.21c}$$

where $\hat{e}_k = y_k - \mathbf{x}_k \hat{\mathbf{B}}$ are residuals and $\hat{\mathbf{B}} = \left(\sum_{s^W} d_k^W v_k \mathbf{z}_k \mathbf{x}_k^T\right)^{-1}\left(\sum_{s^W} d_k^W v_k \mathbf{z}_k y_k\right)$ are model coefficients. In the conducted simulations and application to real data, it was assumed that stratified simple random sampling would be the second stage of web sample selection. Application of (1.21a) in such a case is described in the section devoted to a simulation study.

Instrumental covariates $\mathbf{z}_k^T$ descriptive of nonresponse may be used for weight adjustments $F\left(\hat{\lambda}_w^T \mathbf{z}_i\right)$ instead of calibration covariates $\mathbf{x}_k^T$ assumed to be descriptive of a target variable. This contradicts the notion of double robustness in stating that only covariates relevant to explaining both sample indicator Q and the target variable are relevant for bias correction. However, instrumental covariates may be the only way to deal with bias in the case of *informative* web sample selection, when the web sample indicator explicitly depends on the target variable:

$$\Pr\left(Q_i = 1 | \mathbf{X}_i = \mathbf{x}_i, d_i^H, d_i^W, y_i\right) = q_i \tag{1.22}$$

In this case, no set of covariates $\mathbf{X}_i$ can provide for bias correction. Kott and Chang (2010) have shown that bias can only be reduced when the target variable is employed as an instrumental variable. Using instrumental variables to handle bias caused by informative selection of web samples is demonstrated in simulations below.

2. Performance of PSA and calibration estimators on simulated web samples

A population of size $N = 10,000$ was simulated with two auxiliary variables and one target binary variable (X_1, X_2, Y). X_1 plays the role of a "demographic" covariate (such as race), X_2 represents a possible *core* covariate (such as a response to the question, "Have you ever been told that you have asthma?") and Y can be thought of as a *detail* target variable correlated with X_2 (such as a response to the question, "Do you take asthma medication?"). The population was unevenly stratified by X_1 and evenly by X_2, as $N\left(X_1 = 0\right) = 0.8N$ and $N\left(X_2 = 0\right) = 0.5N$. Reflecting the natural dependence between core and detail variables in real data, Y was set to 0 for $X_2 = 0$, while for $X_2 = 1$, it was randomly generated from a Bernoulli distribution with probability $p_Y\left(X_1\right)$.

From this population, a reference sample s_H of size $n_H = 1000$ was selected using stratified random sampling without replacement (STSR WOR). Stratification by the "demographic" variable was different than in population $n_H\left(X_1 = 0\right) = 0.6n_H$. Therefore, units of the reference sample were weighted depending on X_1 as $d_i^H\left(X_1\right) = N\left(X_1\right)/n_H\left(X_1\right)$.

The web sample s_W was selected in two stages. At the first stage, a simple random sample of size $n_W = 1000$ was selected, stratified by X_1 as $n_W\left(X_1 = 0\right) = 0.9n_W$. As a result, weights $d_i^W\left(X_1\right)$ were associated with web sample units. In the second stage, an additional Poisson-like selection of the web sample units with probability $p_i^W\left(X_{1i}, X_{2i}, Y_i\right)$ was used for mimicking the web panel recruitment and web survey nonresponse processes.

Simulations were conducted for the following cases defined by the web sample selection probabilities $p_i^W\left(X_{1i}, X_{2i}, Y_i\right)$ and population distributions of the binary target variable $p_Y\left(X_1\right)$. In all cases, the target variable Y depends on X_2 due to the natural correlation between core and detail variables, while web sample identifier Q always depends on demographic variable X_1 because of differences in stratification. Notations for different simulated cases show the resulting dependence of Q and Y on covariates $\left(X_1, X_2\right)$. When dependence on one of the covariates is missing, it is indicated by "$-$" (for example, $Q\left(X_1, -\right)$ in Case 2). In Case 5 of informative web sample selection, target variable Y was added to the regular set of covariates $Q\left(X_1, -, Y\right)$.
Case 1. Uncorrelated Y and Q: $Y\left(-, X_2\right); Q\left(X_1, -\right)$.
$p_Y\left(X_1\right) = 0.2$; $p_i^W\left(X_{1i}, X_{2i}, Y_i\right) = 0.5\left(\left(1 - X_{1i}\right) + 0.5X_{1i}\right)$
Case 2. Y and Q are correlated by X_1: $Y\left(X_1, X_2\right); Q\left(X_1, -\right)$.
$p_Y\left(X_1\right) = 0.2X_1 + 0.6\left(1 - X_1\right)$; $p_i^W\left(X_{1i}, X_{2i}, Y_i\right) = 0.5\left(\left(1 - X_{1i}\right) + 0.5X_{1i}\right)$
Case 3. Y and Q are correlated by X_2: $Y\left(-, X_2\right); Q\left(X_1, X_2\right)$.
$p_Y\left(X_1\right) = 0.2$; $p_i^W\left(X_{1i}, X_{2i}, Y_i\right) = \left(\left(1 - X_{1i}\right) + 0.5X_{1i}\right)\left(0.2\left(1 - X_{2i}\right) + 0.6X_{2i}\right)$
Case 4. Y and Q are correlated by $\left(X_1, X_2\right)$: $Y\left(X_1, X_2\right); Q\left(X_1, X_2\right)$.
$p_Y\left(X_1\right) = 0.2X_1{+}0.6\left(1 - X_1\right)$; $p_i^W\left(X_{1i}, X_{2i}, Y_i\right) = \left(\left(1 - X_{1i}\right) + 0.5X_{1i}\right)\left(0.2\left(1 - X_{2i}\right) + 0.6X_{2i}\right)$
Case 5. Informative sampling: $Y\left(-, X_2\right); Q\left(X_1, -, Y\right)$.
$p_Y\left(X_1\right) = 0.4$; $p_i^W\left(X_{1i}, X_{2i}, Y_i\right) = 0.5\left(0.7\left(1 - Y_i\right) + Y_i\right)$
In all of the cases, resulting size of the web samples fluctuated around $n_W \sim 400$.
$N_{sim} = 800$ reference and web samples were drawn from the simulated population in all cases and the finite population mean $\bar{Y} = \sum_U Y_i/N$ was estimated using PSA and calibration estimators (1.13) and (1.15). Both estimators result in adjusting web sample weights by $\hat{v}_i$ (1.17), so estimates of the population mean become:

$$\hat{\bar{y}}^W = \sum\nolimits_{s_W} d_i^W \hat{v}_i y_i \Big/ \sum\nolimits_{s_W} d_i^W \hat{v}_i \tag{2.1}$$

Weight adjustments $\hat{v}_i$ for the calibration estimator were calculated using *calib* and *gencalib* functions of the R package *sampling*, made available by Tille and Matei (2015). It is possible to select one of the four calibration adjustment functions $F\left(\hat{\lambda}_w^T \mathbf{x}_i\right)$ by specifying parameter `method =c("linear","raking","truncated", "logit")`. The exponential form of adjustment function $F\left(\hat{\lambda}_w^T \mathbf{x}_i\right) = \exp\left(\hat{\lambda}_w^T \mathbf{x}_i\right)$ was selected, because it corresponds to post-stratification by "raking," a method widely used by survey practitioners.
Double robustness was demonstrated by comparing the performance of estimators depending on all available covariates $\left(X_1, X_2\right)$ with calibration estimators employing limited sets of calibration covariates, either $\left(X_1, -\right)$ or $\left(-, X_2\right)$. The simple estimator of the mean from the web sample data, ignoring all weights and adjustments, is also presented for comparison. The generalized calibration estimator using the target variable as the instrumental variable is presented to demonstrate its advantage in the case of informative web sample selection.
Different estimators of the mean were compared by calculating their relative biases over the simulations:

$$\text{RB}\left(\hat{\bar{y}}^W\right) = \left(\sum_{s=1}^{N_{sim}} \hat{\bar{y}}_s^W - \bar{Y}\right) \Big/ \bar{Y} \tag{2.2}$$

Results, presented in Table 1, show almost identical performance of PSA and calibration estimators utilizing the same set of covariates, even though the PSA estimator models web sample indicator Q, while the calibration estimator calibrates on population totals of covariates correlated with target variable Y.

They also support the relaxed interpretations of double robustness proposed at the end of Section 1. For example, in Case 1 when variables Y and Q depend on different covariates, even a simple MEAN$(-,-)$ estimator produces unbiased estimates. In Case 2, variables Y and Q depend on the same covariate X_1. Estimator CAL$(X_1,-)$, which models $Y(X_1)$ and leaves residuals $E(Y) = f(X_2)$ dependent on X_2 and does not model Q at all, produces unbiased estimates. Estimator PSA$(-,X_2)$, not presented in Table 1, was also used for estimation in all simulated cases. Though it ignored dependence of $Q\,(X_1,X_2)$ on X_1 in Case 3, it still produced unbiased estimates because the outcome variable was independent of X_1. Other cases, like estimator CAL$(-,X_2)$ in Case 3, demonstrate more rigorous "standard" interpretation of double robustness, requiring correctness of one of the models. In Case 4, both variables Y and Q depend on (X_1,X_2). In this case, no estimator ignoring either one of these variables could produce unbiased estimates. This provides implicit support for both interpretations of double robustness.

Table 1: Relative bias (2.2) of PSA and calibration estimators depending on model/calibration covariates. Simulations conducted for five cases of different dependences of sample indicator Q and target variable Y on designated "*core*" and "demographic" covariates.

Estimator $\hat{y}$	MEAN	CAL	CAL	PSA	CAL	CAL
Simulation	$(-,-)$	$(X_1,-)$	$(-,X_2)$	(X_1,X_2)	(X_1,X_2)	(X_1,Y)
$Y\,(-,X_2)\,;Q\,(X_1,-)$	0.005	0.007	0.005	0.006	0.006	0.024
$Y\,(X_1,X_2)\,;Q\,(X_1,-)$	0.110	0.001	0.066	-0.005	-0.001	-0.007
$Y\,(-,X_2)\,;Q\,(X_1,X_2)$	0.508	0.510	0.005	0.008	0.006	NA
$Y\,(X_1,X_2)\,;Q\,(X_1,X_2)$	0.668	0.503	0.068	-0.002	-0.003	-0.868
$Y\,(-,X_2)\,;Q\,(X_1,-,Y)$	0.323	0.322	0.222	0.223	0.222	-0.009

Table 2: Coverage of the population mean by the estimated 95% confidence intervals of the calibration and PSA estimators. Simulations are conducted for five cases of different dependences of sample indicator Q and target variable Y on designated "*core*" and "demographic" covariates.

Estimator $\hat{y}$	CAL	CAL	PSA	CAL	CAL
Simulation	$(X_1,-)$	$(-,X_2)$	(X_1,X_2)	(X_1,X_2)	(X_1,Y)
$Y\,(-,X_2)\,;Q\,(X_1,-)$	0.93	0.93	0.91	0.93	0.94
$Y\,(X_1,X_2)\,;Q\,(X_1,-)$	0.94	0.81	0.92	0.92	0.91
$Y\,(-,X_2)\,;Q\,(X_1,X_2)$	0.33	0.93	0.91	0.92	NA
$Y\,(X_1,X_2)\,;Q\,(X_1,X_2)$	0.0	0.74	0.93	0.91	0.0
$Y\,(-,X_2)\,;Q\,(X_1,-,Y)$	0.2	0.31	0.30	0.31	0.89

Coverage of the population mean by the estimated 95% confidence intervals of the calibration and PSA estimators in the conducted simulations is presented in Table 2. Variances were estimated using expression (1.21a), assuming a two-stage sample selection process. If the first stage of web sample selection is considered stratified simple random sampling, both individual d_k and pairwise d_{kl} sampling weights are expressed through population size N_h and sample counts n_h in strata. Chapter 11 of Sarndal and Lundstrom (2005) presents detailed expressions for variance estimates in formula (11.7-11.9). Weight adjustments v_k corresponding to Poisson sampling at the second stage of web sample selection for PSA and calibration estimators were estimated using (1.17). Calculated confidence intervals always provide for close to nominal coverage when model/calibration covariates include

those correlating with both sample indicator Q and target variable Y. Modest undercoverage may be explained by a possible underestimation of standard errors.
The last simulated case of informative sample selection $Y\left(-, X_2\right); Q\left(X_1, -, Y\right)$ is a notable exception to the described standard behavior of the relative bias and coverage of the final population mean. No set of covariates can help to obtain a reasonable inference, unless the target variable itself is used as an instrumental variable of the generalized calibration estimator. However, this estimator has a much larger variance and is very unstable for all other simulated cases.
As shown in (1.20), covariate selection can be validated by estimating residual correlation between the target variable and the web sample indicator, conditional on the selected covariates. The results of *t* statistic for testing of the null hypothesis of these correlations for simulated data are presented in Table 3.

Table 3: t-statistics of null-hypothesis testing of conditional correlation $\beta_Q\left(\mathbf{X}\right) = \text{cov}\left(Y, Q|\mathbf{X}\right)$ (1.20) for different simulations scenarios

Simulation	$\beta_Q\left(X_1\right)$	$\beta_Q\left(X_2\right)$	$\beta_Q\left(X_1, X_2\right)$
$Y\left(-, X_2\right); Q\left(X_1, -\right)$	0.04	0.04	0.03
$Y\left(X_1, X_2\right); Q\left(X_1, -\right)$	0.0	0.92	-0.02
$Y\left(-, X_2\right); Q\left(X_1, X_2\right)$	2.68	0.05	0.04
$Y\left(X_1, X_2\right); Q\left(X_1, X_2\right)$	5.03	1.11	0.25
$Y\left(-, X_2\right); Q\left(X_1, -, Y\right)$	2.63	2.24	2.24

Comparing Tables $1 - 3$ clearly shows that the t-statistic is a good indicator of biased estimates for a given set of model covariates. For example, in Case 2 of the simulations when Q and Y are correlated only by X_1, the t-statistic for $\beta_Q\left(X_1\right)$ is close to 0, the calibration estimator using just X_1 has a low relative bias of 0.001 and almost nominal coverage of 0.94. At the same time, $\beta_Q\left(X_2\right) = 0.92$ indicates a relatively large residual correlation between Q and Y, pointing to substantial (0.066) relative bias and lack of coverage (0.81) of the corresponding estimator.

3. Calibration estimator in application to the real NHIS and web sample data: Selection of calibration covariates, inference, and validation

The calibration estimator (1.15) was used to estimate population means for two detail target variables y^{still} and y^{myr}, corresponding to the NHIS questions: “Do you still have asthma?” and “Did you have an asthma attack in the past year?”. These detail variables are related to the same core variable X_2^{ever}, corresponding to the gateway question: “Have you ever been told by a doctor that you have asthma?”.
Though not the case in reality, in the conducted experiment, variables y^{still} and y^{myr} were available for both web and reference samples. This fact was used to justify the selection of model covariates by testing the null hypothesis of independence (1.19, 1.20) of the web sample indicator Q and target variable. While y^{still} was independent of Q conditionally on just X_2^{ever}, in the case of y^{myr} conditional independence was achieved by employing additional covariates $\mathbf{X}_1$. Calibration on (X_2^{ever}) should be optimal for y^{still} and insufficient for y^{myr}, while calibration on $(X_2^{\text{ever}}, \mathbf{X}_1)$ should be excessive for y^{still} and optimal for y^{myr}:

Model	y^{still}	y^{myr}
$\mathbf{X}^{\text{still}} = (X_2^{\text{ever}})$	Optimal (*t* = -0.44)	Insufficient
$\mathbf{X}^{\text{myr}} = (X_2^{\text{ever}}, \mathbf{X}_1)$	Excessive	Optimal (*t* = 0.47)

Here $\mathbf{X}_1$ combines variables corresponding to the following NHIS questions:
Would you say your health in general is excellent, very good, good, fair, or poor?
Did you use a computer to look up health information in the last year?
Did you use a computer to schedule an appointment with a health care provider in the last year?
Have you delayed getting care for any reason in the last year?
Has it happened that you could not afford getting health care for any reason in the last year?
Inferences from the web sample data in demographic domains obtained using the calibration estimator were compared to direct estimates from the reference NHIS sample. Relative errors were calculated similarly to relative bias (2.2) using estimates from the reference sample $\hat{\hat{y}}_{ref}$ as the "true" value. An analogue of simulated coverage of the "true" value could be obtained by calculating *t*-statistics of pairwise tests comparing estimates from web and random NHIS samples:

$$t_y^{stat} = \frac{\hat{\hat{y}}_{web} - \hat{\hat{y}}_{ref}}{\sqrt{\text{var}\left(\hat{\hat{y}}_{web}\right) + \text{var}\left(\hat{\hat{y}}_{ref}\right)}} \tag{3.1}$$

$\text{var}\left(\hat{\hat{y}}_{web}\right)$ is calculated using (1.21a) and $\text{var}\left(\hat{\hat{y}}_{ref}\right)$ is the regular variance of direct estimates from complex surveys [Sarndal et al. (1992)]. The value $\left|t_y^{stat}\right| < z_{1-0.05/2} = 1.96$ indicates the absence of statistically significant differences between domain estimates $\hat{\hat{y}}_{web}$ and $\hat{\hat{y}}_{ref}$ within 95% confidence limits.
Calibration estimates of $\hat{\hat{y}}_{still}$ and $\hat{\hat{y}}_{myr}$ in demographic domains from web sample data using optimal, excessive, and insufficient sets of covariates are compared to estimates from the regular NHIS sample in Table 4.

Table 4: Relative error and pairwise *t*-test for domain estimates obtained using optimal and excessive sets of calibration covariates according to conditional correlation criteria (1.20). *Estimates shown in red are significantly different between web and reference samples.

Target variable	$\hat{\hat{y}}_{still}$				$\hat{\hat{y}}_{myr}$			
Model	$\mathbf{X}^{\text{still}} = (X_2^{\text{ever}})$ Optimal		$\mathbf{X}^{\text{myr}} = (X_2^{\text{ever}}, X_1)$ Excessive		$\mathbf{X}^{\text{myr}} = (X_2^{\text{ever}}, X_1)$ Optimal		$\mathbf{X}^{\text{still}} = (X_2^{\text{ever}})$ Insufficient	
Domain	Relative error	Pairwise *t*-test	Relative error	Pairwise *t*-test	Relative error	Pairwise *t*-test	Relative error	Pairwise *t*-test
All	-0.02	-0.4	-0.28	-7.1^*	-0.11	-1.5	0.3	4.1^*
Sex								
Male	0.22	2.4	-0.01	-0.1	0.29	1.6	0.68	4.0
Female	-0.13	-2.6	-0.41	-8.4	-0.28	-3.5	0.13	1.5
Age								
18 - 34	-0.07	-0.8	-0.39	-4.8	-0.25	-1.8	0.36	2.3
35 - 54	0.23	2.6	-0.22	-2.7	-0.05	-0.4	0.52	3.6
55 and over	-0.19	-2.5	-0.23	-2.9	-0.05	-0.4	0.03	0.3
Race and ethnicity								
Non-Hispanic white	-0.16	-3.1	-0.37	-7.1	-0.28	-3.3	0.06	0.6
Non-Hispanic black	0.21	1.5	-0.19	-1.4	0.19	0.8	0.80	3.2
Non-Hispanic other	-0.03	-0.1	-0.10	-0.4	0.49	1.0	0.52	1.3
Hispanic	0.53	3.1	0.02	0.1	0.23	0.9	0.95	3.5
Education								
High school or less	-0.31	-4.5	-0.34	-4.5	-0.10	-0.8	-0.06	-0.6
Associate degree, some college	0.04	0.5	-0.26	-3.2	-0.24	-2.1	0.22	1.7
Bachelor's degree or higher	0.36	3.5	-0.22	-2.4	0.08	0.5	1.02	5.2
Income								
\$0 – \$49,999	-0.49	-9.2	-0.65	-12.8	-0.52	-6.5	-0.30	-3.5
\$50,000 – \$99,999	0.09	0.9	-0.22	-2.3	-0.05	-0.3	0.52	2.9
\$100,000 and over	0.27	2.1	-0.08	-0.6	0.16	0.7	0.59	2.5
Average	0.03	-0.4	-0.25	-3.8	-0.03	-0.9	0.40	2.0

Estimates in some of the demographic domains from the web and NHIS samples are statistically different, even for optimally selected covariates. However, optimal covariate selection substantially improves estimates for complete samples and has a very positive effect on errors and *t* statistics averaged over all domains. Estimates $\hat{\hat{y}}^{\text{myr}}\left(\mathbf{X}^{\text{still}}\right)$ based on an insufficient set of calibration covariates show persistent positive shifts, while estimates $\hat{\hat{y}}^{\text{still}}\left(\mathbf{X}^{\text{myr}}\right)$ using an excessive set of calibration covariates are smaller on average than estimates from the reference NHIS sample. The last observation is quite unexpected, given the general understanding that adding excessive covariates to the nonresponse adjustment model may increase variance of the estimates, but should not cause any extra bias. This result requires further investigation.

Estimates of standard errors (not presented in Table 4) do not show noticeable dependence on selected calibration covariates.

4. Conclusions and prospects of estimating general population characteristics from web sample data

This paper proposed using the PSA estimator for making inferences from a combination of web and reference samples data. If the propensity score model is correct and the web sample size is sufficiently large, it converges to the regular Horwitz-Thompson estimator from the reference sample. Its functional similarity to the calibration estimator, calibrating on estimates of covariates' totals from the reference sample, suggests that the variance expression for the calibration estimator in Sarndal and Lundstrom (2005) can be used for the PSA estimator. Inferences by both estimators checked in the simulation study proved to be very close, even though the PSA estimator depends on modeling the web sample indicator, while covariates of the calibration estimator must correlate with the target variable. This observation, and also the expression for bias of both estimators as the correlation between residuals of propensity and target variable models (1.18), suggests that the conditional correlation of the web sample indicator and target variables can be used as criteria for variable selection. Successfully tested in simulations, this criteria appeared to work in the course of the NCHS experiment. This is good evidence that web sample data can be used for inferences, at least for the two asthma-related NHIS variables under consideration, if reliable criteria for selecting model covariates are available.

Unfortunately, correlation between the target and web indicator variables cannot be easily calculated in practice because the target variable is usually not available for a reference sample. It remains the subject of future research to find some analogue of this criteria that can be applied in reality for model checking. Better estimates from web samples can be achieved by extending the class of estimators through the use of mixed models, nonparametric regression, and machine learning, as suggested by Breidt and Opsomer (2016). Such extensions of the basic PSA and calibration estimators can be developed in the application to web samples.

Estimation from nonrandomly selected web samples is prone to substantial biases. Reliability of the proposed estimation methodology using combined web and reference samples ultimately depends on the ability of the shared core and demographic covariates to explain correlation between the web sample indicator and target variables of interest. Thoughtful allocation of the corresponding questions between web and reference surveys and the development of reliable estimation methodologies may facilitate the estimation from futuristic data structures collected from a random reference survey bonded to a constellation of web surveys by strong covariates, as shown in Figure 1.

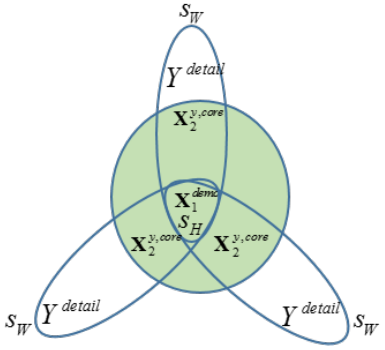

Figure 1: Web surveys complementing traditional random survey

References

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4(3):251–260.

Breidt, F. J. and Opsomer, J. D. (2016). Model-assisted survey estimation with modern prediction techniques. *Submitted to Statistical Science*.

Brick, J. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3):329–353.

Chmura, L., Rivers, D., Bailey, D., Pierce, C., and Bell, S. (2013). Modeling a probability sample? An evaluation of sample matching for an internet measurement panel. In *Presented at AAPOR 2013 Conference*.

Da Silva, D. and Opsomer, J. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35:165–176.

Dever, J. A., Rafferty, A., and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods*, 2(2):47–62.

Deville, J. and Sarndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

DiSogra, C., Cobb, C., Chan, E., and Dennis, J. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pages 4501–4515.

Haziza, D. and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1):129–145.

Kim, J. and Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35(4):501–514.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 32(2):133–142.

Kott, P. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491):1265–1275.

Little, R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54(2):139–157.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Little, R. and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22:1589–1599.

Phipps, P. and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6:772–794.

Sarndal, C. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.

Sarndal, C. and Lundstrom, S. (2005). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Sarndal, C., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Schenker, N., Raghunathan, T., and Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, 29(5):533–545.

Tille, Y. and Matei, A. (2015). *Package "sampling"*. `https://cran.r-project.org/web/packages/sampling/sampling.pdf`.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137.