Could Nonresponse Be Biasing Trends Of Health Estimates?

Beatrice Baribeau[1]
[1]Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6

Abstract
The Canadian Community Health Survey (CCHS), a household survey conducted by Statistics Canada, is experiencing steady declines in response rates. The targeted minimum response rate has not been achieved the last two years. Response rates for telephone interviews are particularly low. With nonresponse on the rise, the effect of nonresponse bias is of concern.

In this paper, the impact of nonresponse bias is simulated through the 2007 CCHS telephone respondent data. The group representing nonrespondents are shown to have characteristics different than the remaining respondents. These differences significantly affect health estimates. Furthermore, this study demonstrates these differences cannot be fully corrected via weighting. Under study assumptions, this paper provides evidence that health trends could be affected by bias from nonresponse.

Key Words: Nonresponse, bias, health survey, telephone frame

1. Introduction

The Canadian Community Health Survey (CCHS) is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. It surveys a large sample of respondents and is designed to provide reliable estimates at the health region level (120 sub-provincial domains)[1]. There are 21 key variables related to health for which estimates are produced.

The survey collects data via computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI). Like many surveys, response rates for both modes of collection have been on a steady decline. Telephone interviewing experiences consistently lower response rates than face-to-face interviewing. Figure 1 provides the response rate by mode of collection from 2001. The survey is undergoing a redesign and cost constraints are driving the exploration for more cost effective means of data collection.

[1] Excerpt from the Canadian Community Health Survey 2013User Guide.

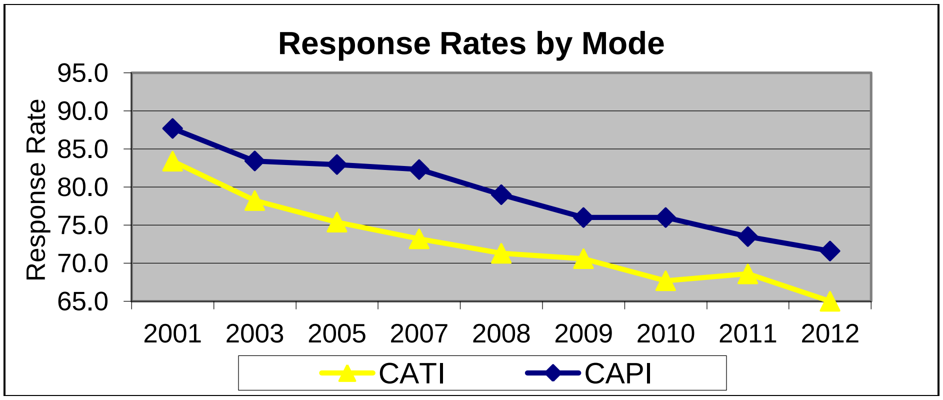

Figure 1: The steady decline in response rates over the last nine survey occasions.

1.1 Motivation for the Study

Nonresponse reduces the sample size and therefore affects the precision of estimates. It can also lead to nonresponse bias if the set of respondents have different health characteristics than the nonrespondents and these differences are not fully adjusted for through weighting procedures. The magnitude of the nonresponse bias is a function of the magnitude of the difference in variables of interest, between the respondents and the nonrespondents as well as the magnitude of the nonresponse. Modifying slightly the expression provided by Montaquila and Olson we can represent nonresponse bias as:

Nonresponse Bias = f(Rate, Difference between Respondents and Nonrespondents) [2] (1)

Nonresponse bias increases as the rate of nonresponse increases or as the differences in characteristics between the respondents and nonrespondents increases. From Figure 1, the nonresponse is increasing which implies a potential increase in nonresponse bias.

1.2 Study Details

While nonresponse bias is nearly impossible to measure, there are methods that can be employed to estimate it. Three methods were contemplated at the outset of this study. The first option was a nonresponse follow-up of a subsample of nonrespondents. This option was excluded due to budget and time constraints. The second option was to link the entire sample to an auxiliary source containing data on both respondents and nonrespondents in order to determine the differences between the two groups. This type of analysis only provides insight into characteristics that are available on the auxiliary source and not the survey variables themselves. Therefore potential bias in survey variables remains unknown. The third option, the method employed in this paper, was to simulate nonresponse on a dataset of respondents and then compare differences between the subset of respondents to the full sample in order to estimate bias. A year with a higher response rate was selected so that the estimate of bias due to nonresponse represented the potential bias due to increased nonresponse since the selected year. The validity of the results rely on the assumption that the simulated nonrespondents (later referred to as the "proxy nonrespondents") effectively represent current nonresponse. Therefore the simulated nonrespondents, or proxy nonrespondents, needed to be determined in manner not

[2] "Practical Tools for Nonresponse Bias Studies", 2012 SRMS/AAPOR Webinar by Montaquila & Olson.

dissimilar to the real nonresponse mechanism. Unfortunately, a limitation of this method is that it fails to estimate the nonresponse experienced in the selected year's data. Although the study is limited by its assumptions, it provided insight in a cost effective and timely manner.

The dataset used in this paper is the 2007 CCHS sample of respondents. The focus of the study is the CATI portion of the sample. The response rates for the CATI portion are lower, making it more susceptible to increased nonresponse bias. The CATI response rate of 2007 was 73% while for the most recent year it was 64%. The study attempts to estimate the potential bias from the nine percent decrease in response. To simulate nonresponse in a manner similar to what is currently experienced; respondents were classified as "easy-to-reach" respondents and "difficult-to-reach" respondents. The groups were created by using the number of call attempts required to obtain response. Figure 2, charts the change in the estimate of daily fruit and vegetable consumption throughout collection. When the estimate is calculated based only on respondents from the first collection call attempt, the estimate is at its lowest value. As more responses are obtained through collection, the estimate steadily increases. Three quarters of all respondents are reached within the first 10 call attempts, yet the remaining quarter of respondents continues to pull the estimate upward. The steady change in the estimate through collection indicates that easy-to-reach respondents have different characteristics than difficult-to-reach respondents.

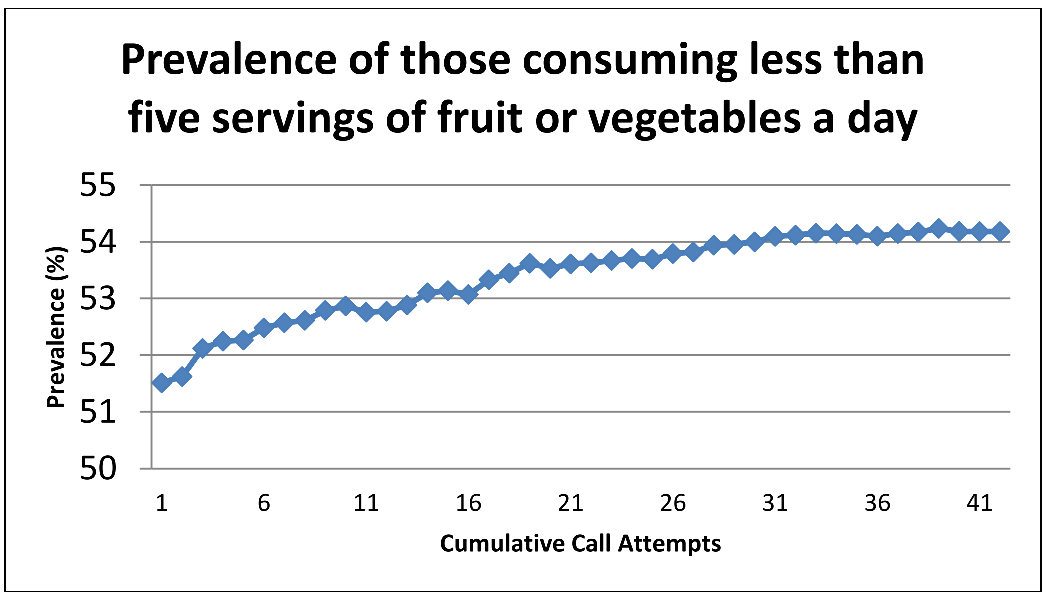

Figure 2: The change in estimated prevalence of a key indicator through collection.

The difficult-to-reach respondents, those requiring more call attempts before response was obtained, were thought to be more reluctant respondents and therefore more likely to represent the nine percent of the sample that might previously have been respondents but now would be nonrespondents. Imposing a maximum cut-off of 13 call attempts to define the easy-to-reach respondents yielded a 60% response rate, a little lower than the current national CATI response rate. The difficult-to-reach respondents then become the proxy nonrespondents for the study.

As shown in formula 1, for nonresponse bias to exist, a difference between respondents and nonrespondents is required. To determine whether differences exist, the subsampled respondents (the easy-to-reach respondents) were compared to the proxy nonrespondents (the difficult-to-reach respondents). After determining that differences exist, the subsampled respondents were compared to the full sample of respondents to establish whether the omission of the proxy nonrespondents is enough to alter estimates. Finally, because weighting steps can adjust for nonresponse, the subsampled respondents were reweighted in a manner similar to the usual weighting methodology to see whether weighting could fully adjust for differences due to the additional proxy nonresponse.

2. Study Results

2.1 Statistical Testing of Differences

As shown in Figure 2, there is evidence of differences between the easy-to-reach respondents and the difficult-to-reach respondents in terms of some survey variables of interest. Whether or not these differences are significant remains to be established. In terms of sociodemographic variables, it is well known that certain groups are more likely to respond, and respond more quickly, than other groups. For instance, younger males are harder to reach than older persons; male or female. These traits may also be correlated to some health factors, so in the following tables, comparison of estimates between easy-to-reach respondents and the proxy nonrespondents is made within age group and sex or within the sub-provincial health regions for the survey.

Table 1 shows the number of significant differences in domain estimates between the easy-to-reach respondents and the proxy nonrespondents for the 21 key health indicators. Estimates that had fewer than 10 respondents in the numerator of the prevalence rate were excluded from analyses.

Table 1: Comparison of easy-to-reach respondents vs. difficult-to-reach respondents based on 21 key health indicators

Domain	*Total number of estimates*	*Percentage of estimates found to be significant ($\alpha=0.05$)*
Canada	21	62%
Five age groups by gender	192	10%
120 sub-provincial geographic domains	1 079	10%

Easy-to-reach respondents have different characteristics than difficult-to-reach respondents for 13 out of the 21 key variables when compared at the national level. Even with comparisons limited to within age and sex group or within geographic domain, significant differences remain. Difficult-to-reach respondents comprise a small proportion of all respondents therefore converting them to nonrespondents may not have a significant impact on estimates. Table 2 compares the estimates based on the full set of respondents to estimates from just the easy-to-reach respondents. It may seem counterintuitive that the number of significant differences between Table 1 and Table 2 has increased. While the two groups compared in Table 2 are more homogeneous than in Table 1 (easy-to-reach respondents are represented in both comparison groups in Table 2, the dependence was accounted for in comparisons) the sample sizes being compared are larger in Table 2, resulting in more useable estimates and smaller CVs. For this reason, more differences are detected.

Table 2: Comparison of estimates for easy-to-reach respondents vs. estimates for the entire sample based on 21 key health indicators

Domain	*Total number of estimates*	*Percentage of estimates found to be significant (α=0.05)*
Canada	21	62%
Five age groups by gender	204	11%
120 sub-provincial geographic domains	2 239	13%

The significant differences in Table 2 show that difficult-to-reach respondents are different enough and/or have large enough contributions to estimates that treating them as nonrespondents could result in bias. Furthermore, unless these differences are treated through weighting, this indicates the current CCHS estimates could be subject to bias due to the increase in nonresponse. This assertion relies on the validity of difficult-to-reach respondents being a good proxy for “extra” nonrespondents.

Preliminary investigations showed that the difficult-to-reach respondents had sociodemographic characteristics different than the easy-to-reach respondents. If these differences are correlated to health they could explain significant differences of survey variables found in Tables 1 and 2, despite comparing within age/sex group or geographic area. Of interest is to determine the potential for nonresponse bias currently experienced in the CCHS that is attributable to the decrease in response rates. The CCHS data undertakes many weighting adjustments aimed at improving representivity, and characteristics like known sociodemographics totals are used in the adjustments. A nonresponse adjustment that distributes the weights of the nonrespondents amongst the respondents based on characteristics correlated to response that are in common between respondents and nonrespondents is performed. Calibration is used to adjust the weights of respondents to ensure survey totals are consistent with known totals from auxiliary sources (such as population counts). If the auxiliary variables used in calibration are correlated to response or the survey variables, then calibration can improve estimates. If these steps effectively adjust weights in order to account for nonresponse then estimates may not show bias. The numbers in the following table are based on estimates after weighting adjustments similar to those done in production have been applied. The results reflect the estimates that would be obtained if the easy-to-reach respondents were the set of respondents obtained in the given year. If the manner for simulating nonrespondents in the paper reflects the mechanism for the additional nonresponse currently experienced then these differences represent the error that may be part of the currently produced survey estimates.

Table 3 shows that significant differences remain after the easy-to-reach respondents have been reweighted to represent both the easy-to-reach respondents and the proxy nonrespondents. The reweight has reduced the number of significant differences when compared to Table 2; it compensates in part for the loss of the proxy nonrespondents, but not fully.

Table 3: Comparison of estimates for easy-to-reach respondents vs. estimates for the entire sample based on 21 key health indicators after the samples have been reweighted

Domains	*Total number of estimates*	*Percentage of estimates found to be significant (α=0.05)*
Canada	21	24%
Five age groups by gender	204	7%
120 sub-provincial geographic domains	2 239	7%

2.1 Trend Analysis of the Differences

Table 3 demonstrates that significant differences are detected between the reweighted estimates based on the sample of easy-to-reach respondents and the entire sample of respondents in 2007. Whether the differences are persistent among certain variables or systematic in direction (i.e., new estimates are consistently higher or consistently lower) is unclear. In order to determine if the differences had a systematic element, a variable of change in estimate, $\hat{C}$, was created. $\hat{C}$ can be expressed as:

$$\hat{C} = (\hat{Y}_{original} - \hat{Y}_{new}) \qquad (2)$$

Where $\hat{Y}_{original}$ represents an estimate of interest based on the full sample of respondents and $\hat{Y}_{new}$ represents the estimate based on the reweighted easy-to-reach respondents. If $\hat{C}$ is positive then there was a decrease in the estimate when the proxy nonrespondents are excluded. When $\hat{C}$ is negative, then the new estimate is higher. $\hat{C}$ can be calculated by any domain of interest for all variables of interest. The following boxplots reveal that certain variables are consistently lower or higher across domains when the proxy nonrespondents are excluded. Not all differences are significant, but regardless of whether the difference is enough to be significant, there is a trend of bias. Ideally, if the change in estimates as represented by $\hat{C}$ was due only to sampling error (in the subsampling of nonrespondents) then boxplots of $\hat{C}$ would be centered around zero, ideally with little variance.

Figure 3 shows boxplots of $\hat{C}$ for 10 different domains within each variable. In Figure 3, the domains of interest are the five age groups by sex. Looking at the variable *regular drinker* (listed as “alcohol” in Figure 3), for all five age groups and both sexes, $\hat{C}$ has a positive value. This result implies that the original estimate for regular drinkers is consistently higher than the new estimate. Another way to express this is that the proxy nonrespondents are less likely to be regular drinkers across all age groups by sex. Furthermore, current weighting techniques fail to correct for this difference. Other variables with a systematic trend are *consumption of less than five servings of fruits and vegetables*, *having high blood pressure*, *having poor health*, *being a regular smoker* and *having very good health*. The highlighted boxplots have a minimum of eight of the 10 domain estimates being consistently higher or lower.

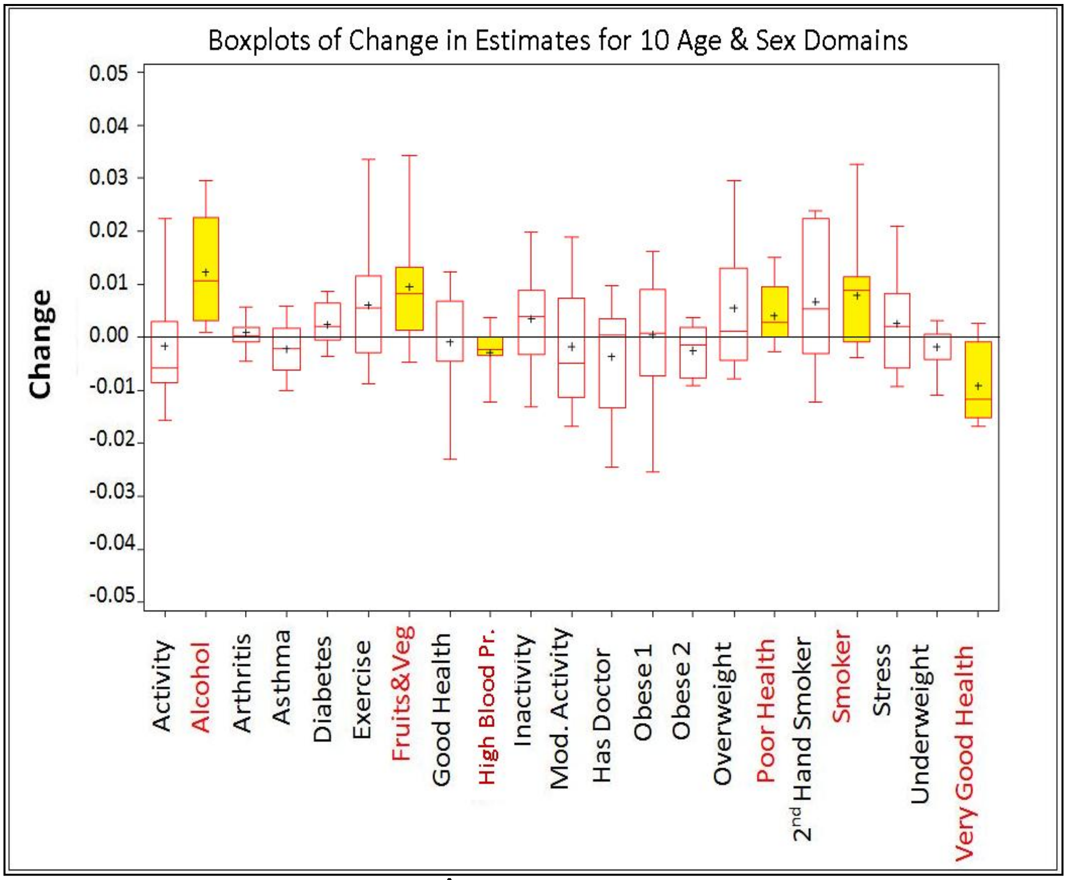

Figure 3: The change in estimate, Ĉ, calculated for 10 domains of age group by sex within each of the 21 variables of interest.

Figure 4 plots the spread in Ĉ for the same 21 variables but by province. Variables such as *consumption of less than five servings of fruits and vegetables*, *having poor health* and *having very good health* again demonstrate bias. At the provincial level, the variables *has a doctor* and *exposure to second hand smoke* are consistently affected, with eight of the 10 estimates being affected in the same direction.

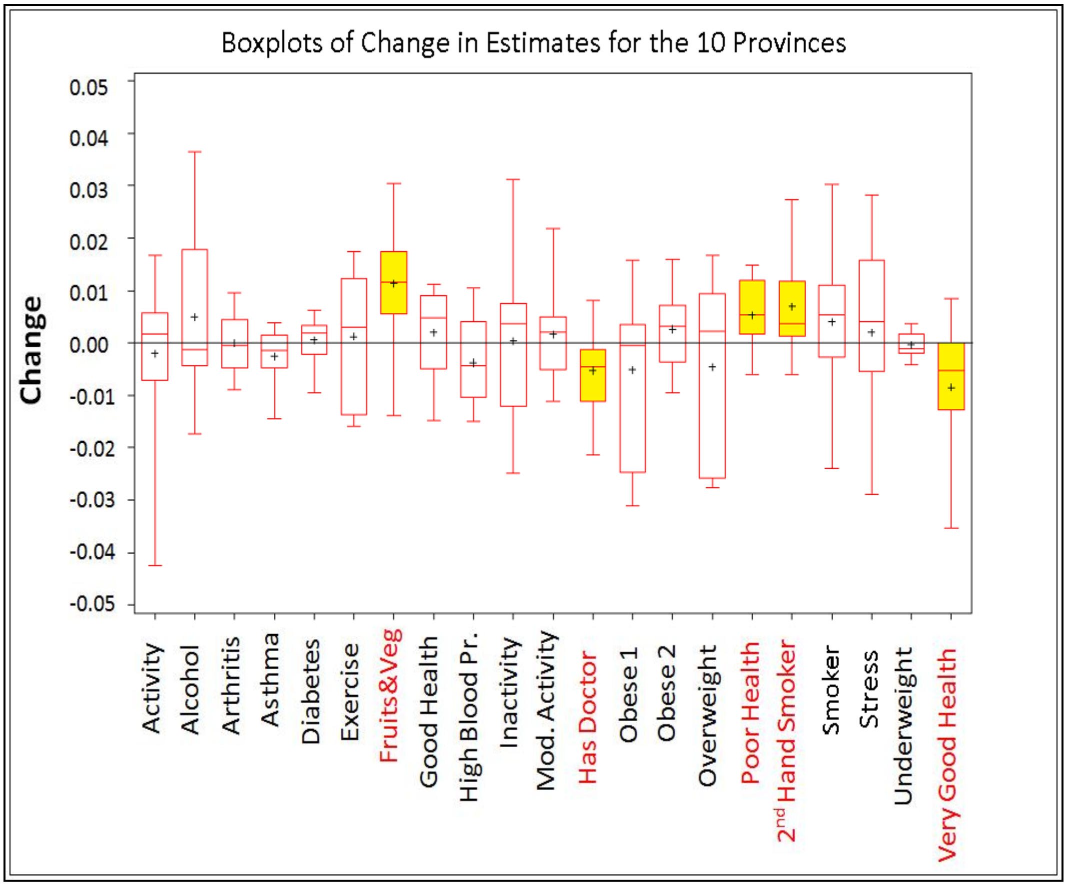

Figure 4: The change in estimate, Ĉ, calculated for 10 geographic domains (province) within each of the 21 variables of interest.

3. Conclusions

Nonresponse bias is a function of the magnitude of nonresponse as well as the magnitude of the difference between respondents and nonrespondents for survey variables. This paper demonstrated that the difficult-to-reach respondents, or proxy nonrespondents, had different characteristics than the easy-to-reach respondents. These differences could result in nonresponse bias if they are not corrected through weight adjustments. Although re-weighting the reduced sample of easy-to-reach respondents helped to bridge the gap in differences of survey estimates, it did not fully capture the proxy nonresponse. Some variables were systematically biased, despite efforts to re-weight. The implication that increased bias exists in current survey estimates hinges on how well the proxy nonrespondents in the study represent the current increase in nonresponse. If characteristics of the difficult-to-reach respondents provide a good estimate of the characteristics of the extra nonrespondents in the current year, then changes in trend estimates of variables may not represent true population changes but an increase in nonresponse bias. The desired shift to more CATI collection may come with a reduction in costs, but also an increase in nonresponse bias that is not fully adjusted through weighting. The potential for nonresponse bias due to the nonrespondents of 2007 has not been evaluated in this paper.

Acknowledgements

The author would like to thank François Brisebois, Marie-Claude Duval and Steven Thomas for their ideas in setting up this study and input throughout the process.

References

The Canadian Community Health Survey User Guide (2013), Statistics Canada.

Duval, M-C. (2012). Biais et étude de non-réponse dans l'ESCC, Internal document, Statistics Canada.

Montaquila, J.M., Olson, K.M. (2012). Practical Tools for Nonresponse Bias Studies, SRMS/AAPOR Webinar.