Methodology Branch

Direction de la méthodologie

Household Survey Methods Division

Division des méthodes d'enquêtes auprès des ménages

Statistics Canada Statistique Canada

Canada

WORKING PAPER
METHODOLOGY BRANCH

Income Imputation for the Canadian Community Health Survey

HSMD-2013-003E

Chi Wai Yeung and Steven Thomas

Household Survey Methods Division
Statistics Canada

April 2013

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada

ABSTRACT

The Canadian Community Health Survey (CCHS) is a cross-sectional survey that collects health information on the Canadian population. Total household income is asked during the interview because of its potential association with health variables. Evidence suggests that there is a potential for nonresponse bias for total household income. Imputation is required to reduce this bias and preserve the income distribution so proper analysis between income and health variables can be done. This paper begins by exploring potential sources for income imputation before describing the final approach in detail. Simulation results are presented and show that the imputation strategy worked well at ensuring that imputed values are often within the same quintile as what would have been reported. Simulations of the resulting estimates show that analyses that include imputed income values are different from analyses based on only income respondents, which suggests a possible nonresponse bias reduction.

RÉSUMÉ

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête transversale qui vise à recueillir des renseignements sur l'état de santé de la population canadienne. Le revenu total du ménage est recueilli lors de l'interview puisque celui-ci est un facteur pouvant potentiellement être associé aux variables liées à la santé. Selon toute indication, il peut y exister un biais de non-réponse pour le revenu total du ménage. Un processus d'imputation est donc requis pour réduire ce biais et préserver la distribution du revenu afin de pouvoir effectuer des analyses appropriées impliquant le revenu et des variables sur la santé. Des sources possibles pour l'imputation du revenu sont d'abord examinées dans ce document, puis l'approche retenue est décrite plus en détail. Les résultats des simulations sont présentés et montrent que la stratégie d'imputation adoptée réussit à produire des valeurs se situant souvent dans le même quintile que ce qui aurait été la valeur déclarée. Les estimations produites à partir des simulations montrent que les résultats d'analyses incluant les revenus imputés sont différents de ceux produits seulement à partir des données rapportées, ce qui sous-entend une réduction potentielle du biais de non-réponse.

Income Imputation for the Canadian Community Health Survey

Chi Wai Yeung and Steven Thomas
Statistics Canada, Household Survey Methods Division

1. Introduction

Income questions are asked during the Canadian Community Health Survey (CCHS) interview because of their potential association with health variables. The income variable most commonly used in analyses is total household income (variable INC_3). Because of the sensitive nature of reporting income, the response rate for this variable is usually between 65% and 70% among CCHS respondents.

In addition to the low response rate, a study conducted in 2009 (Sarafin 2009) suggested that the income nonrespondents had different health characteristics than the respondents. For example, a higher proportion of income respondents rated their general health and mental health to be very good or excellent. Because of this difference in health profiles and the low response rate, analyses that are based exclusively on income respondents may be biased. It was therefore decided to impute total household income starting with the release of the 2011 CCHS data file.

This paper describes the imputation process for total household income of the CCHS. Section 2 describes the income module of the CCHS while section 3 provides an overview of the potential sources of data that can be used in the imputation process. Section 4 summarizes the income imputation methods of some Statistics Canada surveys. Section 5 provides details of the derivation of modeled household income, which is used during the actual imputation process. The imputation strategy is described in section 6. The performance of the imputation process is evaluated with a simulation, which is the topic of section 7. The impact of imputing income is described in section 8 when the proportions of some health indicators are compared before and after imputation. Section 9 is the conclusion.

2. Income on the CCHS

Although many health expenses are covered by health insurance, there is still a relationship between health and income. The CCHS collects income information because of this relationship but it should not be considered the official source of income statistics. Various questions related to income are collected including income sources, personal income of the respondent and household income. Household income is the variable of interest in most analyses and is collected through the following question. It appears on the data files as INC_3:

> **INC_3:** ***What is your best estimate of the total income received by all household members, from all sources, before taxes and deductions, in the past 12 months?***

When this variable cannot be reported either because of an unwillingness to respond or the respondent is not able to respond, a series of questions that attempt to determine an income range are offered as a means to estimate the income for the household:

INC_5A: ***Can you estimate in which of the following groups your household income falls? Was the total household income in the past 12 months...?***

1 Less than $50,000 include income loss (Go to 5B)
2 $50,000 and more (Go to 5C)

INC_5B: Please stop me when I have read the category which applies to your household. Was it...?

1 Less than $5,000
2 $5,000 to less than $10,000
3 $10,000 to less than $15,000
4 $15,000 to less than $20,000
5 $20,000 to less than $30,000
6 $30,000 to less than $40,000
7 $40,000 to less than $50,000

INC_5C: ***Please stop me when I have read the category which applies to your household. Was it...?***

1 $50,000 to less than less than $60,000
2 $60,000 to less than less than $70,000
3 $70,000 to less than less than $80,000
4 $80,000 to less than less than $90,000
5 $90,000 to less than less than $100,000
6 $100,000 to less than less than $150,000
7 $150,000 and over

The unweighted response rate for INC_3 amongst respondents to the CCHS is 66.4% in 2011. The unweighted response rates at the provincial level in 2011 are presented in Table 1. Unweighted response rates to the variables INC_5B or INC_5C among nonrespondents to INC_3 are presented in Table 2.

Table 1: Response rates of INC_3 in 2011 by province

N.L	PEI	N.S	N.B	QC	ON	MB	SK	AB	BC	YT	NWT	NU
65.8%	62.9%	68.8%	70.3%	70.2%	65.5%	63.8%	56.4%	67.4%	66.6%	72.5%	66.4%	63.3%

Table 2: Response rates to income range among nonrespondents to INC_3 by province

N.L	PEI	N.S	N.B	QC	ON	MB	SK	AB	BC	YT	NWT	NU
66.6%	56.8%	56.0%	50.1%	54.0%	37.7%	47.4%	51.8%	40.4%	41.2%	37.6%	52.6%	42.9%

3. Potential Sources for Income

In researching an income imputation strategy, several sources were investigated. Two possible sources for income are tax files from the Canada Revenue Agency and the income module from the National Household Survey or the Census long form.

Canada Revenue Agency's T1 Family File

Income is available on the T1 Family File (T1FF), which provides a recent household income value as reported to the Canada Revenue Agency. There are some challenges with using this file. First is the difference in reference periods between the CCHS and the T1FF. Under the current CCHS production schedule, the most recent T1FF has reference period *t-2* for CCHS reference year *t*. In addition, while the T1FF's reference period is the calendar year, the CCHS income module asks about income for the 12 months prior to the date of the interview. Second, it is known that the T1FF has an under-coverage problem for those 18 to 24 years old. This group is often not reported under a parent's tax form and would often not have income of their own to report. Third, linkage to the CCHS is hampered by the fact that proper identification is not mandatory for the CCHS. Finally, any linkage to tax information would have to have permission from the respondent. Unfortunately, the CCHS does not ask permission to link to tax files and this information would not be publishable on any standard share file.

The Census and the National Household Survey

Income has traditionally been collected as part of the Census long form and more recently in 2011 as part of the National Household Survey. It is a valuable source of income for the CCHS since the information is collected for all household members aged 15 and over. The issues with using this data source are similar to using the tax files. First, the permission to link to Census information is not explicitly asked during the CCHS interview. Second, potential linkage variables such as names and addresses are not available for all CCHS respondents. This makes linking to the correct census household difficult. Third, the file would only be available every 5 years and would be delayed by as much as two years after the Census took place. Finally, it was not clear if any of the information coming from the Census would be of higher quality than the data coming from the T1FF.

4. Income Imputation of Other Statistics Canada Surveys

Along with researching potential data sources, possible methods for imputing were examined by looking at the methods used by other survey's for imputing income. Two other Statistics Canada surveys that impute income are the Survey of Household Spending (SHS) and the National Household Survey (NHS). Their processes were considered as potential options for the CCHS imputation process. Overviews of their imputation methods are provided here mainly to show that their processes cannot be directly replicated in CCHS. The surveys' documentation should be consulted for more details.

Survey of Household Spending (SHS)

There are two main differences between the SHS and the CCHS in terms of income imputation. First of all, the SHS asks for permission to link to tax information while the CCHS does not. As a result, the SHS has greater flexibility in linking and publishing imputed income from tax data. Secondly, the SHS collects detailed expenditures that are highly correlated to income. These expenditures are used by the SHS in modeling household income when it is not reported. The CCHS does not have such information.

Census of Population
Starting with the 2006 long form Census and more recently with the National Household Survey, permission to link to tax data is asked of all respondents aged 15 or over (Bankier 2006). The CCHS does not have such permission and is more limited in publishing imputed income from tax data. Another difference between the Census and the CCHS is that the Census collects labor activity status (e.g. employed vs. not working) from every member of the household. Labor status is highly correlated to income. Unfortunately, such information is available only for the CCHS selected personal respondent instead of all household members. As a result, the Census income imputation method cannot be directly applied to the CCHS.

5. Modeled Household Income

The CCHS income imputation uses a nearest neighbor donor imputation strategy. The nearest neighbor measure is based on a modeled household income which can be calculated for every responding household using a four step process. Details of the modeling process are given below.

Step 1: Model Person-level income

The first step of the imputation process is to model personal income for *all* household members. Modeling household income directly is challenging because of its relationship with household structure. During the CCHS interview, detailed information on age, gender, and education is collected for *every* member of a household. Having such detailed information is useful in modeling personal income values. A preliminary value for household income can be obtained by summing personal income from all household members.

Besides age, gender, and education of the individuals, some household characteristics such as source of income and whether the dwelling is owned or rented are significant in modeling personal income. Another observation is that there are important differences in personal income between males and females. Income values for females are more heterogeneous than males. Many females work full-time while others do not work at all. On the other hand, most males work full-time. Therefore, we use separate regression models for males and females. In addition, examination of residual plots shows non-constant variances if we directly model personal income. This violates the assumptions of linear regression models and thus log-transformed personal income is used as the dependent variable in the models. Even though there are small differences in the models between years, typical models for males and females are respectively:

$$\begin{aligned}&\log(male\ personal\ income) = \\ &\beta_1 age\ group + \beta_2 education + \beta_3 main\ source\ of\ household\ income + \beta_4 health\ region + \\ &\beta_5 martial\ status + \beta_6 household\ size + \beta_7 dwelling\ owned\ or\ rented + \\ &\beta_8 home\ during\ the\ day\ on\ weekday + \beta_9 kids\ in\ household\end{aligned}$$

$$\log(female\ personal\ income) = \beta_1 age\ group + \beta_2 education + \beta_3 main\ source\ of\ household\ income + \beta_4 health\ region + \beta_5 martial\ status + \beta_6 household\ size + \beta_7 dwelling\ owned\ or\ rented + \beta_8 home\ during\ the\ day\ on\ weekday$$

Once modeled, the predicted values are transformed back to the original scale to estimate personal income.

Step 2: Sum person-level values to obtain household income

A preliminary household income value is derived by summing personal incomes from all household members from step 1.

Step 3: Derive median tax value

As stated above, there are dissemination issues with using information obtained by directly linking to the T1FF. On the other hand, the *median* tax value by postal code and household size is also correlated with household income, and there is no permission or dissemination concern with using median tax value. As a result, median tax value from the most updated T1FF will be used to improve the preliminary household income estimate. Under the current CCHS production schedule, T1FF from reference period *t -2* is used for reference period *t* in CCHS.

An issue with deriving household size on the T1FF is that people are not required to file tax until they earn income. Therefore, there is under-coverage of people between 18 and 24 years old on the T1FF. This has an impact when we compare household size between the tax file and the CCHS. To remedy this, only people aged 25 or more are counted when obtaining household size on both the CCHS and the T1FF. To avoid having too many categories, the number of people aged 25+ in a household is limited to one, two, and three or more.

Median tax value is derived first by postal code and the number of people aged 25+. If there are fewer than 5 units within each class, we compute median tax value from just the postal code. If there are still fewer than 5 units, we look at the first 3 digits of the postal code and the number of people aged 25+. We want a minimum of 5 units to ensure some quality of the median value. For the 2011 CCHS (2009 T1FF), the following table shows the level at which the median value is obtained:

Table 3: Level at which median tax value is obtained on the 2009 T1FF

Level	No. of units (Percentage)
Postal code x No. of people aged 25+	52,433 (82.5%)
Postal code	8,200 (12.9%)
First 3 digits of postal code x No. of people aged 25+	2,909 (4.6%)
Total	63,542 (100%)

Step 4: Improve modeled value from step 2

As a final step, a regression model is used to combine the median tax value from step 3 and the preliminary household income from step 2. The use of regression eliminates the issue of different reference periods between the CCHS and the T1FF.

In addition to median tax value and preliminary household income, some health variables are used as predictors in the model. Even though we only have health variables from the selected personal respondent, his / her health is significant in predicting total household income. For example, if the selected personal respondent has an activity limitation, his / her income may be lower than another identical person without such limitation. Furthermore, other household members may need to take time off from work or work shorter hours to help the selected respondent who has activity limitation. This reduces the earning power of other household members.

It should be mentioned that initially using health variables in the model was avoided because we did not want to artificially emphasize the relationships between such health outcomes and income with the imputation process. However, the use of donor imputation (instead of direct imputation such as regression imputation) should introduce enough noise to prevent artificially emphasizing those relationships.

Examination of the residual plot shows that log transformation is again needed. There are small differences in the final model between years. In 2012, the model looks like:

$$\begin{aligned}\log(INC\_3) = \; & \beta_1 \log(preliminary\ household\ income) + \beta_2 \log(median\ tax) \\ & + \beta_3 presence\ of\ Alzheimer's\ disease + \beta_4 immigrant\ flag \\ & + \beta_5 heavy\ drinker + \beta_6 activity\ limitation + \ \beta_7 daily\ smoker \\ & + \beta_8 general\ health\end{aligned}$$

The final modeled household income in 2012, which is derived for both respondents and non-respondents to household income, is then:

$$\begin{aligned}& Final\ modeled\ household\ income \\ & \quad = \exp\{\hat{\beta}_1 \log(preliminary\ household\ income) + \hat{\beta}_2 \log(median\ tax) \\ & \quad + \hat{\beta}_3 presence\ of\ Alzheimer's\ disease + \hat{\beta}_4 immigrant\ flag \\ & \quad + \hat{\beta}_5 heavy\ drinker + \hat{\beta}_6 activity\ limitation + \ \hat{\beta}_7 daily\ smoker \\ & \quad + \hat{\beta}_8 general\ health\}\end{aligned}$$

6. Imputation Process

Following the suggestion of the Technical Committee on Household Surveys, a nearest neighbor (NN) donor imputation method is used to impute income. Under mild assumptions, NN imputation preserves distribution of income (Chen & Shao 2000). This is important since an objective of imputing income is to preserve its distribution so analyses between income and health variables are not biased.

Statistics Canada's generalized system BANFF is used to implement the NN donor imputation. The modeled household income defined above is used as the distance measure to define which pair of donor-recipient is the closest. As much as possible, within each imputation class, we require at least 30% of the units to be donors as well as a minimum of 10 donors before imputation can take place. The intent is to prevent the same donor from being used too many times which could affect the household income distribution.

Imputation classes are formed differently depending on the amount of information the income nonrespondents provide. Units who provide INC_5A and INC_5B / INC_5C give enough information to derive full income ranges. On the other hand, some units provide INC_5A but not INC_5B or INC_5C. Only broad income ranges of less than $50,000 and $50,000 or more can be derived for these units. The remaining income nonrespondents do not have any income ranges.

For units who provide a detailed income range, imputation classes are based on the full income range and household size (1, 2, 3, 4, and 5+), at the national level. If the detailed income range by household size domain does not have the required percentage or number of donors, we will collapse household size 3, 4, and 5+ to form household size 3+. If after collapsing there are still not enough donors, imputation classes will then be formed by just the detailed income range.

For units who provide a broad income range, imputation classes are formed based on the rough income range and household size (1, 2, 3, 4, 5+), at the national level. The rules of collapsing are similar to those who provide the detailed income range. In other words, we first collapse household size 3, 4, and 5+ before using only the rough income range.

For people who do not provide any income range information, imputation classes are formed based on health region and household size. The rules of collapsing are similar to the units who provide their income ranges.

For some data products such as 2 month rapid response, it is possible that there are still not enough donors in the health region. In this case, province and household size (1, 2, 3, 4, and 5+) are used before the same rules of collapsing are applied to the province.

The following list summarizes the steps of forming imputation classes:

- Step 1 (units with detailed income range): detailed income range by household size (1, 2, 3, 4, and 5+)
- Step 2 (units with detailed income range): detailed income range by household size (1, 2, 3+)
- Step 3 (units with detailed income range): detailed income range
- Step 4 (units with broad income range): broad income range by household size (1, 2, 3, 4, and 5+)
- Step 5 (units with broad income range): broad income range by household size (1, 2, 3+)
- Step 6 (units with broad income range): broad income range
- Step 7 (units without income range): health region by household size (1, 2, 3, 4, and 5+)
- Step 8 (units without income range): health region by household size (1, 2, 3+)
- Step 9 (units without income range): health region
- Step 10 (units without income range): province by household size (1, 2, 3, 4, and 5+)
- Step 11 (units without income range): province by household size (1, 2, 3+)

- Step 12 (units without income range): province

7. Simulation

The quality of the imputation process is assessed through a simulation based on 2011 CCHS data. An important objective of the imputation process is to correctly capture the income distribution. We are mainly interested in how well the imputation preserves income quintiles instead of the exact income. This section gives some details of the simulation set up before providing the results.

Simulation Set up

The 2009 report by Sarafin suggests that nonresponse to total household income is not uniform but correlated with some health variables. As a result, we set up the simulation in the following manner:

A. Using the significant variables from the report, we derive the probability of income nonresponse for every unit using logistic regression. We denote this probability p_i for unit i. The mean of p_i is around 33.5% for 2011.

B. We then generate a separate random number for each income respondent. If the random number is less than p_i then the reported total household income is temporarily blanked out.

C. Among units with total household income blanked out in step B, we further blank out 50% of the detailed income range. Then 75% of the units with detailed income range blanked out have their broad income range discarded. This setup is done to evaluate the performance of imputation with broad income range and imputation without any income range.

D. Among units with total household income blanked out in step B, we further blank out 75% of personal income (INC_8A). This extra randomness is needed to study the robustness of the modeled household income and its impact on the imputation.

E. Using the dataset after step D, we construct modeled household income using the 4 step process outlined above.

F. Nearest neighbor imputation is carried out in BANFF using the modeled household income from step E as distance measure.

G. We classify all income respondents into quintiles based on the reported income, regardless of whether they were blanked out in step B. We call these the reported quintiles. For these same units, we also classify them into quintiles based on income after imputation[1]. We call these the imputed quintiles. We then compare the reported quintiles and the imputed quintiles to assess how many units change quintiles.

H. Repeat steps A to G multiple times.

Simulation Results

The simulation was run 200 times and the results are reported in the following quintile transition tables. For each reported quintile (column), the figures are percentages of units that move to each of the imputed quintile. Each column sums to 100%. Using the 2nd column of the next table as an example, 94.7% of units with reported income in the 2nd quintile have imputed value in the 2nd quintile. Intuitively, we would like high percentages in the diagonal elements. The diagonal elements are highlighted for ease of comparison.

[1] For units not blanked out, income after imputation is still the reported income

Table 4 shows the transition of units imputed within detailed income ranges. In other words, it is the overall result of forming imputation class using steps 1 to 3 described in section 6. The result is excellent because the boundaries of the income range are quite narrow and correspond well to the quintiles.

Table 4: Quintile transition table for units imputed within detailed income range

		Reported quintile				
		1st	**2nd**	**3rd**	**4th**	**5th**
Imputed quintile	**1st**	100%	5.3%	0%	0%	0%
	2nd	0%	94.7%	1.0%	0%	0%
	3rd	0%	0%	99.1%	0.9%	0%
	4th	0%	0%	0%	99.1%	0%
	5th	0%	0%	0%	0%	100%

Table 5 shows the transition of units imputed within broad income ranges. In other words, it is the overall result of forming imputation class using step 4 to 6 described in section 6. The result is not as good as the first transition table because the broad income ranges are not as detailed and do not correspond well to the quintiles.

Table 5: Quintile transition table for units imputed within partial income range

		Reported quintile				
		1st	**2nd**	**3rd**	**4th**	**5th**
Imputed quintile	**1st**	64.5%	18.8%	0.1%	0%	0%
	2nd	35.5%	81.2%	0.9%	0%	0%
	3rd	0%	0%	47.8%	25.6%	14.4%
	4th	0%	0%	27.4%	36.9%	24.8%
	5th	0%	0%	23.9%	37.6%	60.7%

Table 6 shows the transition of units imputed without any income range information. In other words, it is the overall result of forming imputation class using steps 7 to 12 from section 6. The result is not as good as the transition table for units imputed within income ranges but is deemed acceptable. Many users are especially interested in the low income and high income families and the result below shows that the 1st and 5th quintiles have decent quality.

Table 6: Quintile transition table for units imputed without income range

Imputed quintile		Reported quintile				
		1st	2nd	3rd	4th	5th
	1st	61.0%	10.2%	3.8%	2.3%	1.4%
	2nd	25.1%	48.6%	16.2%	9.7%	4.9%
	3rd	7.4%	22.6%	40.7%	19.5%	10.7%
	4th	3.6%	11.4%	23.3%	37.6%	22.3%
	5th	2.9%	7.2%	16.1%	31.0%	60.6%

Table 7 shows the transition of all imputed units. In other words, it is the weighted average of the first three transition tables.

Table 7: Quintile transition table for all units that are imputed

Imputed quintile		Reported quintile				
		1st	2nd	3rd	4th	5th
	1st	80.8%	8.9%	1.4%	0.9%	0.5%
	2nd	13.9%	75.7%	6.7%	3.6%	1.9%
	3rd	2.8%	8.5%	70.7%	11.0%	5.8%
	4th	1.4%	4.3%	12.2%	68.2%	11.4%
	5th	1.1%	2.7%	9.0%	16.3%	80.4%

Finally table 8 shows the transition of all units, including those people whose values are not temporarily blanked out. It gives an idea of the distortion to the entire dataset as a result of imputation. Since most diagonal elements have value around 90%, we can conclude that the imputation process preserves income quintiles to a large extent.

Table 8: Quintile transition table for all units

Imputed quintile		Reported quintile				
		1st	2nd	3rd	4th	5th
	1st	93.5%	6.5%	0.5%	0.3%	0.2%
	2nd	4.7%	88.3%	2.9%	1.2%	0.6%
	3rd	1.0%	2.8%	89.7%	4.2%	1.9%
	4th	0.5%	1.4%	4.0%	89.0%	3.8%
	5th	0.4%	0.9%	3.0%	5.4%	93.6%

8. Compare Health Indicators before and after Imputation

Income imputation was set up to reduce bias from analyses excluding income nonrespondents. To assess the impact of the imputThe proportions before imputation are computed from only the income respondents. The rates after imputation are computed from all units. The quintiles[2] are re-derived after imputation so the same unit can have different quintiles after imputation.

[2] The quintiles make use of INCDRCA, which takes into consideration low income cut-off of different family size and population size

The differences are tested to see if they are statistically significant using bootstrap weights and the BOOTVAR macro. The 2009-2010 file is used since it is the most recent two year file available. The two year file was used instead of an annual file to give a larger sample size and more power to the statistical tests. The following tables present the proportions before and after imputation alongside the p-value of testing if the rates are significantly different. Indicators that are significantly different at α = 0.05 are highlighted for ease of reference.

Table 9: Comparing health indicators for the first income quintiles, before and after imputation

	Proportion from 1st income quintile		
Health indicator	**Before**	**After**	**p-value**
Current smoker	0.272	0.259	0.0000
Exposure to second-hand smoke at home	0.158	0.149	0.0004
Arthritis	0.212	0.198	0.0000
Asthma	0.100	0.098	0.3447
Diabetes	0.093	0.086	0.0000
High blood pressure	0.220	0.204	0.0000
Have regular medical doctor	0.808	0.813	0.1203
Self rated health - poor or fair	0.210	0.197	0.0000
Self rated health - good	0.323	0.320	0.4636
Self rated health - very good	0.292	0.300	0.0144
Self rated health - excellent	0.176	0.183	0.0264
Heavy drinker	0.137	0.140	0.3085
Life stress - quite a lot	0.236	0.234	0.4829

Table 10: Comparing health indicators for the second income quintiles, before and after imputation

	Proportion from 2nd income quintile		
Health indicator	**Before**	**After**	**p-value**
Current smoker	0.218	0.216	0.5265
Exposure to second-hand smoke at home	0.126	0.127	0.6856
Arthritis	0.179	0.182	0.3495
Asthma	0.085	0.086	0.5076
Diabetes	0.071	0.075	0.0376
High blood pressure	0.194	0.198	0.2178
Have regular medical doctor	0.847	0.854	0.0138
Self rated health - poor or fair	0.128	0.133	0.0345
Self rated health - good	0.328	0.324	0.3607
Self rated health - very good	0.350	0.349	0.7772
Self rated health - excellent	0.194	0.193	0.7951
Heavy drinker	0.135	0.130	0.1870
Life stress - quite a lot	0.210	0.211	0.8970

Table 11: Comparing health indicators for the third income quintiles, before and after imputation

	Proportion from 3rd income quintile		
Health indicator	**Before**	**After**	**p-value**
Current smoker	0.210	0.201	0.0069
Exposure to second-hand smoke at home	0.114	0.106	0.0096
Arthritis	0.142	0.144	0.5637
Asthma	0.079	0.081	0.2618
Diabetes	0.058	0.057	0.6369
High blood pressure	0.153	0.155	0.4758
Have regular medical doctor	0.850	0.851	0.6653
Self rated health - poor or fair	0.094	0.099	0.0337
Self rated health - good	0.277	0.285	0.0323
Self rated health - very good	0.409	0.393	0.0002
Self rated health - excellent	0.220	0.223	0.4178
Heavy drinker	0.140	0.137	0.3665
Life stress - quite a lot	0.212	0.206	0.1159

Table 12: Comparing health indicators for the fourth income quintiles, before and after imputation

	Proportion from 4th income quintile		
Health indicator	**Before**	**After**	**p-value**
Current smoker	0.183	0.187	0.2088
Exposure to second-hand smoke at home	0.092	0.100	0.0036
Arthritis	0.120	0.128	0.0023
Asthma	0.075	0.075	0.7651
Diabetes	0.044	0.047	0.1117
High blood pressure	0.138	0.150	0.0000
Have regular medical doctor	0.857	0.856	0.8295
Self rated health - poor or fair	0.068	0.078	0.0000
Self rated health - good	0.258	0.257	0.7033
Self rated health - very good	0.420	0.416	0.2529
Self rated health - excellent	0.253	0.250	0.3841
Heavy drinker	0.166	0.163	0.2974
Life stress - quite a lot	0.239	0.227	0.0020

Table 13: Comparing health indicators for the fifth income quintiles, before and after imputation

	Proportion from 5th income quintile		
Health indicator	**Before**	**After**	**p-value**
Current smoker	0.154	0.156	0.4050
Exposure to second-hand smoke at home	0.070	0.077	0.0004
Arthritis	0.123	0.125	0.3895
Asthma	0.075	0.074	0.3144
Diabetes	0.041	0.045	0.0199
High blood pressure	0.143	0.143	0.9871
Have regular medical doctor	0.871	0.873	0.3661
Self rated health - poor or fair	0.055	0.058	0.0979
Self rated health - good	0.221	0.233	0.0001
Self rated health - very good	0.428	0.424	0.3168
Self rated health - excellent	0.296	0.285	0.0003
Heavy drinker	0.170	0.167	0.1005
Life stress - quite a lot	0.268	0.261	0.0161

The tables above show that several indicators change after imputation. They give some indication of the bias from analyses that exclude income nonrespondents. In other words, they show the impacts and importance of having an imputation process in place.

9. Conclusion

A previous study has shown that imputation is needed to reduce bias in analyses of CCHS data involving total household income. This paper provides an overview of the imputation process implemented starting with the release of the 2011 CCHS data file. Different potential sources of income data such as the T1FF and the Census are reviewed. In addition, the imputation methods of Survey of Household Spending and the Census of Population are shown to be not suitable for the CCHS. A nearest neighbor donor imputation strategy is proposed using the suggested modeled household income as distance function. Simulation shows that to a large extent, the income distribution is preserved after imputation. Comparisons of analyses before and after imputation show significant differences for several health indicators, which confirm the need of having the imputation process in place.

References

Bankier, M. (2006), "Editing a Mixture of Canadian 2006 Census and Tax Data", Working Paper 8, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Germany.

Chen, J. and Shao, J (2000), "Nearest Neighbor Imputation for Survey Data", Journal of Official Statistics, Vol. 16, No. 2, 2000, pp. 113-131.

Sarafin, C. (2009), "CCHS Health Indicators by Income Quintiles", Statistics Canada Internal Report.

Thomas, S. and Yeung, C.W. (2012), "Income imputation for the Canadian Community Health Survey", Presentation to the Technical Committee on Household Surveys http://method/BiblioStat/Research/TechCom/HouseholdSurveys/Minutes/meeting54_e.htm.