Methodological Issues in Analyzing Small Populations Using CCHS Cycles Based on the Official Language Minority Studies

Ewa Makvandi, MSc,¹ Louise Bouchard, PhD,² Pierre-Jerôme Bergeron, PhD,³ Golnaz Sedigh, MSc¹

ABSTRACT

Statistical analyses for small populations or small domains of interest can be challenging. To obtain reliable estimates, only very large surveys such as the Canadian Community Health Survey can be considered. However, despite its good geographical and temporal coverage, the analysis of small populations in smaller regions (e.g., health regions) and in regards to specific health issues remains challenging. We will look at the methodological issues in analysis of small populations in relation to sampling and non-sampling survey errors that affect the precision and accuracy of the estimates. Francophone minorities in Canada will be used to illustrate the issues throughout the paper.

KEY WORDS: Health survey; minority health; language; sample size; methods

La traduction du résumé se trouve à la fin de l'article.

Can J Public Health 2013;104(6)(Suppl. 1):S55-S59.

Statistical analyses for small populations or small domains of interest can be challenging. To obtain reliable estimates, only very large surveys can be considered. The Canadian Community Health Survey (CCHS), a cross-sectional multi-stage survey that collects information related to health status, health care utilization and health determinants for the Canadian population aged 12 and older living in private dwellings in the 10 provinces and 3 territories (with the exclusion of populations on Indian Reserves, Canadian Forces Bases, and some remote areas) has a potential for such analyses. However, despite its good geographical and temporal coverage, the analysis of small populations remains challenging. This paper examines some of the methodological issues of utilizing the CCHS surveys for small populations.

Statistical challenges encountered by Réseau de recherche appliquée sur la santé des francophones de l'Ontario (RRASFO) researchers when studying health status of Francophone minority populations in Canada provided motivation for this paper. Since these challenges can be easily generalized to those of other small populations, we will use Francophone minority populations to illustrate the issues discussed throughout the paper. Francophone minority status of respondents in the presented examples was determined using the following CCHS variables: language of conversation, mother tongue, language of interview, and preferred language of household interview (see Appendix A for details).

One of the goals of the CCHS is to provide a single data source for health research on small populations and rare characteristics. However, even with large sample sizes in the CCHS core surveys, it is often not possible to examine all of the health dimensions of small populations using a single CCHS cycle. In practice, only two streams of surveys – the annual component and, to some extent, the focused content – can be applied. The annual component is collected from all survey respondents and remains relatively

unchanged over several years. It is designed to provide reliable estimates at the national, provincial and health region* levels. The focused content surveys which examine specific topics in greater detail (e.g., mental health, nutrition, healthy aging) are collected from representative subsamples and designed to provide reliable estimates only at the national and provincial levels. In some cases, the estimation of specific sub-provincial area characteristics is possible due to the presence of provincial buy-ins that add sample units to the surveys and in which the final sampling unit is the household.¹

METHODOLOGICAL ISSUES

In this study, we look at the methodological issues in analysis of small populations in relation to survey errors that affect the precision and accuracy of the estimates.

Sampling error

In survey methodology, errors are often classified as sampling and non-sampling. Both types of errors are of special concern when

Author Affiliations

1. Réseau de recherche appliquée sur la santé des francophones de l'Ontario, University of Ottawa, Ottawa, ON
2. Institute of Population Health, Montfort Hospital Research Institute, Department of Sociology and Anthropology, University of Ottawa, Ottawa, ON
3. Department of Mathematics and Statistics, Faculty of Sciences, University of Ottawa, Ottawa, ON

Correspondence: Ewa Makvandi, Biostatistician, Réseau de recherche appliquée sur la santé des francophones de l'Ontario, University of Ottawa, Ottawa, ON, Tel : 613-265-6595, E-mail: ewa.makvandi@gmail.com

Acknowledgements of sources of support: This research is funded by the Réseau de recherche appliquée sur la santé des francophones de l'Ontario (RRASFO). We thank Jean-Michel Billette from COOL RDC and David Paton from CIHI, Ottawa for their support and valuable comments.

Conflict of Interest: None to declare.

* Health region (HR) refers to administrative areas defined by the provincial ministries of health.

Appendix A. Language diagram

small populations are considered for analyses. Sampling error is an error that arises from estimating a population characteristic by measuring only a portion of rather than the entire population. It is related to the size of the population under study and can be quantified in terms of sampling variance, standard errors, and confidence intervals. The precision and accuracy of estimates also depend on sample size: they decrease in small samples, especially when small samples from small populations are considered. Inadequate sample size is therefore of main concern for the analysis of small populations as it increases sampling error and leads to estimates that may not be robust. Looking at the least complicated probability sampling method, Simple Random Sampling (SRS), Figure 1 shows the sample sizes required for different population sizes in order to obtain an estimate of a true proportion $p=0.5$ within a 0.05 margin of error with a confidence level of 95%. Table 1 shows sample sizes of Francophone minority populations obtained in our studies on national and provincial levels in the core and some optional components of the CCHS (cycles 1.1, 1.2, 2.1, 2.2, 3.1, 4.1, 2008 and 2009). Though the complexity of the survey design would normally require a larger sample size, the sample sizes for the core components of the CCHS allow the performance of a robust analysis at the national level and in the selected provinces (e.g., Ontario), but are too small, for example, for the province of Newfoundland (Table 1). When further subdivision of a sample in order to investigate the subpopulations in relation to various socio-economic or demographic factors is considered, the problems with the precision of estimates intensify. For example, if one wants to estimate a mean difference intake of caffeine for populations of Francophone and Anglophone men, one will obtain an estimate with a confidence interval so wide that meaningful interpretation of the result becomes impossible: -4.87 mg [-65.36 mg, 55.62 mg]. This example, although extreme, demonstrates the common problem when estimation of a specific characteristic of a small population is desired. Analysis on subprovincial levels of geography, such as health

Figure 1. Example of sample size required to estimate proportion P at 95% level of confidence using SRS strategy with true value of $p=0.05$

regions, further precludes examination of selected health measures for small populations or minority groups. In addition, the prevalence of certain characteristics in a population may affect reliability of the estimates: for rare characteristics, producing reliable estimates for the subpopulations with the attributes becomes impossible without increasing the sample size within the subpopulations (i.e., oversampling). For example, if one wants to investigate prevalence of cancer, stroke, or diabetes in Francophone men and women populations or in Francophones aged 65 and older in Ontario's health regions, none of the surveys could be used alone to obtain meaningful results. The sample size in each of the health regions in this case would be less than 20 (Table 2), and therefore not large enough for the results to be released based on Statistics Canada results dissemination rules. For researchers who are interested in studying small populations defined by detailed geography or by some rare socio-demographic or health-related characteristics, providing reliable estimates only at the national or provincial level is often not satisfactory. Economic and health issues of minority groups are often specific not only to each province or regions within the province, but also to health divisions, and estimates based on these divisions (e.g., health regions) are desirable.

Analysis of small populations can also be challenging when comparisons with the rest of the population are desired. In the case of our population of interest, the Francophone minority sample is relatively small compared with the Anglophone sample (5% vs. 95%). As a result, certain differences cannot reach statistical significance due to power constraints, especially when smaller subpopulations are compared. When further partitioning of the population to account for other socio-economic factors is required, the resulting estimates will exhibit increased variability and will be less efficient.² Studying the Francophone minority's aging population in terms of access to health services such as having a regular doctor, language of conversation with a health professional, number of consultations, etc. is challenging due to very low sample sizes: fewer than

Table 1. Sample Size of Francophone Minority Population in the CCHS Cycles Based on Two Definitions: First Language Learned, and Combination of Language Variables

	Cycle 1.1	Cycle 1.2	Cycle 2.1	Cycle 2.2	Cycle 3.1	Cycle 4.1	Cycle 2008	Cycle 2009	Combined Cycles
First language learned*									
Canada w/o QC	27,204	6856	30,661	6127	32,178	14,119	13,944	13,096	144,185
RRASFO's definition†									
Canada w/o QC	5772	1893	6255	1643	5740	3183	3108	2769	30,363
Newfoundland and Labrador	21	<10	37	14	19	17	<10	12	<140
Prince Edward Island	142	46	111	61	96	44	40	42	582
Nova Scotia	236	106	189	65	223	122	95	98	1134
New Brunswick	1779	608	2017	560	1927	1159	1133	988	10,171
Ontario	2384	827	2758	676	2395	1314	1313	1149	12,816
Manitoba	370	99	363	135	330	161	151	146	1755
Saskatchewan	170	45	160	24	209	107	96	78	889
Alberta	380	79	344	50	246	116	153	140	1508
British Columbia	290	79	276	58	295	143	118	116	1375

* Based on the "first language learned and still understood" variable in the CCHS.

† Based on the combination of language variables from the CCHS (Appendix A).

Table 2. Available Sample Size of Francophone Men and Women Population and Francophones Age 65 Years and Older in the HRs in Ontario Based on the Combined CCHS Cycles (1.1, 1.2, 2.1, 2.2, 3.1, 4.1, 2008 and 2009) for Diabetes, Stroke, and Cancer

		North-West	North-East	East	Centre-East	Centre-West	South-West
Diabetes	65+	17	<10	<10	<10	76	141
	Men	<10	<10	<10	<10	70	106
	Women	16	<10	<10	<10	75	162
Stroke	65+	<10	<10	<10	<10	19	32
	Men	<10	<10	<10	<10	16	21
	Women	<10	<10	<10	<10	18	21
Cancer	65+	<10	<10	<10	<10	33	40
	Men	<10	<10	<10	<10	33	28
	Women	<10	<10	<10	<10	31	47

5 observations for each of the listed variables, even after combining the CCHS cycles.

Beyond the sample size issue, analysis for geographical levels smaller than health regions (i.e., neighbourhoods) using the CCHS is also limited due to the design issue. For regions smaller than HRs, no specific weights in the original data exist. Statistics Canada's bootstrapping methodology is based on resampling with replacement for those sampling units that were used for data collection purposes. Robust variance estimation based on bootstrap weights is therefore challenging for any level of subprovincial geography that differs from those sampling units. Estimates for geographical areas smaller than health regions, or for cross-region boundaries, can sometimes be viewed as domain estimates. In contrast to arbitrary domains, estimates for domains that are also strata or post-strata could generally benefit from having weights that are calibrated to their externally determined population sizes and perhaps from other controls used in the design and sampling. Estimation and variance estimation in this case, although not impossible, will most likely be less precise and less stable.

To analyze small populations, one should also take into account other factors that may reduce the useable sample size and increase sampling error of an estimate. These factors (which will not be discussed in detail here) include but are not limited to non-response, partial non-response (missing values), and estimation method.

Estimation by combining the CCHS datasets

Inadequate sample size in a domain can pose a problem for estimation. Various techniques are available to generate estimates in such cases: these include synthetic estimation, composite estimation and others that usually require correlated information from another source, or the use of good models.³ In this paper, we will focus on the method that can be readily applied with the CCHS and is based

on the combination of different CCHS cycles. Combining the cycles, although aimed to improve precision of estimates, may introduce non-sampling issues that will be considered below.

In order to increase the effective sample size of Francophone minorities and the accuracy of the estimates, we combined five to eight consecutive CCHS cycles at micro-data level to obtain datasets that can be analyzed as coming from a single population (a pooled approach). The resulting sample sizes allow conducting more specific and detailed analysis (Table 1). Combining CCHS cycles, although feasible in most cases, has several methodological limitations and should be applied with care. The type of estimate (simple statistics such as proportions or means, or analytic quantities such as population parameters) and the properties of the estimators should be considered when choosing the method for merging the cycles. There are two methods to be considered when combining the cycles: the pooled approach based on pooled estimates from a combined dataset (mentioned above), and the separate approach based on the average of estimates from the different CCHS cycles.⁴ One of the disadvantages of the separate approach is the assumption of independence between cycles. For example, a fraction of the sample in cycle 2.2 is drawn from the list frame of cycle 2.1 dwellings, and the two samples cannot be considered independent. However, in particular settings when quantities of interest or population characteristics vary substantially from survey to survey, the separate approach may be more appropriate to apply. Still, under a few assumptions the pooled approach has several advantages over the separate approach in which a choice of appropriate composite weighting of each survey might be problematic. Considering the above, we have chosen the pooled method for our analyses. The most important assumption one must make when merging cycles from different time periods is the stationarity of the population of interest.⁵ Often, this assumption cannot be applied

Figure 2. Proportion of Francophone minority population in post-secondary level of education and fifth income quintile based on the CCHS cycles 2001 to 2009

to some of the population's characteristics as it will most likely affect the accuracy of the estimates and bias the results. In variables such as education and income, where the obtained proportions differ substantially between the cycles, the trend over time rather than averaging should be considered when interpreting the results (Figure 2).

For studies on language minorities, the linguistic mobility should also be considered. Since the CCHS is a cross-sectional survey, it is impossible to track down linguistic transitions at the individual level. We could also estimate aggregate trends using overall proportions for each reference year, or in the case of the model-based approach, include a time variable in the model to control for the year the survey was conducted.

Another important issue to be considered when merging the cycles is the homogeneity of definitions and variables used in the merged CCHS cycles. To ensure consistency and comparability across cycles, all the variables must be checked for wording, module structures and response categories. If regions are used, the correspondence between health regions in all the cycles must also be verified. For example, in our study on obesity in Francophone minorities, we had to exclude cycles 1.1 and 1.2 from analysis of the BMI variable due to differences in the target population between the two cycles: in cycles 1.1 and 1.2, BMI was reported for respondents aged 20 to 64 years; in later cycles, it was reported for those aged 18 years and over.⁶ Similarly, to investigate food insecurity using combined CCHS cycles, cycles 2008 and 2009 would have to be excluded due to a change in definition of food insecurity in these years. Sometimes, a combination of cycles is impossible as a specific health topic exists only in one of the surveys. To study nutrition status of Francophone minorities, only cycle 2.2 could be used since questions on nutrition were specific to the CCHS survey in 2004 (focused content).

Often, even after combining the data, the analysis of certain variables is still not possible for smaller geographical regions (e.g., health regions) because of the limited sample size (Table 2).

When combining the CCHS cycles, we also have to remember that the obtained population does not represent a real population, but rather an average population over the combined time periods of the individual cycles. For example, in the Ontario report we combined the CCHS cycles from 2001 to 2009; therefore, the population represented in this study is the average of the Francophone and Anglophone populations over nine years from 2001 to 2009,

calculated simply by adding the population in each cycle and dividing it by the number of cycles used.⁷

In addition, although the CCHS surveys examine the most important aspects of vitality in the Canadian population, a standardized questionnaire cannot claim to cover all the particularities of each province or region. Similarly, the issues specific to a given subpopulation are often not addressed by the survey.

To summarize, combining several cycles in order to analyze small populations adds additional uncertainty to the produced estimates, and several factors must be taken into account when choosing the most appropriate approach.

Non-sampling errors

Under-representation of Small Populations in the Sample

Despite good coverage in the provinces (approximately 98% in all the surveys) and a relatively good definition of variables that identify small populations (for example, linguistic variables to identify Francophone minorities), small populations are often under-represented in the CCHS samples due to a sampling strategy and various geographical or demographic factors specific to a given subpopulation. Based on the Census 2006, Francophone minority populations in Canada comprise approximately 5% of the total population; however, in the CCHS samples, the percentage of the population is usually between 4 and 4.5%. The observed underestimation of Francophones in the CCHS samples might be due to the fact that there is a higher proportion of Francophones in rural areas, and Francophone communities in remote areas of Canada are often difficult to reach and are omitted from sampling. It is also possible that the complicated nature of defining Francophone minority populations in Canada (in addition to the language mobility) may result in misclassification of Francophones and lead to further undercoverage of the population. A survey designed to target these populations specifically, such as the Survey on the Vitality of Official-Language Minorities (SVOLM), which focuses on French-speakers outside Quebec and English-speakers in Quebec, should be developed in order to better portray the situation of small populations in Canada.⁸

Recall Bias and Missing Values

Finally, despite the high quality of CCHS data, all health-related information from the CCHS is based on self-reports. Although much of self-reported information has been shown to be valid and reliable, some measures may be subject to social desirability and/or recall bias. The possible effect of these biases on the quality of estimates should be examined. For example, it is not known whether the potential bias from self-reported variables in our study applies equally to Anglophone and Francophone adults, or whether the magnitude or direction of this bias differs by language group. In addition, some important variables such as income, education, and lifestyle variables (consumption of alcohol, smoking, etc.) have more than 10% missing values and the results should be carefully reported as they might underestimate the true proportions of the listed variables, especially for respondents with lower socio-economic status. We do not expect that there is a differential distribution of these variables across the language groups; however, we did not examine the effect of this potential bias on the obtained estimates. Imputation methods could be considered to resolve the issues with missing values for the variables of interest.

CONCLUSION

National surveys such as the CCHS provide researchers with an opportunity to study specific domains and populations of interest. However, despite the large coverage of the survey, many questions regarding a specific population and its characteristics cannot be answered. Designing a survey targeting a particular population is not always an option due to economic constraints. Relying on secondary analysis of big national surveys becomes the only way to investigate and address emerging issues. For now, methods to improve analysis and estimation for these populations, such as combining the cycles, have to be used. Factors affecting the results (sampling and non-sampling errors) must be considered and cautious interpretation of the results exercised. Extending the CCHS to target specific populations would help to extend knowledge of the minority populations in Canada and allow for planning of adequate resources to address their needs.

Finally, analyzing small populations would become more efficient if the CCHS data could be linked to other administrative and hospitalization databases. However, due to insufficient details within these databases, identification of certain populations and linking are impossible. For example, the Discharge Abstract Database (DAD), National Ambulatory Care Reporting System (NACRS), and Vital Statistics collect only limited demographic information on patients that would enable identification of Francophone minority populations in Canada. Other databases such as the Client Profile Database (CPRO) or Registered Persons Database (RPDB), although they provide language information, do not allow investigation of the health status and access to health care services of French-speaking populations due to inadequacy of these variables. Collecting information that would eventually help researchers to better quantify existing needs and health issues of the population of interest needs to be addressed and improved.

REFERENCES

1. Statistics Canada. Canadian Community Health Survey (CCHS) 2007 Microdata files - User Guide. Ottawa, ON: Statistics Canada, 2008. Available at:

- http://www.statcan.gc.ca/cgi-bin/af-fdr.cgi?l=eng&loc=http://www.statcan.gc.ca/imdb-bmdi/document/3226_D7_T9_V4-eng.pdf&teng=User%20Guide&tfra=Guide%20de%20l'utilisateur (Accessed April 12, 2012).
2. Kish L. Cumulating/combining population surveys. *Survey Methodol* 1999;25(2):129-38.
3. Statistics Canada. Survey methods and practices. Ottawa, ON: Statistics Canada, Catalogue no. 12-587-X, 2003.
4. Roberts G, Binder D. Analysis based on combining similar information from multiple surveys. *JSM Proceedings, Survey Research Methods Section* 2009. Alexandria, VA: American Statistical Association, 2009;2138-47.
5. Thomas S, Wannell B. Combining cycles of the Canadian Community Health Survey. *Health Reports* 2009. Ottawa, ON: Statistics Canada, 2009.
6. Gagnon-Arpin I, Makvandi E, Imbeault P, Batal M, Bouchard L. Le surplus de poids chez les francophones et les anglophones. *Can J Public Health* 2013;104(6)(Suppl. 1):S21-S25.
7. Bouchard L, Batal M, Imbeault P, Gagnon-Arpin I, Makvandi E, Sedigh G. La santé des francophones de l'Ontario : un portrait régional tiré des Enquêtes sur la santé dans les collectivités canadiennes (ESCC). Rapport préparé pour le Bureau des services de santé en français, Ministère de la Santé et des Soins de longue durée, 2012. Available at: http://www.rasfo.ca/images/docs/publications/2012/rapport_escs_ontario-final.pdf (Accessed April 11, 2012).
8. Corbeil JP, Grenier C, Lafreniere S. Minorities Speak Up: Results of the Survey on the Vitality of the Official-Language Minorities. Ottawa: Statistics Canada, 2006, Catalogue no. 91-548-X.

RÉSUMÉ

Les analyses statistiques pour les petites populations ou petits domaines d'intérêt peuvent constituer un défi. Pour obtenir des estimations fiables, seules les très grandes enquêtes telles que l'Enquête sur la santé dans les collectivités canadiennes peuvent être considérées. Toutefois, en dépit de sa couverture géographique et temporelle, l'analyse de petites populations au niveau des régions (régions sociosanitaires par exemple) et en ce qui concerne les questions de santé spécifiques, reste difficile. Nous nous pencherons sur les questions méthodologiques de l'analyse de petites populations en ce qui concerne les erreurs d'enquête (échantillonnage et non dues à l'échantillonnage) qui influent sur la précision et l'exactitude des estimations. La population francophone en situation minoritaire au Canada sera utilisée pour illustrer ces problèmes.

MOTS CLÉS : enquête sur la santé; santé des minorités; langue; taille de l'échantillon; méthodes