

Age-standardizing proportion estimates from combined cycles of the Canadian Community Health Survey

Tristan Watson¹, Kathy Kornas², Laura C. Rosella³

ABSTRACT

Population-based survey data from the Canadian Community Health Survey (CCHS) is commonly used by health system organizations to estimate the prevalence of behavioral, health status, and other risk factors, in order to inform health needs and priorities for people in different populations. However, two issues typically present challenges. First, a single cycle of the CCHS may not have a large enough sample size to reliably compute estimates for a small region or population. Second, proportion estimates between two populations might be distorted by the different underlying age structure of the two groups. Thomas and Wannell (2009) discuss a ‘pooled approach’ method for combining CCHS cycles, in order to increase the power and sample size for analysis. We will further explore the application of this pooled approach and demonstrate how to produce age-standardized proportion estimates and confidence intervals generated with bootstrap weights from pooled CCHS data using procedures in SAS.

KEY WORDS: Complex Survey Data; Canadian Community Health Survey; Age Standardizing; SAS; Survey Methods

RÉSUMÉ

Les données de l'enquête sur la santé dans les collectivités canadiennes (ESCC) sont fréquemment utilisées par les organismes de santé pour estimer la prévalence de facteurs de risque comportemental, de statut de santé et autres et ainsi mieux comprendre les besoins et priorités de santé des membres de différentes populations. Cependant, deux questions se posent souvent. Premièrement, un seul cycle de l'ESCC ne produit pas forcément un échantillon assez grand pour calculer avec fiabilité des estimations pour une petite région ou population donnée. Deuxièmement, les estimations de prévalence brutes pour deux populations peuvent être faussées par la structure d'âge sous-jacente des deux groupes. Thomas et Wannell (2009) ont proposé une « approche de regroupement » pour combiner plusieurs cycles de l'ESCC et ainsi augmenter la puissance d'une analyse et la taille de l'échantillon. Nous illustrons l'application de cette approche de regroupement et montrons comment produire des estimations de prévalence normalisées selon l'âge et des intervalles de confiance générés avec des poids bootstrap à partir des données regroupées de l'ESCC avec des procédures SAS

MOTS CLÉS : Données d'enquête complexe ; Enquête sur les communautés et la santé canadienne ; standardisation de l'âge ; SAS ; Méthodes d'enquête

1. INTRODUCTION

1.1 The Use of Canadian Community Health Survey for Health System Research

The Canadian Community Health Survey (CCHS) is a cross-sectional population-based survey that collects information related to health status, health care utilization and other risk factors in Canada (Statistics Canada, 2018). The CCHS is administered by Statistics Canada and is designed to provide reliable estimates at the health region level every two years. The CCHS is representative of Canadians aged 12 years and older who are living in private dwellings (~98% of the Canadian population). Excluded from the sampling frame are individuals living in long-term care institutions, Indian Reserves and on Crown Lands, full-time members of the Canadian Forces, and some remote areas. The detailed survey methodology of the CCHS is described elsewhere (Béland, 2002). CCHS proportion estimates are often used by health system organizations (e.g., regional health organizations) for the surveillance of health needs and priorities for people in

¹ Tristan Watson, ICES - UofT, University of Toronto, 155 College Street, Toronto, Ontario, Canada M5T 3M7; tris.watson@gmail.com

² Kathy Kornas, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario, Canada M5T 3M7; kathy.kornas@utoronto.ca

³ Laura C. Rosella, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario, Canada M5T 3M7; laura.rosella@utoronto.ca

different populations (Rosella et al., 2014). In order for these proportion estimates to be representative of the population, users must incorporate the CCHS survey weights into their complex survey data analysis (Lewis, 2016).

However, two common issues typically present challenges when using CCHS data for health system research. First, a single cycle of the CCHS may not have a large enough sample size to compute accurate estimates for a small population or geographic unit. Second, proportion estimates between two populations might be distorted by the different underlying age structure of the two groups.

1.2 Organization of the Paper

The plan of the paper is as follows. Section 2 will review the ‘pooled method’ for combining multiple cycles of CCHS, in order to increase sample size. Section 3 will explain the method for direct-standardizing proportion estimates using combined cycles of CCHS, in order to control for the effect of age when making group comparisons. Section 4 will conclude the paper. Appendix 1 and 2 will contain a link to the relevant SAS code examples for implementing these methods explained in Section 2 and 3, respectively.

2. COMBINING CCHS ACROSS SURVEY CYCLES

Thomas and Wannell (2009) discuss a ‘pooled approach’ method for combining CCHS cycles, in order to increase the power and sample size for analysis: “In its most basic form, pooling involves taking the individual data files with the corresponding weights and using a simple merge or set statement in SAS to create one data file”. The authors also explain the various limitations that need to be considered when assessing the appropriateness of combining cycles. This method can be implemented using statistical software. We will further explore how to use the ‘pooled approach’ using SAS software in four steps. Appendix 1 has SAS code examples for applying the pooled approach in SAS.

First, the user must identify relevant and homogenous CCHS variables across cycles. The different CCHS cycles contain hundreds of variables. Relevant CCHS variables are those that are identified for use in the analysis. A CCHS variable is homogenous if the survey question and category responses are the same, despite being labelled differently in different survey cycles. For example, in CCHS cycle 1.1, the variable for whether a person has a chronic condition is labeled CCCAF1, where 1 = “Yes”, 2 = “No”, and 9 = “Not Stated” (Statistics Canada, 2003). In CCHS cycle 2.1, the variable for whether a person has a chronic condition is labeled CCCCCF1, where 1 = “Yes”, 2 = “No”, and 9 = “Not Stated” (Statistics Canada, 2004). It is often useful to create a table similar to Table 1 to compare relevant variables across the different survey cycles. In doing so, researchers should consider that across the different CCHS cycles there may have been changes to the survey question, category responses, and survey design. This could cause variables across cycles to be inconsistent if combined. For instance, one survey cycle could ask a question to respondents older than 17, but in another cycle the same question may be asked to respondents over age 35. In addition, a particular CCHS variable may not be available across all the different cycles. Hence, it is important to read the various user guides and documentation on the summary of changes for each CCHS cycle, which are available through Statistics Canada.

Second, transform and harmonize CCHS variables across the different cycles. In the final dataset, transform the relevant variables across the various CCHS cycles into one common variable name. For example, in Table 1, we could transform CCCAF1 in CCHS cycle 1.1 and CCCCCF1 in CCHS cycle 2.1 into CCHS_chronic. When there is one consistent variable name for a variable across the different cycles, then the variable is considered harmonized.

Table 1 – Example: Identifying CCHS variables across CCHS cycles

CCHS variable concept	CCHS 1.1 Cycle variable	CCHS 2.1 Cycle variable	Combined CCHS Surveys
Has a chronic condition	CCCAF1	CCCCCF1	CCHS_chronic

Third, combine the CCHS cycles including the bootstrap weights. In SAS, this step is done by using a set statement to combine two cycles, which leads to stacking a file for one cycle on top of the file for another cycle. Forth, the CCHS survey and bootstrap weights need to be rescaled, if you are interested in estimating population totals. If survey weights

from three surveys are summed together, this would overestimate the population total by a factor of three. One solution to this problem is to divide the survey and bootstrap weights by the number of cycles combined. This approach would represent “the average population (or period estimate), which covers the combined time periods of the individual cycles” (Thomas & Wannell, 2009).

As with any method, the pooled approach has strengths and limitations. The strength of this method is that it helps to increase the sample size for analysis. In addition, the pooled approach helps to provide a large enough sample size to generate small regional estimates. However, this method can only be used with CCHS cycles that have homogenous variables and similar survey design. Importantly, due to substantial changes to the CCHS survey methodology, caution should be taken when comparing data from previous cycles to data released for the 2015 cycle onwards (Statistics Canada, 2018). The pooled approach can be applied for CCHS surveys cycles from 2001 to 2014, but this method may not be appropriate for CCHS survey cycle 2015+ because of changes to the survey design and weighting. The key limitation with the pooled approach is that it creates an artificial average population that assumes that the population estimates are stable over time (across survey cycles) (Makvandi, Bouchard, Bergeron, & Sedigh, 2013). If the variable of interest has changed considerably over the time period of the combined cycles, the combined estimates could be very misleading. This could mask variability over time. Hence, we recommend that studies should first generate population proportion and counts by CCHS survey cycle, in order to check this stability assumption.

3. DIRECT AGE-STANDARDIZING PROPORTION ESTIMATES

In order to produce direct age-standardized proportions for CCHS variables, it requires the age distribution of the standard population, and the survey-weighted age-specific proportions of the CCHS study population (Namrata Bains, 2009). The age distribution of the standard population is used to create age-specific standard population weights. The standard population weights are age-specific proportions of the standard population calculated from the age-specific stratum count divided by the sum of all age-specific strata. An example of this calculation is shown in Table 2 using the 2015 Canada Population (Statistics Canada, n.d.).

Table 2 – 2015 Canada Population 20+ Age Distribution

Age Group	Standard Population Count	Standard Population Weights
20-34	9,656,910	0.320
35-49	7,166,967	0.237
50-64	7,641,229	0.253
65 and over	5,722,237	0.190
Total	30,187,343	1

For each age group category in the CCHS study population, the age-specific proportions for each level of the CCHS categorical variable needs to be estimated. It is worth noting that almost all CCHS variables are categorical and the method presented here assumes this. For example, if we combined the “Not stated” level with the “No” level, the CCHS variable for chronic condition would have two levels: yes and no. The proportion for each level of the CCHS variable is calculated from the survey-weighted count of individuals in each level of the categorical variable, divided by the survey-weighted count of the total study population. An example of this calculation is shown in Table 3 using contrived numbers. By incorporating the CCHS survey weights in the analysis, the estimated proportions, or survey-weighted crude proportions, are representative “of the covered population, and not just the sample itself” (Statistics Canada, 2018).

Table 3 – Example: Calculating the Proportion for Each Level a Categorical CCHS Variable

Chronic condition	Survey-weighted Counts (represented population)	Survey-weighted Proportions
Yes	67,000	0.508
No	65,000	0.492
Total	132,000	1

In order to calculate age-specific proportions, we can cross tabulate the age-specific groups by the levels of the CCHS variable. For each age-specific group, this will produce the proportions for each level of the CCHS variable. An example of this calculation is shown in Table 4 using the contrived age and chronic condition specific weighted counts and the standard population weights in Table 2. For each age group category, the survey-weighted crude proportions are multiplied by the standard population weights. The direct age-standardized proportion is calculated by summing the product of the survey-weighted proportions and the standard population weights by each level of the CCHS variable in Table 4 to produce column four in Table 5. The age-standardized proportions represent what would have been expected if the study population had the age distribution of the standard population.

Table 4 – Example: Calculating the Product of the Survey-weighted Proportions and Standard Population Weights for each Age Group Category

Age Group	Chronic condition	Survey-weighted Counts	Survey-weighted Proportions (SWP)	Standard Population Weights (PW)	SWP * PW
20-34	Yes	10,000	0.455	0.320	0.145
	No	12,000	0.545	0.320	0.174
	Total	22,000	1		
35-49	Yes	20,000	0.526	0.237	0.125
	No	18,000	0.474	0.237	0.112
	Total	38,000	1		
50-64	Yes	20,000	0.513	0.253	0.130
	No	19,000	0.487	0.253	0.123
	Total	39,000	1		
65 and over	Yes	17,000	0.515	0.190	0.098
	No	16,000	0.485	0.190	0.092
	Total	33,000	1		

Table 5 – Example: Calculating Direct Age-Standardized Proportions

Chronic condition	Survey-weighted Counts	Survey-weighted Proportions	Direct Age-standardized Proportions
Yes	67,000	0.508	0.498
No	65,000	0.492	0.502
Total	132,000	1	1

In summary, there are five main steps needed to produce direct age-standardized proportions using CCHS data:

1. Select a standard reference population and age group categories.
2. Calculate the standard population weights.
3. For each age group category, calculate the survey-weighted proportions for each level of the CCHS variable.
4. For each age group category, the survey-weighted crude proportions are multiplied by the standard population weights.
5. In order to get direct-standardized proportions, sum product of the survey-weighted proportions and the standard population weights by each level of the CCHS variable.

The link to the SAS macro program example with corresponding steps to produce direct age-standardized proportions is available in Appendix 2. The SAS program also contains additional steps to include a method to produce logit confidence intervals rather than Wald confidence intervals. The Wald confidence interval typically needs very large domain sample sizes, in order to be appropriate for proportions. The logit confidence interval is one alternative to Wald confidence interval with better coverage for proportions (Dunnigan, 2008; SAS, 2010b). Our approach uses the variance of the survey-weighted proportion as an estimate of the variance of the age-standardized proportion; this approach has an unknown effect on the quality of the confidence interval estimate. Another option for the variance calculation is to directly use the bootstrap replicate estimates of the age-standardized proportion. A review of how SAS computes the bootstrap variance estimates using balance repeated replication method has been described elsewhere (SAS, 2010a). A more

generalized review of direct age-standardization, age group category considerations, and the underlying math is described elsewhere (Namrata Bains, 2009).

4. FINAL COMMENTS

This paper has explained a method for combining CCHS cycles and direct-age standardizing proportions, and illustrated the application of these methods using SAS. Users could consider applying this approach to increase the CCHS sample size for analysis and to make group proportion comparisons adjusting for the effect of age structure. By combining CCHS cycles and direct-age standardizing the proportions, the user moves further away from real and observed data, to imaginary and adjusted data (Kaufman, 2017). Imaginary and adjusted data can be useful, when properly used and interpreted in relation to observed data. Combined CCHS cycles represent an artificial average population across the time period cycles of combined individual cycles (Thomas & Wannell, 2009). This artificial average population is not observed in the real world. The usefulness of the artificial average population will depend on the validity of the underlying assumptions. By combining CCHS cycles, assumptions are made about the homogeneity of the variables and the stationarity of the variable of interest across the different CCHS cycles (Makvandi et al., 2013). These assumptions can be assessed by carefully considering the CCHS variables and assessing population characteristics in each single CCHS cycle. By direct-age standardizing a combined CCHS dataset, these proportions estimates for CCHS variables are re-weighted to the age-group specific proportions of the standard population. It is often more informative to make comparison between groups while adjusting for the effect of age. Unadjusted or crude proportions should be presented with age-adjusted proportions, since age-adjusted proportions can be affected by the choice of age group categories and the standard population. When feasible, unadjusted age-specific proportions should be calculated across the different CCHS cycles, in order to see if age is acting as an effect modifier in the variable of interest over time (Namrata Bains, 2009). Although we focused on describing the pooled approach and calculating direct age standardized proportions specifically for the CCHS, the application of these methods is relevant to other population-based cross sectional surveys administered around the world, such as the National Health Interview Survey and the Health Survey for England.

ACKNOWLEDGMENTS

The authors would like to thank Lenka Mach from Statistics Canada for her review and helpful feedback on earlier drafts of this manuscript.

APPENDIX 1

The link contains a SAS code example for pooling two cycles of CCHS:

https://github.com/watsonr/cchs_age_standard/blob/master/pooling_cchs

APPENDIX 2

The link contains a SAS code example for direct age-standardizing proportion estimates using CCHS data:

https://github.com/watsonr/cchs_age_standard/blob/master/cchs_age_standard_code

REFERENCES

- Béland, Y. (2002). Canadian Health Survey - Methodology Overview. *Health Reports*, 13(3), 9–14.
- Dunnigan, K. (2008). Confidence Interval Calculation for Binomial Proportions. *Midwest SAS Users Group*, 12. Retrieved from <http://www.mwsug.org/proceedings/2008/pharma/MWSUG-2008-P08.pdf>
- Kaufman, J. S. (2017). Statistics, adjusted statistics, and maladjusted statistics. *American Journal of Law and Medicine*, 43(2–3), 193–208. <https://doi.org/10.1177/0098858817723659>
- Lewis, T. H. (2016). *Complex Survey Data Analysis with SAS*. <https://doi.org/10.1201/9781315366906>
- Makvandi, E., Bouchard, L., Bergeron, P. J., & Sedigh, G. (2013). Methodological issues in analyzing small populations using CCHS cycles based on the official language minority studies. *Canadian Journal of Public Health*.
- Namrata Bains. (2009). *Standardization of Rates*. Retrieved from <https://www.apheo.ca/membership/documents/loadDocument?id=1077&download=1#upload/membership/document/standardization-report-nambains-finalmarch16.pdf>
- Rosella, L. C., Fitzpatrick, T., Wodchis, W. P., Calzavara, A., Manson, H., & Goel, V. (2014). High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. *BMC Health Services Research*, 14(1), 532. <https://doi.org/10.1186/s12913-014-0532-2>
- SAS. (2010a). Balanced Repeated Replication (BRR). Retrieved October 22, 2019, from https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveyfreq_a0000000212.htm
- SAS. (2010b). Confidence Limits for Proportions. Retrieved October 22, 2019, from https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveyfreq_a0000000221.htm
- Statistics Canada. (n.d.). Table 17-10-0005-01 Population estimates on July 1st, by age and sex. <https://doi.org/doi.org/10.25318/1710000501-eng>
- Statistics Canada. (2003). *CCHS Cycle 1.1: Data Dictionary*. Retrieved from http://www.statcan.gc.ca/eng/statistical-programs/document/3226_D3_T9_V1-eng.pdf
- Statistics Canada. (2004). *CCHS Cycle 2.1: Data Dictionary*. Retrieved from http://www.statcan.gc.ca/eng/statistical-programs/document/3226_D3_T9_V2-eng.pdf
- Statistics Canada. (2018). Canadian Community Health Survey - Annual Component (CCHS). Retrieved October 22, 2019, from <http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=795204>
- Thomas, S., & Wannell, B. (2009). Combining cycles of the Canadian Community Health Survey. *Health Reports / Statistics Canada, Canadian Centre for Health Information = Rapports Sur La Santé / Statistique Canada, Centre Canadien d'information Sur La Santé*, 20(1), 53–58.