cchsflow: An open science approach to transform & combine population health surveys into one dataset

## Introduction

You are a public health epidemiologist who would like to report the change in body mass index (BMI) in your health unit over the past 15 years. You review the codebook for the Canadian Community Health Survey (CCHS) and note that BMI is collected. BMI *seems* like a straightforward measure that is routinely collected worldwide.(Statistics Canada, 2001) Indeed, BMI is included in all CCHS cycles. You examine the documentation and find the variable HWTAGBMI in the CCHS 2001 corresponds to body mass index, but that in other cycles, the variable name changes to HWTCGBMI, HWTDGBMI, HWTEGBMI, etc. On reading the documentation, you notice that some cycles round the value to one decimal, whereas other cycles round to two digits. Furthermore, some cycles don’t calculate BMI for respondents under the age of 20 or over the age of 64 years. Also, some cycles calculate BMI only if height and weight are within specific ranges. After spending hours on the task, you talk with a colleague in a neighbouring health unit. They did the same task a few years ago. You share your Stata code by email and compare notes, only to realize that you both had different approaches, each with errors.

We created a process called *cchsflow* to minimize the amount of time public health epidemiologists and others spend cleaning and transforming CCHS variables across multiple survey cycles. We sought to use an open science approach for the development of *cchsflow*. Open science is the movement to improve research reproducibility, accessibility, and collaboration.(Ross & Krumholz, 2013) Public health practice strives for these qualities and, therefore, our field can potentially benefit from the same tools that are used to support open science. An example of the open science toolkit is versioning software and cloud-based repositories such as GitHub and GitLab that allow people to colloborate and share programming code.

Currently, *cchsflow* harmonizes 160 variables for 1,092,951 survey respondents of the CCHS Public Use Microdata File (PUMF) from 2001 to 2014. *cchsflow* uses open science tools to allow users the ability to contribute to the package, including making suggestions and requests, and identify errors. People can also “fork” the package meaning they can use the *cchsflow* approach to harmonize other databases. *cchsflow* uses R language package since R is the most commonly used open statistical programming language. The core of *cchsflow*, however, are references files that could be used in other programming languages.

Even this paper was created using the open science principles. This paper was written using R Markdown - a notebook that allows R code to be executed within a document. Both the *cchsflow* R package and this paper’s notebook are available on [GitHub](https://github.com/Big-Life-Lab/cchsflow/blob/paper-writeups/papers/CJPH/cjph-paper.Rmd) which allows readers to make comments, suggestions or note errors. Readers can execute or modify all examples in this paper in R.

## Background

### Cleaning and transforming CCHS datasets

Data cleaning, including transforming variables into harmonized or common variables, is typically the most time-consuming part of data analyses. According to Dasu & Johnson, 80% of data analysis is spent on data cleaning.(Dasu & Johnson, 2003) With the CCHS, data cleaning and harmonization issues arise when combining CCHS surveys for longitudinal analysis. Currently, there is no standardized method or tool used to combine CCHS survey cycles. Health units across Canada that use the CCHS do their own data cleaning and preparation, taking time away from other data analysis.

### Open science and its benefits to public health practice

Open science is defined as “transparent and accessible knowledge that is shared and developed through collaborative networks”.(Vicente-Saez & Martinez-Fuentes, 2018) Included in open science is: open data, data that is publicly accessible such as the CCHS with Statistics Canada’s new Open License (Statistics Canada, 2020); open source, the use of open access programs such as data science languages including R, Python and Julia; and open methodology, program code that is publicly accessible and shared through on-line repositories.(McKiernan et al., 2016; Stodden, Guo, & Ma, 2013) In public health, there has been a marked trend toward open data and sharing code - notably during the COVID-19 pandemic.(Moorthy, Restrepo, Preziosi, & Swaminathan, 2020)

Adopting open science practices comes with well-described benefits.(Donoho, 2017; Hicks & Irizarry, 2018) McKiernan et al. found that open science is associated with increased research exposure in both media and in citations; and an increase in collaboration, funding, and job opportunities.(McKiernan et al., 2016) For public health professionals, an open science approach and toolkit facilitates collaboration as it allows for data & coding methods to be shared between different health units. Additional benefits include improved transparency, accessibility, efficiency, reduced coding errors, and faster analyses. As in other sectors, public health practitioners can use open science tools to potentially improve and compress many time consuming, repetitive, and inconsistent analyses tasks.

## Methods

*cchsflow* follows the approach of the Open Source Initiative and open software for research.(Journal of Open Source Software, 2018; Open Source Initiative, 2020) *cchsflow* was created in R with provisions to support other program languages such as Stata or SAS.(Manuel, Yusuf, Vyuha, & Bennett, 2020) The package currently supports the first 10 cycles of the CCHS PUMF surveys from 2001 to 2014, in which the variables of each were harmonized and transformed to use the same set of variables. In *cchsflow*, variables were renamed to the variable names used in CCHS cycles from 2007 to 2014 as they have shown to be the most consistent.

### Selection of variables

Variables included in *cchsflow* fall into three categories: health behaviours, sociodemographic information, and health status. There are provisions and instructions on how users can contribute or request the addition of new variables.

In Gochman’s handbook on health behaviour research, he defines health behaviours as “overt behavioral patterns, actions and habits that relate to health, to health restoration and to health improvement”.(Gochman, 1997) This definition encompasses a wide variety of behaviours such as, smoking, alcohol, diet, and physical activity.(Conner & Norman, 2017) For the development of *cchsflow*, the aforementioned behaviours were examined using existing CCHS variables, as well as derived variables. With respect to sociodemographic information, variables such as age, sex, immigration status, country of birth, time spent in Canada, ethnicity, education (individual and highest family), income (adjusted for province and inflation), home ownership, and marital status were collected and added to the *cchsflow* library. According to Shah, health status can be defined as “[t]he degree to which a person is able to function physically, emotionally, socially, with or without aid from the healthcare system.”.(Shah, 2003) Within this definition of health status lies health conditions, human function, wellbeing, and deaths. Variables relating to health conditions, human function, and wellbeing were collected. Deaths were not collected as CCHS data included only living respondents.

### Variable mapping

CCHS variables were transformed across 10 survey cycles. For many variables, the only difference between cycles were their variable name. As such, only a name change was required to standardize a variable across the 10 cycles.

Changes in the number and type of categories was also common. For example, in the 2001 and 2003 CCHS survey cycles, there were 15 age categories; while in CCHS survey cycles from 2005 to 2014, there were 16 age categories. There were two options for such variable category changes. The first option was to create a harmonized variable by collapsing categories into common forms. The second option was to maintain separate variables. For age we also included a third option to maximize age information by deriving a new continuous age variable was derived that takes the midpoint of each age category for all cycles.

There were also changes to question wording, missing categories, and inclusion and exclusion criteria. Variables were not included in all cycles or all health regions. Harmonized variables were included when there was a consensus amongst developers that the differences across cycles where small. *Notes* are included when any difference was identified, with a default to print all notes during transformations.

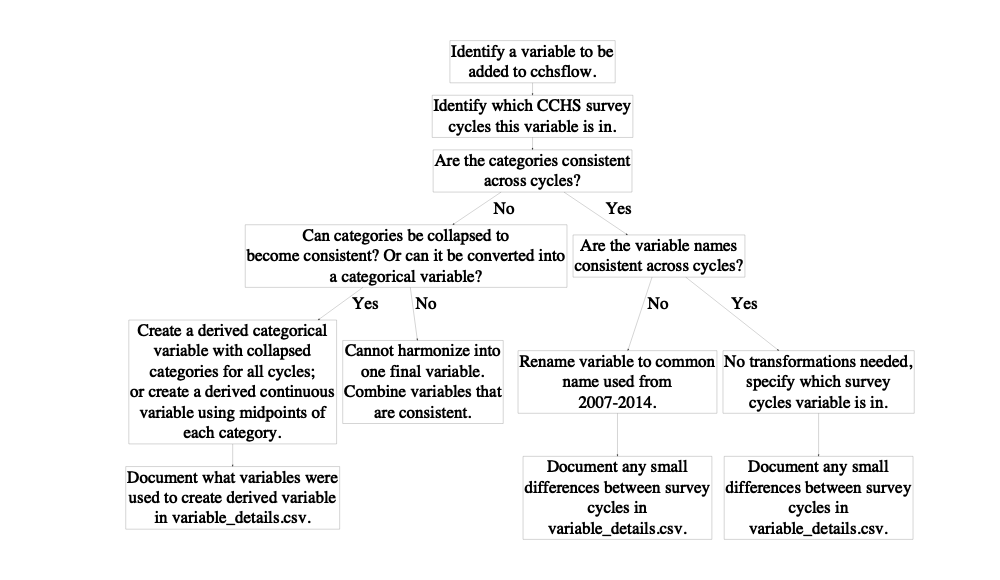
### Transformation of variables through specification worksheets

Two worksheets are included in the *cchsflow* packages that contain variable information and metadata: *variables.csv* specifies all the variables in the package. *variable\_details.csv* specifies CCHS data that contain the variables, the variable type, and the category structure.

*rec\_with\_table()*, short for “recode with table”, is the key function to transform variables. *rec\_with\_table()* uses the two worksheets to create a transformed data from a CCHS cycle. Once all CCHS survey cycles have been transformed, they can be combined to create one large transformed data set that spans across the 10 CCHS survey cycles. The two CSV worksheets also has variable labels and other metadata that can be added to the data using the *rec\_with\_table()* function.

### Derived variables

CCHS includes derived variables that were created using multiple responses and variables. BMI is an example of an original CCHS derived variable that was calculated using self-reported height and weight. Several new derived variables were included, and there are provisions and instructions for adding additional variables. These variables were based on derived variables used in previous studies and include smoking pack-years, binge drinking, and diet pattern.(Manuel et al., 2016)



**Figure 1**: Flowchart of how CCHS variables were added to *cchsflow*. Users can add variables using the same approach.

### Documentation

Open source, web-based documentation is available at <https://big-life-lab.github.io/cchsflow/>, and includes a searchable reference of all transformations, vignettes with examples of how to perform transformations, collaboration principles, and a development roadmap.



**Figure 2:** The homepage for the *cchsflow* website.

## Results

The *cchsflow* package is available on the Comprehensive R Archive Network (CRAN), a network of servers that contain documentation for R packages.(Manuel et al., 2020; *The Comprehensive R Network*, 2020) The package contains the following items: the *variables.csv* worksheet, the *variable\_details.csv* worksheet, the various functions, and subsets of 200 respondents for each CCHS cycle.

**Figure 3a** illustrates the command line to install the CRAN version of *cchsflow*, while **Figure 3b** illustrates the command to install the development version of *cchsflow*, which is a more up to date version of the package.

install.packages("cchsflow")

**Figure 3a:** The command line to install the *cchsflow* package that is currently saved on CRAN.

devtools::install\_github("Big-Life-Lab/cchsflow")

**Figure 3b:** The command line to install the development version of *cchsflow* from Github.

### Recode with table

The *rec\_with\_table()* function is used to recode or transform variables based on the information from the two specification worksheets. The function has the ability to transform an entire dataset, or a subset of variables. **Figure 4a** illustrates how to load the *cchsflow* package, the 2001 CCHS data, and then transform all variables in *cchsflow* to their harmonized version. The *cchsflow* package comes with a subsample of CCHS data for 2001 to 2013 versions, made possible with Statistics Canada new Open License (Statistics Canada, 2020). **Figure 4b** illustrates how to transform a subset of variables from the 2001 survey cycle.

library(cchsflow)  
cchs2001 <- read.csv("~/data/cchs2001.csv")  
transformed\_cchs <- rec\_with\_table(cchs2001)

**Figure 4a:** The command lines to load the *cchsflow* package, load the 2001 CCHS PUMF data and then transform the all variables in the worksheets using the *rec\_with\_table()* function.

library(cchsflow)  
cchs2001 <- read.csv("~/data/cchs2001.csv")  
transformed\_cchs2001 <- rec\_with\_table(cchs2001, c("DHH\_SEX", "DHHGAGE\_cont"))

**Figure 4b:** The command lines to transform the sex & age variables using the *rec\_with\_table()* function.

## Discussion

*cchsflow* R package harmonizes and transforms CCHS data from 2001 to 2014.(Manuel et al., 2020) *cchsflow* provides public health epidemiologists and others the ability to more robustly analyze over 1 million respondents across a 13-year period to examine trends in health indicators. The use of an open science approach improves collaboration, transparency, and efficiency when transforming variables. The package allows public health professionals that use CCHS to spend less time on data cleaning and spend more time on analysis such as surveillance and health status reporting.

### Comparison to other projects

*cchsflow* facilitates the use of CCHS metadata for public health. Metadata is increasingly recognized as helpful data infrastructure to support open science and data harmonization. Metadata is “data about data” and includes information about variable and category labels, variable types, and provenance (how the data was collected and transformed).(Duval, Hodgins, Sutton, & Weibel, 2002; McGilvray, 2008) The CCHS comes with Data Documentation Initiative (DDI) metadata.(Vardigan, Heus, & Thomas, 2008) DDI metadata is used worldwide for over 10 thousand different surveys and research projects.(Data Documentation Initiative (DDI), 2020) There are also initiatives such as Maelstrom that is used by other Canadian health surveys to improve the use of metadata.(Bergeron, Doiron, Marcon, Ferretti, & Fortier, 2018)

A consistent approach to calculate health indicators is a long-standing public health goal. *cchsflow* uses an open science approach to build from and support several health-related indicator and harmonization projects that use CCHS data including the Canadian Institute for Health Information indicator library, the Public Health Agency of Canada health inequality reports, and Ontario’s Public Health Indicator Working Group. (Association of Public Health Epidemiologists in Ontario, 2018; Canadian Institute for Health Information, 2020; Pan-Canadian Public Health Network, 2018) These initiatives typically include the definition of indicators, but it is uncommon to publish how to calculate indicators using CCHS data, especially across CCHS cycles. *cchsflow* an provides open, reusable, accessible, and collaborative approach to calculate indicators using the CCHS.

To our knowledge it is uncommon for public health practitioners to use metadata beyond variable labels in their analyses workflow.(Manuel & Fisher, 2019) Barriers to using metadata in public health include the lack of well-organized metadata in public health data and the lack of metadata analysis tools such as *cchsflow*. It is commendable that DDI documents are included with CCHS, but not all metadata is included or consistent. Support for variable transformation is robustly supported in newer versions of DDI that are not yet available for the CCHS. *cchsflow* uses DDI documents to create the worksheets with the added benefit of harmonizing and transforming metadata across CCHS cycles. *cchsflow* also supports the use of Predictive Modeling Markup Language (PMML).(Grossman et al., 1999) Our team uses *cchsflow*’s PMML metadata in public health planning tools that calculate life expectancy and chronic disease risk for public health regions.(Project Big Life, 2020)

### Limitations and challenges

While the CCHS has many consistent variables across survey cycles, there are differences between cycles that can sometimes be irreconcilable or difficult to harmonize. Within *cchsflow*, variables with irreconcilable differences were either transformed into a new derived variable or kept as separate variables that can be only be used in select cycles. Along with variables with irreconcilable differences, there are variables in *cchsflow* that were not asked in all CCHS cycles. This means for some variables, data does not span across the length of the CCHS cycles available in *cchsflow*. A possible solution is to impute missing variables, where missing data is replaced with values on based on other respondents and responses to other variables.

*cchsflow* has been used in a range of projects, but care must be taken to understand how specific variable transformation and harmonization with *cchsflow* affects each use of CCHS data. Across survey cycles, almost all CCHS variables have had at least some change in wording and category responses. Furthermore, there have been changes in survey sampling, response rates, weighting methods and other survey design changes that affect responses. Combining CCHS datasets across survey cycles will result in misclassification error and other forms of bias that affects studies in different ways.

### Collaboration with other users

Collaboration is facilitated using GitHub, the most popular online code repository with over 45 million users. Github is based on the Git version-control system which, in turn, is a cornerstone of open software development.(Dabbish, Stuart, Tsay, & Herbsleb, 2012)

The open-access nature of the *cchsflow* package also allows users to add other CCHS variables that might benefit others. There is complete transparency on how the package was developed as the entire source code for the package is publicly accessible on the repository. Along with being transparent, putting the *cchsflow* package on GitHub offers users of the package the opportunity to provide feedback on how to further improve the package. In the issues section of the GitHub repository, users can submit bug reports where they can identify issues they are encountering while using the package. Users can also request variables to be added in this section. Users can also add new variable transformations using the worksheets and then create a “pull request” to add the transformation to the main *cchsflow* package. All *cchsflow* documentation (including this paper write up) are also open access and available on the GitHub repository.

GitHub provides benefit to users in that it provides them an opportunity to implement better practices in their own code (Dabbish et al., 2012). The implementation of GitHub in the development of *cchsflow* allows public health professionals across Canada to collaborate and share potential variables that can be useful for health surveillance and health status reporting.

### Roadmap

A roadmap, also known as next steps or future plans, are recommended for open software projects. *cchsflow* includes a roadmap and milestones on the [project website](https://github.com/Big-Life-Lab/cchsflow/projects). At the time of writing, the roadmap includes the adding the “share” version of CCHS that is used in Statistics Canada Regional Data Centres and other settings, the ability to compare variable frequency across survey cycles, and improved metadata support. *cchsflow* has been forked by related projects to support other datasets. The expanded use of *cchsflow* for related projects is a hallmark of open science and demonstration of how open science leads to expanded science and public health resources.

### Conclusion

Open science and the open science toolkit are transforming science and have been invaluable in public health emergencies such as COVID-19. *cchsflow*’s open science also allows public health professionals to collaborate and share their work with other colleagues, saving time recoding and cleaning health datasets. By implementing open science practices, *cchsflow* aims to minimize the amount of time needed to clean and prepare CCHS data for the many CCHS users in health units across Canada.

## References

Association of Public Health Epidemiologists in Ontario. (2018). *Core Indicators Work Group*. Retrieved from <https://www.apheo.ca/core-indicators-work-group>

Bergeron, J., Doiron, D., Marcon, Y., Ferretti, V., & Fortier, I. (2018). Fostering population-based cohort data discovery: The maelstrom research cataloguing toolkit. *PloS One*, *13*(7), e0200926. <https://doi.org/https://doi.org/10.1371/journal.pone.0200926>

Canadian Institute for Health Information. (2020). *Indicator Library*. <https://indicatorlibrary.cihi.ca/display/HSPIL/Indicator+Library?desktop=true&_ga=2.214437141.328439497.1597831848-1908179218.1597831848>.

Conner, M., & Norman, P. (2017). *Health behaviour: Current issues and challenges*. <https://doi.org/10.1080/08870446.2017.1336240>

Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012). Social coding in GitHub: Transparency and collaboration in an open software repository. *Proceedings of the acm conference on computer supported cooperative work, cscw*. <https://doi.org/10.1145/2145204.2145396>

Dasu, T., & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. <https://doi.org/10.1002/0471448354>

Data Documentation Initiative (DDI). (2020). *Document, Discover and Interoperate*. Retrieved from <https://ddialliance.org/>

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-Lib Magazine*, *8*(4), 1082–9873.

Gochman, D. S. (1997). *Handbook of Health Behavior Research I* (p. 506). Springer US.

Grossman, R., Bailey, S., Ramu, A., Malhi, B., Hallstrom, P., Pulleyn, I., & Qin, X. (1999). Management and mining of multiple predictive models using the predictive modeling markup language. *Information and Software Technology*. <https://doi.org/10.1016/S0950-5849(99)00022-1>

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *The American Statistician*, *72*(4), 382–391.

Journal of Open Source Software. (2018). *Journal of Open Source Software*. Retrieved from <https://joss.readthedocs.io/en/latest/index.html>

Manuel, D., & Fisher, S. (2019). *A toolkit has emerged to support open science for health service and policy research*. Presented at the Convention for Canadian Society for Epidemiology and Biostatistics, Ottawa, Canada.

Manuel, D. G., Perez, R., Sanmartin, C., Taljaard, M., Hennessy, D., Wilson, K., … Rosella, L. C. (2016). Measuring Burden of Unhealthy Behaviours Using a Multivariable Predictive Approach: Life Expectancy Lost in Canada Attributable to Smoking, Alcohol, Physical Inactivity, and Diet. *PLoS Medicine*. <https://doi.org/10.1371/journal.pmed.1002082>

Manuel, D., Yusuf, W., Vyuha, R., & Bennett, C. (2020). *Cchsflow: Transforming and harmonizing cchs variables*. Retrieved from <https://github.com/Big-Life-Lab/cchsflow>

McGilvray, D. (2008). *Executing data quality projects: Ten steps to quality data and trusted information (tm)*. Elsevier.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., … Yarkoni, T. (2016). *How open science helps researchers succeed*. <https://doi.org/10.7554/eLife.16800>

Moorthy, V., Restrepo, A. M. H., Preziosi, M.-P., & Swaminathan, S. (2020). Data sharing for novel coronavirus (covid-19). *Bulletin of the World Health Organization*, *98*(3), 150.

Open Source Initiative. (2020). *Open Source Initiative*. Retrieved from <https://opensource.org/>

Pan-Canadian Public Health Network. (2018). *Key Health Inequalities in Canada: A National Portrait*.

Project Big Life. (2020). *Project Biglife Planning Tool Guide*. Retrieved from <https://big-life-lab.github.io/pbl-planning-tool-guide/>

Ross, J. S., & Krumholz, H. M. (2013). *Ushering in a new era of open science through data sharing: The wall must come down*. <https://doi.org/10.1001/jama.2013.1299>

Shah, C. P. (2003). *Public Health and Preventive Medicine in Canada* (5th ed., p. 595). Elsevier.

Statistics Canada. (2001). *CCHS Cycle 1.1 (2000-2001), Public Use Microdata File Documentation* (p. 77).

Statistics Canada. (2020). *Statistics Canada Open Licence*. Retrieved from <https://www.statcan.gc.ca/eng/reference/licence>

Stodden, V., Guo, P., & Ma, Z. (2013). Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. *PLoS ONE*, *8*(6), 2–9. <https://doi.org/10.1371/journal.pone.0067111>

*The Comprehensive R Network*. (2020). Retrieved from <https://cran.r-project.org/>

Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, *3*(1).

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*. <https://doi.org/10.1016/j.jbusres.2017.12.043>