



# *cchsf*low: an open science approach to transform and combine population health surveys

Warsame Yusuf<sup>1</sup> · Rostyslav Vyuha<sup>1</sup> · Carol Bennett<sup>1</sup> · Yulric Sequeira<sup>1</sup> · Courtney Maskerine<sup>1,2</sup> · Douglas G. Manuel<sup>1,2,3,4,5</sup>

Received: 7 July 2020 / Accepted: 22 December 2020 / Published online: 24 March 2021  
© The Author(s) 2021

## Abstract

**Setting** The Canadian Community Health Survey (CCHS) is one of the world’s largest ongoing cross-sectional population health surveys, with over 130,000 respondents every two years or over 1.1 million respondents since its inception in 2001. While the survey remains relatively consistent over the years, there are differences between cycles that pose a challenge to analyze the survey over time.

**Intervention** A program package called *cchsf*low was developed to transform and harmonize CCHS variables to consistent formats across multiple survey cycles. An open science approach was used to maintain transparency, reproducibility and collaboration.

**Outcomes** The *cchsf*low R package uses CCHS survey data between 2001 and 2014. Worksheets were created that identify variables, their names in previous cycles, their category structure, and their final variable names. These worksheets were then used to recode variables in each CCHS cycle into consistently named and labelled variables. Following, survey cycles can be combined. The package was then added as a GitHub repository to encourage collaboration with other researchers.

**Implication** The *cchsf*low package has been added to the Comprehensive R Archive Network (CRAN) and contains support for over 160 CCHS variables, generating a combined data set of over 1 million respondents. By implementing open science practices, *cchsf*low aims to minimize the amount of time needed to clean and prepare data for the many CCHS users across Canada.

## Résumé

**Contexte** L’Enquête sur la santé dans les collectivités canadiennes (ESCC) est l’une des plus grandes enquêtes transversales sur la santé de la population, avec plus de 130 000 sondés tous les deux ans et plus de 1,1 million de sondés depuis son début en 2001. Tant que l’enquête reste relativement cohérent, il y a des différences entre des cycles qui posent une challenge majeure pour analyser l’enquête au fil du temps.

**Intervention** Un paquet de programme appelé *cchsf*low a été développé pour transformer et harmoniser les variables CCHS aux formats cohérents à travers plusieurs cycles de sondage. Une approche de science ouverte était utilisée pour maintenir la transparence, la reproductibilité et la collaboration.

**Résultats** Le paquet *cchsf*low R développé utilisait les données d’enquête de l’ESCC entre 2001 et 2014. Les feuilles de calcul ont été créées pour identifier des variables, leurs noms dans des cycles précédents, leurs structures de catégories et leurs noms de variables finales. Ces feuilles de calcul ont ensuite été utilisées pour recoder les variables dans chaque cycle de l’ESCC pour générer les ensembles de données harmonisés qui peuvent être combinés dans un ensemble de données constamment étiqueté

✉ Warsame Yusuf  
waryusuf@ohri.ca

<sup>1</sup> Ottawa Hospital Research Institute, Civic Campus, ASB 2-012, 1053 Carling Avenue, Ottawa, ON K1Y 4E9, Canada

<sup>2</sup> Department of Family Medicine, University of Ottawa, Ottawa, Ontario, Canada

<sup>3</sup> ICES, Ottawa and Toronto, Ottawa, Ontario, Canada

<sup>4</sup> Statistics Canada, Ottawa, Ontario, Canada

<sup>5</sup> School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada

pour l'analyse. Le paquet a ensuite été ajouté comme un entrepôt de GitHub pour encourager la collaboration avec les autres chercheurs.

**Implication** Le paquet *cchsflow* a été ajouté au Comprehensive R Archive Network (CRAN) et contient un appui pour plus de 160 variables de l'ESCC, générant un ensemble de données de plus d'un million de sondés. En exécutant les pratiques de sciences ouvertes, *cchsflow* vise à minimiser le temps requis pour nettoyer et préparer les données pour les plusieurs utilisateurs du CCHS à travers le Canada.

**Keywords** Health surveys · Data analysis · Data science · Population health

**Mots-clés** Enquêtes de santé · analyse des données · science des données · santé de la population

## 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.

A process called *cchsflow* was created to minimize the amount of time public health epidemiologists and others spend cleaning and transforming CCHS variables across multiple survey cycles. An open science approach was sought for the development of *cchsflow*. Open science is the movement to improve research reproducibility, accessibility, and collaboration (Ross and Krumholz 2013). Public health practice strives for these qualities and, therefore, the 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 collaborate 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 to 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 reference 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](#) which allows readers to make comments and suggestions or note errors. Readers can execute or modify all examples in this paper in R.

## Background

### Cleaning and transforming CCHS data

Data cleaning, including transforming variables into harmonized or common variables, is typically the most time-consuming part of data analyses. According to Dasu and Johnson, 80% of data analysis is spent on data cleaning (Dasu and Johnson 2003). With the CCHS, data cleaning and harmonization issues arise when combining CCHS surveys. 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 and Martinez-Fuentes 2018). Included in open science is: open data, data that are publicly accessible such as the CCHS with Statistics Canada's new

Open Licence (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 online repositories (McKiernan et al. 2016; Stodden et al. 2013). In public health, there has been a marked trend toward open data and sharing code—notably during the COVID-19 pandemic (Moorthy et al. 2020).

Adopting open science practices comes with well-described benefits (Donoho 2017; Hicks and 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 and coding methods to be shared between different health units. Additional benefits include improved transparency, accessibility, and 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 analysis tasks. In light of the COVID-19 pandemic, open science allows public health researchers to quickly aggregate and analyze data across many health units to guide policy-makers in making informed public health decisions.

## Methods

*cchsflow* follows the approach of the Open Source Initiative and open software for research. The developers of *cchsflow* are public health researchers who collaborate with federal, provincial and local public health units. *cchsflow* was developed following publication of several peer-reviewed reports created with Public Health Ontario, ICES and the Ontario Public Health Association (Journal of Open Source Software 2018; Manuel et al. 2012; Open Source Initiative 2020). One of the developers of *cchsflow* (DGM) is a part-time employee at Statistics Canada. The *cchsflow* package is not a Statistics Canada product, nor is the package supported by Statistics Canada. However, analysts at Statistics Canada use *cchsflow* and have contributed to variable transformations.

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.

Many variables in *cchsflow* are used in peer-reviewed studies of our development team and other researchers (Manuel et al., 2012, 2016, 2018, 2020a). Occupation variables, for example, were incorporated from peer-reviewed occupation

studies (Nowrouzi-Kia et al. 2019). Depression variables are an example of variables for which there was not consistent use in peer-reviewed literature, but which were added in consultation with mental health researchers. Open discussion with the mental health researchers is included in the package development (<https://github.com/Big-Life-Lab/cchsflow/pull/64>). Anyone can participate in the discussions when new variables are added. *cchsflow* was created in R with provisions to support other program languages such as Stata or SAS.

## Selection of variables

Variables included in *cchsflow* fall into three categories: health behaviours, socio-demographic information, and health status. At the time of writing, there are 160 variables, 30 subjects and 6 sections. There are provisions and instructions on how users can contribute or request the addition of new variables.

Health behaviours variables include smoking, alcohol, diet, and physical activity (Conner and Norman 2017). There are derived variables such as smoking pack-years (*pack\_years\_der*) that are not available in the original CCHS data files.

Socio-demographic variables include 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. Harmonized occupation variables were created (*LBFA\_31A*, *LBFA\_31A\_A* and *LBFA\_31A\_B*) by reviewing studies that used the CCHS to study occupation (Nowrouzi-Kia et al. 2019). References to these papers are included in the notes section of the variable transformation.

Health status variables include chronic disease, the Health Utility Index, need for help for activities of daily living (ADL), mental health, and other measures. There is a new derived variable for the number of ADL requiring assistance (*ADL\_score\_5*) that is not available in the original CCHS data.

## Variable mapping

CCHS variables were transformed across 10 survey cycles. For many variables, the only difference between cycles was 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 were 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, a third option was also added

to maximize age information by deriving a new continuous age variable, one 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 among developers that the differences across cycles were small. *Notes* were 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 and *variable\_details.csv* specifies CCHS data that contain the variables, the variable type, and the category structure.

*cchsflow* was created using the *recodeflow* R package—also developed by the authors. Within *recodeflow*, the *rec\_with\_table()* function—short for “recode with table”—transforms 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 have 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, such as smoking pack-years, binge drinking, and diet pattern (Manuel et al. 2016). There are provisions and instructions for adding additional derived variables.

### 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.

## 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. 2020c). 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 1 illustrates how variables were added to *cchsflow*, while Figure 2 illustrates the homepage of the *cchsflow* package at <https://big-life-lab.github.io/cchsflow/>.

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.

### 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 data set, or a subset of variables. Figure 4a illustrates how to load the *cchsflow* package and 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 2014 versions, made possible with Statistics Canada new Open Licence (Statistics Canada 2020). Figure 4b illustrates how to transform a subset of variables from the 2001 survey cycle.

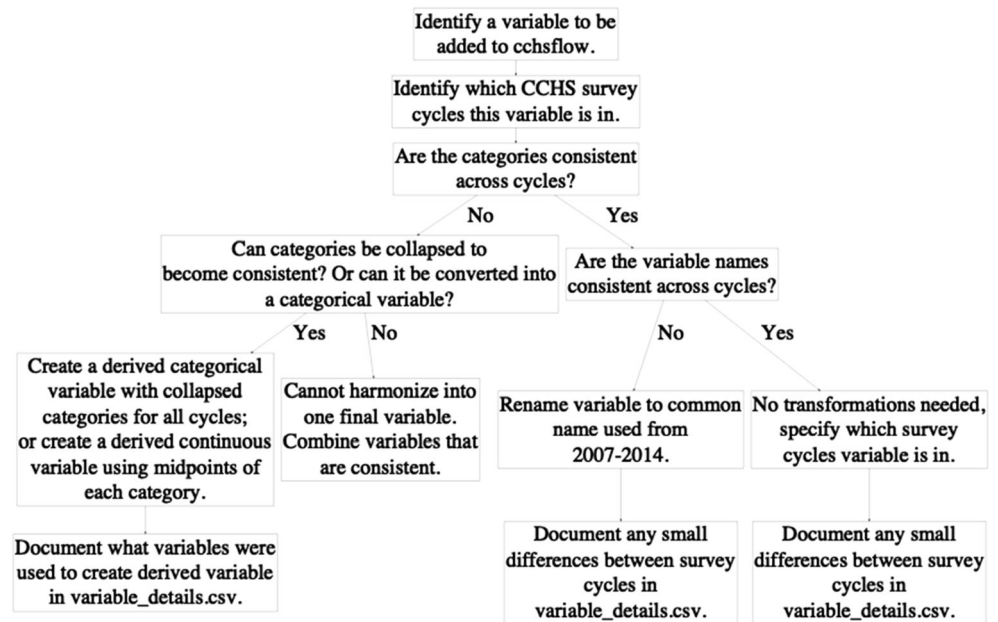
## Discussion

*cchsflow* harmonizes and transforms CCHS data from 2001 to 2014 (Manuel et al. 2020c). *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 who 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

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

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



(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.

Observational Health Data Sciences and Informatics (OHDSI) is an open science network that creates a common dictionary and software tools to support studies across different information systems (ODESI 2020). The focus of OHDSI

is hospital data. Investigators in different hospitals generate their own code to harmonize their hospital data into common, standard definitions.

*cchsflow* facilitates the use of CCHS metadata that come with the survey but are not commonly used by public health practitioners (Manuel and Fisher 2019). The CCHS comes with Data Documentation Initiative (DDI) metadata. DDI metadata are used worldwide for over ten thousand different surveys and research projects (Data Documentation Initiative (DDI) 2020). There are also initiatives such as Maelstrom that

## cchsflow

lifecycle maturing crn v1.6.0 GitHub v1.6.0 License MIT doi 10.17605/OSF.IO/HKUY3

*cchsflow* supports the use of the Canadian Community Health Survey (CCHS) by transforming variables from each cycle into harmonized, consistent versions that span survey cycles (currently, 2001 to 2014).

The CCHS is a population-based cross-sectional survey of Canadians that has been administered every two years since 2001. There are approximately 130,000 respondents per cycle. Studies use multiple CCHS cycles to examine trends over time and increase sample size to examine sub-groups that are too small to examine in a single cycle.

The CCHS is one of the largest and most robust ongoing population health surveys worldwide. The CCHS, administered by Statistics Canada, is Canada's main general population health survey. Information about the survey is found [here](#). The CCHS has a [Statistic Canada Open Licence](#).

## Concept

Each cycle of the CCHS contains over 1000 variables that cover the four main topics: sociodemographic measures, health behaviours, health status and health care use. The *seemingly* consistent questions across CCHS cycles entice you to combine them together to increase sample size; however, you soon realize a challenge...

Imagine you want to use BMI (body mass index) for a study that spans CCHS 2001 to 2014. BMI *seems* like a straightforward measure that is routinely-collected worldwide. 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 HWTGCGBMI, HWTGDBMI, HWTGEBMI, 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 < age 20 or > 64 years. Also, some cycles calculate BMI only if height and weight are within specific ranges. These types of changes occur for almost all CCHS variables.

**Fig. 2** The homepage for the *cchsflow* website



## Links

Download from CRAN at <https://cloud.r-project.org/package=cchsflow>

Browse source code at <https://github.com/Big-Life-Lab/cchsflow>

Report a bug at <https://github.com/Big-Life-Lab/cchsflow/issues>

Calculators at <https://www.projectbiglife.ca>

## License

MIT + file LICENSE

## Community

[Contributing guide](#)

[Code of conduct](#)

## Developers

Doug Manuel  
Author, copyright holder

Warsame Yusuf  
Author, maintainer

```

a
install.packages("cchsflow")

b
devtools::install_github("Big-Life-Lab/cchsflow")

```

**Fig. 3** **a** The command line to install the *cchsflow* package that is currently saved on CRAN. **b** The command line to install the development version of *cchsflow* from GitHub

are used by other Canadian health surveys to improve the use of metadata (Bergeron et al. 2018). Metadata are increasingly recognized as helpful data infrastructure to support open science and data harmonization. Metadata are “data about data” and include information about variable and category labels, variable types, and provenance (how the data were collected and transformed) (McGilvray 2008).

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 are included or consistent. 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 Model Markup Language (PMML) (Grossman et al. 1999). The Project Big Life team uses *cchsflow*'s PMML metadata in public health planning tools (Manuel et al. 2020a).

### Limitations and challenges

While the CCHS has many consistent variables across survey cycles, there are differences between cycles that can 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 only be used in select cycles. Along with variables with differences, there are variables in *cchsflow* that were not asked in all CCHS cycles. This means that for some variables, data do not span across the length of the CCHS cycles available in *cchsflow*. A possible solution is to impute missing variables, where missing data are replaced with values based on other respondents and responses to other variables.

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 data across survey cycles will result in misclassification error and other forms of bias that affect 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 et al. 2012).

The open-access approach *cchsflow* allows users to add other CCHS variables that might benefit others. There is full transparency on how the package was developed with the entire source code for the package publicly accessible. Along with being transparent, sourcing 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 or add new variable transformations themselves. New variables are added using a “pull request” that is then reviewed by the package maintainers before being merged with 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 allows 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

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

```

a
library(cchsflow)
cchs2001 <- read.csv("~/data/cchs2001.csv")
transformed_cchs <- rec_with_table(cchs2001)

b
library(cchsflow)
cchs2001 <- read.csv("~/data/cchs2001.csv")
transformed_cchs2001 <- rec_with_table(cchs2001, c("DHH_SEX", "DHHGAGE_cont")
)

```

reporting. Projects that examine health surveillance and health status reporting such as the Public Health Agency of Canada health inequality reports (Pan-Canadian Public Health Network 2018) can benefit from *cchsf*’s repository of harmonized variables.

## Roadmap

A roadmap, also known as next steps or future plans, is recommended for open software projects. *cchsf* includes a roadmap and milestones on the [project website](#). At the time of writing, the roadmap includes 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. *cchsf* has been forked by related projects to support other data sets. The expanded use of *cchsf* for related projects is a hallmark of open science and a demonstration of how open science leads to expanded science and public health resources.

## Conclusion

*cchsf*’s open science approach allows public health professionals to collaborate and share their work with other colleagues, saving time spent on recoding and cleaning health data. By implementing open science practices, *cchsf* aims to minimize the amount of time needed to clean and prepare CCHS data for the many CCHS users in health units across Canada.

**Author contributions** All authors contributed to the study conception and design. Data collection, variable selection, and software development were performed by Warsame Yusuf, Douglas G. Manuel, Rostyslav Vyuha, Carol Bennett, Yulric Sequeira, and Courtney Maskerine. The first draft of the manuscript was written by Warsame Yusuf and Douglas G. Manuel, and all authors commented on previous versions of the manuscript. All authors have read and approved the final manuscript.

**Funding** This study was funded by CIHR (FRN 162222).

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Association of Public Health Epidemiologists in Ontario. (2018). *Core Indicators Work Group*. <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/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](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*. CSCW: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. <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*. <https://ddialliance.org/>.
- Donoho, D. (2017). 50 years of data science. *J Comput Graph Stat*, *26*(4), 745–766. <https://doi.org/10.1080/10618600.2017.1384734>.
- 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. *Inf Softw Technol*. [https://doi.org/10.1016/S0950-5849\(99\)00022-1](https://doi.org/10.1016/S0950-5849(99)00022-1).
- Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *Am Stat*, *72*(4), 382–391.
- Journal of Open Source Software. (2018). *Journal of Open Source Software*. <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., Bennett, C., Rosella, L., Taljaard, M., Roberts, M., ... Manson, H. (2012). *Seven more years: The impact of smoking, alcohol, diet, physical activity and stress on health and life expectancy in Ontario*. [report]. Institute for Clinical Evaluative Sciences; Public Health Ontario. <https://www.ices.on.ca/Publications/Atlases-and-Reports/2012/Seven-More-Years>.
- Manuel, D. G., Perez, R., Sanmartin, C., Taljaard, M., Hennessy, D., Wilson, K., et al. (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. G., Tuna, M., Bennett, C., Hennessy, D., Rosella, L., Sanmartin, C., et al. (2018). Development and validation of a cardiovascular disease risk-prediction model using population health surveys: the cardiovascular disease population risk tool (cvdport). *CMAJ*, *190*(29), E871–e882. <https://doi.org/10.1503/cmaj.170914>.
- Manuel, D. G., Wilton, A. S., Bennett, C., Dass, R., Laporte, A., & Holford, T. R. (2020a). Smoking patterns based on birth-cohort-specific histories from 1965 to 2013, with projections to 2041. *Health Rep*, *31*(11), 16–31.
- Manuel, D. G., Bailey, L., Sequeira, Y., & Bennett, C. (2020b). *Project BigLife Planning Tool Guide*. <https://big-life-lab.github.io/pbl-planning-tool-guide/>.
- Manuel, D., Yusuf, W., Vyuha, R., & Bennett, C. (2020c). *cchsf*: Transforming and harmonizing cchs variables. <https://github.com/Big-Life-Lab/cchsf>.

- 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., et al. (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). *Bull World Health Organ*, 98(3), 150.
- Nowrouzi-Kia, B., Baig, A., Li, A., Casole, J., & Chai, E. (2019). Occupational injury trends in the Canadian workforce: An examination of the Canadian Community Health Survey. *International Journal of Critical Illness and Injury Science*, 9(1), 29.
- ODESI. (2020). *Ontario Data Documentation, Extraction Service and Infrastructure*. <https://odesi.ca>
- Open Source Initiative. (2020). *Open Source Initiative*. <https://opensource.org/>.
- Pan-Canadian Public Health Network. (2018). Key health inequalities in Canada: a national portrait.
- 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>.
- Statistics Canada. (2001). *CCHS Cycle 1.1 (2000-2001), Public Use Microdata File Documentation* (p. 77).
- Statistics Canada. (2020). *Statistics Canada Open Licence*. <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>.
- Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *J Bus Res*. <https://doi.org/10.1016/j.jbusres.2017.12.043>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.