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

**Authors:** Warsame Yusuf, Methodologist, Ottawa Hospital Research Institute. Douglas G. Manuel, Senior Scientist & Physician, Ottawa Hospital Research Institute. Rostyslav Vyuha, Research Assistant, Ottawa Hospital Research Institute. Carol Bennett, Epidemiologist, Ottawa Hospital Research Institute.

## Abstract

**Setting:** The Canadian Community Health Survey (CCHS) is one of the world’s largest cross-sectional population health surveys with over 130 000 respondents every two years or over 1.X million respondents since its inception in 2001. While the survey remains relatively consistent over the years, there are differences between cycles that pose a major challenge to analyse the survey over time.

**Intervention:** We developed a process to transform & harmonize CCHS variables to consistent formats across multiple survey cycles. We sought to use a current open science approach to maintain transparency, reproducibility and open collaboration.

**Outcomes:** The *cchsflow* R package was developed that 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 to generate harmonized datasets that can be combined into one consistently labelled dataset for analysis. The package was then added as a GitHub repository to encourage collaboration with other researchers.

**Implication:** The *cchsflow* package has been added to the Comprehensive R Network; and contains support for over 140 CCHS variables, generating a dataset of over 1 million respondents. By implementing open science practices, *cchsflow* aims to minimize the amount of time needed to clean & prepare data for the many CCHS users across Canada.

**Keywords:** Health Surveys, Data Analysis, Data Science, Population Health