CASE STUDY: RISK OF CARDIOVASCULAR DISEASE AMONG OSTEOARTHRITIS PATIENTS

Data Source: Canadian Community Health Survey (CCHS) cycles 1.1, 2.1 and 3.1.

Background: The Canadian Community Health Survey (CCHS) is a nationwide cross-sectional survey. This survey gathers health-related data for the Canadian population 12 years of age and over living in the 10 provinces and 3 territories, covering about 97% of the target population. In this case study, we will use Public Use Microdata Files (PUMF) from cycles 1.1, 2.1 and 3.1 that contain data collected in years 2000-2001, 2003 and 2005, respectively. Various measures were taken to protect the confidentiality of the participants of the survey.

The survey sampling weight provided corresponds to the number of individuals represented by the respondent for the target population. Incorporation of these weights will ensure an appropriate representation of the covered population, and hence these need to be considered to produce meaningful statistical estimates. Due to confidentiality concerns, only survey weights are made available on PUMF, but neither design information nor bootstrap weights for estimating variances are provided. Using the survey weight will provide correct point estimates, but in the absence of bootstrap weights and necessary design information, estimated variability measures calculated assuming simple random sampling will not be accurate and often be under estimated.

Research Question:

This case study aims to familiarize participants with analyzing the PUMF version of the CCHS dataset (combined data from cycles 1.1, 2.1 and 3.1). To that end, participants are asked to use this PUMF data to first create an ‘analytic dataset’ (with only the appropriate variables and records useful for analyses from all cycles), and then use that dataset to estimate crude and adjusted measures of association between osteoarthritis and self-reported heart diseases.

Questions to consider:

1. Within Canadian adults (20-64 years of age), is having osteoarthritis associated with the developing heart disease? For the purpose of this case study, assume that, from the literature, we know that the following variables are risk factors for the outcome and confounders in the above relationship: age, sex, ethnicity, education, household income, body mass index (BMI), access to a regular medical doctor, smoking habit, alcohol drinking habit, high-blood pressure, and diabetes. Also, assume that physical

activity is suspected to be an intermediate factor between osteoarthritis and heart disease.

2. Does the relationship between osteoarthritis and heart disease vary (a) between participants living in the northern parts of Canada versus those living in the southern parts, (b) between men and women, (c) by marital status, or (d) by recency of immigration?
3. Do the results change when missing values (i.e., invalid responses) for the 'household income' are imputed? Which assumptions do you have to make to perform such an analysis?
4. With the information provided in the PUMF, what would be your interpretation of the analysis results? What are the limitations of this study? What additional information would be helpful in reaching a more meaningful conclusion?

Variables:

In order to create an 'analytic dataset,' reviewing the corresponding data documentation (e.g., data dictionary, topical index and user guide associated with the data) for further details of the following variables is strongly recommended (e.g., check 'Universe'). It is often a good idea to cross-tabulate variables with the 'Age' variable (from the same cycle) to double check if the question was restricted to particular age groups. Similarly, cross-tabulating with the 'Province' variable often helps identify variables that were created from an 'optional CCHS component.' Note that, unless stated otherwise in the research question, some of the following variables may not be relevant for the relationship of interest. Also, there are no identifying information in cycle 1.1, 2.1 and 3.1 in this public-use data that will enable us to identify whether the same person was surveyed in multiple cycles. Therefore, for the purposes of this study (for simplicity), we will assume that the lists of subjects surveyed in different cycles were different.

	Variable names in 3 cycles			
Variable Concept	**CCHS 1.1**	**CCHS 2.1**	**CCHS 3.1**	**Comments** (see notes below)
Has heart disease	CCCA_121	CCCC_121	CCCE_121	Outcome. Only "YES" and "NO" are considered valid responses. (1)
Has arthritis or rheumatism	CCCA_051	CCCC_051	CCCE_051	Those who answered 'NO' are considered as 'NOT APPLICABLE' in the next variable 'kind of arthritis'.

Kind of arthritis	CCCA_05A	CCCC_05A	CCCE_05A	Useful for creating the exposure variable. Response “OSTEOARTHRITIS” will create the exposed group, and “NOT APPLICABLE” will create the unexposed group. (2)
Age	DHHAGAGE	DHHCGAGE	DHHEGAGE	Recode into categories that make sense and apply to all 3 cycles. (3)
Sex	DHHA_SEX	DHHC_SEX	DHHE_SEX	
Marital Status	DHHAGMS	DHHCGMS	DHHEGMS	Recode into categories that make sense and apply to all 3 cycles. (1)
Cultural / racial origin	SDCAGRAC	SDCCGRAC	SDCEGCGT	(1)
Immigrant status	SDCAFIMM	SDCCFIMM	SDCEFIMM	Those who answered ‘NO’ is considered as ‘NOT APPLICABLE’ in the next variable ‘Length of time in Canada since immigration’. (4)
Length of time in Canada since immigration	SDCAGRES	SDCCGRES	SDCEGRES	(4)
Highest level of education - respondent	EDUADR04	EDUCDR04	EDUEDR04	Recode into categories that make sense and apply to all 3 cycles. (1)
Total household income from all sources	INCAGHH	INCCGHH	INCEGHH	Recode into categories that make sense and apply to all 3 cycles. (1)
Body mass index	HWTAGBMI	HWTCGBMI	HWTEGBMI	Recode into categories 3 categories: underweight (<18.5), healthy weight (between 18.5 and 25), overweight (>25). (1)
Physical activity index	PACADPAI	PACCDPAI	PACEDPAI	(1)
Has a regular medical doctor	TWDA_5	HCUC_1AA	HCUE_1AA	(1)

Type of smoker	SMKADSTY	SMKCDSTY	SMKEDSTY	Recode into categories that make sense and apply to all 3 cycles. (1)
Type of drinker	ALCADTYP	ALCCDTYP	ALCEDTYP	Recode into categories that make sense and apply to all 3 cycles. (1)
Has high blood pressure	CCCA_071	CCCC_071	CCCE_071	(1)
Has diabetes	CCCA_101	CCCC_101	CCCE_101	(1)
Has emphysema or chronic obstructive pulmonary disease (COPD)	CCCA_91B	CCCC_91B	CCCE_91F	(1)
Daily consumption - total fruits and vegetables	FVCADTOT	FVCCDTOT	FVCEDTOT	Recode into categories 3 categories, 0-3, 4-6 and 6+ daily serving. (1)
Self-perceived stress	GENA_07	GENC_07	GENE_07	Recode into categories that make sense and apply to all 3 cycles. (1)
Province	GEOAGPRV	GEOCGPRV	GEOEGPRV	Recode Northwest Territories, Nunavut, Yukon as 'north' and the rest of the provinces/territories as 'south'. (1)
Sampling weight - master weight	WTSAM	WTSC_M	WTSE_M	Divide them by 3 to get a nationally representative sample (on average).

Note:

1. The following are considered invalid responses, and hence may be considered as missing values: "NOT APPLICABLE", "DON'T KNOW", "REFUSAL", "NOT STATED" unless otherwise stated. For a complete case analysis, all of these records could be excluded from the study.
2. Responses "RHEUMATOID ARTHRITIS" and "OTHER" will be excluded from the study.
3. According to study eligibility criteria, the study will be restricted to participants 20-64 years of age.

4. Useful for creating immigration status (potential categories: “not immigrant,” “recent immigrant,” “immigrated more than 10 years ago”)

Data Access:

Data and Documentation Files:
https://www.dropbox.com/sh/dntqkl6wv54ypop/AACPOf6pnGh4sgithHJRQyYYa?dl=1

The zip files (inside the download) contain the unedited original Statistics Canada public-use datafiles. The data files are also provided in RData formats (converted from the original datasets), suitable for opening in R. Note that, other than the format of the data, the content and the corresponding documentations are unedited. These data files should be the same as the data that can be accessed through the Data Liberation Initiative (DLI) member universities. Associated documentations (e.g., data dictionary, topical index and user guide associated with the data) and licence agreements are provided in the respective 'documentation' folders within the zip files. Finally, note that the use of these files must be done with respect to the terms and conditions of the Statistics Canada Open Licence (link: https://www.statcan.gc.ca/eng/reference/licence). Please consult the licence agreements provided here before downloading.

Number of records and variables:

	CCHS 1.1	**CCHS 2.1**	**CCHS 3.1**
Number of records	130,880	134,072	132,221
Number of variables	614	1,068	1,284