cmajOPEN

Research

Harmonization of the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project: a descriptive analysis

Isabel Fortier PhD, Nataliya Dragieva MSc, Matilda Saliba PhD, Camille Craig MSc, Paula J. Robson PhD; with the Canadian Partnership for Tomorrow Project's scientific directors and the Harmonization Standing Committee*

Abstract

Background: The Canadian Partnership for Tomorrow Project is a multistudy platform integrating the British Columbia Generations Project, Alberta's Tomorrow Project, the Ontario Health Study, CARTaGENE (Quebec) and the Atlantic Partnership for Tomorrow's Health. This paper describes the process used to harmonize the Health and Risk Factor Questionnaire data and provides an overview of the key information required to properly use the core data set generated.

Methods: This is a descriptive analysis of the harmonization process that was developed on the basis of the Maelstrom Research guidelines for retrospective harmonization. Core variables (DataSchema) to be generated across cohorts were defined and the potential for cohort-specific data sets to generate the DataSchema variables was assessed. Where relevant, algorithms were developed and applied to process cohort-specific data into the DataSchema format, and information to be provided to data users was documented.

Results: The Health and Risk Factor Questionnaire DataSchema (version 2.0, October 2017) comprised 694 variables. The assessment of harmonization potential for the variables over 12 cohort-specific data sets resulted in 6799 (81.6%) of the variables being considered as harmonizable. A total of 307 017 participants were included in the harmonized data set. Through the cohort data portal, researchers can find information about the definitions of variables, harmonization potential, algorithms applied to generate harmonized variables and participant distributions.

Interpretation: The harmonization process enabled the creation of a unique data set including data on health and risk factors from over 307 000 Canadians. These data, in combination with complementary data sets, can be used to investigate the impact of biological, environmental and behavioural factors on cancer and chronic diseases.

Large prospective population-based cohorts are crucial to obtaining the statistical power required to investigate direct and interactive effects of genetic, social, environmental and lifestyle factors on health outcomes.[1] Very large cohorts have been established in the last few decades. Some, like UK Biobank,[2] Lifelines[3] and EpiHealth,[4] are designed as single studies, while others are based upon collaborative efforts in which partner studies collect and integrate a common set of core information. Examples of cohorts in the latter category include the European Prospective Investigation into Cancer and Nutrition,[5] the Health and Retirement Study[6] and the Canadian Partnership for Tomorrow Project.[7,8] This integrative design is valuable for leveraging innovative research, while supporting the specific interests of partner study investigators. However, it can be challenging to recruit compatible populations, implement coordinated participant follow-up, collect and integrate large sets of common variables and offer timely access to data and samples collected by partner studies. Ethicolegal and governance models, financial and time constraints, infrastructures and technical resources accessible, and specific population or social contexts often vary across partner studies, leading to unavoidable heterogeneity. Thus, while a collaborative approach allows the identification of common data elements

Competing interests: None declared.

*Lists of the scientific directors and committee members appear at the end of the article.

This article has been peer reviewed.

Correspondence to: Isabel Fortier, ifortier@maelstrom-research.org

***CMAJ Open* 2019. DOI:10.9778/cmajo.20180062**

to be collected, it is important to implement a rigorous process to harmonize, integrate, document, disseminate and provide access to the core data generated by partner studies.[9,10]

As described by Dummer and colleagues,[8] the Canadian Partnership for Tomorrow Project is a multistudy cohort of more than 307 000 participants, created to support a broad range of research activities related to the determinants of cancer and other chronic diseases.[7,8] This strategic partnership brought together 5 Canadian cohorts across 8 provinces: the British Columbia Generations Project,[11] Alberta's Tomorrow Project,[12] the Ontario Health Study,[13] CARTaGENE (Quebec)[14] and the Atlantic Partnership for Tomorrow's Health (Nova Scotia, New Brunswick, Prince Edward Island, and Newfoundland and Labrador).[15] When the Canadian Partnership for Tomorrow Project was established in 2008, investigators agreed to collect core data about health and risk factors, physical measures and biological samples.[8] However, because of technical, methodological and financial constraints, study participants' sampling and recruitment procedures as well as data collection tools and methods used (e.g., online or paper questionnaire) varied across cohorts and over time. This resulted in the creation of 12 data sets including participants with specific characteristics and information collected with distinct instruments. To facilitate integrated data analysis, it was thus suitable to implement a rigorous harmonization process to generate, where possible, a set of core (or harmonized) data across these data sets. Dummer and colleagues' paper[8] provides an overview of the general cohort profile; the present paper is a descriptive analysis of the methods used to harmonize a subset of the cohort data. The paper describes the methodological approach, tools and process used to harmonize the specific health and risk factor data collected by cohorts and provides an overview of the key information required to properly use the core data set generated (version 2.0, October 2017).

Methods

Before the recruitment of the study participants,[8] a proposed list of core variables to be collected across regional cohorts and a questionnaire to be used to collect information were generated using a consensus approach involving the cohort investigators and harmonization team.[8] The working group included epidemiologists, social scientists, statisticians, geneticists, physicians and lawyers with expertise in different domains including, but not limited to, the following: environmental exposures, nutrition, genomics, cancer and chronic diseases. Through a series of consensus workshops, the working group discussed the scientific focus of the cohort, selected generic domains of interest (e.g., diabetes, alcohol intake, etc.), proposed variables and standard questionnaires to be used to collect information, compared the proposed variables with information collected by similar cohorts, obtained agreement on a final set of core variables to be collected and agreed on the specific questions to be used to collect data. Selection of variables, definition of variable categories and question wording were informed by existing tools,[16–18] standards[19,20] and questionnaires.[2,5,21] The proposed Health and Risk Factor Questionnaire was used as a reference to develop the regional questionnaires but adjusted by the cohort investigators to comply with maturing cohort-specific designs and data collection modes, leading to the generation of similar, but not identical, instruments. Variations in questionnaires and participant sampling led to the creation of 12 cohort-specific data sets, including distinct subgroups of participants. The Maelstrom Research guidelines for retrospective data harmonization[9] were used to guide retrospective harmonization of the cohort-specific data sets and generate the core set of harmonized Health and Risk Factor Questionnaire data (version 2.0, October 2017) required to support the Canadian Partnership for Tomorrow Project's activities. Figure 1 provides an overview of the data implementation, collection and harmonization procedures.

After initiation of the data collection, a Harmonization Standing Committee supported the retrospective harmonization process achieved by the Maelstrom Research team. The committee members changed over the years, but the committee included at least 9 members (epidemiologists, biostatisticians and data managers): 2 representatives from Maelstrom Research (principal investigator and project coordinator), 1 data manager from each of the regional cohorts, 1 of the regional cohorts' princial investigators and a representative of the project coordination centre. Members of the committee had monthly conference calls during which the Maelstrom Research team reported the evolution of the harmonization process, raised questions to be discussed and presented information to be reviewed by the cohorts. Cohort representatives were then responsible for reviewing the harmonization outputs for their study (e.g., algorithms generated, variables documented and inconsistent data content identified). Regular meetings with regional cohort investigators allowed for discussion of strategic issues related to the harmonization process and resolution of key problems encountered. The retrospective harmonization process included the following steps.

Step 1: assemble information on cohort-specific data sets

For each data set, study participants' sampling and recruitment processes as well as methods used to collect data were documented (Table 1). Cohorts' questionnaires and code books were assembled by the harmonization team and the 12 data sets were uploaded on a central Opal server.[22] The content (e.g., adequate list of variables and participants) of each cohort-specific data set was explored. Univariate distributions were generated for all discrete and continuous variables, and the occurrence of outliers and missing values was reviewed. Cross references between variables were also generated (e.g., occurrence of prostate cancer and sex). Inconsistencies identified were reported to the data managers of the regional cohorts and, when applicable, were corrected.

Step 2: define core variables to be generated and evaluate harmonization potential

After review of the variables included in the cohort-specific data sets provided, as well as the list of variables selected before participant recruitment, a final list of core variables to

be generated across data sets (DataSchema) was created by the Maelstrom Research team and reviewed by the Harmonization Standing Committee members. Each DataSchema variable was defined according to standardized specifications that included its name, label, scientific meaning (e.g., reported lifetime occurrence of diabetes diagnosed by a doctor), format (e.g., text, decimal), categories and units. Variables were defined to facilitate harmonization across cohort-specific data sets, while aiming to limit harmonization to data considered compatible. The ability of the cohort-specific data

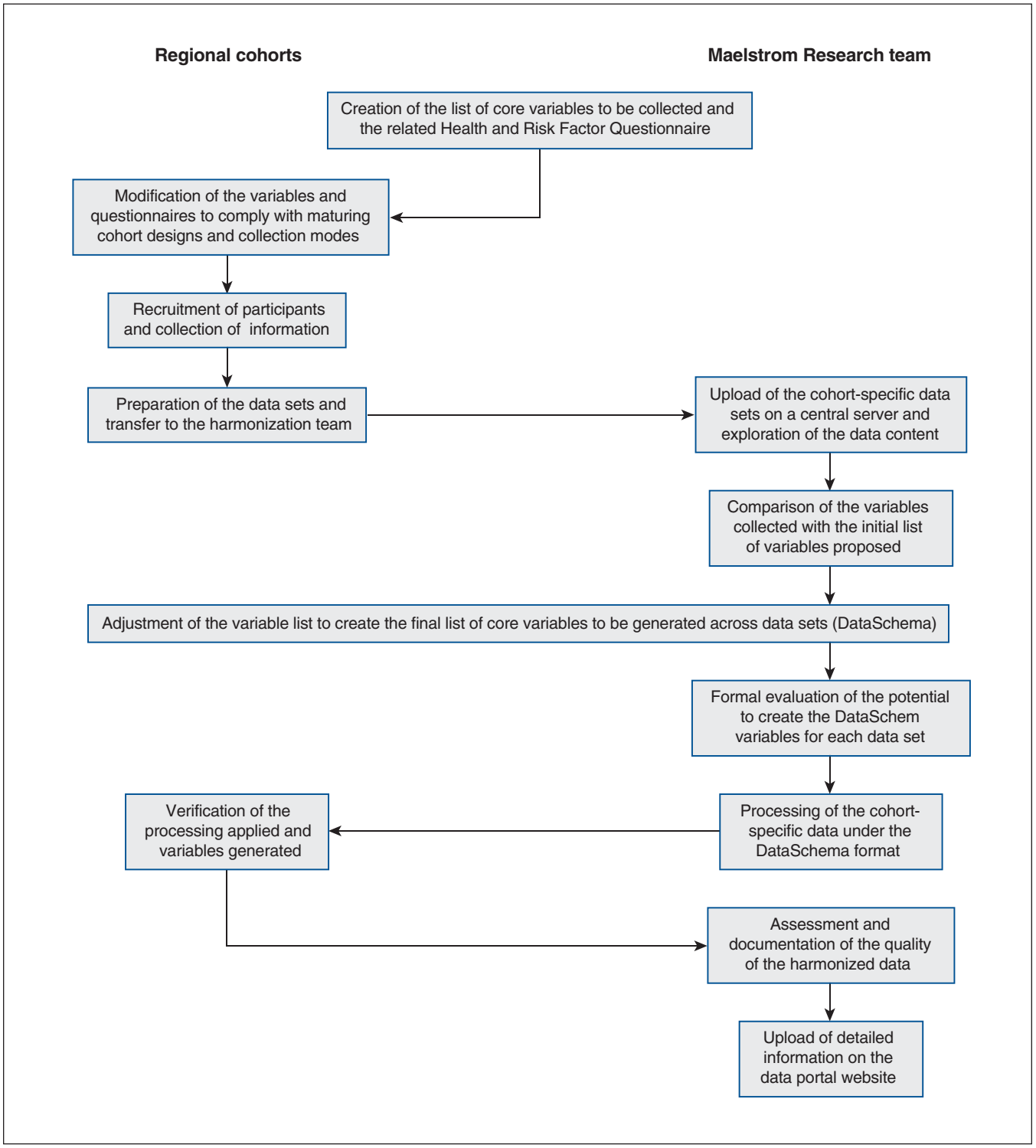

Figure 1: Flowchart describing the data preparation, collection and harmonization process by the regional cohorts and the Maelstrom Research harmonization team.

Table 1: Number of participants, recruitment profile, inclusion/exclusion criteria and data collection methods for the Health and Risk Factor Questionnaire for each data set

Data set (yr collected)	No. of participants	Source of participants	Invitation to participate	Inclusion/exclusion criteria	Data collection mode; procedures; and language
ATL-DS1 (2010–2016)	22 491	Open community; employees	Advertisements	≥ 18 yr; Nova Scotia, New Brunswick, Prince Edward Island, Newfoundland and Labrador	Online and paper (manual data entry) forms; auto-administered; English, French
ATL-DS2 (2009–2014)	11 038	Open community; health/administrative registries (Nova Scotia); employees	Letters (Nova Scotia); advertisements	≥ 18 yr; Nova Scotia, New Brunswick, Prince Edward Island	Electronic (on-site) and paper (manual data entry) forms; auto-administered, face-to-face support available for some participants; English, French
ATP-DS1 (2009–2013)	12 239	Open community; participants of existing ATP cohort study	Letters; phone; random digit dialling	35–69 yr; plans to reside in Alberta for at least 1 yr; no history of cancer other than nonmelanoma skin cancer	Paper form (automatic data entry); auto-administered; English
ATP-DS2 (2009–2015)	26 545	Open community; participants of existing ATP cohort study; employees; commercial mailing lists	Letters; phone; random digit dialing; advertisements	35–69 yr; plans to reside in Alberta for at least 1 yr; no history of cancer other than nonmelanoma skin cancer	Paper form (automatic data entry); auto-administered; English
BCGP-DS1 (2013–2016)	3489	Open community; employees; friends/ family referral	Letters; advertisements	35–69 yr; all province	Online form; auto-administered; English
BCGP-DS2 (2011–2016)	17 479	Open community; employees; friends/ family referral	Letters; advertisements	30–74 yr; all province	Paper form (automatic data entry); auto-administered; English
BCGP-DS3 (2009–2010)	7840	Open community; employees; friends/ family referral	Letters; advertisements	30–74 yr; all province	Electronic (on-site) forms; auto-administered, face-to-face support available; English. Paper (manual data entry) forms; auto-administered; English
CaG-DS1 (2009–2010)	19 958	Health/administrative registries	Letters; phone	40–69 yr; Montréal, Québec, Sherbrooke and Saguenay regions	Electronic (on-site) form; auto-administered, face-to-face support available; English, French
CaG-DS2 (2012–2015)	22 475	Health/administrative registries	Letters; phone	40–69 yr; Montréal, Québec, Sherbrooke, Saguenay, Gatineau and Trois-Rivières regions	Online form; auto-administered; English, French
OHS-DS1 (2010–2013)	141 725	Open community; employees; commercial mailing lists; friends/ family referral	Letters; emails; advertisements; incentives offered	35–65 yr; all province	Online form; auto-administered; English, French
OHS-DS2 (2012–2015)	13 768	Open community; employees; commercial mailing lists; friends/ family referral	Letters; emails; advertisements	30–74 yr; all province	Online form; auto-administered; English, French
OHS-DS3 (2009–2010)	7970	Open community; employees; commercial mailing lists; friends/ family referral	Letters; advertisements	35–69 yr; Mississauga, Owen Sound, Sudbury	Electronic (on-site) form; auto-administered and interview; English, French, others

Note: ATL-DS1 = Atlantic Partnership for Tomorrow's Health (PATH) data set 1, ATL-DS2 = Atlantic PATH data set 2, ATP-DS1 = Alberta's Tomorrow Project data set 1, ATP-DS2 = Alberta's Tomorrow Project data set 2, BCGP-DS1 = British Columbia Generations Project data set 1, BCGP-DS2 = British Columbia Generations Project data set 2, BCGP-DS3 = British Columbia Generations Project data set 3, CaG-DS1 = CARTaGENE data set 1, CaG-DS2 = CARTaGENE data set 2, OHS-DS1 = Ontario Health Study data set 1, OHS-DS2 = Ontario Health Study data set 2, OHS-DS3 = Ontario Health Study data set 3.

sets to generate each of the DataSchema variables was then assessed by research assistants with backgrounds in epidemiology or biostatistics. For each cohort-specific data set, a harmonization status was assigned to each DataSchema variable. The status was deemed complete when the data set allowed construction of the DataSchema variable as defined, using a direct mapping (same question wording and categories) or through transformation/combination of cohort-specific variable(s), and it was deemed incomplete when this was impossible. An incomplete status was attributed when information was not collected or incompatible with the DataSchema variable defined. Table 2 provides an example of harmonization status for the variable "Ever had colonoscopy." At the end of the process, only DataSchema variables created by at least 2 of the cohort-specific data sets were retained.

Step 3: process cohort-specific data under a common format

Transformation algorithms were developed and applied using Opal and R software for each variable with complete status.[22] Algorithms were generated by the research assistants responsible for the evaluation of the harmonization potential and reviewed by the Harmonization Standing Committee. Cohort-specific data were processed into DataSchema format to create 12 cohort-specific harmonized data sets. These data sets were finally aligned to form the Canadian Partnership for Tomorrow Project Health and Risk Factor Questionnaire harmonized data set (version 2.0, October 2017).

Step 4: estimate the utility and quality of the harmonized data set(s) generated

Descriptive statistics outlining the harmonization potential across the cohort-specific data sets (coverage of complete harmonization status) were generated. The number of data sets allowing generation of each DataSchema variable and the number of harmonized variables generated for each cohort-specific data set were calculated. To explore harmonized data content, univariate distributions were generated for each variable included in the integrated core data set, and total number of participants, logic of distributions (e.g., smokers v. nonsmokers), occurrence of outliers and missing values were reviewed. Cross references between selected variables were generated (e.g., sitting height < standing height). Inconsistencies identified were reported to Harmonization Standing Committee members and, when possible, were corrected. Finally, variabilities in participant distributions and missing values across data sets were explored. The potential impact of sampling and recruitment process as well as methods used to collect data were examined and documented to help understand data content (e.g., exclusion of participants with a cancer history at recruitment resulted in lower cancer rates for 2 harmonized data sets).

Step 5: disseminate and preserve final harmonization products (data and metadata)

The Canadian Partnership for Tomorrow Project Web-based data portal (https://portal.partnershipfortomorrow.ca) was set up using the Maelstrom Research cataloguing tool kit, particularly the MICA software application,[23] to collate and disseminate information on the regional cohorts' designs, the definitions of the DataSchema variables, the harmonization potential across data sets, the algorithms applied to generate each harmonized variable and the distributions of participants.

Statistical analysis

Descriptive statistics were used to explore the content of the cohort-specific and harmonized data sets. Participant distributions were documented to provide the users of the Canadian Partnership for Tomorrow Project with the minimal basic information required to understand and use the resource and to identify needs to further explore data content.

Table 2: DataSchema variable "Ever had colonoscopy"; example of cohort-specific questions and related harmonization potentials

DataSchema variable: ever had colonoscopy Description: indicator of whether the participant has ever had a colonoscopy Format: categorical Categories: 0 = never had a colonoscopy, 1 = ever had a colonoscopy	
Specific questions used by participating cohorts	Harmonization status
Format A (used in 1 data set) Have you ever had a colonoscopy? [ ] No; [ ] Yes; [ ] Prefer not to answer; [ ] Don't know	Complete (direct match to the DataSchema variable, but the "Prefer not to answer" and "Don't know" categories classified as missing)
Format B (used in 6 data sets) When was the last time you had a colonoscopy? A colonoscopy is an exam where a long tube is used to examine the entire colon. Before the procedure is done, you are usually given a sedative. [ ] Less than 6 mo ago; [ ] 6 mo to less than 1 yr ago; [ ] 1 yr to less than 2 yr ago; [ ] 2 yr to less than 3 ys ago; [ ] 3 or more yr ago; [ ] Never; [ ] Don't know; [ ] I prefer not to answer	Complete (algorithm combining multiple categories developed)
Format C (used in 5 data sets) Have you ever had a colonoscopy or sigmoidoscopy? These are tests where a long tube is inserted into the rectum to view the bowel for early signs of cancer and other health problems. [ ] Yes; [ ] No; [ ] Don't know; [ ] Prefer not to answer	Incomplete (colonoscopy and sigmoidoscopy are integrated in the question)

Ethics approval

Participants from all the Canadian Partnership for Tomorrow Project cohorts provided informed consent at recruitment. The central and regional cohorts' data access committees approved the harmonization project and granted access to the data required to generate the harmonized data set.

Results

Cohort-specific data

The number of participants varied across cohort-specific data sets, ranging from 3489 to 141 725 participants per data set, for a total of 307 017 (Table 1). Differences were observed in participants' age and sex distributions across data sets (Figure 2). Cohort participants' age and sex distributions also differed from those of the Canadian population (Table 3). Information on participant distributions for all DataSchema variables is available on the data portal[24] and an overview of the representativeness of participants in relation to the Canadian population is provided in Dummer and colleagues.[8] Relatively few quality issues were identified in the cohort-specific data sets provided. The most frequent problems were related to content format (e.g., trailing white spaces, line terminations and inappropriate representation of missing values), missing information, skip patterns (e.g., nonsmokers reporting a given number of cigarettes smoked/wk) and outliers (e.g., participant having 72 brothers).

Harmonization potential

The final DataSchema (version 2.0, October 2017) comprised 694 variables. An overview of the number of variables by area of information is presented in Table 4 and the full list of variables is provided in Appendix 1 (available at www.cmajopen.ca/content/7/2/E272/suppl/DC1). Unusual formats or duplications can be observed in some of the DataSchema variables. For example, both the International Physical Activity Questionnaire long and short forms were used by the cohorts. Because of the divergence reported between the questionnaire's versions,[19,25] distinct sets of harmonized variables were created for each version.

Of the 8328 harmonization status assessments for the 694 variables across the 12 data sets, 6799 (81.6%) were deemed complete and 1529 (18.4%) incomplete. The harmonization status across all variables is in Appendix 2 (available at www.cmajopen.ca/content/7/2/E272/suppl/DC1). Almost half of the algorithms developed for the variables with complete status comprised direct mapping (identical categories between the cohort-specific variables and DataSchema variable), whereas the other half necessitated a more complex procedure to account for variability in question wording, categories, skip

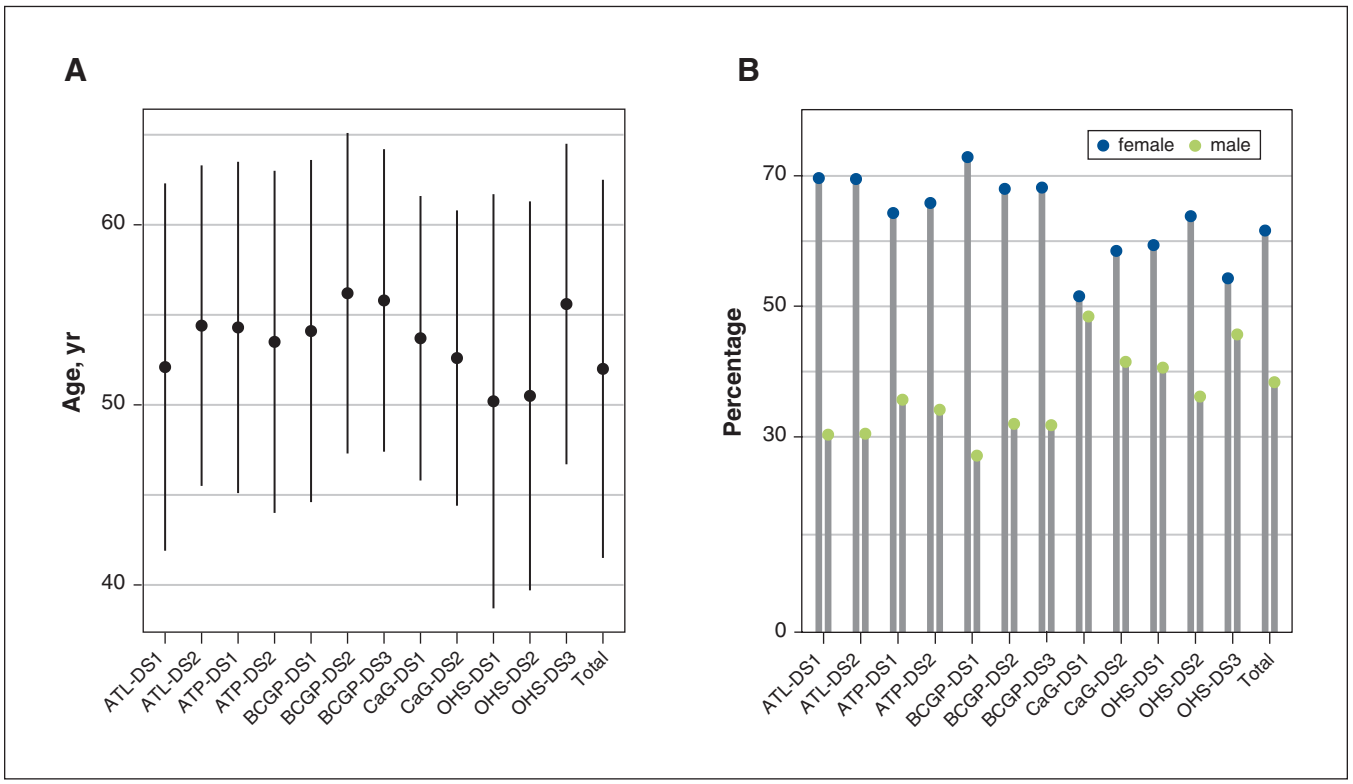

Figure 2: Overview of the (A) age distributions (mean and standard deviations) and (B) sex distributions across integrated and cohort-specific data sets. Note: ATL-DS1 = Atlantic Partnership for Tomorrow's Health (PATH) data set 1, ATL-DS2 = Atlantic PATH data set 2, ATP-DS1 = Alberta's Tomorrow Project data set 1, ATP-DS2 = Alberta's Tomorrow Project data set 2, BCGP-DS1 = British Columbia Generations Project data set 1, BCGP-DS2 = British Columbia Generations Project data set 2, BCGP-DS3 = British Columbia Generations Project data set 3, CaG-DS1 = CARTaGENE data set 1, CaG-DS2 = CARTaGENE data set 2, OHS-DS1 = Ontario Health Study data set 1, OHS-DS2 = Ontario Health Study data set 2, OHS-DS3 = Ontario Health Study data set 3.

Table 3: Sex and age distributions in the Canadian Partnership for Tomorrow Project cohort and in the Canadian population

	Cohort, %	Canadian population, %*
Sex		
Women	61.6	50.4
Men	38.4	49.6
Age, yr		
30–35	5.7	12.0
35–44	20.4	23.2
45–54	30.7	26.0
55–64	28.6	23.3
65–74	14.6	15.4

*Calculated using data from the following webpage: Statistics Canada. Table 17-10-0005-01: Population estimates on July 1st, by age and sex. Available: www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501 (accessed 27 Mar. 2019).

patterns and response types provided. Most of the variables that received an incomplete status received this status because information was not included in the questionnaire(s).

More than 60% (n = 432) of the DataSchema variables were created across at least 11 data sets, and 27.1% (n = 188) were created across all data sets. For each data set, the total number of DataSchema variables generated varied between 488 (70.3%) and 652 (93.9%), except for 1 data set that generated 322 variables (46.4%) because the questionnaire version used was shorter (see also Appendix 2).

Harmonized data content

Participant distributions for all DataSchema variables (integrated and data set-specific) can be obtained on the cohort data portal.[24] Figure 3 presents examples of participant distributions across data sets pertaining to the lifetime occurrence of cancer, smoking status, education level and familial income. Percentages observed varied across data sets. For example, the percentage of smokers ranged from 5.2% for the British Columbia Generation Project data set-1 to 19.0% for the CARTaGENE data set-1. The percentages of missing data were 0%–10.9%, 0.3%–11.9%, 0%–7.1% and 2.9%–18% for the occurrence of cancer, smoking status, education level and familial income, respectively. Distributions also varied across study-specific data sets for most of the harmonized variables, including sociodemographic and socioeconomic characteristics, lifestyle habits and health outcomes. Box 1 provides examples of factors that potentially explain variability and that need to be considered, where relevant, when analyzing and interpreting the Health and Risk Factor Questionnaire variables.

Data dissemination and management

Information regarding the regional cohort profiles (e.g., participants' characteristics, source of recruitment and timelines of data collections) as well as variable-specific harmonization outputs can be found on the data portal.[24] The harmonization status results are provided on the portal with links to the definitions and characteristics of the DataSchema variables, the algorithms used to process cohort-specific data into each DataSchema variable and the descriptive statistics (e.g., means and participant distributions across categories) for the core Health and Risk Factor Questionnaire data set and the 12 harmonized cohort-specific data sets. As the portal is regularly updated, online information references the most recent version of the data sets and varies from the information (version 2.0, October 2017) reported in this paper.

Interpretation

The present paper describes the process used to harmonize the Health and Risk Factor Questionnaire data (version 2.0, October 2017) and provides an overview of the information to be considered when analyzing the harmonized data generated. Similar to the European Prospective Investigation into Cancer and Nutrition[5] study and the Health and Retirement Study,[6] the Canadian Partnership for Tomorrow Project is a multistudy initiative aiming to integrate a core set of data across individual partner cohorts that have specific characteristics and that have collected additional regional data to support particular research activities. A number of multistudy cohorts have successfully generated core sets of harmonized data. However, none have provided online access to information that is as comprehensive, user-friendly and interactive as the Canadian Partnership for Tomorrow Project's portal, which includes the designs of the cohorts, definitions of variables, harmonization potential/process and participant distributions. The methods and open source software developed by Maelstrom Research while harmonizing the Canadian Partnership for Tomorrow Project data are now available to the research community (www.maelstrom-research.org) and are being used by an increasing number of multistudy projects around the world.

As a multistudy effort, the Canadian Partnership for Tomorrow Project has advantages and disadvantages. It optimizes the impact of each of the individual cohorts by allowing users to obtain the very large numbers of participants needed to generate sufficient statistical power to investigate relatively rare events and explore interactions between selected risk factors (e.g., genetic and environmental factors). It also probably increases exposure heterogeneity and enables users to undertake more refined subgroup analysis and comparison, cross validation or replication across data sets. However, usage and interpretation of the harmonized data require understanding and consideration of cohort-specific characteristics and participant profiles as well as the protocol used to generate the harmonized data. For example, it would be suboptimal to consider the variable "lifetime occurrence of cancer" without acknowledging that at initial recruitment 1 of the partner cohorts (Alberta's Tomorrow Project) excluded participants with a history of cancer other than nonmelanoma skin cancer. To support external researchers making use of the harmonized data set, it is thus essential to provide open access to

Table 4: DataSchema variables by area of information

Area of information (no. of variables)	Total no. of variables in the area of information (% of all DataSchema variables)
Sociodemographic and economic characteristics	109 (15.7)
Language (34), labour force and retirement (25), ethnicity and religion (15), birthplace (14), family and household structure (10), income, possessions, and benefits (3), education (2), marital/partner status (1), residence (1), sex/gender (1), twin (1), age/birthdate (1), other sociodemographic characteristics (1)	
Lifestyle and health behaviours	95 (13.7)
Tobacco (45), alcohol (26), physical activity (14), nutrition (6), sleep (4)	
Health status and functional limitations	1 (0.1)
Perception of health (1)	
Diseases (*International Statistical Classification of Diseases and Related Health Problems, 10th Revision*)	373 (53.7)
Neoplasms (179), digestive system (46), respiratory system (40), musculoskeletal system (31), circulatory system (30), skin and subcutaneous tissues (18), endocrine, nutritional and metabolic diseases (11), mental and behavioural disorders (9), nervous system (9)	
Medication and supplements	40 (5.8)
Medication and supplements (40)	
Nonpharmacological interventions	20 (2.9)
Other nonpharmacological interventions (6), surgical interventions (6), biosample analyses (6), radiological interventions (2)	
Health and community care utilization	4 (0.6)
Visit to health professionals (4)	
Reproduction	23 (3.3)
Menstruation, menopause and andropause (7), gravidity, pregnancy outcomes, parity (female) and fertility (male) (7), pregnancy, labour and delivery (4), contraception and family planning (3), breastfeeding (1), reproductive and sexual problems (1)	
Physical measures	9 (1.3)
Anthropometry (6), physical characteristics (3)	
Physical environment	11 (1.6)
Chemical exposure (6), radiation exposure (5)	
Administrative information	9 (1.3)
Questionnaire- and interview-related information (5), identifiers (2), date and time (1), information related to physical measures and biosamples (1)	
Total	694 (100)

detailed information concerning study designs and harmonized data, including variable definitions, harmonization potential, algorithms applied to generate harmonized variables and participant distributions across data sets. The cohort data portal, and the support provided by the cohort investigators and the harmonization team, aim to ensure access to transparent, consistent and practical information so that users can understand and properly interpret and analyze the harmonized data.

Limitations

The quality of the harmonized data depends on the quality of the data set provided by each of the participating cohorts. Although the cohorts used rigorous approaches to collect and manage data, each has strengths and weaknesses.[11–15] Weaknesses include limitations in the cohort-specific sampling frames and recruitment processes; the fact that questionnaires were not necessarily formally validated even if they were piloted and tested; and the influence of data collection modes and procedures on the prevalence of missing values.

Various factors can contribute to the heterogeneity observed in data content across harmonized data sets. As expected, heterogeneity may be explained by differences in participants' characteristics, but it may also be associated with variability in the methods (e.g., online v. paper forms) and questionnaire versions used to collect data or in the cohort-specific data management procedures applied (e.g., data cleaning protocol). Our rigorous harmonization process helped to

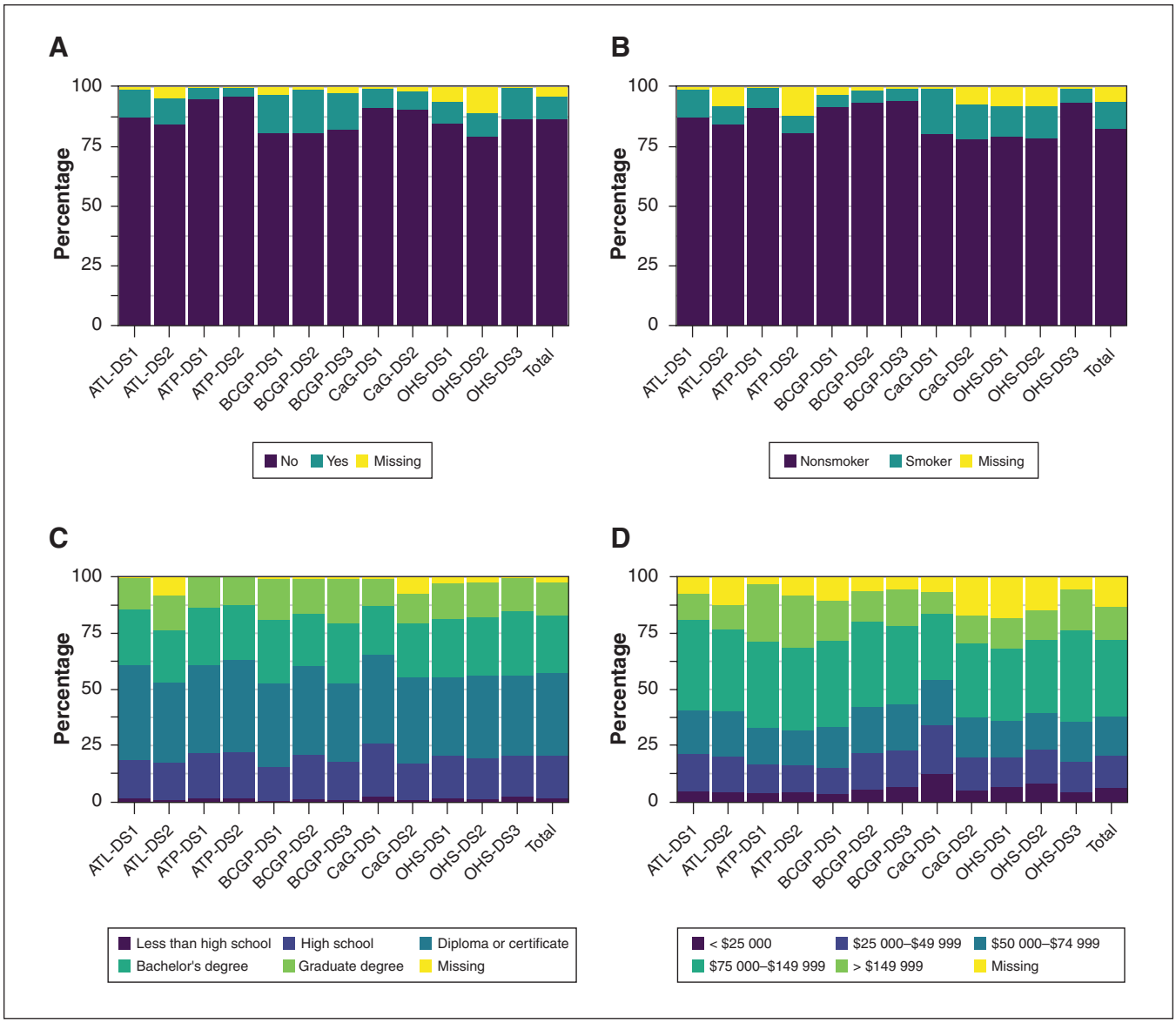

Figure 3: Distribution of participants according to (A) lifetime occurrence of cancer, (B) current smoking status, (C) education level and (D) familial income. Note: ATL-DS1 = Atlantic Partnership for Tomorrow's Health (PATH) data set 1, ATL-DS2 = Atlantic PATH data set 2, ATP-DS1 = Alberta's Tomorrow Project data set 1, ATP-DS2 = Alberta's Tomorrow Project data set 2, BCGP-DS1 = British Columbia Generations Project data set 1, BCGP-DS2 = British Columbia Generations Project data set 2, BCGP-DS3 = British Columbia Generations Project data set 3, CaG-DS1 = CARTaGENE data set 1, CaG-DS2 = CARTaGENE data set 2, OHS-DS1 = Ontario Health Study data set 1, OHS-DS2 = Ontario Health Study data set 2, OHS-DS3 = Ontario Health Study data set 3.

ensure that data were harmonized only when questions were considered compatible. Therefore, several DataSchema variables were created for only a restricted number of cohort-specific data sets. However, harmonization was often deemed possible even if the order of the questions or variable categories varied across questionnaires or if the data were collected using various methods (e.g., questionnaire completed at home or at a data collection centre with support from cohort staff). Variables were also harmonized without taking into account the proportion of missing data. Evaluating the specific impact of these decisions on the content of each variable requires extensive work, and the methodological approach to be undertaken will vary from one variable to another. However, it is imperative to further explore the content and quality of the harmonized data. We hope to be able to achieve refined analyses for a selected number of variables and establish close collaborations with data users to gather feedback on potential improvements required to optimize the data and online resources offered. On the basis of feedback, maturing versions of the core data and related documentation will be released. Cohort investigators also plan to improve the standardization of data collection tools and procedures to minimize the heterogeneity of additional data to be collected.

There were over 307 000 participants in the cohort. However, the subsamples of participants to be included in future statistical analyses depend on the level of completion of the

Box 1: Examples of factors to be considered, where relevant, when analyzing and interpreting the harmonized Health and Risk Factor Questionnaire variables

Cohort-specific design

- Participant sampling frames; participant recruitment methods; inclusion/exclusion criteria (e.g., prevalent cases of cancer excluded at recruitment for 2 data sets)

Cohort-specific data collection and management procedures

- Timing of data collection (e.g., year or date data were collected), location of collection (e.g., at a collection centre or at home), mode of collection (e.g., paper or online questionnaire), support (e.g., assisted by an interviewer), language, data preparation protocols applied (e.g., whether outliers are kept or deleted)

Harmonization process

- Definition of the core variable generated across studies (DataSchema)
- Harmonization status attributed (e.g., integrate data generated from questions with different orders, wording, categories, response types and skip patterns)
- Data processing applied to harmonize data and estimate quality (e.g., treatment of missing values)

questionnaire and the harmonization potential. Participant samples will thus vary according to the research questions addressed. For example, only 77 000 participants would contribute to an analysis requiring usage of information related to time per day the participant typically spends in the sun. Selection of targeted subsamples of participants could introduce bias. In addition, as reported by Dummer and colleagues,[8] the voluntary nature of recruitment for the cohort affects the external validity of the study and limits generalizability. The data set should not be considered as fully representative of the Canadian population. The cohort participants are more educated and slightly more affluent than the general population, but the prevalence of common chronic diseases and obesity in the cohort appears to be similar to national rates.[8] However, even given its limited generalizability, the cohort represents a powerful resource to support a wide range of analyses (e.g., impact of genetic, behavioural and environmental risk factors on health outcomes) and generate innovative scientific knowledge.

Finally, to fully support investigating the impacts of, and interactions between, biological, environmental and behavioural factors, longitudinal data will need to be gathered and the Health and Risk Factor Questionnaire data will need to be used in combination with complementary data sets. The harmonization process is ongoing to generate additional core data collected at baseline (data related to mental health and physical measures and biological samples) and during the first follow-up of participants. Some of these harmonized data sets are already available and documented on the data portal. Although based on the approach described in this paper, the harmonization process required to generate these additional data sets (e.g., physical measures, genotypes) varies. The challenges faced by investigators using these additional data sets can also differ from those encountered by investigators using questionnaire data.

Conclusion

The harmonization process we implemented enabled the creation of a unique data set including health and risk factors data from over 307 000 Canadians. The data content will be explored in collaboration with users, and maturing versions of the Health and Risk Factor Questionnaire data set will be regularly released to provide the national and international research community with an invaluable scientific resource.

References

1. Burton PR, Hansell AL, Fortier I, et al. Size matters: Just how big is BIG?: quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009;38:263-73.
2. Collins R. What makes UK Biobank special? *Lancet* 2012;379:1173-4.
3. Scholtens S, Smidt N, Swertz MA, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol* 2015;44:1172-80.
4. Lind L, Elmståhl S, Bergman E, et al. EpiHealth: a large population-based cohort study for investigation of gene-lifestyle interactions in the pathogenesis of common diseases. *Eur J Epidemiol* 2013;28:189-97.
5. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;5:1113-24.
6. Sonnega A, Faul JD, Ofstedal MB, et al. Cohort profile: the Health and Retirement Study (HRS). *Int J Epidemiol* 2014;43:576-85.
7. Borugian MJ, Robson P, Fortier I, et al. The Canadian Partnership for Tomorrow Project: building a pan-Canadian research platform for disease prevention. *CMAJ* 2010;182:1197-201.
8. Dummer TJB, Awadalla P, Boileau C, et al.; CPTP Regional Cohort Consortium. The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. *CMAJ* 2018; 190:E710-7.
9. Fortier I, Raina P, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol* 2017;46:103-5.
10. Rolland B, Reid S, Stelling D, et al. Toward rigorous data harmonization in cancer epidemiology research: one approach. *Am J Epidemiol* 2015;182:1033-8.
11. Dhalla A, McDonald TE, Gallagher RP, et al. Cohort profile: the British Columbia Generations Project (BCGP). *Int J Epidemiol* 2018 Aug. 28 [Epub ahead of print]. doi: 10.1093/ije/dyy160.
12. Robson PJ, Solbak NM, Haig TR, et al. Design, methods and demographics from phase I of Alberta's Tomorrow Project cohort: a prospective cohort profile. *CMAJ Open* 2016;4:E515-27.
13. Ontario Health Study (home). Available: www.ontariohealthstudy.ca/ (accessed 2018 Oct. 1).
14. Awadalla P, Boileau C, Payette Y, et al.; CARTaGENE Project. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int J Epidemiol* 2013;42:1285-99.
15. Sweeney E, Cui Y, DeClercq V, et al. Cohort profile: the Atlantic Partnership for Tomorrow's Health (Atlantic PATH) study. *Int J Epidemiol* 2017;46:1762-1763i.
16. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol* 2011;174:253-60.
17. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;39:1383-93.
18. Fortier I, Doiron D, Little J, et al.; International Harmonization Initiative. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. *Int J Epidemiol* 2011;40:1314-28.
19. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc* 2003; 35:1381-95.
20. ISCO-88. Geneva: International Labour Organization; updated 2004 Sept. 18. Available: www.ilo.org/public/english/bureau/stat/isco/isco88/index.htm (accessed 2018 Oct. 1).
21. Tremblay M, Wolfson M, Connor Gorber S. Canadian Health Measures Survey: rationale, background and overview. *Health Rep* 2007;18(Suppl):7-20.
22. Doiron D, Marcon Y, Fortier I, et al. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. *Int J Epidemiol* 2017;46:1372-8.
23. Bergeron J, Doiron D, Marcon Y, et al. Fostering population-based cohort data discovery: the Maelstrom Research cataloguing toolkit. *PLoS One* 2018;13:e0200926.

24. Canadian Partnership for Tomorrow Project (CPTP) Portal (home). Toronto. Available: https://portal.partnershipfortomorrow.ca/ (accessed 2018 Oct. 1).
25. Hallal PC, Victora CG, Wells JCK, et al. Comparison of short and full-length international physical activity questionnaires. *J Phys Act Health* 2004;1:227-34. doi: 10.1123/jpah.1.3.227.

Affiliations: Research Institute of the McGill University Health Centre (Fortier, Dragieva, Saliba); Centre hospitalier de l'Université de Montréal (CHUM) Research Centre (Craig), Montréal, Que.; CancerControl Alberta and Cancer Strategic Clinical Network (Robson), Alberta Health Services; Department of Agricultural, Food and Nutritional Science (Robson), Faculty of Agricultural, Life and Environmental Sciences, University of Alberta, Edmonton, Alta.

Contributors: Isabel Fortier was responsible for developing the concept of the paper and ensuring the validity of its contents. Isabel Fortier, Nataliya Dragieva, Matilda Saliba and Camille Craig drafted the paper. Paula Robson contributed to the development of the concept and revised the manuscript for important intellectual content. All authors approved the final version to be published and agree to be accountable for all aspects of the work.

Canadian Partnership for Tomorrow Project scientific directors: Philip Awadalla (Ontario Institute for Cancer Research, Toronto, Ont.), Trevor J.B. Dummer (Centre of Excellence in Cancer Prevention, School of Population and Public Health, Faculty of Medicine, University of British Columbia, Vancouver, BC), Jason M.T. Hicks (Dalhousie University, Halifax, NS), Sébastien Jacquemont (Department of Pediatrics, Faculty of Medicine, University of Montreal; Sainte-Justine University Health Centre, Montréal, Que.), Bartha Maria Knoppers (Centre of Genomics and Policy, McGill University, Montréal, Que.), Nhu Le (British Columbia Cancer Research Centre, Vancouver, BC), Anne-Marie Mes-Masson (Institut du cancer de Montréal, Centre de recherche du CHUM; Department of Medicine, Faculty of Medicine, University of Montreal, Montréal, Que.), John McLaughlin (Public Health Ontario, Toronto, Ont.), Céline Moore (Canadian Partnership for Tomorrow Project; Canadian Partnership Against Cancer, Toronto, Ont.), Anne-Monique Nuyt (Department of Pediatrics, Faculty of Medicine, University of Montreal; Sainte-Justine University Health Centre, Montréal, Que.), Louise Parker (Dalhousie University, Halifax, NS), John J. Spinelli (BC Cancer Agency Research; School of Population and Public Health, Faculty of Medicine, University of British Columbia, Vancouver, BC), Jennifer Vena (CancerControl Alberta, Alberta Health Services, Edmonton, Alta.).

Harmonization Standing Committee: Julie Bergeron (Research Institute of McGill University Health Centre, Montréal, Que.), Tanya Flanagan (Canadian Partnership for Tomorrow Project; Canadian Partnership Against Cancer, Toronto, Ont.), Chenwei Gao (Ontario Institute for Cancer Research, Toronto, Ont.), Calvin Lai (British Columbia Cancer Research Centre, Vancouver, BC), Tharsiya Martin (Ontario Institute for Cancer Research, Toronto, Ont.), Kelly McDonald (Ontario Institute for Cancer Research, Toronto, Ont.), Treena McDonald (British Columbia Cancer Research Centre, Vancouver, BC), Anouar Nechba (Research Institute of McGill University Health Centre, Montréal, Que.), Yves Payette (Sainte-Justine University Health Centre, Montréal, Que.), Will Rosner (CancerControl Alberta, Alberta Health Services, Edmonton, Alta.).

Funding: Funding and in-kind support for the Canadian Partnership for Tomorrow Project was provided by the Canadian Partnership Against Cancer, Health Canada, the Ontario Ministry for Research and Innovation, the Ontario Institute for Cancer Research, Alberta Health Services, Alberta Health and the Alberta Cancer Prevention Legacy Fund, the Alberta Cancer Foundation, Genome Quebec, Genome Canada and the BC Cancer Foundation. Harmonization activities were also funded by the Research Institute of the McGill University Health Centre and the Montreal General Hospital Foundation.

Acknowledgements: The authors thank the participants in the Canadian Partnership for Tomorrow Project across the 5 regional cohorts for their participation and commitment. The authors also thank the many people who work to maintain the cohort, including the staff of the National Coordination Centre (Canadian Partnership Against Cancer); the Maelstrom Research team; and the staff of the BC Generations Project, Alberta's Tomorrow Project, the Ontario Health Study, CARTaGENE and the Atlantic Partnership for Tomorrow's Health.

Supplemental information: For reviewer comments and the original submission of this manuscript, please see www.cmajopen.ca/content/7/2/E272/suppl/DC1.