Article details: 2018-0062	
Title	A retrospective harmonization effort to generate the Health and Risk Factor Questionnaire data of the Canadian Partnership for Tomorrow Project cohort study
Authors	Isabel Fortier PhD, Nataliya Dragieva MS, Matilda Saliba PhD, Camille Craig MSc, Paula J. Robson PhD; for the the Canadian Partnership for Tomorrow Project Scientific Directors and Principal Investigators and of the Harmonization Standing Committee
Reviewer 1	Ilona Csizmadi
Institution	Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, Calif.
General comments (author response in bold)	This manuscript describes the process involved in the harmonization and integration of variables pertaining to sociodemographic characteristics, lifestyle behaviors, and physical measures and risk factors across five cohorts that have been brought together in the Canadian Partnership for Tomorrow Project (CPTP). The manuscript will be an important resource for investigators planning to use CPTP data. Ideally it will also encourage use of the data amongst researchers unfamiliar with CPTP. In its current state, however, the manuscript provides little insight into the rigor, precision, and the conceptual framework of the work. Since the manuscript will be widely read, it would be important to provide more details on the process of harmonization and the final product to optimize its value as a resource to researchers. **Thank you for the comments, we hope the new version of the paper address these concerns.** Areas requiring clarification and/or additional details are itemized below: 1. Abstract: The interpretation should provide insight into the potential value of the 'harmonized dataset' as stand-alone data. **The text was modified.** 2. Related to the comment above, do the harmonized variables represent a comprehensive dataset requiring little reliance on supplementation from data in the individual cohorts – presumably a requirement for maximum study power as stated on page 8 under Interpretation: '''It optimizes the impact of each of the individual cohorts by allowing, for the subsample of harmonized data, to obtain the very large numbers of participants needed to generate sufficient statistical power to address complex questions and to facilitate more refined subgroup analysis.' Based on the knowledge of the harmonized variables can the authors elaborate on the types of studies for which this is true, e.g., common exposures such as smoking, alcohol intake and cancer outcomes? Alternatively, what are the limitations of the approach, e.g., types of studies which cannot be addressed by the current harmonized data and what might be done in the future to address the limitation. **The text was modified to address this comment and a limitations section was added.** 3. What distinguishes the 12 datasets from each other – how were these generated? One would expect a dataset from each cohort but OHS and BCGP have three and the other cohorts have two. Would it be important for investigators to be aware of the differences? While DS11 and DS12 are different according to descriptive variables listed in Table 3 there may be more subtle differences that should be described to readers. **Additional information was added to the text and as supplementary material.** 4. How were the 'core variables' and the minimum variable details decided on? Was the consensus guided by a priori defined hypothetical research questions, e.g., minimum details required for research related to common cancers and their risk factors. Or, was the consensus guided by minimizing missing data – to optimize sample size? **Information has been added to the paper.** 5. Did the harmonization process consider the chronological timing and/or temporal order of the data collection? **The years of data collection and specific dates of questionnaire completion for each participant were documented and are available for data users. However, the specific timing of collection was not taken into account in the evaluation of the harmonization potential.** 6. Page 5, lines 10 and 11: 'Variables selected are … needed to allow the generation of 'valid and relevant information'? Valid based on what criteria? Relevant to what? Was this achieved for all variables in the CPTP DataSchema? This might be more easily understood if an example was taken through each step of the harmonization process – perhaps in a figure that illustrates each key step. **An example was added, and the text was modified.** 7. Is there a weblink to CPTP DataSchema (Version 2.0 – Oct. 2017). Is this equivalent to a data dictionary? Perhaps the DataSchema page for a variable that is suggested for illustration in Item 6 could be shown. **The full list of variables and harmonization potential was added as supplementary material. Information can as well be obtained on the data portal and links were added to the paper.** 8. The use of the Lambda statistic is confusing. Do the values in Table 4 under the column 'Lambda statistic' -represent PREs as described on page 5, lines 9 to 15? This doesn't seem particularly informative and doesn't add much to the manuscript. Perhaps more complex examples could be used to justify the analysis or elaborate on why the authors feel this is informative. **The section was deleted** 9. A paper by Rolland et al., Am J Epidemiol 2015;182:1033-1038, describing the importance of understanding the process of harmonization could be referenced. **Reference to the paper was added.** 10. The STROBE Vet statement checklist is appended to the manuscript but not completed by the authors. **We have not been able to find any appropriate reporting guidelines for descriptive analysis of retrospective harmonization procedures. We thus used the check list provided in the Maelstrom Research guidelines for rigorous retrospective data harmonization (Fortier et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol. 2016 Jun 6).** Tables and figures: 1. Figure 1. Is an awkward presentation of data --> histogram might be better? **The figure was deleted, and the information is now provided in other formats.** 2. Please expand on Items within the broad categories in Table 2, e.g., Sociodemographic and economic characteristics, screening, lifestyle factors. **Information was added.**

	Minor edits/typos: 1. Page 6, Line 10: End of sentence [ref] – insert reference. 2. Table 3, Non-smoker comma misplaced, 25,3041 -->253,041; please add total Ns under each sub-cohort and dataset. **The text was modified.**
Reviewer 2	Darren Brenner
Institution	Cancer Epidemiology and Prevention Research, Alberta Health Services, Calgary, Alta.
General comments (author response in bold)	This descriptive paper provides an overview of the processes developed to document, harmonize and integrate data across the regional cohorts within the Canadian Partnership for Tomorrow Project (CPTP). The manuscript also presents the specific attributes of the data generated from health and risk factor questionnaires (HRFQ). The paper is an important piece of work as the number of publications and inquiries from the CPTP cohort continues to expand over the next number of years. A few recommendations for clarification are included. **Thank you for the comments, we hope the new version of the paper in combination the CPTP data portal will provide enough information to the users.** Minor points Introduction 1. The authors state that "this integrative approach can be particularly valuable for leveraging innovative research" How is this approach more valuable or feasible for innovative research than a single design large cohort? Consider revising this statement or clarifying the meaning of this statement. **The text was modified.** 2. The authors state that "Therefore, CPTP investigators combined prospective and retrospective harmonization approaches to generate shared data across their member cohorts" **Information was added to the paper to clarify the approach.** Methods 3. "These resources were used to develop the regional data collection tools but were adjusted to comply with study-specific designs, constraints and infrastructures, leading to the generation of similar, but not identical study-specific collections tools and datasets." Can the authors either provide additional metrics on the census approach? Or a reference to support these choices? **Additional information was included in the text. The reasons for modifying the core questionnaire were various, but mainly related to the need to adapt the questionnaire to the tools and procedure locally used to collect data. Even if CPTP can be considered as a harmonized study, it is important to highlight that each participant cohort was independent, and the sampling frame, recruitment procedures, etc. were different across cohorts. The role of the harmonization team (and objective of this paper) was only to explore potential to generate the core variables and document procedures applied, to offer users information to understand and properly use the harmonized data generated.** 4. The authors state that "inferentially equivalent" information was sought across studies. What metric or parameters were used to define this concept. **The wording used was not appropriate, the sentence was modified.** 5. "Data accuracy was assessed through checking the logical reasoning of the responses provided - Is that actually a measure of accuracy?" Participants can provide responses that would appear reasonable, but not necessarily accurate – consider revising. **The text was revised to clarify this issue.** 6. Additional details on QC – "based on pre-defined rules developed in collaboration with regional cohort investigators". Can the team cite where to find these rules, presumably they can be found online? **The text was modified to clarify.** 7. Data consistency was assessed by exploring variabilities in data distributions across datasets. When possible, the reasons explaining inconsistencies were documented (e.g. exclusion of participants with a cancer history at recruitment resulting in lower cancer rates for two harmonized datasets). This implies an expected range across datasets - What is the expected range across datasets? **The text was modified to clarify.** 9. Missing reference [ref] page 6 line 10 **The reference was added.** 10. Describe what teleform is in text or as a footnote in tables as a reader might not be aware. **The text was modified to clarify.** Results 11. The results state that there are 0-20% problems with responses across variables? How many had closer to 20%? Was an upper threshold considered before a question would be dropped from the CPTP dataset? **The text was modified to simplify the reading, but additional information is available on the portal or can be provided to granted users.** 12. Inclusion/eligibility criteria for the cohorts should be included in a table or supplementary table which would also provide context to the observed disparities in responses or missingness across studies. Related to this – a brief description or citation for the differences in the within study datasets would be helpful. **Text was modified and a table added to clarify.**
Reviewer 3	Ruth Hall
Institution	Institute for Clinical Evaluative Sciences, Toronto, Ont.
General comments	This is an important paper to have in the literature for researchers to cite when using the CPTP data. The paper also has relevance to future work related to database harmonization. The authors provide a clear description of the methods used to

(author response in bold)	create a high quality harmonized dataset. The authors may want to consider providing a bit more information on the Maelstrom guidelines used to guide this work so the reader does not need to refer to the referenced article. Page 5, second paragraph is missing the reference number. The inclusion of examples was helpful. Is it worth including a comment a low Lambda value is desirable? Is there a website url available to within the paragraph on page 7? **Thank you for the comments, the text was modified to clarify the methods and procedures used. Examples and link to the website were also added. The reference was added.**