User Guide for the Crowdsourcing: Impacts of the COVID-19 on Canadians – Your Mental Health

Public Use Microdata File (PUMF)

June 2020

1. Introduction

The data collection series Crowdsourcing: Impacts of COVID-19 on Canadians is designed to assess the quality and viability of a more timely collection model using willing participants and web-only collection. The Crowdsourcing: Impacts of COVID-19 on Canadians – Your Mental Health is the second iteration in the continuing series of crowdsourcing cycles. The overall goal of the crowdsourcing initiative is to invite all members of the Canadian population to participate in a data collection exercise on a voluntary basis. The main topic of this second crowdsourcing was to determine how Canadians are reacting to the COVID-19 crisis and the impact it has had on their mental health.

In the context of this product, the term *crowdsourcing* refers to the process of collecting information via an online questionnaire. Open advertising was used to obtain participants who chose to self-select by completing the questionnaire. As such, the crowdsourcing data was collected through a completely non-probabilistic approach which does not involve a random selection of respondents like other traditional Statistics Canada surveys. Therefore, results pertain only to the participants and cannot be used to draw conclusions about the larger population of Canadians.

The following sections of the document provide a summary of different methodological considerations relevant to this crowdsourcing as well as information on how the Public-Use Microdata File (PUMF) was created, how it should be used, and what its limitations are.

2. Confidentiality

Statistics Canada is prohibited by law from releasing any information it collects that could identify any person, business, or organization, unless consent has been given by the respondent or as permitted by the Statistics Act. Various confidentiality rules are applied to all data that are released or published to prevent the publication or disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data.

The approach for creating this PUMF is to balance the requirements for maintaining the confidentiality of participants by minimizing disclosure risks, while providing the most useful data to users. The production of a PUMF includes many safeguards to prevent the identification of any one person. Confidentiality of the participants to the crowdsource is ensured mainly by the reduction of information. Some variables that were collected (ex: postal code) do not appear on the PUMF due to suppressions (see Appendix A) while response categories have been restricted or collapsed for other variables in order to reduce disclosure risk (see Appendix B).

3. Methodology

3.1 Collection period

Collection for this crowdsourcing started on April 24, 2020 and closed approximately two and a half weeks later on May 11.

3.2 Target population

The target population for the Crowdsourcing: Impacts of COVID-19 on Canadians was all Canadians aged 15 and up, living in one of the ten provinces or three territories during the collection period.

3.3 Data sources and collection tool

Participation in this crowdsourcing initiative was voluntary. Prompts to participate were done through social media as well as a variety of outside partners like other government agencies, private and public organizations, associations, and news channels. Data were collected directly from participants via a self-administered online questionnaire found on an anonymous portal on Statistics Canada's website (i.e. the crowdsourcing application). Data collection was available in English and in French and the questionnaire took approximately five minutes to complete. The questionnaire followed standard practices and wording used in a computer-assisted interviewing environment, such as the automatic control of flows that depend upon answers to earlier questions and the use of edits to check for logical inconsistencies and capture errors. The computer application for data collection was tested extensively.

3.4 Verifications

The following validation rules were implemented during processing of the data to comply to the target population of the crowdsourcing.

- If the age of the participant was missing or less than 15 years old, the record was set as out of scope.
- If the postal code was missing, started with an invalid letter (not assigned to a Canadian province or territory), did not have a number in the second character, had the following first three digits (A9A, H0H, N1N, X0X, X1X), or was A0A 0A0 or N0N 0N0, the record was set to out of scope.
- For this crowdsourcing initiative, a three category gender variable was derived. Write-in responses to the gender questions were coded to "Male", "Female", or "Gender diverse" if the participant provided a valid response in the "Gender-specify" category. Participants with a missing or an invalid response provided in the "Gender-specify" category were considered out-of-scope and were removed from the file.
- While the three category gender variable is retained on the analytical master file, a derived binary variable for gender was also created for inclusion on the PUMF. This was done in order to match available control totals for benchmarking, as well as to safeguard the confidentiality of participants due to the small number of "Gender-specify" responses.
- This binary gender variable was derived by randomly assigning responses of "Gender diverse" to either "Male" or "Female".

No imputation was performed and questions that were left unanswered by participants should be excluded from analysis. For each question, only a very small percentage of participants (less than 1%) did not provide an answer.

3.5 Sample design

Crowdsourcing is a non-probabilistic approach to collecting data which does not use a sample design. Unlike probability-based surveys which select a sample of units using a controlled random mechanism, crowdsourcing participants provide their information on a voluntary basis. They are not sampled with a known probability of selection and therefore, a survey weight cannot be calculated.

3.6 Coverage

In the absence of a sample design, the coverage of the population cannot be measured. It is possible that some participants who provided their data were not actually part of the target population. It is also possible that some participants completed more than one questionnaire. In the crowdsourcing context, it is acceptable to have individuals participate more than once during the collection period. This could be the case if their opinion has changed for example. Verifications were performed to detect an abnormally large number of responses from one person and none were found.

Canadians not in tune with the various channels used to advertise the crowdsourcing as well as individuals with lower propensity to participate in surveys and data collection exercises are not well represented by the collected data. This translates into over/underrepresentation for some groups of the population with respect to some of the demographic information that was collected. For example, males, youth (less than 25 years old), seniors (65 years and older), and people from Quebec and Alberta were underrepresented. On the other hand, women, young (25 to 34 years old) and middle-aged (35 to 44 years old) adults, as well as people from Nova Scotia, Ontario, and Yukon were overrepresented. To generalize results observed from the participants to the population, one would have to make the assumption that participants to the crowdsourcing are representative of those not participating, which is an assumption that cannot be validated in practice.

3.7 Benchmarking

Benchmarking the crowdsourced data to known population totals can, under certain circumstances, partly reduce some of the self-selection and coverage bias but cannot adequately correct for it completely. To compensate for the over/underrepresentation of the participants, benchmarking was done in a similar way as calibration or post-stratification is done in a probability-based survey. In this case, benchmarking helped to correct for differing participation rates across province/territory, sex, and age group.

Demographic projections of the number of people by province, sex, and age groups as of February 2020 were used to calculate a benchmarking factor for every participant. The initial goal was to benchmark by province and sex for the following age groups: 15-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+. However, some collapsing of those age groups had to be done because of the small number of participants in some of them. The final collapsing strategy used to produce the benchmarking factors was as follows:

- No collapsing needed in Nova Scotia, Quebec and Ontario for both sexes.
- No collapsing needed for women in Newfoundland and Labrador, New Brunswick, Manitoba, Saskatchewan, Alberta, and British Columbia.

- Collapsing of men in Newfoundland and Labrador, New Brunswick, Manitoba, Saskatchewan, Alberta, British Columbia, and Yukon as well as women in Prince Edward Island and Yukon to the following three age groups: 15-34, 35-49, 50+.
- Collapsing men of Prince Edward Island to the following two age groups: 15-44, 45+.
- Collapsing all age groups for men and women of the Northwest Territories.
- Collapsing all participants of Nunavut into one group, irrespective of age and sex.

The benchmarking factors were standardized in a way that their sum totals up to the number of participants instead of the population size in each benchmarking group.

4. Guidelines to analysis

4.1 Benchmarking factors

The microdata on the PUMF are unweighted. It is the responsibility of data users to apply the standardized benchmarking factors in any results they wish to produce. Benchmarking factors for this crowdsourced data were calculated to correct for differing participation rates across province/territory, sex, and age group. If these factors are not used, the results derived from the microdata will not correspond to those that would be produced by Statistics Canada.

Standardized benchmarking factors should be used to produce results in the same way weights are used to produce estimates from a probabilistic survey. **However, because of the non-probabilistic nature of crowdsourcing and the calculation of a standardized benchmarking factor, results should be limited to proportions only and data from the crowdsourcing should not be used to calculate totals**. Furthermore, results cannot be used to draw conclusions about the Canadian population.

4.2 Data quality indicators

Given the non-probabilistic nature of the crowdsourcing data collection and the absence of a sample design, a probability of selection and a survey weight are not available for this product. Even though a standardized benchmarking factor has been calculated, it should not be used to calculate measures of precision commonly associated with probabilistic surveys (e.g. coefficients of variation, margins of error, confidence intervals) as the results would not be valid.

4.3 Rounding guidelines

It is strongly recommended that users adhere to the following guidelines regarding the rounding of results produced from crowdsourced data.

a) Proportions and ratios are to be computed from unrounded components (i.e. numerators and/or denominators) and then are to be rounded themselves to three decimals using normal rounding. In normal rounding to a single digit, if the final or only digit to be dropped is 0 to 4 the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is increased by 1.
b) Rounding of totals and differences in aggregates does not apply as crowdsource data should only be used to calculate proportions.
c) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used, with the result that results to be published or otherwise released differ from

corresponding results published by Statistics Canada, users are strongly advised to indicate the reasons for the differences in the documents to be published or released.

d) Under no circumstances are unrounded results to be published or otherwise released by users. Unrounded results imply greater precision than actually exists.

4.4 Minimum sizes

For reliability purposes in the calculation of proportions, it is recommended that the numerator should have at least 10 participants displaying the characteristic of interest while the denominator should have at least 30 participants in the domain of interest. Given the non-probabilistic nature of crowdsourcing, the user should not publish proportions of 0% or 100%.

4.5 Other considerations

a) Given the non-probabilistic nature of crowdsourcing, the user is reminded that the data cannot be used to draw conclusions for the Canadian population. The smaller the domain of interest is, the more likely the results are to be biased and not representative. The user is advised to proceed with extreme caution.

b) Given the rapidly evolving situation with regards to the COVID-19 pandemic, appropriate context should be provided about the data collection period as well as the nature and extent of restrictions in place at that time when reporting results.

c) The content of the questionnaire asked participants to reflect how they felt at the time of collection. This should be made clear when reporting results.

Appendix A

The following is the list of variables removed from the file in the creation of the PUMF.

Variables removed

Variable	Description
DEM_05	Age of participant
DEM_15	Postal code
AGEGR_5	Age group in increments of 5
DEM1_35	Landing year for immigrant
DEM1_30A	Born in Canada
DEM1_30B	Canadian citizen
DEM1_30C	Landed immigrant or permanent resident
VISMIN	Visible minority group
IIDENT	Indigenous identity group
IMMYR	Years since immigration
CMA	Census metropolitan area
CT	Census tract

Appendix B

Recoded or capped variables

Variable Description	Recoding done
Province of residence (original: Prov_C) (recoded: PProv)	Territories (prov_c in (60, 61, 62)) were grouped together under PProv=63
Rural/Urban indicator (original: RURURB) (recoded: PRURURB)	All cases from Territories were recoded to PRURURB=9
Immigrant Status (original: IMMST) (recoded: PIMMST)	Immigrant and non-permanent resident were grouped together
might lose main job or main self-employment income (original: LM_30) (recoded: PLM_30)	Group categories 6 and 7 into one: 6 – Have lost my job or business within the last 4 weeks 7 – I did not work at a job or business in the last 4 weeks

Variables created for the PUMF

Variable	Description	Value
PCSizMiz	Community Size and Metropolitan Influence Zones	1 – 1,500,000 + 2 – 500,000 – 1,499,999 3 – 100,000 – 499,999 4 – 10,000 – 99,999 5 – Non-CMACA 9 – Unknown