Datasheet for the 2019-2020 Canadian Community Health Survey (CCHS)*

Comprehensive Overview of Dataset Characteristics and Applications

Jiaxuan Song

December 2, 2024

The 2019-2020 Canadian Community Health Survey (CCHS) is a cross-sectional survey conducted by Statistics Canada to provide insights into the health status, health behaviors, and healthcare utilization of Canadians. This datasheet outlines the dataset's characteristics, including its probabilistic sampling design, demographic coverage, and preprocessing methods. The dataset is a vital resource for public health research, supporting tasks like trend analysis, health policy evaluation, and the study of socio-economic health disparities. It adheres to strict confidentiality and ethical guidelines, ensuring the responsible use of data for evidence-based decision-making.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The 2019-2020 CCHS dataset was created to provide reliable and timely information about the health status, health behaviors, and health care use of Canadians. Its purpose was to support public health research, inform policy-making, and address gaps in data on health determinants such as mental health, physical activity, and nutrition, while offering insights into health disparities across different regions and populations in Canada.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

*Code and data are available at:

- The 2019-2020 CCHS dataset was created by Statistics Canada, a federal government agency responsible for producing statistics to help better understand Canada's population, economy, and society. It was conducted on behalf of federal, provincial, and territorial health departments to support public health research and policy development.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The 2019-2020 CCHS dataset was funded by the Government of Canada through Statistics Canada, with additional financial support from federal, provincial, and territorial health departments. There is no specific grant or grant number associated with its creation, as it was part of the regular activities of Statistics Canada to support public health monitoring and research.

4. *Any other comments?*

    - The 2019-2020 CCHS is a vital resource for understanding health trends and disparities in Canada. Its comprehensive design and collaborative funding ensure its utility for policymakers, researchers, and public health officials, making it a cornerstone for evidence-based health decisions.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The instances in the 2019-2020 CCHS dataset represent individual respondents, each providing self-reported data on various aspects of their health, lifestyle, and demographics. The dataset comprises one primary type of instance: people, with each record corresponding to a unique individual. These instances capture a wide range of health-related variables, including physical and mental health status, healthcare utilization, and social determinants of health, providing a comprehensive view of the health behaviors and outcomes of Canadians.

2. *How many instances are there in total (of each type, if appropriate)?*

    - The 2019-2020 Canadian Community Health Survey (CCHS) collected data from approximately 110,000 respondents over the two-year period.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please*

describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

- The 2019-2020 CCHS dataset is a sample from the larger Canadian population aged 12 and older living in private dwellings across all provinces and territories. It excludes individuals living on reserves, in institutions, or in certain remote areas, as well as full-time members of the Canadian Forces. The sample was designed to be representative of the Canadian population through a multistage stratified sampling method. Geographic and demographic representativeness was ensured by using weighted adjustments to account for survey design, non-response, and population estimates. Validation involved comparing survey weights and outcomes to external benchmarks from the Census and other data sources. While the survey is highly representative for its intended coverage, it does not include certain populations, such as those living on reserves or in institutions, which limits its generalizability to those groups.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance in the 2019-2020 CCHS dataset consists of structured data representing responses to survey questions. These include:

Demographics: Age, gender, marital status, education level, and income.

Health Status: Physical and mental health conditions, self-reported health, and chronic diseases.

Health Behaviors: Smoking, alcohol consumption, physical activity, and dietary habits.

Healthcare Utilization: Access to and use of healthcare services, such as visits to doctors and specialists.

Social Determinants: Factors like employment status, housing, and social support.

The data is processed into features, as it has been cleaned and coded to standardize responses, such as categorizing age groups or scaling survey weights. While it is not "raw," the dataset preserves detailed and granular information suitable for analysis.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - The 2019-2020 CCHS dataset does not have a single predefined label or target associated with each instance, as it is designed for multiple purposes rather than a specific predictive task. However, researchers can define labels or targets based on the dataset's variables depending on their analysis goals. For example:

- **Health Status Target:** Self-reported health ratings (e.g., excellent, good, fair, poor) can be used as a target for studying health determinants.
- **Behavioral Targets:** Variables like smoking status or physical activity levels can serve as targets in behavioral health studies.
- **Healthcare Access Targets:** Frequency of healthcare visits can be used to analyze healthcare utilization patterns.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Yes, some information may be missing from individual instances in the 2019-2020 CCHS dataset due to the following reasons:

Non-Response: Respondents may choose not to answer certain questions due to privacy concerns, discomfort, or lack of knowledge about the question.

Survey Skip Patterns: Certain questions are only asked if a respondent answers previous questions in a way that triggers follow-up. For example, questions about specific health conditions may only appear if a respondent reports having such conditions.

Inaccessibility: Some data may be missing for individuals living in remote areas or in populations excluded from the survey, such as institutionalized individuals or those living on reserves.

These gaps are typically documented in the survey methodology, and statistical techniques like imputation or weighting adjustments are used to mitigate their impact on representativeness.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - No, relationships between individual instances in the 2019-2020 CCHS dataset are not made explicit. Each instance represents a single respondent's data and is treated independently. The dataset does not include information about interactions or relationships between respondents, such as social connections or shared environments. Any relationships or correlations must be inferred through aggregate analysis of shared characteristics or patterns across instances (e.g., similar health behaviors within geographic regions).

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- The 2019-2020 CCHS dataset does not come with predefined data splits, as it is designed primarily for descriptive and exploratory analysis rather than predictive modeling. However, users can create their own splits based on the objectives of their analysis. For example:

- **Geographic Splits:** Divide the data by provinces or territories to analyze regional health trends.
- **Temporal Splits:** Use data from specific periods (e.g., 2019 vs. 2020) to assess changes over time.
- **Population Splits:** Separate data by demographics (e.g., age groups, gender, or income levels) for subgroup analyses.
- **Modeling Splits:** If used for predictive modeling, the dataset can be split into training, validation, and testing sets using random sampling or stratified sampling to ensure representativeness.

The rationale for these splits should align with the research or modeling goals, ensuring that the splits maintain the representativeness and integrity of the dataset.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Yes, the 2019-2020 CCHS dataset may contain some sources of noise or potential errors, common in survey data, such as:

Self-Reporting Bias: Responses rely on participants' self-assessment, which can lead to inaccuracies due to memory errors, misunderstanding of questions, or social desirability bias.

Measurement Errors: Differences in how questions are interpreted or recorded by respondents can introduce variability that does not reflect true differences in the population.

Non-Response Bias: Missing data from respondents who decline to answer certain questions or drop out of the survey can lead to bias if their characteristics differ systematically from those who respond.

Survey Weighting Variability: While weights are applied to adjust for representativeness, small errors or assumptions in weight calculations can introduce slight inaccuracies in population estimates.

Redundancies: Some questions or variables may overlap or measure closely related constructs (e.g., similar questions about health status), potentially leading to multicollinearity in analyses.

These issues are typically addressed during data cleaning, processing, and analysis by using imputation, statistical weighting, and sensitivity checks to minimize their impact on conclusions.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The 2019-2020 CCHS dataset is self-contained and does not rely on external resources for its core content, as all data collected and provided is included within the dataset itself. However, users may need to consult external documentation, such as the survey's user guide or metadata files available on the Statistics Canada website, to fully understand the variables, methodology, and structure. Access to the dataset is subject to restrictions and typically requires obtaining the data through Statistics Canada's Research Data Centres (RDCs) or under specific agreements, which may involve fees or licensing conditions. While there are no guarantees that the supporting documentation or access mechanisms will remain constant over time, Statistics Canada provides archival versions of the dataset and accompanying materials to ensure reproducibility. Restrictions on access and use are governed by Statistics Canada's data policies, which require proper authorization and adherence to confidentiality rules.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - Yes, the 2019-2020 CCHS dataset contains data that might be considered confidential, as it includes sensitive personal information about individuals' health, demographics, and behaviors. This data is protected under Canada's **Privacy Act** and **Statistics Act**, which ensure that individuals' privacy is safeguarded. Identifiable information, such as names or contact details, is not included in the publicly available dataset. Additionally, all data is anonymized and aggregated to prevent identification of individuals, and access to detailed microdata is restricted to authorized researchers through secure Research Data Centres (RDCs) under strict confidentiality agreements. These measures ensure compliance with legal and ethical standards for protecting respondents' confidentiality.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - The 2019-2020 CCHS dataset does not contain content intended to be offensive, insulting, or threatening. However, some data might cause anxiety or discomfort if viewed directly, as it includes information on sensitive topics such as mental

health conditions, chronic illnesses, substance use, and healthcare access disparities. These topics could evoke emotional responses, especially for individuals personally affected by such issues. To mitigate potential discomfort, access to detailed data is limited to authorized researchers, and guidelines are provided for ethical use and interpretation of the data.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes, the 2019-2020 CCHS dataset identifies several sub-populations based on demographic and health-related characteristics. These include:

- **Age:** Respondents are categorized into specific age groups (e.g., 12–17, 18–34, 35–64, 65+).
- **Gender:** The dataset collects information on respondents' self-identified gender (e.g., male, female, other, depending on survey scope).
- **Geography:** Respondents are classified by their province or territory and, in some cases, by urban or rural residence.
- **Socioeconomic Status:** Sub-populations are identified by income levels, education attainment, and employment status.
- **Health Status and Behaviors:** Groups are defined based on health conditions, behaviors (e.g., smoking status), and healthcare utilization patterns.

The dataset uses statistical weights to ensure these sub-populations are representative of the Canadian population, allowing for accurate analysis of their distributions. Exact distributions vary by variable and are detailed in the accompanying survey documentation, which provides breakdowns for demographic and regional representativeness.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, it is not possible to identify individuals directly or indirectly from the 2019-2020 CCHS dataset. The dataset is anonymized and stripped of any personally identifiable information (PII), such as names, addresses, or contact details. Additionally, measures are taken to prevent indirect identification, such as:

Data Aggregation: Responses are aggregated to broader categories (e.g., age groups, income ranges) to reduce granularity.

Suppression of Small Counts: Data points that could potentially identify individuals in small sub-populations (e.g., rare health conditions within a small geographic area) are suppressed or not included.

Strict Access Controls: Microdata access is restricted to authorized users in secure environments, such as Statistics Canada's Research Data Centres (RDCs), under strict confidentiality agreements.

These measures ensure compliance with legal and ethical standards, making it extremely unlikely to identify individuals from the dataset.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - Yes, the 2019-2020 CCHS dataset contains sensitive data, including:

Health Data: Information about physical and mental health conditions, healthcare utilization, chronic illnesses, and disabilities.

Demographics: Data on age, gender, income, and education level, which, while anonymized, could be sensitive if misinterpreted or misused.

Behavioral Data: Information about substance use (e.g., smoking, alcohol consumption), physical activity, and diet.

Social Determinants: Data on employment status, housing stability, and social support, which may touch on sensitive socio-economic factors.

Ethnic and Cultural Origins: In some cases, data may include variables related to Indigenous identity or other cultural groups.

All sensitive data is anonymized and collected under strict confidentiality protocols. Access to detailed microdata is restricted, and its use is governed by ethical guidelines to ensure responsible handling of sensitive information.

16. *Any other comments?*

    - The 2019-2020 CCHS dataset is a valuable resource for understanding the health and well-being of Canadians. Its robust design and detailed data collection make it an essential tool for public health research and policy-making. However, its use requires adherence to strict confidentiality and ethical guidelines to ensure the responsible handling of sensitive information and to respect the privacy of respondents.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data in the 2019-2020 CCHS dataset was primarily self-reported by subjects through structured interviews conducted via phone or in person. Respondents provided information on various topics, such as health status, behaviors, and demographics. To ensure data reliability, the survey employed standardized questions and trained interviewers to minimize variability and inaccuracies in responses. Additionally, quality control measures were implemented to monitor inconsistencies, and survey weights were applied to adjust for non-response and maintain the representativeness of the sample. While self-reported data may be subject to biases such as memory errors or social desirability bias, the CCHS employed rigorous methodologies to enhance data quality and validity.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data for the 2019-2020 CCHS was collected through **structured interviews**, conducted via **telephone** or **in-person surveys** by trained interviewers. Responses were manually recorded using **computer-assisted interviewing (CAI) systems**, which ensured consistency in question delivery and data entry.

To validate these mechanisms:

- **Interviewer Training:** Interviewers underwent comprehensive training to ensure uniformity and accuracy in data collection.
- **Standardized Questionnaires:** The survey utilized a pre-tested and standardized questionnaire to minimize interpretation errors and maintain consistency across respondents.
- **Quality Assurance:** Data collection procedures were monitored, and routine checks were performed to identify and address potential errors or inconsistencies during the process.

These measures ensured that the data collection mechanisms produced accurate and reliable results suitable for analysis.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- The 2019-2020 CCHS dataset is a sample from the larger Canadian population, and it employed a **probabilistic multistage stratified sampling strategy**. The sampling process involved:

Stratification: The population was divided into strata based on geographic and demographic characteristics (e.g., province, urban/rural status) to ensure coverage of diverse subpopulations.

Cluster Sampling: Within each stratum, clusters (e.g., households or dwellings) were selected as primary sampling units.

Random Selection: Individuals within the selected clusters were randomly chosen to participate, with probabilities proportional to population size to maintain representativeness.

This approach ensured that the sample was representative of the Canadian population aged 12 and older, excluding certain groups (e.g., those living on reserves, in institutions, or remote regions). The design also incorporated survey weights to correct for non-response and align the sample with population estimates.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The data collection for the 2019-2020 CCHS was conducted by **trained interviewers** employed by Statistics Canada. These interviewers were professional staff who underwent comprehensive training to ensure accurate and standardized data collection. As part of Statistics Canada's workforce, interviewers were compensated with regular salaries and benefits as per federal government employment standards. Compensation details are not publicly specified but align with the salary scales for federal employees involved in data collection roles. Respondents who participated in the survey were not financially compensated, as participation was voluntary and intended to contribute to public health research and policy development.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- The data for the 2019-2020 CCHS was collected over a two-year period, from **January 2019 to December 2020**. This timeframe aligns with the creation of the data associated with the instances, as all survey responses were gathered directly from participants during this period. The data reflects real-time conditions and self-reported information provided by respondents at the time of the survey, ensuring that the collection and creation timelines are consistent.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Yes, the 2019-2020 CCHS underwent ethical review processes in accordance with **Statistics Canada's policies** and the legal requirements outlined in the **Statistics Act**. These processes ensured that the survey adhered to strict ethical standards for data collection, including:

Privacy and Confidentiality Protections: The survey design complied with the **Privacy Act**, ensuring the confidentiality of respondents' personal information and preventing identification in published data.

Voluntary Participation: Respondents were informed about the purpose of the survey and their right to participate voluntarily, with the assurance that their responses would remain confidential.

Informed Consent: Participants provided informed consent before taking part in the survey.

Statistics Canada's procedures do not typically require external institutional review board (IRB) approval, as their surveys are internally governed by established legal and ethical frameworks.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data for the 2019-2020 CCHS was collected directly from individuals through structured interviews conducted by Statistics Canada. These interviews were carried out either by telephone or in person, ensuring that the data came directly from respondents. No third parties or external sources were involved in the data collection process.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Yes, the individuals participating in the 2019-2020 CCHS were notified about the data collection. Before participating, respondents were informed about the survey's purpose, how their data would be used, and their rights as participants, including the voluntary nature of their participation and the confidentiality of their responses. This information was typically provided verbally by interviewers during the introduction to the survey, and in some cases, written materials or consent forms were shared.The exact language of the notification can be found in Statistics Canada's

privacy and survey materials, which outline participants' rights under the **Privacy Act** and the **Statistics Act**.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Yes, the individuals participating in the 2019-2020 CCHS provided informed consent before their data was collected. Consent was requested verbally by trained interviewers at the start of the survey. Respondents were informed about the purpose of the survey, how their data would be used, the voluntary nature of their participation, and the confidentiality measures in place. By agreeing to proceed with the survey, participants provided their consent.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Yes, individuals participating in the 2019-2020 CCHS were informed that their participation was voluntary and that they could decline to answer any specific question or withdraw from the survey at any point during the data collection process. However, once the data was anonymized and integrated into the dataset, it was no longer linked to the individual, making it impossible to revoke consent retroactively for specific uses.

This approach ensures compliance with confidentiality and privacy standards while maintaining the integrity of the dataset.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Yes, Statistics Canada conducts a thorough analysis of the potential impact of its datasets, including the 2019-2020 CCHS, on data subjects. This process is part of its compliance with the **Privacy Act** and the **Statistics Act**, which include measures to assess and mitigate risks related to privacy and data protection.

Key aspects of this impact analysis include:

- **Confidentiality Safeguards:** Ensuring all data is anonymized to prevent identification of individuals.
- **Access Controls:** Restricting access to sensitive data through secure environments such as Research Data Centres (RDCs).

- **Ethical Use:** Requiring researchers to agree to confidentiality agreements and adhere to ethical guidelines for data use.
- **Transparency:** Informing respondents about data collection, use, and protections through clear communication.

The outcomes of these analyses affirm that the dataset adheres to strict privacy and security standards, minimizing risks to data subjects.

12. *Any other comments?*

    - The 2019-2020 CCHS dataset is a robust and well-protected resource for public health research, designed with privacy and ethical considerations at its core. Its adherence to strict confidentiality protocols and comprehensive data protection measures ensures that it can be used responsibly for generating insights while safeguarding the privacy and trust of respondents.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, the 2019-2020 CCHS dataset underwent preprocessing and cleaning to ensure data quality and usability. Key steps included:

Data Validation: Responses were checked for inconsistencies or errors during and after collection to ensure accuracy.

Anonymization: Personally identifiable information was removed to protect respondent confidentiality.

Weighting Adjustments: Survey weights were applied to correct for sampling design and non-response, ensuring the dataset accurately represents the Canadian population.

Categorization: Certain variables, such as age and income, were grouped into categories (e.g., age ranges, income brackets) for ease of analysis.

Handling Missing Data: Standard methods such as imputation or exclusion were used to address missing responses while minimizing bias.

Variable Recoding: Some responses were recoded or standardized for consistency, such as harmonizing yes/no answers across questions.

These steps ensured the dataset's integrity and representativeness for public health analysis.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - The "raw" data from the 2019-2020 CCHS is not publicly available. Statistics Canada processes and anonymizes the data before releasing it to ensure confidentiality and compliance with privacy laws. Researchers accessing the dataset through secure environments, such as Research Data Centres (RDCs), work with the preprocessed data, which has been cleaned and prepared for analysis.

Statistics Canada retains the raw data internally for quality control and validation purposes, but it is not accessible to the public or researchers to protect respondent confidentiality.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Statistics Canada utilized the **Social Survey Processing Environment (SSPE)**, a suite of SAS-based statistical software programs, custom applications, and manual processes, to preprocess and clean the data for the 2019-2020 CCHS. The SSPE facilitates systematic processing steps, including data receipt, cleanup, recoding, flow edits, coding, consistency checks, variable conversion, creation of derived variables, and the assembly of final dissemination files.

While the SSPE is an internal tool developed by Statistics Canada and is not publicly available, researchers can access detailed information about the data processing methodologies employed in the CCHS through official documentation provided by Statistics Canada. This documentation offers insights into the processing steps and quality assurance measures applied to the data.

4. *Any other comments?*

    - The preprocessing and cleaning of the 2019-2020 CCHS dataset through Statistics Canada's robust Social Survey Processing Environment (SSPE) ensure high-quality, reliable data for public health research. While the software itself is not publicly accessible, Statistics Canada provides comprehensive documentation detailing the processes, offering transparency into the methods and ensuring confidence in the dataset's usability and integrity.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - Yes, the 2019-2020 CCHS dataset has been widely used for various public health and policy research tasks. Examples include:

Health Trend Analysis: Researchers and policymakers have used the dataset to track trends in physical and mental health, chronic diseases, and health behaviors across Canada.

Public Health Interventions: The data has informed the development and evaluation of public health programs, such as initiatives targeting smoking cessation, physical activity, and mental health support.

Health Disparities Research: The dataset has been used to analyze health inequities across different demographic groups, including age, gender, income levels, and geographic regions.

Impact Studies: It has supported studies on the impact of socio-economic factors and healthcare access on health outcomes.

The dataset is a critical resource for evidence-based decision-making, enabling researchers and government agencies to address key health challenges in Canada.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - There is no centralized repository linking all papers or systems that use the 2019-2020 CCHS dataset. However, research utilizing this dataset is commonly found in academic journals, university repositories, and public health reports. Organizations like Statistics Canada and Health Canada, as well as research data networks, publish analyses and studies based on the CCHS. These platforms serve as key sources for accessing work related to the dataset, which is widely used for public health and policy research.

3. *What (other) tasks could the dataset be used for?*

    - The 2019-2020 CCHS dataset could be used for a variety of tasks, including:

Health Policy Evaluation: Assessing the effectiveness of public health policies and programs targeting chronic diseases, mental health, or preventive healthcare.

Socioeconomic Analysis: Exploring the relationship between social determinants of health (e.g., income, education) and health outcomes to identify areas for intervention.

Regional Health Studies: Analyzing health disparities across provinces, territories, or urban and rural populations.

Behavioral Health Research: Studying health-related behaviors, such as physical activity, smoking, or alcohol consumption, and their impact on long-term health outcomes.

Predictive Modeling: Developing models to predict health risks or healthcare utilization patterns based on demographic and behavioral factors.

Public Health Trends: Tracking changes in health status and behaviors over time or comparing data across different CCHS cycles.

Mental Health Research: Investigating mental health prevalence, barriers to access, and the effectiveness of mental health services.

This dataset provides a comprehensive foundation for addressing a wide range of public health and policy questions.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - The 2019-2020 CCHS dataset has certain characteristics that users should consider to avoid potential risks or harms. The dataset excludes specific populations, such as those living on reserves, in institutions, or in remote areas, which could lead to underrepresentation and analyses that overlook these groups. Additionally, many variables rely on self-reported data, which may introduce biases like over- or under-reporting. Proper application of survey weights is crucial to ensure representativeness, as failure to do so could result in biased outcomes. The dataset also contains sensitive health and socioeconomic data, which, if misinterpreted, could perpetuate stereotypes or stigmatization. Furthermore, the dataset reflects a specific timeframe (2019-2020) and societal context, including the impacts of the COVID-19 pandemic, so findings should be situated within this context to avoid misattributing trends. To mitigate these risks, users should acknowledge the dataset's limitations, interpret results responsibly, and apply appropriate analytical practices.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - Yes, the 2019-2020 CCHS dataset is not suitable for certain tasks due to its design and limitations:

Individual-Level Predictions: The dataset is anonymized and aggregated, making it inappropriate for tasks requiring identification or tracking of individuals, such as personalized health predictions or interventions.

Representation of Excluded Populations: The dataset excludes groups such as individuals living on reserves, in institutions, or in certain remote areas. It should not be used to make claims about these populations, as they are not represented.

Legal or Diagnostic Purposes: The data is self-reported and not verified by medical records or legal documents, so it should not be used for tasks requiring legally or clinically verified information.

Causal Inferences Without Context: While the dataset is valuable for identifying associations, it is not designed to establish causation without careful consideration of confounding variables and study design.

Stereotyping or Stigmatization: The dataset should not be used to draw conclusions that could reinforce stereotypes or stigmatize specific groups, especially when analyzing sensitive variables such as health behaviors or socioeconomic status.

Users should apply the dataset responsibly, respecting its limitations and intended purposes to ensure ethical and accurate analyses.

6. *Any other comments?*

    - The 2019-2020 CCHS dataset is a powerful resource for understanding public health trends and informing policy decisions. However, users must handle it with care, considering its design limitations, ethical implications, and the context of its data collection. Proper application of analytical methods and acknowledgment of exclusions or biases are essential to ensure the dataset is used responsibly and for its intended purposes.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, the 2019-2020 CCHS dataset is made available to third parties, such as researchers and policymakers, outside of Statistics Canada. Access to the dataset is provided through secure environments, like Research Data Centres (RDCs), under strict confidentiality agreements. Users must adhere to Statistics Canada's data access policies, which include requirements for ethical use, protection of respondent privacy, and compliance with the terms of the **Statistics Act**. Publicly available versions of the data are anonymized and aggregated to ensure confidentiality while still supporting research and analysis.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The 2019-2020 CCHS dataset is distributed through secure access channels managed by Statistics Canada. Authorized researchers can access the dataset via the following methods:

Research Data Centres (RDCs): Secure physical or virtual environments where approved researchers can analyze the microdata.

Public Use Microdata File (PUMF): A highly aggregated and anonymized version of the dataset available for download from the Statistics Canada website.

The dataset does not have a Digital Object Identifier (DOI), as it is distributed under Statistics Canada's access and usage guidelines rather than through a traditional academic repository. Researchers must apply for access through appropriate channels, ensuring adherence to confidentiality and ethical standards.

3. *When will the dataset be distributed?*

    - The 2019-2020 Canadian Community Health Survey (CCHS) data was released in stages:

- 2019 Data: Released on August 6, 2020.
- 2020 Data: Released on September 8, 2021.
- Combined 2019/2020 Data: Released on April 19, 2022.

These releases are available through Statistics Canada's official channels.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The 2019-2020 CCHS dataset is distributed under Statistics Canada's copyright and terms of use, which govern its access and usage. Users must adhere to strict guidelines, including maintaining confidentiality, using the data solely for statistical and research purposes, and avoiding any attempts to identify individuals. While the Public Use Microdata File (PUMF) is freely accessible, detailed microdata available through Research Data Centres (RDCs) may require an approved research proposal and could involve associated fees. Full terms and conditions for accessing the dataset are outlined on Statistics Canada's official website.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No, third parties have not imposed any IP-based or other restrictions on the data associated with the 2019-2020 CCHS instances. The dataset is solely governed by Statistics Canada under its copyright and terms of use. Access and usage are regulated by Statistics Canada's policies, ensuring that the data is used responsibly and ethically. Any fees or conditions associated with accessing detailed microdata are determined by Statistics Canada and are not influenced by third parties.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - The 2019-2020 Canadian Community Health Survey (CCHS) dataset is subject to **Crown copyright** under Canadian law, as it is produced by Statistics Canada, a government agency. This means that Statistics Canada retains the rights to the dataset, and its use is governed by specific terms and conditions outlined by the agency. Users must adhere to these terms, which include maintaining confidentiality, using the data solely for statistical and research purposes, and avoiding any attempts to identify individuals. While the Public Use Microdata File (PUMF) is freely accessible, accessing more detailed microdata through Research Data Centres (RDCs) may require an approved research proposal and could involve associated fees.

7. *Any other comments?*

    - The 2019-2020 CCHS dataset is a valuable resource for public health research and policy-making, governed by strict confidentiality and usage guidelines to ensure responsible use. Researchers must adhere to Statistics Canada's licensing terms and regulatory requirements, which ensure data integrity, privacy protection, and ethical application. Users should familiarize themselves with these terms and any associated conditions before accessing the dataset.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The 2019-2020 CCHS dataset is supported, hosted, and maintained by **Statistics Canada**, the national statistical agency of Canada. Statistics Canada is responsible for the dataset's storage, updates, and accessibility, ensuring its integrity and compliance with privacy and confidentiality standards. They also provide access through secure platforms like Research Data Centres (RDCs) and offer guidance and documentation to users for appropriate and effective utilization of the data.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - To contact the owner, curator, or manager of the 2019-2020 Canadian Community Health Survey (CCHS) dataset, you can reach out to Statistics Canada through the following channels:

- **Email:** infostats@statcan.gc.ca
- **Toll-Free Phone:** 1-800-263-1136

- **International Phone:** 1-514-283-8300
- **TTY (for the hearing impaired):** 1-800-363-7629
- **Fax:** 1-514-283-9350
- **Mailing Address:** Statistics Canada 150 Tunney's Pasture Driveway Ottawa, Ontario K1A 0T6

These contact details are provided by Statistics Canada for inquiries related to their datasets and services.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - As of now, there is no specific erratum issued for the 2019-2020 Canadian Community Health Survey (CCHS) dataset. However, Statistics Canada maintains a commitment to data accuracy and transparency.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - The 2019-2020 CCHS dataset is not regularly updated, as it represents a specific survey cycle completed over two years. However, if any errors or issues are identified, **Statistics Canada** may release corrected versions or errata. Updates or corrections are managed by Statistics Canada and communicated through their official channels, such as:

- **Statistics Canada Website:** Updates or revisions to the dataset are posted on the dataset's page or in related publications.
- **Public Notifications:** Major corrections are often communicated through press releases, newsletters, or official announcements.
- **Research Data Centres (RDCs):** Updates to microdata accessed through RDCs are communicated directly to approved users.

For ongoing health data collection, subsequent survey cycles (e.g., 2021, 2022) may be released as separate datasets, providing new instances rather than updates to the 2019-2020 dataset. Users should monitor Statistics Canada's website for the latest developments.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

- The 2019-2020 CCHS dataset adheres to the **Statistics Act** and **Privacy Act**, which govern data retention and privacy in Canada. While respondents were not explicitly informed of a fixed retention period, Statistics Canada retains anonymized data indefinitely for statistical and research purposes, ensuring it cannot be linked back to individuals. Personal identifiers are removed, and strict data management policies are in place to securely store and control access to the data. These practices comply with legal and ethical standards while preserving the dataset's utility for future research, with oversight provided by Statistics Canada's robust data governance framework.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the 2019-2020 CCHS dataset will continue to be hosted and maintained by **Statistics Canada** as part of their commitment to providing access to historical data for research and policy analysis. These datasets are securely stored and made available through platforms such as Research Data Centres (RDCs) and public archives.

If a version of the dataset becomes obsolete due to updates or corrections, Statistics Canada communicates this through their official website, data release bulletins, or notifications to authorized users in RDCs. Researchers are advised to regularly check the Statistics Canada website for any updates or announcements regarding dataset versions to ensure they are using the most accurate and up-to-date information.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - The 2019-2020 CCHS dataset is managed exclusively by **Statistics Canada**, and there is no mechanism for external parties to directly extend, augment, or contribute to the dataset. This restriction ensures the integrity, consistency, and confidentiality of the data.

Researchers and organizations can, however, build on the dataset by conducting analyses or creating derivative works, such as reports, models, or new datasets derived from the original data. These outputs remain independent and are not incorporated into the original dataset managed by Statistics Canada.

If external contributions or analyses are shared publicly, they are not validated or verified by Statistics Canada. Validation of such contributions is the responsibility of the contributing party or their institution. Statistics Canada does not provide a formal process for distributing

or communicating external contributions, as they do not integrate external data into the original dataset. Users are encouraged to refer to official releases for authoritative information.

8. *Any other comments?*

    - The 2019-2020 CCHS dataset is a carefully curated resource that prioritizes data integrity, confidentiality, and usability for public health research. While contributions to the dataset itself are not permitted, its design encourages independent analyses and derivative works that can inform policy and research. Users should rely on official Statistics Canada releases for validated and authoritative data while responsibly building on the dataset within the bounds of its terms of use.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.