

Datasheet for the paper of analyzing the factors that may affect the COVID-19 results in Toronto*

Qingya Li

26 April 2022

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 dataset was created to estimate the COVID-19 cases in Toronto between 2020 to 2022. It contained the demographic, geographic, and severity information for the cases.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - The dataset was created by Toronto Public Health
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 creation was funded by the Government of Canada
4. *Any other comments?*
 - The dataset was extracted from the provincial Case & Contact Management System (CCM).

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 that comprise the dataset represent the neighbourhoods in Toronto. The city of Toronto is geographically divided into distinct neighborhoods including Islington-City Centre West, Pelmo Park-Humberlea, New Toronto and so on.
2. *How many instances are there in total (of each type, if appropriate)?*
 - There are 140 instances in total.
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).*

*Code and data are available at: <https://github.com/QingyaLi/What-factors-can-affect-COVID-19-results->

- The dataset contains all COVID-19 cases in Toronto.
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 consists of 18 variables that describe the information about the cases.
 5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - Each of the instance has an ID assigned by Toronto Public Health .
 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.*
 - No, there is no missing values.
 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.*
 - There is no relationship between individual instances.
 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.*
 - There is no recommended data splits.
 9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
 - There is no errors, no sources of noise or redundancies in the dataset.
 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 dataset is self-contained.
 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.*
 - There is no considered confidential data.
 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.*
 - There is no offensive, insulting, threatening or might otherwise cause anxiety 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.*
 - Outbreak-associated variable is divided into outbreak associated and sporadic. Age group is grouped by 10, source of infection is determined including household contact, close contact, outbreaks, travel, community and no information. Classification is categorized as either confirmed or probable. Client gender is categorized as female, male, non-binary, transgender and unknown. Outcome is either fatal or resolved.
 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
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.*
- Client Gender may be considered sensitive since it included the following options: female, male, non-binary, transgender and unknown.
16. *Any other comments?*
- No

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 are extracted from the the provincial Case & Contact Management System (CCM) and refreshed weekly.
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 collected by the report from Toronto Public Health.
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 dataset is not a sample from a large set.
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)?*
 - Toronto Public Health was involved. No compensation.
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 dataset is collected between 2020-2022 and refreshed weekly. The data are extracted at 8:30 AM on the Tuesday and released on Wednesday.
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.*
 - No, there was no ethical review process.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
 - I obtained the dataset from Toronto Open Data Portal website.
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.*

- The noticed was not provided.
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 were consent. The consents were not available.
 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).*
 - A mechanism to revoke their consent in the future or for certain uses was not provided.
 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.*
 - No, it has not been conducted.
 12. *Any other comments?*
 - No

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.*
 - We used the `select` function to keep the variables that we needed, including “Outbreak Associated”, “Age Group”, “Neighbourhood Name”, “Source of Infection”, “Classification”, “Client Gender”, “Outcome”, “Ever Hospitalized”, “Ever in ICU”, “Ever Intubated”. Then we used `filter` to exclude the “ACTIVE” cases from “Outcome” because we wanted to focus on the fatal and resolved cases. Also, we used `filter` to exclude the other options in “Client Gender” except males and females. We created a new variable called “result” to distinguish these two cases with `mutate` and `as.factor` functions. 1 indicated the COVID-19 cases were resolved and 0 represented fatal cases. Moreover, we `rename` the rest variables into lower letters: “Outbreak Associated” as “outbreak associated”, “Age Group” as “age”, “neighbourhood” as “Neighbourhood Name”, “infection” as “Source of Infection”, “Classification” as “classification”, “Client Gender” as “gender”, “Outcome” as “outcome” “Ever Hospitalized” as “ever hospitalized”, “Ever in ICU” as “ever in ICU”, and “Ever Intubated” as “ever intubated”.
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 saved in inputs/data/COVID-19 cases.csv
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
 - The R Software is available at <https://www.R-project.org/>
4. *Any other comments?*
 - No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
 - No
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.*
 - No
3. *What (other) tasks could the dataset be used for?*
 - The dataset could be used for predicting the number of confirmed COVID-19 cases in Toronto.
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?*
 - NO
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
 - The dataset is not appropriately used for estimating the COVID-19 cases outside of Toronto.
6. *Any other comments?*
 - No

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.*
 - No
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
 - The dataset will be distributed by Github.
3. *When will the dataset be distributed?*
 - The dataset will be distributed on April 26, 2022.
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.*
 - MIT license.
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.*
 - None
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.*
 - No
7. *Any other comments?*

- No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
 - Qingya Li
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - Github: <https://github.com/QingyaLi?tab=repositories>, email: qingya.li@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
 - No erratum.
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)?*
 - No
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.*
 - No limitations so far.
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.*
 - No
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.*
 - there is no mechanism
8. *Any other comments?*
 - No

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.