Datasheet for 'Outbreaks in Toronto Healthcare Institutions'*

Kevin Cai

28 November 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Extract of the questions from (**gebru2021datasheets?**).

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 track and manage outbreaks of gastroenteric and respiratory infections in Toronto healthcare institutions. It addresses the need for real-time, structured data on institutional outbreaks to support public health responses and decision-making.
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.*
    - Toronto government funded the of the dataset.
4. *Any other comments?*
    - No.

Composition

*Code and data are available at: [https://github.com/kevicai/toronto-healthcare-outbreak-prediction](https://github.com/kevicai/toronto-healthcare-outbreak-prediction.

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.*

    - Each instance represents a recording of an outbreak event in a Toronto healthcare facility. The dataset includes the institution name and location, the type of outbreak, the causative agents (primary and secondary), and the start and end time of the outbreak.

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

    - There are 5387 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).*

    - The dataset contains all reported outbreaks from Toronto healthcare institutions.

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 features such as the institution name, address, outbreak setting, type of outbreak, causative agents, and the dates of outbreak onset and conclusion.

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

    - Each instance has a label that indicates whether the outbreak is active or not.

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.*

    - The dataset includes instances where the second causative agent is "None" if only one causative agent is identified.

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 are no explicit relationships between instances, as each outbreak is treated as an independent observation.

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.*

    - No.

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

    - No.

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 relies on external sources, specifically Toronto Public Health's outbreak reporting system, which is publicly accessible. The data is guaranteed to be updated weekly, and archival versions are not specified.

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.*

    - No.

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.*

    - No.

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.*

    - No.

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.

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.

- No.

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.*

    - Data was acquired through mandatory reporting by healthcare institutions to Toronto Public Health, where outbreaks are monitored and tracked.

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?*

    - Data was collected via public health reporting systems and APIs from the City of Toronto Open Data Portal, and verified through public health monitoring and oversight.

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

    - N/A.

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 was conducted by Toronto Public Health staff and healthcare workers at the outbreak institutions, but their compensation is unknown.

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.*

    - Data was collected continuously from January 2016 to the present and is updated weekly.

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
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 was obtained from Toronto Public Health and the City of Toronto Open Data Portal.
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 dataset is publicly available through the City of Toronto Open Data Portal. Specific notifications to individuals about data collection were not required as the data is aggregated and anonymized.
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.*
    - Consent was not required due to the nature of the dataset.
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).*
    - Consent was not required due to the nature 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.*
    - No.
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.*

- Yes, the data has been cleaned and preprocessed. Instances with missing or irrelevant data were removed, and categorical variables were renamed for clarity. Additionally, dates were standardized, and the duration of outbreaks was calculated based on the start and end dates.

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 is available from the City of Toronto Open Data Portal. Link to the raw dataset on the City of Toronto Open Data Portal.

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 preprocessing was done using Python and the pandas package. The code used for cleaning and analysis is available at GitHub repository.

4. *Any other comments?*

   - No.

Uses

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

   - Yes, the dataset has been used for analysis in outbreak prediction modeling, as part of research on factors influencing the duration of outbreaks in healthcare institutions.

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.*

   - The code and data used for the analysis are available at GitHub repository.

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

   - The dataset could be used for further research on outbreak prediction, healthcare preparedness, and policy recommendations. It could also be used to explore the impact of different infectious agents on outbreak duration or for developing risk models for healthcare settings.

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 dataset contains aggregated information on outbreaks and does not include any personally identifiable data, so there are minimal risks regarding unfair treatment or harms. However, consumers of the dataset should be mindful of the limitations in the dataset, such as missing values or the inability to link data to individual patients, which could affect the accuracy of certain analyses.

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

    - The dataset should not be used for any analysis that requires personally identifiable information, as the data is anonymized and aggregated. It should not be used to draw conclusions about specific individuals or make personal health-related decisions.

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.*

    - Yes, the dataset is publicly available through the City of Toronto Open Data Portal and can be accessed by third parties.

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 is available through the City of Toronto Open Data Portal as a downloadable CSV file. It does not currently have a DOI but is available via the public open data platform.

3. *When will the dataset be distributed?*

    - The dataset is refreshed weekly and distributed via the City of Toronto Open Data Portal.

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 dataset is distributed under the Open Government License - Toronto, which allows users to copy, modify, and distribute the data for any lawful purpose. Link to license details.

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-party IP-based restrictions apply to the dataset.

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?*

    - The dataset will be maintained by Toronto Public Health and updated weekly via the City of Toronto Open Data Portal.

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

    - The contact email for Toronto Public Health is edau@toronto.ca.

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

    - No.

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 dataset is updated weekly by Toronto Public Health. Consumers can track updates through the City of Toronto Open Data Portal.

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 dataset does not relate to people.

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 dataset are archived and available for reference through the City of Toronto Open Data Portal.

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 dataset is open and available for public use through the City of Toronto Open Data Portal. Individuals wishing to contribute to or augment the dataset would need to work through Toronto Public Health or the City of Toronto's data governance procedures.

8. *Any other comments?*

    - No.

1 References