Datasheet for ‘Allergy Dataset’*

Zijun Meng

Invalid Date

I created a datasheet for the allergy dataset created by Hill et. al (2016). By using R (2023), tidyverse (2019) and arrow (2024), I converted the csv file into a Parquet file. The datasheet aims to enhance data accessibility and utility for researchers exploring the epidemiology of pediatric allergies. The conversion process and datasheet creation are detailed to facilitate replication and further analysis.

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.*
    - To analyze the epidemiological characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children within a large primary care network, addressing the need for accurate disease prevalence data through provider-based diagnoses rather than participant reporting.
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 developed by David Hill, Robert Grundmeier, Gita Ram, and Jonathan Spergel from the Children’s Hospital of Philadelphia (CHOP) (Hill et al. 2016).

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.*

[*] Code and data are available at: https://github.com/FrankMengZJ/datasheet-allergy .Ther paper was reviewed by Shuyuan Zheng.

- Each instance represents a healthcare provider-diagnosed case of eczema, asthma, allergic rhinitis, or food allergy in the pediatric population covered by the CHOP care network.

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

    - The study analyzed data from two cohorts: a birth cohort of 29,662 children and a cross-sectional cohort of 333,200 children.

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).*

    - It contains all.

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 dataset likely consists of features derived from electronic health records, including patient demographic information (age, gender), diagnostic codes (indicating specific food allergies), results from allergy tests, and possibly patient-reported outcomes. This structured data facilitates rigorous analysis while ensuring patient confidentiality.

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

    - In the context of a dataset focused on food allergies, labels could include the specific allergen(s) diagnosed, severity of allergic reactions (mild, moderate, severe), or a binary indicator of the presence of an allergic condition. These labels enable the study of prevalence, patterns, and risk factors associated with pediatric food allergies.

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. Missing information could occur due to various reasons, such as incomplete patient records, the absence of specific test results, or redaction to protect patient privacy.

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.

8. *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.*

   - No

9. *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

10. *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

11. *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.*

    - The dataset does identify sub-populations, particularly by age, as it involves a study on the epidemiology of eczema, asthma, allergic rhinitis, and food allergy among children within a large primary care network.

12. *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.

13. *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, it contains health-related data, specifically healthcare provider-diagnosed conditions of eczema, asthma, allergic rhinitis, and food allergy in children.

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 associated with each instance were directly obtained from electronic medical records (EMRs) within The Children's Hospital of Philadelphia care network. This includes healthcare provider-diagnosed conditions of eczema, asthma, allergic rhinitis, and food allergy. The data were not reported by subjects nor indirectly inferred but were based on clinical diagnoses made by healthcare providers, ensuring accuracy and reliability. Validation/verification of the data was implied through the use of International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes and a manual chart review process to estimate the accuracy of the coded EMR data, with a high degree of accuracy reported for the conditions studied.

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 were collected through a healthcare system's EMRs, utilizing the hospital and primary care networks' existing infrastructure. The validation of these mechanisms is inherent in the medical practice and the electronic record-keeping system's operational standards within the hospital network.

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 study defined two cohorts from over one million children in the care network: a closed birth cohort and a cross-sectional cohort, based on specific inclusion criteria (e.g., age at healthcare establishment and follow-up period). This strategy allowed for a comprehensive analysis of disease prevalence and incidence, although it may not explicitly describe as a sampling strategy from a statistical perspective, it effectively targeted the population of interest.

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 presumably carried out by healthcare providers and the research team at The Children's Hospital of Philadelphia. Compensation details for healthcare providers in this context would typically follow standard employment rather than specific compensation for data collection activities.

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 were collected from outpatient healthcare encounters before the subjects' 18th birthday, between January 1, 2001, and December 31, 2013. This timeframe corresponds with the actual healthcare encounters and diagnoses, ensuring the data's contemporaneity with its creation.

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.*

   - The Institutional Review Board of The Children's Hospital of Philadelphia reviewed the study and determined it did not meet the definition of "human subject" research, thus exempting it from requiring ethics approval. However, the chart review portion of the study was approved under a specific protocol, indicating an ethical review process was conducted.

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 were collected directly from the individuals' EMRs within the healthcare system, not via third parties or external sources.

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.*

   - Given the data were extracted from EMRs, notification to individuals specifically for this study's purposes is not explicitly mentioned but would be governed by the healthcare system's general privacy practices and patient consent procedures for medical care.

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 for the use of EMR data for research purposes typically falls under general consent procedures for medical care and research at the institution, although specific consent mechanisms for this study are not detailed.

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).*

- Not specified in the provided document summary. In general, patients have rights regarding their health information and can request changes or restrict access under healthcare privacy laws, but specifics for this study were 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.*

    - An analysis of the potential impact of the dataset on data subjects is not explicitly mentioned in the provided document summary. Given the dataset's de-identified nature, the focus would likely be on minimizing privacy risks while enabling research benefits.

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.*

    - The study applied data preprocessing and cleaning by utilizing ICD-9 codes for accurate disease identification and excluding non-relevant conditions. A manual chart review ensured the reliability of the data used.

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 document doesn't specify if "raw" EMR data was saved separately. Due to privacy concerns, it's likely that only processed, de-identified data was used for the study, with raw data remaining within the secure EMR system.

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

    - There's no information on specific software for data preprocessing. The study's data preparation likely involved standard medical record systems and expert review, without distinct software tools being mentioned.

Uses

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

    - Yes, the dataset has been used in epidemiological studies to understand food allergy prevalence, risk factors, and outcomes, contributing to peer-reviewed publications and health policy recommendations.

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.*

   - https://zenodo.org/records/44529

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, an academic database. https://zenodo.org/records/44529

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

   - Website. DOI: 10.5281/zenodo.44529

3. *When will the dataset be distributed?*

   - 2016

4. *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

Maintenance

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

   - Gebru et al. (2021)

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

   - Email

3. *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

4. *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.*

- Yes

5. *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.*

    - 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.

Hill, David, Robert Grundmeier, Gita Ram, and Jonathan Spergel. 2016. “Allergy Dataset.” Zenodo. https://doi.org/10.5281/zenodo.44529.