Datasheet for '2023 Natality Data for the United States'*

Analysis of US Birth Records for Public Health Insights

Yunkyung Ko

Invalid Date

The 2023 Natality Dataset provides birth record data for the United States, covering demographic, medical, and geographic variables. The dataset enables research on public health, maternal outcomes, and disparities in birth trends. The datasheet documents the creation, composition, and potential uses of the dataset while addressing its limitations and ethical considerations.

Extract of the questions from Centers for Disease Control and Prevention (CDC) (2016).

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 serve as a standardized source for analyzing trends in maternal and infant health across the United States (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). Its primary purposes include: - Public health researching and reporting By capturing demographic, medical, and geographic variables, it supports studies on key public health indicators. The target of studies include rates of preterm births and low birth weight, and disparities in maternal health outcomes by race, ethnicity, or socioeconomic status (Centers for Disease Control and Prevention, National Center for Health Statistics 2023). - Policy development and evaluation Policymakers make use of this dataset to evaluate the effectiveness of health interventions and programs, such as those under the health people of age 20 to 30 (Centers for Disease Control and Prevention 2023d). It identifies geographic regions or sub-populations that require targeted healthcare intiatives or funding, and assess trends over time to guide future health policies and resource allocation (Centers for Disease Control and Prevention 2023d). - Addressing data gaps in maternal and infant health The dataset solves for the inconsistencies in reporting across states by providing

*Code and data are available at: https://github.com/koyunkyung/infant_health.

a nationwide coverage of natality data (Big Data Framework 2023). It standardizes information collected from birth certificates, ensuring compatibility for cross-state comparisons and longitudinal studies (Centers for Disease Control and Prevention 2023c).

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 the National Center for Health Statistics (NCHS), a division of the Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). The NCHS, which is responsible for collecting, standardizing, and disseminating natality data, supports statistical reporting at both national and state levels (Centers for Disease Control and Prevention 2023c). This effort was part of the Vital Statistics Cooperative Program, which monitors key health trends and provides data for annual reports on births in the United States (Centers for Disease Control and Prevention 2023b). It furthermore supports a broad range of research tasks including the understanding of long-term health implications of birth conditions by linking with other datasets (Centers for Disease Control and Prevention 2023b).

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 of the 2023 Natality Dataset was funded by the U.S. Department of Health and Human Services (HHS) through the Vital Statistics Cooperative Program (VSCP) (Centers for Disease Control and Prevention 2023e). This program is administered by the Centers for Disease Control and Prevention (CDC) and financially supports state and local offices to collect, process, and transmit natality data to the National Center for Health Statistics (NCHS) (Centers for Disease Control and Prevention 2023e). This funding aims to ensure the nationwide consistency and quality of vital statistics reporting (Centers for Disease Control and Prevention 2023c).

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 dataset represent individual live birth records from the United States, as documented on standardized birth certificates (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). Each record contains detailed information about the birth, including demographic data (e.g., maternal age, race, education), medical details (e.g., prenatal care, delivery method, complications), and geographic identifiers (e.g., state and country of birth) (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). These records are uniform in structure

and do not include multiple types of instances, ensuring consistency across all data points (Centers for Disease Control and Prevention 2023a).

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

The dataset contains a total of 3,605,081 instances, each representing a single live birth recorded in the United States during the year 2023 (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). These instances include both U.S. resident births (3,596,017) and a smaller number of births to non-residents (9,064), showing a wide coverage of all registered live births nationwide (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023).

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 2023 Natality Dataset is a census of all registered live births in the U.S, ensuring relatively complete nationwide coverage with low sampling error (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). Data is collected uniformly through mandatory birth certificates and validated by the Vital Statistics Cooperative Program (Centers for Disease Control and Prevention 2023a). While rare unregistered births may introduce minimal undercoverage, systematic quality checks ensure data consistency and relatively accurate representation across geographic, demographic, and medical dimensions (Centers for Disease Control and Prevention 2023a).

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 consists of structured data derived from standardized birth certificates (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). The data includes both raw fields (e.g., mother's age, infant's birth weight) and processed features (e.g., gestational age in weeks, calculated Apgar scores) (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). These fields capture demographic, medical, and geographic information, such as: - *Demographics: maternal age, race, education, marital status, residency* - Medical details: birth weight, gestational age, prenatal care visits, complications during pregnancy or delivery, method of delivery (e.g., cesarean) - *Geographic information: state and country of birth

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

With no explicit label or target variable, several fields can serve as target variables for analysis or modeling, depending on the research question. For instance, outcomes such as `low birth weight (<2500 grams)`, `preterm birth (<37 weeks gestation)`, or `Apgar score below 7` can be used as target variables in studies predicting adverse birth outcomes (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023).

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.*

Commonly missing fields in this dataset include details about the father (e.g., age, education) or specific prenatal care data (National Bureau of Economic Research (NBER) 2023). It may often be due to incomplete reporting by healthcare providers or individuals (BMJ Medicine 2023). Moreover, certain medical details, such as complications during pregnancy, may be underreported becuase of inconsistencies in documentation practices across states or hospitals (BMJ Medicine 2023). These gaps typically arise from data collection challenges rather than intentional omissions and can be addressed through data validation and impuation methods where feasible (Mann, Carl J. 2003).

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.*

    - TBD

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.*

    - TBD

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

    - TBD

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.*

- TBD

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.*

    - TBD

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.*

    - TBD

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.*

    - TBD

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.*

    - TBD

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.*

    - TBD

16. *Any other comments?*

    - TBD

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.*

    - TBD

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?*

- TBD

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

   - TBD

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)?*

   - TBD

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.*

   - TBD

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.*

   - TBD

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

   - TBD

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.*

   - TBD

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.*

   - TBD

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).*

    - TBD

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.*
    - TBD
12. *Any other comments?*
    - TBD

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

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - TBD
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.*
    - TBD
3. *What (other) tasks could the dataset be used for?*
    - TBD

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?*

    - TBD

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

    - TBD

6. *Any other comments?*

    - TBD

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.*

    - TBD

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

    - TBD

3. *When will the dataset be distributed?*

    - TBD

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.*

    - TBD

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.*

    - TBD

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.*
    - TBD
7. *Any other comments?*
    - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - TBD
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - TBD
3. *Is there an erratum? If so, please provide a link or other access point.*
    - TBD
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)?*
    - TBD
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.*
    - TBD
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.*
    - TBD
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.*
    - TBD

8. *Any other comments?*
    - TBD

References

Big Data Framework. 2023. "Understanding Data Quality." 2023. https://www.bigdataframework.org/knowledge/understanding-data-quality/.

BMJ Medicine. 2023. "Strengths and Limitations of Observational Studies." *BMJ Medicine* 2 (1): e000399. https://doi.org/10.1136/bmjmed-2023-000399.

Centers for Disease Control and Prevention. 2023a. *Guidelines for Birth Certificate Data Specifications*. Centers for Disease Control; Prevention. https://www.cdc.gov/nchs/data/dvs/Guidelinesbirthspecs1101acc.pdf.

———. 2023b. *Natality Information Help - CDC WONDER*. Centers for Disease Control; Prevention. https://wonder.cdc.gov/wonder/help/natality.html.

———. 2023c. *National Vital Statistics System (NVSS)*. Centers for Disease Control; Prevention. https://www.cdc.gov/nchs/nvss/index.htm.

———. 2023d. *Program Evaluation Framework*. Centers for Disease Control; Prevention. https://www.cdc.gov/evaluation/php/evaluation-framework/index.html.

———. 2023e. *Vital Statistics Cooperative Program: Contracts and Data Collection*. Centers for Disease Control; Prevention. https://www.cdc.gov/nchs/data/series/sr_01/sr01_062.pdf.

Centers for Disease Control and Prevention (CDC). 2016. "Facility Worksheet for the Live Birth Certificate." https://www.cdc.gov/nchs/data/dvs/facility-worksheet-2016-508.pdf.

Centers for Disease Control and Prevention (CDC), and National Center for Health Statistics (NCHS). 2023. *User Guide to the Natality Public Use File, 2023 Data Set*. Centers for Disease Control; Prevention (CDC). https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/natality/UserGuide2023.pdf.

Centers for Disease Control and Prevention, National Center for Health Statistics. 2023. "Births: Provisional Data for 2023." Centers for Disease Control; Prevention. https://www.cdc.gov/nchs/data/vsrr/vsrr035.pdf.

Mann, Carl J. 2003. "Observational Studies: Cohort and Case-Control Studies." *Emergency Medicine Journal* 20 (1): 54–60. https://doi.org/10.1136/emj.20.1.54.

National Bureau of Economic Research (NBER). 2023. "2023 Natality Data for the United States." National Center for Health Statistics (NCHS). https://data.nber.org/nvss/natality/csv/2023/natality2023us.csv.