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

Analysis of US Birth Records for Public Health Insights

Yunkyung Ko

November 29, 2024

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 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 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:

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

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 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.*

Relationships between individual instances are not explicitly defined in this dataset and each record represents a standalone live birth (Centers for Disease Control and Prevention 2023a). Every individual instance is not linked to other births, even in cases of familial connections (e.g., siblings born in different years) (Centers for Disease Control and Prevention 2023a). However, relationships can be inferred indirectly through shared geographic or demographic attributes, such as births occurring within the same country or among mothers of similar age and race (Centers for Disease Control and Prevention (CDC) and National Center for Health

Statistics (NCHS) 2023). These inferred connections are not explicitly encoded, but can be analyzed statistically for patterns or trends.

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 Centers for Disease Control and Prevention (CDC) does not prescribe specific data splits for this dataset (Centers for Disease Control and Prevention 2023a), but researchers are encouraged to determine appropriate data partitioning strategies. For instance, they can divide data into training, validation, and testing sets based on their specific analytical objectives and methodological requirements (Labs 2023). This approach may ensure the data to be utilized effectively in addressing diverse research questions.

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

The 2023 Natality Dataset contains relatively high-quality data due to its collection and standardization by the National Center for Health Statistics (NCHS). The NCHS employs rigorous data cleaning protocols including checks for internal consistency, cross-referencing with prior datasets for anomalies, and imputing missing values where appropriate (Disease Control and Prevention 2023). However, there are potential errors, noise, and redundancies that arise from the complexities of data collection across states and facilities:

- Incomplete Data:

  Certain fields, such as prenatal care visits, father's information (e.g., age, education), and specific medical details, may be missing (National Bureau of Economic Research (NBER) 2023). These gaps often result from incomplete reporting by healthcare providers or omitted information from parents (Centers for Disease Control and Prevention 2023e).

- Reporting Variability:

  Differences in how states and hospitals document medical complications, delivery methods, or maternal health conditions can lead to inconsistencies (Mann, Carl J. 2003). For example, complications during labor may be underreported in some jurisdictions due to differences in medical coding practices or provider discretion (Mann, Carl J. 2003).

- Data Noise:

  Some data points may include noise due to typographical errors, such as invalid maternal ages (e.g., extreme outliers) or implausible gestational ages (BMJ Medicine 2023). These errors, though infrequent, can affect statistical analyses if not addressed.

- Geographical and Temporal Variations:

  Geographic disparities in healthcare access and reporting capabilities can lead to systematic biases, such as underrepresentation of certain rural areas or delays in data submission (Centers for Disease Control and Prevention 2023e).

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 '2023 Natality Data for the United States' is largely self-contained, as it contains extensive birth records collected through standardized birth certificates across all United States and territories (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). The dataset does not rely on external resources like websites, tweets, or other datasets for its core content. However, researchers can enhance analyses by linking this dataset to other publicly available resources, such as socioeconomic data from the American Community Survey(ACS) or environmental data from the EPA (Centers for Disease Control and Prevention 2023b).

For access to more specific geographic data, such as specific countries or smaller regions, researchers may need to submit a formal data request to the National Center for Health Statistics (NCHS) (Centers for Disease Control and Prevention 2023b). These requests are subject to review and approval to ensure compliance with privacy and ethical standards (Centers for Disease Control and Prevention 2023e). Such restrictions are in place to prevent the misuse of sensitive data and to protect individual privacy (Centers for Disease Control and Prevention 2023e).

a) Guarantees of Continuity

The datsaet itself is archived and maintained by the National Center for Health Statistics (NCHS), ensuring its availability over time (Centers for Disease Control and Prevention 2023e). External resources for linkage, like ACS data, are also federally maintained, providing similar guarantees of continuity (Bureau 2023).

b) Archival Versions

The NCHS maintains archival versions of the natality dataset, preserving the data as it existed at the time of release (National Bureau of Economic Research (NBER) 2023). These archives include documentation to support reproducibilty.

c) Restrictions

While the dataset is free of charge and unrestricted for general use, requests for detailed geographic data may require additional permissions. Users must also follow ethical guidelines when linking with external datasets to prevent privacy violations or misuse (Centers for Disease Control and Prevention 2023e).

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.*

The dataset does not contain data that is protected by legal privilege or doctor-patient confidentiality, as it is publicly available and anaonymized to safeguard individual privacy (National Bureau of Economic Research (NBER) 2023). Personally identifiable information (PII), such as names, Social Security numbers, or exact residential addresses, is excluded to comply with federal privacy laws and regulations, including the Health Insurance Portability and Accountability Act (HIPAA) (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023).

However, the dataset includes sensitive demographic and health-related information, such as maternal age, race, education, and detailed medical data (e.g., complications during pregnancy, prenatal care visits). While this information is anonymized, it is still considered sensitive due to its potential to identify individuals when combined with external datasets. To mitigate risks, the National Center for Health Statistics (NCHS) applies strict data handling policies, and access to specific geographic data, such as detailed county-level information, requires additional permissions and adherence to ethical guidelines (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). These measures ensure that the dataset is suitable for public health research while minimizing risks associated with confidentiality breaches.

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 dataset does not contain data that is inherently offensive, insulting, or threatening. However, certain fields may include sensitive demographic or medical information, such as maternal age, race, ethnicity, and health outcomes (e.g., complications during pregnancy, infant mortality). When misinterpreted or presented without proper context, this data could inadvertently reinforce stereotypes or cause anxiety, particularly in discussions around health disparities or adverse outcomes (Centers for Disease Control and Prevention 2023e).

For instance, data highlighting disparities in preterm birth rates among different racial or ethnic groups could be misused or misrepresented in ways that perpetuate stigma or bias (BMJ Medicine 2023). To address this, researchers and policymakers using the dataset are encouraged to approach analyses with sensitivity, ensuring that findings are contextualized and presented in a manner that respects the dignity and privacy of all individuals represented (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023).

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.*

Sub-populations identified by the dataset based on demographic and medical attributes (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023) include:

- Age:

The maternal age variable allows for sub-populations such as adolescent mothers (<20 years old), mothers of advanced maternal age (35+ years), and age-specific birth trends.

- Race and Ethnicity:

Maternal race and ethnicity are recorded, enabling analyses of sub-populations such as non-Hispanic, White, Black, Hispanic, and Asian mothers. This enables studies on health disparities, such as differences in preterm birth or low birth weight rates among racial and ethnic groups.

- Geographic Location:

Births are categorized by state and country, allowing for sub-populations based on urban or rural settings and regional disparities in healthcare access or outcomes.

- Medical Factors:

The dataset identifies sub-populations based on medical characteristics, such as mothers with complications during pregnancy, preterm births, or low birth weight infants.

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.*

It is not possible to directly identify individuals from the dataset, as it is anonymized and excludes personally identifiable information (PII) such as names, addresses, and Social Security numbers (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). However, indirect identification could be theoretically possible if the dataset is combined with external information. For instance, a record with unique combinations of attributes—such as maternal age, race, specific birth complications, and detailed geographic location—could potentially be matched to external data sources like local news reports or publicly available registries (Centers for Disease Control and Prevention 2023e). To mitigate such risks, the dataset limits access to granular geographic details (e.g., county-level data) and enforces strict ethical guidelines for use (Centers for Disease Control and Prevention 2023e). These measures significantly reduce the likelihood of re-identification while preserving the dataset's utility for research.

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.*

The dataset contains sensitive data, primarily related to health and demographics. Sensitive fields include (Centers for Disease Control and Prevention 2023e):

- Race and Ethnic Origins:

  Records of maternal race and ethnicity are included in the dataset. While this information is critical for analyzing disparities in birth outcomes, these fields must be interpreted carefully to avoid reinforcing stereotypes, as disparities often reflect systemic inequalities rather than intrinsic differences.

- Health Data:

  - Maternal Health: Data on complications during pregnancy, delivery methods (e.g., cesarean sections), prenatal care visits
  - Infant Outcomes: Metrics like Apgar scores, birth weight, gestational age, whether the birth was preterm

  These fields are sensitive as they relate to medical conditions that could reveal vulnerabilities or personal health histories if misused.

- Geographic Location:

  Even though the dataset provides state-level and limited geographic information, such as country or metropolitan area, access to detailed geographic data is possible after special permission. Geographic information, when combined with other datasets, could lead to indirect identification or stigmatization of specific communities with poor health outcomes.

In order to minimize the risk of misuse or harm to individuals and communities, researchers must handle these fields responsible, contextualizing analyses to avoid reinforcing biases or stigmatizing groups (Centers for Disease Control and Prevention 2023e). Access to more specific geographic details is restricted to approved uses, ensuring compliance with federal privacy regulations, including HIPAA and ethical guidelines from the National Center for Health Statistics (NCHS) (Centers for Disease Control and Prevention 2023d).

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 in the dataset was acquired from standardized birth certificates across the United States. These certificates are filled out at the time of birth by healthcare providers, including physicians, midwives, and hospital staff, based on direct observations during the delivery process and information reported by the mother or other reliable sources (e.g., medical records) (Centers for Disease Control and Prevention 2023e).

- Directly Observable Data:

Certain data points, such as the infant's weight at birth, gestational age, and Apgar scores, are measured or observed by healthcare professionals.

- Reported by Subjects:

Information such as the mother's age, education level, race/ethnicity, and prenatal care visits is often provided by the mother or documented from her medical history.

- Indirectly Inferred Data:

Some fields, such as gestational age or complications during delivery, may be derived from medical evaluations, clinical records, or healthcare provider notes.

The following methods were used to ensure data quality and consistency (Centers for Disease Control and Prevention 2023a):

- Standardization:

The dataset adheres to the National Center for Health Statistics (NCHS) guidelines, ensuring that data collection and categorization are uniform across all states.

- Validation:

Healthcare professionals validate critical fields such as birth weight and gestational age by cross-referencing clinical records. Data entry undergoes quality checks by state vital statistics offices before submission to the NCHS.

- Verification:

Regular training for data collectors and feedback loops between hospitals and state agencies help minimize errors and improve reliability.

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 in the dataset was collected through a combination of manual and automated mechanisms, standardized by the National Vital Statistics System (NVSS) under the National Center for Health Statistics (NCHS). The key mechanisms and procedures include (Centers for Disease Control and Prevention 2023c):

1. Manual Data Entry:

Birth certificates are completed by healthcare providers (e,g,, physicians, nurses, or midwives) and administrative staff in hospitals or birthing centers. These individuals manually document details such as maternal demographics, birth outcomes, and medical procedures based on observations, patient interviews, and clinical records.

2. Standardized Electronic Systems:

Many states utilize Electronic Birth Registration Systems (EBRS) to digitize the process (EBRS Online 2024). These systems allow hospital staff to enter birth certificate data electronically, reducing manual errors (EBRS Online 2024). The EBRS software enforces validation rules, such as checking for logical consistency (e.g., gestational age aligning with birth weight) and completeness (e.g., ensuring no mandatory fields are left blank) (Online 2024).

3. Data Transmission:

Once entered, data is transmitted from state vital statistics offices to the NCHS through secure channels (Centers for Disease Control and Prevention 2023c). This ensures uniform data collection across all states and territories. The NCHS applies additional validation algorithms to identify and flag inconsistencies, such as outliers in birth weight or maternal age (Centers for Disease Control and Prevention 2023e).

4. Validation Procedures:

Training: Healthcare providers and staff responsible for data collection are trained periodically to ensure accuracy and compliance with NCHS guidelines.

Audits: Routine quality assurance audits are conducted to compare entered data with source documents (e.g., medical records) and identify discrepancies.

Feedback Loops: States receive feedback reports from the NCHS to address anomalies and improve future data collection processes.

By leveraging a combination of manual human curation, electronic systems, and rigorous validation protocols, the dataset maintains a relatively high standard of accuracy and reliability, making it a trusted resource for public health research and policy development.

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

As part of the National Vital Statistics System (NVSS), the dataset includes all reported live births based on data collected through birth certificates, ensuring extensive coverage of the target population (Centers for Disease Control and Prevention 2023c). While the dataset itself is exhaustive and not a sample, researchers often create subsets or samples for specific analyses. These subsets can be generated using the following strategies:

- Deterministic Sampling:
  - Researchers may filter data based on predetermined criteria (Medium 2024), such as geographic region (e.g., focusing on births in a specific state) or maternal age groups (e.g., teenage mothers or mothers over 40).
  - This approach ensures that the subset aligns with the research objectives but may not represent the entire population.
- Probabilistic Sampling:
  - For statistical analyses, probabilistic methods such as stratified random sampling can be employed (Alexander 2023). For example, researchers might stratify the data by race/ethnicity or socioeconomic status to ensure representation across key demographic groups.
  - Sampling weights, often provided in supplementary datasets, can adjust for any oversampling or undersampling to maintain population-level estimates (Alexander 2023).
- Exclusions Due to Missing or Invalid Data:
  - In practice, records with missing, incomplete, or invalid fields may be excluded from specific analyses (Central 2013). For instance, births with unrecorded gestational ages or birth weights might be omitted in studies focusing on preterm births.

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 process involved healthcare providers, administrative staff, and state-level vital statistics offices (Centers for Disease Control and Prevention 2023e). Healthcare providers, including physicians, nurses, and midwives, recorded information during and after delivery, while hospital administrative staff assisted in completing birth certificates. State-level vital statistics offices ensured data entry and validation before transmitting the records to the National Center for Health Statistics (NCHS) (Centers for Disease Control and Prevention 2023e). These individuals were compensated as part of their professional roles within

healthcare and administrative systems, and no additional payments were made specifically for data collection (Centers for Disease Control and Prevention 2023a). The NCHS also provided training and feedback to ensure standardized data quality across all contributors (Centers for Disease Control and Prevention 2023e).

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 was collected continuously throughout the calendar year of 2023, with records submitted by hospitals, birthing centers, and state vital statistics offices shortly after each birth (Centers for Disease Control and Prevention 2023e). The data collection timeframe aligns with the creation timeframe of the instances, as birth certificates are typically completed and validated within days or weeks of the birth event (Centers for Disease Control and Prevention 2023a). This real-time collection ensures that the dataset accurately reflects births occurring in 2023 without significant delays or retrospective adjustments.

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 National Vital Statistics System (NVSS), where strict privacy and data protection regulations are operated, is where the birth certificate data was collected through (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). While the dataset itself does not require an Institutional Review Board (IRB) review due to its aggregated and de-identified nature, the underlying processes for collecting and managing birth certificate data adhere to ethical standards outlined by the National Center for Health Statistics (NCHS) and state-level public health agencies (Disease Control and (CDC) 2024b). These processes ensure compliance with federal and state laws, such as the Public Health Service Act, to protect individual confidentiality and data integrity (Disease Control and (CDC) 2024b). Additional ethical safeguards include restricted access to identifiable data and routine audits to maintain data security (Disease Control and (CDC) 2024b).

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 not collected directly from individuals but obtained via third parties, specifically through state vital statistics offices. These offices collect data from birth certificates completed by healthcare providers and administrative staff at hospitals and birthing centers (Centers for Disease Control and Prevention 2023e). The birth certificate information includes both observed data and data reported by the mother, ensuring comprehensive coverage of birth-related variables (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023).

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.*

Individuals are indirectly notified about the data collection as part of the birth registration process, which is legally mandated in the United States (Disease Control and (CDC) 2024b). Hospitals and healthcare providers inform mothers that birth-related data is recorded on official birth certificates. Notices regarding the use of this data for statistical purposes, including anonymization and public health research, are typically included in state guidelines or hospital documentation (Centers for Disease Control and Prevention 2023e).

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 data collection and use is implied as part of the legal requirement for birth registration in the United States (Disease Control and (CDC) 2024b). Parents or guardians provide the necessary information to healthcare providers, who then complete the birth certificate (Centers for Disease Control and Prevention 2023e). Explicit consent for the statistical use of de-identified data is not required, as the data is anonymized and used solely for public health purposes under federal law (Disease Control and (CDC) 2024b).

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).*

Because the data is de-identified and collected as part of a legal process, individuals do not have a mechanism to revoke their consent for data collection (Disease Control and (CDC) 2024b). However, stringent privacy protections ensure that the data is used only for statistical and public health purposes (Canada 2024).

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.*

The National Center for Health Statistics (NCHS) conducts routine data protection impact analyses to assess the risks and benefits of using the dataset (Disease Control and (CDC) 2024b). These analyses ensure that the de-identification process effectively protects individuals' privacy while maintaining the dataset's utility for public health research ((GDPR) 2024). Outcomes of these reviews have led to stringent privacy policies and restricted access to personally identifiable information (Disease Control and (CDC) 2024b).

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.*

Preprocessing and cleaning were performed on the raw birth certificate data before compiling the 2023 Natality Dataset. The following are the steps included (Centers for Disease Control and Prevention 2023a):

- Validation and Cleaning:
  - Removal of records with missing or invalid values for critical fields such as birth weight, gestational age, or maternal age.
  - Logical consistency checks to ensure values align with realistic ranges (e.g., birth weights matching gestational age ranges).
- Standardization:
  - Data fields were standardized based on the guidelines provided by the National Center for Health Statistics (NCHS) to ensure uniformity across states.
  - Categorical variables (e.g., race, method of delivery) were encoded into predefined formats to facilitate consistency in analysis.
- Derived Variables:
  - Additional fields, such as age groups or preterm birth indicators, were derived from the raw data for easier stratification in research.

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 retained alongside the processed dataset to ensure transparency and allow for future analysis (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). Researchers and authorized personnel can request access to the raw, unprocessed data through the NCHS Vital Statistics Restricted Data Program. Details on accessing raw data can be found here: https://www.cdc.gov/nchs/data/data_access_and_resources_booklet_web.pdf

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 and cleaning processes were carried out using state-level Electronic Birth Registration Systems (EBRS) and federal validation tools provided by the NCHS (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). While the exact software used at the state level varies, the guidelines and validation tools employed by the NCHS are publicly documented (Centers for Disease Control and Prevention 2023e). More information on these tools can be accessed here: https://www.cdc.gov/nchs/nvss/births.htm

Uses

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

The datasets has already been widely used for public health research and reporting. For instance, it has been used to analyze trends in preterm births, low birth weight rates, and disparities in maternal health outcomes across demographic groups (Centers for Disease Control and Prevention 2023d). Policymakers and public health agencies have utilized the dataset to monitor key health indicators and evaluate the effectiveness of maternal and infant health programs (Centers for Disease Control and Prevention 2023d).

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 dataset is commonly referenced in public health reports published by the Centers for Disease Control and Prevention (CDC) and other research institutions. A repository of related publications can be accessed through the CDC WONDER system: https://wonder.cdc.gov/natality.html. Additional scholarly articles citing the dataset are available via platforms like PubMed https://pubmed.ncbi.nlm.nih.gov/ and Google Scholar https://scholar.google.com/.

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

- Epidemiology: Studying correlations between maternal risk factors (e.g., obesity, smoking) and infant health outcomes.
- Policy Evaluation: Assessing the impact of interventions like prenatal care programs or smoking cessation campaigns.
- Healthcare Resource Allocation: Identifying geographic areas with higher rates of adverse outcomes to guide resource distribution.
- Machine Learning Models: Developing predictive models for maternal and infant health risks.

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?

There are some considerations that dataset consumers of 2023 Natality Data should consider (BMJ Medicine 2023):

- Sampling Bias: While the dataset includes all registered births, populations underrepresented in healthcare systems may have incomplete or inaccurate data.
- De-Identification: Data anonymization limits the ability to link records to external datasets, potentially reducing its utility for longitudinal studies.
- Data Cleaning: Removal of records with missing values might unintentionally exclude marginalized populations, introducing bias. To mitigate these risks, researchers should document preprocessing steps transparently and consider applying statistical adjustments to address biases.

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.*

The dataset is distributed to third parties, including researchers, public health organizations, and policymakers, through the Centers for Disease Control and Prevention (CDC) and its associated systems, such as the CDC WONDER platform https://wonder.cdc.gov/. This distribution ensures that the dataset is accessible for public health analysis and academic research.

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 primarily distributed through the CDC WONDER platform as downloadable files or via an interactive query system. Some datasets may also be made available through restricted-use agreements for more detailed analysis. While the dataset does not have a specific Digital Object Identifier (DOI), it is assigned unique identifiers through the NVSS (Disease Control and (CDC) 2024a). Access it here: https://wonder.cdc.gov/

3. *When will the dataset be distributed?*

The dataset is distributed on an annual basis, typically several months after the close of the calendar year to allow for validation and quality assurance (Disease Control and (CDC) 2024c). The 2023 Natality Dataset was made available in mid-2024 (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023), following the standard release schedule.

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 public use terms as defined by the CDC (Disease Control and (CDC) 2024a). Users must adhere to the terms outlined in the CDC's Data Use Agreement, which restricts the misuse of data for reidentification or commercial exploitation (Disease Control and (CDC) 2024a). Detailed terms are available here: /https://www.cdc.gov/nchs/data/data_access_and_resources_booklet_web.pdf

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. However, access to restricted-use versions of the dataset, containing additional variables or identifiers, requires approval and is subject to stringent use agreements to ensure data confidentiality (Disease Control and (CDC) 2024a).

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 specific export controls apply to the public-use version of the dataset. However, restricted-use data may be subject to additional controls to comply with U.S. privacy laws, such as the Health Insurance Portability and Accountability Act (HIPAA) (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). These controls ensure the dataset remains compliant with national data privacy regulations (Disease Control and (CDC) 2024a).

Maintenance

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

The 2023 Natality Dataset is hosted and maintained by the National Center for Health Statistics (NCHS) as part of the National Vital Statistics System (NVSS) (Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) 2023). The NCHS ensures the dataset is updated, accurate, and accessible to researchers and public health professionals (Disease Control and Prevention 2023).

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

The NCHS can be contacted through its official support channels. For inquiries about the dataset, email: nchsquery@cdc.gov, or visit their website: https://www.cdc.gov/nchs/nhis/contact.htm

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

Errata are issued as needed to correct errors in previously released datasets. Updates and corrections are communicated through the NCHS website: https://www.cdc.gov/nchs/products/errata.htm

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 annually to include new data, correct errors, and improve data quality. Updates are performed by state vital statistics offices and the NCHS (Disease Control and Prevention 2023). Changes are communicated to users through public notices on the NCHS website https://www.cdc.gov/nchs/index.html and platforms like CDC WONDER https://wonder.cdc.gov/.

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.*

As the dataset is anonymized and used for statistical purposes, there are no strict limits on its retention. However, personally identifiable information (PII) is removed to ensure compliance with privacy laws (Disease Control and (CDC) 2024b), such as the Public Health Service Act, and original PII-containing records are securely stored by state authorities (ScienceDirect 2024).

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 remain publicly accessible for historical analysis but are not actively updated. Researchers are encouraged to use the latest version for current analyses. Notices about dataset updates and versioning are provided on the NCHS website: https://www.cdc.gov/nchs/nvss/index.htm

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 formal mechanism for external contributions or augmentation of the dataset, as it is based on standardized birth certificate data collected through the NVSS. Researchers interested in collaborative efforts can propose supplemental analyses or request access to restricted-use datasets (Disease Control and (CDC) 2024b) by applying through the NCHS Restricted Data Access Program: https://www.cdc.gov/nchs/dqs/index.html

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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.

Bureau, U. S. Census. 2023. *American Community Survey (ACS)*. U.S. Census Bureau. https://www.census.gov/programs-surveys/acs.

Canada, Statistics. 2024. "Trust Centre: Laws and Regulations." https://www.statcan.gc.ca/en/trust/laws.

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), 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.

Central, PubMed. 2013. "Sampling Strategies for Public Health Data: Applications and Challenges." *PubMed Central (PMC)* 10 (4): 123–35. https://pmc.ncbi.nlm.nih.gov/articles/PMC3668100/.

Disease Control, Centers for, and Prevention (CDC). 2024a. "Data Access and Resources Booklet." https://www.cdc.gov/nchs/data/data_access_and_resources_booklet_web.pdf.

———. 2024b. "How NCHS Protects Your Privacy." https://www.cdc.gov/nchs/policy/how-nchs-protects-your-privacy.html.

———. 2024c. "National Vital Statistics System: Data Release Schedule." https://www.cdc.gov/nchs/nvss/dvs_data_release.htm.

Disease Control, Centers for, and Prevention. 2023. *Vital Statistics Rapid Release: Natality*.

Centers for Disease Control; Prevention. https://www.cdc.gov/nchs/nvss/vsrr/natality.htm.

EBRS Online. 2024. "About Electronic Birth Registration Systems (EBRS)." https://ebrs.online/about/.

(GDPR), General Data Protection Regulation. 2024. "Data Protection Impact Assessment (DPIA) Template." https://gdpr.eu/data-protection-impact-assessment-template/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." https://arxiv.org/abs/1803.09010.

Labs, Sharp Sight. 2023. "Training, Validation, and Test Sets." 2023. https://www.sharpsightlabs.com/blog/training-validation-and-test-sets/.

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.

Medium. 2024. "Speeding up Your Apps: A Guide to Deterministic Sampling." https://medium.com/design-bootcamp/speeding-up-your-apps-a-guide-to-deterministic-sampling-1269804d60c7.

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.

Online, EBRS. 2024. "Electronic Birth Registration Systems (EBRS) Dashboard." https://ebrs.online/dashboard/.

ScienceDirect. 2024. "Public Health Service Act." https://www.sciencedirect.com/topics/computer-science/public-health-service-act.