HHS Public Access

Author manuscript

*Birth Defects Res*. Author manuscript; available in PMC 2025 January 01.

Published in final edited form as:

*Birth Defects Res*. 2024 January ; 116(1): e2267. doi:10.1002/bdr2.2267.

# Leveraging Automated Approaches to Categorize Birth Defects from Abstracted Birth Hospitalization Data

**Suzanne M. Newton**[1], **Samantha Distler**[1], **Kate R. Woodworth**[1], **Daniel Chang**[2], **Nicole M. Roth**[1], **Amy Board**[1], **Hailee Hutcherson**[3], **Janet D. Cragan**[1], **Suzanne M. Gilboa**[1], **Van T. Tong**[1]

[1]Division of Birth Defects and Infant Disorders, Centers for Disease Control and Prevention,

[2]Eagle Global Scientific, LLC;

[3]G[2]S Corporation

## Abstract

**Background:** The Surveillance for Emerging Threats to Pregnant People and Infants Network (SET-NET) collects data abstracted from medical records and birth defects registries on pregnant people and their infants to understand outcomes associated with prenatal exposures. We developed an automated process to categorize possible birth defects for prenatal COVID-19, hepatitis C, and syphilis surveillance. By employing keyword searches, fuzzy matching, natural language processing (NLP), and machine learning (ML), we aimed to decrease the number of cases needing manual clinician review.

**Methods:** SET-NET captures *International Classification of Diseases, 10th Revision, Clinical Modification* (ICD-10-CM) codes and free text describing birth defects. For unstructured data, we used keyword searches, then conducted fuzzy matching with a cut-off match score of ≥90%. Finally, we employed NLP and ML by testing three predictive models to categorize birth defect data.

**Results:** As of June 2023, 8,326 observations containing data on possible birth defects were submitted to SET-NET. The majority (n=6,758 [81%]) were matched to an ICD-10-CM code and 1,568 (19%) were unable to be matched. Through keyword searches and fuzzy matching, we categorized 1,387/1,568 possible birth defects. Of the remaining 181 unmatched observations, we correctly categorized 144 (80%) using a predictive model.

**Conclusions:** Using automated approaches allowed for categorization of 99.6% of reported possible birth defects, which helps detect possible patterns requiring further investigation. Without employing these analytic approaches, manual review would have been needed for 1,568 observations. These methods can be employed to quickly and accurately sift through data to inform public health responses.

SMN: snewton@cdc.gov.

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Conflict of Interest: The authors declare no conflict of interest.

**Keywords**

Birth defects; Natural language processing; Machine learning; Automation

## Introduction

The Surveillance for Emerging Threats to Pregnant People and Infants Network (SET-NET) collects data abstracted from electronic medical records and birth defects registries on pregnant people and their infants in multiple United States (U.S.) jurisdictions to understand outcomes associated with prenatal exposures, including Coronavirus Disease 2019 (COVID-19), hepatitis C, and syphilis (Woodworth et al., 2021). SET-NET utilizes a complementary approach to birth defects surveillance by monitoring infant outcomes through pregnancy-infant linked longitudinal surveillance, which can provide quick insights to inform clinical decision making and public health efforts.

Large surveillance systems that are meant to be rapid and hypothesis-generating may need to rely on *International Classification of Diseases, 10th Revision, Clinical Modification* (ICD-10-CM) codes rather than extensive chart review by trained clinicians to identify maternal and infant outcomes, which is traditionally how active birth defects surveillance systems have operated. SET-NET contains tens of thousands of birth outcomes for pregnant people exposed to COVID-19, hepatitis C, or syphilis, which would require a large amount of time from analysts to prepare the data for review and from clinicians to manually review each outcome to synthesize and categorize individual birth defect findings for dissemination. Machine learning algorithms have previously been shown to accurately predict manual review by clinicians of the classification of Zika-associated birth defects and autism cases in surveillance data, and automated approaches have the potential to improve the timeliness of those data to inform clinical and public health action (Lee et al., 2019; Lusk et al., 2020).

We sought to develop and evaluate an automated process to categorize possible birth defects resulting from COVID-19, hepatitis C, and syphilis exposure *in utero* by type and organ system to quickly identify potential patterns requiring further investigation. By employing keyword searches, fuzzy matching, natural language processes (NLP), and machine learning (ML) to rapidly categorize birth defects, we aimed to decrease the number of cases needing manual review and reduce time burden on clinicians and analysts as well as maintain high validity in case categorization.

## Materials and Methods

### Population

Within SET-NET, pregnancy outcome data were collected on 136,607 pregnant persons with lab-confirmed Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), hepatitis C, or syphilis infection during pregnancy from 26 U.S. jurisdictions (Arkansas, Arizona, California, Chicago, Georgia, Houston, Illinois, Iowa, Kansas, Los Angeles County, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Nebraska, Nevada, New Jersey, New York state, Ohio, Pennsylvania, Puerto Rico, South Carolina, Tennessee, Utah,

Washington) (Woodworth et al., 2021). 125,650 infants were born to these persons between 2018 and 2022. Birth defect data included in this surveillance network are those reported at birth hospitalization among liveborn infants and could be obtained from electronic medical records, birth defects registries, or birth certificates. The methods below were applied at the birth defect level.

### Clean birth defect diagnosis codes

SET-NET captures ICD-10-CM codes and free text describing possible birth defects at birth hospitalization. One jurisdiction submits Metropolitan Atlanta Congenital Defects Program (MACDP) six-digit codes (Centers for Disease Control and Prevention). These data are cleaned by removing duplicates and missing diagnosis codes, and by matching the ICD-10-CM code format of a birth defect code lookup table.

### Categorize birth defects using cleaned ICD-10-CM codes

In order to synthesize and categorize SET-NET birth defect data, ICD-10-CM Q00-Q99 (Q) codes are mapped to one of 13 categories using a lookup table, which was developed by clinicians and published on GitHub (Centers for Disease Control and Prevention, 2023b). Categories include ICD-10-CM organ system level groupings of congenital malformations, deformations and chromosomal abnormalities (Centers for Disease Control and Prevention, 2023a). In addition, our analysis includes two additional categories, ‘Not a birth defect of interest/unable to categorize’ (e.g., Q38.1, Ankyloglossia) and ‘Not a birth defect’ to categorize ICD-10-CM codes that do not fall within congenital malformations, deformations, and chromosomal abnormalities (e.g., P91.6, Hypoxic ischemic encephalopathy). The lookup table was developed based on guidelines from the National Birth Defects Prevention Network (NBDPN), MACDP, European Surveillance of Congenital Anomalies (EUROCAT), and clinical subject matter expertise (SME) on birth defects for the purposes of SET-NET surveillance, with some variations given differences in surveillance system methodology.

Some defects were categorized as “Not a birth defect of interest” for SET-NET purposes but may be considered relevant in other birth defects surveillance systems that collect additional information. For example, SET-NET excludes all ICD-10-CM codes for undescended testicles, whereas MACDP would include for instances where surgical intervention was required. Conversely, some codes were included for SET-NET purposes regardless of supporting evidence but would only be collected in other systems if supporting data were also present (e.g., ICD-10-CM code Q70.3 for webbed toes is included in SET-NET; however, MACDP does not include webbing between the second and third toes).

### Clean free text descriptions of birth defects

For data that did not match an ICD-10-CM Q code in the lookup table, we first cleaned the free text fields in Python by removing any special characters, setting text strings to lower case, and separating text string in camel case (e.g., “HeartDisease”). This allowed us to match the format of the lookup table’s descriptions of birth defects.

### Categorize birth defects using keyword searches of free text

We then searched for ICD-10-CM codes that did not fall within congenital malformations, deformations and chromosomal abnormalities by identifying text beginning with a letter other than “Q” and followed by any number and categorized them as ‘Not a birth defect’. For the remaining unmatched observations, we used keyword searches for common birth defect descriptions, such as “cleft lip and palate” and “atrioventricular septal defect” to assign appropriate ICD-10-CM Q codes.

### Categorize birth defects using fuzzy matching

For the remaining observations yet to be categorized, we determined how similar the text was to the ICD-10-CM Q code description using the Levenshtein distance (i.e., fuzzy string matching), and computed a match score, with 100% indicating an exact match. Our clinical SME reviewed these matches and set a match score cut-off of 90% or above to indicate a true match as matches below this cut-off resulted in inaccuracies in our dataset. It’s possible that the ideal cut-off may vary depending on the quality of other datasets. Fuzzy string matching was employed using the FuzzyWuzzy module in Python (Cohen, 2020).

### Categorize birth defects using natural language processing and machine learning

For the remaining uncategorized text below the 90% cut-off, we employed NLP and ML. Common stop words were removed. We compared three predictive models (Naïve Bayes, Multi-layer Perceptron [MLP] Classifier, and Random Forest) with no max features, using all contiguous single, double, and triple words derived from the free text strings (known as unigrams, bigrams, and trigrams), and we used five-fold cross validation on each model. The three models were selected for their ease of implementation and variety in methods used for prediction. Data were split into training (75%) and validation (25%) datasets. We measured each model’s accuracy using the weighted average F1 score, which is a measure of model performance that combines sensitivity (recall) and positive predictive value (precision) weighted by the number of true cases in each birth defect category. Our clinical SME reviewed the categorized cases in our test dataset to confirm the correct categories and determine accuracy of the best-performing model on our validation dataset. All code was developed in Python version 3.9.12 and has been published on Github (Centers for Disease Control and Prevention, 2023b). The scikit-learn module was used for developing our predictive models (Fabian Pedregosa, 2011).

## Results

As of June 2023, 8,326 distinct observations containing possible birth defects were submitted to SET-NET from 26 jurisdictions (Figure 1). The majority (n=6,758 [81%]) were matched to an ICD-10-CM Q code and 1,568 (19%) did not match an ICD-10-CM Q code in our lookup table. Through identification of ICD-10-CM non-Q codes (i.e., A00-P99 or R00-Z99) and keyword searches for common birth defect descriptions, we categorized 1,264 observations, and through fuzzy matching we categorized an additional 123. The MLP Classifier model performed the best (weighted average F1 score=0.82) on our validation dataset (Table 1); therefore, it was employed on our test dataset of the remaining 181 unmatched observations. Of these, our clinical SME determined that 144/181 (80%) were

correctly categorized using the MLP Classifier model by comparing the model output to the free text submitted by a jurisdiction. Twenty-six observations were incorrectly categorized as “Not a birth defect of interest/Unable to categorize”. For seven of these 26 observations the correct category should have been “Congenital malformations and deformations of the musculoskeletal system” and for ten of these the correct category should have been **“**Not a birth defect” (Supplementary Table). Without employing our analytic approaches of keyword searches, fuzzy matching, NLP, and ML, manual clinician review would be needed for 1,568 (19%) of our observations to identify and categorize possible birth defects.

## Discussion

Manual review for classification of specific birth defects is resource and time intensive to ensure accuracy of the surveillance data. Based on recent response efforts such as COVID-19, surveillance platforms are needed that can quickly detect patterns of possible birth defects for further investigation. Employing analytic methods including keyword searches, fuzzy matching, NLP, and ML enabled us to quickly and accurately categorize 99.6% of our data into birth detects categories, which allowed for more efficient use of our clinician’s time. Without the use of our analytic methods, 1,568 (19%) of observations containing possible birth defects would be uncategorized and would require time-intensive manual review. Synthesized and categorized birth defect data from SET-NET can be used to identify potential patterns that may indicate a need for further investigation into the implications of COVID-19, hepatitis C, or syphilis infection during pregnancy. Analytic methods that can quickly sift through data to find sentinel events such as birth defects are important, as they may highlight a need for rapid public health action, particularly for public health responses with little data on the impact of a pathogen on maternal and child health, such as COVID-19 (Neelam et al., 2023). These methods have been fine-tuned and used on data submitted to SET-NET on a quarterly basis. Model performance has remained consistent across data submissions.

There are two primary limitations to consider in the context of this analysis. First, the reliance on solely ICD codes for birth defects surveillance can result in potential under-ascertainment of birth defects, reporting errors, and failure to identify some birth defects with high accuracy (Salemi et al., 2018). For example, Salemi and colleagues reported that almost half of reduction deformities of the lower limb were false positives in a passive surveillance system using ICD-9-CM and ICD-10-CM diagnosis codes without medical record review (Salemi et al., 2016). While SET-NET data sources include medical records and birth defects registries, these may be inconsistent across the jurisdictions. SET-NET is meant to rapidly detect patterns, and validation of findings of concern should be performed through more consistent and rigorous birth defects surveillance. Second, due to the large spectrum of birth defects, we trained our model to predict broad categories, including ICD-10-CM organ system level groupings (e.g., congenital malformations of the nervous system), which do not include the ICD-10-CM Q code for the specific birth defect reported. However, through the first two steps in our automated approach (keyword searches and fuzzy matching), we were able to identify the specific ICD-10-CM code for 69% of the 585 reported birth defects, streamlining manual review by clinicians. Therefore, our predictive model was needed for less than a third of our observations.

By implementing several automated processes, we rapidly synthesized and categorized possible birth defects in infants prenatally exposed to COVID-19, hepatitis C, and syphilis. These methods allow for detection of patterns of possible birth defects, whereas additional studies may elucidate potential causal relationships between exposures to pathogens during pregnancy and adverse pregnancy outcomes. There is a push for health departments to leverage interoperable standards from electronic health records that can increase timeliness of case reporting. Aspects of our approach could be layered with others for timely case reporting or could help with preliminary case categorization prior to extensive clinical review (Public Health Informatics Institute). The analytic methods used in this study are available on GitHub and could be adapted for other large surveillance datasets to synthesize and categorize birth defect data for identifying potential patterns of concern, which could expand public health research and action (Centers for Disease Control and Prevention, 2023b).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We thank the following persons for their contributions to this project: Arkansas Department of Health, Arizona Department of Health Services, California Department of Public Health, Chicago Department of Public Health, Georgia Department of Public Health, Houston Health Department, Illinois Department of Public Health, Iowa Department of Health and Human Services/University of Iowa, Kansas Department of Health and Environment, Los Angeles County Department of Public Health, Maryland Department of Health, Massachusetts Department of Public Health, Michigan Department of Health and Human Services, Minnesota Department of Health, Missouri Department of Health and Senior Services, Nebraska Department of Health and Human Services, Nevada Department of Health and Human Services, New Jersey Department of Health and Senior Services, New York State Department of Health, Ohio Department of Health, Pennsylvania Department of Health, Puerto Rico Department of Health, South Carolina Department of Health and Environmental Control, Tennessee Department of Health, Utah Department of Health and Human Services, Washington State Department of Health.

**Funding:**

This study was performed as regular work of the Centers for Disease Control and Prevention (CDC). This work is supported by the Epidemiology and Laboratory Capacity for Prevention and Control of Emerging Infectious Diseases Cooperative Agreement (CK19-1904). Staffing support for this work was funded by CDC to a contract to Eagle Global Scientific (200-2019-06754). Research by Iowa Department of Health and Human Services/University of Iowa reported in this publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UM1TR004403.

## Data availability:

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## References

Centers for Disease Control and Prevention. (2023a). International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). Retrieved May 22 from https://www.cdc.gov/nchs/icd/icd-10-cm.htm

Centers for Disease Control and Prevention. (2023b). SET-NET GitHub Repository. Retrieved May 31 from https://github.com/cdcgov/SET-NET

Centers for Disease Control and Prevention, Emory University, Georgia Mental Health Institute. Metropolitan Atlanta Congenital Defects Program (MACDP). Retrieved May 26 from https://www.cdc.gov/ncbddd/birthdefects/macdp.html

Cohen A (2020). FuzzyWuzzy. In (Version 0.18.0) https://pypi.org/project/fuzzywuzzy/

Fabian Pedregosa GV, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu, Prettenhofer Peter, Weiss Ron, Dubourg Vincent, Vanderplas Jake, Passos Alexandre, Cournapeau David, Brucher Matthieu, Perrot Matthieu, Duchesnay Edouard. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Lee SH, Maenner MJ, & Heilig CM (2019). A comparison of machine learning algorithms for the surveillance of autism spectrum disorder. PloS One, 14(9), e0222907. 10.1371/journal.pone.0222907

Lusk R, Zimmerman J, VanMaldeghem K, Kim S, Roth NM, Lavinder J, Fulton A, Raycraft M, Ellington SR, & Galang RR (2020). Exploratory analysis of machine learning approaches for surveillance of Zika-associated birth defects. Birth Defects Research, 112(18), 1450–1460. https://doi.org/10.1002/bdr2.1767 [PubMed: 32815300]

Neelam V, Reeves EL, Woodworth KR, O'Malley Olsen E, Reynolds MR, Rende J, Wingate H, Manning SE, Romitti P, Ojo KD, Silcox K, Barton J, Mobley E, Longcore ND, Sokale A, Lush M, Delgado-Lopez C, Diedhiou A, Mbotha D, … Gilboa SM (2023). Pregnancy and infant outcomes by trimester of SARS-CoV-2 infection in pregnancy-SET-NET, 22 jurisdictions, January 25, 2020-December 31, 2020. Birth Defects Res, 115(2), 145–159. 10.1002/bdr2.2081 [PubMed: 36065896]

Public Health Informatics Institute. Birth Defects Surveillance. Retrieved October 13 from https://phii.org/birth-defects-surveillance/

Salemi JL, Rutkowski RE, Tanner JP, Matas JL, & Kirby RS (2018). Identifying Algorithms to Improve the Accuracy of Unverified Diagnosis Codes for Birth Defects. Public Health Rep, 133(3), 303–310. 10.1177/0033354918763168 [PubMed: 29620432]

Salemi JL, Tanner JP, Sampat D, Anjohrin SB, Correia JA, Watkins SM, & Kirby RS (2016). The Accuracy of Hospital Discharge Diagnosis Codes for Major Birth Defects: Evaluation of a Statewide Registry With Passive Case Ascertainment. J Public Health Manag Pract, 22(3), E9–e19. 10.1097/phh.0000000000000291

Woodworth KR, Reynolds MR, Burkel V, Gates C, Eckert V, McDermott C, Barton J, Wilburn A, Halai UA, Brown CM, Bocour A, Longcore N, Orkis L, Lopez CD, Sizemore L, Ellis EM, Schillie S, Gupta N, Bowen VB, … Gilboa SM (2021). A Preparedness Model for Mother-Baby Linked Longitudinal Surveillance for Emerging Threats. Matern Child Health J, 25(2), 198–206. 10.1007/s10995-020-03106-y [PubMed: 33394275]

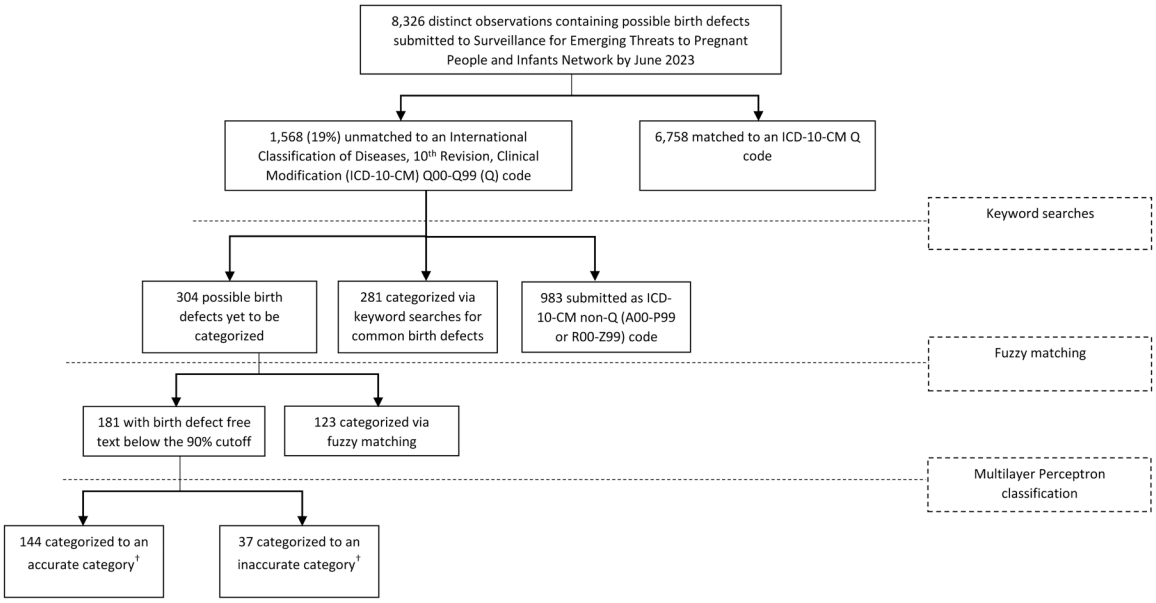


**Figure 1.**

Categorization process of distinct observations containing possible birth defects, Surveillance for Emerging Threats to Pregnant People and Infants Network (SET-NET), June 2023

†Categories include ICD-10-CM congenital malformations, deformations and chromosomal abnormalities' organ system level categories in addition to two created categories, 'Not a birth defect of interest/unable to categorize' and 'Not a birth defect'.

**Table.**

Comparison of models for predicting birth defect categories using text descriptions of birth defects, Surveillance for Emerging Threats to Pregnant People and Infants Network (SET-NET), June 2023

| | | Weighted average F1 score[†] | | | Percent correctly categorized[‡] |
|---|---|---|---|---|---|
| | | Training | Validation | Five-fold Cross-Validation | Test data |
| Unigrams | Multinomial naïve Bayes | 0.77 | 0.75 | 0.71 | |
| | Multilayer Perceptron Classifier[§] | 0.89 | 0.82 | 0.78 | 0.80 |
| | Random Forest[¶] | 0.37 | 0.37 | 0.35 | |
| Bigrams | Multinomialnaïve Bayes | 0.70 | 0.66 | 0.62 | |
| | Multilayer Perceptron Classifier | 0.83 | 0.74 | 0.72 | |
| | Random Forest[¶] | 0.29 | 0.29 | 0.28 | |
| Trigrams | Multinomial naïve Bayes | 0.57 | 0.53 | 0.50 | |
| | Multilayer Perceptron Classifier | 0.72 | 0.61 | 0.58 | |
| | Random Forest[¶] | 0.18 | 0.18 | 0.17 | |

[†]The weighted average F1 score measures model performance by combining sensitivity (recall) and positive predictive value (precision) weighted by the number of true cases in each birth defect category.

[‡]Birth defect categories from the best performing model were reviewed by a clinical subject matter expert to determine accuracy of the model.

[§]Multilayer Perceptron Classifier was the best performing model during training, validation, and cross-validation, and is outlined in the black box.

[¶]Random Forest model included 500 decision trees.