Journal of Biomedical Informatics 87 (2018) 68–78

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Social media mining for birth defects research: A rule-based, bootstrapping approach to collecting data for rare health-related events on Twitter

Ari Z. Klein*, Abeed Sarker, Haitao Cai, Davy Weissenbacher, Graciela Gonzalez-Hernandez

Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

ARTICLE INFO

Keywords:
Natural language processing
Social media mining
Birth defects
Patient-reported pregnancy outcomes
Cohort discovery
Epidemiology

ABSTRACT

Background: Although birth defects are the leading cause of infant mortality in the United States, methods for observing human pregnancies with birth defect outcomes are limited.

Objective: The primary objectives of this study were (i) to assess whether rare health-related events—in this case, birth defects—are reported on social media, (ii) to design and deploy a natural language processing (NLP) approach for collecting such sparse data from social media, and (iii) to utilize the collected data to discover a cohort of women whose pregnancies with birth defect outcomes could be observed on social media for epidemiological analysis.

Methods: To assess whether birth defects are mentioned on social media, we mined 432 million tweets posted by 112,647 users who were automatically identified via their public announcements of pregnancy on Twitter. To retrieve tweets that mention birth defects, we developed a rule-based, bootstrapping approach, which relies on a lexicon, lexical variants generated from the lexicon entries, regular expressions, post-processing, and manual analysis guided by distributional properties. To identify users whose pregnancies with birth defect outcomes could be observed for epidemiological analysis, inclusion criteria were (i) tweets indicating that the user's child has a birth defect, and (ii) accessibility to the user's tweets during pregnancy. We conducted a semi-automatic evaluation to estimate the recall of the tweet-collection approach, and performed a preliminary assessment of the prevalence of selected birth defects among the pregnancy cohort derived from Twitter.

Results: We manually annotated 16,822 retrieved tweets, distinguishing tweets indicating that the user's child has a birth defect (true positives) from tweets that merely mention birth defects (false positives). Inter-annotator agreement was substantial: $\kappa = 0.79$ (Cohen's kappa). Analyzing the timelines of the 646 users whose tweets were true positives resulted in the discovery of 195 users that met the inclusion criteria. Congenital heart defects are the most common type of birth defect reported on Twitter, consistent with findings in the general population. Based on an evaluation of 4169 tweets retrieved using alternative text mining methods, the recall of the tweet-collection approach was 0.95.

Conclusions: Our contributions include (i) evidence that rare health-related events are indeed reported on Twitter, (ii) a generalizable, systematic NLP approach for collecting sparse tweets, (iii) a semi-automatic method to identify undetected tweets (false negatives), and (iv) a collection of publicly available tweets by pregnant users with birth defect outcomes, which could be used for future epidemiological analysis. In future work, the annotated tweets could be used to train machine learning algorithms to automatically identify users reporting birth defect outcomes, enabling the large-scale use of social media mining as a complementary method for such epidemiological research.

1. Introduction

According to the United States Centers for Disease Control and Prevention (CDC), birth defects are the leading cause of infant mortality in the United States [1], likely because the etiology of the majority of birth defects remains unknown [2]. Closing this knowledge gap has been challenging because methods for studying birth defects are limited; for example, pregnant women are largely excluded from clinical trials [3,4], animal reproductive studies may not translate to human risk factors [5,6], and *pregnancy exposure registries* [7] have suffered

* Corresponding author at: 421A Blockley Hall, University of Pennsylvania, 423 Guardian Dr., Philadelphia, PA 19104, United States.
E-mail addresses: ariklein@pennmedicine.upenn.edu (A.Z. Klein), abeed@pennmedicine.upenn.edu (A. Sarker), hcai@pennmedicine.upenn.edu (H. Cai), dweissen@pennmedicine.upenn.edu (D. Weissenbacher), gragon@pennmedicine.upenn.edu (G. Gonzalez-Hernandez).

https://doi.org/10.1016/j.jbi.2018.10.001
Received 12 April 2018; Received in revised form 26 September 2018; Accepted 3 October 2018
Available online 04 October 2018
1532-0464/

from selection bias (e.g., enrolling women who have had prenatal testing with normal results) [8], lack of internal comparator groups [8], and short follow-up periods, which can lead to an under-assessment of birth defects because not all are recognized at birth [9]. Given these methodological limitations, additional methods for observing pregnancies with birth defect outcomes should be explored to complement existing methods for studying birth defects.

In recent work [10], we took the first step towards exploring whether social media mining could be used to complement pregnancy exposure registries as a novel method for observing pregnancies. Considering that 21% of American adults and, more specifically, 36% of Americans between ages 18–29 use Twitter [11], the promise of valuable information directly from the population of interest motivated us to develop and deploy a natural language processing (NLP) and machine learning pipeline that automatically collects and stores the Twitter *user timelines*—all publicly available posts over time by that user—of women who have reported a pregnancy on Twitter. Because pregnancy is a common event, the use of only 14 query patterns was sufficient to retrieve a large set of tweets containing at least 60% true positives. Indeed, most public health applications of social media mining [12] have focused on health-related events that, like pregnancy, impact a relatively large proportion of the population, including influenza epidemics [13], alcohol, tobacco, and drug use [14], and adverse drug reactions [15]. In contrast, because the prevalence of birth defects is only 3% [16], collecting social media data for such rare health-related events, as we propose here, introduces significant challenges to the methods commonly used.

Supervised machine learning algorithms require annotated training data, but, as we will demonstrate in Section 2.1.1, collecting social media data for manual annotation of rare health-related events involves grappling with a high degree of data sparsity and an extremely low signal-to-noise ratio. Our objectives for this study were (i) to assess whether rare health-related events—in this case, birth defects—are reported on social media, (ii) to design and deploy an NLP approach to collecting such sparse data for manual annotation, and (iii) to utilize the annotated data to discover a cohort of women whose pregnancies with birth defect outcomes could be observed on social media for future epidemiological analysis. We mined more than 432 million tweets by 112,647 users who have publicly announced a pregnancy on Twitter. Through an application to birth defects, this paper presents a generalizable NLP-based approach to iteratively preparing an annotated data set of rare health-related events reported on social media, which would enable the use of social media mining as a complementary method for studying such events on a large scale.

2. Methods

2.1. Data collection

To retrieve the small number of tweets that mention possible birth defects from the more than 432 million tweets in our database, we developed a rule-based, bootstrapping approach. The approach relies on a lexicon, lexical variants generated from the lexicon entries, regular expressions, post-processing, and manual analysis guided by distributional properties. The primary goal was to collect a set of tweets that could be categorized—manually distinguishing tweets (possibly) indicating that the user's child has a birth defect (true positives) from tweets that merely mention birth defects (false positives)—for analysis and future use as training data for automatic processing methods. Considering the excessive noise associated with the sparsity of tweets that mention birth defects, we presumed that a rule-based, bootstrapping approach would be more effective than a learning-based approach for collecting an extensive and usable set of tweets for annotation. For this paper, the annotated tweets will direct us to the timelines of the users who posted them, for an inclusion/exclusion analysis to discover a cohort of women whose pregnancies with birth defect outcomes could be observed on Twitter for future epidemiological analysis.

2.1.1. Preliminary query approach

A priori, we did not know how users on social media would linguistically express birth defects, so, initially, we designed a query aimed at maximizing the *recall* of the retrieved tweets—that is, returning the highest possible number of true positives in the database. To identify keywords for the query, we drew upon published reports and guidelines by the National Birth Defects Prevention Network [17], the CDC's Metropolitan Atlanta Congenital Defects Program [18], the Illinois Department of Public Health's Division of Epidemiologic Studies [19], and EUROCAT [20]. When provided, we used the International Classification of Diseases (ICD) codes to look up the birth defects in the Unified Medical Language System (UMLS) [21]. We manually compiled a lexicon of more than 500 keywords. As a heuristic, we attempted to account for birth defects that may be expressed in a variety of clinical, colloquial (e.g., body organs and systems), abstract (e.g., *malformation*, *anomaly*, *defect*, *abnormality*, *disorder*), or, considering Twitter's constraints on message length, abbreviated ways.

Approximately seven million (1.61%) tweets in the database matched our initial set of keywords—far too many for manual processing, and effectively detecting, users' birth defect outcomes. In order to further constrain the semantic space of the retrieved tweets, we required that tweets must also match (variants of) query patterns that indicate a personal experience, such as "my baby" or "I have a child." This additional filtering significantly reduced the number of retrieved tweets, returning approximately 140,000 from the set of keyword-constrained tweets; however, upon manually studying a random sample of 1500 of them, we identified only five true positives. Thus, at this point, the *precision* of the query was 0.003, where *Precision = True Positives/(True Positives + False Positives)*. Because this level of precision would yield too few true positives even if it were possible to manually annotate all 140,000, we decided to shift our focus to obtaining a more substantial proportion of true positives, perhaps at the expense of recall. The results from the preliminary query underscore the methodological limitations of using basic keywords and query patterns—a common data collection approach for mining social media for more prevalent events—for detecting rare events on social media.

2.1.2. Final query approach

In an effort to make the query more precise, we modified the initial lexicon entries to be based primarily on clinical expressions of birth defects, and subsequently added other entries through the bootstrapping approach we present in this section. Our final lexicon (Penn Social Media Lexicon of Birth Defects), available in Supplementary Material, contains approximately 650 entries (single-word or multi-word terms). The words in the entries are expressed as root forms, and we semi-automatically generated lexical variants of the root words, including misspellings, alternative spellings, and inflections (e.g., plurals, possessives, parts of speech). Considering the morphological complexity of clinical terms and their low frequency of use in a non-clinical context, health-related events are likely to be misspelled on social media, so accounting for lexical variants is especially important for detecting rare health-related events. Fig. 1 provides an overview of our workflow for mining social media to discover a cohort of women whose pregnancies with birth defect outcomes could be observed on Twitter for epidemiological analysis. We will describe this workflow in the remainder of Section 2.

We took a data-centric approach to automatically generate some of

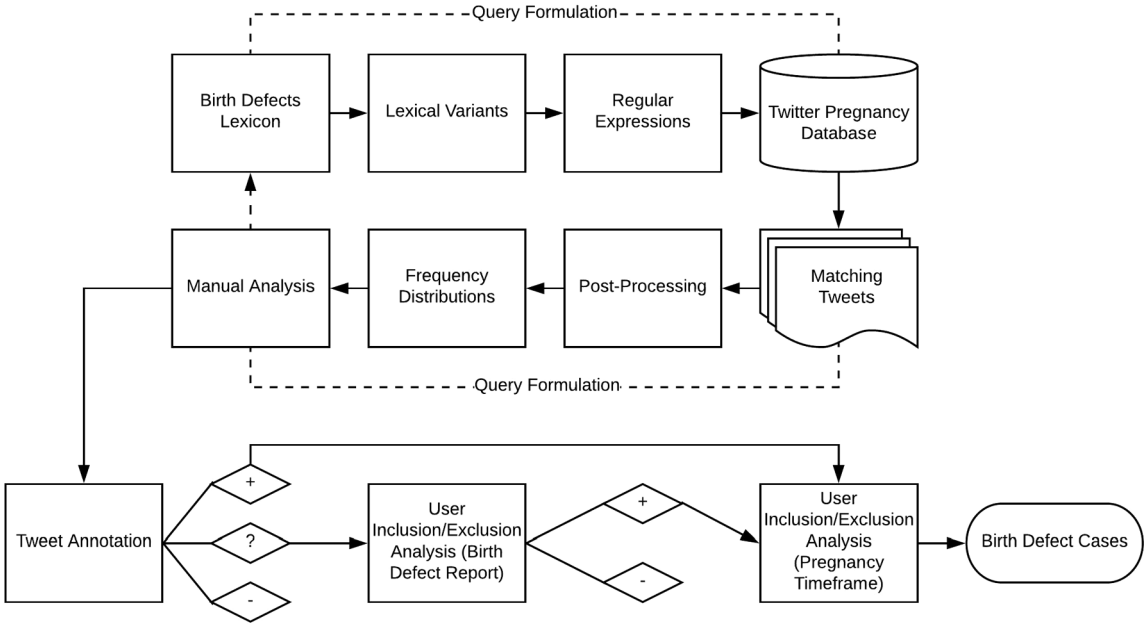

Fig. 1. The workflow for mining social media for birth defect cases. We compiled a lexicon of possible birth defects, generated lexical variants of the lexicon entries, and handcrafted regular expressions to retrieve tweets that mention these birth defects from users who publicly announced a pregnancy on Twitter. We post-processed the tweets and generated frequency distributions of the matching lexicon entries to guide us through a manual analysis of the tweets, which informed an iterative refinement of the query until it yielded a usable set of tweets for annotating "defect" (+), "possible defect" (?), and "non-defect" (−) tweets. To identify users whose pregnancies with birth defect outcomes could be observed on Twitter for epidemiological analysis, inclusion criteria were (i) tweets indicating that the user's child has a birth defect, and (ii) accessibility to the user's tweets within the timeframe of the pregnancy with a birth defect outcome.

Table 1
Samples of lexicon entries and their semi-automatically generated lexical variants used to retrieve tweets that mention birth defects.

Lexicon entries	Lexical variants
down syndrome	downs syndrom, down's syndorme, dwn symdrome
bladder exstrophy	bladdar extrophies, baldder estrophy
diaphragm hernia	diaphragmatic hernea, diaphram henia, diapragm hernai
cyst kidney	cystic kidneys, cysts kidny, cycst kideny
microcephaly	microcephalus, microcephalic, microcephali
tracheoesophageal fistula	tracheo-oesophageal fistulae

the lexical variants for words in the Penn Social Media Lexicon of Birth Defects [22]. This approach can be used to generate lexical variants for entries in any lexicon. First, for each word in an entry, we used a large dense vector model, learned from a set of unlabeled tweets collected for prior work [23], to identify the top 1000 most semantically similar words. Since variants appear in contexts similar to those of the original words, they tend to appear close in the semantic space represented by the vector model. Then, we filtered out lexically dissimilar words by using their Levenshtein distance—a measure of the similarity between two strings computed as the number of deletions, insertions, or substitutions required to transform one string into the other. To filter out words, we used only those above a similarity threshold of 0.80, also called the Levenshtein ratio (LR). LR is calculated as $LR = (lensum - lendist)/lensum$, where *lensum* is the sum of the lengths of the two strings, and *lendist* is the Levenshtein distance between the two strings. Table 1 provides samples of lexicon entries and their lexical variants, including those that were automatically generated. The complete set of variants is available in Supplementary Material.

In the interest of exploration and illustration, we included entries in the Penn Social Media Lexicon of Birth Defects that capture a broad range of birth defects [17–21], which might not be included in all study designs. While the majority of included birth defects would be considered "major" malformations, some of them might be considered "minor" structural defects, such as those that are thought to be positional (e.g., plagiocephaly), maturational (e.g., umbilical hernia), or transient (e.g., a small ventricular septal defect that spontaneously closes). We also included entries for chromosomal anomalies (e.g., trisomy disorders) that might be excluded in studies of teratology. We decided to make the Penn Social Media Lexicon of Birth Defects

Table 2
Tweets are retrieved for the lexicon entry "cleft palate" if they match the combinations of text strings described by the regular expression.

Regular expression	Sample matches
r'\b(?:clefts|cleff|clefty|clefted|cleftee|cleft|clef)(?:\b (?:the|this|that|these|those|a|an|his|her|of|and|(?:w/ |with|w)|their|between|in|on|under|both)\b|\W) *(?:palate|pallettes|palats|pallate|palat|pallet|pallats|pallatt|palet|palates|pallete|palets|pal- lat|paletes|pallett|pallates|pallets|palete|pallattes|pallatts|palletes|pallette|palletts|pallatte)'	Cleft Palate, cleft on her palate cleft-palate #cleftpalate #CleftPalateAwareness

inclusive, allowing epidemiological research that utilizes the lexicon, beyond the scope of this paper, to exclude entries for birth defects deemed not of interest.

We implemented the query as a set of hand-crafted, complex *regular expressions*—search patterns used to match combinations of text strings. Table 2 illustrates the final regular expression form that we applied to most of the lexicon entries. The regular expression begins by specifying that the first character of "cleft" (or any of the variants of "cleft," as indicated by the "|" separator) can be adjoined only with a space or non-alphanumeric character (e.g., punctuation), as indicated by the leading word boundary, "\b". The regular expression goes on to allow particular *stopwords*—common words that would not modify the meaning of the entries—to occur between the words of multi-word entries. The "\b" on both sides of the stopwords—the leading and trailing boundaries—indicates that the stopwords can be adjoined on either side only with a space or non-alphanumeric character. The "|\W" following the trailing boundary of the stopword cluster indicates that stopwords or ("|") a space or non-alphanumeric character ("\W") can occur between the words of multi-word entries. Finally, the asterisk ("*") indicates that a stopword, space, or non-alphanumeric character can occur zero or more times between the words of multi-word entries. Implemented with this regular expression, none of the entries in the Penn Social Media Lexicon of Birth Defects have inclusive relations. Table 2 provides sample combinations of "cleft palate" that would be matched by the regular expression.

As Fig. 1 illustrates, we iteratively refined the lexicon, variants, regular expressions, and data processing in a bootstrapping manner. After each time the query was tested, we generated frequency distributions of the matching lexicon entries (and their variants) and used them to guide us through a manual analysis of the results. We discarded frequently matching entries that, when included, introduced a significant amount of noise—hundreds and sometimes thousands of instances where the entry did not refer to its associated birth defect. We also noticed that some of the entries that are abbreviations of specific birth defects were matching arbitrary strings in URLs—for example, "CHD" (congenital heart defect) in *https://chd.abc/xyz*—*even* with word boundaries (which allow for adjacent non-alphanumeric characters) on both sides of the entries. Thus, in addition to ignoring retweets (indicated by "RT" at the beginning of the tweet), we ended up ignoring URLs in post-processing the retrieved tweets. To address a similar issue, we also ignored usernames (i.e., strings that begin with "@").

In manually studying the data, we also discovered co-occurring patterns that our query was not formulated to match. For example, we found that, by using word boundaries on both sides of the entries, in an effort to reduce noise, we were not capturing birth defects in hashtags in which the end of the entry was adjoined with other words (e.g., *#downsyndromeawareness)*—nontraditional textual representations that social media affords. However, removing trailing boundaries caused some of the entries to become significantly noisier; for example, "club foot" returned tweets containing "club football." For such entries, we re-added trailing word boundaries, as ad hoc rules in post-processing. As this example illustrates, modifying the matching rules, or adding entries to the lexicon, oftentimes required negotiating concomitant noise. With our final query formulation, we reached an estimated precision of 0.07—a more than 23-fold increase over our preliminary precision of 0.003—which appeared to be tolerable for proceeding to retrieval and manual annotation for distinguishing true positives. Using this query formulation against the 432 million tweets included in 112,647 user timelines in our database, we collected 16,822 tweets, which were manually annotated in whole. We will describe the annotation process next.

2.2. Annotation

We collected data for manual annotation three times over a period of four months, with minor variations to the query between the first and second collection times. In total, we collected 16,822 tweets, which were annotated by two professionally trained annotators, with overlapping annotations for 15,547 tweets. We analyzed linguistic patterns in a sample of the retrieved tweets and used this analysis to inform the development of annotation guidelines, which were used to help the annotators distinguish three classes of tweets: "defect," "possible defect," and "non-defect." The complete annotation guidelines are available in Supplementary Material. Table 3 (in Section 3.1) provides examples of annotated tweets. We can summarize the three annotated classes as follows:

- *Defect*: The tweet refers to a person who has a birth defect and identifies that person as the Twitter user's child.
- *Possible Defect*: The tweet is ambiguous about whether a person referred to has a birth defect and/or is the Twitter user's child.
- *Non-defect*: The tweet does not indicate that a person referred to has or may have a birth defect and is or may be the user's child.

2.3. Inclusion/exclusion analysis

The annotations directed us to the timelines of the users who posted them, for an inclusion/exclusion analysis to discover a cohort of women whose pregnancies with birth defect outcomes could be observed on Twitter for future epidemiological analysis. First, we analyzed the timelines of users who posted a "possible defect" tweet (without also posting a "defect" tweet) to determine whether they are the parent of a child with a birth defect. Then, we analyzed the timelines of (i) these included users and (ii) users who posted a "defect" tweet, to determine, for each, whether their timeline encompasses at least part of the timeframe of the pregnancy with a birth defect outcome. Users were excluded from the cohort if we could not determine that they were the parent of a child with a birth defect, or if there were no tweets available during the pregnancy with a birth defect outcome.

2.3.1. "Possible defect" tweets

We excluded users who posted "possible defect" tweets for whom we could not determine, based on analyzing the contextual posts in their timeline, that they are the parent of a child with a birth defect. Many "possible defect" tweets mention the name of someone who has a birth defect, and, thus, are ambiguous as to whether the tweet is referring to the user's child. For such tweets, we began by simply searching the user's timeline for that proper name and examining matching tweets to see if they provide evidence that the referent of the name is the user's child. For many of the other "possible defect" tweets, we began by inspecting a window of tweets surrounding the "possible defect" tweet, assuming, for example, that a tweet such as *he has pyloric stenosis* might be preceded by a tweet such as *my son is having surgery tomorrow*. In some cases, a broader analysis of the timeline was necessary. Based on the number of "possible defect" tweets that we decided to include for further analysis, we assessed the value of retaining this "placeholder" class for future classification.

2.3.2. Pregnancy timeframe

To discover a cohort of women whose pregnancies with birth defect outcomes could be observed on Twitter for epidemiological analysis, we further excluded users whose timeline does not include at least part of the timeframe of the pregnancy with a birth defect outcome. A pregnancy report (e.g., *During my pregnancy*…) does not mean necessarily

that a user was pregnant within the span of her collected timeline, since Twitter's public application programming interface (API) places a limit on how many past tweets can be collected for a user, or a user simply may not have been active on Twitter during her pregnancy. In other cases, the user may be posting tweets during pregnancy, but the pregnancy may not be the one with a birth defect outcome. In cases where we had identified the name of the user's child that had a birth defect, we found that searching for that name towards the beginning of the timeline—usually an indicator that the child had already been born—was an efficient way of excluding data.

2.4. Evaluation

We conducted a preliminary assessment of the prevalence of selected birth defects among the pregnancy cohort derived from Twitter. The purpose of this assessment was not to evaluate Twitter as an alternative source for determining prevalence, but rather to gain insight into how social media could complement the limited methods currently available for studying specific birth defects. In comparing the prevalence of selected birth defects reported on Twitter to national surveillance data, we also sought to contextualize the extent to which there may be birth defect cases in our collection that (i) our approach has not detected or (ii) users have not reported. For the denominator, we statistically derived an estimate of users in the database who have given birth. For the numerator, we used the number of each detected birth defect case that satisfied our inclusion criteria. Based on an effort to identify potential birth defect cases in our collection that our approach has not detected, we measured the recall of the query.

2.4.1. Calculating the prevalence of selected birth defects

For selected birth defects, we calculated their prevalence per 10,000 pregnancy outcomes (live births, stillbirths, miscarriages, and elective terminations) represented in our database, where a "pregnancy outcome" is defined as a user who was posting tweets during a pregnancy that should have ended by the time of this study. Although the prevalence of birth defects in the general population is typically calculated using *live birth* as the denominator, we have not addressed the challenge of distinguishing pregnancy outcomes on social media—a limitation of some surveillance programs as well [24]. In the numerator, we included cases of the selected birth defect for all outcomes, excluding prenatal diagnoses for pregnancies that were ongoing at the time of this analysis. For each selected birth defect, we divided the number of cases by the estimated number of total pregnancy outcomes, and then multiplied by 10,000. In the remainder of this section, we will describe how we estimated the denominator—the number of total pregnancy outcomes represented in our database.

In related work [25], we developed a deterministic system that automatically estimates the prenatal period, and we used a version of it in this study to help estimate the proportion of users in the database who have given birth. The rule-based system is built on handcrafted regular expression patterns that capture pregnancy-related temporal information—in particular, the baby's gestational age, due date, and birth date. The system draws on SUTIME [26] to normalize the temporal expressions in matching patterns and define rules for estimating the beginning and end of pregnancy. In this study, we applied the system to identify users who were pregnant within the span of their timeline, and we used the automatically derived due dates to help determine if a user's pregnancy should have ended by the time of this analysis. We calculated the recall of the timeframe detection system to estimate the number of total pregnancy outcomes represented in the database.

First, we ran the system over the 112,647 timelines in the database. Next, we statistically derived a stratified random sample (n) from the population of matching timelines (N), using the following formula for estimating a proportion for a small, finite population: $n = m/(1 + (m - 1)/N)$, where $m = (z^2_{\alpha/2}\hat{p}(1 - \hat{p}))/\varepsilon^2$. We chose a 95% confidence interval ($z_{0.025} = 1.96$), a maximum error (ε) of 0.05, and a sample proportion ($\hat{p}$) of 0.5. Assuming that the timelines of completed pregnancies were collected earlier and, thus, contain more tweets in the database than the timelines of ongoing pregnancies, the sample was stratified so that timelines of varying tweet lengths were proportionally distributed. Then, to estimate the number of total pregnancy outcomes in N, we evaluated the true proportion of pregnancy outcomes (p) in the n samples and multiplied the N population of users by p. Finally, to estimate the number of total pregnancy outcomes represented in the database, we calculated the recall of the system—by running it on a set of users manually determined to have been pregnant in their timeline and given birth—and divided the estimated number of pregnancy outcomes in N by the system's recall.

2.4.2. Discovering missed tweets and calculating query recall

We experimented with approaches for identifying "defect" and "possible defect" tweets that the query missed (i.e., false negatives), and annotated the set of tweets retrieved by these approaches to calculate the query's recall, where *Recall = True Positives/(True Positives + False Negatives)*. False negatives were unlikely to be represented in a relatively small random sample of the database, given its size and the sparse nature of birth defects, so we attempted to increase our chances of discovering them by using text mining methods deliberately designed to cast a wider net. As our first approach, we simply modified the query's matching rules by removing the leading word boundary ("\b") restriction from the regular expressions, which allowed the beginning of lexicon entries to be adjoined with a space or non-alphanumeric character. To refine this set of retrieved tweets for manual annotation, we filtered it by generating frequency distributions of the matching entries and manually identifying entries that introduced a significant number of true negatives (i.e., tweets that the initial query correctly did not retrieve) into the results; for example, without the restricting boundary, the entry *hole heart* retrieved *whole heart*. We re-added the leading word boundary to the noisy entries (e.g., "\bhole heart"), and re-ran the query over the initial results. This filtering technique yielded a set of 552 potential false negative tweets, which were then manually annotated.

We also used a "fuzzy matching" algorithm to detect tweets in our database containing text strings that are lexically similar to the entries in the Penn Social Media Lexicon of Birth Defects (e.g., *hole heart*) or the linguistic patterns that were matched by the regular expressions in our set of 16,822 annotated tweets (e.g., *hole in her heart*). Prior to performing the similarity measurements, we pre-processed the tweets in our database and annotated data set by lowercasing them and removing stopwords. For each lexicon entry/linguistic pattern, we ran a sliding window of size w_n through all the tweets in the database, and computed the lexical similarity between the words in the sliding window and the lexicon entry/linguistic pattern using the LR measure we discussed in Section 2.1.2. w_n was set as the number of words in a given lexicon entry/linguistic pattern. Tweets with a maximum similarity above the threshold of $(t - (k \times w_n))/100$ were kept for further analysis, where $t = 95$ and $k = 2$ was chosen for lexicon entries, and $t = 90$ and $k = 1$ was chosen for linguistic patterns, which were typically longer than the lexicon entries. These parameters were chosen based on preliminary analysis and were targeted towards maximizing recall while limiting the number of true negatives.

This approach returned more than 20,000 potentially relevant tweets that the initial query did not retrieve. To refine this set for manual annotation, we generated frequency distributions of the word *bigrams* and *trigrams* (i.e., contiguous sequences of n words, where $n = 2$ and $n = 3$, respectively) that were "fuzzily" matched, and removed tweets containing (i) frequent n-grams that we manually identified as marking true negatives (e.g., *whole heart*, which "fuzzily" matches *hole heart* as in *I love him with my whole heart*), or (ii) co-occurring phrases that are strongly indicative of true negatives (e.g., swear words, names of politicians and celebrities). This filtering technique reduced the set of potential "defect" and "possible defect" tweets to 3617. These 3617 tweets and the 552 tweets from the query modification approach described above amount to 4169 potential false negatives that were manually annotated. Table 5 (in Section 4.2) provides examples of tweets that were missed by the initial query, and indicates the text strings that were matched by the methods described in this section.

3. Results

3.1. Annotation

Two annotators annotated 16,822 tweets, posted by 5923 unique users, with overlapping annotations for 15,547 (92.42%) tweets. Their inter-annotator agreement was $\kappa = 0.79$ (Cohen's kappa), considered "substantial agreement" [27]. The first author of this paper resolved the disagreements through independent annotation. In total, 765 (4.55%) tweets were annotated as "defect," 877 (5.21%) tweets were annotated as "possible defect," and 15,180 (90.24%) tweets were annotated as "non-defect." That is, our data collection approach yielded 9.76% true positives (i.e., an approximate precision of 0.10), and, thus, enabled manual annotation. As we discussed in Section 2.1.1, basic unsupervised retrieval methods yielded only 0.33% true positives, so would have required annotating more than 500,000 tweets to detect these "defect" and "possible defect" tweets. In total, 287 (4.85%) users posted at least one "defect" tweet, 359 (6.06%) users posted at least one "possible defect" tweet (without also posting a "defect" tweet), and 5277 (89.09%) users posted at least one "non-defect" tweet (without also posting a "defect" or "possible defect" tweet). Table 3 provides examples of tweets that were annotated as "defect" (+), "possible defect" (?), and "non-defect" (−).

Many of the tweets that were annotated as "defect," such as tweet 1, explicitly refer to the user's child (e.g., *my son*) and indicate that the child has a birth defect; however, some "defect" tweets do not as explicitly state that the user is the parent of a child or that the child has a birth defect. For example, 2 and 3 mention a child (2 more explicitly than 3), but the main text does not indicate that the user is the parent of the child; rather, we can infer from *mother* or *mom* in the hashtags that the user is the parent. Conversely, 4 more explicitly indicates a parent-child relationship (through a metaphor, *my little miracle*), but it requires piecing together the elements of the tweet to infer that the child has a birth defect. Tweet 5 does not even explicitly mention a child, but it does indicate that the user is a mother, and modifying *mummy* with *heart* (together with the hashtags) suggests that the child has a birth defect.

Tweets such as 6 were annotated as "possible defect" because they indicate a parent-child relationship (implied through a genetic concern about hip dysplasia), but are ambiguous about whether the child has a birth defect. The user indicates that her child might have a birth defect (*we're finding out*), but the child has not yet received a diagnosis; nonetheless, 6 is more suggestive of a birth defect than tweets that merely mention routine tests. In addition, 7 is ambiguous about whether the referent of the child referred to by name is the user's child. Similarly, while 8 indicates that *he* has a birth defect, it is unclear if the antecedent of *he* is the user's child. In 9, we took the name of the birth defect to be a metonym for a child who has the birth defect, but we did not know if the user is the parent. Finally, 10 explicitly states that a child has a birth defect, but it is unclear if *your* is being used here to imply a self-reference—that is, to indicate that the *user's* child has a cleft lip.

Tweets such as 11–15 were annotated as "non-defect" because they not do indicate that a person referred to has or may have a birth defect and is or may be the user's child. Tweet 11 does not refer to a specific individual, and while 12 refers to individuals with a birth defect, it does not imply that the user has had a personal experience with them, hence lacking sufficient reason to believe that there is a parent-child relationship. Tweet 13 seems to resemble a "defect" tweet because it explicitly indicates a parent-child relationship (*my son*) and states that the child has Down syndrome (*suffering from*); however, the source attribution (*South African mom*) indicates that the tweet is not about the *user's* child. Similarly, the context surrounding *our #clubfoot kids* in 14

Table 3
Sample tweets retrieved and annotated as "defect" (+), "possible defect" (?), and "non-defect" (−). The bold text indicates the string that was matched by the regular expression. For ethical considerations, names are redacted and, in most cases, the tweet text is slightly modified.

	Tweet	Class
1	My son was born 5 weeks early and we found out about **te fistula** after.	+
2	yes he was born with a **cleft lip** but he is still perfect #motherslove #[name] 💝👶	+
3	Girl says, as I push stroller w/1 arm & carry full basket in other: 'you're so strong!' Me, to myself: 'you don't know the half of it!' **#downsyndromemom**	+
4	My little miracle, we are so blessed to have you **#hypoplasticleftheartsyndrome #hlhs**	+
5	I'm a heart mummy - **#CHD #CHDawarenessweek**	+
6	[name] has hip clicks so we're finding out if he has **hip dysplasia.** I am nervous bc I had it	?
7	Waiting to hear from pediatric neurosurgery to make an appointment for [name] to confirm **craniosynostosis**.	?
8	He was born with **hypospadias** that fixed itself so he's going to get circumsized in 2 weeks. 😐😐😐	?
9	Thought I'd never encounter **Prune Belly** syndrome. Today showed me otherwise	?
10	What it's like to learn that your baby has a **cleft lip**	?
11	**Gastroschisis** looks really scary. Probably shouldn't look at pics online, but I can't help it.	−
12	Kids with **Down Syndrome.**	−
13	South African mom: abortion would've spared my son suffering from **Down syndrome**	−
14	Its fun with feet in #Kenya, as our **#clubfoot** kids run and play clubfoot free today! #runfree2030 #WorldClubfootDay	−
15	I hate that I have an **umbilical hernia** while I'm pregnant. It really scares me. 😞	−

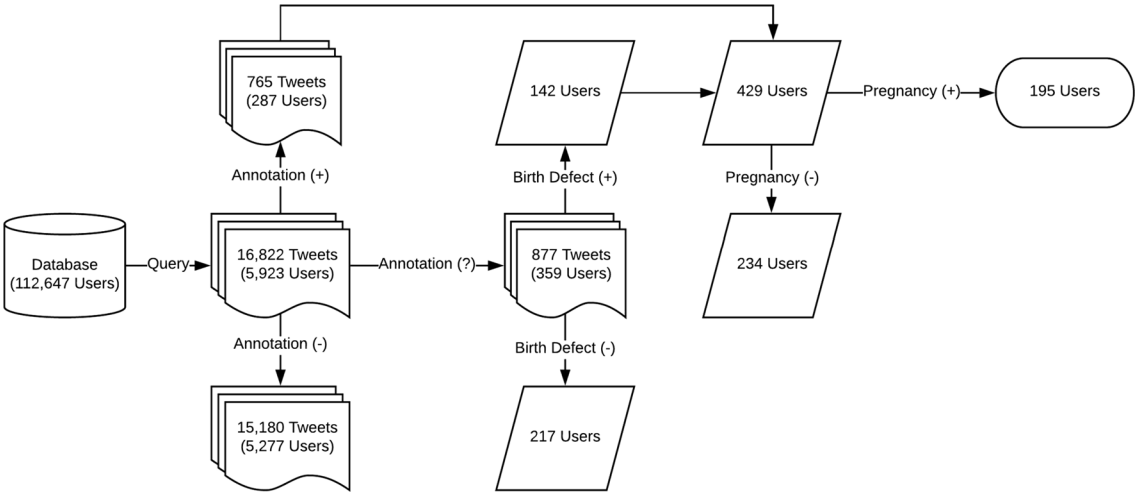

Fig. 2. The results of mining and annotating social media for data on birth defects. Of the 16,822 tweets retrieved from our refined query, 765 tweets (posted by 287 users) were annotated as "defect" (+), 877 tweets (posted by 359 users) were annotated as "possible defect" (?), and 15,180 tweets (posted by 5277 users) were annotated as "non-defect" (−). Of the 359 users who posted a "possible defect" tweet, 142 were determined to indeed have a child with a birth defect, which, added to the 287 users who posted a "defect" tweet, amounts to 429 users reporting a birth defect outcome. Of these 429 users, we found 195 for whom at least some tweets were available within the timeframe of the pregnancy with a birth defect outcome. Of these 195 users, 186 were determined to have a pregnancy outcome at the time of this study.

suggests that the antecedent of *our* is not a parent, but rather a sort of charity organization. Finally, 15 indicates that, in this case, the *user* has an umbilical hernia, not the child.

3.2. Inclusion/exclusion

In the timelines of the 359 users who posted "possible defect" tweets (without also posting a "defect" tweet), we determined that 142 (39.6%) of the users indeed have a child with a birth defect. Of these 142 timelines and the timelines of the 287 users who posted at least one "defect" tweet (429 total timelines), we verified that 195 (45.5%) encompass at least part of the timeframe of the pregnancy with a birth defect outcome. In 10 (5.1%) of these 195 cases, the pregnancy outcome was fetal or neonatal loss. In 9 (4.6%) of these 195 cases, the birth defect was a prenatal diagnosis and the pregnancy was ongoing, meaning that 186 (95.4%) of the 195 users had pregnancy outcomes. Fig. 2 summarizes the results of the annotation and analysis processes, and the frequencies of the specific birth defects reported by the 195 users are provided in Supplementary Material.

For the majority of the 234 users that were excluded because their timelines did not contain tweets within the timeframe of the pregnancy with a birth defect outcome, we determined that, in most of these cases, the users were reporting a birth defect for a pregnancy that occurred prior to the earliest tweets available through Twitter's public API. In other cases, we found that the users were "bots"—Twitter accounts reposting other users' public tweets. At the tweet level, posts by bots resemble true positives, so they are particularly difficult to distinguish for our automatic tweet-level classification system [7]. For a minority of the 234 timelines, we determined that the users indeed were pregnant within the span of their timeline, but that the pregnancy during which we had access to the user's tweets was not the one with a birth defect outcome.

3.3. Evaluation

As explained in Section 2.4.1, we used an automatic system [25] in order to help estimate the proportion of users in the database who had given birth by the time of this study, and used this number as the denominator when calculating the prevalence of selected birth defects reported by the 186 Twitter users with pregnancy outcomes (within our final cohort of 195). The system detected a pregnancy timeframe for 23,743 (21.08%) of the 112,647 timelines in our database. To estimate the proportion of these 23,743 users who have given birth, we evaluated the true proportion of pregnancy outcomes in a stratified random sample of 379 timelines, and found 287 (75.7%) pregnancy outcomes; thus, we estimate that 75.7% (17,973) of the 23,743 users have given birth. To estimate the total number of pregnancy outcomes represented in the database, though, we must account for the timelines *not* detected by the automatic timeframe detection system—that is, the recall of the system. We ran the system on the timelines of the 186 users (within our final cohort of 195) who had been posting tweets during a pregnancy that should have ended by the time of this analysis, and detected a pregnancy timeframe for 86 (46.2%) of the 186 timelines; thus, the 17,973 pregnancy outcomes represent 46.2% of the total outcomes. To estimate the total number of pregnancy outcomes represented in the database, we divided the 17,973 pregnancy outcomes by the system's recall, resulting in 38,903 total pregnancy outcomes as the denominator. Table 4 compares the estimated prevalence of selected birth defects reported on Twitter, per 10,000 births, with their prevalence in the United States population [28,29].

To reiterate, the purpose of this prevalence assessment was not to evaluate Twitter as an alternative source for determining prevalence; rather, it was used in part to help shed light on the extent to which there may be birth defect cases in our collection that (i) our approach has not detected or (ii) users have not reported. We annotated a total of 4169 tweets in the database that were not retrieved by our approach but were

Table 4
A comparison of prevalence, per 10,000 births, for selected birth defects, illustrating the extent to which there may be specific birth defect outcomes among our Twitter cohort that (i) our data collection approach is not adequately detecting or (ii) users are not reporting.

Birth defect	United States[a]	Social media[b]	Social media (Adj.)[c]
Central nervous system defects			
Anencephaly	1.7	0.3	0.4
Encephalocele	0.8	0.3	0.4
Holoprosencephaly	2.1	0.5	0.8
Spina Bifida	3.5	0.5	0.8
Eye and ear defects			
Congenital Cataracts	1.5	0.3	0.4
Anotia/Microtia	1.5	0.3	0.4
Cardiovascular defects			
Congenital Heart Defect[d]	81.4	11.3	17.5
Orofacial defects			
Cleft Lip with Cleft Palate	5.9	0.8	1.2
Cleft Lip without Cleft Palate	3.2	1.5	2.4
Cleft Palate without Cleft Lip	6.1	1.3	2.0
Gastrointestinal defects			
Biliary Atresia	0.6	0.3	0.4
Genitourinary defects			
Bladder Exstrophy	0.3	0.3	0.4
Hypospadias	64.7[e]	0.5[f]	0.8[g]
Musculoskeletal defects			
Diaphragmatic Hernia	2.8	0.8	1.2
Gastroschisis	4.5	1.8	2.8
Limb Reduction Deformities	4.2	0.3	0.4
Clubfoot	13.4	2.3	3.6
Craniosynostosis	5.0	1.8	2.8
Chromosomal defects			
Trisomy 13	1.0	0.3	0.4
Trisomy 18	2.4	1.0	1.6
Turner Syndrome	2.1[h]	0.5[i]	0.8[j]
Trisomy 21 (Down Syndrome)	13.0	4.1	6.4

[a] The denominator used to calculate the prevalence of birth defects in the U.S. is live births.

[b] The denominator used to calculate the prevalence of birth defects reported on social media is the estimated *total* number of pregnancy outcomes (38,903), including live births, stillbirths (fetal deaths), miscarriages (spontaneous abortions), and elective terminations (induced abortions).

[c] Based on the estimated rate of live-birth outcomes in the U.S. (64.58%) [30], the original total of 38,903 is adjusted for this column to 25,124—the estimated number of *live* births.

[d] The majority of congenital heart defects reported on social media were not specified (e.g., *hole in heart*, *CHD*), so we aggregated all of the cardiovascular defects into a single class.

[e] The denominator used to calculate the prevalence of hypospadias in the U.S. is male live births.

[f] As an estimate of male births, the denominator used to calculate the prevalence of hypospadias reported on social media is half of the estimated number of total pregnancy outcomes.

[g] As an estimate of male births, the denominator used to calculate the prevalence of hypospadias reported on social media is half of the estimated number of live births.

[h] The denominator used to calculate the prevalence of Turner syndrome in the U.S. is female live births.

[i] As an estimate of female births, the denominator used to calculate the prevalence of Turner syndrome reported on social media is half of the estimated number of total pregnancy outcomes.

[j] As an estimate of female births, the denominator used to calculate the prevalence of Turner syndrome reported on social media is half of the estimated number of live births.

identified as potentially missed true positives by the methods described in Section 2.4.2. We identified 91 of them as false negatives—66 (11.96%) from the query modification approach (out of 552 tweets retrieved), and 25 (0.69%) from the "fuzzy matching" approach (out of 3617 tweets retrieved). Table 6 (in Section 4.2) provides examples of these false negatives. Thus, based on this evaluation, we calculated the recall of the query as 0.95 for the birth defects in this study.

4. Discussion

4.1. Principal results

Our study demonstrates that rare health-related events are indeed reported on Twitter, but also that utilizing social media for studying such events on a large scale requires, first, addressing the methodological challenges of collecting sparse data. Our rule-based, bootstrapping approach allowed us to collect an extensive set of tweets, while filtering out the excessive noise that would have prevented us from creating an annotated data set to be used for training machine learning algorithms. In this study, we used the annotated tweets to discover a cohort of women whose pregnancies with birth defect outcomes could be observed on Twitter for epidemiological analysis. Although mining the users' timelines to study the etiology of birth defects is beyond the scope of this paper, Table 5 provides examples of risk factors [31] reported in the timelines of the 195 users who met our inclusion criteria, highlighting the research opportunities ultimately enabled by our data collection approach. Tweets 1–5, which indicate Twitter users' tobacco and illicit drug use, illness, and medication intake [32], were posted in the prenatal period.

Among the 195 women identified as posting tweets during a pregnancy with a birth defect outcome, the majority were found to be active on Twitter even before they were pregnant or could have been aware of their pregnancy. Thus, our social media mining approach can enable a unique opportunity of observing risk factors in the periconceptional period and the early period of the first trimester, during which the users posted tweets 1 and 2 in Table 5. By allowing us to observe pregnancies from conception, social media mining may help shed light on how the selection bias associated with pregnancy exposure registries [8] has impacted the assessment of birth defect risks.

Because Twitter users also tend to post information after their pregnancy, as in tweet 7, social media mining can provide a cost-effective means of long-term follow-up after birth, which is usually cost prohibitive. Some of the birth defects that were detected in the present study were reported on Twitter more than a year following the postpartum period. By offering a longer-term follow-up period, social media mining may help provide insight on whether short-term follow-up periods have led pregnancy exposure registries to under-assess the risks of birth defects, considering that not all birth defects are present at birth [9].

Our data collection pipeline collects all public tweets of users who report a pregnancy on Twitter, so our database can provide a "ready-made" population from which to select internal comparator groups. By allowing us to compare pregnancies with and without reported birth defect outcomes, social media mining may help determine if the lack of internal comparator groups [8] has led pregnancy exposure registries to over-assess the risks of birth defects.

Our results suggest that congenital heart defects (CHDs) are the most common type of birth defect reported on Twitter, consistent with findings in the general population [28], in which nearly 1% of infants are born with CHDs [29]. CHDs are the leading cause of infant mortality due to birth defects [33], and the causes of CHDs remain largely

Table 5
Sample tweets containing information about birth defect risk factors, reported in the timelines of the 195 users who met our inclusion criteria. For ethical considerations, names are redacted and, in sensitive cases, the tweet text is modified in an effort to prevent the specific user from being identified.

	Tweet	Risk factor	Birth defect outcome
1	I needed a cigarette but it's so cold so I'm currently sitting in the truck chain smoking	tobacco use	gastroschisis
2	haven't smoked like this in a long time I'm too high rn	illicit drug use	Down syndrome; congenital heart defect
3	My first High Risk appt. today to get an in depth ultrasound of [name], and get some info on what to do about my high blood pressure.	hypertension	pyloric stenosis
4	Type 1 diabetes keeps things in check with frequent monitoring! Full day! Fetal monitoring for a non stress test than a growth ultrasound.	diabetes mellitus	hip dysplasia
5	I can't decide whether my migraine is from all my stress or from my [medication].	medication exposure	hydrocephalus
6	I was doing an online survey this morning and I had to enter into a new age demographic. #35	maternal age	Hirschsprung's disease
7	@[username] just watched you announce ur pregnant congratulations! This is my little man he has achondroplasia like me	family history of birth defect	achondroplasia

Table 6
Sample false negative tweets discovered by "fuzzy matching" and modifying the regular expressions. The bold text indicates the string that was ("fuzzily") matched by the corresponding lexicon entry/linguistic pattern listed. For "fuzzy matching," some false negative tweets (e.g., 4–6) contain birth defects that differ semantically from the lexicon entry/linguistic pattern on which the matching tweet was based. For ethical considerations, names are redacted and the tweet text is slightly modified.

	Tweet	Entry/Pattern
1	**#ISupportPottersSyndrome** #RestMyLittleAngel #[name]	potter syndrome
2	Our boy's having surgery for **#sagittalcraniosynostosis**	craniosynostosis
3	Excited to raise awareness for [name]!! **#hydranencephalyawarness**	anencephaly
4	my baby has a **hole in her** spine so it's medically recommended	hole in heart
5	[name] was **born without the** left ventricle in her heart.	born without toe
6	made appointment for my 22-month-old with **congenital kidney disorder**	congenital skin disorder
7	he had a twin but he didn't survive **he had a heart** problem:(	hole and a heart
8	the hospital ped told me my baby had **spin bifida** and needed testing	spina bifida
9	Pray for my friend who is in labor with her son that has **tristomy 18**	trisomy 18
10	my baby could possibly have a **whole in his heart**	hole in his heart

unknown [34]. Our social media mining approach, by directing us to all the publicly available tweets by users with birth defect outcomes, can provide a unique opportunity of exploring unknown causes of CHDs. In sum, there are a variety of ways in which social media mining could complement existing methods for studying birth defects.

This study verifies that a combination of automatic NLP and manual analysis methods can be used to collect data for events that are too sparse and noisy on social media for more basic data retrieval methods. Because our bootstrapping approach relies on generalizable techniques—a lexicon, lexical variants, regular expressions, post-processing, and manual analysis guided by distributional properties—it can be applied to studying other rare health-related events on social media. Our approach to harnessing social media data may be particularly valuable for studying health-related events that, like birth defects, have limited methods currently available.

4.2. Error analysis

To determine why some "defect" and "possible defect" tweets were not retrieved by the initial query, we conducted a brief error analysis of the false negative tweets discovered in our evaluation. Table 6 provides examples of true positives that were missed by the initial query. This error analysis provides methodological insight for expanding the query in the future development of automatic retrieval methods. The majority of the errors in the query can be attributed to the leading word boundary in the regular expression, which, while filtering out a significant amount of noise, caused the query to miss birth defects in hashtags in which the beginning of the lexicon entry was adjoined with words that the query was not formulated to match—for example, non-clinical words, such as *I support* in tweet 1 (*#ISupportPottersSyndrome*), clinical words, such as *sagittal* in 2 (*#sagittalcraniosynostosis*), or prefixes, such as *hydra* in 3 (*#hydranencephaly*). Removing the leading word boundary from the regular expressions (and re-adding it to the noisy lexicon entries) yielded a level of precision (0.12) that warrants its removal as an approach for expanding the query in future implementations.

Other sources of error include birth defects that were expressed colloquially or abstractly in ways that were not learned in the process of manually studying the data. Colloquial expressions tend to describe the birth defect, such as *hole in her spine* in tweet 4 and *born without the left ventricle in her heart* in 5. Abstract expressions indicate a type of birth defect without specifying the concrete problem, such as *congenital kidney disorder* in 6 and *heart problem* in 7. Interestingly, 4, 5, and 6, discovered by "fuzzy matching," refer to birth defects that differ semantically from the lexicon entry/linguistic pattern on which the matching tweets were based. Finally, some of the errors were caused by misspellings that were not accounted for in our semi-automatic process of generating lexical variants, such as *spin bifida* in 8, *tristomy 18* in 9, and *whole in his heart* in 10. Although we did discover tweets 4–10 through "fuzzy matching," this approach's low level of precision (0.01) makes it unsuitable for automatically expanding the query, as it would result in too much noise. For now, we will manually add such matches to our dictionary of variants.

4.3. Limitations

The results of this study, while promising, do point to some methodological limitations. While the high recall of our query (0.95) in part reflects our extensive retrieval approach, it may also reflect the challenge of detecting non-clinical expressions of birth defects, or health-related events in general, in social media. Our comparison of prevalence for selected birth defects between social media and the general population suggests that, despite the high recall, we may be falling methodologically short in our retrieval of tweets that mention particular birth defects. Our efforts to identify false negatives were limited to discovering *lexically* similar tweets, so we may be mostly "finding what

we are looking for," and there may remain latent indications of birth defects on social media that would require more advanced text mining methods to detect. This is a common limitation of rule-based, lexical-matching approaches to data collection. Alternatively, if we are to assume that the recall score is accurate, then social media users may be under-reporting such rare health-related events on Twitter [35]—an issue in pregnancy exposure registries as well [8]. Despite possible under-reporting and the relatively small cohort discovered in this initial study, additional users are being constantly added to the database over time. Thus, automating the identification of birth defect cases on social media is deemed essential for large-scale epidemiological research, which we will address in future work.

5. Conclusions

In this paper, we presented (i) evidence that rare health-related events—in this case, birth defects—are reported on Twitter, (ii) a generalizable, systematic NLP approach to collecting sparse tweets for manual annotation, (iii) a semi-automatic method to identify undetected tweets (false negatives), and (iv) a collection of publicly available Twitter timelines of pregnant users with birth defect outcomes, which could be used for future epidemiological analysis. In future work, we will seek to address the methodological challenges of expanding the query and utilizing the annotated data to train machine learning algorithms to automatically identify users reporting birth defect outcomes on social media. In deploying machine learning algorithms on unlabeled tweets, our rule-based data collection approach would serve as a fully-automated pre-filtering module in an end-to-end social media pipeline. In general, the ability to prepare such an annotated data set, in the face of sparsity and noise, enables the training and deployment of machine learning algorithms for the large-scale use of social media mining as a complementary method for studying rare health-related events, which may have limited methods currently available.

Conflict of interest

The authors declared that there is no conflict of interest.

Acknowledgements

The University of Pennsylvania's Institutional Review Board (IRB) determined this study to be "exempt" as human subjects research. This work was funded in part by AbbVie Inc., and by the National Institutes of Health (NIH) National Library of Medicine (NLM) award number R01LM011176. The content is solely the responsibility of the authors, and does not necessarily represent the views of the NIH, NLM, or AbbVie Inc. The authors would like to acknowledge AbbVie Inc. for reviewing a completed version of this manuscript. AbbVie Inc. was not otherwise involved in preparing this manuscript or conducting the research reported in this manuscript. The authors would like to acknowledge Karen O'Connor and Alexis Upshur for their efforts in the annotation and analysis processes.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2018.10.001.

References

[1] T.J. Mathews, M.F. MacDorman, M.E. Thoma, Infant mortality statistics from the 2013 period linked birth/infant death data set, Natl. Vital. Stat. Rep. 64 (9) (2015) 2000–2013 PMID:24979974.
[2] M.L. Feldkamp, J.C. Carey, J.L.B. Byrne, S. Krikov, L.D. Botto, Etiology and clinical presentation of birth defects: population based study, BMJ 357 (2017), https://doi.org/10.1136/bmj.j2249 PMID:28559234.
[3] M.C. Blehar, C. Spong, C. Grady, S.F. Goldkind, L. Sahin, J.A. Clayton, Enrolling pregnant women: issues in clinical research, Women's Health Issues 23 (1) (2013) e39–e345, https://doi.org/10.1016/j.whi.2012.10.003 PMID:23312713.
[4] R.I. Hartman, A.B. Kimball, Performing research in pregnancy: challenges and perspectives, Clin. Dermatol. 34 (3) (2016) 410–415, https://doi.org/10.1016/j.clindermatol.2016.02.014 PMID:27265080.
[5] R.M. Ward, Difficulties in the study of adverse fetal and neonatal effects of drug therapy during pregnancy, Semin. Perinatol. 25 (3) (2001) 191–195, https://doi.org/10.1053/sper.2001.24567 PMID:11453616.
[6] D.L. Kennedy, K. Uhl, S.L. Kweder, Pregnancy exposure registries, Drug Saf. 27 (4) (2004) 215–228, https://doi.org/10.2165/00002018-200427040-00001 PMID:15003034.
[7] U.S. Department of Health and Human Services, Food and Drug Administration. Reviewer Guidance: Evaluating the Risks of Drug Exposure in Human Pregnancies, 2005. < https://www.fda.gov/downloads/Drugs/%E2%80%A6/Guidances/ucm071645.pdf > (accessed 2017 December 13).
[8] S. Sinclair, M. Cunnington, J. Messenheimer, J. Weil, J. Cragan, R. Lowensohn, M. Yerby, P. Tennis, Advantages and problems with pregnancy registries: observations and surprises throughout the life of the International Lamotrigine Pregnancy Registry, Pharmacoepidemiol. Drug Saf. 23 (8) (2014) 779–786, https://doi.org/10.1002/pds.3659 PMID:4406353.
[9] R.E. Gliklich, N.A. Dreyer, M.B. Leavy, Registries for Evaluating Patient Outcomes: A User's Guide, third ed., Agency for Healthcare Research and Quality, Rockville, MD, 2014.
[10] A. Sarker, P. Chandrashekar, A. Magge, H. Cai, A. Klein, G. Gonzalez, Discovering cohorts of pregnant women from social media for safety surveillance and analysis, J. Med. Internet. Res. 19 (10) (2017) e361, , https://doi.org/10.2196/jmir.8164 PMID:29084707.
[11] S. Greenwood, A. Perrin, M. Duggan, Social Media Update 2016, 2016. < http://www.pewinternet.org/2016/11/11/social-media-update-2016/ > (accessed 2017 December 13).
[12] M.J. Paul, A. Sarker, J.S. Browenstein, A. Nikfarjam, M. Scotch, K.L. Smith, G. Gonzalez, Social media mining for public health monitoring and surveillance, in: Proceedings of the Pacific Symposium on Biocomputing; 2016 January 4-8, Hawaii, United States, 2016, pp. 468–479. < https://psb.stanford.edu/psb-online/proceedings/psb16/intro-smm.pdf > .
[13] A. Alessa, M. Faezipour, A review of influenza detection and prediction through social networking sites, Theor. Biol. Med. Model. 15 (1) (2018) 2, https://doi.org/10.1186/s12976-017-0074-5 PMID:29386017.
[14] H.W. Meng, S. Kath, D. Li, Q.C. Nguyen, National substance use patterns on Twitter, PLoS One 12 (11) (2017) e0187691, https://doi.org/10.1371/journal.pone.0187691 PMID:29107961.
[15] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, G. Gonzalez, Utilizing social media data for pharmacovigilance: a review, J. Biomed. Inform. 54 (2015) 202–212, https://doi.org/10.1016/j.jbi.2015.02.004 PMID:25720841.
[16] L. Rynn, J. Cragan, A. Correa, Update on overall prevalence of major birth defects: Atlanta, Georgia, 1978–2005, MMWR Morb. Mortal Wkly. Rep. 57 (1) (2008) 1–5 PMID:18185492.
[17] National Birth Defects Prevention Network, Guidelines for Conducting Birth Defects Surveillance, 2004. < https://www.nbdpn.org/docs/NBDPN_Guidelines2012.pdf > (accessed 2017 December 13).
[18] Metropolitan Atlanta Congenital Defects Program, Executive summary, Birth Defects Res. A Clin. Mol. Teratol. 79 (2) (2007) 66–93, https://doi.org/10.1002/bdra.20351.
[19] J.E. Fornoff, T. Shen, Birth Defects and Other Adverse Pregnancy Outcomes in Illinois 2005-2009: A Report on County-Specific Prevalence, 2013. < http://www.dph.illinois.gov/sites/default/files/publications/ers14-03-birth-defects-inillinois-2005-2009-041516.pdf > (accessed 2017 December 18).
[20] EUROCAT, Guide 1.4: Instruction for the Registration of Congenital Anomalies, 2013. < http://www.eurocat-network.eu/content/Section%203.3-%2027_Oct2016.pdf > (accessed 2017 December 18).
[21] U.S. National Library of Medicine, UMLS Reference Manual, 2009. < https://www.ncbi.nlm.nih.gov/books/NBK9676/pdf/Bookshelf_NBK9676.pdf > (accessed 2017 December 18).
[22] A. Sarker, G. Gonzalez-Hernandez, An unsupervised and customizable misspelling generator for mining noisy health-related text sources. < https://arxiv.org/abs/1806.00910 > .
[23] A. Sarker, G. Gonzalez, A corpus for mining drug-related knowledge from Twitter chatter: language models and their utilities, Data Brief 10 (2017) 122–131, https://doi.org/10.1016/j.dib.2016.11.056.
[24] C.T. Mai, C.H. Cassell, R.E. Meyer, J. Isenburg, M.A. Canfield, R. Rickard, R.S. Olney, E.B. Stallings, M. Beck, S. Shahrukh Hasmi, S.J. Cho, R.S. Kirby, Birth defects data from population-based birth defects surveillance programs in the United States, 2007 to 2011: highlighting orofacial clefts, Birth Defects Res. A Clin. Mol. Teratol. 100 (11) (2014) 895–904, https://doi.org/10.1002/bdra.23329.
[25] M. Rouhizadeh, A. Magge, A. Klein, A. Sarker, G. Gonzalez, A rule-based approach to determining pregnancy timeframe from contextual social media postings, Proceedings of the Eighth International Conference on Digital Health; 2018 April 23–26, Association for Computing Machinery, Lyon, France, 2018, pp. 16–20, , https://doi.org/10.1145/3194658.3194679.
[26] A.X. Chang, C.D. Manning, SUTIME: a library for recognizing and normalizing time expressions, Proceedings of the Eighth International Conference on Language Resources and Evaluation; 2012 May 21-27, European Language Resources Association, Istanbul, Turkey, 2012, pp. 3735–3740.
[27] A.J. Viera, J.M. Garrett, Understanding interobserver agreement: the kappa

statistic, Fam. Med. 37 (5) (2005) 360–363 PMID:15883903.
[28] C.T. Mai, J. Isenburg, P.H. Langlois, C.J. Alverson, S.M. Gilboa, R. Rickard, M.A. Canfield, S.B. Anjohrin, P.J. Lupo, D.R. Jackson, E.B. Stallings, A.E. Scheuerle, R.S. Kirby, Population-based birth defects data in the United States, 2008–2012: presentation of state-specific data and descriptive brief on variability of prevalence, Birth Defects Res. A Clin. Mol. Teratol. 103 (11) (2015) 972–993, https://doi.org/10.1002/bdra.23461.
[29] Md. Reller, M.J. Strickland, T. Riehle-Colarusso, W.T. Mahle, A. Correa, Prevalence of congenital heart defects in Metropolitan Atlanta, 1998–2005, J. Pediatr. 153 (6) (2008) 807–813, https://doi.org/10.1016/j.jpeds.2008.05.059 PMID:18657825.
[30] S.J. Ventura, S.C. Curtin, J.C. Abma, S.K. Henshaw, Estimated pregnancy rates and rates of pregnancy outcomes for the United States, 1990–2008, Natl. Vital Stat. Rep. 60 (7) (2012) 1–21 PMID:22970648.
[31] B.S. Harris, K.C. Bishop, H.R. Kemeny, J.S. Walker, E. Rhee, J.A. Kuller, Risk factors for birth defects, Obstet. Gynecol. Surv. 72 (2) (2017) 123–135, https://doi.org/10.1097/OGX.0000000000000405 PMID:28218773.
[32] A. Klein, A. Sarker, M. Rouhizadeh, K. O'Connor, G. Gonzalez, Detecting personal medication intake in Twitter: an annotated corpus and baseline classification system, Proceedings of the BioNLP 2017 Workshop; 2017 Aug 4, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 136–142.
[33] Q. Yang, H. Chen, A. Correa, O. Devine, T.J. Mathews, M.A. Honein, Racial differences in infant mortality attributable to birth defects in the United States, 1989–2002, Birth Defects Res. A Clin. Mol. Teratol. 76 (10) (2006) 706–713, https://doi.org/10.1002/bdra.20308 PMID:17022030.
[34] K.J. Jenkins, A. Correa, J.A. Feinstein, L. Botto, A.E. Britt, S.R. Daniels, M. Elixson, C.A. Warnes, C.L. Webb, Noninherited risk factors and congenital cardiovascular defects: current knowledge, Circulation 115 (23) (2007) 2995–3014, https://doi.org/10.1161/CIRCULATIONAHA.106.183216 PMID:17519397.
[35] S. Golder, G. Norman, Y.K. Loke, Systematic review on the prevalence, frequency and comparative value of adverse events data in social media, Br. J. Clin. Pharmacol. 80 (4) (2015) 878–888, https://doi.org/10.1111/bcp.12746 PMID:26271492.