CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

ARTIGO
ARTICLE

Mining in Twitter for adverse events from malaria drugs: the case of doxycycline

O uso do Twitter como minerador de eventos adversos de medicamentos de combate à malária: o caso da doxiciclina

El uso de Twitter como localizador de eventos adversos con medicamentos de combate a la malaria: el caso de la doxiciclina

Felipe Vieira Duval [1]
Fabrício Alves Barbosa da Silva [1]

doi: 10.1590/0102-311X00033417

Abstract

During the post-marketing period, when medicines are used by large population contingents and for longer periods, unexpected adverse events (AE) can occur, potentially altering the drug's risk-benefit ratio enough to demand regulatory action. AE are health problems that can occur during treatment with a pharmaceutical product, which in the drug's post-marketing period can require a significant increase in health care and result in unnecessary and often fatal harm to patients. Therefore, a key objective for the health system is to identify AE as soon as possible in the post-marketing period. Some countries have pharmacovigilance systems responsible for collecting voluntary reports of post-marketing AE, but studies have shown that social networks can be used to obtain more and faster reports. The current project's main objective is to build a totally automated system using Twitter as a source to detect both new and previously known AE and conduct the statistical analysis of the resulting data. A system was thus built to collect, process, analyze, and assess tweets in search of AE, comparing them to U.S. Food and Drug Administration (FDA) data and the reference standard. The results allowed detecting new and existing AE related to the drug doxycycline, showing that Twitter can be useful in pharmacovigilance when employed jointly with other data sources.

Drug and Narcotic Control; Biological Ontologie; Natural Language Processing; Social Media; Database

Correspondence
F. V. Duval
Rua 35, Qd 73, Condomínio Colinazul nº 7, Niterói, RJ 24342-086, Brasil.
felipeduval@gmail.com

[1] *Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.*

Introduction

During the post-marketing period, when medicines are used by large population contingents and for longer periods, adverse events (AE) can occur that can alter the drug's risk-benefit ratio enough to require regulatory action. AE are defined as health problems that can emerge in a user or patient during treatment with a pharmaceutical product, potentially resulting from medication errors, deviation in the drugs' quality, adverse drug reactions (ADR), drug-drug interactions, and intoxications [1].

According to the World Health Organization (WHO), pharmacovigilance is defined as *"as the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem"* [2]. Pharmacovigilance is responsible for identifying, assessing, and monitoring the occurrence of drug-related AE, with the aim of guaranteeing that the benefits outweigh the risks caused by them [1]. To achieve this objective, the main instrument in pharmacovigilance is spontaneous reporting, informing government agencies on AE that have occurred with the drugs' use.

In Brazil, pharmacovigilance activities are shared by the state and municipal health surveillance agencies and the Brazilian Health Regulatory Agency (Anvisa) [2,3]. The rate of AE reports received by Anvisa is low [4], often far lower than the target proposed by the international literature, which suggests 300 reports per million inhabitants [5]. It is thus necessary to use other sources to detect AE.

AE can be identified during the drug's study phase prior to marketing, known as the clinical phase. Clinical tests occur in three distinct phases, known as phases I, II, and III, conducted with healthy volunteers and a limited number of patients. In addition, patient selection and treatment generally differ from actual clinical practice [6,7]. AE detected later, in the post-marketing period (also known as phase IV), may require a significant increase in health care and result in unnecessary and often fatal harm to patients [8]. Therefore, the discovery of AE as soon as possible in the post-marketing period is a key objective for health systems and especially for pharmacovigilance systems.

Computational methods commonly referred to as "signal detection" allow drug safety evaluators to analyze large data volumes to identify risk signals for potential AE, and also serve as an essential component of pharmacovigilance. For example, the U.S. Food and Drug Administration (FDA) routinely uses a signal tracking process to calculate statistics, reporting associations for all the millions of drug combinations and events in its system for communicating AE [8]. These signals alone are not sufficient to establish a causal relationship, but they are considered early warnings that require in-depth assessment by specialists to establish causality.

Dedicated resources for subsequent assessment of each of the multiple signals normally generated by detection algorithms is not feasible. Resources deployed for false leads can undermine a pharmacovigilance system [9]. Automated strategies are thus imperative to reduce the amounts of false-positives and set priorities in order to allow assessing only the most promising signals.

The article's main contribution is thus the proposal for TweetAEMiner (Tweet Adverse Event Miner), an automated pharmacovigilance system capable of identifying new and existing drug-AE associations with the use of text mining.

Text mining consists of techniques to retrieve textual information, extract information, and process natural language with algorithms and methods for discovering knowledge, data mining, and machine learning [10].

Twitter was used in the current project as a text mining source. It is an unconventional database due to greater ease and speed in accessing its data. Examples of other unconventional databases that have been used recently in epidemiological surveillance are search logs [11,12,13] and social networks [14,15].

Most of the previous studies on text mining in pharmacovigilance have focused on electronic health records and medical case reports [16,17]. Harpaz et al. [18] provide an in-depth study on the existing approaches to the post-marketing phase, exploring various resources such as electronic health records and spontaneous AE reporting systems. Social networks have also been used recently for this purpose. Leaman et al. [19] analyzed users' comments in social networks and showed that they contain information on medicines that can be extracted for subsequent analysis. In a recent study, Yates & Goharian [20] analyzed the value of users' commentary in revealing unknown AE, assessing ADR extracted from the SIDER database (http://sideeffects.embl.de/), which contains information on known AE.

Most studies that use Twitter as a data source and that focus on the medical field seek information other than AE. Some studies have used Twitter for this purpose [22,23,24] and have shown that the use of tweets can lead to real-time pharmacovigilance. Freifeld et al. [23] used Twitter to assess the level of agreement between tweets that mentioned AE (Proto-AE – posts with resemblance to AE) and spontaneous reports from the FDA Adverse Event Reporting System (FAERS). The study used 6.9 million tweets with the names of drugs, of which 4,401 were identified as Proto-AEs and showed that Twitter had almost three times more Proto-AE than the FDA reports [23].

Studies that search for AE in Twitter generally collect data from just a few months to find known ADR, use one or no ontology (a data model that represents a set of concepts and relationships within a domain) to do so, and have manual stages in the pipeline (a sequence of operations in which the exit from one stage/operation serves as the entry to the next operation in the sequence). This article uses an automatic pipeline for collecting, storing, and processing tweets that use a complete ontology totally focused on the search for ADR.

Due to limitations on the number of words that can be searched for in Twitter, this study focused on ADR from drugs for malaria, which was the neglected disease with the most tweets in 2014. Among these drugs, an analysis was done of AE related to doxycycline as found in tweets and compared to consolidated AE reports received by the FDA. However, the system described in this article can be adapted to monitor multiple diseases and drugs simultaneously.

Materials and methods

TweetAEMiner collects tweets continuously using Twitter's API (application programming interface) with predetermined words (diseases or drugs). These tweets are stored in the database. The system periodically initiates the tweets' processing and analysis. The system is currently configured for processing and analysis on Sundays, when a new week begins on the epidemiological calendar [25], but this periodicity can be altered easily if necessary. The tweets are processed with a natural language processor (NLP), and the data output from this processing is submitted to statistical analysis. Finally, the results are assessed against a reference standard.

The system generates a list of specific signals that are assessed against a reference standard. One signal corresponds to a "drug-AE" association identified by the pipeline.

Figure 1 shows the four stages in the pipeline: data extraction, processing, analysis, and assessment. Besides the stages, Figure 1 also shows the database used to store the tweets and the reference standard.

Extraction

Twitter has two API to collect tweets: REST API (http://dev.twitter.com/rest/public) and Streaming API (http://dev.twitter.com/docs/api/streaming). The two API only allow access to recent tweets, so those actually collected will be useful for future research. The material has been collected since early 2014 using the above-mentioned API.

As an initial approach, tweets were collected that were related to neglected diseases such as malaria, dengue, Chagas disease, tuberculosis, and leishmaniasis [26]. The queries were later expanded to other diseases, also including non-neglected diseases such as AIDS.

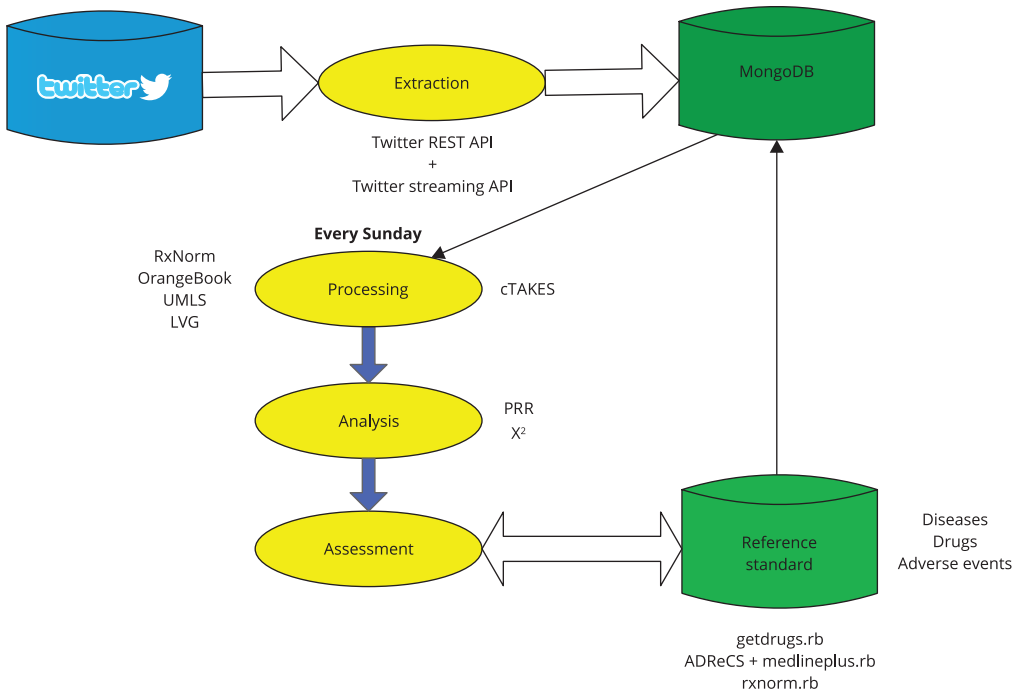
A preliminary analysis of the collected data indicated that malaria was the disease with the most tweets. Although some of these diseases still lack an associated drug, the messages referring to them may be useful in other projects, as for example in epidemiological studies.

Given the limited number of words that can be searched for in the respective social network, we only collected tweets on drugs used to treat malaria.

The website http://www.drugs.com was used to obtain the names of drugs related to malaria. The site allows finding names of both brand names and generic drugs. To facilitate the search for these data, a program was developed that relates the associated drugs to the name of each disease. Nineteen drugs were used for malaria, of which 10 were brand name drugs (Plaquenil, Malarone, Doryx, Lar-

Figure 1

TweetAEMiner methodology.



Note: system's pipeline. Yellow shows the four stages in the process; green shows the databases used to store the tweets and as the reference standard; blue shows Twitter.

iam, Daraprim, Aralen, Fansidar, Morgidox, Ocudox, and Oraxyl) and 9 were generics (atovaquone, proguanil, doxycycline, mefloquine, pyrimethamine, sulfadoxine, hydroxychloroquine, chloroquine, and primaquine). Among these drugs, the one with the most tweets in 2014 was doxycycline, as shown in Table 1, and was thus chosen as the target for analysis.

TweetAEMiner was developed to allow the pipeline's portability to other types of texts besides tweets, with a minimum of effort. Suffice it to adjust the extraction component to some text source other than Twitter.

Reference standard

The reference standard was developed to be a widely accepted database with all the currently known AE. This meant mainly using *Adverse Drug Reaction Classification System* (ADReCS) [27], an ontology of terms for adverse reactions that uses medical sources. A linkage between diseases and their drugs was added to this ontology.

These sources were used to create a database with the target diseases, the drugs used in their treatment, and each one's AE.

At present, only tweets in English are being processed, since all the sources used in the reference standard consist exclusively of words in English.

Table 1

Numbers of tweets citing drugs used to treat malaria in 2014.

Drugs	n
Morgidox	0
Ocudox	0
Oraxyl	0
Daraprim	35
Sulfadoxine	61
Proguanil	98
Aralen	122
Doryx	173
Atovaquone	191
Fansidar	193
Pyrimethamine	216
Primaquine	359
Lariam	671
Hydroxychloroquine	819
Malarone	890
Plaquenil	982
Mefloquine	1,312
Chloroquine	2,912
Doxycycline	14,333

Processing

After extraction, the tweets are submitted to a NLP. Various NLP are used in medicine, such as Medlee [28], cTAKES [29], and MetaMap [30]. cTAKES was chosen as an open code NLP used to extract information from free text, using different vocabularies from various medical sources.

cTAKES is used in a program that processes stored tweets, generating as output diseases, drugs, and the associated adverse reactions as well as other medical information found in the text.

Although TweetAEMiner uses tweets rather than spontaneous reports, the messages are filtered in order to have at least a drug and an AE, discarding those without them. The approach is similar to that of Proto-AE by Freifeld et al. [23].

This study uses a drug-based approach [31], chosen because we did not know the number of tweets with a given AE, as well as to determine the number of tweets with AEs and the drugs related to the target disease. With this approach, it is more appropriate to consider a tweet with the drug's name than to collect any tweet that may not be related to drugs.

Analysis

After processing the tweets, a measure of disproportionality analysis is used for the data to be analyzed. Disproportionality analysis (DPA) in pharmacovigilance is the main class of analytical methods for spontaneous reporting systems (SRS) [18]. SRS are reports that include one or more drugs, with one or more AE, and possibly some basic demographic data. These methods identify relevant associations in SRS databases, with a focus on projections of low data dimensionality, more specifically 2x2 contingency tables. Both the FDA and WHO use DPA methods to find these associations [18]. This measure was used to classify drug-AE pairs identified in the previous processing stage. The analytical method can vary according to the data that are processed. SRS based on ADR most frequently perform signal detection using disproportionality measures.

The basic task for a DPA method is classification of the tables in order of "interest". Different DPA methods focus on different statistical measures of association as their measure of "interest". Table 2 presents the formulas for the most commonly used measures of association, together with their probabilistic interpretation, in which "¬drug" denotes the reports that do not include the target drug.

A particular drug that causes a specific AE more than any other will normally have the highest measure of association. If an AE and a drug are stochastically independent, the measure of association receives a value of 1. Since each AE from an individual drug occurs in a small proportion of all the reports, we generally have a << b or a << c and c << d, and in practice these measures tend to have identical values and interpretations. For example, a value of 3 indicates that there are three times more reports involving a drug-AE pair than expected if there were no association between the two [32].

The associations are calculated using the frequentist approach proportional reporting ratio (PRR) for disproportionality analysis. Bayesian measures tend to produce extreme values that are less extreme than PRR when the number of cases is very small. However, when the sensitivity, specificity, and predictive power of these measures were compared using Dutch data in 2002 [33], no important differences were found when at least three cases were reported. In addition, PRR has already been used in various studies to detect ADE in spontaneous reporting systems, [32,34,35] and it is one of the principal measures used in the European Union. Together with PRR, the 95% confidence interval (95%CI) was calculated and the χ^2 test was performed to validate the signals generated, as is performed by the SRS used by the European Union, called EudraVigilance [34].

Assessment

TweetAEMiner verifies in the data analysis whether there was some signal (a "drug-AE" association) as in EudraVigilance, calculating the measure of disproportionality, PRR, together with its 95%CI and the use of the χ^2 test.

Since PRR is a highly sensitive method, it can generate many false positives, especially if the number of reports is low. To reduce this, one of the criteria used is to calculate the 95%CI.

The 95%CI for the Napierian logarithm of PRR is estimated as , in which "se" is the standard error of the mean of the natural logarithm of PRR [33,36]. If PRR is shown with the 95%CI, it will be considered a disproportionality signal when [34]: lower limit of the interval ≥ 1; number of cases ≥ 3.

Table 2

Common measures of association in spontaneous reporting systems (SRS) analyses.

Measure of association	Formula	Value	Probabilistic interpretation
Relative reporting ratio (RRR)	$\frac{t \,.\, a}{m \,.\, n}$	35.57355	$\frac{Pr(AE \mid drug)}{Pr(AE)}$
Proportional reporting ratio (PRR)	$\frac{(a(t-n))}{c \,.\, n}$	37.36421	$\frac{Pr(AE \mid drug)}{Pr(AE \mid drug)}$
Reporting odds ratio (ROR)	$\frac{a \,.\, d}{c \,.\, b}$	37.57431	$\frac{\frac{Pr(AE \mid drug)}{Pr(AE \mid drug)}}{\frac{Pr(AE \mid drug)}{Pr(AE \mid drug)}}$
Information component	$\log_2$(RRR)	5.12573	$\log_2 \frac{Pr(AE \mid drug)}{Pr(AE)}$

AE: adverse events.

Note : the letters "a", "b", "c", and "d" are values from the 2x2 contingency table for a drug and an AE. The letters "m", "n", and "t" are sums, as exemplified in Duval et al. [26].

Another signal detection measure used together with PRR is the χ^2 statistic, a test of independence of categorical variables used as an alternative measure of the contingency table's heterogeneity with a drug D and an AE [34].

If PRR is shown with the χ^2, it will be considered a disproportionality signal when: PRR ≥ 2; $\chi^2 \geq 4$; number of cases ≥ 3.

Besides analysis of the tweets, the FDA data were also analyzed to compare the signals generated in the two. The signals detected in each of the analyses were grouped in three types:
(a) Type A: generated by the criterion of the confidence interval for PRR, that is, when the lower limit of the 95%CI for PRR is greater than or equal to 1 and the number of tweets/reports is greater than or equal to 3;
(b) Type B: generated by the χ^2 criterion, that is, PRR ≥ 2; $\chi^2 \geq 4$ and the number of tweets/reports is ≥ 3;
(c) Type C: when there are both type A and B signals.

Results

One of the article's main results was the development of an automatic tool to collect and analyze AE in Twitter. Among the 19 malaria drugs that were used to filter the tweets, doxycycline yielded the most messages, as shown in Table 1, and was thus chosen for the analysis. Assessment of the results included a comparison of the analysis of data obtained by the TweetAEMiner and FDA data obtained by the htpps://open.fda.gov website.

Analysis of Twitter data

Calculation in the disproportionality analysis used the PRR measure, only considering the tweets that cited some AE. The synonyms for ADR in the ADReCS were also used in the count to build the contingency tables.

Table 3 shows the PRR report for the drug doxycycline with the drug's known AE in the reference standard and which had at least one tweet.

In some situations, when the number of tweets with the target drug and AE in question is greater than zero and the number of tweets with the AE but without the target drug is equal to zero, the PRR cannot be calculated. It is thus arbitrarily assigned "99.9" in the "PRR" column in Table 3 to reflect the presence of a possible signal. In these cases, the limits of the confidence interval are not calculated, as can be seen in the columns "PRR(-)" and "PRR(+)".

Signals were detected for two possible new AEs: alopecia and rosacea. Both also appear in the FDA data in the same period, as shown in Table 4. In the FDA, more than 200 AE are reported.

Analysis of FDA data

Analysis of the FDA data is done in the same way as with Twitter, but using the FDA reports during the same period with the 19 drugs.

Unlike Twitter, the drug with the most reports in FDA was hydroxychloroquine. Oraxyl was the only drug with no reports in 2014 (Table 5). Since the reports focus specifically on the detection of AE, it is normal for their analysis to produce a large number of signals. Doxycycline, for example, was reported with more than 200 different AE, 138 of which generated signals.

Generation of type A, B, and C signals

No type A signals were generated by Twitter. The FDA generated a total of 51 type A signals, 40 of which are not in the reference standard. The 11 AE of signals that were in the reference standard are abdominal pain, discomfort, hypersensitivity, malaise, muscle spasms, myalgia, nausea, rash, erythematous rash, urticaria, and vomiting.

Two type B signals were generated by Twitter for the AE upper abdominal pain and tension, both present in the reference standard. Two other type B signals were also generated that are not in the

Table 3

Proportional reporting ratio (PRR) report for adverse events (AE) with the drug doxycycline (Twitter).

AE	PRR(-) *	PRR **	PRR(+) ***	χ^2	Tweets	FDA #
Abdominal discomfort	Not calculated	99.9	Not calculated	2.356	11	
Abdominal distension	Not calculated	99.9	Not calculated	1.071	5	
Abdominal pain upper	Not calculated	99.9	Not calculated	6.434	30	
Abscess	Not calculated	99.9	Not calculated	0.428	2	
Anaemia	0.197	1.634	13.568	0.166	6	
Anaphylactic reaction	0.038	0.272	1.933	1.529	2	YES
Angioedema	0.108	0.233	0.504	12.807	12	
Anorexia	0.017	0.272	4.353	0.764	1	
Anxiety	1.812	4.466	11.007	10.022	82	YES
Aphthous stomatitis	1.079	4.493	18.716	4.035	33	
Arthralgia	0.314	0.953	2.894	0.006	14	
Back pain	0.427	0.657	1.012	2.897	70	
Blood pressure increased	Not calculated	99.9	Not calculated	2.356	11	
Bronchitis	0.113	0.272	0.654	7.651	10	
Candidiasis	Not calculated	99.9	Not calculated	2.999	14	
Cough	0.567	1.634	4.706	0.664	24	
Decreased appetite	Not calculated	99.9	Not calculated	0.428	2	
Dermatitis	Not calculated	99.9	Not calculated	0.214	1	YES
Diarrhoea	0.214	0.681	2.169	0.336	10	
Discomfort	0.017	0.272	4.353	0.764	1	YES
Dyspepsia	Not calculated	99.9	Not calculated	0.642	3	
Dysphagia	0.055	0.272	1.349	2.293	3	
Ear infection	0.113	0.272	0.654	7.651	10	
Emotional distress	Not calculated	99.9	Not calculated	0.214	1	YES
Fungal infection	1.417	4.539	14.543	6.159	50	
Gingivitis	Not calculated	99.9	Not calculated	0.214	1	
Haemolytic anaemia	0.017	0.272	4.353	0.764	1	
Headache	0.165	0.327	0.648	8.937	18	
Hypersensitivity	0.482	0.754	1.179	1.211	72	
Hypertension	Not calculated	99.9	Not calculated	1.713	8	
Infection	2.664	4.341	7.076	32.958	271	
Inflammation	Not calculated	99.9	Not calculated	1.499	7	
Influenza	0.229	0.256	0.285	528.852	557	
Injury	0.172	0.363	0.767	6.032	16	YES
Insomnia	0.088	0.182	0.377	20.92	12	
Intracranial pressure increase	Not calculated	99.9	Not calculated	0.214	1	
Leukopenia	Not calculated	99.9	Not calculated	0.428	2	
Malaise	Not calculated	99.9	Not calculated	0.856	4	YES
Muscle spasms	Not calculated	99.9	Not calculated	2.356	11	YES
Myalgia	0.085	0.817	7.852	0.024	3	
Nasal congestion	Not calculated	99.9	Not calculated	0.214	1	
Nasopharyngitis	0.009	0.091	0.872	5.37	1	
Nausea	0.943	3.949	16.54	3.245	29	
Oedema	0.009	0.091	0.872	5.37	1	
Oesophageal ulcer	Not calculated	99.9	Not calculated	0.642	3	YES
Oesophagitis	Not calculated	99.9	Not calculated	0.642	3	
Oropharyngeal pain	0.039	0.163	0.683	6.316	3	
Pain	1.556	2.465	3.905	12.485	181	

(continues)

Table 3 (continued)

AE	PRR(-) *	PRR **	PRR(+) ***	χ^2	Tweets	FDA #
Photosensitivity reaction	Not calculated	99.9	Not calculated	1.928	9	YES
Pigmentation disorder	0.049	0.545	6.005	0.199	2	
Rash	0.974	2.451	6.17	3.048	45	
Rhinorrhoea	Not calculated	99.9	Not calculated	0.214	1	
Sinusitis	0.172	0.272	0.432	27.638	36	
Stevens-Johnson syndrome	0.036	0.091	0.229	32.26	6	
Stomatitis	Not calculated	99.9	Not calculated	0.428	2	
Swelling	1.383	10.076	73.414	6.272	37	
Tension	Not calculated	99.9	Not calculated	4.716	22	
Thrombocytopenia	Not calculated	99.9	Not calculated	0.428	2	
Tooth abscess	0.038	0.272	1.933	1.529	2	
Toothache	Not calculated	99.9	Not calculated	1.499	7	
Ulcer	Not calculated	99.9	Not calculated	3.643	17	
Urticaria	0.064	0.117	0.213	55.349	15	
Vomiting	Not calculated	99.9	Not calculated	3.428	16	YES

FDA: U.S. Food and Drug Administration.

Note: When a signal is detected by χ^2, the cell is filled in red; when a signal is detected by the 95% confidence interval (95%CI) for PRR, the cell is filled in orange. The FDA column is filled in green if the signal appeared in both Twitter and FDA.

* Lower limit of the 95%CI for PRR;

** PRR value for the AE;

*** Upper limit of the 95%CI for PRR;

Shows that there was a signal for this AE in FDA in the same period of 2014.

reference standard for the AEs: alopecia and rosacea. Of these signals, only rosacea also occurred in the FDA data, which had a total of 24 type B signals, of which only menorrhagia is found in the reference standard.

Twitter generated a total of six type C signals for AEs: anxiety, aphthous stomatitis, fungal infection, infection, pain, and swelling. All are present in the reference standard of AE for doxycycline. Of these signals, only anxiety occurred in the FDA data, which had a total of 63 signals, eight of which were present in the reference standard: anaphylactic reaction, anxiety, dermatitis, emotional distress, injury, esophageal ulcer, photosensitivity reaction, and maculopapular rash, plus another 55 signals that are not found in the reference standard.

Discussion

In order to build a system capable of collecting, storing, and processing tweets related to drugs, a collector was first implemented using the API from Twitter itself. Since this API does not allow the acquisition of old messages, TweetAEMiner is already collecting tweets citing various drugs and diseases that were not the target of this article, but can be useful in future studies.

The disease with the most tweets was dengue, but since there are no drugs to treat it, the test study for the tool focused on drugs for malaria, the disease with the second most messages.

Tweets were collected throughout the year 2014 citing drugs related to malaria. Some of these drugs did not present any tweets, like Morgidox, Ocudox, and Oraxyl. Doxycycline was the drug that yielded the most tweets (14,333, without including similar drugs), as shown in Table 4. Other drugs either did not present a significant enough number of messages for any analysis or did not have any AE associated with them.

There is no consensus on the best approach for disproportionality analysis: frequentist or Bayesian [37]. Both approaches are used in international research. The FDA uses *Multi-Item Gamma-Poisson Shrinker* (MGPS) [18], a Bayesian method. The frequentist method PRR was used in the

Table 4

Comparison of numbers of adverse events (AE) found in tweets and in the U.S. Food and Drug Administration (FDA) reports for the malaria drug doxycycline in the year 2014.

AE	Tweets	FDA reports
Abdominal discomfort	11	21
Abdominal distension	5	10
Abdominal pain upper	30	32
Abscess	2	-
Alopecia	155	18
Anaemia	6	33
Anaphylactic reaction	2	12
Angioedema	12	-
Anorexia	1	-
Anxiety	82	86
Aphthous stomatitis	33	-
Arthralgia	14	48
Back pain	70	29
Blood pressure increased	11	16
Bronchitis	10	33
Candidiasis	14	-
Cough	24	48
Decreased appetite	2	36
Dermatitis	1	11
Diarrhoea	10	96
Discomfort	1	17
Dyspepsia	3	11
Dysphagia	3	17
Ear infection	10	-
Emotional distress	1	47
Fungal infection	50	-
Gingivitis	1	-
Haemolytic anaemia	1	-
Headache	18	119
Hypersensitivity	72	29
Hypertension	8	22
Infection	271	19
Inflammation	7	14
Influenza	557	16
Injury	16	54
Insomnia	12	29
Intracranial pressure increase	1	-
Leukopenia	2	-
Malaise	4	91
Muscle spasms	11	42
Myalgia	3	32
Nasal congestion	1	-
Nasopharyngitis	1	20
Nausea	29	200
Oedema	1	12
Oesophageal ulcer	3	18

(continues)

Table 4 (continued)

AE	Tweets	FDA reports
Oesophagitis	3	-
Oropharyngeal pain	3	23
Pain	181	122
Photosensitivity reaction	9	18
Pigmentation disorder	2	-
Rash	45	90
Rhinorrhoea	1	15
Rosace	27	9
Sinusitis	36	18
Stevens-Johnson syndrome	6	-
Stomatitis	2	-
Swelling	37	9
Tension	22	-
Thrombocytopenia	2	12
Tooth abscess	2	-
Toothache	7	-
Ulcer	17	-
Urticaria	15	47
Vomiting	16	137

Table 5

Number of adverse events (AE) in reports on malaria drugs in 2014.

Drugs	n
Oraxyl	0
Primaquine	24
Fansidar	34
Sulfadoxine	36
Aralen	48
Lariam	110
Daraprim	128
Pyrimethamine	198
Mefloquine	319
Malarone	385
Proguanil	429
Morgidox	533
Ocudox	533
Chloroquine	621
Doryx	640
Atovaquone	1,040
Doxycycline	6,079
Plaquenil	7,664
Hydroxychloroquine	10,564

European Union at the time our analysis was done, and the ROR method (*reporting odd ratios*) is now used. Meanwhile, the WHO uses *Bayesian Confidence Propagation Neural Network* (BCPNN) [18], which is a Bayesian version of information component. Based on these observations, we opted to conduct the first analysis with PRR, since it was simpler than the other methods.

The analysis in Twitter detected signals for eight known AE for doxycycline: abdominal pain upper, anxiety, aphthous stomatitis, fungal infection, infection, pain, swelling, and tension. Two other AE were detected that had not been related previously to doxycycline: alopecia and rosacea. Of the known AE for doxycycline detected by analysis of the tweets, only anxiety was also found in the analysis of the FDA data. It would be interesting to make this comparison for a longer period of time to verify whether the signals generated by Twitter for these eight AE tend to increase, remain constant, or decrease. If these signals continue to appear only in tweets, it would potentially indicate that people are using this social network more than formal reports of AE.

A comparison of Tables 3 and 4 shows the existence of three AEs present in the reference standard and that only generated signals in Twitter, since there were no associated FDA reports. They are: aphthous stomatitis, fungal infection, and tension. This shows that AEs that do not appear in the reports could also be detected in Twitter, since they are also AE for doxycycline.

When investigating the two AE that were not in the reference standard (rosacea and alopecia) and that were detected by Twitter, we found that they also appeared in the FDA reports for the same period. There are reports not only that doxycycline can cause baldness, but also that it might be used to prevent it. On the AE rosacea, the vast majority of the tweets and reports indicated that the drug was used for its treatment, and that it was implicated as the cause [38].

Both alopecia and rosacea appear in the FDA reports, but only rosacea generated a signal in the data analysis. This is further evidence that the use of multiple data sources lends greater sensitivity to the automatic signal detection system, because if one considers only rare events, the analysis of multiple data sources is necessary to achieve the necessary statistical power and population heterogeneity to detect differences in the effectiveness of drugs in subpopulations, taking genetic, ethnic, and clinical differences into account [39].

The fact that alopecia is not in the reference standard means that it may be a potential new AE. This signal was not detected by FDA, only by Twitter, suggesting that this social network was able to detect signals that escape other sources.

Importantly, all the results of the analyses are signals, and not claims of a cause-and-effect relationship between the drug and the AE. In no way can such claims be made automatically, and subsequent studies led by specialists are needed to use these signals as initial warnings to justify more in-depth assessment.

Importantly, PRR and χ^2 are measures of association, not of causality. Thus, some events may not have generated signals, even though they are related to the target drugs, and this occurs in the analyses of both Twitter and FDA. Neither of the two analyses generated signals for all the AE in the reference standard, as shown in Table 3.

Although the FDA reports focus precisely on identifying ADEs, the vast majority of the 138 signals were generated for AE not in the reference standard (40 type A, 23 type B, and 55 type C). In other words, only 20 AE were already associated with doxycycline in the reference standard.

The study's results corroborate the idea that Twitter is useful for pharmacovigilance, but not as a stand-alone data source, rather as a complementary source. The social network proved capable of generating both new signals and those already in the reference standard, besides signals that were not obtained by analysis of the FDA data.

An emerging belief in pharmacovigilance research is that the combination of information from multiple data sources can lead to more effective and precise discovery of AE [8]. Depending on the data sources used and the ways they are combined, it is believed that the resulting system can lead to increased statistical significance in the results or facilitate new discoveries that are not possible based on single data sources. This hypothesis recently received preliminary confirmation [8], but further research is necessary. The use of multiple pipelines with the data processing, assessment, and analysis stages, each with different data sources, would be a way of corroborating the hypothesis and serve as an important future direction for research.

Besides being corroborated as additional source, another important factor is the availability of Twitter data, which allows real-time access for the data analysis, while pharmacovigilance networks usually take time to make their data available. The FDA, for example, publishes data by quarter, but these data are not necessarily made public after three months. The data for the months of January, February, and March are only made public halfway into the next quarter. The analysis of Twitter data proved useful for building a more complete pharmacovigilance system. Through analysis of these data, AE were detected that were not in the reference standard (alopecia and rosacea), and of these, alopecia was not in the signals generated by the FDA. Still, further analyses are needed to corroborate these results in order to include other drugs and other surveillance periods. It would also be interesting to conduct an analysis based on another method, such as MGPS, which is used by the FDA.

Contributors

F. V. Duval participated in the analysis and interpretation of the data, writing the article and is responsible for all aspects of the work in ensuring the accuracy and integrity of any part of the work. F. A. B. Silva collaborated in the conception and design of the article, critical review relevant of the intellectual content and final approval of the version to be published.

Additional informations

ORCID: Felipe Vieira Duval (0000-0003-4476-1277); Fabrício Alves Barbosa da Silva (0000-0002-8172-5796).

Acknowledgments

The authors wish to thank Brazilian Graduate Studies Coordinating Board (Capes) for the financial support.

References

1. Mendes M, Pinheiro R, Avelar K, Teixeira J, Silva G. História da farmacovigilância no Brasil. Rev Bras Farm 2008; 89:246-51.
2. World Health Organization. Pharmacovigilance. https://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/ (accessed on 01/Oct/2018).
3. Balbino EE, Dias MF. Farmacovigilância: um passo em direção ao uso racional de plantas medicinais e fitoterápicos. Rev Bras Farmacogn 2010; 20:992-1000.
4. Mota DM. Evolução e resultados do sistema de farmacovigilância do Brasil [Dissertação de Mestrado]. Porto Alegre: Faculdade de Medicina, Universidade Federal do Rio Grande do Sul; 2017.
5. Meyboom RH, Egberts AC, Gribnau FW, Hekster YA. Pharmacovigilance in perspective. Drug Saf 1999; 21:429-47.
6. Venulet J, ten Ham M. Methods for monitoring and documenting adverse drug reactions. Int J Clin Pharmacol Ther 1996; 34:112-29.
7. Cardoso MA, Amorim MAL. A farmacovigilância e sua importância no monitoramento das reações adversas a medicamentos. Revista Saúde e Desenvolvimento 2013; 4:33-56.
8. Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, et al. Combining signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013; 20:413-9.
9. Hauben M, Bate A. Data mining in drug safety: side effects of drugs essay. Side Effects of Drugs Annual 2007; 29:xxxiii-xlvi.
10. Hotho A, Nürnberger A, Paaß G. A brief survey of text mining. https://pdfs.semanticscholar.org/9643/0cc91ed91fd2d4042fa6fcb7ecf4005d77a7.pdf (accessed on Sep/2018).
11. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009; 457:1012-4.
12. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infecti Dis 2009; 49:1557-64.
13. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis 2014; 8:e2713.

14. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PLoS One 2011; 6:e19467.
15. Lampos V, Cristianini N. Nowcasting events from the social web with statistical learning. ACM Trans Intell Syst Technol 2012; 3:72.
16. First Workshop on Computational Methods in Pharmacovigilance held during the Medical Informatics in Europe (MIE) Conference, Pisa, Italy, 29 August 2012. Drug Saf 2012; 35:1191-200.
17. Gurulingappa H, Rajput AM, Toldo L. Extraction of adverse drug effects from medical case reports. J Biomed Semantics 2012; 3:15.
18. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012; 91:1010-21.
19. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Uppsala: Association for Computational Linguistics; 2010. p. 117-25.
20. Yates A, Goharian N. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In: Serdyukov P, Braslavski P, Kuznetsov SO, Kamps J, Rüger S, Segalovich EA, et al., editors. Advances in information retrieval. Berlin: Springer; 2013. p. 816-9.
21. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010; 6:343.
22. Ginn R, Pimpalkhute P, Nikfarjam A, Patki A, O'Connor K, Sarker A, et al. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. http://www.nactem.ac.uk/biotxtm2014/papers/Ginnetal.pdf (accessed on 01/Oct/2018).
23. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Saf 2014; 37:343-50.
24. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: SHB'12 Proceedings of the 2012 International Workshop on Smart Health and Wellbeing. https://dl.acm.org/citation.cfm?id=2389713 (accessed on 01/Oct/2018).
25. Portal Sinan. Calendário epidemiológico. http://portalsinan.saude.gov.br/calendario-epidemiologico (accessed on Oct/2018).
26. Duval F, Caffarena E, Cruz O, Silva F. Mining for adverse drug events on twitter. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval. http://www.scitepress.org/PublicationsDetail.aspx?ID=hxPWwh5Sjzw=&t=1 (accessed on 01/Oct/2018).
27. Cai M-C, Xu Q, Pan Y-J, Pan W, Ji N, Li Y-B, et al. ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. Nucleic Acids Res 2015; 43:D907-13.
28. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. Nat Lang Eng 1995; 1:83-108.
29. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010; 17:507-13.
30. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. Bethesda: National Center for Biotechnology Information, U.S. National Library of Medicine; 2001. p. 17-21.
31. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salamé G, et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf 2009; 18:1176-84.
32. Dias P, Ribeiro CF, Marques FB. Medidas de desproporcionalidade na deteção de sinal em farmacovigilância. Revista Portuguesa de Farmacoterapia 2014; 6:28-32.
33. van Puijenbroek EP, Diemont WL, van Grootheest K. Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions. Drug Saf 2003; 26:293-301.
34. EudraVigilance Expert Working Group. Guideline on the use of statistical signal detection methods in the Eudravigilance data analysis system. London: European Medicines Agency; 2006.
35. Evans S, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001; 10:483-6.
36. Rothman K, Greenland S. Introduction to categorical statistics. In: Rothman K, Greenland S, editors. Modern epidemiology. 2nd Ed. Philadelphia: Lippincott Williams & Wilkins; 1998. p. 231-52.
37. Klarreich E. In search of bayesian inference. Commun ACM 2014; 58:21-4.
38. Valentín S, Morales A, Sánchez JL, Rivera A. Safety and efficacy of doxycycline in the treatment of rosacea. Clin Cosmet Investig Dermatol 2009; 2:129-40.
39. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. J Am Med Inform Assoc 2013; 20:453-61.

Resumo

Durante o período de pós-comercialização, quando medicamentos são usados por grandes populações e por períodos de tempo maiores, eventos adversos (EA) inesperados podem ocorrer, o que pode alterar a relação risco-benefício dos medicamentos o suficiente para exigir uma ação regulatória. Eventos adversos são agravos à saúde que podem surgir durante o tratamento com um produto farmacêutico, os quais, no período de pós-comercialização do medicamento, podem requerer um aumento significativo de cuidados de saúde e resultar em danos desnecessários aos pacientes, muitas vezes fatais. Portanto, o quanto antes, a descoberta de EA no período de pós-comercialização é um objetivo principal do sistema de saúde. Alguns países possuem sistemas de vigilância farmacológica responsáveis pela coleta de relatórios voluntários de EA na pós-comercialização, mas estudos já demonstraram que, com a utilização de redes sociais, pode-se conseguir um número maior e mais rápido de relatórios. O objetivo principal deste projeto é construir um sistema totalmente automatizado que utilize o Twitter como fonte para encontrar EA novos e já conhecidos e fazer a análise estatística dos dados obtidos. Para isso, foi construído um sistema que coleta, processa, analisa e avalia em busca de EA, comparando-os com dados da Agência Americana de Controle de Alimentos e Medicamentos (FDA) e do padrão de referência construído. Nos resultados obtidos, conseguimos encontrar EA novos e já existentes relacionados ao medicamento doxiciclina, o que demonstra que o Twitter, quando utilizado em conjunto com outras fontes de dados, pode ser útil para a farmacovigilância.

Controle de Medicamentos e Entorpecentes; Ontologia Biológica; Processamento de Linguagem Natural; Mídias Sociais; Base de Dados

Resumen

Durante el período de poscomercialización, cuando grandes poblaciones consumen medicamentos durante períodos más prolongados de tiempo, se pueden producir eventos adversos (EA) inesperados, lo que puede alterar la relación riesgo-beneficio de los medicamentos. Esta situación es suficiente para exigir una acción regulatoria. Los EA son agravios a la salud que pueden surgir durante el tratamiento con un producto farmacéutico, los cuales, durante el período de poscomercialización del medicamento, pueden requerir un aumento significativo de cuidados de salud y resultar en lesiones innecesarias para los pacientes, muchas veces fatales. Por lo tanto, el hallazgo anticipado de EA durante el período de poscomercialización es un objetivo primordial del sistema de salud. Algunos países cuentan con sistemas de vigilancia farmacológica, responsables de la recogida de informes voluntarios de EA durante la poscomercialización, pero algunos estudios ya demostraron que, con la utilización de las redes sociales, se puede conseguir un número de informes mayor y más rápido. El objetivo principal de este proyecto es construir un sistema totalmente automatizado que utilice Twitter como fuente para encontrar nuevos EA y ya conocidos, además de realizar un análisis estadístico de los datos obtenidos. Para tal fin, se construyó un sistema que recoge, procesa, analiza y evalúa tweets *en búsqueda de eventos adversos, comparándolos con datos de la Agencia Americana de Control de Alimentos y Medicamentos (FDA) y del estándar de referencia construido. En los resultados obtenidos, conseguimos encontrar nuevos eventos adversos y ya existentes, relacionados con el medicamento doxiciclina, lo que demuestra que Twitter, cuando es utilizado junto a otras fuentes de datos, puede ser útil para la farmacovigilancia.*

Control de Medicamenteos y Narcóticos; Ontologías Biológicas; Procesamiento de Lenguaje Natural; Medios de Comunicación Sociales; Base de Datos

Submitted on 24/Feb/2017
Final version resubmitted on 22/Sep/2018
Approved on 18/Oct/2018