Review

Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review

Jérémy Lardon[1,2*], PhD; Redhouane Abdellaoui[3,4*], MS; Florelle Bellet[5], PharmD; Hadyl Asfari[2,6], PharmD; Julien Souvignet[2,6], MS; Nathalie Texier[4], PharmD; Marie-Christine Jaulent[6], PhD; Marie-Noëlle Beyens[7], MD; Anita Burgun[8,9], MD, PhD; Cédric Bousquet[2,6], PharmD, PhD

[1]Université Paris 13, Sorbonne Paris Cité, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), (Unité Mixte de Recherche en Santé, UMR_S 1142), F-93430, Villetaneuse, France, Sorbonne Universités, University of Pierre and Marie Curie (UPMC) Université Paris 06, Unité Mixte de Recherche en Santé (UMR_S) 1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Institut National de la Santé et de la Recherche Médicale (INSERM), U1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Paris, France

[2]Service de Santé Publique et de l'Information Médicale (SSPIM), Department of Public Health and Medical Informatics, Centre Hospitalier Universitaire (CHU) University Hospital of Saint Etienne, Saint-Etienne, France

[3]Institut National de la Santé et de la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1138, équipe 22, Centre de Recherche des Cordeliers, Université Paris Descartes, Sorbonne Paris Cité, F-75006, Paris, France

[4]Kappa Santé, Paris, France

[5]Centre de Pharmacovigilance, Centre Hospitalier Universitaire (CHU) University Hospital of Saint Etienne, Saint-Etienne, France

[6]Université Paris 13, Sorbonne Paris Cité, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), (Unité Mixte de Recherche en Santé, UMR_S 1142), F-93430, Villetaneuse, France, Sorbonne Universités, University of Pierre and Marie Curie (UPMC) Université Paris 06, Unité Mixte de Recherche en Santé (UMR_S) 1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Paris,, Institut National de la Santé et de la Recherche Médicale (INSERM), U1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006, Paris, France

[7]Centre de Pharmacovigilance, Centre Hospitalier Universitaire (CHU) University Hospital of Saint Etienne, Saint-Etienne, France

[8]Institut National de la Santé et de la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1138, équipe 22, Centre de Recherche des Cordeliers, Université Paris Descartes, Sorbonne Paris Cité, F-75006, Paris, France

[9]Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Européen Georges-Pompidou (HEGP), Department of Medical Informatics, Paris, France

[*]these authors contributed equally

Corresponding Author:
Jérémy Lardon, PhD
Université Paris 13, Sorbonne Paris Cité, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), (Unité Mixte de Recherche en Santé, UMR_S 1142), F-93430, Villetaneuse, France
Sorbonne Universités, University of Pierre and Marie Curie (UPMC) Université Paris 06, Unité Mixte de Recherche en Santé (UMR_S) 1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006
Institut National de la Santé et de la Recherche Médicale (INSERM), U1142, Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS), F-75006
Campus des Cordeliers, Institut National de la Santé et de la Recherche Médicale (INSERM) U 1142 - Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé (LIMICS) Esc D - 2ème étage
15 rue de l'École de Médecine
Paris, 75006
France
Phone: 33 144279190
Fax: 33 144279192
Email: jeremy.lardon@chu-st-etienne.fr

Abstract

Background: The underreporting of adverse drug reactions (ADRs) through traditional reporting channels is a limitation in the efficiency of the current pharmacovigilance system. Patients' experiences with drugs that they report on social media represent a new source of data that may have some value in postmarketing safety surveillance.

Objective: A scoping review was undertaken to explore the breadth of evidence about the use of social media as a new source of knowledge for pharmacovigilance.

Methods: Daubt et al's recommendations for scoping reviews were followed. The research questions were as follows: How can social media be used as a data source for postmarketing drug surveillance? What are the available methods for extracting data? What are the different ways to use these data? We queried PubMed, Embase, and Google Scholar to extract relevant articles that were published before June 2014 and with no lower date limit. Two pairs of reviewers independently screened the selected studies and proposed two themes of review: manual ADR identification (theme 1) and automated ADR extraction from social media (theme 2). Descriptive characteristics were collected from the publications to create a database for themes 1 and 2.

Results: Of the 1032 citations from PubMed and Embase, 11 were relevant to the research question. An additional 13 citations were added after further research on the Internet and in reference lists. Themes 1 and 2 explored 11 and 13 articles, respectively. Ways of approaching the use of social media as a pharmacovigilance data source were identified.

Conclusions: This scoping review noted multiple methods for identifying target data, extracting them, and evaluating the quality of medical information from social media. It also showed some remaining gaps in the field. Studies related to the identification theme usually failed to accurately assess the completeness, quality, and reliability of the data that were analyzed from social media. Regarding extraction, no study proposed a generic approach to easily adding a new site or data source. Additional studies are required to precisely determine the role of social media in the pharmacovigilance system.

(J Med Internet Res 2015;17(7):e171) doi: 10.2196/jmir.4304

KEYWORDS

pharmacovigilance; adverse drug reaction; Internet; Web 2.0; social media; text mining; scoping review; adverse event

Introduction

Pharmacovigilance is defined by the World Health Organization as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" [1]. It comprises postmarketing safety surveillance activities to monitor drug benefit/risk ratios and to identify new potential adverse drug reaction (ADR) signals in real-life conditions. An ADR is defined as "[...] a response to a medicinal product which is noxious and unintended and which occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease or for the restoration, correction or modification of physiological function. (WHO, 1972)" [2]. Not all ADRs are identified during clinical trials because of their limited duration and the numbers and types of patients. ADRs need to be followed up after drug approval [3] and, therefore, are burdens for health systems that can potentially lead to hospitalization [4] or death [5].

Pharmacovigilance is mainly based on the spontaneous reporting of ADRs. Initially, only health professionals were allowed to report ADRs. Subsequently, however, a number of studies [6-10] demonstrated the value of patients as reporters. Hughes and Cohen stated that drug user reporting could be a complementary source of knowledge [11]. Currently, a number of countries consider direct patient reporting to be a valuable source in pharmacovigilance [12] and have implemented regulations and solutions for patients' spontaneous reporting to health authorities. Although patients have increased the number of reporters, ADR underreporting remains a limitation of the current pharmacovigilance system [13-15]. Moreover, other sources for pharmacovigilance are now considered, such as via the secondary use of electronic health records [16-23].

The recent Web 2.0 and social media expansions have been accompanied by a rapid growth in the number of discussions on the Internet regarding drug uses. Social media constitutes a new data source for postmarketing drug safety surveillance [24] and may be of interest in identifying signals because of their high volume and availability.

The use of Internet discussions as an additional data source relies on methods to parse, extract, structure, collect, and organize relevant information from the Web pages for analysis. The use of many sources, the large amount of data, and the heterogeneity of data require multiple steps to obtain analyzable corpora. Methods derived from big data and natural language processing (NLP) need to be considered. Recent works have proposed solutions to address these issues and to standardize the process of extracting information from Web pages in social media.

In addition, questions remain about the quality of information available in users' Web 2.0 discussions. Whereas electronic health records and health professionals' ADR reports are structured and well documented, there are no requirements regarding writing and structuring descriptions of pharmacovigilance-related events on social media, and information may be scarce or incomplete.

We performed a scoping review of relevant previously published studies to assess how social media can be used as a data source for postmarketing drug surveillance. This type of literature review aims at providing an overview of the type, extent, and quantity of research available on this topic. Our overview describes the methods used to manage the data from the corpus of Web users' messages and the obtained results, and identifies potential research gaps and future needs.

Methods

Overview

We used the scoping review methodology described by Arksey and O'Malley [25] and further refined by Levac et al [26] and Daudt et al [27]. This methodology divides reviews into six stages: (1) identifying the research question, (2) identifying relevant studies, (3) selecting studies, (4) charting the data, (5) collating, summarizing, and reporting the results, and (6)

consultation with stakeholders. Although the sixth stage is optional, we followed the recommendations to consider it a required component.

Stage 1: Identifying the Research Question

The focus of this scoping review is the use of social media as a new source of data in pharmacovigilance. We use the common definition of social media as media-based or user-generated content. Consequently, we did not consider the news media. To define the search question, we first selected a sample of available publications and found two types: (1) reviews of Web forums conducted by pharmacovigilance specialists on the one hand and (2) technical articles on information extraction approaches authored by computer scientists on the other. In this article, we use the term "identification" to denote the manual process of pulling up social media pages and reviewing them for reports of ADRs. The term "extraction" is used to describe the algorithms that automatically extract ADR information from social media. In the following sections, terms such as "messages," "social media," "discussion," and "page" refer to Web content.

The research question regarding the use of social media or pharmacovigilance is twofold:

1. Theme 1: What is the relevant information for ADR signals that have been issued from social media? The identification theme focuses on the first question and evaluates the information contained in patients' narrations on social media.

2. Theme 2: What are the methods used to extract information from social media? The extraction theme commits to describing the automated tools and methods that have been used to access structured and valuable pharmacovigilance information.

Stage 2: Identifying Relevant Studies

Two electronic databases—PubMed and Embase—were searched for English and French articles. The PubMed database was searched twice, as follows:

1. With the following keywords (query #1) to investigate the pharmacovigilance and social media dimensions: pharmacovigilance, adverse reaction, adverse event (AE), drug, medication, pharmaceutical product, social media, Web 2.0, social network, Twitter, Facebook, blog, forum, fora, message board, comment, and user feedback. An outline view of this request is presented in Figure 1.

2. Medical Subject Heading (MeSH) terms (query #2)—pharmacovigilance, natural language processing, Adverse Drug Reaction Reporting Systems, and Internet—associated with the following keywords in the title or the abstract: surveillance, Twitter, Facebook, Doctissimo (the main French health-related discussion forum), social media, social network, online health community, online discussion, medical data mining, online, patient forum, and natural language processing.

Query #3 was specially designed for the Embase database based on query #1. The details of the three queries are given in Table 1.

The upper date limit of June 2014 was applied, with no lower date limit, considering that articles published in the early days of social networks could be of interest.

As an iterative process and in accordance with the scoping review methodology, all references from the studies selected in stage 3 were screened, as were all of the publications that cited the selected studies. To broaden the scope of the search, Google Scholar was also used to search for citations.

Table 1. Full search strategy for each database.

Database	Query	Query text
PubMed		
	Query #1 (keywords)	(pharmacovigilance[MeSH[a] Terms] OR pharmacovigilance[All Fields] OR ADR[b][All Fields] OR ADE[c][All Fields] OR (("adverse reaction"[All Fields] OR "adverse event"[All Fields] OR "side effect"[All Fields]) AND (drug[All Fields] OR medication[All Fields] OR pharmaceutical product*[All Fields]))) AND ("social media"*[All Fields] OR "Web 2.0"[TIAB[d]] OR "Web 2.0"[TIAB] OR "social media" [TIAB] OR "social network*" OR Twitter OR Facebook OR blog OR forum* OR fora OR message board* OR comment* OR (user feedback*))
	Query #2 (MeSH terms)	((((("pharmacovigilance"[MeSH]) OR surveillance[Title])) AND (((((Twitter[Title/Abstract]) OR Facebook[Title/Abstract]) OR Doctissimo[Title/Abstract])) OR (((((((((social media[Title/Abstract]) OR social networks[Title/Abstract]) OR "online health community"[Title/Abstract]) OR "online discussion"[Title/Abstract]) OR medical data mining[Title/Abstract]) OR online[Title/Abstract]) OR patient forum[Title/Abstract]) OR natural language processing[MeSH Terms]) OR "natural language processing"[Title/Abstract]))) OR (((((("Adverse Drug Reaction Reporting Systems"[MeSH]) AND (((((Twitter[Title/Abstract]) OR Facebook[Title/Abstract]) OR Doctissimo[Title/Abstract])) OR (((((((((social media[Title/Abstract]) OR social networks[Title/Abstract]) OR "online health community"[Title/Abstract]) OR "online discussion"[Title/Abstract]) OR medical data mining[Title/Abstract]) OR online[Title/Abstract]) OR patient forum[Title/Abstract]) OR natural language processing[MeSH Terms]) OR "natural language processing"[Title/Abstract])))) OR ((((((((((social media[Title/Abstract]) OR social networks[Title/Abstract]) OR "online health community"[Title/Abstract]) OR "online discussion"[Title/Abstract]) OR medical data mining[Title/Abstract]) OR online[Title/Abstract]) OR patient forum[Title/Abstract]) OR natural language processing[MeSH Terms]) OR "natural language processing"[Title/Abstract])) AND Adverse Drug Reaction Reporting Systems[MeSH Terms])) OR (("Adverse Drug Reaction Reporting Systems"[MeSH]) AND "Internet"[Mesh]))
Embase	Query #3	"pharmacovigilance"/de OR ADR OR ADE OR ("adverse reaction"/de OR "adverse event" OR "side effect"/de AND ("drug"/de OR "medication"/de OR "pharmaceutical product")) AND ("social media"/de OR "Web 2.0":ab,ti[e] OR "Web 2.0":ab,ti OR "social media":ab,ti OR "social network"/de OR Twitter OR Facebook OR blog OR forum OR fora OR "message board" OR comment OR "user feedback")

[a]MeSH: Medical Subject Heading

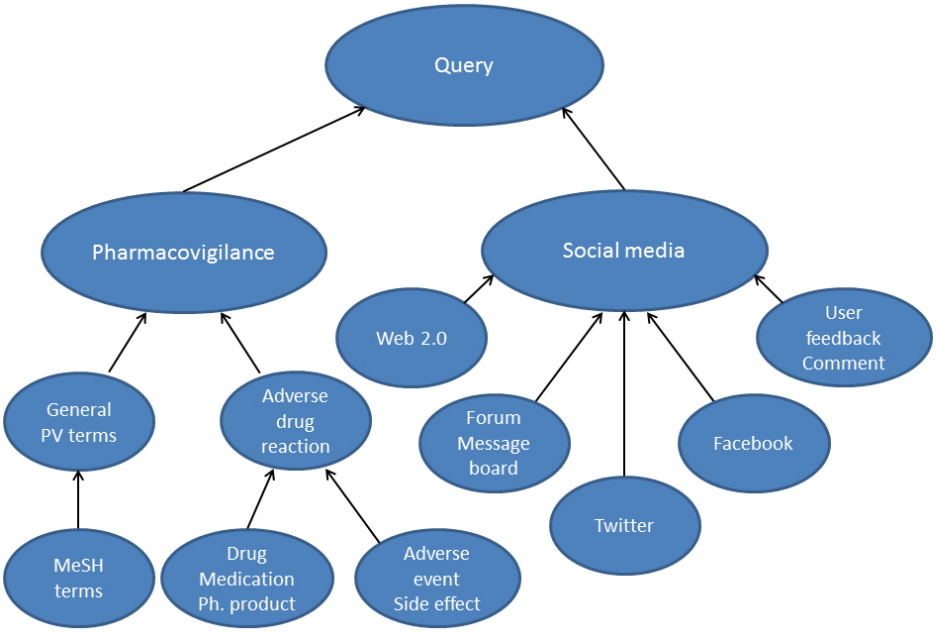
[b]ADR: adverse drug reaction

[c]ADE: adverse drug event

[d]TIAB: title and abstract

[e]ab, ti: abstract, title

Figure 1. Structure of the search queries.



Stage 3: Selecting Studies

Four authors (JL, RA, CB, and FB) independently screened the titles and abstracts (when available) of the query results to identify relevant articles. Disagreements about exclusions were discussed until a consensus was reached.

Abstracts were excluded if they met at least one of the following criteria:

1. Not related to drugs.

2. Not related to ADR reporting or ADR detection (eg, efficacy or effectiveness of a study's design).

3. Not related to patients' reporting (eg, a safety study in animals, signal detection on a pharmacovigilance database).

4. No, or insufficient, results on the use of social media.

5. The study was a review.

6. Soliciting reporting: the study used data from a source where patients were asked to report the ADRs. Patients' behavior is different depending on whether they are asked to report ADRs or they describe ADRs spontaneously without knowing that there may be further analysis of what they write [28].

7. Editorial: the study did not encompass a result but was an expression of the author's opinion about the usefulness of social media as a new source of knowledge for pharmacovigilance.

Stage 4: Charting the Data

Two pairs of reviewers independently identified a set of characteristics that could be used to describe the articles in each theme (FB and HA for theme 1, ie, identification; JL and RA for theme 2, ie, extraction). In addition to the basic elementary metadata, a number of characteristics were recorded for each theme. They are listed in Table 2.

The reviewers independently extracted data from the articles that were assigned to them.

Table 2. Article characteristics overview.

Characteristics	Theme 1	Theme 2
Year of publication	✓	✓
Language used in the studied texts	✓	✓
Type of data source, for example, forums or Twitter	✓	✓
Presence of an anonymization step	✓	✓
Volume of data analyzed	✓	✓
List of studied drugs	✓	✓
Coding ADRs[a] (medical lexicon)	✓	✓
Keywords the authors used to identify sources or posts of interest	✓	
Use of semiautomated processes (mixed methods)	✓	
Main results	✓	
Whether reported ADRs were highly informative or not	✓	
Seriousness of reported ADRs	✓	
Reference source was used for comparison with reported ADRs	✓	
Identification of potential unexpected ADRs or unexpected frequency of known ADRs	✓	
Analysis of the influence of other media, for example, television, radio, or the press, as a potential cause of increased ADR reporting in social media	✓	
If the authors mentioned the use of a crawler		✓
Implemented methods of preprocessing		✓
Lay language lexicon or tools used		✓
Authors attempted to identify the relationship between the drug and the event		✓
Authors used a machine-learning approach		✓
Evaluation of the extraction methods with metrics		✓
Comparison with external pharmacovigilance databases		✓
Whether the system enabled evaluating the unexpectedness of any extracted ADRs		✓

[a]ADR: adverse drug reactions

Stage 5: Collating, Summarizing, and Reporting Results

This work aimed to describe the methods and the results of the two themes. The results for theme 1, "identification," were related to studies based on the manual search and are presented in terms of methods and quality of data. The second theme, "extraction," was related to the studies that promoted an automated approach to extracting information from raw data.

We summarized the "methods" sections of this last set of studies to describe each step of the methods presented.

Stage 6: Consultation

Following Daubt et al's recommendation [27], the research was multidisciplinary and multi-professional. The overall expertise covered pharmacovigilance, pharmacoepidemiology, public health, medical informatics, statistics, and data mining.

This helped us identify additional expectations regarding pharmacovigilance and social media, such as misuses, counterfeit drugs, drug-drug and food-drug interactions, and ADRs in specific populations such as pregnant women. It also permitted us to identify potential stakeholders—health care professionals, regulatory agencies, pharmaceutical companies, and patients—and establish the necessity of measuring the impact of mining social media, the interest in integrating this approach in a practical way in addition to classical reporting systems, and how we can be confident about the findings.

Results

Overview of Results

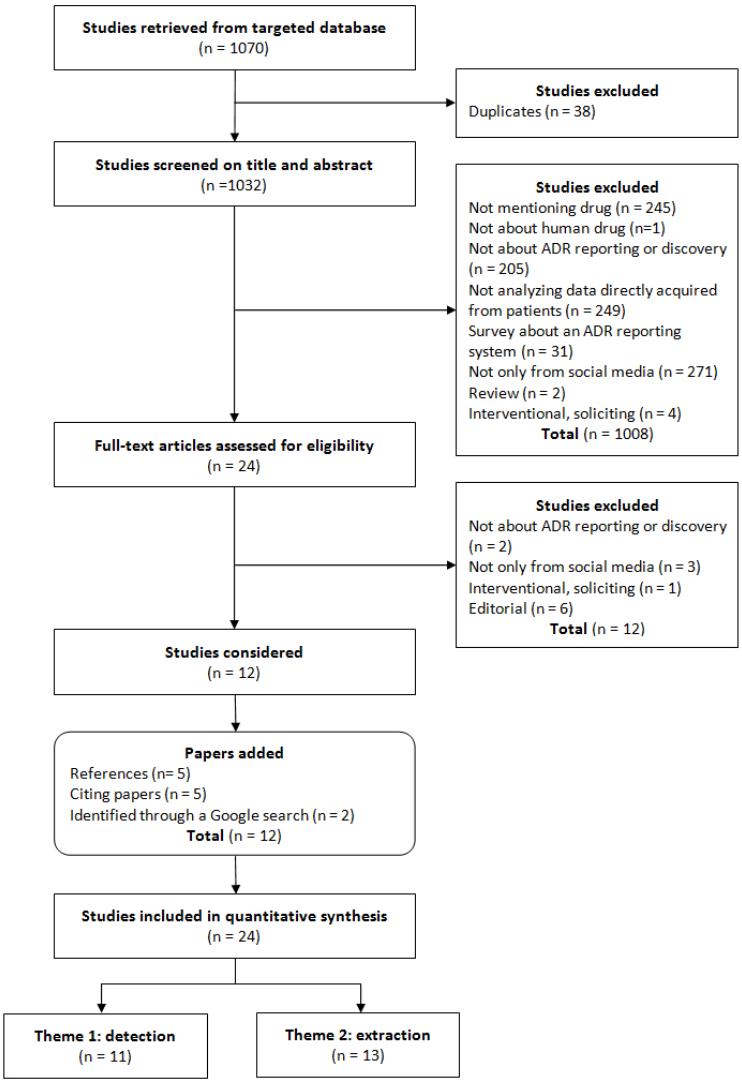
Figure 2 depicts the full review process and shows the number of citations excluded at each step.

A total of 1032 publications were identified in PubMed and Embase after duplicates (n=38) were removed. After applying the exclusion criteria to the titles, abstracts, and full texts, 11 citations were relevant to the research question at the end of the screening process (stage 3).

An additional 2 publications were added based on our personal previous knowledge, and 11 studies were identified by the references cited in the publications that were initially selected or by checking other articles that cited these publications on Google Scholar—7 via references and 4 via citing papers.

A total of 24 publications were finally included in the chart process. Of these, 11 (46%) were analyzed for theme 1 (identification) [11,28-37] and 13 (54%) for theme 2 (extraction) [38-50]. The detailed results of charting the data are displayed in Multimedia Appendix 1 and Multimedia Appendix 2.

Figure 2. Flowchart of our mapping process and study selection.



Theme 1: Identification

Overview

A total of 11 studies described a manual (or mixed) approach for identifying drug-ADR pairs in patients' narratives that were posted on social media. The majority of these studies were performed in the United States (6/11, 55%) or in France (3/11, 27%). Of these 11 studies, 4 (36%) were published in 2014 and 2 (18%) before 2010—in 2007 [36] and 2009 [34].

In 3 out of the 11 (27%) studies, the authors used the term "adverse event" rather than "adverse drug reaction" to refer to problems reported by patients in social media [30,32,35]. Pages et al justified this in their methodology by stating that the events reported by patients "were not analyzed by health professionals to assess the causal relationship" with the drug [35].

Table 3 details the steps identified in the studies to conduct the manual analysis: (1) selection of data sources, (2) data collection, (3) identification of drug-ADR/AE pairs, and (4) results evaluations.

Table 3. Main steps for identifying adverse drug reactions from social media.

Step	Description
Step 0: Selection of data sources	This step consists of identifying and selecting the most relevant websites to answer the research question. They can be identified using a combination of keywords (eg, generic or brand-name drug, disease, ADR[a]/AE[b]) in Web search engines.
Step 1: Data collection	Potentially relevant patient narratives or posts are identified by entering keywords into the search engine hosted by the selected websites (manual identification only) or using a semiautomated process. Data may be imported into software (after anonymization) with the aim of additional analyses.
Step 2: Identification of drug-ADR/AE pairs	The manual identification of drug-ADR/AE pairs is performed by reading the patients' narratives or posts that were initially collected.
Step 3: Results evaluation	This step consists of manually evaluating the frequency and the seriousness of the ADRs or AEs that were identified in patients' narratives or posts. The results can be compared, after coding, with those of other sources (Summary of Product Characteristics [SPC], clinical trials, pharmacovigilance databases, or literature) to identify potential new ADRs or an unexpected frequency of a known ADR.

[a]ADR: adverse drug reaction

[b]AE: adverse event

Analyzed Data Sources

The main data source was online forums. Three authors also reported on the analysis of Tweets or blogs [29,30,32]. The selected websites are often devoted to consumers' health. Patients' comments, mostly written in English, were identified by keywords (eg, brand-name and generic drugs, diseases, ADR) in the search engine hosted by the selected websites. In 2 of 11 (18%) studies, a hybrid (semiautomated) process was performed to identify potentially relevant posts [30,33].

The volume of analyzed data varied according to the studies—from 96 comments [31] to 61,401 Tweets [30]. Most often (ie, 8/11, 73%), the studies analyzed hundreds of narratives or posts.

Scope of the Surveillance

Out of 11 studies, 9 (82%) were designed to identify all of the ADRs/AEs that were potentially associated with one or more preselected drugs. Among these 9 studies, 7 (78%) focused on a class of drugs (eg, statins [31] or antineoplastic [33,35], psychotropic [10], or antiparkinsonian agents [36]), a recently marketed drug (eg, dabigatran [37]), or a drug that was removed from the market for pharmacovigilance reasons (eg, benfluorex [28]). Out of the 11 studies, 1 (9%) focused on two specific life-threatening ADRs—Stevens-Johnson syndrome and toxic epidermal necrolysis—and aimed at identifying any potentially associated drugs [29]. Another (1/11, 9%) was designed to identify and analyze predefined drug-AE pairs [34].

The rate of detected ADRs or AEs among the analyzed patients' comments varied according to the studies and was difficult to compare considering the methodological heterogeneity. Whereas Kmetz et al found a relatively low rate of reported AEs among patients' posts containing mentions of targeted drugs—0.3% of all brand mentions and 3.3% of brand mentions that contained side effects keywords [32]—Butt et al identified a large number of Internet descriptions for two rare and serious ADRs [29].

The informativeness of patients' comments was more or less evaluated in 8 of the 11 (73%) selected studies. In 5 of these 8 (63%) publications, information concerning patients' characteristics (ie, age, gender, and medical history), suspected drugs (ie, indications, dosages, and date of treatment initiation), or concomitant medications was often not available [31,32,34,35,37].

The presence of chronological criteria (ie, time to onset of ADRs, dechallenge, or rechallenge) was mentioned in only 3 of 11 (27%) studies and varied significantly according to the websites that were analyzed [10,31,36].

In 6 of 11 (55%) studies, the authors verified if the ADRs that were identified in social media were expected or not and compared them with those that had been reported in clinical trials (3/6, 50%) [33,36,37], in pharmacovigilance databases (2/6, 33%) [30,35], or in the studied drugs' Summary of Product Characteristics (SPC) (1/6, 17%) [31].

Out of 11 studies, 2 (18%) reported using a standard terminology—Medical Dictionary for Regulatory Activities

(MedDRA) [30]—or Problem Intervention Documentation (PI-Doc) [36], a German classification system, for coding the ADRs or AEs that were reported in social media. Freifeld et al [30] found that the AEs/ADRs identified in social media had similar profiles to those that were spontaneously reported through official channels, whereas Pages et al described the qualitative differences between the data sources [35].

Some studies identified potential unexpected ADRs [31,35,37] or unexpected frequencies of known ADRs [33,36]. Furthermore, Abou Taam et al retrospectively identified one case of severe valvulopathy 7 months before benfluorex was withdrawn because of this toxicity [28]. In their study, Butt et al compared patients' unsolicited Internet descriptions of severe cutaneous ADRs with experiences that had been previously collected in face-to-face interviews of survivors of these ADRs [29]. The authors identified new themes from Internet narratives, including fears and concerns of patients who had experienced the condition. According to the authors, patients also reported on more sensitive issues, such as sexual dysfunction, on the Internet rather than in face-to-face interviews.

The ADRs and AEs reported in social media were often less serious than those that were spontaneously reported through official channels [31,35], but they had impaired the patients' quality of life and their adherence to treatment [30,33,34]. Patients' comments were also related to complications from the ADRs, contraindications, drug-drug and food-drug interactions, and storage of drugs [37]. Users shared their experiences with individuals who were taking the same drug or had had similar adverse events, or with health care providers to obtain information or advice.

Abou Taam et al evaluated the impact of media coverage on benfluorex's withdrawal in France by analyzing patients' comments at three periods: before, during, and after the withdrawal of the drug. They found messages reflecting anxiety, anger, and other feelings, with drastic changes in consumers' perceptions following media coverage [28].

Theme 2: Extraction

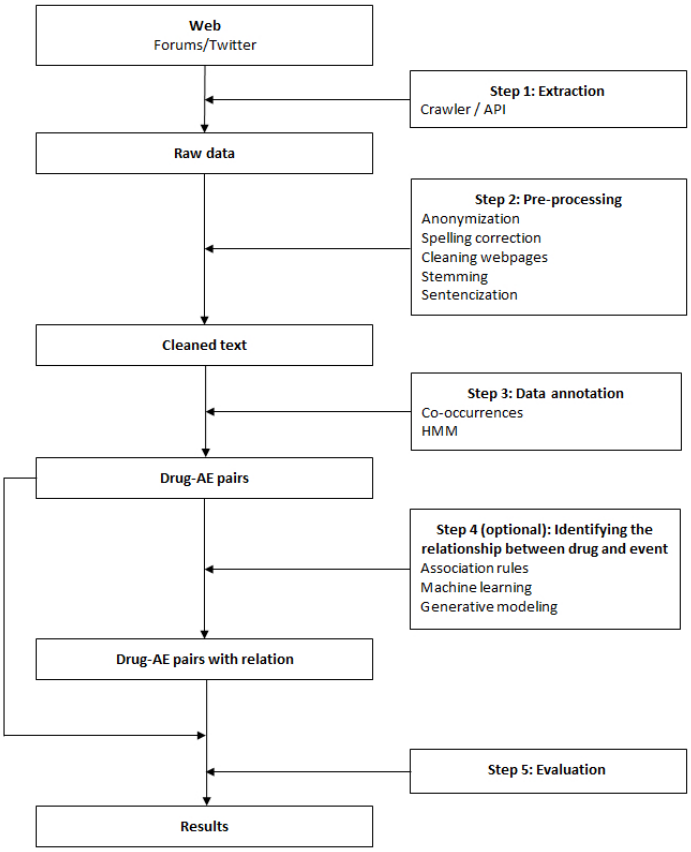
Overview

The 13 studies that were selected for the review had been recently published—2010 for the oldest [42].

Figure 3 shows a synthesis of the complete steps and presents five distinct parts: (1) data extraction, (2) preprocessing, (3) data annotation, (4) identifying the relationship between drug and event, and (5) results evaluation.

Multimedia Appendix 3 summarizes the use of these different steps in the papers.

Figure 3. Main steps for extraction of adverse drug reactions (ADRs) from social media.



Choice of the Source

The main data source was forum discussions in 12 studies out of 13 (92%) [34,35,37,38,40-44,47-49]. Out of the 13 studies, 1 (8%) [39] was about extracting narratives from Tweets.

Each study examined narratives that had been written in English, except for that of Hadzi-Puric and Grmusa [40]. The data volume was heterogeneous and varied from millions of messages [41] or billions of Tweets [39] to a more limited number of

messages, such as the 1290 messages included in the study by Hadzi-Puric and Grmusa [40].

The list of studied drugs was also heterogeneous. Out of the 13 selected studies, 11 (85%) focused on a limited number of drugs, such as lipid-modifying drugs in Li [43]. The other studies aimed at detecting signals of a large number of drugs, as in Liu and Chen's work [44], which considered all of the drugs from the Unified Modeling Language System (UMLS) and the US Food and Drug Administration's (FDA) Adverse Drug Event Reporting System (FAERS).

Data Extraction

The operating method to extract data from Web forums and social media depended on the nature of the source. For Web forums, 8 of 13 (62%) articles used an adapted Web crawler to collect Web pages, and then a Web scraper to extract the messages that were embedded in these Web pages [38,42-44,46-49].

Web scraping can be done through two approaches: (1) by taking the whole code of the page and cleaning it by eliminating the HTML tags and other unwanted elements or (2) by targeting the patients' messages using the HTML structure. The first approach was chosen by Benton [38]. In this work, approximately 48% of the tokens, defined as strings of characters delimited by whitespace in the original HTML pages, were retained to generate the corpus.

When the source was Twitter [39], specific application programming interfaces (APIs) were available for extracting data. These APIs provided some structured information, such as the date of the message or the pseudonym of the author, which are benefits to data quality, but the narratives still had to be processed with NLP.

Preprocessing Data

Using raw data extracted from social media or Web forums was not straightforward. "Preprocessing" the data was necessary and consisted, for example, of clarifying abbreviations and checking spelling mistakes.

As shown in Table 4, a number of types of transformations were performed on the extracted data.

Table 4. Transformations performed on the extracted data.

Transformation	Rationale and methods
Anonymization	Anonymization is required to remove patients' personal data to comply with medical confidentiality. Benton's team trained a classifier to determine if a token had to be anonymized or not [38]. Liu and Chen, only, did not extract the author pseudonyms [44], but they did not apply anonymization to the narratives.
Spelling correction	To maximize the detection of information in the corpus, spelling mistakes and typing errors that are common in texts extracted from social networks have to be corrected. The analyzed texts were extracted from social networks or public forums and included many abbreviations and typing errors. Li [43] applied this method to medical words that were often misspelled in messages.
Cleaning Web pages	Web pages consist of hundreds of tags that are invisible to users. When the crawler extracted a complete Web page code, a cleaning step was necessary to refine the content, as with Benton et al [38] and Liu and Chen [44].
Stemming	Reducing inflected words to their root helps to detect different forms of a word. This process reduces words to their word stem, base, or root forms, and these roots were then used for analysis. Different algorithms can be used by the «stemmer» [38,42,45,47,50]. For example, Benton et al [38] and Leaman et al [42] used the «Porter stemmer».
Sentencization/ Tokenization	Breaking the text up into segments of words, sentences, and paragraphs allows for analyzing the sentences and locutions in the corpus. Liu and Chen [44] used sentences at the information extraction level. Similarly, Benton et al [38] and Leaman et al [42] relied on a window of, respectively, 20 and 5 tokens in which the drug and the event co-occurred.
	Sentencization and tokenization are also documented in Liu and Chen [44], Nikfarjam and Gonzalez [45], and Yeleswarapu et al [50].

Annotation

Annotation of the corpus, for instance, identification of adverse events and drugs in messages, was performed in all of the studies reviewed in this theme. Annotation was realized by (1) machine-learning algorithms [39,44] and (2) final statistical evaluation [38-50].

Out of 13 studies, 9 (69%) used standard medical terminology, including Cerner Multum's Drug Lexicon, UMLS, side effect resource (SIDER), Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART), and MedDRA. Of the 13 studies, 8 (62%) took into account lay language. Among these 13 studies, 7 (54%) used lay vocabulary. Of the 13 that were originally selected, 3 (23%) studies [38,44,48] used a consumer health vocabulary [51], 1 of the 13 (8%) [49] used MedSyn [52], and 3 of the 13 (23%) [40,42,43] used a custom-built vocabulary. Of the 13 studies, 3 (23%) [39,40,44] mapped lay language to medical terminologies using MetaMap [53].

Relationships Between Drugs and Events

The relationship between the drug and the medical term was then analyzed. This relationship could have been an indication (ie, the drug was taken to treat the symptom or the disease), a cause (ie, the drug caused the pathology, in this case, an ADR), or a question about a potential causal relationship.

The methods were classified into two categories. The first category corresponds to methods that assessed a relationship between the medication and the event (ie, machine learning, association rules), which were used in 7 studies out of 13 (54%), with machine learning being used in 5 of 13 (38%) publications.

When this approach was used, the evaluation was done thanks to cross-validation (3/13, 23%). The second category corresponds to exploratory analysis to identify main safety themes from the corpus of messages (ie, statistically significant co-occurrences) [38,40].

Results Evaluation

In the studies that used a computerized approach, the evaluation of results was based on precision (9/13, 69%), recall (9/13, 69%), f-measure (6/13, 46%), accuracy (3/13, 23%), both true- and false-positive rates (1/13, 8%), log-likelihood ratios (1/13, 8%), support (1/13, 8%), confidence (1/13, 8%), leverage (1/13, 8%), and Bayesian confidence propagation neural network (BCPNN) scores and variance (1/13, 8%).

From the initial data volume, authors selected test sets on which they evaluated their systems. The sizes of the test samples were much smaller than the initial volumes of the extracted data. For example, Hadzi-Puric and Grmusa [40] and Li [43] used the whole initial data volume, whereas Yates [49] used only 480 posts out of the initial 400,000 extracted posts.

The pharmacovigilance database that was used for comparison was the FAERS [54] in 4 of the 13 (31%) studies. In 7 other studies out of 13 (54%), the annotators had varying expertise levels—from medical school students to pediatric clinicians and those with PhDs. Benton [38] referred to the tables and notes contained on the drug labels. Overall, only Li [43] did not document the constitution of a new gold standard or the use of an existing standard.

A majority of studies (7/13, 54%) were not only about expected ADRs, but also about discovering relationships between drugs and adverse events that had not been documented on the drug labels or in the literature.

Discussion

Gaps

This scoping review revealed some gaps among the selected studies that could be challenging to fill.

Although some studies that were related to the identification theme concluded that patients' comments posted in social media contained interesting data for pharmacovigilance (ie, potential unexpected ADRs, patients' risk perceptions, effects on adherence), they usually failed to accurately assess the completeness, quality, and reliability of these data. We could highlight the near absence of accessible information related to chronology (ie, time to onset, dechallenge, rechallenge) or differential diagnosis that would be necessary to assess the causal relationship between a drug and an AE. Moreover, evaluations of the seriousness and unexpectedness of ADRs was available in only a few studies. Finally, we retrieved no study that took into account exposures during pregnancy and only one study that partly focused on drug-drug or food-drug interactions.

More than the quality of the information shared in social media, issues can be raised about the reliability of this information. Indeed, social media users adopt pseudonyms, which may allow malicious persons to spread false rumors using multiple pseudonyms with limited risk of being identified as the origin of the rumor.

Furthermore, a user can post the same message twice or more on the same forum or on different forums using the same or different pseudonyms with no malevolent intent simply to maximize their chances of obtaining an answer. Consequently, it would be interesting to identify these duplicates. We found only one study in which an algorithm that addressed data redundancy was implemented but not described [40], and removing duplicates was seldom reported as an issue, for example, in Pages et al [35].

Regarding the extraction theme, we identified a set of processing steps that are used to process social media data after the Web crawling step and that could be recommended:

1. Anonymization: this was performed in only 2 studies out of 13 (15%), suggesting that privacy of data was not a major issue for the authors, who considered using pseudonyms to be sufficient for preserving confidentiality; nevertheless, it should be considered in every study that includes personal identifiers.

2. Preprocessing step: checking spelling errors and typographical errors; stemming, sentencization, and tokenization to process social media data.

3. Annotation and use of existing medical terminology.

Because none of the selected articles reported on a method that encompassed all of the steps we considered key, we assume that refining current methods and tools is desirable to improve the quality of processed data.

We also noticed that implementation did not follow a generic approach, which would be necessary for easily adding new sites or data sources. This is understandable in the context of a research project, but genericity should be addressed if more sites are intended to be included in the general pharmacovigilance process.

Finally, from the studies returned by our citation database queries, no study used comments on video-sharing websites as a source of data.

Limitations

The methodology has at least two limitations. First, when we constituted the research team, we were not exhaustive regarding the stakeholders we included in this review. For example, we lack stakeholders from regulatory agencies, from the pharmaceutical industry, and from patients or patient associations. The second limitation relates to the citation searches. We limited ourselves to PubMed and Embase. Although both of these resources offer a wide range of citations, we potentially missed some citations in the field, as illustrated by the fact that we selected two additional articles that were not found by the queries or by screening the citations. Moreover, the query itself was not trivial because the field is still a new research area. Finally, by using PubMed and Embase, we could not find any analyses of ADRs using social media that were conducted confidentially within company safety departments.

Perspectives

Emerging evidence on the effectiveness of social media for surveillance suggests that mining messages posted on social media may be helpful for complementing pharmacovigilance systems. Examples of information retrieval from social media have previously been shown in other domains. For instance, it has been demonstrated that Tweets and restaurant reviews might aid in identifying and taking action on localized foodborne illnesses [55-57].

Adverse drug reactions are serious, underreported public health problems with high health and financial costs. A number of authors often described the cases of ADRs reported in social media as insufficiently informative to effectively assess a causal association with the drugs, compared with classical reporting in which quality criteria are available [58,59]. Moreover, extracting ADRs from social media presents specific technical constraints, given the unstructured information, compared with electronic health records or pharmacovigilance databases. Finally, spelling errors and patients' expressions [60,61] make extraction even more difficult.

However, our study confirms that there is a sufficient volume of data on pharmacovigilance in social media to work with and that quantity may eventually support the pharmacovigilance process despite variable quality. Whereas some websites may collect huge amounts of poorly documented posts, others such as PatientsLikeMe [60] collect very complete and high-quality data on drug treatments and, therefore, present very interesting possibilities for improving our knowledge on ADRs based on reliable information. It is thus necessary to further evaluate the quality of the different websites to fulfill the expectations of a new data source for pharmacovigilance.

Among the conceivable solutions for increasing reliability, we can suggest the use of comments' metadata (eg, pseudonym, date, and eventually the location given in the profile) to detect duplicated posts from the same author.

Indeed, the objective is to use social media as an additional source of data to expedite signals of potential ADRs. Local pharmacovigilance departments nationwide collect data on adverse events to track cases and interpret data for surveillance. Social media may help to detect the misuse or abuse (including overdose) of drugs [62,63] and adverse effects that would otherwise go unreported (eg, ADRs that are not serious but can impair the patients' quality of life and the adherence to treatment).

In order to verify the reliability of data retrieved online, comparison of this data with established sources, like FAERS or SIDER, as realized by several authors to derive reference material, can also be useful to detect new knowledge and improve quality of documentation of already described ADRs.

Social media may also provide new information on polypharmacy in real life, especially on the concomitant use of prescription drugs and self-medication drugs, and its consequences for patients, such as drug-drug interactions.

Nevertheless, it is necessary to verify how this new data source could be integrated into regular pharmacovigilance systems, with the aim of detecting, verifying, or validating signals.

Moreover, a number of authors highlighted the necessity of considering the context associated with the drug prescription, including whether any ADRs have been described in the media or discussed by regulatory agencies, to interpret the findings. Through the example of benfluorex, Abou Taam et al [28] analyzed narratives that were posted on French websites and reported drastic changes in consumers' risk perceptions following media coverage. As such, social media may be analyzed to assess consumers' behaviors and their risk perceptions and, finally, guide public communication campaigns.

Finally, a broader use of the Internet may include additional sources, such as soliciting reporting studies [64] we excluded, crowdsourcing [65] that may be complementary to social media, or Web search queries [66].

Conclusions

We conducted a scoping review to explore the potential interest in social media as a new source of data in pharmacovigilance and to define the methods for extracting data from this source. The exploratory aspect of the scoping review helped to give us an overview of this field, and this was a mandatory first step when we began our own work in the field. We are currently developing methods and tools within the Adverse Drug Reactions from Patient Reports In Social Media (ADR-PRISM) and *Vigilance dans les Forums sur le Médicament* (Vigi4MED) projects to collect data from social media and to evaluate the data's potential interest for pharmacovigilance. This scoping review was beneficial for identifying gaps in previous studies and designing our work plan.

Among the studies that were related to extraction, the oldest one was published in 2010, which shows that this field is still new and suggests that we can expect numerous further developments and improvements to come.

Finally, it appears that there are still outstanding questions about the data collected from social media and that there is sufficient room for improving extraction systems. Depending on the measured characteristics of social media as a new data source for pharmacovigilance and the headway in extracting ADRs, pharmacovigilants will have to define the role of social media in the classical pharmacovigilance system.

Acknowledgments

This work was funded by the grant AAP-2013-052 from the French agency for drug safety, *Agence Nationale de Sécurité du Médicament et des Produits de Santé* (ANSM), through the Vigi4MED project, and by the *Direction Générale des Entreprises* (DGE) and territorial collectivities (*Ile de France and Haute Normandie*) under the 16th *Fond Unique Interministériel* (FUI) request for proposals through the ADR-PRISM project.

Conflicts of Interest

Kappa Santé, the company that developed the Detec't tool, which extracts data from messages related to potential ADRs in social media, employs Redhouane Abdellaoui and Nathalie Texier. The other authors have no conflicts of interest with the subject matter discussed in the manuscript.

Multimedia Appendix 1

Characteristics of included studies from theme 1 (identification).

[XLSX File (Microsoft Excel File), 26KB-Multimedia Appendix 1]

Multimedia Appendix 2

Characteristics of included studies from theme 2 (extraction).

[XLSX File (Microsoft Excel File), 17KB-Multimedia Appendix 2]

Multimedia Appendix 3

Preprocessing transformations and analyses used in theme 2 (extraction) studies.

[XLSX File (Microsoft Excel File), 13KB-Multimedia Appendix 3]

References

1. World Health Organization. Geneva, Switzerland: World Health Organization Pharmacovigilance URL: http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/ [accessed 2015-06-23] [WebCite Cache ID 5zXGgxwBG]
2. World Health Organization. Geneva, Switzerland: World Health Organization English Glossary URL: http://www.who.int/medicines/areas/coordination/English_Glossary.pdf [accessed 2015-06-23] [WebCite Cache ID 6XyLFSz6M]
3. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood) 2011 Apr;30(4):581-589 [FREE Full text] [doi: 10.1377/hlthaff.2011.0190] [Medline: 21471476]
4. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. BMJ 2004 Jul 3;329(7456):15-19 [FREE Full text] [doi: 10.1136/bmj.329.7456.15] [Medline: 15231615]
5. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 1998 Apr 15;279(15):1200-1205. [Medline: 9555760]
6. Mitchell AS, Henry DA, Sanson-Fisher R, O'Connell DL. Patients as a direct source of information on adverse drug reactions. BMJ 1988 Oct 8;297(6653):891-893 [FREE Full text] [Medline: 3140967]
7. Medawar C, Herxheimer A. A comparison of adverse drug reaction reports from professionals and users, relating to risk of dependence and suicidal behaviour with paroxetine. Int J Risk Saf Med 2003;16(1):5-19.
8. van Grootheest K, de Jong-van den Berg L. Patients' role in reporting adverse drug reactions. Expert Opin Drug Saf 2004;3(4):363-368. [Medline: 15268652]
9. Blenkinsopp A, Wilkie P, Wang M, Routledge PA. Patient reporting of suspected adverse drug reactions: a review of published literature and international experience. Br J Clin Pharmacol 2007 Feb;63(2):148-156. [doi: 10.1111/j.1365-2125.2006.02746.x] [Medline: 17274788]
10. van Hunsel F, Talsma A, van Puijenbroek E, de Jong-van den Berg L, van Grootheest K. The proportion of patient reports of suspected ADRs to signal detection in the Netherlands: case-control study. Pharmacoepidemiol Drug Saf 2011 Mar;20(3):286-291. [doi: 10.1002/pds.2092] [Medline: 21351310]
11. Hughes S, Cohen D. Can online consumers contribute to drug knowledge? A mixed-methods comparison of consumer-generated and professionally controlled psychotropic medication information on the internet. J Med Internet Res 2011;13(3):e53 [FREE Full text] [doi: 10.2196/jmir.1716] [Medline: 21807607]
12. Margraff F, Bertram D. Adverse drug reaction reporting by patients: an overview of fifty countries. Drug Saf 2014 Jun;37(6):409-419. [doi: 10.1007/s40264-014-0162-y] [Medline: 24748428]
13. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc 2003;10(2):115-128 [FREE Full text] [Medline: 12595401]
14. Tubert P, Bégaud B, Péré JC, Haramburu F, Lellouch J. Power and weakness of spontaneous reporting: a probabilistic approach. J Clin Epidemiol 1992 Mar;45(3):283-286. [Medline: 1569425]
15. Tubert P, Bégaud B. Random models for margins of a 2 x 2 contingency table and application to pharmacovigilance. Stat Med 1991 Jun;10(6):991-999. [Medline: 1876790]

16. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 2008;15(1):87-98 [FREE Full text] [doi: 10.1197/jamia.M2401] [Medline: 17947625]
17. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salamé G, EU-ADR group. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf 2009 Dec;18(12):1176-1184. [doi: 10.1002/pds.1836] [Medline: 19757412]
18. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010;160(Pt 1):739-743. [Medline: 20841784]
19. Islamaj Doğan R, Névéol A, Lu Z. A context-blocks model for identifying clinical relationships in patient records. BMC Bioinformatics 2011;12 Suppl 3:S3 [FREE Full text] [doi: 10.1186/1471-2105-12-S3-S3] [Medline: 21658290]
20. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics 2012;3(1):15 [FREE Full text] [doi: 10.1186/2041-1480-3-15] [Medline: 23256479]
21. Warrer P, Hansen EH, Juhl-Jensen L, Aagaard L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. Br J Clin Pharmacol 2012 May;73(5):674-684. [doi: 10.1111/j.1365-2125.2011.04153.x] [Medline: 22122057]
22. Eriksson R, Jensen PB, Frankild S, Jensen LJ, Brunak S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. J Am Med Inform Assoc 2013;20(5):947-953 [FREE Full text] [doi: 10.1136/amiajnl-2013-001708] [Medline: 23703825]
23. Li Y, Salmasian H, Vilar S, Chase H, Friedman C, Wei Y. A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. J Am Med Inform Assoc 2014;21(2):308-314 [FREE Full text] [doi: 10.1136/amiajnl-2013-001718] [Medline: 23907285]
24. Micoulaud-Franchi JA. [One step more toward pharmacovigilance 2.0. Integration of web data community for a pharmacovigilance more alert] [Article in French]. Presse Med 2011 Sep;40(9 Pt 1):790-792. [doi: 10.1016/j.lpm.2011.07.001] [Medline: 21802246]
25. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
26. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
27. Daudt HML, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. BMC Med Res Methodol 2013;13:48 [FREE Full text] [doi: 10.1186/1471-2288-13-48] [Medline: 23522333]
28. Abou TM, Rossard C, Cantaloube L, Bouscaren N, Roche G, Pochard L, et al. Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator®) withdrawal in France. J Clin Pharm Ther 2014 Feb;39(1):53-55. [doi: 10.1111/jcpt.12103] [Medline: 24304185]
29. Butt TF, Cox AR, Oyebode JR, Ferner RE. Internet accounts of serious adverse drug reactions: a study of experiences of Stevens-Johnson syndrome and toxic epidermal necrolysis. Drug Saf 2012 Dec 1;35(12):1159-1170. [doi: 10.2165/11631950-000000000-00000] [Medline: 23058037]
30. 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 May;37(5):343-350 [FREE Full text] [doi: 10.1007/s40264-014-0155-x] [Medline: 24777653]
31. Kheloufi F, Jean-Pastor MJ. Sharing experience with adverse drug reactions on internet: Which adverse drug reactions involving statins are described by patients on internet-based-forums? Fundam Clin Pharmacol 2014 May;28(Supplement s1):56 Abstracts of the 18th Annual Meeting of French Society of Pharmacology and Therapeutics, 81st Annual Meeting of Society of Physiology, 35th Pharmacovigilance Meeting, 15th APNET Seminar, 12th CHU CIC Meeting and 9th Annual Meeting of Physiology, Pharmacology and Therapeutics, , 22–24 April 2014, Poitiers, France [FREE Full text]
32. Kmetz J. Visible Technologies. 2013 Dec 26. Adverse event reporting: What pharmaceutical companies need to know URL: http://www.visibletechnologies.com/blog/adverse-event-reporting-pharma/ [accessed 2015-06-23] [WebCite Cache ID 6SftTDhuQ]
33. Mao JJ, Chung A, Benton A, Hill S, Ungar L, Leonard CE, et al. Online discussion of drug side effects and discontinuation among breast cancer survivors. Pharmacoepidemiol Drug Saf 2013 Mar;22(3):256-262 [FREE Full text] [doi: 10.1002/pds.3365] [Medline: 23322591]
34. Moncrieff J, Cohen D, Mason JP. The subjective experience of taking antipsychotic medication: a content analysis of Internet data. Acta Psychiatr Scand 2009 Aug;120(2):102-111. [doi: 10.1111/j.1600-0447.2009.01356.x] [Medline: 19222405]
35. Pages A, Bondon-Guitton E, Montastruc JL, Bagheri H. Undesirable effects related to oral antineoplastic drugs: comparison between patients' internet narratives and a national pharmacovigilance database. Drug Saf 2014 Aug;37(8):629-637. [doi: 10.1007/s40264-014-0203-6] [Medline: 25027671]
36. Schröder S, Zöllner YF, Schaefer M. Drug related problems with Antiparkinsonian agents: consumer Internet reports versus published data. Pharmacoepidemiol Drug Saf 2007 Oct;16(10):1161-1166. [doi: 10.1002/pds.1415] [Medline: 17486665]

37. Vaughan Sarrazin MS, Cram P, Mazur A, Ward M, Reisinger HS. Patient perspectives of dabigatran: analysis of online discussion forums. Patient 2014;7(1):47-54. [doi: 10.1007/s40271-013-0027-y] [Medline: 24030706]
38. Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, et al. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. J Biomed Inform 2011 Dec;44(6):989-996 [FREE Full text] [doi: 10.1016/j.jbi.2011.07.005] [Medline: 21820083]
39. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing (SHB '12). New York, NY: ACM; 2012 Presented at: 2012 International Workshop on Smart Health and Wellbeing (SHB '12); October 29-November 2, 2012; Maui, HI. [doi: 10.1145/2389707.2389713]
40. Hadzi-Puric J, Grmusa J. Automatic drug adverse reaction discovery from parenting websites using disproportionality methods. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis Mining (ASONAM 2012). Washington, DC: IEEE Computer Society; 2012 Presented at: 2012 International Conference on Advances in Social Networks Analysis Mining (ASONAM 2012); August 26-29, 2012; Istanbul, Turkey. [doi: 10.1109/ASONAM.2012.144]
41. Jiang Y, Liao QV, Cheng Q, Berlin RB, Schatz BR. Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. AMIA Annu Symp Proc 2012;2012:417-426 [FREE Full text] [Medline: 23304312]
42. 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 (BioNLP '10). Stroudsburg, PA: ACL; 2010 Presented at: 2010 Workshop on Biomedical Natural Language Processing (BioNLP '10); July 15, 2010; Uppsala, Sweden.
43. Li YA. DSpace@MIT. Cambridge, MA: Massachusetts Institute of Technology; 2011 Feb. Medical data mining: improving information accessibility using online patient drug reviews URL: http://dspace.mit.edu/handle/1721.1/66437 [accessed 2015-06-23] [WebCite Cache ID 6Sfu51j6d]
44. Liu X, Chen H. AZDrugMiner: An information extraction system for mining patient-reported adverse drug events in online patient forums. In: Proceedings of the 2013 International Conference on Smart Health (ICSH'13). Berlin, Heidelberg, Germany: Springer; 2013 Presented at: 2013 International Conference on Smart Health (ICSH'13); August 3-4, 2013; Beijing, China. [doi: 10.1007/978-3-642-39844-5_16]
45. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. AMIA Annu Symp Proc 2011;2011:1019-1026 [FREE Full text] [Medline: 22195162]
46. Sampathkumar H, Chen XW, Luo B. Mining adverse drug reactions from online healthcare forums using hidden Markov model. BMC Med Inform Decis Mak 2014;14:91 [FREE Full text] [doi: 10.1186/1472-6947-14-91] [Medline: 25341686]
47. Wu H, Fang H, Stanhope SJ. Exploiting online discussions to discover unrecognized drug side effects. Methods Inf Med 2013;52(2):152-159. [doi: 10.3414/ME12-02-0004] [Medline: 23450374]
48. Yang CC, Jiang L, Yang H, Tang X. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In: Proceedings of the ACM SIGKDD Workshop on Health Informatics. New York, NY: ACM; 2012 Presented at: ACM SIGKDD Workshop on Health Informatics; August 12, 2012; Beijing, China.
49. Yates A, Goharian N, Frieder O. Extracting adverse drug reactions from forum posts linking them to drugs. In: Proceedings of the 2013 ACM SIGIR Workshop on Health Search Discovery. New York, NY: ACM; 2013 Presented at: 2013 ACM SIGIR Workshop on Health Search Discovery; August 1, 2013; Dublin, Ireland. [doi: 10.1145/2484028.2484220]
50. Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R. A pipeline to extract drug-adverse event pairs from multiple data sources. BMC Med Inform Decis Mak 2014;14:13 [FREE Full text] [doi: 10.1186/1472-6947-14-13] [Medline: 24559132]
51. Consumer Health Vocabulary Initiative. URL: http://www.consumerhealthvocab.org/ [accessed 2015-06-23] [WebCite Cache ID 6Sfy1bTCY]
52. Yates A, Goharian N. ADRTrace: Detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In: Proceedings of the 35th European Conference on Advances in Information Retrieval (ECIR'13). Berlin, Heidelberg, Germany: Springer-Verlag Berlin; 2013 Presented at: 35th European Conference on Advances in Information Retrieval (ECIR'13); March 24-27, 2013; Moscow, Russia p. 24-27. [doi: 10.1007/978-3-642-36973-5_92]
53. MetaMap. URL: http://metamap.nlm.nih.gov/ [accessed 2015-06-23] [WebCite Cache ID 6SfyBf7Xq]
54. US Food and Drug Administration. Silver Spring, MD: US FDA FDA Adverse Event Reporting System (FAERS) URL: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm [accessed 2014-09-18] [WebCite Cache ID 6Sfxrss3N]
55. Harris JK, Mansour R, Choucair B, Olson J, Nissen C, Bhatt J, Centers for Disease Control and Prevention. Health department use of social media to identify foodborne illness - Chicago, Illinois, 2013-2014. MMWR Morb Mortal Wkly Rep 2014 Aug 15;63(32):681-685 [FREE Full text] [Medline: 25121710]
56. Kuehn BM. Agencies use social media to track foodborne illness. JAMA 2014 Jul;312(2):117-118. [doi: 10.1001/jama.2014.7731] [Medline: 24963655]
57. Chapman B, Raymond B, Powell D. Potential of social media as a tool to combat foodborne illness. Perspect Public Health 2014 Jul;134(4):225-230. [doi: 10.1177/1757913914538015] [Medline: 24990140]

58. Edwards IR, Lindquist M, Wiholm BE, Napke E. Quality criteria for early signals of possible adverse drug reactions. Lancet 1990 Jul 21;336(8708):156-158. [Medline: 1973481]
59. Franzen W. Can social media benefit drug safety? Drug Saf 2011 Sep 1;34(9):793. [doi: 10.2165/11595510-000000000-00000] [Medline: 21830842]
60. Smith CA, Wicks PJ. PatientsLikeMe: Consumer health vocabulary as a folksonomy. AMIA Annu Symp Proc 2008:682-686 [FREE Full text] [Medline: 18999004]
61. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med 2002;41(4):289-298. [Medline: 12425240]
62. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through Twitter. J Med Internet Res 2013;15(9):e189 [FREE Full text] [doi: 10.2196/jmir.2741] [Medline: 24014109]
63. Hanson CL, Burton SH, Giraud-Carrier C, West JH, Barnes MD, Hansen B. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. J Med Internet Res 2013;15(4):e62 [FREE Full text] [doi: 10.2196/jmir.2503] [Medline: 23594933]
64. Härmark L, van Puijenbroek E, van Grootheest K. Intensive monitoring of duloxetine: results of a web-based intensive monitoring study. Eur J Clin Pharmacol 2013 Feb;69(2):209-215 [FREE Full text] [doi: 10.1007/s00228-012-1313-7] [Medline: 22688722]
65. Gottlieb A, Hoehndorf R, Dumontier M, Altman RB. Ranking adverse drug reactions with crowdsourcing. J Med Internet Res 2015;17(3):e80 [FREE Full text] [doi: 10.2196/jmir.3962] [Medline: 25800813]
66. Yom-Tov E, Gabrilovich E. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. J Med Internet Res 2013;15(6):e124 [FREE Full text] [doi: 10.2196/jmir.2614] [Medline: 23778053]

Abbreviations

ab,ti: abstract, title
ADE: adverse drug event
ADR: adverse drug reaction
ADR-PRISM: Adverse Drug Reactions from Patient Reports In Social Media
AE: adverse event
ANSM: the French drug safety agency (French acronym of Agence Nationale de Sécurité du Médicament et des Produits de Santé)
AP-HP: Assistance Publique-Hôpitaux de Paris
API: application programming interface
BCPNN: Bayesian confidence propagation neural network
CHU: University Hospital Center (French acronym of Centre Hospitalier Universitaire)
COSTART: Coding Symbols for Thesaurus of Adverse Reaction Terms
DGE: block grants for investment expenditures (French acronym of Direction Générale des Entreprises)
FAERS: FDA Adverse Drug Event Reporting System
FDA: Food and Drug Administration
FUI: French acronym of Fond Unique Interministériel
HEGP: Hôpital Européen Georges-Pompidou
INSERM: the French National Institute for Health and Medical Research (French acronym of Institut National de la Santé et de la Recherche Médicale)
LIMICS: Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé
MedDRA: Medical Dictionary for Regulatory Activities
MeSH: Medical Subject Heading
NLP: natural language processing
PI-Doc: Problem Intervention Documentation
SIDER: side effect resource
SPC: Summary of Product Characteristics
TIAB: title and abstract
SSPIM: Department of Public Health and Medical Informatics (French acronym of Service de Santé Publique et de l'Information Médicale)
UMLS: Unified Modeling Language System
UMR_S: Unité Mixte de Recherche en Santé
UPMC: University of Pierre and Marie Curie
Vigi4MED: French acronym of Vigilance dans les Forums sur le Médicament

Edited by G Eysenbach; submitted 30.01.15; peer-reviewed by N Dasgupta, M Chary; comments to author 27.02.15; revised version received 09.04.15; accepted 22.04.15; published 10.07.15

Please cite as:
Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, Jaulent MC, Beyens MN, Burgun A, Bousquet C
Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review
J Med Internet Res 2015;17(7):e171
URL: http://www.jmir.org/2015/7/e171/
doi: 10.2196/jmir.4304
PMID: 26163365