Published in final edited form as:
Drug Saf. 2014 October ; 37(10): 777–790. doi:10.1007/s40264-014-0218-z.

Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art

Rave Harpaz[1], **Alison Callahan**[1], **Suzanne Tamang**[1], **Yen Low**[1], **David Odgers**[1], **Sam Finlayson**[1], **Kenneth Jung**[1], **Paea LePendu**[1], and **Nigam H. Shah**

[1]Center for Biomedical Informatics Research, Stanford University, Stanford, CA

Abstract

Text mining is the computational process of extracting meaningful information from large amounts of unstructured text. Text mining is emerging as a tool to leverage underutilized data sources that can improve pharmacovigilance, including the objective of adverse drug event detection and assessment. This article provides an overview of recent advances in pharmacovigilance driven by the application of text mining, and discusses several data sources—such as biomedical literature, clinical narratives, product labeling, social media, and Web search logs—that are amenable to text-mining for pharmacovigilance. Given the state of the art, it appears text mining can be applied to extract useful ADE-related information from multiple textual sources. Nonetheless, further research is required to address remaining technical challenges associated with the text mining methodologies, and to conclusively determine the relative contribution of each textual source to improving pharmacovigilance.

1. Introduction

Text Mining is defined as the *process of extracting meaningful information from large amounts of unstructured text using computational methods* [1, 2]. For pharmacovigilance, we define “meaningful information” as information that can support adverse drug event (ADE) detection and assessment. Because text mining provides a mechanism to transform free-text into computable knowledge, text mining is emerging as a way to explore, analyze, query, and manage underutilized safety information about drugs.

Pharmacovigilance presently relies on the analysis of clinical trials and spontaneous reports, and to some degree on the review of biomedical literature. The analysis is typically performed by domain experts on a manual case-by-case basis. Recently, statistical techniques have been incorporated into routine pharmacovigilance and applied to spontaneous reports[3, 4] and clinical trials[5] to identify signals of ADEs. Nonetheless, well recognized limitations[6, 7] inherent to the type and diversity of data sources employed

Corresponding author: Rave Harpaz, 1265 Welch Road, Stanford CA 94305-5479, T 650-723-4155, F 650-725-7944, rharpaz@stanford.edu.

Conflict of Interest

Nigam H Shah is a Science Advisor to Apixio Inc (www.apixio.com), and Kyron Inc (www.kyron.com). Rave Harpaz is an employee of Oracle Health Sciences. Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu and Nigam H. Shah have no other conflicts of interest that are directly relevant to the content of this article.

in routine pharmacovigilance, along with increased public concern over the safe use of drugs, have stimulated several worldwide research and legislative initiatives[8, 9] with the objective of improving pharmacovigilance. It is widely accepted that progress in pharmacovigilance depends on a comprehensive approach that examines ADE-related information from a diverse set of potentially complementing data sources.

With the passing of the Food and Drug Administration (FDA) Amendments Act (FDAAA) of 2007[10], research in pharmacovigilance has centered on the expanded secondary use of electronic health records (EHRs)[11–13]. In recent years, other sources such as the biomedical literature, product labels, content from social media, and the logs of information seeking activities on the Web have been researched[7] to support holistic pharmacovigilance (Figure 1). Each source provides a unique vantage point, and each source has unique strengths and limitations.

EHRs hold the promise of active surveillance, have the ability to quantify the incidence or risk of ADEs, can identify patients at risk, and have the potential to provide more accurate and earlier ADE detection. The biomedical literature is a burgeoning information source that through case reports, clinical studies, and observational studies, has enabled safety evaluators to assess potential ADEs. In contrast with the prevailing manual use, it is possible to computationally harness the biomedical literature for various pharmacovigilance purposes, including signal detection[14, 15]. Product labels contain a broad array of information, ranging from adverse drug reactions to drug efficacy, risk mitigation, contraindications, drug interactions, and more. Several initiatives have emerged to computationally extract information from product labels in order to create a knowledgebase of known ADEs[16, 17]. The resulting knowledgebase can be used for ADE assessment, to derive benchmarks for signal detection, to prioritize and filter ADEs under investigation, and to detect class effects. Lastly, there are calls[18, 19] to investigate the use of online patient generated data, which hold the promise of earlier ADE detection for certain types of events (e.g., more common or milder events). The proposed data sources include the social media, e.g., patients' experiences with medications that are explicitly shared via online health forums and social networks, and the implicit health information contained in the search logs of popular search engines.

The key challenge in using the aforementioned data sources for pharmacovigilance is that a large proportion of their content is stored as *free-text*. Unlike data that is typically stored in relational databases, free-text is unstructured, is subject to the complexities and variability of natural language, and challenging to deal with algorithmically.

In this article we provide an overview of recent advances in pharmacovigilance driven by the application of text mining. We begin our discussion with a brief review of text mining as a process and its application to biomedicine. We then cover the state of the art, centering on data sources that are currently researched for pharmacovigilance: biomedical literature, product labeling, clinical narratives, social media, and Web search logs. The manuscript is organized according to the key data sources, and concludes with a perspective highlighting future directions. Rather than exhaustively listing all relevant work, we present a synopsis of papers that reflect exemplary recent research, thus highlighting the motivation for using a particular data source, the role of text mining, the approaches used for text mining, and the associated challenges.

2. Text Mining Overview

To meet the challenges posed by unstructured text, text mining employs a wide range of statistical, machine learning, and linguistic techniques that are associated with natural language processing (NLP). It is beneficial to think of text mining as a process that uses tools, methods, and heuristics developed by those who research the processing of natural language. Depending on the use case, text mining workflows can use NLP methods of differing degrees of sophistication. Therefore, unlike classic NLP, which employs sophisticated language models and computationally expensive syntactic and semantic analyses to extract meaning from text, text mining leans towards the implementation of simpler but less costly approaches that scale to large data sets.

A text mining process typically starts with several pipelined NLP subtasks that are used to format the text in preparation for the statistical analysis or pattern discovery phase. The subtasks include a set of foundational *low-level* syntactic tasks, and a set of *high-level* tasks that build on the low-level tasks and involve semantic processing. Common subtasks and a representative pipeline are illustrated in Figure 2. A brief description of these tasks is provided in Table 1, and a comprehensive review thereof is provided by Friedman[20] and Nadkarni[21] et al. While the exact set of components included in a text mining pipeline is application specific, the key ingredients relevant to biomedical text mining appear to be *named entity recognition* (NER) and *relation detection* (defined in Table 1).

Structured domain knowledge in the form of biomedical ontologies plays a key role in NER and other text mining subtasks. Two major public resources for biomedical ontologies are the National Library of Medicine (NLM) Unified Medical Language System (UMLS)[22] and the National Center for Biomedical Ontologies (NCBO) BioPortal[23]. The UMLS Metathesaurus is a compendium of over 150 controlled vocabularies (or ontologies) and contains close to three million biomedical concepts that are associated with synonyms, semantic groups, and relationships between concepts. Similarly, BioPortal is an open repository of over 380 biomedical ontologies that are made available in computationally useful forms. BioPortal also supports a wide range of Web services that enable investigators to use ontologies for text mining applications.

The majority of biomedical text mining applications rely on *dictionary-based* approaches for NER, which draw on comprehensive vocabularies such as those in the UMLS or BioPortal. Such reliance mirrors the findings from the 2008 i2b2 NLP challenge[24], where the organizers write that “Most of the factual and objective pieces of information were identified by simple rule-based systems armed with dictionaries of terms and negation extraction modules”. Vocabularies typically used for pharmacovigilance are displayed in Table 2. A dictionary for NER is typically created by using the entries of one or more of these vocabularies. A central challenge in dictionary-based NER is customizing these dictionaries to maximize their effectiveness for specific applications[25–28].

The simplest and fastest approach to relation detection is co-occurrence analysis, which makes the assumption that terms that co-occur in the same text tend to be related[29, 30]. The degree of co-occurrence can be quantified statistically to rank and eliminate weak co-occurrences. More accurate though slower relation detection can be achieved by applying rule-based or machine-learning-based approaches[29, 30]. Rule-based approaches make use of general knowledge about biomedical entities or language structure to find explicit statements in the text about relationships of interests. A simple rule-based approach might search for hard coded patterns in the text, e.g., <drug> *induces* <disease> or <drug> *treats* <disease>. More sophisticated approaches use linguistic and semantic analyses via part of speech (POS) tagging and parse trees[31, 32]. Machine-learning-based approaches draw on classifiers that operate over POS tags, parse trees, N-grams, terms frequencies, and other textual constructs. Machine-learning-based approaches typically achieve better results but require a large amount of manually annotated training data that is costly to acquire.

An expanding set of tools are currently available for the NER step in biomedical text mining. A popular tool for NER is NLM's MetaMap[33] that maps word phrases in text to UMLS concepts and assigns a score to each mapping. It uses a configurable set of NLP steps such as tokenization, shallow parsing, POS tagging, negation detection, and word sense disambiguation to perform the task. Similarly, NCBO's Annotator[34] is a Web service that recognizes biomedical ontology terms in text, using ontologies available in BioPortal. NegEx[35] and ConText[36] are popular tools for negation detection, and in the case of ConText, also for the identification of experiencer (e.g., patient or other) and temporality (e.g., recent or historical). Both NegEx and ConText use trigger terms to qualify the value of a concept, e.g., '*ruled out*' to locate negated concepts in text. Other tools for biomedical text mining are available at the Online Registry of Biomedical Informatics Tools (ORBIT)[37], some of which are also available packaged into an NLP toolkit by the iDASH center[38].

3. Biomedical Literature

MEDLINE is NLM's publicly available electronic database of approximately 20 million biomedicine and life sciences articles. It is an expanding data collection currently comprised of about 340,000 ADE specific articles, with roughly 13,000 new ADE-related articles indexed each year (Figure 3). Each article is annotated by trained NLM indexers with key MeSH subject headings, subheadings, and supplementary concepts, referred to as MeSH annotations. The subject headings may capture clinical manifestations (CM), drugs, or drug classes. Subheadings are used to narrow the scope of the main headings, e.g., the main/subheading combinations "Cyclooxygenase Inhibitors/adverse effects" and "Myocardial Infarction/chemically induced" used to index an article, suggest that the article content is about the ADE relationship between COX-2 inhibitors and myocardial infarction.

Computational approaches to extract ADEs from MEDLINE either use MeSH annotations, or process free-text in article titles and abstracts (Figure 4). Using MeSH annotations has the advantage that they are readily available, human-curated, are based on the full text of articles (not just abstracts), and do not face the difficulties associated with processing free-text. Conversely, processing the free-text in article titles and abstracts is not limited by the scope and granularity of the MeSH vocabulary or by the NLM annotation rules.

Avillach et al.[15] devised an ADE identification process based entirely on MeSH annotations. The subheadings “chemically induced” and “adverse effects”, and the “pharmacological action” MeSH relationship were used to link drugs and CMs in an article as potential ADEs. Their approach was evaluated against a reference set of 61 drug-event pairs that comprised the EU-ADR gold standard[39]. By establishing a threshold of three articles whose MeSH annotations contained a drug-event pair of interest, Avillach et al. achieved a sensitivity of 90% and a specificity of 100%. While MeSH subheadings and MeSH relations provide important cues to pair ADE-related drugs and CMs, they do not directly specify which of the annotated drugs is related to an annotated CM. Thus, relying exclusively on MeSH annotations may pair up unrelated drugs and events. Shetty et al.[14] applied a set of successive filters to refine the process of extracting ADEs using MeSH annotations. Annotated drug-CM pairs where the CM is the drug indication were removed based on information from product labels. A machine learning approach operating on MeSH derived features was applied to classify drug-CM pairs as either ADE-related or not. Then, by applying Disproportionality Analysis (DPA) to the filtered set of drug-CM pairs they demonstrate that their method detects ADEs with over 70% sensitivity and 40% positive predictive value as assessed against a reference set of ADEs derived from the “Warnings” section of drug labels. They also show that 54% of the associations analyzed could have been detected before the warnings were issued, and that the Rofecoxib–myocardial Infarction association could have been identified using MEDLINE several years prior to the Rofecoxib recall.

In contrast, several projects relied exclusively on processing abstracts’ free-text. Through a series of reports Gurulingappa et al.[40] describe a methodology to detect ADEs from case-report abstracts. Controlled vocabularies derived from DrugBank and MedDRA were used for NER. A manually curated set of 2972 abstracts containing 12,046 drug-event relations[41] was used to train and test a machine learning ADE relation detection methodology that resulted in an F-score of 0.87. DPA was applied to the extracted drug-event pairs, and evaluated against a set of 62 ADEs communicated through recent label changes. The results were compared with signal detection applied to the FDA Adverse Event Reporting System (FAERS) and the Yellow Cards spontaneous reporting system (SRS). Their method was able to predict 30% of the label changes versus 48% using the FAERS or the Yellow Cards SRS. They also note that by the taking the union of signals from all the three sources 76% of the label changes were predicted.

Xu and Wang[42] demonstrate that MEDLINE can be used to boost signals from FAERS. A key assumption in their work is that drug-event pairs that co-occur in both FAERS and MEDLINE likely represent true relationships. A UMLS derived lexicon of drugs and events was used to perform NER. The Stanford parser[43] was used to syntactically parse sentences in over 20 million abstracts. The original FAERS signal scores were squared when the underlying drug-event pairs also appeared in MEDLINE. Using SIDER[44] as a gold standard, they demonstrate improved precision but lower recall compared to using the original FAERS signals.

Duke et al.[45] propose an approach to predict drug-drug interactions (DDIs) associated with myopathy by cross-referencing information from MEDLINE and EHRs. They used a rule-based NLP technique to identify pairs of drugs that interact with the same CYP450 enzymes in MEDLINE abstracts, inferring a DDI when such pairs were identified. They identified 13,197 DDIs involving 232 FDA approved drugs, of which 3670 pairs were found to be co-prescribed in the EHR dataset. Five of these pairs were found to increase the risk for myopathy relative to the prescription of either drug individually.

Given the current state of the art, it is unclear which of the two approaches—MeSH annotations versus processing abstracts—is better for extracting ADEs. There are currently no studies that directly compare the two approaches; thus outlining an area that may merit future research. Nevertheless, there are studies that considered the use of *both* approaches concurrently. Wang et al.[46] propose a classifier that operates on both MeSH annotations and textual information from the abstracts and titles. They demonstrate that the use of both types of features leads to improved performance for identifying ADE relationships.

4. Product Labeling

A key requirement of effective pharmacovigilance efforts is an accurate knowledgebase of known ADEs, indications, and other drug-related information that is ideally in a machine-readable format. For example, there is a critical need for accurate reference sets ("gold standards") to evaluate signal detection and risk estimation methodologies. An authoritative database of known ADEs would enable the derivation of such reference sets[16]. The knowledgebase could also aid drug safety evaluators to asses or prioritize ADEs under consideration, and is a core component of medication-related decision support systems that are being developed to promote medication safety[47]. A public database that meets these needs is currently not available but there are efforts to create one by harvesting information extracted from product labels.

Product labels, also referred to as package inserts, are an authoritative source of information about the risks, benefits, and pharmacological properties of drugs. The DailyMed Website[48] maintained by the NLM and the FDA provides downloadable electronic versions of product labels called Structured Product Labels (SPLs) for most drugs sold in the US. However, the SPLs provide structure only for the sections of the label (e.g., Indications and Usage, Clinical Pharmacology, Warnings, Precautions, Adverse Reactions). The content of the individual sections is still in free-text format.

The side effect resource (SIDER)[44] is a publicly available database containing ADEs text-mined from several public sources including the SPLs. The original version used a custom dictionary derived from the UMLS to perform NER on the Adverse Reaction section of SPLs. SIDER has been used in numerous studies as a reference set to evaluate signal detection algorithms. However, the level of credibility attributed to ADEs can vary based on the location of their mentions, e.g., in Boxed Warning, Warnings, or Adverse Reactions sections. In addition, product labels contain findings reported in clinical trials, many of which lack validation. Therefore, it is inadvisable to use SIDER or any other SPL-based extraction as a reference set without further verification or quantification of the degree of confidence in a specific drug-event pair.

Duke et al.[17] developed a SPL processing tool called the Structured Product Label Information Coder and ExtractoR (SPLICER). Tagging of adverse events (AEs) is accomplished by a set of specific rules tailored to the different sections and formatting structures (e.g., tables, lists) of the SPL. Event frequency and other qualifiers are also extracted. SPLICER demonstrated high accuracy in AE extraction, with a sensitivity of 93% and PPV of 95%, and was used in various projects including the formation of the OMOP gold standard[49, 50] and in assessing the labeling consistency of bio-equivalent drugs[51]. While not yet publicly available, the tool as well as the resulting knowledgebase can be obtained by contacting the authors.

In related work Fung et al.[47] from the NLM proposed a text mining pipeline called the Structured Product Labels eXtractor (SPL-X) to extract indications, noting that AEs could be extracted using the same approach. SPL-X uses open-source tools such as NegEx and MetaMap to identify medical concepts from the Indications section. SPL-X demonstrated precision and recall of 0.95 and 0.77 respectively, noting that the main sources of error were ambiguous terms (e.g., '*strain'* incorrectly identified as muscle strain when it was bacterial-related) and negation detection.

Smith et al.[52] focused on identifying the challenges associated with the extraction and representation of ADEs and indications from publically available sources including MeSH annotations, NDF-RT relationships ('induces', 'may treat'), and SPLs processed using the KnowledgeMap Concept Identifier[53]—an NLP tool developed at Vanderbilt University. The authors highlight complex logical and temporal sentence structures in SPL, such as "Do not take drug X after event Y occurs due to increased chance of event Z" or "useful in preventing XYZ in the setting of condition ABC", which standard NLP approaches fail to handle properly. They find that the three data sources (MeSH, NDF-RT and SPLs) agree on less than 1% of the indication and ADE relationships extracted. They attribute this problem to the mappings and granularity of concepts in the UMLS used to encode the extracted terms. For example, myalgia and musculoskeletal pain are assigned different UMLS concept codes and will be identified as part of two different ADEs.

5. Clinical Narratives

EHRs contain a longitudinal record of clinical data from routine clinical care. While some information is structured, a significant portion of the EHR remains in narrative formats. Much of the information that is critical to risk assessments such as signs and symptoms, disease status and severity, and medical history are typically only in narrative text. In comparison, coded discharge diagnoses and claims data (also used in pharmacovigilance) have relatively low sensitivity for detecting ADEs[26, 54], weaker coverage of symptomatology, and are vulnerable to inaccuracies as they are oriented toward billing. Consequently, clinical narratives offer tremendous potential for pharmacovigilance.

Clinical narratives introduce unique challenges in comparison to other biomedical corpora. For example, physicians document the relevant medical history of the patient as well as their family members. They also document the process of elimination inherent in differential diagnosis, noting conditions that are ruled out or symptoms that the patient denies. It has been estimated that more than 40% of conditions are negated within clinical narratives and can be detected using algorithms like NegEx[35]. Thus, clinical NLP tools should take into account negated mentions as well as other contextual cues that indicate historical information or that relate to the experiencer (e.g., the patient's current problem list versus those of his or her family members)[36]. Furthermore, unlike other data sources in which each document can be taken independently, clinical narratives are tied to a patient who accumulates many such documents over time. Thus, temporal order matters: a drug mentioned in one encounter could be significantly linked to an event mentioned later in a different document. Finally, clinical narratives are subject to biases introduced by local documentation procedures as well as the working of the health system (e.g. a note only gets written when a patient interacts with the health system)[55].

Some of the pioneering applications of clinical narratives for pharmacovigilance can be traced back to the use of classic medical NLP systems such as MedLEE[56]—designed to identify clinical concepts and their modifiers (e.g., negation, body location, time of occurrence, certainty of finding) in clinical narratives, and map them to UMLS concepts.

In an early feasibility study, Wang et al.[57] applied MedLEE to 25,074 discharge summaries from New York Presbyterian Hospital (NYPH) to identify ADEs associated with 7 drugs. Of 132 ADEs identified by statistical analysis over MedLEE's output, 31% were known. Follow-up studies using MedLEE and NYPH narratives focused on approaches to address confounding—a major methodological challenge in observational research. Haerian et al.[58] used an expert-generated list of known risk factors for the events investigated, to identify and exclude patients with predisposing conditions. Based on manual review of 275 random cases their approach yielded a sensitivity of 93.8% and a specificity of 91.8%. Li[59] and Harpaz[60] et al. used regression models to estimate confounding-adjusted association statistics, as well as to automatically select model variables (potential confounders). Li et al.[59] applied their method to 264,155 MedLEE-processed patient records and based on a manually curated set of known ADEs associated with rhabdomyolysis and pancreatitis, their method resulted in a precision of 83.3% and 60.8% respectively, exceeding the performance of four competing methodologies.

Although applications using clinical narratives began with NLP systems such as MedLEE, which provide a linguistic analysis of clinical text and hence incur an increased computational cost, recent studies have demonstrated the use of simpler but less computationally intensive techniques.

LePendu et al.[61] describe a highly scalable workflow to process clinical text as well as count patients corresponding to specific conditions, and demonstrate its efficacy for pharmacovigilance. The workflow was applied to 11 million clinical narratives (spanning 18 years and 1.8 million patients) from the Stanford Translational Research Integrated Database Environment (STRIDE)[62] to recognize present, positive mentions of medical terms. Building on these term-mentions, associations were estimated by matching patients using propensity scores and by keeping track of the temporal ordering of drug/indication/event mentions. Their approach was evaluated using the EU-ADR reference set[39] (augmented with additional test cases) and resulted in an area under the receiver operating characteristic curve (AUC)—which is a measure of the ability to distinguish a true association from a negative control—of 0.84. They also demonstrate that six of nine investigated ADEs could have been identified earlier than the date an official alert was issued.

In subsequent work also using STRIDE, Iyer et al.[63] demonstrate the applicability of the same workflow for detecting adverse DDIs and Jung et al.[64] train a highly accurate classifier for detecting off-label drug uses. The use of drugs for unapproved indications, called off-label use, is problematic because such uses have not been evaluated for safety and efficacy. Iyer et al. estimated the strength of the association between a particular drug combination and a particular event by comparing the number of patients who experienced the event and were taking both drugs with the number of patients who did not experience the event and were taking either of the two drugs. The approach was evaluated against a reference set made of 1165 drugs, 14 events, and 1698 drug-drug-event test cases compiled from existing knowledge sources, and resulted in an AUC of 0.82, slightly better than that of adverse DDI detection based solely on FAERS. Jung et al. found 403 novel off-label uses that were validated in independent data sources, and sort them by risk of adverse events and the cost of the drug to prioritize these off-label uses for further investigation.

Finally, a novel way to use EHRs, including clinical narratives, for pharmacovigilance is in combination with other data sources. Harpaz et al.[65, 66] demonstrate that combining ADE signals from EHRs and FAERS leads to a substantial improvement in signal detection accuracy.

A central challenge to continued progress in the field is limited access to clinical narratives. Typically only researchers affiliated with a medical center can access clinical notes, and care organizations are reluctant to share clinical notes even when they are de-identified[67]. There is also a need for broader availability of curated clinical datasets to accelerate methodological research.

6. Social Media

The rapid expansion of online social media, such as forums, blogs, and social networks is changing the way we gather information about diseases and treatment options, as well as how we share our personal health experiences with others. The Pew Research Center's survey, *The Social Life of Health Information*[68], found that 2% of patients and 6% of caregivers share their experiences online, and that 18% of all internet users, 31% of all patients with chronic conditions, and 38% of caregivers look at online drug reviews. This increasing presence of social media is offering new opportunities for public health surveillance that are internet-based, patient-generated, unsolicited, and up-to-date. The main technical, policy, and privacy challenges associated with the use of social media for pharmacovigilance have been recently discussed in an editorial by Edwards and Lindquist[69].

As early as 2002, Medawar et al.[70] reviewed posts to an online discussion board to validate a relationship between suicidality and the antidepressant paroxetine. The authors concluded that the user reports contained clear evidence of an association that a SRS in place at that time had not detected. In 2005, two researchers wrote a letter to the FDA about numerous reports in FAERS for bisphosphonates associated with severe bone, joint, and muscle pain[71]. Several years later, a public FDA alert lead to the further investigation by hospital staff of one case from Massachusetts General Hospital[72]. When the hospital staff contacted the patient, the patient pointed them to similar reports on the website Askapatient.com. The hospital carried out a follow-up survey specific to the question of joint, bone and muscle pain associated with bisphosphonate, which was completed by almost 400 Askpatient.com bisphosphonate users in a 3 month period. About 60% reported muscle and joint pain along with fatigue[72]. Also using Askapatient.com, Moncrieff et al.[73] analyzed 223 comments related to antipsychotic medications. The authors reported that although ADEs related to this type of medication are common, studies have found that clinicians under-report ADEs related to antipsychotic medication.

From a text mining perspective, the colloquial language employed by online users presents a particularly significant challenge. Leaman et al.[74] proposed an approach to extract AEs from posts in DailyStrength using a custom lexicon as a basis for NER. Colloquial terms and their clinical equivalents were manually curated from a sample of posts (e.g., “zoned out” to mean “somnolent”). Mentions of conditions were categorized as AEs, indications, or beneficial outcomes through a rule-based approach that uses cues from nearby terms (e.g., “taking for seizures” implying that seizures was the indication). The authors reported an F-measure of 74%, noting that the main sources of error were due to colloquial phrases not included in their original lexicon and due to ambiguous terms (e.g., “worrying about a low” where low is the event). In related work, Yang et al.[75] utilized the Consumer Health Vocabulary[76] to map lay language onto medical lexicons for AE extraction.

Another major challenge is the ability to distinguish real experiences from hearsay, non-personal experiences, or media stimulated reports. Liu et al.[77] proposed an ADE extraction approach for forum data called AZDrugMiner, which uses a set of machine learning methods (operating on POS tags and parse trees) for relation extraction and for distinguishing real experiences from other reports. They conclude that applying these methods improves the accuracy of ADE extraction.

Since the work of Leaman et al.[74], text mining has been applied to DailyStrength.com[74, 78, 77], Yahoo Wellness Groups[79], Askapatient.com, Medications.com, WebMD.com[80], parenting websites[81], and various disease specific forums such as for cancer[82] and diabetes[77].

Research on using microblogs (e.g., Twitter) and other general purpose social networking sites (e.g., Facebook) is the newest category for ADE detection via text mining. Similar to forum data, colloquial text and non-experiential reporting are common, but additional challenges related to data volume must be addressed. For example, real-time surveillance using Twitter feeds would require processing 58 million Tweets each day[83]. Additional considerations include the Tweet’s short character length, which restricts the amount of information that is posted, as well as custom source-specific constructions such as “#”, “FF”, or “RT”.

Although a practical system based on general purpose social media has yet to be demonstrated, pilot studies have established proof of concept. Bian et al.[84] and Jiang et al.[85] proposed methods to mine Twitter. Both approaches focused on a small set of drugs, employed MetaMap to perform NER, and used machine learning methods to identify posts of real experiences based on semantic features generated by MetaMap (e.g., presence and frequency of UMLS semantic types such as '*disease or syndrome*'), along with other features such as the number and type of pronouns mentioned (assumed indicative of real experiences). Bian et al. focused on experimental drugs (studied in clinical trials at the time of evaluation) and collected 2 billion Tweets corresponding to 18 months of data, which were analyzed on a high performance computing platform. Pimpalkhute et al.[86] focused on methods to address the challenges posed by colloquial language. For four commonly prescribed drugs, Prozac, Paxil, Seroquel, and Olanzapine, they describe a method to generate the most likely lexical variants based on edit distances and filtering using the Metaphone phonetic algorithm. All these studies reported reasonable accuracy in extracting ADEs, and despite the use of a small sample of tweets, identified a large number of posts about ADEs.

7. Web Search Logs

A 2009 study by the Centers for Disease Control and Prevention estimated that 61% of adults search the Web for health and medical related information[87]. Another study by the Pew Research Center in early 2013 reported that 72% of Internet users claimed to search online for health information, and that 8 in 10 online health inquiries start at a search engine[88]. Search logs are used in the Google Flu Trends project, demonstrating that statistics of influenza-related search terms recorded by search engines can be used to track rates of influenza[89]. Similarly, it is conceivable that analyzing the volume and content of search queries about medications and medical conditions may provide early clues about ADEs as patients engage search engines in an effort to learn about medications that they are using and medical conditions they experience.

White et al. present two studies[90, 91] that examine the feasibility of a signal detection system based on search logs. Both studies were based on the analysis of search queries mentioning drugs, symptoms, and medical conditions that were issued to the Google, Bing, and Yahoo! search engines by 80 million consenting (and anonymized) users over a period of 18 months prior to the time of analysis. The first study[90] characterized the discriminatory power of signal detection via search logs by using known drug interactions (and controls) for 62 drug pairs associated with hyperglycemia. The study also demonstrated that the analysis of search logs could identify a drug interaction between paroxetine and pravastatin reported to cause hyperglycemia in advance of its publication[92] (though the interaction is yet to be confirmed by regulatory agencies). The second study[91] focused on ADEs associated with single drugs, and revised the signal detection methodology to incorporate temporal information and to include several safeguards to counter confounding effects. Using the OMOP gold standard[50], the authors demonstrate that the accuracy of signal detection based on search logs is comparable to that of FAERS, and that by jointly leveraging signals from FAERS and search logs, the accuracy of signal detection can be improved by 19% over the use of each data source independently. To analyze the search logs, terms corresponding to the drugs, conditions, and symptoms of interest in query logs were identified using sets of synonyms automatically generated from medical ontologies available through BioPortal. The synonyms were supplemented with consumer-oriented search terms derived from Bing's query-click logs by identifying all results clicked on after a certain query, and then identifying other query terms that lead to the same pages (e.g., "bleeding stomach ulcers" for the event upper gastrointestinal bleeding).

Users may search on medications, symptoms, and disorders for a variety of reasons, beyond the reason of having taken a medication and experiencing symptomatology. A central challenge in the use of search logs for pharmacovigilance, as with the use of social media, is the ability to reliably distinguish users who are experiencing adverse effects, versus engaged in an information seeking exploration of conditions and medications. However, given the increasing use of search engines for understanding medical conditions, the analysis of search logs for early warning of a drug's adverse events is an exciting area of research.

8. Conclusion

The content of many new data sources that can support a more robust and holistic approach to pharmacovigilance is in free-text format. Furthermore, the majority of these data sources were not created primarily for pharmacovigilance, thus necessitating text-mining. The availability of structured biomedical domain knowledge and relatively mature text-processing tools offer a viable solution to effectively processing free-text, and create a unique opportunity to leverage textual data sources for pharmacovigilance.

Given the diversity of text genres employed, text mining for pharmacovigilance is not a homogeneous undertaking and each data source comes with unique challenges. For example, the limited access to several of the data sources, e.g., EHRs and online content, is one of the main impediments. This is why the formation of research alliances, such as the Observational Health Data Science Initiative (www.ohdsi.org) for the case of EHRs, can accelerate research in this field by aligning the strategic interests of multiple stakeholders. The use of online content, such as social media and Internet search logs poses several technical challenges that go beyond those associated with traditional NLP, including the processing of colloquial language, and distinguishing experiential from non-experiential reported ADEs. The biomedical literature introduces the trade-off between using MeSH annotations versus performing NLP on the abstracts, and it is unclear which of the approaches is better. Last, there is a need for consistent ways to evaluate text mining methodologies—a problem that is general to the field of pharmacovigilance [4, 7]. There is also the issue of the research focus: for example, most existing efforts focus on single drug ADEs. Given that a large proportion of the population are on multi-drug regimens, pharmacovigilance research efforts should consider the use of text mining methodologies for the study and safety profiling of multi-drug combinations.

Other textual corpora mined for pharmacovigilance or potential candidates thereof that we were unable to cover, include: narratives of spontaneous reports [93], regulatory documents such as new drug applications (NDA) to the FDA[94], European public assessment reports for medicines[95], regular safety summaries[96, 97], labeling information from drugs@FDA, and clinical trial report narratives[98].

The ultimate goal of pharmacovigilance is to identify ADEs as early as possible and with high fidelity. Therefore, it is critical to understand how the use of different data sources will advance this goal. To our knowledge, strategies for combining the safety information generated by different sources are yet to be established. It is clear that the main use cases for leveraging multiple data modalities are (1) more efficient dissemination of safety evidence, and (2) improved signal detection via evidence combination. In this regard, questions that need to be researched are: the relative value of each data source for the two use cases, the relative utility of the data sources over each other for the surveillance of specific events or drug classes, and whether the use of some data sources should be reserved for hypothesis generation while others reserved for confirmation.

Lastly, pharmacovigilance is an evolving discipline and text mining can play a key role in its transformation. While text mining does vary in complexity, a large body of research has demonstrated that with existing text-processing tools it is possible to extract useful safety-related information from the aforementioned textual sources. We hope that this review demystifies text mining and outlines the opportunities as well as the challenges that lie ahead.

Acknowledgments

The writing of this manuscript was supported by NIH grant U54-HG004028 for the National Center for Biomedical Ontology, and by NIGMS grant GM101430-01A1.

References

1. Kroeze, JH.; Matthee, MC.; Bothma, TJD. Differentiating data- and text-mining terminology; Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology. 954024: South African Institute for Computer Scientists and Information Technologists; 2003. p. 93-101.
2. Witten, IH. "text mining". In: Singh, MP., editor. Practical handbook of internet computing. Boca Raton, Florida: Chapman & Hall/CRC Press; 2005. 14-1 - -22.
3. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf. 2002; 25(6):381–392. [PubMed: 12071774]
4. Harpaz R, Dumouchel W, Lependu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. Clinical pharmacology and therapeutics. 2013; 93(6):539–546. [PubMed: 23571771]
5. DuMouchel W. Multivariate Bayesian Logistic Regression for Analysis of Clinical Study Safety Issues. Statist Sci. 2012; 27(3):319–339.
6. Honig PK. Advancing the science of pharmacovigilance. Clinical pharmacology and therapeutics. 2013; 93(6):474–475. [PubMed: 23689213]
7. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical pharmacology and therapeutics. 2012; 91(6):1010–1021. [PubMed: 22549283]
8. Prescription Drug User Fee Act (PDUFA V). [Accessed Apr 2014] http://www.fda.gov/ForIndustry/UserFees/PrescriptionDrugUserFee/ucm272170.htm.

9. REGULATION (EU) No 1235/2010 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. 2010 Dec 15. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000492.jsp.
10. [Accessed Apr 2014] Food and Drug Administration Amendments Act (FDAAA) of 2007. http://www.fda.gov/regulatoryinformation/legislation/federalfooddrugandcosmeticactfdcact/significantamendmentstothefdcact/foodanddrugadministrationamendmentsactof2007/default.htm.
11. Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The New Sentinel Network - Improving the Evidence of Medical-Product Safety. New England Journal of Medicine. 2009; 361(7):645–647. [PubMed: 19635947]
12. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. Annals of Internal Medicine. 2010; 153(9) 600-W206.
13. Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. PharmacoepidemiolDrug Saf. 2011; 20(1):1–11.
14. Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. Journal of the American Medical Informatics Association. 2011; 18(5):668–674. [PubMed: 21546507]
15. Avillach P, Dufour JC, Diallo G, Salvo F, Joubert M, Thiessard F, et al. Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EUADR project. Journal of the American Medical Informatics Association : JAMIA. 2013; 20(3):446–452. [PubMed: 23195749]
16. Boyce RD, Ryan PB, Noren GN, et al. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Safety. 2014:1–11.
17. Duke JD, Friedlin J. ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. AMIA Annual Symposium proceedings. 2010; 2010:177–181. [PubMed: 21346964]
18. Innovative Medicines Initiative. 9th Call for Proposals 2013. http://www.imi.europa.eu/sites/default/files/uploads/documents/9th_Call/Calll_9_Text.pdf.
19. [Accessed Apr 2014] FDA Science Board Subcommittee: Review of the FDA/CDER Pharmacovigilance Program (Prepared for the FDA Science Board May 2011). http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/ScienceBoardtotheFoodandDrugAdministration/UCM276888.pdf.
20. Friedman, C.; Elhadad, N. Natural Language Processing in Health Care and Biomedicine. In: Shortliffe, EH.; Cimino, JJ., editors. Biomedical Informatics. Springer London: 2014. p. 255-284.
21. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. Journal of the American Medical Informatics Association : JAMIA. 2011; 18(5):544–551. [PubMed: 21846786]
22. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of information in medicine. 1993; 32(4):281–291. [PubMed: 8412823]
23. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research. 2009; 37(Web Server issue):W170–W173. [PubMed: 19483092]
24. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association : JAMIA. 2011; 18(5):552–556. [PubMed: 21685143]
25. Gurulingappa, H.; Klinger, R.; Hofmann-Apitius, M.; Fluck, J., editors. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. 2nd Workshop on Building and evaluating resources for biomedical text mining (7th edition of the Language Resources and Evaluation Conference); 2010.
26. Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. Journal of the American Medical Informatics Association. 2010; 17(6):671–674. [PubMed: 20962129]

27. Xu R, Musen MA, Shah NH. A Comprehensive Analysis of Five Million UMLS Metathesaurus Terms Using Eighteen Million MEDLINE Citations. AMIA Annual Symposium proceedings. 2010; 2010:907–911. [PubMed: 21347110]

28. Wu ST, Liu H, Li D, Tao C, Musen MA, Chute CG, et al. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. Journal of the American Medical Informatics Association : JAMIA. 2012; 19(e1):e149–e156. [PubMed: 22493050]

29. Rodriguez-Esteban R. Biomedical Text Mining and Its Applications. PLoS Comput Biol. 2009; 5(12):e1000597. [PubMed: 20041219]

30. Cohen KB, Hunter L. Getting Started in Text Mining. PLoS Comput Biol. 2008; 4(1):e20. [PubMed: 18225946]

31. Coulet A, Garten Y, Dumontier M, Altman RB, Musen MA, Shah NH. Integration and publication of heterogeneous text-mined relationships on the Semantic Web. Journal of biomedical semantics. 2011; 2(Suppl 2):S10. [PubMed: 21624156]

32. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2012:410–421. [PubMed: 22174296]

33. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA. 2010; 17(3):229–236. [PubMed: 20442139]

34. Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on translational bioinformatics. 2009; 2009:56–60. [PubMed: 21347171]

35. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001; 34(5):301–310. [PubMed: 12123149]

36. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics. 2009; 42(5):839–851. [PubMed: 19435614]

37. [Accessed Apr 2014] Online Registry of Biomedical Informatics Tools. http://orbit.nlm.nih.gov/.

38. iDASH Center. http://idash.ucsd.edu/nlp/natural-language-processing-nlp-ecosystem.

39. Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, et al. A Reference Standard for Evaluation of Methods for Drug Safety Signal Detection Using Electronic Healthcare Record Databases. Drug Safety. 2013; 36(1):13–23. DOI. [PubMed: 23315292]

40. Gurulingappa H, Toldo L, Rajput AM, Kors JA, Taweel A, Tayrouz Y. Automatic detection of adverse events to predict drug label changes using text and data mining techniques. Pharmacoepidemiology and drug safety. 2013; 22(11):1189–1194. [PubMed: 23935003]

41. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of biomedical informatics. 2012; 45(5):885–892. [PubMed: 22554702]

42. Xu R, Wang Q. Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection. BMC bioinformatics. 2014; 15(1):17. [PubMed: 24428898]

43. The Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml.

44. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Molecular systems biology. 2010; 6:343. [PubMed: 20087340]

45. Duke JD, Han X, Wang Z, Subhadarshini A, Karnik SD, Li X, et al. Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. PLoS Comput Biol. 2012; 8(8):e1002614. [PubMed: 22912565]

46. Wang W, Haerian K, Salmasian H, Harpaz R, Chase HS, Friedman C. A Drug-Adverse Event Extraction Algorithm to Support Pharmacovigilance Knowledge Mining from PubMed Citations. Proceedings of the AMIA Annual Symposium. 2011:1464–1470.

47. Fung KW, Jao CS, Demner-Fushman D. Extracting drug indication information from structured product labels using natural language processing. Journal of the American Medical Informatics Association. 2013; 20(3):482–488. [PubMed: 23475786]

48. [Accessed Apr 2014] DailyMed. http://dailymed.nlm.nih.gov/.

49. Friedlin J, Duke J. Applying Natural Language Processing to Extract Codify Adverse Drug Reaction in Medication Labels. http://omop.fnih.org/OMOPWhitePapers2010.

50. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf. 2013; 36(Suppl 1):S33–S47. [PubMed: 24166222]

51. Duke J, Friedlin J, Li X. Consistency in the safety labeling of bioequivalent medications. Pharmacoepidemiology and drug safety. 2013; 22(3):294–301. [PubMed: 23042584]

52. Smith JC, Denny JC, Chen Q, Nian H, Spickard Iii A, Rosenbloom ST, et al. Lessons Learned from Developing a Drug Evidence Base to Support Pharmacovigilance. Applied Clinical Informatics. 2013; 4(4):596–617. [PubMed: 24454585]

53. Denny JC, Smithers JD, Miller RA, Spickard A. “Understanding” Medical School Curriculum Content Using KnowledgeMap. Journal of the American Medical Informatics Association. 2003; 10(4):351–362. [PubMed: 12668688]

54. 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 Affairs. 2011; 30(4):581–589. [PubMed: 21471476]

55. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. Journal of the American Medical Informatics Association. 2013; 20(e2):e232–e238. [PubMed: 24001516]

56. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association : JAMIA. 2004; 11(5):392–402. [PubMed: 15187068]

57. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal of the American Medical Informatics Association : JAMIA. 2009; 16(3):328–337. [PubMed: 19261932]

58. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of Pharmacovigilance-Related Adverse Events Using Electronic Health Records and Automated Methods. Clinical pharmacology and therapeutics. 2012; 92(2):228–234. doi:http://www.nature.com/clpt/journal/v92/n2/suppinfo/clpt201254s1.html. [PubMed: 22713699]

59. 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. Journal of the American Medical Informatics Association : JAMIA. 2014; 21(2):308–314. [PubMed: 23907285]

60. Harpaz, R.; Haerian, K.; Chase, HS.; Friedman, C. Mining electronic health records for adverse drug effects using regression based methods. Proceedings of the 1st ACM International Health Informatics Symposium; 1883008: ACM; Arlington, Virginia, USA. 2010. p. 100-107.

61. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clinical pharmacology and therapeutics. 2013; 93(6):547–555. [PubMed: 23571773]

62. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annual Symposium proceedings. 2009; 2009:391–395. [PubMed: 20351886]

63. Iyer SV, Harpaz R, Lependu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. Journal of the American Medical Informatics Association : JAMIA. 2013

64. Jung K, LePendu P, Chen WS, Iyer SV, Readhead B, Dudley JT, et al. Automated Detection of Off-Label Drug Use. PLoS ONE. 2014; 9(2):e89324. [PubMed: 24586689]

65. Harpaz, R.; DuMouchel, W.; LePendu, P.; Shah, NH. Empirical Bayes Model to Combine Signals of Adverse Reactions; Proc of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13); p. 1339-1347.

66. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. Journal

of the American Medical Informatics Association : JAMIA. 2013; 20(3):413–419. [PubMed: 23118093]

67. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. Journal of biomedical informatics. 2013; 46(5):765–773. [PubMed: 23810857]
68. The Social Life of Health Information, Pew Research Center. http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011.
69. Edwards IR, Lindquist M. Social Media and Networks in Pharmacovigilance. Drug Safety. 2011; 34(4):267–271. [PubMed: 21417499]
70. Medawar C, Herxheimer A, Bell A, Jofre S. Paroxetine, Panorama and user reporting of ADRs: Consumer intelligence matters in clinical practice and post-marketing drug surveillance. The International Journal of Risk and Safety in Medicine. 2002; 15(3):161–169.
71. Wysowski DK, Chang JT. Alendronate and risedronate: Reports of severe bone, joint, and muscle pain. Archives of Internal Medicine. 2005; 165(3):346–347. [PubMed: 15710802]
72. DeMonaco HJ. Patient- and physician-oriented web sites and drug surveillance: Bisphosphonates and severe bone, joint, and muscle pain. Archives of Internal Medicine. 2009; 169(12):1164–1166. [PubMed: 19546419]
73. Moncrieff J, Cohen D, Mason JP. The subjective experience of taking antipsychotic medication: a content analysis of Internet data. Acta Psychiatrica Scandinavica. 2009; 120(2):102–111. [PubMed: 19222405]
74. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. 2010:117–125.
75. Yang, CC.; Yang, H.; Jiang, L.; Zhang, M. Social media mining for drug safety signal detection. Proceedings of the 2012 international workshop on Smart health and wellbeing; 2389714: ACM; Maui, Hawaii, USA. 2012. p. 33-40.
76. Consumer Health Vocabulary. http://consumerhealthvocab.org/.
77. Liu, X.; Chen, H. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums. In: Zeng, D.; Yang, C.; Tseng, V.; Xing, C.; Chen, H.; Wang, F-Y., et al., editors. Smart Health. Lecture Notes in Computer Science. Springer Berlin: Heidelberg; 2013. p. 134-150.
78. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. AMIA Annual Symposium proceedings. 2011:1019–1026. [PubMed: 22195162]
79. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Annual Symposium proceedings. 2011:217–226. [PubMed: 22195073]
80. Liu, J.; Li, A.; Seneff, S. Automatic drug side effect discovery from online patient-submitted reviews: Focus on statin drugs. The First International Conference on Advances in Information Mining and Management; 2011.
81. Hadzi-Puric, J.; Grmusa, J., editors. Automatic Drug Adverse Reaction Discovery from Parenting Websites Using Disproportionality Methods. Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on; 2012 26–29 Aug; 2012.
82. 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 of Biomedical Informatics. 2011; 44(6):989–996.
83. Statistic Brain. http://www.statisticbrain.com/twitter-statistics/.
84. Bian, J.; Topaloglu, U.; Yu, F. Towards large-scale twitter mining for drug-related adverse events. Proceedings of the 2012 international workshop on Smart health and wellbeing; 2389713: ACM; Maui, Hawaii, USA. 2012. p. 25-32.
85. Jiang, K.; Zheng, Y. Mining Twitter Data for Potential Drug Effects. In: Motoda, H.; Wu, Z.; Cao, L.; Zaiane, O.; Yao, M.; Wang, W., editors. Advanced Data Mining and Applications. Lecture Notes in Computer Science. Springer Berlin: Heidelberg; 2013. p. 434-443.

86. Pimpalkhute P, Patki A, Nikfarjam A, Gonzalez G. Phonetic Spelling Filter for Keyword Selection in Drug Mention Mining from Social Media. AMIA TBI Summit. 2014

87. Centers for Disease Control and Prevention (CDC). [Accessed Apr 2014] Use of the Internet for Health Information: United States2009. http://www.cdc.gov/nchs/data/databriefs/db66.htm.

88. Pew Research Center. Pew Internet & American Life Project: Health Online 2013. http://www.pewinternet.org/~/media/Files/Reports/2013/Pew%20Internet%20Health%20Online%20report.pdf.

89. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232) 1012-U4. Doi.

90. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. Journal of the American Medical Informatics Association. 2013

91. White RW, Harpaz R, Shah NH, DuMouchel W, Horvitz E. Toward Enhanced Pharmacovigilance using Patient-Generated Data on the Internet Clinical. Pharmacology & Therapeutics. 2014 (in press).

92. Tatonetti NP, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, et al. Detecting Drug Interactions From Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels. Clin Pharmacol Ther. 2011; 90(1):133–142. [PubMed: 21613990]

93. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. Journal of the American Medical Informatics Association. 2011; 18(5):631–638. [PubMed: 21709163]

94. New Drug Application (NDA). [Accessed Apr 2014] http://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/newdrugapplicationnda/default.htm.

95. European Public Assessment Reports. http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&mid=WC0b01ac058001d125.

96. [Accessed Apr 2014] World Health Organization Pharmaceuticals Newsletter. http://www.who.int/medicines/publications/newsletter/en/.

97. Potential Signals of Serious Risks/New Safety Information Identified from the FDA Adverse Event Reporting System (FAERS). http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/UCM082196.

98. [Accessed Apr 2014] Clinical Trial Reports. http://www.fda.gov/downloads/regulatoryinformation/guidances/ucm129456.pdf.

Key Points

- Text mining is needed in order to leverage several textual data sources that have the potential to improve pharmacovigilance.
- Despite the challenges associated with processing free-text a large body of research has demonstrated that with existing tools it is possible to extract useful safety-related information from these textual sources.
- Key challenges remain in fully realizing the potential of these data sources for improving pharmacovigilance, and for understanding their precise value for pharmacovigilance.

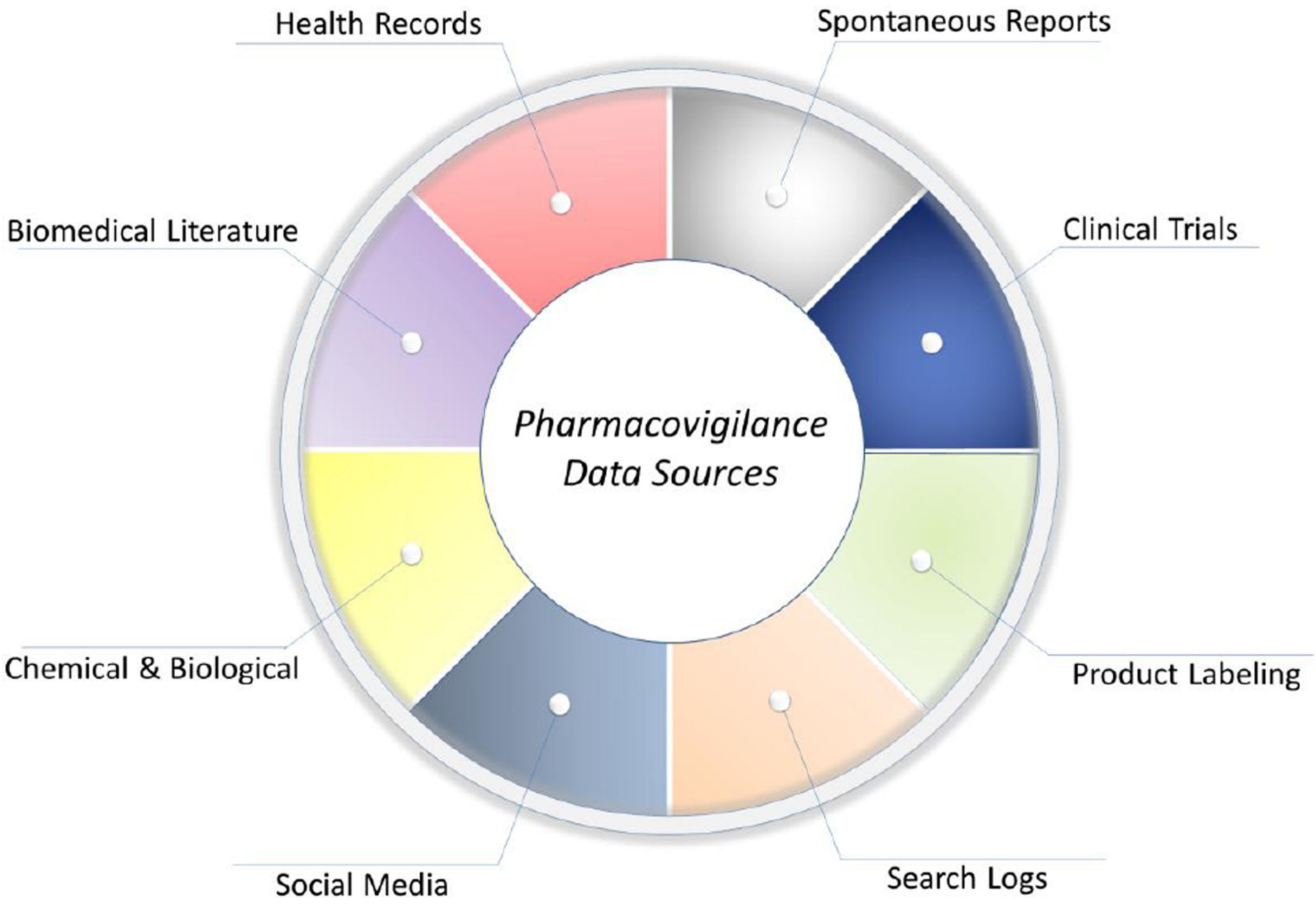


Figure 1.
Data sources currently used or researched to support holistic pharmacovigilance.

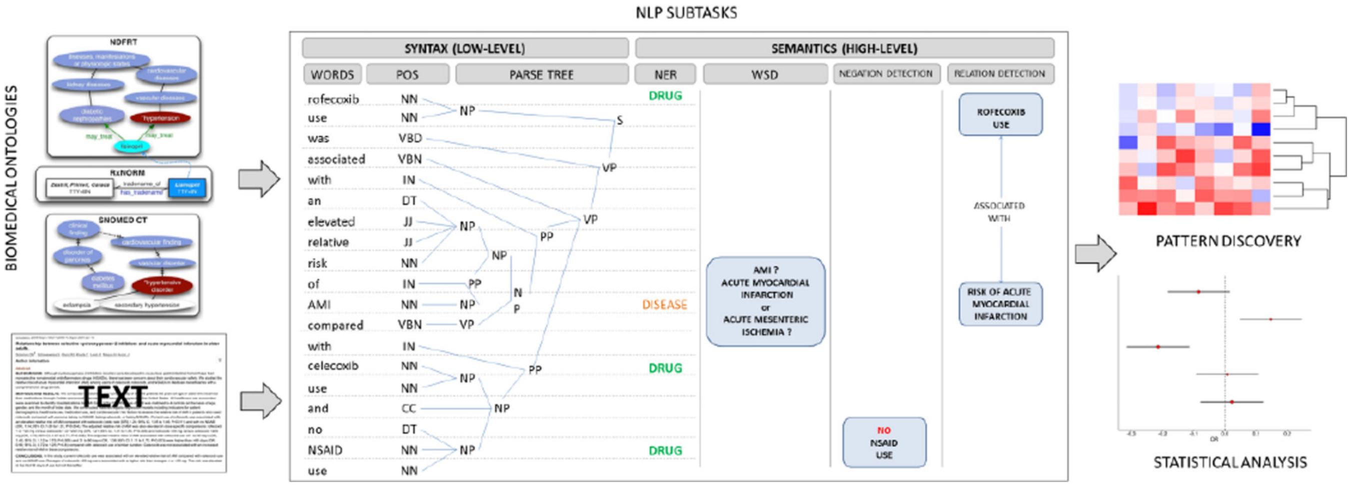


Figure 2.

An example of a biomedical text mining pipeline and common NLP subtasks. The pipeline uses as input the textual corpora to be processed as well as structured domain knowledge in the form of biomedical ontologies. The NLP steps are used to process text in preparation for the statistical analysis or pattern discovery phase. The NLP low-level subtasks include segmentation of documents into sections and sentences and tokenization of sentences into words and punctuation, followed by part-of-speech (POS) tagging and parsing. NLP high-level subtasks operate over the output of the low-level subtasks, and include named entity recognition (NER), word sense disambiguation (WSD), negation detection, temporal inference, and relation detection. The subsequent pattern discovery and statistical analysis can be used, for example, to uncover ADE associations. **POS tags: DT = determiner; IN = preposition or subordinating conjunction; JJ = adjective; NN = noun; VBD = past tense verb; VBN = past participle verb. Parse tags: NP = noun phrase; PP = prepositional phrase; VP = verb phrase.

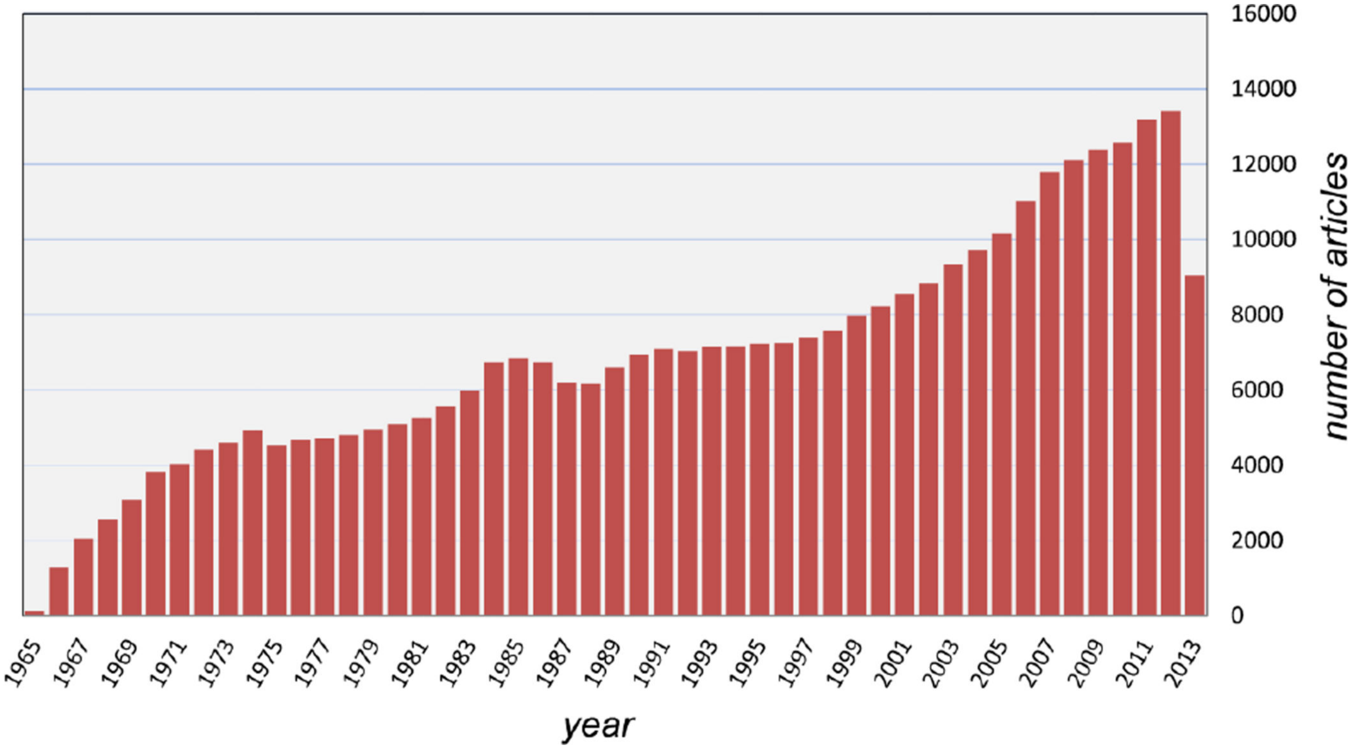
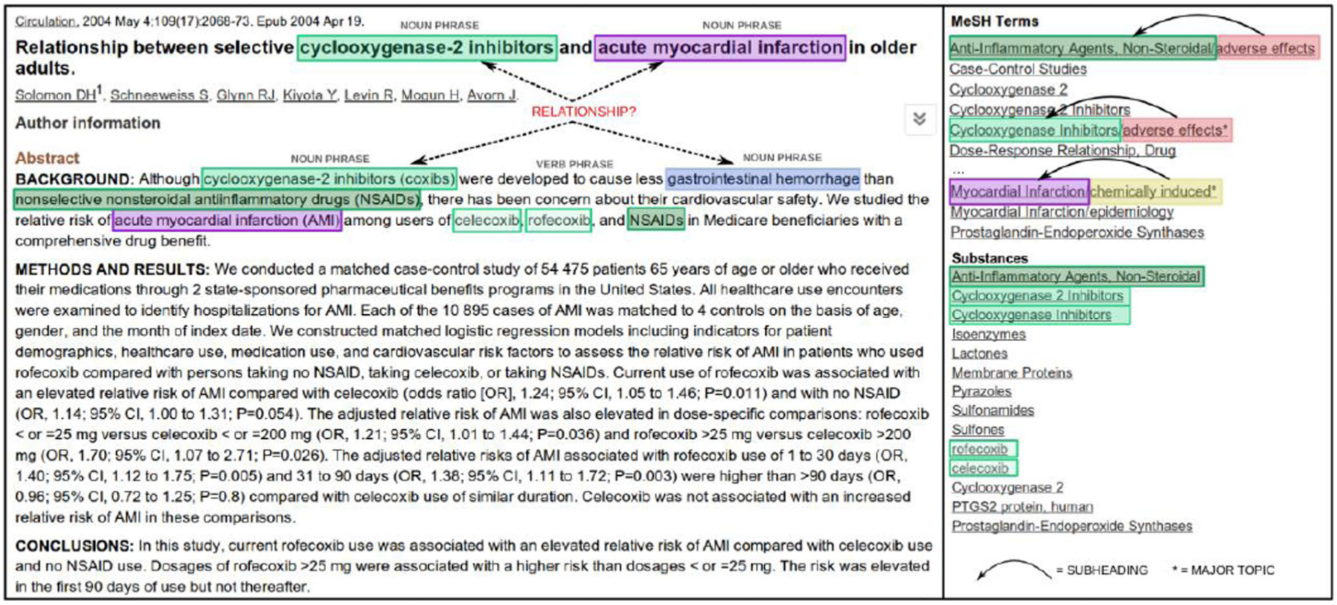


Figure 3.

The growth in number of ADE-related MEDLINE indexed articles over time. The values were obtained through PubMed using the MeSH query: "adverse effects"[Subheading] AND "chemically induced"[Subheading] AND "Chemicals and Drugs Category"[Mesh]. At the time of query (Feb 2014) only a subset of articles published in the year 2013 were indexed.



Circulation. 2004 May 4;109(17):2068-73. Epub 2004 Apr 19.

Relationship between selective cyclooxygenase-2 inhibitors and acute myocardial infarction in older adults.

Solomon DH[1], Schneeweiss S, Glynn RJ, Kiyota Y, Levin R, Mogun H, Avorn J.

Author information

Abstract

BACKGROUND: Although cyclooxygenase-2 inhibitors (coxibs) were developed to cause less gastrointestinal hemorrhage than nonselective nonsteroidal antiinflammatory drugs (NSAIDs), there has been concern about their cardiovascular safety. We studied the relative risk of acute myocardial infarction (AMI) among users of celecoxib, rofecoxib, and NSAIDs in Medicare beneficiaries with a comprehensive drug benefit.

METHODS AND RESULTS: We conducted a matched case-control study of 54 475 patients 65 years of age or older who received their medications through 2 state-sponsored pharmaceutical benefits programs in the United States. All healthcare use encounters were examined to identify hospitalizations for AMI. Each of the 10 895 cases of AMI was matched to 4 controls on the basis of age, gender, and the month of index date. We constructed matched logistic regression models including indicators for patient demographics, healthcare use, medication use, and cardiovascular risk factors to assess the relative risk of AMI in patients who used rofecoxib compared with persons taking no NSAID, taking celecoxib, or taking NSAIDs. Current use of rofecoxib was associated with an elevated relative risk of AMI compared with celecoxib (odds ratio [OR], 1.24; 95% CI, 1.05 to 1.46; P=0.011) and with no NSAID (OR, 1.14; 95% CI, 1.00 to 1.31; P=0.054). The adjusted relative risk of AMI was also elevated in dose-specific comparisons: rofecoxib < or =25 mg versus celecoxib < or =200 mg (OR, 1.21; 95% CI, 1.01 to 1.44; P=0.036) and rofecoxib >25 mg versus celecoxib >200 mg (OR, 1.70; 95% CI, 1.07 to 2.71; P=0.026). The adjusted relative risks of AMI associated with rofecoxib use of 1 to 30 days (OR, 1.40; 95% CI, 1.12 to 1.75; P=0.005) and 31 to 90 days (OR, 1.38; 95% CI, 1.11 to 1.72; P=0.003) were higher than >90 days (OR, 0.96; 95% CI, 0.72 to 1.25; P=0.8) compared with celecoxib use of similar duration. Celecoxib was not associated with an increased relative risk of AMI in these comparisons.

CONCLUSIONS: In this study, current rofecoxib use was associated with an elevated relative risk of AMI compared with celecoxib use and no NSAID use. Dosages of rofecoxib >25 mg were associated with a higher risk than dosages < or =25 mg. The risk was elevated in the first 90 days of use but not thereafter.

MeSH Terms

Anti-Inflammatory Agents, Non-Steroidal/adverse effects
Case-Control Studies
Cyclooxygenase 2
Cyclooxygenase 2 Inhibitors
Cyclooxygenase Inhibitors/adverse effects*
Dose-Response Relationship, Drug
...
Myocardial Infarction/chemically induced*
Myocardial Infarction/epidemiology
Prostaglandin-Endoperoxide Synthases

Substances

Anti-Inflammatory Agents, Non-Steroidal
Cyclooxygenase 2 Inhibitors
Cyclooxygenase Inhibitors
Isoenzymes
Lactones
Membrane Proteins
Pyrazoles
Sulfonamides
Sulfones
rofecoxib
celecoxib
Cyclooxygenase 2
PTGS2 protein, human
Prostaglandin-Endoperoxide Synthases

Figure 4.

Processing free-text in abstracts and titles (left) *versus* using MeSH annotations (right) for a MEDLINE article with PMID 15096449. Occurrences of drugs (*e.g.* rofecoxib), drug classes (*e.g.* NSAIDs) and conditions (*e.g.* myocardial infarction) are colored by identity. Processing the free-text in the abstract and title involves recognizing named entities and determining the kind of relationship that holds between the named entities (dotted lines with arrows). ADE detection via MeSH annotations relies on the ‘adverse effects’ (highlighted in red) and ‘chemically induced’ (in yellow) subheadings, as well as the MeSH Substances entries, to infer ADE relationships between the entities identified in the respective MeSH subject headings (solid lines with arrows).

Table 1

NLP subtasks

Task	**Description**
Segmentation	Splitting a document along sentence and section boundaries
Tokenization	Splitting sentences up into their parts – individual words and punctuation
Part of speech (POS) tagging	Assigning grammatical parts of speech to individual tokens *e.g.* ‘drug’ is a noun, ‘administers’ is a verb, ‘quickly’ is an adjective, ‘the’ or ‘a’ are determiners
Parsing	Determining the grammatical structure of sentences and the relationship between groups of words that together form noun phrases, verb phrases, clauses etc. Shallow parsing, often used instead of deep parsing, only identifies the constituents (e.g., noun phrases) but not the internal structure of the sentence.
Named entity recognition (NER)	Identifying terms or phrases of interest (‘entities’) in the text. NER may go beyond just recognizing terms to also categorizing, normalizing, and mapping them to standardized vocabularies, e.g, identifying ‘rofecoxib’ as a drug, and ‘myocardial infarction’ as a medical condition
Negation detection	Determining whether a named entity is present or absent, e.g. ‘patient does *not* exhibit symptoms of …’, ‘patient was ruled out for myocardial infarction’
Word sense disambiguation (WSD)	Determining which sense of a homograph (words with identical spellings but different meanings) is appropriate in the context of the sentence
Temporal inference	Establishing temporal order of events from text e.g. ‘adverse event occurred *after* prescription of drug’
Relation detection	Determining whether two or more named entities recognized in the text form specific relationships, e.g. ‘drug A *treats* disease B’, ‘drug A *induces* disease B’

Table 2

Main terminologies used in text mining for pharmacovigilance.

Name (abbreviation)	Source/availability	Description
Anatomical Therapeutic Chemical Classification System (ATC)	World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC) http://www.whocc.no/atc	Coding system for drugs based upon the body system they act on as well as their chemical properties and therapeutic effects. Drugs are assigned unique identifiers based on five 'levels' of classification: body system, therapeutic group, pharmacological group, chemical group, and specific chemical substance.
Current Procedural Terminology (CPT)	American Medical Association (AMA) http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page?	Medical terminology used to code medical procedures and services under public and private health insurance programs
International Statistical Classification of Diseases (ICD9, ICD10)	World Health Organization (WHO) http://www.who.int/classifications/icd/en/	A disease classification system designed to group similar diseases, such as 'diseases of the nervous system', 'neoplasms'. Used as a standard diagnostic tool for epidemiology, health management and clinical purposes
Logical Observation Identifiers Names and Codes (LOINC)	Regenstrief Institute http://loinc.org/	A standard for coding laboratory and clinical test results
Medical Dictionary for Regulatory Activities (MedDRA)	Maintenance and Support Services Organization (MSSO) http://www.meddra.org/	A hierarchically organized terminology intended for regulatory communication and classifying adverse event information associated with medical products. It includes terms for symptoms, diseases, indications, medical procedures, and family history.
Medical Subject Headings (MeSH)	National Library of Medicine (NLM) http://www.nlm.nih.gov/mesh/	A hierarchically organized controlled vocabulary of medical terms, as well as synonyms and alternative terms, used for indexing articles in MEDLINE.
National Drug File - Reference Terminology (NDF-RT)	U.S. Department of Veterans Affairs, Veterans Health Administration (VHA) http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/	A hierarchical drug classification that groups drugs by their properties including ingredients, chemical structure, dose form, physiologic effect, mechanism of action, pharmacokinetics, interactions and diseases.
RxNorm	National Library of Medicine https://www.nlm.nih.gov/research/umls/rxnorm/	A normalized vocabulary for generic and branded drugs that associates drugs with their ingredients, strength and forms. Used mainly to support semantic interoperation between drug terminologies and pharmacy knowledge bases.
Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT)	International Health Terminology Standards Development Organization http://www.ihtsdo.org/snomed-ct/	A hierarchically organized multilingual medical terminology of over 311,000 terms and synonyms including diagnoses, procedures, and anatomy, as well as pharmaceuticals and biologics. Used for clinical documentation and reporting, and the core terminology for EHRs.