International Journal of Information Management 48 (2019) 228–237

Contents lists available at ScienceDirect

International Journal of Information Management

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

Exploring adverse drug reactions of diabetes medicine using social media analytics and interactive visualizations

Si Li[a], Chia-Hui Yu[b,*], Yichuan Wang[c], Yedurag Babu[d]

[a] Xinhua College of Sun Yat-Sen University, China
[b] Graduate Institute of Technology, Innovation & Intellectual Property Management, National Chengchi University (NCCU), Taiwan, ROC
[c] Newcastle University Business School, Newcastle University, United Kingdom
[d] The Home Depot, USA

ARTICLE INFO

Keywords:
Social media analytics
Adverse drug reactions (ADRs)
Data visualization
Big data analytics
Online health community

ABSTRACT

The aim of this study is to propose an automatic and real-time social media analytics framework with interactive data visualizations to support effective exploration of knowledge about adverse drug reaction (ADR) surveillance. This proposed framework has been prototypically implemented on the basis of social media data. A longitudinal diabetes patient online community data (AskaPatient.com) as well as FDA Adverse Event Reporting Systems (FAERS) data as a benchmark were used to evaluate our proposed approach's performance. Based on the results, our approach significantly increases the precision and accuracy for ADR extraction. The number of ADR cases, the time when the ADRs occurred, and the rating of Glucophage have been visualized that resulted by mining a collection of 870 ADRs posted in Askapatents.com over a certain time period (from 2001 to 2015). The results have important implications for pharmaceutical companies and hospitals wishing to monitor ADRs of medicines.

1. Introduction

The idea of obtaining new insights from social media through data analytics techniques to improve healthcare quality has attracted considerable attention from both academics and practitioners in the information systems field (Almansoori et al., 2014; Chen, Chiang, & Storey, 2012; Karami, Dahl, Turner-McGrievy, Kharrazi, & Shaw, 2018). Evidence from a practical report described by Demi and Cooper Advertising and DC Interactive Group (2012) shows that 60% of physicians believe that the transparency and authenticity of social media actually helps improve the quality of care delivered to patients. In addition, 41% of online users agree that their choice of a specific doctor, hospital, or medical facility will be influenced by word of mouth on social media. This implies that social media is a vital base for improving healthcare quality (Chretien & Kind, 2013). To improve healthcare quality through social media analytics, developing effective methods of aggregating, analyzing, and visualizing data to transfer disparate data into knowledge and business intelligence is urgently needed.

In health care, an online health community (OHC) is one kind of social networking platform where patients can discuss their health concerns and share their experience with other patients or health professionals. Actively mining the content from OHCs could potentially reveal drug safety concerns before regulators discover them through more passive methods via official channels such as the Food and Drug Administration (FDA) (Liu & Chen, 2015). In fact, the massive amounts of patient-generated content such as patient subjective opinions about adverse drug reactions (ADRs), medicine recommendations and ratings, and self-reported health profile stored in online health communities could be a valuable source for identifying adverse events and risks associated with drugs due to its active and real-time nature (Akay, Dragomir, & Erlandsson, 2015; Wang & Hajli, 2017).

However, the traditional data analysis approaches that manually identify new perspectives from patient data are not feasible or efficient. Research using social media analytics to explore ADRs has increasingly grown in recent years (e.g., Fan & Gordon, 2014; Frost, Okun, Vaughan, Heywood, & Wicks, 2011; Nguyen et al., 2017). ADR surveillance could benefit from social media analytics in various ways, including helping track and predict the course of illness through a population, and adding value to the current practice of pharmacovigilance and pharmaceutical product development (Liu & Chen, 2015; Sarker & Gonzalez, 2015; Yang, Kiang, & Shang, 2015). Despite the benefits of using social media analytics for ADR surveillance, there remain two major research gaps, which are the driving force behind our study.

First, there have been intensive attempts to employ traditional

* Corresponding author.
E-mail addresses: si.li.2018@live.rhul.ac.uk (S. Li), 101359502@nccu.edu.tw (C.-H. Yu), ycwang825@gmail.com (Y. Wang), Yeduragbabu@gmail.com (Y. Babu).

https://doi.org/10.1016/j.ijinfomgt.2018.12.007
Received 1 May 2017; Received in revised form 28 October 2018; Accepted 18 December 2018
Available online 23 January 2019

statistical methods such as regression models or data mining approaches to discover the pattern of ADRs and to identify the association between a drug and an adverse drug reaction (e.g., Harpaz et al., 2012; Yang et al., 2015), and to develop data extraction and classification approaches to automatically obtain meaningful ADR information (Sarker & Gonzalez, 2015). However, ADR information stored in online health communities is rarely discovered, analyzed and visualized with an integrated social media analytics framework which includes key components: automatic information extraction, real-time text analytics and automatic topic modeling on forum and social media data.

Second, previous works have mainly focused on improving the accuracy and efficiency of data extraction and mining methods and algorithms, but very little work has been emphasized on data visualization (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018; Wang, Kung, & Byrd, 2018). Indeed, many of the developed data mining frameworks lack the capability to present the results in a visual, interactive, and real-time way (Galletta, Carnevale, Bramanti, & Fazio, 2018; Yang & Ng, 2011). Therefore, it is difficult for healthcare practitioners to interpret the results from social media to obtain new insights into ADR surveillance (Ward, Marsolo, & Froehle, 2014).

To this end, the main goal of this paper is to propose an automatic and real-time social media analytics framework with interactive visualizations to support effective exploration of knowledge about ADR surveillance. This proposed framework has been prototypically implemented on the basis of web data. A case study on a longitudinal diabetes patient social media platform (i.e., AskaPatient.com) as well as FDA Adverse Event Reporting Systems (FAERS) data as a benchmark evaluates our approach's performance. Specifically, this framework captures the contents provided from patient drug experiences to unveil the variations in adverse event surveillance of a particular diabetes drug among two different data sources with respect to time, gender of patients, age group of patients and associations.

The remainder of the paper is organized as follows: the next section reviews the prior studies in social media analytics and their applications of relevant data analysis techniques. In section 3, we propose a research framework of identifying ADRs through social media analytics and interactive data visualization. Section 4 evaluates the performance of the proposed framework using datasets collected from two online sources pertaining to ADRs. The paper is concluded with the implications regarding research and practice.

2. Literature review

In this section, we provide a brief description of ADRs and review the prior research using the state-of-the-art social media analytics approach for ADRs surveillance, as summarized in Table 1.

2.1. Adverse drug reactions

Adverse drug reactions (ADRs) are a significant cause of admission to hospital and mortality in many countries. ADRs are defined as "an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product" (Edwards & Aronson, 2000, p. 1255). ADRs have been monitored by the Food and Drug Administration (FDA) in the United States. In general, the surveillance of ADRs begins with the pre-marketing stage, in which each drug is investigated by clinical trials. Th surveillance continues in the post-marketing stage through self-reporting from pharmaceutical companies, hospitals, and consumers.

Currently, ADR surveillance relies on FAERS which is restricted by its passive nature and only covers a fraction of knowledge available. It is estimated that the reporting rate of ADRs in this system is lower than 10% (Hazell & Shakir, 2006; Ji et al., 2011). This leads to more than 2 million injuries, hospitalizations, and deaths per year in the US alone and an increase in costs estimated at $75 billion annually (Harpaz et al., 2012). Although there is a vast body of research on developing data analytics frameworks for health social media content, there is limited work on the specific area of ADR surveillance. Earlier works on ADR surveillance has applied text mining and data extraction techniques to understand drug interaction, effectiveness, and side effects by using the data sources spontaneous reporting systems (Bate & Evans, 2009), medical case reports (Gurulingappa et al., 2012), and Electronic Health Records (Lin, Chen, Brown, Li, & Yang, 2017). However, ADR information stored in social media is rarely discovered. To complement insufficiency of ADR knowledge, there is an urgent need to intelligently extract new insights to understand drugs' side effects from patient-centric online communities on social media (Liu & Chen, 2015). In the next section, we review the prior studies regarding the use of social media analytics on ADR surveillance.

2.2. Social media analytics for ADR surveillance

Social networking sites such as online communities provide platforms where members can actively discuss products they have bought and extensively generate subjective opinions, recommendations and ratings for those shopping experiences (Wang & Yu, 2017). To extract knowledge from these user-generated contents, social media analytics have emerged as a powerful analysis tool discovering hidden insights from millions of online sources (Aswani, Kar, Ilavarasan, & Dwivedi, 2018; Chau & Xu, 2012; Zeng, Chen, Lusch, & Li, 2010). Social media analytics is viewed as "informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application" (Zeng et al., 2010, p. 14). A set of analytical techniques has been proposed to automatically collect, extract, and analyze social media data. For example, Park, Huh, Oh, and Han (2012) propose a social network-driven inference framework to determine the accuracy and reliability of self-reported customer profiles through utilizing the individuals' social circles and communication patterns within their circles, whereas Chau and Xu (2012) explore the potential consumer pattern and behavior from blog contents related to the topics of interest and interactions between bloggers and communities through a social media analytics framework. The outcomes of these studies seem striking, which use their analytics frameworks to identify the interaction activities of major contributors and their impact on information dissemination, thereby increasing customer satisfaction.

2.2.1. Information extraction and classification techniques

The most important component of social media analytics is to extract spans of text including the drugs, adverse reactions, diseases, and symptoms of diseases and classify them based on their type. The three most commonly used information extraction techniques are text classification, lexicon-based entity recognition and ADR relation extraction. Text classification methods such as support vector machines (SVMs) and Naïve Bayes have been widely used in recent pharmacovigilance studies (Bian, Topaloglu, & Yu, 2012; Chee, Berlin, & Schatz, 2011). Chee et al. (2011) developed ensemble classifiers with SVM and Naïve Bayes to classify drugs into FDA's watch list and non-watch list based on messages extracted from online health forums, whereas Bian et al. (2012) used SVM to filter out noise in tweets. Huh, Yetisgen-Yildiz, and Pratt (2013) have applied text classification methods to determine whether a thread in an online health forum needs moderators' help. By using the tags and tag clouds, O'Grady et al. (2012) have assessed the credibility of messages posted in online health forums.

Lexicon-based entity recognition in social media ADR surveillance research aims to extract ADR mentions from patient discussions of both treatments and medical events. Over 50% of the previous studies have adopted lexicon-based entity recognition approaches (see a systemic review of pharmacovigilance study by Sarker et al., 2015) because of the wide availability of medical lexicons and knowledge bases in the

Table 1
Summary of related ADR studies with social media data.

Prior study	Data source	Research Focus	Data extraction and classification	Results
Bian et al. (2012)	Two billion Tweets in Twitter	ADRs	• Textual features: Bag-of-words model • Semantic features: use Metamap to discover Unified Medical Language System Metathesaurus concepts • Text classification: SVM	For ADR classification, • Mean area under the curve value: 0.74 • Prediction accuracy: 0.74
Chee et al. (2011)	Health forums in Yahoo! groups	Risky drugs	• Lexicons: UMLS, MedEffect, and SIDER Text classification: SVM and naïve Bayes	The ensemble classifier is able to identify risky drugs for FDA scrutiny
Cocos, Fiks, and Masino (2017)	Combined Twitter datasets	ADRs	• Use ark-twokenize-py Python module Recurrent neural network (RNN) model to classify the text	F-measure of 0.755 for ADR identification
Liu and Chen (2015)	The American Diabetes Association (ADA) online community	ADRs	• Medical named entity extraction • Adverse drug event (ADE) extraction • Bag-of-words (BOW) features and transductive SVMs	Drug entity extraction • 93.9% in precision • 91.7% in recall • 92.5% in F-measure Medical event entity extraction • 87.3% in precision • 80.3% in recall • 83.5% in F-measure
Nguyen et al. (2017)	LiveJournal, Reddit, and Twitter	ADRs	• The lexicon of known ADRs from SIDER • Word embedding techniques (Word2vec)	Correlation coefficients, between 0.29 and 0.59 demonstrates capability of proposed approaches to aid in the discovery of meaningful patterns from social media data
Nikfarjam and Gonzalez (2011)	Daily strength	ADRs	• Lexical pattern-matching (2400 comments for pattern building) • Association rule mining; Apriori algorithm	• Precision: 70.01% • Recall: 66.32% • F-measure: 67.96%
Sarker and Gonzalez (2015)	Three datasets: Twitter (TW), daily strength (DS), and ADE corpus	ADRs	• Lexicons: UMLS, WordNet, MedEffect, SIDER, and COSTART • Text classification: SVM	Achieved detection of sentences with ADR mentions with F-scores of 0.812, 0.538 and 0.678 for the ADE, TW and DS data sets 0.812
Segura-Bedmar, Martínez, Revert, and Moreno-Schneider (2015)	84,000 messages extracted from a Spanish health forum	ADRs	• Distant-supervision method • Kernel method	• Precision: 48%; • Recall: 59%

healthcare domain. Prior studies often use the Unified Medical Language System (UMLS), spontaneous reporting systems (SRSs), the FDA's Adverse Event Reporting System (FAERS) and MedEffect (the adverse drug event reporting system in Canada) as a lexicon source since consumers' health vocabulary often differs from that of medical professionals. Meanwhile, the Consumer Health Vocabulary, a lexicon linking UMLS standard medical terms to patients' colloquial language, has been adopted in many studies to interpret medical terms in online patient discussions (Benton et al., 2011). However, pure lexicon-based approaches do not address some important challenges. Consumers do not always use technical terms found in the existing lexicons. Instead, they use creative phrases, descriptive symptom explanations, and idiomatic expressions.

In addition to the lexicon-based entity recognition, some studies have focused on identifying the relations between ADRs and drugs. Following the study by Nikfarjam and Gonzalez (2011), a popular approach for the discovery of drug-ADR pairs, in lexicon-based and other techniques, has been the use of association rule mining by which associations between entities are discovered. In general, following the identification of ADRs and drugs, association rule mining is used to identify whether a drug and ADR pair is associated or not. Frequent occurrence of drug-ADR pair mentions in close proximity within user posts are considered to be indications of ADRs associated with the drugs, and these associations are detected by association rule mining in unannotated data.

2.2.2. Data visualization

Data visualization is one of the components in big data analytics architecture. This component is capable of automatically generating real-time reports and presenting those using visual dashboards/systems with interactive features (Stieglitz et al., 2018; Wang & Hajli, 2017). There has been growing attention paid to understanding the visualization of clinical results as clinical decisions usually needs to be made quickly and accurately. Without the visualization of clinical information, physicians must iterate through the time-consuming steps (e.g. searching for a particular drug reaction or reviewing the most common ADR of a drug) in order to organize the collective findings into an informed clinical decision (Duke, Li, & Grannis, 2010). Although some visual analytics systems have been proposed to visualize behavioral data in social media (Chae et al., 2014; Weiler, Grossniklaus, & Scholl, 2016), very little work has been focused on visualizing the clinical data meaningfully as a way of communication results.

In fact, healthcare and patient data can be visualized in three ways: (1) yielding general clinical summaries for patients such as patient name, medical history, date of visit, medication list, reason(s) for visit, and treatment procedures. These reporting, when it is presented through visual metrics, provide concise and timely insights that help healthcare practitioners make evidence-based, diagnostic and treatment decisions (Fihn et al., 2014; Ghosh & Scott, 2011); (2) extrapolating trends and patterns for specific treatments, drugs or patient behavior through historical reporting, statistical analyses, and time series comparisons and presenting in interactive dashboards and charts (Roski, Bo-Linn, & Andrews, 2014); and (3) detecting and reporting real-time events such as alerts and proactive notifications, and operational key performance indicators and sending to interested users for further actions (Wang, Kung, Byrd, 2018).

3. Research method

The objective of this study is to design, implement, and apply a framework to explore ADRs from patient data in online health communities. Fig. 1 illustrates our proposed framework for exploring ADRs through social media analytics and displaying the results through interactive visualization tools. From a theoretical standpoint, this framework is based on the concept of information lifecycle management (ILM) which defines as the tools used to align the business value of

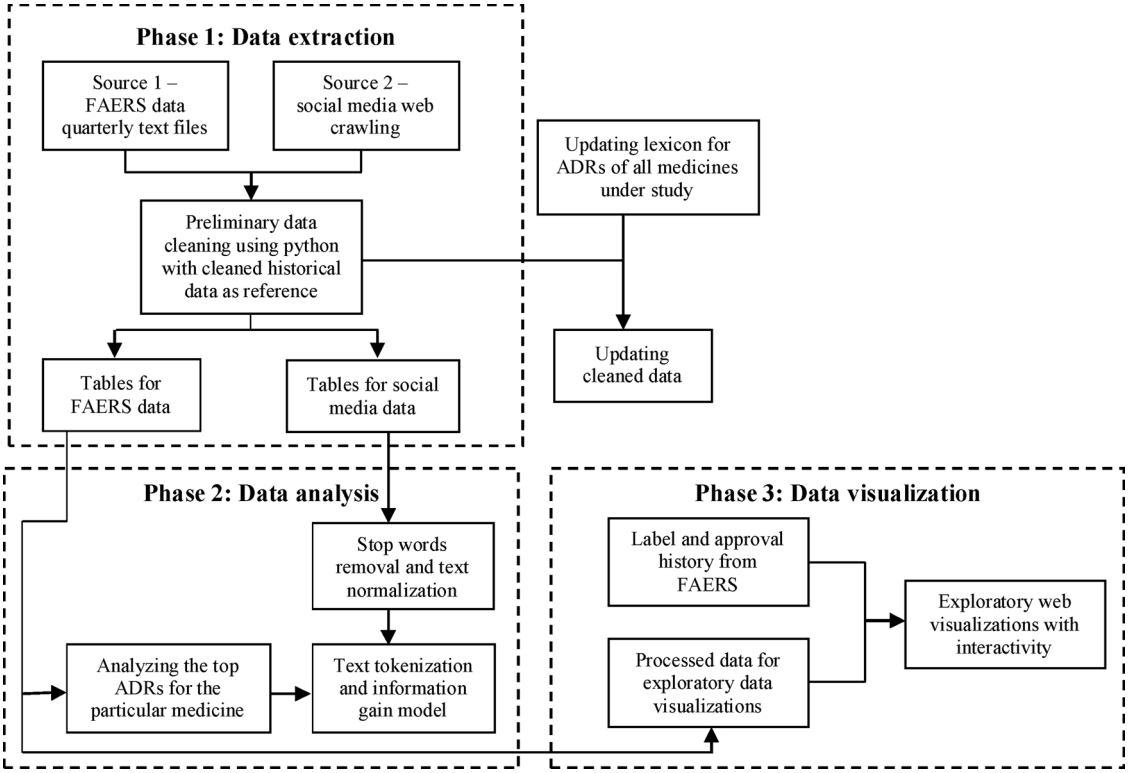

Fig. 1. Research framework of identifying ADRs through using data analytics and interactive data visualization.

information with the most appropriate and cost-effective infrastructure from the time when information is created through its final disposition (Storage Networking Industry Association, 2009; Wang, Kung, Wang, & Cegielski, 2018). Through this data lifecycle, the framework incorporates various analytical techniques into the information cycle stages (i.e., data capture, analysis, and visualization) in our design. In the following, we explain the key components of the framework in detail.

3.1. Data extraction

To understand the associations and patterns for ADRs, we collected patient-provided comments on a specific diabetes drug (i.e., Glucophage) from the FAERS and online health community (i.e., AskaPatient.com). Glucophage is the trade name of Metformin and is considered to be one of the first-choice drugs for type 2 diabetes (Okayasu et al., 2012). The data sources are summarized in Table 2.

The FAERS datasets are available for the public form. This data source has publically available raw data consisting of reports of adverse events. Each data point is an individual report of an adverse drug event. These datasets can be downloaded as ".txt" files. Preliminary data cleaning was performed using python programming language with cleaned historical data as reference. The updated new data is inserted into an Excel table. The cleaned historical data table is maintained in the main PostgreSQL database. This table has data manipulations which are applied on each quarterly text files so that the same manipulations can be reused in the future.

AskaPatient (http://www.askapatient.com/) is one of the largest online patient-based health communities and is designed for patients to record experiences and side effects of all sorts of medications since 2000. AskaPatient.com is uniquely ideal for the purpose of our study because of the availability of threads (drug reviews and comments) and ratings. These features enable an online community to generate distributed knowledge that occurs when active users are engaged in building, distributing, storing, using, and interacting with knowledge (Kazmer et al., 2014). For example, participants in the AskaPatient share their opinions and experiences about specific diseases with respect to diagnosis, symptom management, drug usage and treatment as well as a five-point scale rating. Such a sharing process allows the construction of meaningful medical information from a patient standpoint, so we can extract from them through an analytics framework to formulate healthcare patient safety improvement strategies.

Moreover, AskaPatient provides users with a unique forum in which to share and compare medication experiences. A snippet of the posts on Askapatient.com for the medicine Glucophage® is shown in Fig. 2. We captured the data from AskaPatient using a web crawler after permission from their administrators. A web crawler was used to capture data from webpages and extract relevant information in patient comments. Collected information includes rating, reason, side effects, comments from patient, gender, age, duration/dosage, and posting date. Among these variables, ratings were used to evaluate whether this medicine helps patient to improve their disease, ranging from 1 (low; I would not

Table 2
Data source in our study.

Type of source	Total number of comments/documents captured	Patient data being captured
AskaPatient.com	n = 1,562	• Ratings on drugs • Free text descriptions of drug regarding side effect, dose usage, and symptoms
FDA Adverse Event Reporting Systems	n = 47,213	• Reports about a given drug's ADRs • Patient demographic and administrative information (age, gender, weight)

RATING	REASON	SIDE EFFECTS FOR GLUCOPHAGE	COMMENTS	SEX	AGE	DURATION/ DOSAGE	DATE ADDED
▼▲				F M	▼▲	▼▲	▼▲
1	PCOS	Headaches, muscle pains, suicidal thoughts, behavioral changes similar to being high or drunk, loss of love for loved ones, loss of feeling in hands, dizziness, lightheaded.	This drug ruined my life. Please, do NOT take Metformin.	F	15	2 months 1000MG 3X D	4/26/2015
1	Type 2 diabetes , pcos	I had chest pain, diahreha, vomiting and nausea. I also had severe head aches and the medicine didn't help. I've lost my appetite for everything and everything hurts me. This medication has ruined my life		F	15	3 months 2000mg	3/1/2015
4	Diabetes Mellitus			M	50	8 years 1000 2X D	2/21/2014 Email

Fig. 2. Posts by patients on Askapatient.com for the medicine Glucophage®.

recommend taking this medicine) to 5 (high; this medicine cured me or helped me a great deal).

These datasets from two different sources are stored in a PostgreSQL relational database for the next stage. The datasets are stored in different tables. This information allows us to explore the associations and patterns among side effect symptoms, gender, age, and rating.

3.2. Data analysis

Data analysis tools were used to process all kinds of data (e.g., rating and comments) and perform appropriate analyses to harvest insights. This is particularly important for transforming social media data into meaningful information that supports evidence-based decision making and useful practices for healthcare organizations. In this study, the Askapatient.com dataset was cleaned up and modeled to extract ADRs at this stage. The python programming language is capable of providing support for advanced Natural Language Processing (NLP) and Modeling with enhanced capabilities for data processing and manipulations. Particularly, in this study, the python pattern module developed by the CLiPS was used in this phase.

3.2.1. Step 1: stop words removal and text normalization

Patients' ADR comments on social media tend to be colloquial and sometimes the words do not provide much information. These words are generally called as stop words that have to be removed by using a stop word removal technique. Since the ADR data provided by the internet users on Askapatient.com is text data, text normalization has an important role to play in the data analysis. During normalization, the entire text is transformed to lower case and the Porter's stemming algorithm is applied.

3.2.2. Step 2: text tokenization and information gain model

The text is tokenized into 3 g. This is to endure that text with good information gain are not excluded from the model. After this tokenization, the information gain model is applied. Based on the model outcome, the important parameters which have more information gain than a threshold level are selected. Thus tokens up to 3 g are considered and the top few tokens are selected. Table 3 shows the top-ranked ADRs and associated keywords we found for Glucophage on AskaPatient.com. We not only consider single words, but also bigrams and trigrams. A snippet from the lexicon is given in Table 4 below.

3.2.3. Step 3: information extraction for the top ADRs

Based on the information gain that each *n* gram provides in accurately predicting the ratings given by Askapatient.com (This rating being an integer number between 1 [very dissatisfied with the medicine] and 5 [very satisfied]), the presence of these top ADRs in particular reviews are captured by forming binary flags for each of these side effects. If the particular side effect is present in the review, then the corresponding binary flag in the data is updated as 1, else 0.

3.3. Data visualization

The iterative visualizations page is made possible by dc.js, which is a Cool Javascript library with native cross-filter support for exploration of multidimensional datasets. Other Javascript libraries used in this study are d3.js, a highly popular Javascript library which has revolutionized data visualization, and crossfilter.js, which enabled fast multidimensional filtering for coordinated views. Interactive visualizations are made from the different data sets combined. These exploratory data visualizations are performed using the dc.js JavaScript coding library. The data allows us to dig deep into the lowest level of data, making the reasons for any fluctuations clearer.

4. Results and evaluation

4.1. The results of ADR extraction

Based on our text mining analysis, the main ADR, as reported by patients who are taking Glucophage in Askapatient.com, are shown in Table 5. The total number of patients who reported about the ADRs associated with Glucophage was 870. We compare the results gained from the Askapatient.com dataset with the FDA adverse events summary for Glucophage. This summary is based on 145,556 reports filed with the FDA between 2004 and 2012, as shown in Table 6. Our discovery confirms the findings of FDA reports, showing all these ADRs come from the top 10 list of FDA reports. In our analysis, we found diarrhea to be the top ADR when taking Glucophage. However, the top ADRs in FDA reports, such as blood glucose increase and weight loss were not reported by many patients in AskaPatient.com.

4.2. The results of data visualization

Data visualization is one of the critical big data analytics features that aims to extrapolate meaning from data and perform visualization of the information (Wang & Byrd, 2017). In this study, the data visualization component generates the outputs such as various visualization side effects reports and real-time interactive information derived

Table 3
Top ranked ADRs and associated keywords for Glucophage.

ADRs	Keywords						
Diarrhoea	loos	stool	diarrhea	restroom	diarrea		
Fatigue	lethargi	energi	tire	exhaust	weak	fatigu	
Digestive Problems	stomach	bowel	bloat	abdomin	consti	gas	flatul
Pain	pain						
Nausea/Dizziness/Vomiting	nausea	dizzi	fog	lightheaded	vomit		
Head Ache	head	headach	ach				
Decreased appetite	appetit	food	appietit	hungri but			

Table 4
A snippet from the lexicon.

Words/bigrams/trigrams	Category	Weight
loos	Diarrhoea	0.0203
lethargi	Fatigue	0.0114
(u'loos', u'bowel')	Diarrhoea	0.0095
(u'loos', u'stool')	Diarrhoea	0.0093
vomit	Nausea/Dizziness/Vomiting	0.0088
(u'stomach', u'pain', u'and')	Stomach Problems	0.0076
kidney	Kidney Problems	0.0076
(u'hair', u'loss')	Hair Loss	0.0073
(u'diarrhea', u'in')	Diarrhoea	0.0072
fart	Flatulance	0.0072
restroom	Diarrhoea	0.0072
(u'bowel', u'movement')	Stomach Problems	0.0071
muscl	Muscle Pain	0.0070
pain	Pain	0.0060
(u'abdomin', u'bloat')	Stomach Problems	0.0057
(u'constant', u'diarrhea')	Diarrhoea	0.0057
(u'constant', u'nausea')	Nausea/Dizziness/Vomiting	0.0057
(u'energi', u'loss')	Fatigue	0.0057

Table 5
ADRs for Glucophage reported by AskaPatient.com.

ADRs	# of ADRs reported on AskaPatient.com	% out of total posts
Diarrhea	213	37%
Digestive Problems	210	37%
Nausea/Dizziness/ Vomiting	149	26%
Headache	88	15%
Fatigue	75	13%
Decreased appetite	70	12%
Pain	65	11%

Table 6
Top 10 ADRs for Glucophage reported by FDA between 2004 and 2012.

ADRs	# of ADRs reported by FDA
Blood glucose increase	3762
Nausea	3339
Weight decrease	2472
Diarrhoea	1998
Vomiting	1778
Decreased appetite	1624
Myocardial infarction	1598
Dizziness	1394
Dyspnoea	1351
Renal failure acute	1309

from the data analysis components. Three key results are included: 1) Posts from patents on AskaPatient.com and general summaries reporting, such as gender distribution and age distribution (see Fig. 3); 2) Primary reasons for taking Glucophage including diabetes (do not mention type 1 or type 2), hyperinsulinemia, insulin resistance, polycystic ovary syndrome (PCOS), pre-diabetic, trying to conceive, type 2 diabetes, weight loss, and others (see Fig. 3); 3) Time series comparisons in terms of average rating (Fig. 4 on the top left), total number of cases (Fig. 4 on the top right), and variation of number of case for each ADR (see Fig. 4 on the bottom left). From the visualizations, there is a fluctuation in the number of ADR cases reported from 2006 to 2008 and a dramatic drop in the number of ADR cases reported in 2013. These fluctuations may be explained by the variability in medicine quality and variations in medicine formulations of Glucophage. For example, after the manufacturing change or additions in 2013, the number of ADRs reported by patients in Askapatient.com has substantially declined.

Time series comparisons can not only be visually examined for changes in trends and patterns, but also displayed by specific groups. For example, Fig. 5 displays the number of case, rating, variation of ADR for females who are aged between 30 and 40, while Fig. 6 shows ADR for males who are aged between 30 and 40. To compare two groups, we found the most common ADR for females who are aged between 30 and 40 is digestive problem. Males who are aged between 30 and 40 regularly suffer from diarrhea when taking Glucophage.

4.3. Evaluations

We used text analysis evaluation metrics including accuracy, precision, recall and F1 score to evaluate the performance of our model. These metrics have been widely used in information extraction and health social media studies. The accuracy parameter is self-explanatory, while precision is a measure of hits versus errors. If the classifier has a low precision, negative cases are being misclassified as positive. Recall is a measure of hits versus misses. If the classifier has a low recall, not all positive cases are being caught. The F1 score is simply the harmonic mean of precision and recall. K-fold cross-validation performs K tests on a given classifiers, each time partitioning the given dataset into different subsets for training and testing, and returns the average.

To evaluate the performance of ADR classifiers, we randomly selected 200 comments from the dataset and established definitions and content coding for labelling ADRs. Then we used the k-nearest neighbour classifier, single-layer averaged perceptron, and support vector machines, as well as the Naïve Bayes classifier to evaluate our performance. From the results of Table 7 we found the Naïve Bayes model with multi-nominal weighting can achieve higher accuracy and precision than other classifiers. The Naïve Bayes classifier is based on the probability that a feature occurs in a class, independent of other features, using Bayes' theorem. With the multinomial method, feature weights are used (0.0–1.0). With the binomial method, a feature is part of a document (1) or not (0). The model also showed consistent results under many cross validations. Since we have a relatively low number of features (436) compared to the other text classification models, the over-fitting problem might not be present. Regarding ADR extraction, our approach achieved approximately 95% accuracy, 87% precision, 73% recall, and 79% in F1 score. Based on the evaluation results, our approach significantly increases the precision and accuracy for ADR extraction.

Date Posted	Gender	Age	Rating	Side Effects
Side Effects for Glucophage				
4/26/2015	F	15	1	Headaches, muscle pains, suicidal thoughts, behavioral changes similar to being high or drunk, loss of love for loved ones, loss of feeling in hands, dizziness, lightheaded.
3/1/2015	F	15	1	I had chest pain, diahreha, vomiting and nausea. I also had severe head aches and the medicine didn't help. I've lost my appetite for everything and everything hurts me. This medication has ruined my life
2/21/2014	M	50	4	
12/21/2013	F	35	5	I got pregnant!
12/2/2013	F	24	5	I lost 60 lbs on it not even trying. Definitely curbs sweets cravings. Horrible explosive diarrhea with cramping tho. But overall makes my periods 100% regular and on time.. and I conceived.
4/7/2013	F	38	3	Initially all was good. I have however developed a rash on both legs. I have also lost 20kg's since been diagnosed with type 2 diabetes 3 months ago and started taking Glucophage !
3/5/2013	F	28	5	
1/2/2013	F	38	5	First week it was rough. I adjusted my eating to be more low carb and lost 40 pounds in 3 months. My blood sugar is perfect now. Love it.
12/31/2012	F	27	4	First I got extremely nauseous, my stomach was cramping, I went to the bathroom a lot, and I was extremely tired. Now those are pretty much over.
6/29/2012	M	55	1	It turned me into a ticking poop bomb. I could have sudden and explosive diarea at any time, as well as constant upset stomach. I stopped taking it and became regular again, and I felt a lot better.

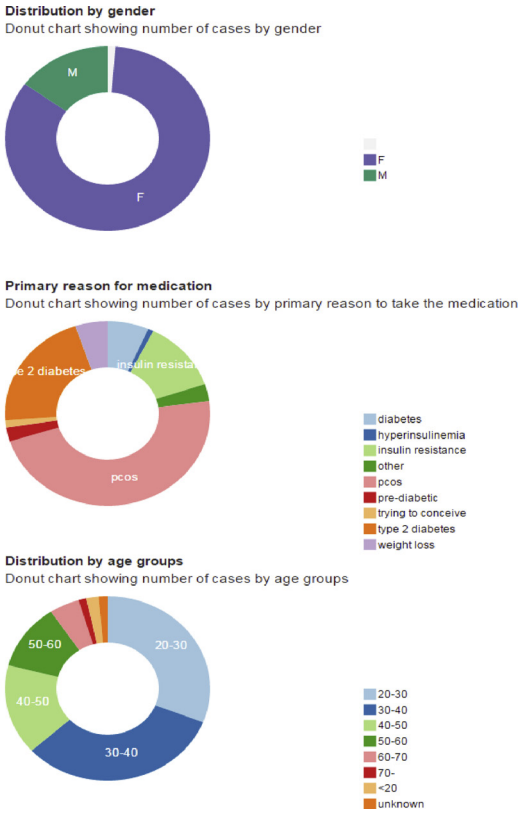

Fig. 3. Interactive visualization of ADR results.

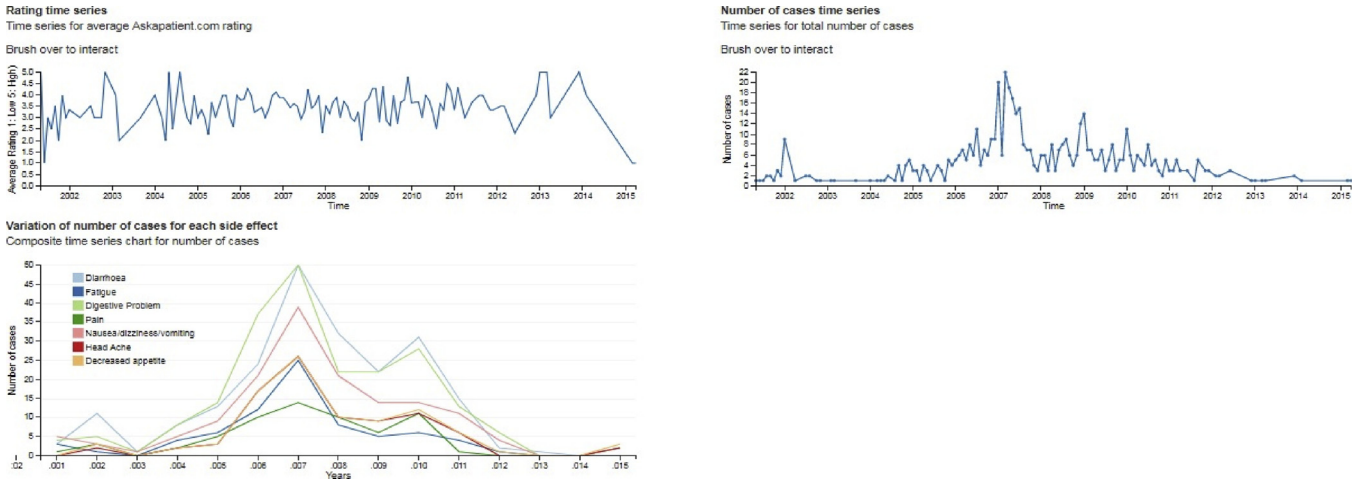

Fig. 4. Time series comparisons.

5. Discussions and implications

In line with the ADR publications, our research supports that social media data could be a valuable resource to explore the side effects for specific drugs. We compare our findings to previous studies on ADR of Glucophage/ Metformin in medicine literature. From their analysis of ADRs, diarrhea is the most common symptom for Glucophage/ Metformin, and can be accompanied by vomiting and nausea. Our discovery of ADRs from social media data confirms the findings of several studies listed in Table 8. It is, however, important to note that, this study has visualized the number of ADR cases, the time when the ADRs occurred, and the rating of Glucophage that resulted by mining a collection of 870 ADRs posted in Askapatents.com over a certain time period (from 2001 to 2015). Displaying this information in the form of visual dashboards has enabled users to understand the patterns and associations of ADRs for Glucophage. Thus, these reports can be utilized to provide a comprehensive view that supports the implementation of evidence-based medicine, provides advanced warnings for ADR surveillance, and guides diagnostic and treatment decisions.

Our research has two implications regarding research and practice. First, our proposed framework can be used for extracting knowledge such as demographic patient data and comments and for performing social media analysis on the data collected. The text mining and data visualization analysis methods applied on the side effect detection can be extended to other medicines. As the framework is generic, it can be easily applied in other applications. New techniques can be readily plugged into the framework for identifying novel patterns and useful knowledge. In practice, business and marketing managers in the

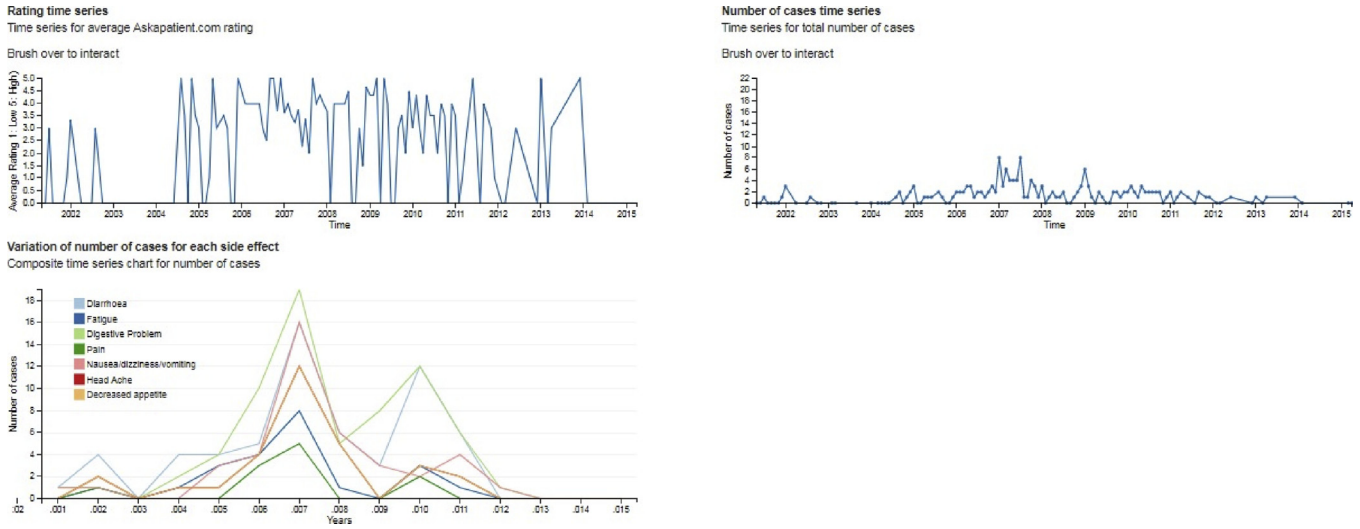

Fig. 5. Time series comparisons (only display females who are aged between 30 and 40).

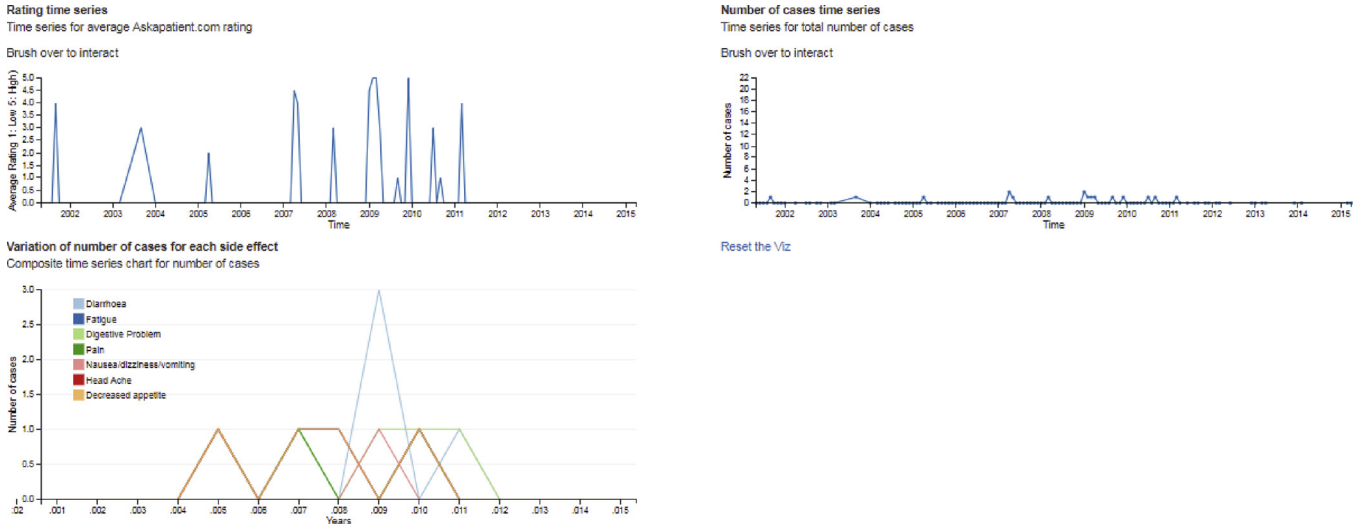

Fig. 6. Time series comparisons (only display males who are aged between 30 and 40).

Table 7
Evaluation results of our research framework.

Run No	k folds	Parameters	Accuracy %	Precision %	Recall %	F1 Score	Std Dev
1	3	Multinomial	94.68%	85.44%	70.98%	77.52%	0.027
2	5	Multinomial	94.83%	85.77%	73.29%	78.92%	0.033
4	10	Multinomial	95.02%	87.18%	72.69%	79.25%	0.019
5	15	Multinomial	95.00%	86.68%	73.17%	79.23%	0.042
6	20	Multinomial	94.98%	86.16%	73.33%	79.16%	0.059
7	3	Binomial	85.68%	65.90%	86.41%	74.77%	0.015
8	5	Binomial	85.57%	65.76%	86.79%	74.80%	0.020
9	10	Binomial	85.53%	65.92%	87.36%	75.11%	0.019
10	20	Binomial	85.51%	66.04%	87.41%	75.21%	0.034

pharmaceutical industry can apply the framework for social media analysis on a wide range of organizations, products, and topics.

Second, we believe our study is timely and important for research and practice in the area of social media analytics. While many previous studies have recognized the potential of web intelligence mining in the social media environment, very few have provided a viable methodology, coupled with case studies, describing how it should be conducted. Particularly, this study has presented the exploration of ADRs in a visual way. As social media data can be displayed in clear and concise figures or dashboards, it can support the decision-making process. Therefore, our study has provided a good start point for future work on the visualization of social media data.

6. Conclusion

A growing body of research has suggested social media analytics is beneficial for understanding the human behaviors, characteristics and patterns of user-generated content related to the company of interest

Table 8
Comparing the results to medicine literature on ADR of Glucophage/ Metformin.

Study	Subjects and Data sources	Top 3 ADRs discovered	
den Hertog et al. (2015)	• 40 patients were enrolled between August 2007 and August 2008. 19 were randomly assigned to metformin and 21 to the control group.	Nausea Diarrhea Anorexia	21% 16% 11%
Forrester (2008)	• N = 1528 (Patients were aged 20 years or greater); 264 ADRs reported • Cases were all ingestions of metformin reported to the Texas Poison Center Network (TPCN) during 2000–2006	Vomiting Nausea Hyperglycemia	8.0% 6.1% 4.2%
Ji et al. (2015)	• Patients (n = 689) were randomized to linagliptin + LD metformin (n = 344) or HD metformin (n = 345) for 14 weeks during 2011-2013	Diarrhea Abdominal pain Nausea	12.1% (LD); 15.8% (HD) 3.4% (LD); 6.2% (HD) 4.4% (LD); 4.7% (HD)
Legro et al. (2007)	• Infertile women with the polycystic ovary syndrome • N = 208; Clinical trials (Subjects were enrolled in the study from November 2002 to December 2004)	Diarrhea Nausea Abdominal pain or discomfort	64.9% 61.5% 59.1%
Okayasu et al. (2012)	• 101 participants admitted to the case hospital and started on metformin during September 1, 2009 and August 31, 2010	Diarrhea Anorexia –	26.7% 3.0% –
Current study	• Online users registered in AskaPatient.com • 870 ADRs posted in Askapatents.com between 2001 and 2015	Diarrhea Digestive Problems Nausea/Dizziness/Vomiting	37% 37% 26%

(Chau & Xu, 2012; Park et al., 2012). In this study, we propose a social media analytics framework and use it to automatically collect comments, analyze the patient reviews on a specific diabetes drug, and develop an interactive data visualization page to present the results. Using this framework, the associations and patterns between ADRs and patient demographics (i.e., gender and age) can be discovered. Our case study on Glucophage demonstrates the usefulness of the framework and reveals patterns which will help answer important questions in the domain of knowledge discovery in online health communities. This framework can be also applied to any medicine to discover its side effects, association and patterns, and even prediction.

Acknowledgement

We would like thank AskaPatient.com for providing us with their forum posts for our analysis.

References

Akay, A., Dragomir, A., & Erlandsson, B. E. (2015). Network-based modeling and intelligent data mining of social media for improving care. *IEEE Journal of Biomedical and Health Informatics, 19*(1), 210–218.

Almansoori, W., Addam, O., Zarour, O., Elzohbi, M., Sarhan, A., Kaya, M., ... Alhajj, R. (2014). The power of social network construction and analysis for knowledge discovery in the medical referral process. *Journal of Organizational Computing and Electronic Commerce, 24*(2-3), 186–214.

Aswani, R., Kar, A. K., Ilavarasan, P. V., & Dwivedi, Y. K. (2018). Search engine marketing is not all gold: insights from Twitter and SEOClerks. *International Journal of Information Management, 38*(1), 107–116.

Bate, A., & Evans, S. J. W. (2009). Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology Drug Safety, 18*, 427–436.

Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., ... Holmes, J. H. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics, 44*(6), 989–996.

Bian, J., Topaloglu, U., & Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. *Proceedings of the 2012 International workshop on smart health and wellbeing*, 25–32.

Chae, J., Thom, D., Jang, Y., Kim, S., Ertl, T., & Ebert, D. S. (2014). Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics, 38*, 51–60.

Chau, M., & Xu, J. (2012). Business intelligence in blogs: Understanding consumer interactions and communities. *MIS Quarterly, 36*(4), 1189–1216.

Chee, B. W., Berlin, R., & Schatz, B. (2011). Predicting adverse drug events from personal health messages. *In AMIA Annual Symposium Proceedings, Vol. 2011*, 217.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Chretien, K. C., & Kind, T. (2013). Social media and clinical care ethical, professional, and social implications. *Circulation, 127*, 1413–1421.

Cocos, A., Fiks, A. G., & Masino, A. J. (2017). Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association, 24*(4), 813–821.

Demi & Cooper Advertising and DC Interactive Group (2012). *Infographic: Rising use of social and mobile in healthcare*Available at:http://thesparkreport.com/infographic-social-mobile-healthcare/.

den Hertog, H. M., Vermeer, S. E., Zandbergen, A. A. M., Achterberg, S., Dippel, D. W., Algra, A., ... Koudstaal, P. J. (2015). Safety and feasibility of Metformin in patients with impaired glucose tolerance and a recent TIA or minor ischemic stroke (LIMIT) trial—A multicenter, randomized, open-label phase II trial. *International Journal of Stroke, 10*(1), 105–109.

Duke, J. D., Li, X., & Grannis, S. J. (2010). Data visualization speeds review of potential adverse drug events in patients on multiple medications. *Journal of Biomedical Informatics, 43*(2), 326–331.

Edwards, I. R., & Aronson, J. K. (2000). Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet, 356*(9237), 1255–1259.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74–81.

Fihn, S. D., Francis, J., Clancy, C., Nielson, C., Nelson, K., Rumsfeld, J., ... Graham, G. L. (2014). Insights from advanced analytics at the veterans health administration. *Health Affairs, 33*(7), 1203–1211.

Forrester, M. B. (2008). Adult metformin ingestions reported to Texas poison control centers, 2000–2006. *Human & Experimental Toxicology, 27*(7), 575–583.

Frost, J., Okun, S., Vaughan, T., Heywood, J., & Wicks, P. (2011). Patient-reported outcomes as a source of evidence in off-label prescribing: Analysis of data from PatientsLikeMe. *Journal of Medical Internet Research, 13*(1), e6.

Galletta, A., Carnevale, L., Bramanti, A., & Fazio, M. (2018). An innovative methodology for big data visualization for telemedicine. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/TII.2018.2842234.

Ghosh, B., & Scott, J. E. (2011). Antecedents and catalysts for developing a healthcare analytic capability. *Communications of the Association for Information Systems, 29* Article 22.

Gurulingappa, H., Mateen-Rajput, A., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics, 45*(5), 885–892.

Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel data mining methodologies for adverse drug event discovery and analysis. *Clinic Pharmacology & Therapeutics, 91*(6), 1010–1021.

Hazell, L., & Shakir, S. A. (2006). Under-reporting of adverse drug reactions. *Drug Safety, 29*(5), 385–396.

Huh, J., Yetisgen-Yildiz, M., & Pratt, W. (2013). Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics, 46*(6), 998–1005.

Ji, L., Zinman, B., Patel, S., Ji, J., Bailes, Z., Thiemann, S., ... Seck, T. (2015). Efficacy and safety of linagliptin co-administered with low-dose metformin once daily versus high-dose metformin twice daily in treatment-naïve patients with type 2 diabetes: A double-blind randomized trial. *Advances in Therapy, 32*(3), 201–215.

Ji, Y., Ying, H., Dews, P., Mansour, A., Tran, J., Miller, R. E., ... Massanari, R. M. (2011). A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. *IEEE Transactions on Information Technology in Biomedicine, 15*(3), 428–437.

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., & Shaw, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management, 38*(1), 1–6.

Kazmer, M. M., Lustria, M. L. A., Cortese, J., Burnett, G., Kim, J. H., Ma, J., ... Frost, J. (2014). Distributed knowledge in an online patient support community: Authority and discovery. *Journal of the Association for Information Science and Technology, 65*(7),

1319–1334.
Legro, R. S., Barnhart, H. X., Schlaff, W. D., Carr, B. R., Diamond, M. P., Carson, S. A., ... Myers, E. R. (2007). Clomiphene, metformin, or both for infertility in the polycystic ovary syndrome. *New England Journal of Medicine, 356*(6), 551–566.
Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., & Yang, H. J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A bayesian multitask learning approach. *MIS Quarterly, 41*(2), 473–495.
Liu, X., & Chen, H. (2015). Identifying adverse drug events from patient social media: A case study for diabetes. *IEEE Intelligent Systems, 30*(3), 44–51.
Nguyen, T., Larsen, M. E., O'Dea, B., Phung, D., Venkatesh, S., & Christensen, H. (2017). Estimation of the prevalence of adverse drug reactions from social media. *International Journal of Medical Informatics, 102*, 130–137.
Nikfarjam, A., & Gonzalez, G. H. (2011). *Pattern mining for extraction of mentions of adverse drug reactions from user comments. AMIA annual symposium proceedings*1019–1026.
O'Grady, L., Wathen, C. N., Charnaw-Burger, J., Betel, L., Shachak, A., Luke, R., ... Jadad, A. R. (2012). The use of tags and tag clouds to discern credible content in online health message forums. *International Journal of Medical Informatics, 81*(1), 36–44.
Okayasu, S., Kitaichi, K., Hori, A., Suwa, T., Horikawa, Y., Yamamoto, M., ... Itoh, Y. (2012). The evaluation of risk factors associated with adverse drug reactions by metformin in type 2 diabetes mellitus. *Biological and Pharmaceutical Bulletin, 35*(6), 933–937.
Park, S. H., Huh, S. Y., Oh, W., & Han, S. P. (2012). A social network-based inference model for validating customer profile data. *MIS Quarterly, 36*(4), 1217–1237.
Roski, J., Bo-Linn, G. W., & Andrews, T. A. (2014). Creating value in health care through big data: Opportunities and policy implications. *Health Affairs, 33*(7), 1115–1122.
Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53*, 196–207.
Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., ... Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics, 54*, 202–212.
Segura-Bedmar, I., Martínez, P., Revert, R., & Moreno-Schneider, J. (2015). Exploring Spanish health social media for detecting drug effects. *BMC Medical Informatics and Decision Making, 15*(2), S6.
Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics—Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156–168.
Storage Networking Industry Association (2009). *The information lifecycle management maturity model.* Storage Networking Industry Association Press.
Wang, Y., & Byrd, T. A. (2017). Business analytics-enabled decision-making effectiveness through knowledge absorptive capacity in health care. *Journal of Knowledge Management, 21*(3), 517–539.
Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research, 70*, 287–299.
Wang, Y., & Yu, C. (2017). Social interaction-based consumer decision-making model in social commerce: The role of word of mouth and observational learning. *International Journal of Information Management, 37*(3), 179–189.
Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change, 126*, 3–13.
Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analytics-enabled transformation model: Application to health care. *Information & Management, 55*(1), 64–79.
Ward, M. J., Marsolo, K. A., & Froehle, C. M. (2014). Applications of business analytics in healthcare. *Business Horizons, 57*(5), 571–578.
Weiler, A., Grossniklaus, M., & Scholl, M. H. (2016). Situation monitoring of urban areas using social media data streams. *Information Systems, 57*, 129–141.
Yang, C. C., & Ng, T. D. (2011). Analyzing and visualizing web opinion development and social interactions with density-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 41*(6), 1144–1155.
Yang, M., Kiang, M., & Shang, W. (2015). Filtering big data from social media–Building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics, 54*, 230–240.
Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems, 25*(6), 13–16.