Artificial Intelligence in Medicine 71 (2016) 43–56

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Detecting signals of detrimental prescribing cascades from social media

Tao Hoang[a,*], Jixue Liu[a], Nicole Pratt[b], Vincent W. Zheng[c], Kevin C. Chang[d], Elizabeth Roughead[b,1], Jiuyong Li[a,1]

[a] School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, Adelaide, South Australia 5095, Australia
[b] School of Pharmacy and Medical Sciences, University of South Australia, City East Campus, North Terrace, Adelaide, South Australia 5000, Australia
[c] Advanced Digital Sciences Center, 1 Fusionopolis Way, #08-10 Connexis North Tower, Singapore 138632, Singapore
[d] Department of Computer Science, University of Illinois at Urbana-Champaign, 201 N Goodwin Ave, Urbana, IL 61801, United States

ARTICLE INFO

Article history:
Received 2 March 2016
Received in revised form 2 June 2016
Accepted 7 June 2016

Keywords:
Sequence mining
Existence uncertainty
Order uncertainty
Drug
Adverse effect
Detrimental prescribing cascade
Social media

ABSTRACT

Motivation: Prescribing cascade (PC) occurs when an adverse drug reaction (ADR) is misinterpreted as a new medical condition, leading to further prescriptions for treatment. Additional prescriptions, however, may worsen the existing condition or introduce additional adverse effects (AEs). Timely detection and prevention of detrimental PCs is essential as drug AEs are among the leading causes of hospitalization and deaths. Identifying detrimental PCs would enable warnings and contraindications to be disseminated and assist the detection of unknown drug AEs. Nonetheless, the detection is difficult and has been limited to case reports or case assessment using administrative health claims data. Social media is a promising source for detecting signals of detrimental PCs due to the public availability of many discussions regarding treatments and drug AEs.
Objective: In this paper, we investigate the feasibility of detecting detrimental PCs from social media.
Methods: The detection, however, is challenging due to the data uncertainty and data rarity in social media. We propose a framework to mine sequences of drugs and AEs that signal detrimental PCs, taking into account the data uncertainty and data rarity.
Results: We conduct experiments on two real-world datasets collected from Twitter and Patient health forum. Our framework achieves encouraging results in the validation against known detrimental PCs (F_1 = 78% for Twitter and 68% for Patient) and the detection of unknown potential detrimental PCs (Precision@50 = 72% and NDCG@50 = 95% for Twitter, Precision@50 = 86% and NDCG@50 = 98% for Patient). In addition, the framework is efficient and scalable to large datasets.
Conclusion: Our study demonstrates the feasibility of generating hypotheses of detrimental PCs from social media to reduce pharmacists' guesswork.

1. Introduction

Adverse effects (AEs) associated with medicines are among the top causes of deaths (100,000 deaths annually), constituting 5% hospitalizations, and with an estimated medical cost of $75 billion in the USA [13,21,37]. Some drug AEs, however, may be misinterpreted as new medical conditions, leading to the use of additional drugs to treat the AEs. This process is referred to as a *prescribing cascade* (PC) [9,27]. We present a well-known example of PC [12,27] in Fig. 1B. In this example, the PC is a sequence of drugs and AEs, i.e., taking the drug d_1 = "Naproxen", suffering from the AE s_1 = "hypertension" caused by d_1, then treating s_1 with another drug d_2 = "Ramipril". The additional treatments d_2 put patients at the risk of additional AEs, and may also exacerbate the existing conditions. As a result, AEs associated with PCs are costly and are difficult to manage. We define a *detrimental prescribing cascade* (DPC) to be a PC that is associated with a subsequent AE. In Fig. 1B, a patient may suffer from the AE s_2 = "chest pain" as a result of the PC ⟨Naproxen → hypertension → Ramipril⟩.

Timely detection of DPCs is essential for minimizing consequences on health and cost. In particular, identifying DPCs would enable contraindications or warnings to be issued and assist the detection of unknown drug AEs. The earlier example of DPC suggests that "Ramipril" should not be taken after the ADR ⟨Naproxen → hypertension⟩ since the AE "chest pain" may occur.

* Corresponding author.
E-mail address: hoatn002@mymail.unisa.edu.au (T. Hoang).
[1] Senior authors.

http://dx.doi.org/10.1016/j.artmed.2016.06.002
0933-3657/

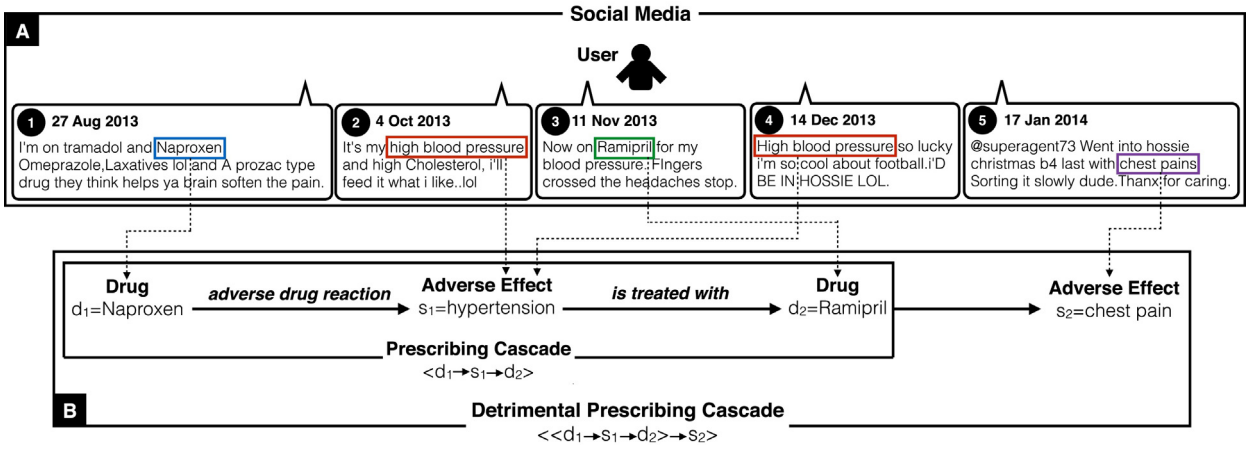

Fig. 1. A signal of DPC on social media. Part B shows an example of a known DPC. Part A presents a user tweeting about the drugs and AEs of the known DPC.

Rather, the ADR should be addressed by dose adjustment, treatment cessation or alternative therapies [32]. Besides, s_2 might be an unknown AEs caused by d_2 or the drug–drug interaction [2,37] between d_1 and d_2, and might be interesting for expert's investigation. In the earlier example of DPC, s_2 = "chest pain" is a known AE of d_2 = "Ramipril" according to *Drugs.com* (http://www.drugs.com/sfx/ramipril-side-effects.html). Most previous works [1,9,12,16] focused solely on discovering PCs, while only Caughey et al. [4] attempted to identify DPCs. Nonetheless, detecting DPCs is difficult with traditional sources and existing approaches. Spontaneous case reports have been the main source of information [28] but the estimated under-reporting rate of drug AEs was shown to be more than 90% [14]. More recently, administrative claims databases have been used [4,16] but the approach has been limited to investigating DPCs on a case-by-case basis. Also, electronic health records [2] have been increasingly utilized for detecting ADRs and drug–drug interactions. Administrative claims databases and electronic health records, while being comprehensive sources, are not publicly accessible. Therefore, an important research gap here is: Can we automatically generate the hypotheses of DPCs for investigation from an open data source?

Social media is a promising data source for detecting signals of DPCs. It was estimated that 11 million people in the United States have posted information about health and treatment issues on social media [5]. Recent studies have demonstrated the availability of such discussions in online health forums (DailyStrength, HealthBoards, etc.) [5,23–25] as well as general social networks (Twitter) [8,17,25,41]. Fig. 1A presents a user tweeting about the DPC mentioned earlier. Interestingly, an analysis revealed that patients tend to discuss their drug AEs on social media before reporting the information to health professionals [40]. Some of the discussed AEs were unknown and interesting for expert's investigation [5]. Furthermore, unlike restrictive traditional sources, the content on social media is publicly available, enabling worldwide detection of DPCs.

In this paper, we investigate the feasibility of detecting DPCs from social media. While previous works have exploited social media for detecting ADRs [5,8,23,24,26,40], drug label changes [6], drug abuse [31] and epidemics [39], none of them has attempted to detect DPCs from social media to the best of our knowledge. Our work reduces pharmacist's guesswork by generating hypotheses that can then be verified using more rigorous but expensive studies [4]. Given a set of users with their posts from social media, benchmark drugs and AEs, we aim to mine sequences of drugs and AEs that signals DPCs with reliable evidence.

Upon investigation, we notice two essential characteristics of social media data for DPC detection: *data uncertainty* and *data rarity*. The lack of context and confusion in unstructured social media posts produce two types of uncertainty in the data. *Existence uncertainty* concerns if a mentioned drug or AE is really consumed or suffered by the user. In the post #4 of Fig. 1A, the user mentions "high blood pressure" without indicating their actual suffering. *Order uncertainty* refers to the unknown actual order of consumed drugs and suffered AEs. In Fig. 1A, the actual occurrence order of "Ramipril" and "hypertension" is unknown as "high blood pressure" is mentioned both before and after "Ramipril". In fact, it is hard to determine the actual time of drug consumption or AE suffering due to the existence uncertainty and the scarcity of temporal evidence in the posts. In addition, DPCs are *rare* in social media as each user may consume and suffer from very different drugs and AEs. We observe from experiments that a DPC often occurs in less than 10 out of 100,000 users.

Mining DPCs is challenging due to the *data uncertainty* and *data rarity* in social media. First, the data uncertainty and data rarity induce difficulties in selecting reliable DPCs as the supporting evidence is scarce. Furthermore, the uncertainty and rarity pose a challenge to the scalable detection of DPCs. Fig. 2 presents a database example of two users with existence and order probabilities for drugs and AEs. An uncertain database may correspond to numerous different *possible worlds*, each of which is a unique combination of alternatives for all uncertain data items and exists with a different probability [35]. For instance, the uncertain database in Fig. 2 corresponds to 10 different possible worlds, each of which has a unique set of drugs, AEs and orders. The probabilities of all the possible worlds in an uncertain database sum to 1. The number of possible worlds, however, grows exponentially with respect to the number of uncertain values. When the total number of drugs and AEs is n, there are at least 2^n possible worlds (considering only existence uncertainty). In the database of Fig. 2, $n = 3$ and the number of possible worlds is $10 > 2^3$. Also, due to the order uncertainty, we might need to consider all possible permutations of drugs and AEs with different probabilities. Imagine a benchmark set of 1400 drugs and 6100 AEs. In the worst case, the search space of DPCs as permutations of drugs and AEs is $(1400 \times 6100)^2$. To make matters worse, since DPCs are rare, very few sequences can be pruned from the search space. Our experiments show that it takes more than one day to mine DPCs from a dataset of 100,000 users using a general sequence mining algorithm for uncertain data [42].

The state-of-the-art method to detect DPCs [4], however, is not adaptable to our problem. First, it focuses on examining the statistical significance of a given DPC signal, where the given signal is extracted from case reports or manually hypothesized by research pharmacists. On the other hand, we aim to automatically generate the DPC signals before investigating their statistical significance. Additionally, the statistical significance measure used in

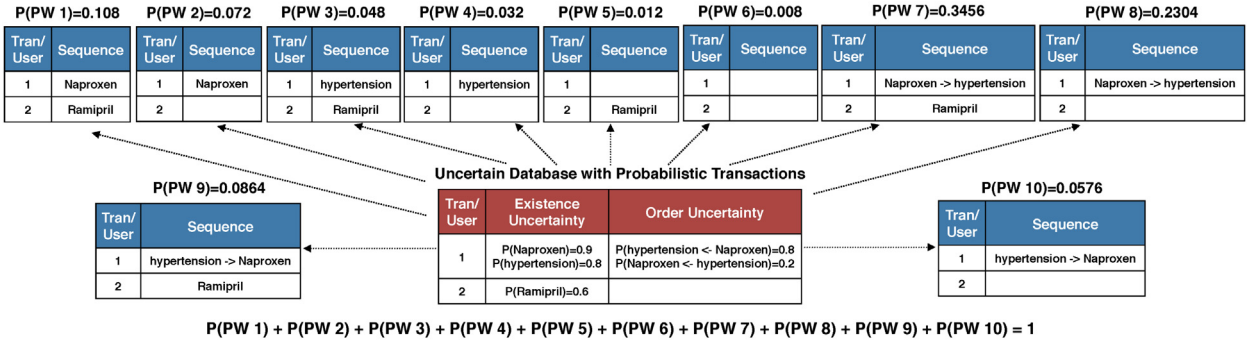

Fig. 2. Possible worlds in an uncertain database. A toy example of ten different possible worlds (blue) of an uncertain database (red) with different existence probabilities that sum to one. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the method, i.e., sequence ratio does not take into account the data uncertainty.

As a result, we develop a framework to detect signals of DPCs from social media as highlighted in Fig. 3, which tackles the above challenges. Our framework consists of two main components. First, we adopt existing techniques to build an uncertain database from social media posts, i.e., recognizing drugs, AEs, and estimating data uncertainty. Then, given the uncertain database, we design a candidate generate-and-test approach based on the Generalized Sequential Pattern (GSP) mining algorithm [34] to identify and prioritize DPCs of interest. In particular, we formulate the criteria to select DPCs of interest that address the data rarity, and extend state-of-the-art probabilistic techniques [35,38,42] to determine the likelihood that a sequence satisfies the criteria. Additionally, we exploit the characteristics in such criteria to prune the search space, which improves the efficiency by more than 100 times. The validation against known DPCs [12,27] demonstrates the feasibility of our method to detect DPCs (F_1 = 78% for the Twitter dataset and 68% for the Patient dataset). Furthermore, our method achieves encouraging results in detecting unknown potential DPCs (Precision@50 = 72% and NDCG@50 = 95% on the Twitter dataset, Precision = 86% and NDCG@50 = 98% for the Patient dataset). All the datasets, source codes and evaluation results are available at http://nugget.unisa.edu.au/PCSocialMedia/.

2. Methods

Fig. 3 illustrates our framework. We first describe how to build an uncertain database from a given set of users with their posts. Then we formulate the criteria to select DPCs and present the algorithm to mine DPCs from the uncertain database.

2.1. Building uncertain database

2.1.1. Recognizing users' drugs and AEs

The very first step towards building an uncertain database is to recognize the drugs and AEs that may be consumed and suffered by each user. For instance, the user in Fig. 1A may take "Naproxen", "Ramipril" and suffer from "hypertension", "chest pain". In order to recognize the drugs and AEs, we need a benchmark set of drugs and AEs. We collect a set of 1430 drugs and 6155 AEs from SIDER 2 [18]. Each drug has an ID (e.g., 5002), a generic name (e.g., "Quetiapine") and a brand name (e.g., "Seroquel"). Each AE has an ID (e.g., C0027497) and multiple names ("nausea", "nauseous", etc.).

We locate mentions of drugs and AEs in each post using dictionary lookup [40]. There are two reasons behind this design decision. First, annotations are expensive to obtain and thus unavailable in this work. Second, using dictionary lookup allows us to map the mentions of drugs and AEs to the corresponding IDs in SIDER 2. Such mapping is required in our method to utilize the ADR and treatment relationships between drugs and AEs in SIDER 2 for mining potential DPCs in the later stage. Particularly, we first use TweetNLP [10] (for tweets only) and Lingpipe (http://alias-i.com/lingpipe/) (for other posts) to split a post into sentences. In each sentence, we remove stop-words. For tweets, we also remove URLs, tokens starting with @ while keeping hashtags (but removing prefix #). For each sentence, we search n-grams against the medical dictionaries to identify mentions of drugs and AEs. A mention refers to a drug or an AE if it matches one of the names of the drug or AE. Each mention only refer to one drug or one AE. For overlapping mentions that refer to different drugs or AEs, we only retain the one of longest length. We discard posts mentioning no drug and AE. The posts are grouped by user.

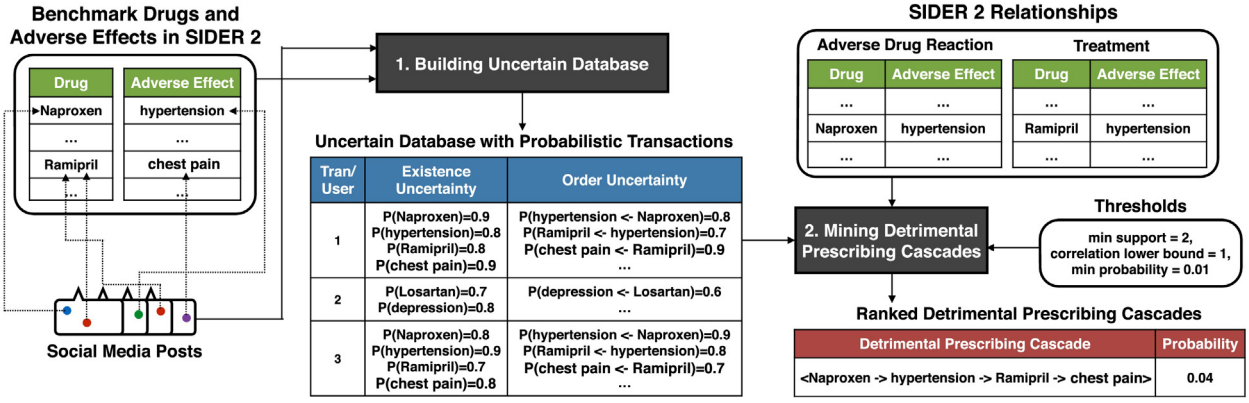

Fig. 3. Our framework to detect DPCs from social media. The framework contains two main components: (1) building an uncertain database given social media posts and benchmark set of drugs and AEs in SIDER 2, (2) mining DPCs given the uncertain database, SIDER 2 relationships and thresholds.

2.1.2. Building probabilistic transactions

Given the drugs and AEs recognized from users' posts, our next step is to construct a set of transactions, from which we will mine the DPCs. We observe that a DPC is usually mentioned across different posts of a user rather than in merely one post. For instance, the drugs and AEs of the DPC in Fig. 1B are scattered in five different posts in Fig. 1A. In addition, the drugs and AEs consumed and suffered by one user are intuitively independent of the others. As a result, each user is a natural choice to be a transaction in the database. From now on, a transaction refers to a user and vice versa. We define a *transaction* as a sequence of drugs and AEs that might be consumed and suffered by the corresponding user.

We notice two types of *data uncertainty* due to the limited context and confusion in social media posts.

Existence uncertainty. In the post #4 of Fig. 1A, for instance, the user only mentions "high blood pressure" without explicitly stating whether the user really suffered from it. We refer to this uncertainty as *existence uncertainty*. Let $P(x \sqsubseteq u)$ denote the probability that a drug or AE x occurs in a transaction u. We observe that users tend to repeatedly mention their consumed drugs and suffered AEs across different posts. In Fig. 1A, for instance, "high blood pressure" is mentioned multiple times in two posts. Logistic function has been widely used in binary classification to measure the probability that an observation belongs to a particular class given the features of the observation [7]. As a result, we utilize the logistic function to approximate the probability that a drug or AE x occurs in a transaction u based on the number of user u's posts mentioning x. In particular, we compute $P(x \sqsubseteq u) = \exp(n_{x,u})/[\exp(n_{x,u}) + 1]$, where $n_{x,u}$ is the number of u's posts in which x is mentioned. In Fig. 1A, we have $P(\text{hypertension} \sqsubseteq u) = e^2/(e^2 + 1) = 0.88$. $P(x \sqsubseteq u)$ increases as x is mentioned in more posts by user u. It is worth noting that some previous studies including [30] took into account whether a post mentions personal experience of AEs based on the context. While such technique might help better quantify the existence uncertainty, it requires annotated data, which are unavailable in our case. Section 4 discusses a possible adaptation of the technique in the presence of annotated data.

Order uncertainty. In Fig. 1A, we are also unsure if the user suffered from "hypertension" before or after taking "Ramipril" as "high blood pressure" is mentioned both before and after "Ramipril". It is hard to determine the actual time of taking "Ramipril" and suffering from "hypertension" due to the existence uncertainty and the scarcity of temporal evidence in the posts. We refer to this uncertainty as *order uncertainty*. Let y be another drug or AE that is different from x. Intuitively, the more we observe y being mentioned after x in posts, the more likely y has occurred after x. Denote $P(y \leftarrow x \sqsubseteq u)$ as the probability that y occurs after x given the existence of x and y in u. For any x and y, we assume that there must be an order between them, i.e., one starting before the other. We again employ the logistic function [7] to estimate $P(y \leftarrow x \sqsubseteq u) = \exp(n_{y \leftarrow x,u} - n_{x \leftarrow y,u})/[\exp(n_{y \leftarrow x,u} - n_{x \leftarrow y,u}) + 1]$, where $n_{x \leftarrow y,u}$ is the number of times we observe two different posts p_y, p_x such that y is mentioned in p_y, x is mentioned in p_x, and p_y was written after p_x. Following the previous PC detection studies [4], we set the maximum time interval between p_x and p_y to 1 year. Our earlier order assumption holds since $P(y \leftarrow x \sqsubseteq u) + P(x \leftarrow y \sqsubseteq u) = 1$. As an example, in Fig. 1A, $P(\text{Ramipril} \leftarrow \text{hypertension} \sqsubseteq u) = e^{1-1}/(e^{1-1} + 1) = 0.5$.

Due to the data uncertainty, we refer to our database as an *uncertain database*, where each transaction in the database is called a *probabilistic transaction*. Fig. 2 presents an example of an uncertain database, in which there are two probabilistic transactions. In the database, for instance, $P(\text{Naproxen} \sqsubseteq u_1) = 0.9$ and $P(\text{hypertension} \leftarrow \text{Naproxen} \sqsubseteq u_1) = 0.8$. Conceptually, an uncertain database is a set of deterministic databases or *possible worlds* [35], each of which is a unique combination of alternatives for all uncertain data items and exists with a different probability [35]. The probabilities of all possible worlds of an uncertain database sum to one. Consider the possible world PW 7 in Fig. 2. In this possible world, the user in transaction 1 takes "Naproxen" then suffers from "hypertension" while the user in transaction 2 takes "Ramipril". The probability of this possible world is thus $0.9 \times 0.8 \times 0.8 \times 0.6 = 0.3456$. A transaction in a possible world is called a *deterministic transaction*.

2.2. Mining DPCs from uncertain database

For the sake of understanding, we first describe the criteria that defines a DPC in a possible world, followed by a generalization to an uncertain database. We then present an algorithm to mine DPCs from an uncertain database.

2.2.1. DPCs in a possible world

A DPC is in nature a sequence of drugs and AEs, e.g., ⟨Naproxen → hypertension → Ramipril → chest pain⟩. For generalization, let d_1, s_1, d_2, s_2 denote four elements of a DPC, i.e., d_1 = "Naproxen", s_1 = "hypertension", d_2 = "Ramipril" and s_2 = "chest pain".

Frequency criterion. First, to ensure the DPC is supported by sufficient evidence, we require it to occur to at least some number of users. Let X be an arbitrary sequence (e.g., a DPC) and denote $sup(X)$, the *support* of X as the number of deterministic transactions containing X in an arbitrary possible world. Since each user may consume and suffer from very different drugs and AEs, the support of a DPC is expected to be low in general. In fact, we observe from experiments that a DPC often occurs in less than 10 out of 100,000 users. This phenomenon was referred to as *rare item problem* in previous works, and has been common in other medical applications, e.g., detecting ADRs [40]. As a result, we require $sup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle] \geq minsup = 2$, i.e., two or more users experience the DPC.

However, a low *minsup* may result in numerous irrelevant sequences. Thus, we exploit the characteristics of a DPC to introduce additional criteria.

PC criterion. Based on the definition of PC [9,27], s_1 is an AE caused by d_1 while d_2 is used to treat s_1. To eliminate irrelevant candidates, we again employ SIDER 2 [18] to enforce the ADR relationship between d_1 and s_1, and the treatment relationship between s_1 and d_2. Particularly, let $I = \{(d, s)\}$ be the set of 18,272 treatment pairs (e.g., "Ramipril" treating "hypertension") and $A = \{(d, s)\}$ the set of 179,102 ADR pairs (e.g., "Naproxen" causing "hypertension") collected from SIDER 2. Hence, we require $(d_1, s_1) \in A$ and $(d_2, s_1) \in I$. Note that the set of treatment pairs and ADR pairs are non-overlapping, i.e., $A \cap I = \emptyset$. In addition, we constrain that $d_1 \neq d_2$.

AE criterion. To ensure the AE s_2 does not occur by chance, we enforce s_2 to be strongly correlated with the PC $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle$. While *confidence* is a common measure for association in data mining, it does not work well in our case due to the data rarity [36]. To address the rarity issue, we employ *lift* as a correlation measure [3]. Lift has been widely used in existing medical applications, e.g., detecting ADRs [6,40]. Let N be the total number of transactions. Lift measures how many times more often s_2 occurs after $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle$ than expected if they were statistically independent.

$$lift[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle \Rightarrow s_2] = \frac{\sup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle] \times N}{\sup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle] \times \sup(s_2)} \quad (1)$$

where $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle \rightarrow s_2$ represents the correlation between the PC and the AE. $lift[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle \rightarrow s_2] = 1$ indicates that the PC and the AE are statistically independent, $lift[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle \rightarrow s_2] > 1$ that they are positively correlated, and $lift[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle \rightarrow s_2] < 1$

that they are negatively correlated. $|lift - 1|$ indicates the correlation strength.

Our goal is to find s_2 such that $lift[\langle d_1 \to s_1 \to d_2\rangle \to s_2] > corrLB = 1$, where $corrLB$ is the correlation lower bound. In addition, we constrain s_2 to have no treatment relationship with d_1 and d_2 based on SIDER 2, i.e., $(d_1, s_2) \notin I$ and $(d_2, s_2) \notin I$. This is to eliminate user's prior medical conditions to be treated by d_1 or d_2. We also require $s_2 \neq s_1$.

2.2.2. DPCs in an uncertain database

Given the criteria to select DPCs in a possible world, we now describe the criteria to select DPCs from an uncertain database, i.e., multiple possible worlds. The key measure that quantifies the likelihood that a sequence is indicative of DPC in an uncertain database is the fraction of the possible worlds in which the sequence satisfies the criteria in Section 2.2.1. Computing such measure, however, requires enumerating all possible worlds, which are exponential in the number of uncertain values. This section presents an approach to compute the measure efficiently with Poisson approximation.

Let $P(X \in \text{DPC})$ denote the probability that a sequence $X = \langle d_1 \to s_1 \to d_2 \to s_2\rangle$ is a DPC in an uncertain database. We define $P(X \in \text{DPC})$ as the fraction of the uncertain database's possible worlds in which X satisfies the criteria to be a DPC. Let $C_X = (d_1, s_1) \in A \wedge d_2 \neq d_1 \wedge (d_2, s_1) \in I \wedge s_2 \neq s_1 \wedge (d_1, s_2) \notin I \wedge (d_2, s_2) \notin I$. Denote W as the set of possible worlds and $\mathbb{1}(C)$ the indicator function, i.e., $\mathbb{1}(C) = 1$ if C is true and 0 otherwise, we have:

$$\begin{aligned} P(X \in \text{DPC}) &= \sum_{w \in W} \mathbb{1}[C_X \text{ istrue} \wedge \text{sup}(X) \geq minsup \wedge lift[\langle d_1 \to s_1 \to d_2\rangle \Rightarrow s_2] > corrLB] \\ &= \mathbb{1}[C_X \text{ istrue}] \sum_{w \in W} \mathbb{1}[\text{sup}(X) \geq minsup \wedge lift[\langle d_1 \to s_1 \to d_2\rangle \Rightarrow s_2] > corrLB] \\ &= \mathbb{1}[C_X \text{ istrue}] P[\text{sup}(X) \geq minsup \wedge lift[\langle d_1 \to s_1 \to d_2\rangle \Rightarrow s_2] > corrLB] \end{aligned} \quad (2)$$

A sequence $X = \langle d_1 \to s_1 \to d_2 \to s_2\rangle$ is a DPC in an uncertain database if $P(X \in \text{DPC}) \geq minprob$, where $minprob > 0$ is a user-specified minimum probability.

The bottleneck in computing $P(X \in \text{DPC})$ is $P[sup(X) \geq minsup \wedge lift[\langle d_1 \to s_1 \to d_2\rangle \to s_2] > corrLB]$. Computing $P[sup(X) \geq minsup \wedge lift[\langle d_1 \to s_1 \to d_2\rangle \to s_2] > corrLB]$ by enumerating all possible worlds is intractable since the number of possible worlds grows exponentially with respect to the number of uncertain values. When the total number of drugs and AEs is n, there are at least 2^n possible worlds (considering only existence uncertainty). For instance, in the database of Fig. 2, $n = 3$ and the number of possible worlds is $10 > 2^3$.

Here we present a polynomial-time method to compute $P[sup(X) \geq minsup \wedge lift[\langle d_1 \to s_1 \to d_2\rangle \to s_2] > corrLB]$. Let $sup(Y)$ be the number of deterministic transactions in an arbitrary possible world in which s_2 does not occur after $\langle d_1 \to s_1 \to d_2\rangle$ given that $\langle d_1 \to s_1 \to d_2\rangle$ has occurred. Also, let $sup(Z)$ be the number of deterministic transactions in an arbitrary possible world in which $\langle d_1 \to s_1 \to d_2\rangle$ does not occur before s_2 given that s_2 has occurred. With some abuse of notation, we denote $msup(X)$ as the number of probabilistic transactions in which $P(X \sqsubseteq u) > 0$. The same interpretation applies for $msup(Y)$ and $msup(Z)$. Remember that N denotes the total number of transactions, $\lfloor . \rfloor$ the floor function and $min(a, b)$ the smaller value between a and b, our further manipulation results in (proof in Section S1 of the Supplementary Information):

$$\begin{aligned} &P[\text{sup}(X) \geq minsup \wedge lift[\langle d_1 \to s_1 \to d_2\rangle \Rightarrow s_2] > corrLB] \\ &= \sum_{i=minsup}^{msup(X)} P[\text{sup}(X) = i] \sum_{j=0}^{\min\left(\lfloor \frac{N}{corrLB} - i \rfloor, msup(Y)\right)} P[\text{sup}(Y) = j] \sum_{k=0}^{\min\left(\lfloor \frac{Ni}{corrLB(i+j)} - i \rfloor, msup(Z)\right)} P[\text{sup}(Z) = k] \end{aligned} \quad (3)$$

In order to compute the expression in Eq. (3), we need to examine the underlying distributions of $sup(X)$, $sup(Y)$ and $sup(Z)$ [42,38,35]. We notice that X exists in a probabilistic transaction with a probability due to the existence and order uncertainty. Let $P(X \sqsubseteq u)$ be the probability that $X = \langle d_1 \to s_1 \to d_2 \to s_2\rangle$ exists in a probabilistic transaction u, which equals to the product of existence and order probabilities of X's individual drugs and AEs in u:

$$P(X \sqsubseteq u) = P(d_1 \sqsubseteq u)P(s_1 \sqsubseteq u)P(d_2 \sqsubseteq u)P(s_2 \sqsubseteq u)P(s_1 \leftarrow d_1 \sqsubseteq u)P(d_2 \leftarrow s_1 \sqsubseteq u)P(s_2 \leftarrow d_2 \sqsubseteq u) \quad (4)$$

Note that the computation of $P(X \sqsubseteq u)$ is applicable for general sequences. Denote U as the set of probabilistic transactions in an uncertain database. Given a sequence X, as X exists in a transaction $u \in U$ with probability $P(X \sqsubseteq u)$, u imitates the behaviour of a Bernoulli trial with success probability $P(X \sqsubseteq u)$. In addition, since all transactions are independent, the number of transactions in a possible world of the uncertain database in which X occurs, $sup(X)$, is a random variable following the Poisson Binomial distribution [19], i.e., $sup(X) \sim \text{PoissonBinomial}(\{P(X \sqsubseteq u)\}_{u \in U})$. Denote $P(Y \sqsubseteq u)$ as the probability that s_2 does not occur after $\langle d_1 \to s_1 \to d_2\rangle$ given that $\langle d_1 \to s_1 \to d_2\rangle$ has occurred. Denote $P(Z \sqsubseteq u)$ as the probability that $\langle d_1 \to s_1 \to d_2\rangle$ does not occur before s_2 given that s_2 has occurred. It follows that $P(Y \sqsubseteq u) = P[\langle d_1 \to s_1 \to d_2\rangle \sqsubseteq u] - P(X \sqsubseteq u)$ and $P(Z \sqsubseteq u) = P(s_2 \sqsubseteq u) - P(X \sqsubseteq u)$. With similar reasoning, we can deduce $sup(Y) \sim \text{PoissonBinomial}(\{P(Y \sqsubseteq u)\}_{u \in U})$ and $sup(Z) \sim \text{PoissonBinomial}(\{P(Z \sqsubseteq u)\}_{u \in U})$.

The probability mass function (pmf) of Poisson Binomial distribution, however, does not have a closed form. We therefore employ Le Cam's theorem [19] to approximate the Poisson Binomial pmf via Poisson distribution efficiently [38]. Here we only demonstrate how to approximate $P[sup(X) = i]$. Similar computation applies for $P[sup(X) = j]$ and $P[sup(Z) = k]$. Let $E[sup(X)]$ be the expectation of the support of X in the uncertain database, which is given by $E[sup(X)] = \sum_{u \in U} P(X \sqsubseteq u)$ [19]. According to Le Cam's theorem, the Poisson Binomial distribution followed by $sup(X)$ can be approximated by a Poisson distribution with mean $E[sup(X)]$. As a result, we have:

$$P[\text{sup}(X) = i] \approx \frac{E[\text{sup}(X)]^i}{i!} \exp(-E[\text{sup}(X)]) \quad (5)$$

Here we use an example to demonstrate how to compute $P(X \in \text{DPC})$. Consider the uncertain database in Fig. 3 and

$X = \langle \text{Naproxen} \rightarrow \text{hypertension} \rightarrow \text{Ramipril} \rightarrow \text{chest pain} \rangle$. Also, suppose $minsup = 2$, $corrLB = 1$ and $minprob = 0.01$. For transaction 1, we have $P(X \sqsubseteq u_1) = 0.9 \times 0.8 \times 0.8 \times 0.9 \times 0.8 \times 0.7 \times 0.9 = 0.26$ based on Eq. (4). Likewise, $P(X \sqsubseteq u_2) = 0$ and $P(X \sqsubseteq u_3) = 0.2$. As a result, $E[sup(X)] = 0.46$ and $msup(X) = 2$. In a similar way, we can compute $E[sup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle]] = 0.68$, $E[sup(s_2)] = 1.7$. Hence, $E[sup(Y)] = 0.68 - 0.46 = 0.22$ and $E[sup(Z)] = 1.7 - 0.46 = 1.24$. In addition, we have $msup(Y) = msup(Z) = 2$. Since C_X is true according to SIDER 2, $P(X \in \text{DPC}) = P[sup(X) \geq 2 \wedge lift[\langle \text{Naproxen} \rightarrow \text{hypertension} \rightarrow \text{Ramipril} \rangle \rightarrow \text{chestpain}] > 1] = 0.04$ according to Eqs. (2)–(5).

2.2.3. The algorithm

Given the criteria to select DPCs from an uncertain database in Section 2.2.2, we now present an algorithm to mine the sequences that signal DPCs from an uncertain database. A straightforward approach is generating all possible sequences $X = \langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle$ and test if each of them satisfies the criteria in Section 2.2.2. However, given that our benchmark set contains 1430 drugs and 6155 AEs, the number of generated sequences can be huge, i.e., $(1430 \times 6155)^2$. Thus, this approach is not computationally feasible. Instead, our algorithm reduces the number of generated sequences by expanding the size of sequences one by one and stop the expansion once any generated sequence violates the criteria. In this section, we describe the details of our algorithm in following.

We design a candidate generate-and-test algorithm named *Probabilistic DPC Miner (PDPCMine)* based on the Generalized Sequential Pattern (GSP) mining algorithm [34] to mine DPCs from an uncertain database. GSP was invented to mine frequent sequences from a deterministic database, i.e., a possible world. The idea of GSP is to iteratively expand sequences of sizes k into $k+1$ and test whether the newly generated sequences are frequent, i.e., $sup(X) \geq minsup$. If a sequence is not frequent, it will be pruned and not further expanded in next iterations. The rationale behind this pruning strategy is the anti-monotonicity property of frequent sequences that super-sequences of an infrequent sequence are also infrequent.

In order to mine the DPCs, we need to modify the GSP algorithm in both the *candidate generation* and *candidate test*. For candidate generation, while GSP assumes deterministic order of items in a sequence, we need to consider all possible permutations of drugs and AEs with different probabilities due to the order uncertainty. In addition, when the sequences are expanded, we need to ensure certain criteria in $C_X = (d_1, s_1) \in A \wedge d_2 \neq d_1 \wedge (d_2, s_1) \in I \wedge s_2 \neq s_1 \wedge (d_1, s_2) \notin I \wedge (d_2, s_2) \notin I$ are met. Furthermore, since each user may consume and experience very different drugs and AEs, generating sequences of drugs and AE at the global-level as in GSP produces a huge number of cross-transaction sequences, i.e., sequences having zero support. We therefore generate sequences at the transaction-level instead. For the candidate test, we utilize $P(X \in \text{DPC})$ as the measure to select sequences instead of support as in GSP. Additionally, we adopt the following frequency-based criterion [35] to reduce the number of sequences for which $P(X \in \text{DPC})$ is to be computed (proof in Section S2 of the Supplementary Information).

Frequency pruning. A sequence $X = \langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle$ is not a DPC in an uncertain database if there exists a subsequence $X' \sqsubseteq X$ that satisfies one of the following conditions:

1. $msup(X') < minsup$, or
2. $\sigma(X') \geq 2e - 1$ and $2^{-\sigma(X')E[\sup(X')]} < minprob$, or
3. $0 < \sigma(X') < 2e - 1$ and $\exp\left(\frac{-\sigma^2 E[\sup(X')]}{4}\right) < minprob$

where $\sigma(X') = (minsup - E[sup(X')] - 1)/E[sup(X')]$

PDPCMine algorithm. Fig. 4 presents an example to demonstrate the four main steps in our algorithm. Given a set of probabilistic transactions U in an uncertain database, a set of treatment pairs I, a set of ADR pairs A, thresholds *minsup*, *corrLB*, *minprob*, we aim to find the set of DPCs, denoted as H_4. Our algorithm goes as follows:

1. Find the set D_1 of drugs and the set S_1 of AEs that are not removed by the frequency pruning.

   Let D_u and S_u be the set of drugs and AEs recognized in user u's posts. Initialize D_1 and S_1 to empty sets. Compute D_1 as

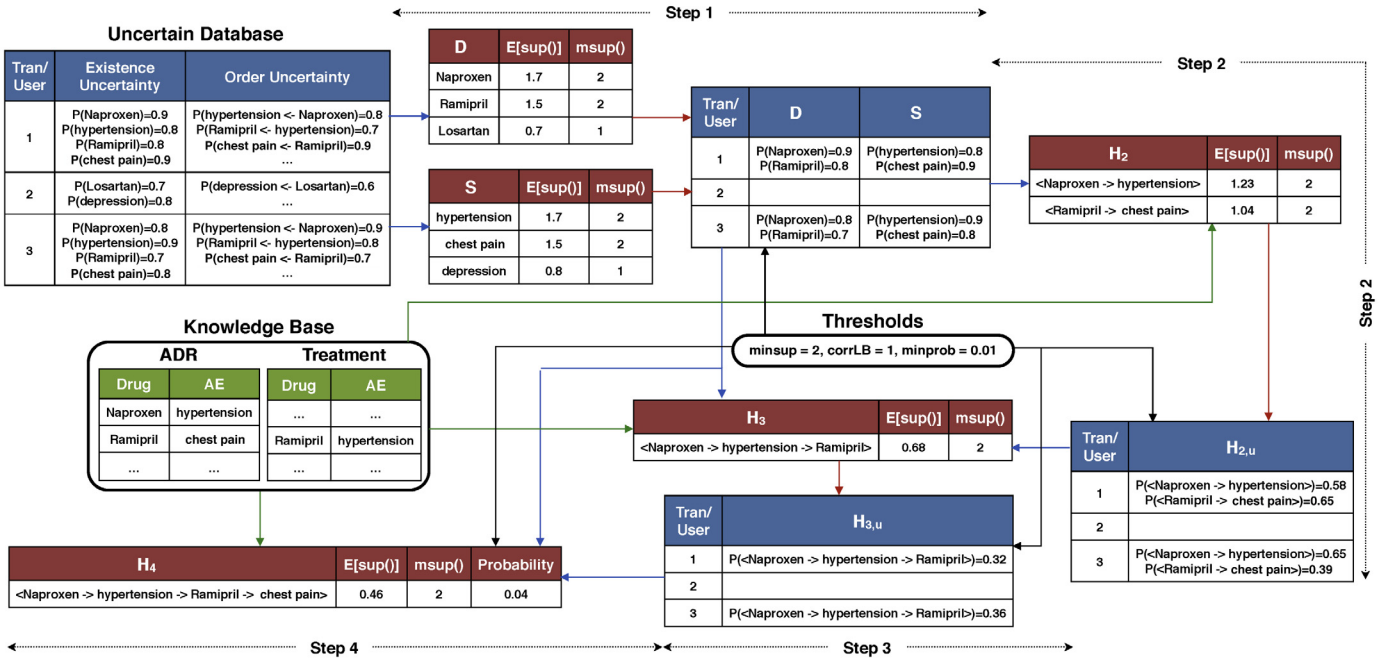

Fig. 4. A step-by-step demonstration of our PDPC-Mine algorithm to mine DPCs from an uncertain database. The inputs and outputs of four main steps in the algorithm are presented when the algorithm is applied on a toy example.

the union of drugs from $\{D_u\}_{u \in U}$ and accumulate $E[sup(d)]$ and $msup(d)$ for each $d \in D_1$ at the same time. Similarly, compute S_1 as the union of AEs from $\{S_u\}_{u \in U}$ and accumulate $E[sup(s)]$ and $msup(s)$ for each $s \in S_1$ at the same time. Then prune D_1 and S_1 based on frequency pruning. Finally, for each user u, set $D_{1,u}$ as the intersection of D_u and D_1, $S_{1,u}$ as the intersection of S_u and S_1.

2. Find the set H_2 of sequences that satisfy the ADR relationship and are not removed by the frequency pruning.

   Initialize H_2 to empty sets. For each user u, select $(d_1, s_1) \in D_{1,u} \times S_{1,u}$ such that $(d_1, s_1) \in A$ and add them to H_2. At the same time, compute $P[\langle d_1 \rightarrow s_1 \rangle \sqsubseteq u] = P(d_1 \sqsubseteq u)P(s_1 \sqsubseteq u)P(s_1 \leftarrow d_1 \sqsubseteq u)$ and accumulate $E[sup[\langle d_1 \rightarrow s_1 \rangle]]$ and $msup[\langle d_1 \rightarrow s_1 \rangle]$. Then prune H_2 based on frequency pruning. Finally, for each user u, set $H_{2,u}$ as the intersection of $D_{1,u} \times S_{1,u}$ and H_2.

3. Find the set H_3 of sequences that satisfy the criteria to be PCs and are not removed by the frequency pruning.

   Initialize H_3 to empty sets. For each user u, select $(d_1, s_1, d_2) \in H_{2,u} \times D_{1,u}$ such that $d_2 \neq d_1$ and $(d_2, s_1) \in I$, and add them to H_3. At the same time, compute $P[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle \sqsubseteq u] = P[\langle d_1 \rightarrow s_1 \rangle \sqsubseteq u]P[d_2 \sqsubseteq u]P[d_2 \leftarrow s_1 \sqsubseteq u]$ and accumulate $E[sup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle]]$ and $msup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle]$. Then prune H_3 based on frequency pruning. Finally, for each user u, set $H_{3,u}$ as the intersection of $H_{2,u} \times D_{1,u}$ and H_3.

4. Find the set H_4 of sequences that satisfy the criteria to be DPCs.

   Initialize H_4 to empty sets. For each user u, select $(d_1, s_1, d_2, s_2) \in H_{3,u} \times S_{1,u}$ such that $s_2 \neq s_1$ and $(d_1, s_2) \notin I$ and $(d_2, s_2) \notin I$, and add them to H_4. At the same time, compute $P[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle \sqsubseteq u] = P[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle \sqsubseteq u]P[d_2 \sqsubseteq u]P[d_2 \leftarrow s_2 \sqsubseteq u]$ and accumulate $E[sup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle]]$ and $msup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle]$. Then prune H_4 based on frequency pruning. For each $X \in H_4$, compute $P(X \in \text{DPC})$. If $P(X \in \text{DPC}) < minprob$, remove X from H_4. Rank the DPCs in H_4 in descending order of $P(X \in \text{DPC})$, then descending order of $E[sup(X)]$.

5. Output H_4 as results.

3. Results and analysis

3.1. Real-world evaluation datasets

We collect two datasets from two different social media sites *Twitter* (https://twitter.com/) (general social media) and *Patient* (http://patient.info/forums) (health forum). Table 2 summarizes the details of the datasets. For the Twitter dataset, since Twitter is huge, we need a finite set of queries to collect the tweets. It would be ideal to utilize the benchmark set of 1430 drugs and 6155 AEs from SIDER 2 as queries to collect relevant tweets. However, due to time and resource constraints, we need to select a smaller set. We first compile a set of 45 popular drugs for common diseases and conditions from *Drugs.com* (http://www.drugs.com/medical_conditions.html) and group them by classes as in Table 1. Then for each drug, we query its various names, i.e., brand name and generic name to collect tweets mentioning the names. Since the set of 45 drugs are popular, the collected tweets are relatively extensive. Besides, we observe that AEs are often mentioned in tweets mentioning drugs and tweets in the same threads. Hence, instead of select AEs to collect tweets, we collected posts containing drugs and also their same-thread posts to reduce the time. In total, we collect 426,217 tweets from November 2007 to August 2015. For the Patient dataset, we enumerate all 32 health topics in the site and collect 194,353 posts from July 2005 to March 2015.

3.2. Validation against known DPCs

We validate our method against the DPCs whose PCs have been reported in previous works. Table 3 summarizes the details of the known DPCs. First, from two known generic PC ⟨NSAIDs → hypertension → Antihypertensives⟩ [12,27] and ⟨ACE Inhibitors → cough → Cough Suppressants⟩ [15], we derive 12 PCs by assigning corresponding drugs to the classes NSAIDs,

Table 1
Benchmark drugs for querying Twitter. The list of 45 popular drugs for common diseases and conditions from *Drugs.com* grouped by classes.

Class	Drug
Antipsychotics	Quetiapine (Seroquel), Risperidone (Risperdal), Olanzapine (Zyprexa), Aripiprazole (Abilify)
Antidepressants	Sertraline (Zoloft), Venlafaxine (Effexor), Trazodone (Desyrel), Paroxetine (Paxil), Citalopram (Celexa), Escitalopram (Lexapro), Bupropion (Wellbutrin), Duloxetine (Cymbalta), Amitriptyline (Elavil)
Anticonvulsants	Lamotrigine (Lamictal), Tiagabine (Gabitril), Gabapentin (Neurontin), Topiramate (Topamax), Pregabalin (Lyrica)
Anticoagulants	Warfarin (Counmadin), Rivaroxaban (Xarelto), Dabigatran (Pradaxa), Apixaban (Eliquis)
Antibiotics	Minocycline (Dynacin), Amoxicillin (Amoxil), Linezolid (Zyvox)
NSAIDs	Aspirin, Ibuprofen (Advil, Motrin), Celecoxib (Celebrex), Meloxicam (Mobic)
Beta-blockers	Propranolol (Inderal), Metoprolol
Alpha-blockers	Prazosin (Minipress)
Alpha-agonist Hypotensive Agents	Clonidine (Catapres)
Cancer	Avastin (Bevacizumab)
Opioid Pain, Narcotic	Codeine
ACE Inhibitors	Lisinopril (Prinivil, Zestril), Ramipril (Altace)
Statins	Rosuvastatin (Crestor), Atorvastatin (Lipitor), Simvastatin (Zocor)
Anemia	Epoetin Alfa (Procrit, Epogen)
Leukotriene Receptor Antagonists	Montelukast (Singulair)
Angiotensin II Receptor Antagonists	Losartan (Cozaar)
Phosphodiesterase Inhibitors	Sildenafil (Viagra)
SSRIs	Fluoxetine (Prozac)
Potent Diuretic	Furosemide (Lasix)

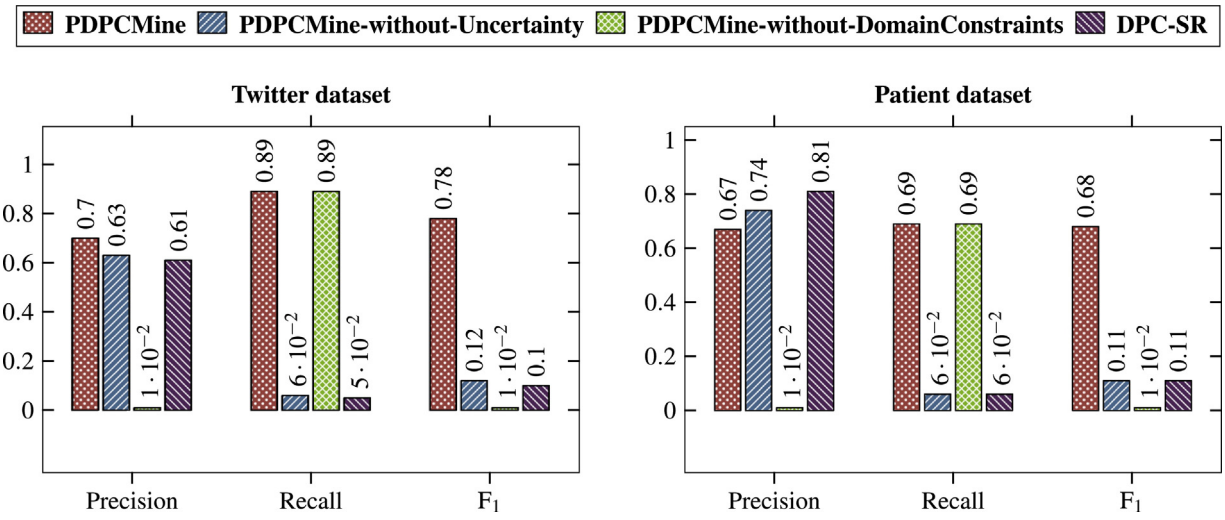

Fig. 5. Performance of our method on validation datasets. The performance of the proposed algorithm *PDPCMine*, its alternative settings and state-of-the-art baseline *DPC-SR* in Precision, Recall and F_1 are plotted. *PDPCMine-without-Uncertainty* does not take into account the data uncertainty while *PDPCMine-without-DomainConstraints* does not require the domain constraints in SIDER 2 to be satisfied. *DPC-SR* utilizes $SR(X)$ as the statistical significance measure to test whether X is a DPC.

Antihypertensives, ACE Inhibitors and Cough Suppressants. We select these PCs since their individual drugs and AEs are mentioned in our two datasets. Then for each PC $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle$, we employ the medical knowledge base *Drugs.com* to obtain the known AEs s_2 of either d_2 (http://www.drugs.com/sfx/) or the drug–drug interaction between d_1 and d_2 (http://www.drugs.com/drug_interactions.php). Since the known AEs are numerous, we only choose those AEs that are common to all the PCs in each generic PC. In total, we generate 406 known DPCs by combining the known PCs and their known AEs. We extract the portions of our datasets that are relevant to the known DPCs in Table 3 for validation. Particularly, we extract 169,799 posts (23,026 users) from the Twitter dataset and 155,955 posts (8707 users) from the Patient dataset. We apply our method in the sub-datasets and recognize only drugs and AEs relevant to the DPCs in Table 3. We empirically set $minsup = 2$, $corrLB = 1$ and $minprob = 10^{-10}$.

We validate the performance of our method using three well-known metrics [22]. The first metric, *Precision*, is the fraction of detected candidates that are known DPCs. The second metric, *Recall*, is the fraction of all known DPCs that are detected. The last metric, F_1 *score*, is the harmonic mean of *Precision* and *Recall*:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

First, we examine the performance of our proposed algorithm *PDPCMine* and its alternative settings. Our purpose is to validate the need to consider the data uncertainty and domain constraints from SIDER 2 in our method. The setting *PDPCMine-without-Uncertainty* does not take into account the data uncertainty in social media. In *PDPCMine-without-Uncertainty*, drugs and AEs having positive existence probabilities are considered existent. Also, between a drug and an AE, only one order exists and the order with greater probability is chosen. The setting *PDPCMine-without-DomainConstraints* does not enforce DPCs to satisfy the domain constraints from SIDER 2.

Additionally, we compare *PDPCMine* with the state-of-the-art baseline. While none of the previous methods on detecting DPCs is directly applicable to our problem, we have created a baseline *DPC-SR* that utilizes the sequence ratio (SR) in [4] to determine whether a sequence $X = \langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle$ is indicative of DPC instead of our probability $P(X \in \text{DPC})$. SR is the state-of-the-art statistical significance measure for testing DPCs in administrative claims databases. SR is computed as the ratio between the number of transactions containing $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle$ and the number of transactions containing $\langle s_2 \rightarrow d_1 \rightarrow s_1 \rightarrow d_2 \rangle$:

$$SR(X) = \frac{\sup[\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle]}{\sup[\langle s_2 \rightarrow d_1 \rightarrow s_1 \rightarrow d_2 \rangle]} \tag{7}$$

Intuitively, SR quantifies the how much more likely s_2 occurs after than before $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle$. However, SR does not account for the data uncertainty in our problem.

Fig. 5 demonstrate the performance of *PDPCMine*, its alternative settings and the state-of-the-art baseline *DPC-SR* on validation datasets. Overall, *PDPCMine* achieves promising performance on both validation datasets, e.g., $F_1 = 0.78$ on Twitter dataset and 0.68 on Patient dataset. This indicates the feasibility of utilizing social media to detect signals of DPCs and suggests that our method can be employed as a tool for detection. Furthermore, *PDPCMine* provides the best trade-off between Precision and Recall in comparison with other algorithms. While *PDPCMine-without-Uncertainty* obtains higher Precision than *PDPCMine* (e.g., 0.74 vs. 0.67 on Patient dataset), it suffers from extremely low Recall (e.g., 0.06 vs. 0.69 on Patient dataset). Likewise, neglecting data uncertainty, the state-of-the-art baseline *DPC-SR* demonstrates similar performance pattern as *PDPCMine-without-Uncertainty* (e.g., Precision = 0.81 and Recall = 0.06 on Patient dataset). While being able to identify many more known DPCs improves the recall significantly, detecting additional unknown DPCs decreases the precision. Specifically, in the Patient dataset, *PDPCMine* detects many more known DPCs (279 vs. 25 DPCs) as well as unknown DPCs (172 vs. 6 DPCs) than *DPC-SR*, leading to lower precision yet much higher

Table 2
Summary of the datasets. The first dataset is collected from Twitter while the second dataset is collected from the Patient health forum.

Dataset	#Posts	#Users/transactions	#Mentioned drugs	#Mentioned AEs	Time span
Twitter	426,217	167,776	877	1673	November 2007–August 2015
Patient	194,353	18,057	603	2168	July 2005–March 2015

Table 3
Known DPCs for validation. The PCs are derived from two well-known generic PCs ⟨NSAIDs → hypertension → Antihypertensives⟩ and ⟨ACE Inhibitors → cough → Cough Suppressants⟩. The drugs and AEs in the PCs are selected to fit the datasets. For each PC $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rangle$, *Drugs.com* is employed to obtain the known AEs s_2 of d_2 or d_1–d_2 drug interaction.

PC			Known AE
d_1	s_1	d_2	s_2
NSAIDs: • Naproxen • Ibuprofen	Hypertension	*Antihypertensives:* – Losartan – Ramipril – Atenolol – Lisinopril – Amlodipine	Tinnitus, gastritis, urticaria, tachycardia, vomiting, rash, numbness, insomnia, psoriasis, chest pain, gastric reflux, renal failure, vertigo, tremor, nausea, arthralgia, edema, cough, weight gain, difficult breathing, constipation, pruritus, alopecia, tingling, drowsiness, impotence, panic attacks, stomach disorder, cramps, liver failure, seizures, dizziness, mental depression, impaired memory
ACE inhibitors: • Ramipril • Lisinopril	Cough	*Cough suppressants:* – Codeine	Seizures, numbness, tingling, pruritus, mental depression, cramps, weight loss, nausea, hypersensitivity, vomiting, stomach disorder, tremor, urticaria, dizziness, gastric reflux, rash, insomnia, edema, hyperglycemia, impaired memory, anxiety, vertigo, asthma, tinnitus, constipation, impotence, chest pain, tachycardia, panic attacks

recall. On the other hand, in comparison with *PDPCMine*, *PDPCMine-without-DomainConstraints* achieves comparable Recall (e.g., 0.89 on Twitter dataset), yet much lower Precision (e.g., 0.01 vs. 0.7 on Twitter dataset). As a consequence, the validation results also demonstrate that incorporating data uncertainty and domain constraints improves the detection efficacy.

3.3. Detection of unknown potential detrimental prescribing cascades

We apply our method to two datasets described in Table 2. We empirically set *minsup* = 2, *corrLB* = 1 and *minprob* = 0.5. Since a large number of DPCs might be found, we evaluate the top K candidates for $K = 10, 20, \ldots, 50$. As a result, we use two well-known metrics for top K ranked results [22]. The first metric, *Precision@K*, is the fraction of top K candidates that are potential. We employ the medical knowledge bases to evaluate whether a candidate is a potential DPC. $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle$ is a potential DPC if s_1 is a known AE of d_1, d_2 is a known treatment of s_1, and s_2 is a known AE of either d_2 or the drug–drug interaction between d_1 and d_2 according to *Drugs.com* (http://www.drugs.com/) or *MedlinePlus* (https://www.nlm.nih.gov/medlineplus/druginformation.html) or *RxList* (http://www.rxlist.com/) or *WebMD* (http://www.webmd.com/drugs/). However, this does not mean other DPCs that are not potential DPCs according to the knowledge bases can not be true DPCs. This is because the knowledge bases are not complete in practice, s_1 may be an unknown AE of d_1, d_2 may be an off-label treatment of s_1, and s_2 may be an unknown AE of d_2 or d_1–d_2 interaction. In other words, detected candidates can be true DPCs and the term "potential" in this context is for the consistency with known DPCs only. Let $h_i = 1$ indicate whether the ith candidate is a potential DPC and 0 otherwise. The second metric, *NDCG@K* (Normalized Discounted Cumulative Gain), takes into account how potential DPCs are ranked against non-potential DPCs.

$$NDCG@K = \frac{1}{Z} \times \sum_{i=1}^{K} \frac{2^{h_i} - 1}{\log_2(i+1)} \tag{8}$$

where Z is the normalizing constants selected so that $NDCG@K = 1$ when the ranking is perfect, i.e., all the potential DPCs ranked higher than non-potential DPCs. The key difference between *Precision@K* and *NDCG@K* is that the latter takes into account the ranking while the former does not. Particularly, *NDCG@K* penalizes potential DPCs ranked lower in the list by reducing the score logarithmically proportional to the ranks of these DPCs. *NDCG@K* has been also used to evaluate the ranking of unknown ADRs [23]. Note that *Recall@K* is not applicable in this case since it is intractable to know all the potential DPCs in a real-world dataset.

Our method detects 88 candidate DPCs from the Twitter dataset and 1305 candidate DPCs from the Patient dataset. Table 4 presents the top 5 potential DPCs detected from each dataset with their probabilities. The complete table of detected potential DPCs can be found in Tables S1 and S2 of the Supplementary Information. In both datasets, we achieve encouraging performance, i.e., Precision@50 = 0.72 and NDCG@50 = 0.95 on Twitter dataset, and Precision@50 = 0.86 and NDCG@50 = 0.98 on Patient dataset.

Fig. 6 shows the performance of *PDPCMine*, its alternative settings and the state-of-the-art baseline *DPC-SR* on the two datasets. We plot the values of Precision@K and NDCG@K as the number of top candidates K varies from 10 to 50. Overall, *PDPCMine* significantly outperforms other algorithms in all metrics. For Precision@50, we perform better than the runner-up by 227% (Twitter dataset) and 53.57% (Patient dataset). For NDCG@50, we improve by 30.14% (Twitter dataset) and 12.64% (Patient dataset). The statistics once again demonstrate the necessity of taking into account data

Table 4
Top 5 potential DPCs in Twitter and Patient datasets. Five DPCs with highest probabilities that are evaluated as potential in each dataset are presented.

Prescribing cascade			Associated AE	Probability	Dataset
d_1	s_1	d_2	s_2		
Metoprolol	Stroke	Simvastatin	Arrhythmia	0.654	
Venlafaxine	Stroke	Simvastatin	Hemorrhage	0.651	
Celecoxib	Hypertension	Lisinopril	Depression	0.642	Twitter
Venlafaxine	Arthritis	Meloxicam	Hypertension	0.639	
Trazodone	Hypertension	Prazosin	Anxiety	0.627	
Doxorubicin	Pulmonary embolism	Warfarin	Myalgia	1	
Lisinopril	Pulmonary embolism	Warfarin	Myalgia	1	
Ciprofloxacin	Pulmonary embolism	Warfarin	Myalgia	1	Patient
Citalopram	Pulmonary embolism	Warfarin	Myalgia	1	
Clopidogrel	Pulmonary embolism	Warfarin	Myalgia	1	

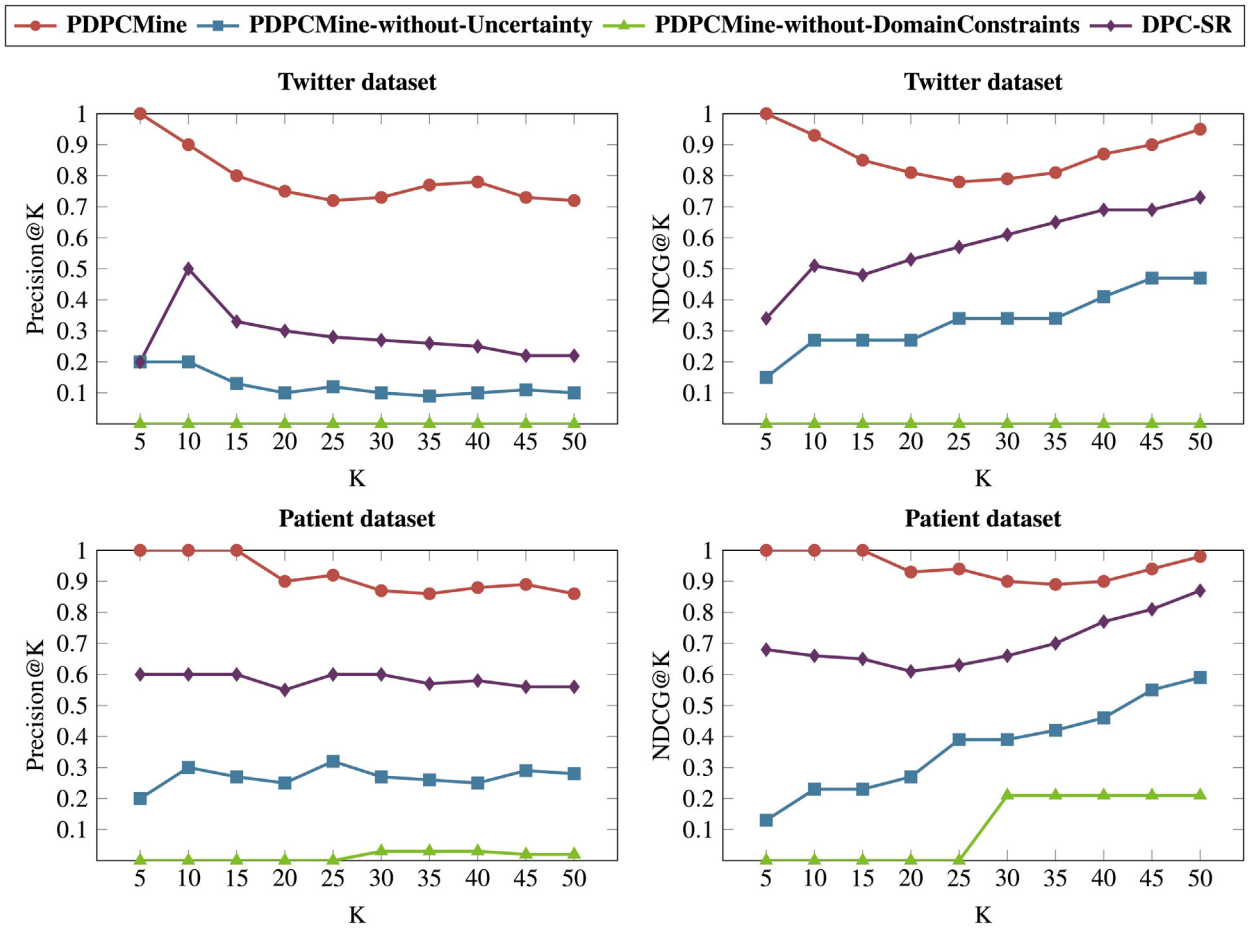

Fig. 6. Performance of our method on detecting unknown DPCs. The performance of the proposed algorithm *PDPCMine*, its alternative settings and the state-of-the-art baseline *DPC-SR* in Precision@K and NDCG@K are plotted. *PDPCMine-without-Uncertainty* does not take into account the data uncertainty while *PDPCMine-without-DomainConstraints* does not require the domain constraints in SIDER 2 to be satisfied. *DPC-SR* utilizes $SR(X)$ as the statistical significance measure to test whether X is a DPC.

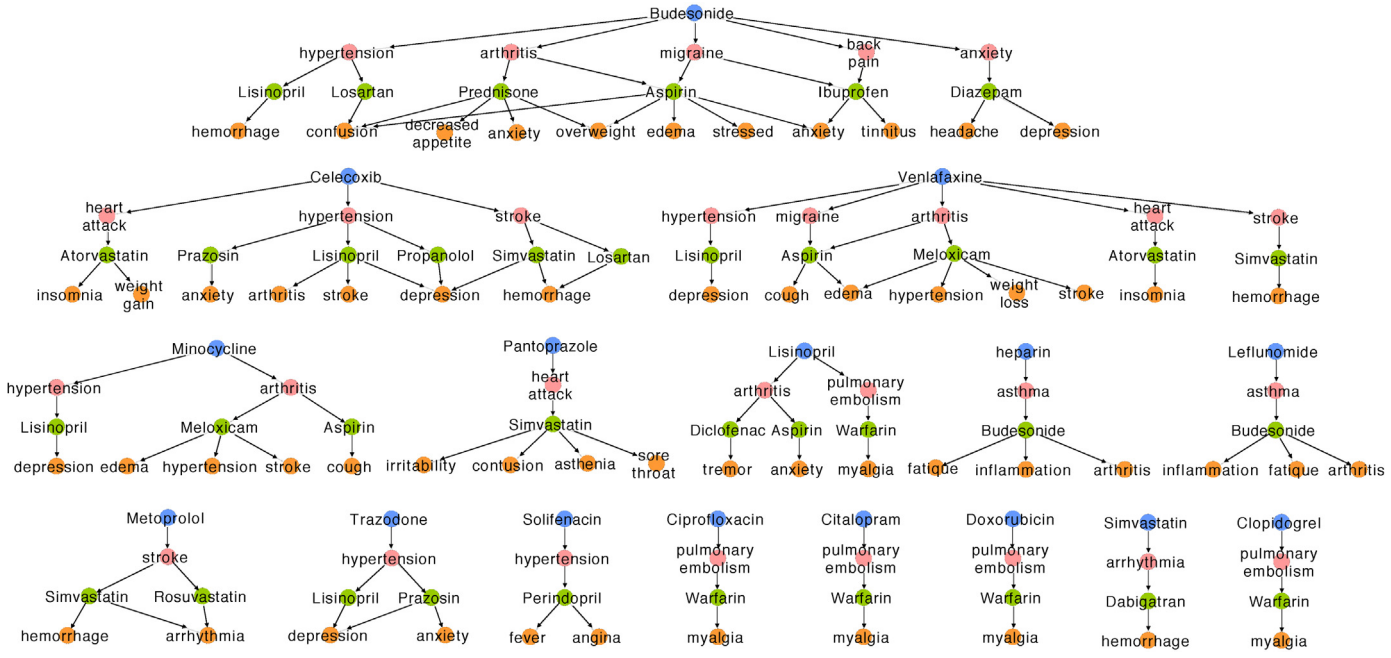

Fig. 7. Top 60 potential DPCs detected from both datasets grouped by d_1. 30 potential DPCs with highest probabilities in each dataset is extracted. Then they are combined and grouped by d_1.

Table 5
Four largest groups of potential DPCs by s_1. 50 potential DPCs with highest probabilities in each dataset is extracted. Then they are combined and grouped by s_1 and drug class. Four groups with the most DPCs are presented.

Prescribing cascade			Associated AE
d_1	s_1	d_2	s_2
NSAIDs (Celecoxib)	Hypertension	ACE Inhibitors (Lisinopril)	Depression, arthritis, stroke
		Beta-blockers (Propanolol)	Depression
Antidepressants (Trazodone)		Alpha-blockers (Prazosin)	Depression, anxiety
		ACE Inhibitors (Lisinopril)	Depression
Antimuscarinics (Solifenacin)		ACE Inhibitors (Perindopril)	Angina, fever
Antibiotics (Minocycline, Linezolid)		ACE Inhibitors (Lisinopril)	Depression
		ARBs (Losartan)	Confusion
Corticosteroids (Budesonide)		ACE Inhibitors (Lisinopril)	Hemorrhage
Corticosteroids (Budesonide)	Arthritis	Corticosteroids (Prednisone)	Confusion, overweight, sciatica
		NSAIDs (Aspirin)	Overweight, anxiety, vomiting, confusion
Antidepressants (Venlafaxine)		NSAIDs (Meloxicam, Aspirin)	Hypertension, edema, stroke, cough, weight loss, anxiety, depression
ACE Inhibitors (Lisinopril)		NSAIDs (Aspirin, Diclofenac)	Tremor, anxiety
Antibiotics (Minocycline)		NSAIDs (Meloxicam, Aspirin)	Hypertension, stroke, cough, edema
Beta-blockers (Metoprolol)	Stroke	Statins (Simvastatin, Rosuvastatin)	Hemorrhage, arrhythmia
		Statins (Simvastatin)	Hemorrhage, depression
NSAIDs (Celecoxib)		ARBs (Losartan)	Hemorrhage
Antidepressants (Venlafaxine)		Statins (Simvastatin, Rosuvastatin)	Hemorrhage
PPIs (Pantoprazole)	Heart attack	Statins (Simvastatin)	Irritability, asthenia, sore throat, contusion
NSAIDs (Celecoxib)		Statins (Atorvastatin)	Weight gain, insomnia
Antidepressants (Venlafaxine)			Insomnia

uncertainty and domain constraints. In addition, in *PDPCMine*, as K increases from 5 to 20, Precision@K and NDCG@K decrease consistently in both datasets. This means that potential DPCs are usually ranked higher than non-potential DPCs. As a result, pharmacists may not need to investigate all the candidates but only those with high probabilities (e.g., in top 20).

The top detected potential DPCs can be organized more compactly in various ways to facilitate further investigation. Fig. 7 (drawn with Cytoscape [33]) groups the top 60 potential DPCs detected from both datasets by d_1. In Fig. 7, each group of potential DPCs is represented as a directed graph, in which the nodes describe drugs and AEs while the directed edges specify the orders of drugs and AEs. The directed graph is presented as a hierarchy, in which d_1, s_1, d_2, s_2 in each potential DPC are ordered from top to bottom, e.g., d_1 = "Budesonide" appears at a higher position than s_1 = "hypertension" in the first group. Also, the groups are ordered top-to-bottom and left-to-right by the number of potential DPCs, e.g., d_1 = "Budesonide" is associated with the most number of potential DPCs, followed by "Celecoxib", "Venlafaxine", etc. This representation may help enhance the caution of drug uses or prescriptions. For instance, some drugs like "Lisinopril" or "Diazepam" should not be used or prescribed in patients under "Budesonide" since those drugs might treat the AEs caused by "Budesonide" and the treatments are likely to introduce additional AEs such as "hemorrhage", "depression", etc. In the drug interaction point of view, some pairs of drugs such as d_1 = "Venlafaxine" and d_2 = "Aspirin" might potentially interact with each other and produce AEs such as s_2 = "coughing up blood", etc. Another way to is to group the potential DPCs by s_1 as in the previous works [15,27]. Table 5 presents the four largest groups, the potential DPCs in each of which share the same s_1. This representation may signal that some medical conditions such as s_1 = "hypertension" might be an AE of a drug rather than being a new disease symptom, i.e., additional treatments might become harmful. In addition, some drugs in potential DPCs share the same classes, e.g., "Meloxicam" and "Aspirin" are both non-steroidal anti-inflammatory drugs (NSAIDs). The drugs in a same class may share essential properties and thus are grouped together for better interpretation and more efficient investigation [15,27]. Based on such signals of DPCs, the pharmacists may take further investigation to verify them and issue warnings or contraindications.

Furthermore, we utilize the medical knowledge bases to examine the potential cause of s_2 in each DPC. For each DPC, we are particularly interested in three factors: (1) whether s_2 is a known AE of d_2, (2) whether d_1 and d_2 are known to interact, and (3) whether s_2 is a known AE of d_1–d_2 interaction. Note that if d_1 and d_2 are not known to interact, s_2 cannot be a known AE of d_1–d_2 interaction. Fig. 8 shows the proportions of s_2 due to various factors according to the knowledge bases and corresponding examples from the detection of two datasets. In Fig. 8, five different categories (as in the figure legend) corresponding to different satisfactions of three factors and respective proportions of DPCs are plotted. In both datasets, it can be observed that in the majority of cases (72% in Twitter dataset and 86% in Patient dataset), s_2 is a known AE of d_2. In approximately half of the cases where s_2 is a known AE of d_2, there is a known interaction between d_1 and d_2 (53% in Twitter dataset and 44% in Patient dataset). When s_2 is a known AE of d_2 and d_1 and d_2 are known to interact, s_2 is a known AE of d_1–d_2 interaction in minority cases (5.3% in Twitter dataset and 21.1%). On the other hand, none of the cases where s_2 is a known AE of d_1–d_2 interaction but not d_2 has been detected. Fig. 8 also presents three examples of detected DPCs corresponding to different scenarios. In the first DPC, "depression" is a known AE of "Lisinopril" while "Minocycline" is not known to interact with "Lisinopril". In the second DPC, "coughing up blood" is a known AE of "Aspirin" as well as a known AE of the interaction between "Venlafaxine" and "Aspirin". In the last DPC, "hypertension" is a known AE of "Meloxicam" but not of "Venlafaxine-Meloxicam" interaction although such interaction is known to exist.

3.4. Running time scalability study

We conduct our experiments on a computer with 4 AMD 6238 12-core 2.6Ghz CPUs of 128GB memory in total. Our method processes the Twitter dataset in 30 min and the Patient dataset in 40 min. On the other hand, it takes more than one day to handle each of the dataset using a general sequence mining algorithm for uncertain data [42]. Furthermore, we examine the running time

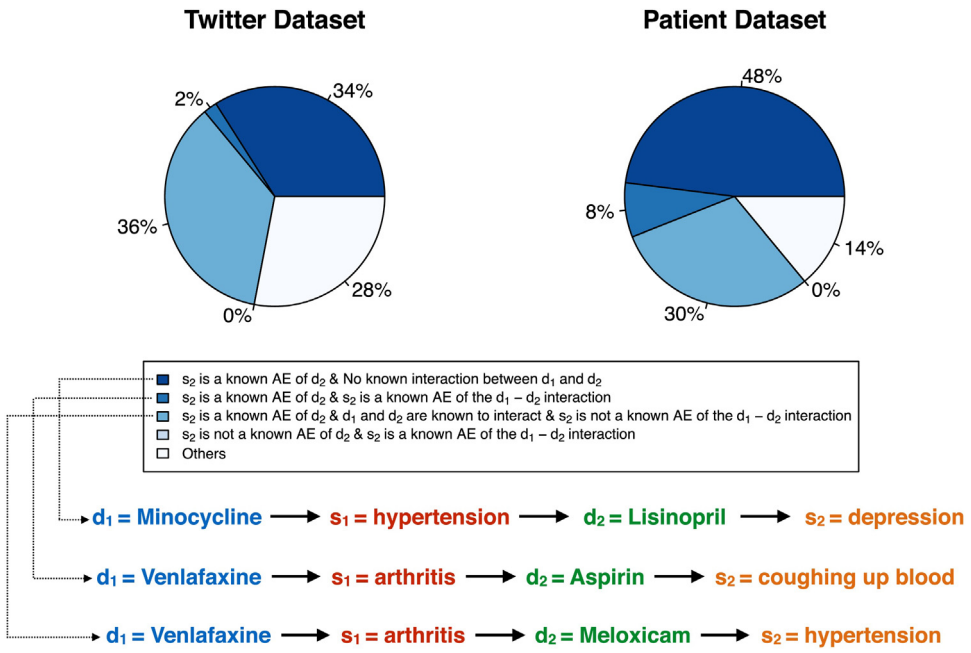

Fig. 8. Proportions of s_2 due to various factors according to medical knowledge bases. The three factors of interest are: (1) whether s_2 is a known AE of d_2, (2) whether d_1 and d_2 are known to interact, and (3) whether s_2 is a known AE of d_1–d_2 interaction. Five categories corresponding to different satisfactions of factors according to the knowledge bases are derived. Then the proportions of DPCs falling into five categories are plotted. Examples of DPCs in the first three categories are presented.

scalability of our method when the thresholds *minsup*, *corrLB* and *minprob* and dataset sizes vary. We sample 10, 100, 1000, 10,000, 100,000 users from the Twitter dataset and 10, 100, 1000, 10,000 users from the Patient dataset.

Fig. 9 presents the running times of our method with various thresholds and dataset sizes. The running times of the method indicate reasonable trends. As the thresholds increase, the running times drop consistently, since more sub-sequences can be pruned at earlier stages of the algorithm. Likewise, the running times rise when the dataset size increases. In addition, the results demonstrate that our method is scalable to large datasets. In fact, the running time of our method grows approximately linear in terms of the dataset size. Furthermore, the results show that the detection is sensitive to the frequency of the candidate sequences. It can be observed that the running time varies more steeply as *minsup* changes than *corrLB* and *minprob*.

4. Discussion

The results demonstrate that PCs might be associated with many AEs, some of which can be very harmful, e.g., the risk of "stroke" associated with ⟨Celecoxib → hypertension → Lisinopril⟩. In addition, in some potential DPCs, associated AEs s_2 may even be more life-threatening than s_1, e.g., ⟨Budesonide → arthritis → Meloxicam → stroke⟩. Hence, the resulting harm may outweigh the benefits of the additional treatments. PCs may potentially lead to *polypharmacy*, the concurrent use of five or more drugs. Inappropriate polypharmacy, however,

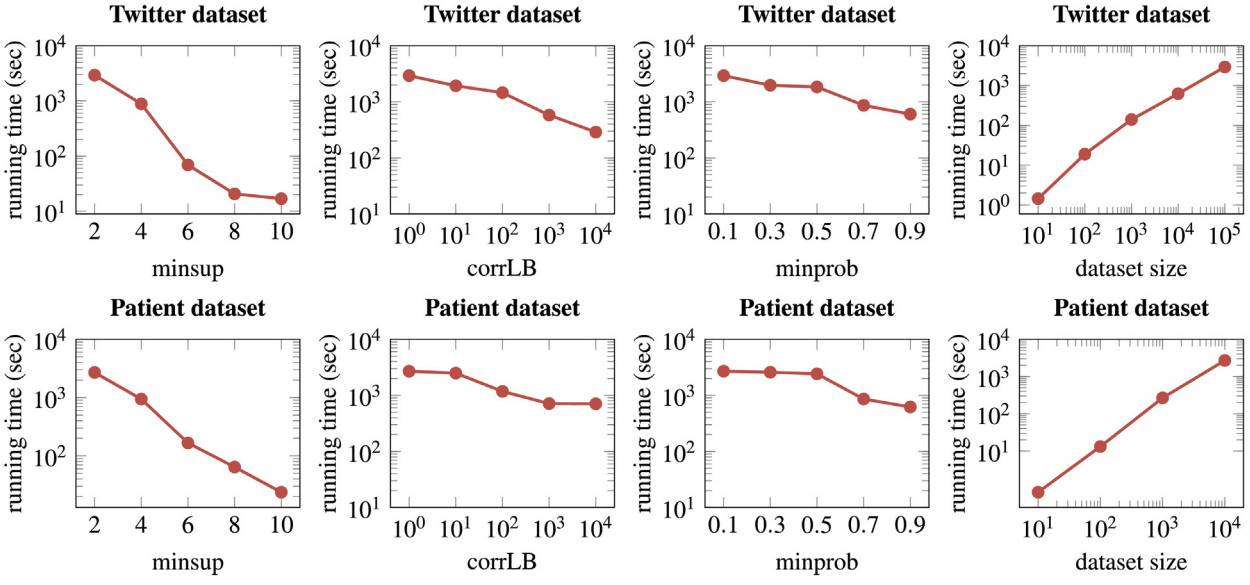

Fig. 9. Running time scalability of our method. The running times of *PDPCMine* with different values of *minsup*, *corrLB*, *minprob* and dataset size are plotted.

increases the risk of drug AEs significantly and has been imposing a substantial burden of death and hospitalization [32]. As a result, PCs should generally be avoided as patients are likely to suffer more as they receive more treatments.

Upon investigating the results, we observe some information errors in SIDER 2 due to automatic data extraction. For example, "drug interaction" is listed as an AE with ID C0687133, while "Losartan" (3961) is listed as a possible treatment of "cough" (C0010200) ("cough" is instead a known AE of "Losartan"). Therefore, as a preprocessing step, we filter such noises from SIDER 2 before utilizing it in the detection. Nevertheless, the filtering may not be exhaustive since the it is solely based on the investigation of the results. Besides, some similar AEs such as "obesity" (C0028754) and "overweight" (C0497406) are listed as different AEs in SIDER 2, which are difficult to be matched by our method and lead to redundant signals.

Similar to the results reported by Golder et al. [11], we find that serious AEs such as "liver failure", "renal failure", "stroke", etc. are rarely mentioned as personal experience in social media. However, DPCs involving such rare AEs as s_2, e.g., ⟨Naproxen → hypertension → Ramipril → liver failure⟩ are still detectable by our method.

While the results demonstrates the feasibility of detecting DPCs from social media, our work can be further improved in various dimensions. First, the signals generated by our framework can be cross-checked against reliable medical database records such as administrative claims database [4] or electronic health records [2] to further refine the hypotheses and enhance their plausibility. In fact, some DPCs may be complicated by treatment recommendations or confounding bias. For example, in the potential DPC ⟨Celecoxib → hypertension → Lisinopril → stroke⟩, while "stroke" is a known AE of "Lisinopril", people with "hypertension" are at increased risk of "stroke", i.e., the causal relationship may be unclear. Therefore, more rigorous causality techniques should be employed to test the causal relationship between the PCs and their associated AEs in reliable medical database records.

Second, signals of DPCs detected from different social media sources can be combined to improve the detection accuracy. In fact, Li et al. [20] has attempted to combine signals of ADRs from spontaneous reporting systems and observational healthcare data and shown that the ADR detection is significantly more accurate. Some sources, however, might be more reliable than the others in providing signals. For instance, health forums might intuitively be more reliable than Twitter since they contain more health-related content and the users' language is generally less informal. As a result, the credibility of different sources need to be taken into account in the signal combination for better accuracy.

In addition, taking into account whether mentions of drugs or AEs that are related to real personal experience may improve the estimation of data uncertainty. For instance, the post "Being an Arsenal fan will give you high blood pressure but I'm proud of my boys" may not describe actual personal experience of "hypertension". Suppose we have sufficient annotated data regarding whether a post describes personal consumption of drug or experience of AE. We can build a classifier to output the likelihood that a user consumes a drug or experiences an AE given a post by adapting the technique in [30]. Then given all the posts of a user, we can again utilize logistic function to estimate the existence uncertainty. The estimation of existence uncertainty may become more accurate since it takes into account the "existence uncertainty per post".

More generally, we can enhance the detection accuracy by exploiting the context surrounding the mentions of drugs and AEs in social media posts. For example, the post "Now on Ramipril for my blood pressure" does not contain a complete mention of the AE "high blood pressure" but the context "Ramipril" is an indicator. Also, our current method utilizes the "can be used to treat" relationship from SIDER 2 to estimate the "is used to treat" relationship. It would be useful to also identify the actual "is used to treat" relationship by leveraging the context as it may reveal off-label drug uses.

Besides, since the lexicon of each AE may not be sufficient, our dictionary lookup to recognize AEs may lead to lower recall in mining DPCs. For instance, "gastric ulcer" (C0038358) might be referred to as "stomach ulcer" or "stomach bleeding" in social media posts but "stomach bleeding" is not in the lexicon. Extending the dictionary would improve the performance, yet require laborious regular update, i.e., not scalable in terms of time and efforts. As a result, in future work, we plan to extend the data-centric extraction approach in [25] to improve the recall of our framework while enable the mapping of drug and AE mentions to corresponding IDs in SIDER 2.

Furthermore, some other enhancements may be useful to consider. While $\langle d_1 \rightarrow s_1 \rightarrow d_2 \rightarrow s_2 \rangle$ is the most common form of a DPC, it can be generalized to a sequence of arbitrary size and partial ordering. From such generalized sequences, sub-sequences of interest can be extracted for further investigation, e.g., DPCs as in our current definition, ADRs, drug interactions, off-label drug uses, etc. Also, the data uncertainty might be better estimated by leveraging the frequency information of ADRs and treatments in SIDER 2. Lastly, our framework can be extended towards detecting DPCs from social media streaming data that are generated in real-time. Handling streaming data would induce additional challenges regarding the running time and memory scalability.

Lastly, it is worth noting that Sarker et al. [29] presented an end-to-end framework for detecting signals of ADRs. There are similarities and differences between our framework and the framework in [29]. For similarities, both frameworks consist of the extraction and statistical analysis components. For differences, firstly, the framework in [29] covers more components such as data collection and dangerous drug flagging. Also, instead of classifying whether a post mentions personal experience of ADRs, our framework quantifies the likelihood of user taking a drug or suffering from an AE for later probabilistic data mining. While our statistical analysis component aims at mining sequences indicative of DPCs, Sarker et al. [29] focuses on identifying ADRs. An interesting future work is to integrate two frameworks for detecting general sequences signalling drug safety issues.

5. Conclusion

PCs can be associated with AEs, which are costly in terms of harm and expense. Therefore, timely detection of DPCs is necessary to reduce harm and improve health. Nevertheless, the existing detection has been conducted on a case-by-case basis and relying on limited case reports or restrictive data sources. In this paper, we investigate the feasibility of detecting DPCs from social media, an open data source containing health-related discussions. We develop a framework to mine sequences of drugs and AEs that signal DPCs, which tackles the data uncertainty and data rarity in social media. We conduct experiments on real-world datasets collected from Twitter (general social media) and Patient (health forum). The validation shows that our framework is able to detect DPCs that have been reported in previous works. In addition, our framework detects and prioritizes unknown potential DPCs with encouraging performance. Furthermore, we show that our framework is efficient and scalable to large datasets. Our study generates hypotheses to reduce pharmacists' guesswork in identifying DPCs.

Acknowledgements

This research has been supported by the Australian Research Council (ARC) DP130104090 and the National Health and Medical Research Council (NHMRC) GNT 1040938.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.artmed.2016.06.002.

References

[1] Avorn J, Gurwitz JH, Bohn RL, Mogun H, Monane M, Walker A. Increased incidence of levodopa therapy following metoclopramide use. J Am Med Assoc (JAMA) 1995;274(22):1780–2.
[2] Banda JM, Callahan A, Winnenburg R, Strasberg HR, Cami A, Reis BY, et al. Feasibility of prioritizing drug–drug-event associations found in electronic health records. Drug Saf 2016;39(1):45–57.
[3] Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Peckham J, editor. Proceedings of the 1997 ACM SIGMOD international conference on management of data, vol. 26. New York, NY, USA: ACM; 1997. p. 255–64.
[4] Caughey GE, Roughead EE, Pratt N, Shakib S, Vitry AI, Gilbert AL. Increased risk of hip fracture in the elderly associated with prochlorperazine: is a prescribing cascade contributing? Pharmacoepidemiol Drug Saf 2010;19(9):977–82.
[5] Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Ann Symp Proc 2011;2011:217–26.
[6] Feldman R, Netzer O, Peretz A, Rosenfeld B. Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. In: Cao L, Zhang C, Joachims T, Webb GI, Margineantu DD, Williams G, editors. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM; 2015. p. 1779–88.
[7] Freedman DA. Statistical models: theory and practice. New York, NY, USA: Cambridge University Press; 2009.
[8] Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in Twitter. Drug Saf 2014;37(5):343–50.
[9] Gill SS, Mamdani M, Naglie G, Streiner DL, Bronskill SE, Kopp A, et al. A prescribing cascade involving cholinesterase inhibitors and anticholinergic drugs. Arch Intern Med 2005;165(7):808–13.
[10] Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, et al. Part-of-speech tagging for Twitter: annotation, features, and experiments. In: Lin D, Matsumoto Y, Mihalcea R, editors. Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: short papers, vol. 2. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 42–7.
[11] Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. Br J Clin Pharmacol 2015;80(4):878–88.
[12] Gurwitz JH, Avorn J, Bohn RL, Glynn RJ, Monane M, Mogun H. Initiation of antihypertensive treatment during nonsteroidal anti-inflammatory drug therapy. J Am Med Assoc (JAMA) 1994;272(10):781–6.
[13] Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012;91(6):1010–21.
[14] Hazell L, Shakir SA. Under-reporting of adverse drug reactions: a systematic review. Drug Saf 2006;29(5):385–96.
[15] Kalisch LM, Caughey GE, Roughead EE, Gilbert AL. The prescribing cascade. Aust Prescr 2011;34(6):162–6.
[16] Kalisch Ellett LM, Pratt NL, Barratt JD, Rowett D, Roughead EE. Risk of medication-associated initiation of oxybutynin in elderly men and women. J Am Geriatr Soc 2014;62(4):690–5.
[17] Kuehn BM. Twitter streams fuel big data approaches to health forecasting. J Am Med Assoc (JAMA) 2015;314(19):2010–2.
[18] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2015, gkv1075.
[19] Le Cam L. An approximation theorem for the Poisson Binomial distribution. Pac J Math 1960;10(4):1181–97.
[20] Li Y, Ryan PB, Wei Y, Friedman C. A method to combine signals from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions. Drug Saf 2015;38(10):895–908.
[21] Lin S-F, Xiao K-T, Huang Y-T, Chiu C-C, Soo V-W. Analysis of adverse drug reactions using drug and drug target interactions and graph-based methods. Artif Intell Med 2010;48(2):161–6.
[22] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval, vol. 1. New York, NY, USA: Cambridge University Press; 2008.
[23] Mukherjee S, Weikum G, Danescu-Niculescu-Mizil C. People on drugs: credibility of user statements in health communities. In: Macskassy SA, Perlich C, Leskovec J, Wang W, Ghani R, editors. Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM; 2014. p. 65–74.
[24] Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In: AMIA annual symposium proceedings 2011. 2011. p. 1019–26.
[25] Nikfarjam A, Sarker A, OConnor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc (JAMIA) 2015;22(3):671–81.
[26] OConnor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G. Pharmacovigilance on Twitter? Mining tweets for adverse drug reactions. In: AMIA annual symposium proceedings 2014. 2014. p. 924–33.
[27] Rochon PA, Gurwitz JH. Optimising drug treatment for elderly people: the prescribing cascade. Br Med J (BMJ) 1997;31(7115):1096.
[28] Rosenberg J, Rochon A, Gill PSS. Unveiling a prescribing cascade in an older man. J Am Geriatr Soc 2014;62(3):580–1.
[29] Sarker A, Ginn R, Nikfarjam A, OConnor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: a review. J Biomed Inform 2015;54:202–12.
[30] Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015;53:196–207.
[31] Sarker A, OConnor K, Ginn R, Scotch M, Smith K, Malone D, et al. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. Drug Saf 2016:1–10.
[32] Scott IA, Hilmer SN, Reeve E, Potter K, Le Couteur D, Rigby D, et al. Reducing inappropriate polypharmacy: the process of deprescribing. JAMA Intern Med 2015;175(5):827–34.
[33] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504.
[34] Srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements. In: Apers PMG, Bouzeghoub M, Gardarin G, editors. Proceedings of the 5th international conference on extending database technology: advances in database technology, vol. 1057. London, UK: Springer-Verlag; 1996. p. 3–17.
[35] Sun L, Cheng R, Cheung DW, Cheng J. Mining uncertain data with probabilistic guarantees. In: Rao B, Krishnapuram B, Tomkins A, Yang Q, editors. Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'10. New York, NY, USA: ACM; 2010. p. 273–82.
[36] Tan P-N, Kumar V, Srivastava J. Selecting the right interestingness measure for association patterns. In: Hand D, Keim D, Ng R, editors. Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM; 2002. p. 32–41.
[37] Tatonetti NP, Patrick PY, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012;4(125), 125ra31.
[38] Tong Y, Chen L, Cheng Y, Yu PS. Mining frequent itemsets over uncertain databases. Proc Very Large Database Endow (PVLDB) 2012;5(11):1650–61.
[39] Velardi P, Stilo G, Tozzi AE, Gesualdo F. Twitter mining for fine-grained syndromic surveillance. Artif Intell Med 2014;61(3):153–63.
[40] Yang CC, Yang H, Jiang L. Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. ACM Trans Manage Inf Syst (TMIS) 2014;5(1):2.
[41] Yates A, Goharian N, Frieder O. Extracting adverse drug reactions from social media. In: Bonet B, Koenig S, editors. AAAI conference on artificial intelligence. Menlo Park, California: AAAI Press; 2015.
[42] Zhao Z, Yan D, Ng W. Mining probabilistically frequent sequential patterns in large uncertain databases. IEEE Trans Knowl Data Eng (TKDE) 2014;26(5):1171–84.