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/

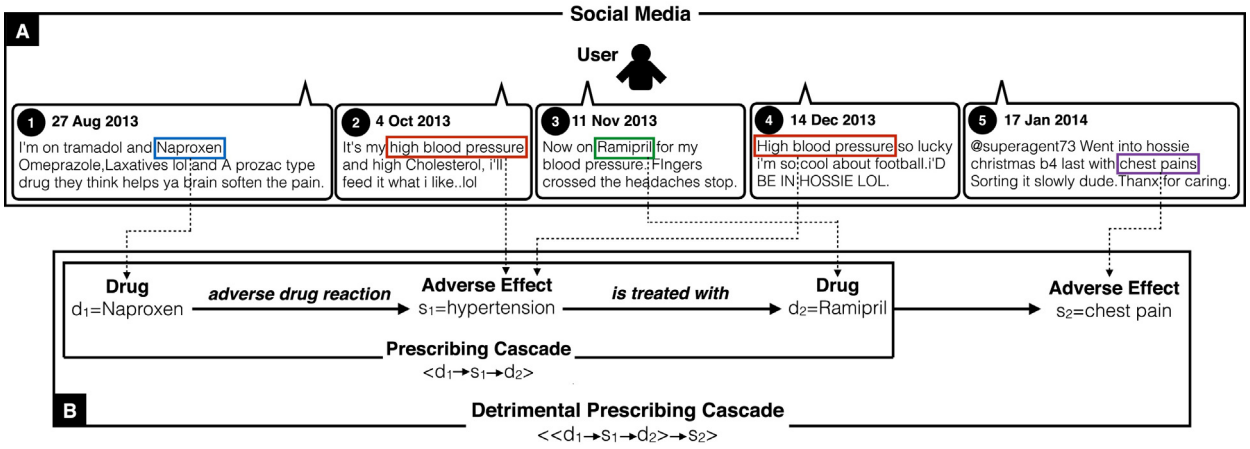

Fig. 1. A signal of DPC on social media. Part B shows an example of a known DPC. Part A presents a user tweeting about the drugs and AEs of the known DPC.

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