Journal of Biomedical Informatics 59 (2016) 42–48

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Early identification of adverse drug reactions from search log data

Ryen W. White [a,*], Sheng Wang [b,1], Apurv Pant [c], Rave Harpaz [d], Pushpraj Shukla [c], Walter Sun [c], William DuMouchel [d], Eric Horvitz [a]

[a] Microsoft Research, Redmond, WA, United States
[b] University of Illinois at Urbana Champaign, Urbana, IL, United States
[c] Bing Predicts Team, Microsoft Bing, Bellevue, WA, United States
[d] Oracle Health Sciences, Bedford, MA, United States

ARTICLE INFO

Article history:
Received 21 August 2015
Revised 7 November 2015
Accepted 12 November 2015
Available online 29 November 2015

Keywords:
Pharmacovigilance
Search log analysis
Adverse drug reactions

ABSTRACT

The timely and accurate identification of adverse drug reactions (ADRs) following drug approval is a persistent and serious public health challenge. Aggregated data drawn from anonymized logs of Web searchers has been shown to be a useful source of evidence for detecting ADRs. However, prior studies have been based on the analysis of established ADRs, the existence of which may already be known publically. Awareness of these ADRs can inject existing knowledge about the known ADRs into online content and online behavior, and thus raise questions about the ability of the behavioral log-based methods to detect new ADRs. In contrast to previous studies, we investigate the use of search logs for the early detection of known ADRs. We use a large set of recently labeled ADRs and negative controls to evaluate the ability of search logs to accurately detect ADRs in advance of their publication. We leverage the Internet Archive to estimate when evidence of an ADR first appeared in the public domain and adjust the index date in a backdated analysis. Our results demonstrate how search logs can be used to detect new ADRs, the central challenge in pharmacovigilance.

1. Introduction

Adverse drug reactions (ADRs) are the fourth leading cause of death in the United States, ahead of pulmonary disease, diabetes, infection with human immunodeficiency virus, and automobile accidents [1–4]. *Pharmacovigilance* centers on the assessment, prevention, monitoring, and detection of ADRs in the post-marketing period (i.e., after the medication has been released to market and is being used by patients). Mining evidence of ADRs from various data sources to identify previously unknown ADRs is a central goal of pharmacovigilance [4].

The United States Food and Drug Administration (FDA) receives information about post-marketing ADRs via spontaneous reports submitted by healthcare professionals. The FDA's Adverse Event Reporting System (FAERS) pools these reports and these data are routinely analyzed to identify signals of new ADRs [5,6]. Significant evidence of ADRs drawn from spontaneous reports in FAERS may lead to deeper investigations followed by regulatory actions such as a drug withdrawal from the market, the issuance of public warnings, and/or enforcement of changes to the label that appears on the packaging (i.e., label changes). Beyond spontaneous reports, other data have also been employed to develop more capable and robust systems for pharmacovigilance purposes. These additional sources include electronic health records and medical insurance claims [7–9], findings published in the biomedical literature [10–12], as well as other sources such as chemical and biological knowledge bases [13,14]. Pharmaceutical companies also perform post-marketing safety surveillance to understand the long-term effects of their products and to discover less frequent ADRs that are not identified in clinical trials.

Non-traditional sources such as logs of search engine activity or social media (e.g., postings on online forums and social networks) contain evidence of health-related issues [15] and may provide new insights in support of early detection of ADRs. These sources are currently being studied as additional inputs for signal detection [4,16–18]. People have been shown to consistently search the Internet for health-related matters. A 2013 study by the Pew Research Center found that 72% of Internet users claimed to search online for health information and that 8 in 10 online health inquiries start at a search engine [19]. Search logs are used in the

* Corresponding author.
E-mail address: ryenw@microsoft.com (R.W. White).
[1] Work done while employed as an intern at Microsoft.

http://dx.doi.org/10.1016/j.jbi.2015.11.005
1532-0464/

Google Flu Trends project, demonstrating that statistics of influenza-related search terms recorded by search engines can be used to provide fast-paced updates on rates of influenza [20]. Recently, search logs have been shown to be effective in identifying ADRs and interactions between medications [21–23], as well as a complement to more traditional methods of mining ADRs based on spontaneous reporting [22].

We consider a set of recent label changes for our study of the early identification of ADRs from logs of Web search activity. Specifically, we consider the medications that are the focus of attention, the ADR added during a label change for that medication, and the date that this label change occurred (hereafter referred to as the *index date*) as ground truth data for our study of the early identification of ADRs. We use anonymized large-scale search engine query log data from consenting users of the Microsoft Bing Web search engine. Search logs may reveal concerns about observed side effects of medications in advance of traditional reporting by physicians and patients. Despite the promise of search services to provide signals about such concerns on a wide scale, analyses of aggregate signals of online human behavior in the absence of more detailed interviews pose multiple statistical challenges. For example, the frequencies of terms used in searches may be significantly influenced by media coverage [24], related pandemics, e.g., H1N1 (swine flu) [25], and changes in search engine ranking functions [26] and data capture policies.

A key challenge in assessing the power of using aggregate online behavior to detect previously unknown ADRs is accounting for the potential leak of existing ADR reports and knowledge onto the Web. ADR information may appear on Websites such as social media before the publication of FDA label changes and affect people's search actions via factors such as information cascades [27,28]. Studies to date have explored the detection of ADRs that were known at analysis time, using reference standards such as those from the Observable Medical Outcomes Partnership (OMOP) [29] and the European Union EU-ADR [8] projects, designed for the retrospective evaluation of ADR detection methods using health records. The public availability and awareness of knowledge about ADRs may affect spontaneous reporting rates for those ADRs or prescription patterns, which in turn could bias retrospective evaluations [16,30]. Interest demonstrated by users via queries to Web search engines using terms associated with ADRs may be prompted by existing online content rather than personal experiences with side effects. To be a truly useful mechanism, pharmacovigilance systems need to accurately predict emerging and unknown ADRs in advance of slower processes involving the curation of medical reports [30,31].

We created as a benchmark a time-indexed reference set of ADRs recently labeled by the FDA (and matching negative controls) [32]. We used this reference set to evaluate the ability of search logs to detect ADRs in advance of their publication by backdating the signal detection analysis to periods prior to their publication. Signals derived from Microsoft Bing search log data collected over a period of three years were used as the basis for our analysis. We combined logged data on searches and snapshots of Web page content from the Wayback Machine provided by the Internet Archive (archive.org), a non-profit organization that stores periodic snapshots of Web content. The Wayback Machine was used to assess conservatively the date of the appearance of any evidence related to knowledge or suspicions of drug–ADR associations in online content. These dates serve as index dates for backdating analyses to limit the influence of existing Web page content associated with ADRs on analyzed queries. Such dates could be earlier than the dates on which our ADRs were added to medication labels by the FDA. If so, we use those earlier dates as the index dates in our analysis.

2. Materials and methods

2.1. Search log data

We used three full years of log data collected from consenting users of the Microsoft Bing search engine during 2011–2013 as the basis for the study. The data collection and portions of the analysis was undertaken as part of the Bing Predicts project within the Microsoft Bing search engine. These logs contained users' search queries, a timestamp for when each query was issued (in the user's local timezone), and a unique identifier for the user which could be used to associate queries with a particular user over time. We used longitudinal analysis of search behavior in these logs as the basis for the early detection of ADRs for medications. Although the logs span a period of three years, any single user appears in the logs for at most 18 months, conforming to the terms of use under which the data were collected.

All data access and analysis was done in accordance with the search engine's published end-user license agreement, which specifies that user data may be used for research purposes and to improve the search experience. Our work was conducted offline, on data collected to support existing business operations, and in no way impacted the presentation of search results or other aspects of the user experience. All data were anonymized (such that users cannot be identified, directly or through identifiers linked to them) prior to data analyses. The Ethics Advisory Committee at Microsoft Research considers these precautions sufficient for triggering the Common Rule, exempting this research from detailed ethics review.

To ensure that we had sufficient data to perform our within-user long-term analysis, we focused on users in our dataset for whom we had observed at least 100 search sessions, yielding 57,101,343 users in total. Our unique user identifiers were based on Web browser cookies and were reset when users cleared their cookies. As such, we focused on users for whom we had more complete data on their long-term behavior. We experimented with different session-count thresholds, ranging from 1 to 200 search sessions. A threshold of 100 sessions yielded strong performance at the early detection task, while still retaining sufficient users to cover a sizeable set of drug–ADR pairs. Sessions were identified using a 30-min inactivity timeout to define session termination, a threshold commonly employed in research on user modeling in search logs [33,34]. Users linked to $\geqslant$1000 search queries on any given day were classified as automated traffic (Internet bots) and removed. In previous work [22], we found that the percentage of a user's queries that contained a medical term within their first month of search activity could help identify healthcare professionals (HCPs). Applying this filter, we removed the 1.45% of users who performed health-related queries for more than 20% of their searches (the same medical query percentage as used to filter HCPs in previous work [22]). We also swept the percentage of HCPs across the range of possible values and found that a threshold of 20% minimized the number of users excluded while still obtaining strong predictive performance in the forecasting of unknown ADRs. The determination of queries as healthcare-related was performed by a proprietary classifier used by the Microsoft Bing search engine to determine when to provide special support (e.g., instant answers on result pages) for health-related queries. Removal of HCPs is important given that health professionals may perform searches for many reasons, including patient care and continuing medical education and awareness. Also, physicians may have awareness of ADR knowledge before such information becomes public, e.g., through anecdotal patient reports or the medical literature, especially important in the prospective setting described in this article. We focus in our efforts on ADR surveillance on the pursuit