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

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

of information from individuals who can serve as primary early-warning sensors based on their experiences with medications. However, there can be value in analyzing the search activities of HCPs for health-related studies [35]. The mean average duration (M) of the logs for non-bot and non-HCP users was 263.57 days (standard deviation (SD) = 198.51 days).

2.2. Ground truth

The ground truth consisted of a set of drug–event pairs, each of which was classified as either a true association or false association based on data collected from FDA-approved product labels. The methodology employed in the generation of this dataset is described in detail in previous work [32]. The positive examples were selected based on safety-related medication label changes approved by the FDA and communicated to the public in the year 2013 through monthly summaries posted at the FDA's MedWatch Website [36]. The set of associations used in this work totaled 74 drug–ADR pairs, split between 15 positives cases (drawn from a larger set of 62 positive cases that match our requirements regarding the timing of when the ADR first emerged, described later) and 59 negative controls (drawn from a larger set of 75 negative pairs) which again were filtered per the terms of our ADR analysis. In addition to the binary labels, for the positive cases the data also contained the month within 2013 in which the FDA label change occurred. Since we did not want to risk using data after the label change, we took the first day of the month in which the label change occurred as the index date. This date was further revised to an even earlier date as needed using the Wayback filtering method described in detail in the next section.

2.3. Wayback filtering

Web page content has been shown to change considerably over time [37] and the content accessed on those pages can influence future queries [38,39]. To control for confounds of Web page content associated with ADR information being present in Websites prior to the index date, we revised the index date of the backdated analysis for a given test case to the earlier of the label change date or a date derived from the Internet Archive using the Wayback Machine. Fig. 1 illustrates the process by which the index date is revised based on the presence of the drug–ADR pair in a Web page preceding the FDA label change date.

For each of the 39 distinct drugs in our set of positive and negative cases, we mined 18 months of query logs from the Microsoft Bing search engine to obtain all clicked results for the union of all queries containing the generic drug name or the brand names under which it was marketed. Table S2 contains the generic and the associated brand names for each drug. We considered all users' Bing search result clicks for these queries, including those that were only visited by a single user in the 18-month timeframe. This provided a broad range of pages that could potentially contain information on the ADR in advance of the FDA label change date. For each of the Web pages in the resultant set, we obtained its HTML content using the Wayback machine, which allowed us to review the content of the site periodically going back in time, in some cases over 10 years. On average, sites in our dataset were crawled by the Wayback Machine with a frequency of approximately once every three months (M = 97.27 days, SD = 109.27 days). This is similar to the quarterly cadence with which the FAERS data are made available publically.

For each test case (drug–ADR pair), given the ADR, we could therefore compute the earliest date that it was mentioned on the previously identified sites. We automatically scraped the Websites for co-occurrences of the drug and ADR in the content of the page over time. Appearances of such information guided a revision of the date to a time when suspicions or concerns regarding the ADR emerged online. In total, 11.70% of all visited search results contained at least one matching ADR for the drug of interest. The ADRs and their associated synonyms are listed in Table S1 (see Supplementary Materials). Synonyms were determined based on manual review of authoritative medical Websites describing the ADR as well a two-step walk on the search engine click graph similar to previous work [40] i.e., originating with the ADR as a query, finding all URLs in logs visited for that query, and then finding a set of queries issued by at least 10 users resulting in visits to the same URLs. Both the ADR and its synonyms were included in the content analysis. In 65.6% of the original set of 62 positive pairs there was at least one clicked result (visited by at least one user) searching for the drug–ADR pair before the FDA label change date, and often well before that date (M = 4.40 years, Max = 10.22 years). This finding demonstrates the value of the additional refinement of the index date to a point in time where knowledge was available in advance of FDA labeling actions.

Positive examples were included in the set of positive cases if one of the following three criteria were met: (1) the earliest date from the Wayback machine analysis was between 2011 and 2013 inclusive; (2) the earliest date of the Wayback machine analysis was after the FDA label change date, or (3) there was no record of the ADR being associated with the drug on any Website, regardless of the date. 15 positive examples met these criteria and were included in our analysis. Negative examples (59 in total) were included based on similar criteria as positives (either (1) or (3)). In both positives and negatives, we excluded pairs where the first mention of the drug–ADR pair was on a Website prior to 2011 since we could not guarantee that we were removing the effect of that information being public from our logs.

2.4. Association analysis

Disproportionality analysis [32] was used to quantify the strength of association between a given drug–ADR pair being analyzed. Specifically, we computed and used the *observed-to-expected* ratio as our measure of association (signal score used in the subsequent ROC analysis). This generates a score for each drug–ADR pair quantifying the extent of deviation from the expected (independence). The higher the score, the stronger the connection between the drug and the ADR. The outcome is one of 38 adverse events ranging from mild (e.g., hallucinations, nocturnal enuresis) to rare and serious (e.g., eosinophilic pneumonia, severe bullous dermatitis). For each drug–ADR pair and date of analysis, the *observed* is the number of users that (1) search for both the drug and event within a specified time interval following the first instance of a search for that drug by that user, and (2) search for the drug and ADR prior to the index date of the analysis. The time interval is based on co-occurrence within the same: *query*, *search session* (defined as noted earlier), *day* (within 24 h), *week* (within 7 days), *two weeks* (within 14 days), *month* (within 30 days), or over *all time* (WholeLife, i.e., the remaining duration of the log for the current user). The average post-drug duration for the WholeLife analysis per-user was 174.98 days (SD = 166.63 days). Since we were interested in causal associations between medications and events, we required that the drug query preceded the ADR query chronologically in order to be included in our calculation. The *expected* rate is the expected number of users that search for both the drug and ADR under the assumption that searches for both is random. We note that performing this type of analysis at the user level is common in other applications of search logs since it avoids skew toward more active users, who could be overrepresented in a query-level analysis [21,22,41].

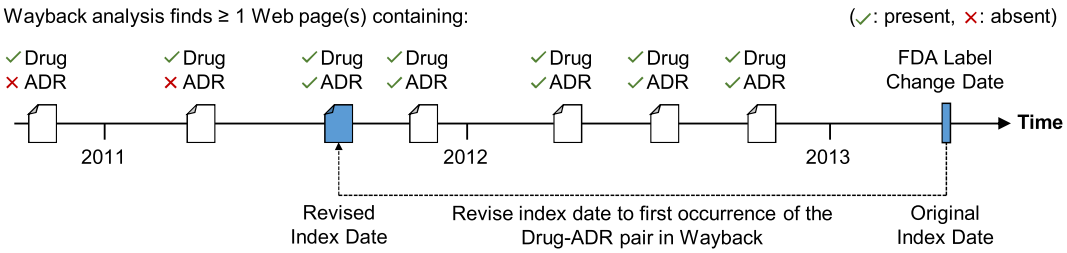

Fig. 1. Revision of the index date based on the co-occurrence of the drug–ADR pair on a Web page in Wayback analysis before the FDA label change date. Drug/ADR presence/absence on at least one Web page at a particular point in time is highlighted with ticks (present) or crosses (absent).

To account for small user counts we used the lower limit of the *observed-to-expected* ratio's 95% confidence interval (based on the Poisson distribution) as the final signal score to be used in the performance (ROC) analysis. Using the lower limit confidence interval statistics instead of point estimates is a common adjustment applied in pharmacovigilance signal detection [6] to reduce false signals. Table S2 (Supplementary Materials) shows the signal scores, user counts, and confidence intervals for the 74 test cases.

2.5. Scoring methodology

Fig. 2 presents a flowchart illustrating the process followed to compute signal statistics from the search engine log data. The flowchart shows the filtering and adjustment of the index dates in the ground truth using the Wayback method described earlier in this section. The positive and negative examples that meet our criteria are included in the evaluation set. Index dates are determined based on drug–ADR co-occurrence on Web pages and/or the relationship between the earliest date that such co-occurrences are observed in the Wayback machine and the label change date from the FDA. Three years of search logs are filtered to remove bots, inactive users, and HCPs; all of whom may bias our analyses in different ways. Disproportionality analysis is performed at each time interval (SameQuery, SameSession, etc.) using the resultant logs and our ground truth comprising 74 test cases (15 positives and 59 negative controls). The signal scores obtained from the disproportionality analysis are used to rank the drug–ADR pairs and compute the AUROC metrics used in evaluating signal detection performance.

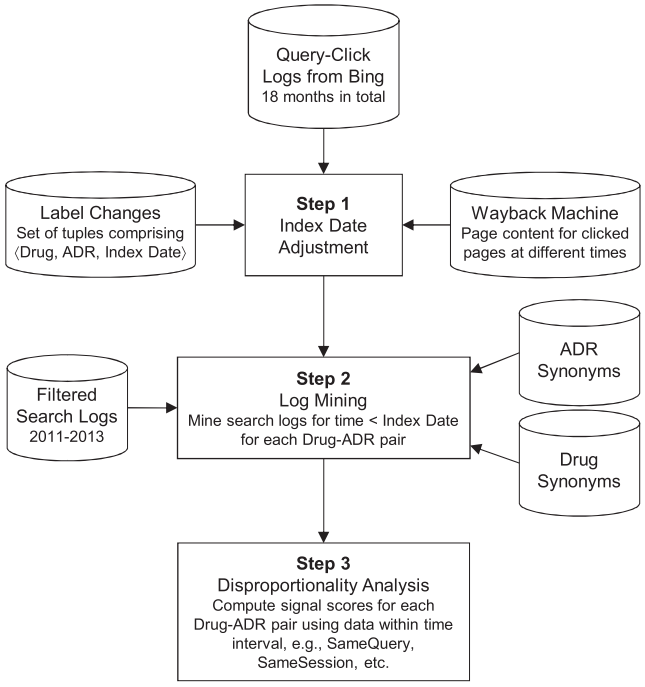

Fig. 2. Flowchart showing the three-step process to compute signal statistics from search logs. FDA label change dates offer useful timing information on the formal identification and actions taken in regards to a new ADR. However, there may be prior availability of information about the drug–ADR pair on Websites in advance of the label change date. The availability of this information online could lead people to query about the ADRs regardless of whether they were experiencing the ADR. To address this potential leak of existing knowledge on searches, we considered the date of revelation of knowledge about ADRs to be the earliest of the label change date or the first online report of the ADR in relation to a medication.

3. Results

3.1. Prospective analysis

The results of our backdated analyses of the 74 test cases comprising our benchmark (see Section 2) are displayed in Table 1. A signal score for each test case was computed via disproportionality analysis (see Section 2) [42]. The discriminatory power of ADR signals generated from our search logs signals was measured using the area under the receiver operating characteristic (ROC) curve (AUROC), a common performance index for signal detection in pharmacovigilance [6,21,22]. Table 1 shows signal detection performance for different time intervals between the first observation of the drug in a user's long-term search history and the subsequent ADR search. For example, SameQuery describes the co-occurrence of the drug and ADR in the same query, whereas for OneDay the ADR search happens within 24 h of the first drug search. The time intervals extend from SameQuery to WholeLife (i.e., the full duration of the available log data for the drug searcher following the first search for the medication of interest in our log data). In ranking the test cases, we apply a correction for small user counts to score each case based on the lower 95% confidence interval of the signal score. Although the original AUROC values are slightly larger, none of the differences in AUROC between applying the original signals and the adjusted signals for rankings are significant

Table 1
Performance of the models (AUROC) for different time intervals following the first observed drug query. The time interval denotes the time in which an ADR query can be observed.

Time interval	AUROC
SameQuery	0.514
SameSession	0.580
OneDay	0.609
OneWeek	0.721
TwoWeeks	0.840
OneMonth	0.811
WholeLife	0.817

using DeLong's test [43] at $p < 0.05$. We use the adjusted signal scores for all of the analyses.

Table 1 shows that the AUROC for longer time intervals (TwoWeeks or more) is higher than shorter time periods (OneWeek or less) (all $p < 0.001$). The differences between the longer time intervals are not statistically significant (all $p > 0.05$). The advantage of considering longer time intervals is the potential to find more drug searchers and more people experiencing ADRs, and hence cover more drug–ADR pairs. Although TwoWeeks has the highest AUROC in Table 1, two weeks may be insufficient to observe ADRs resulting from all drugs; a 30–90 day observation period has been more commonly used in ADR analysis [22,44]. In fact, in previous work we have excluded activity close to the drug search as less likely to be experiential [22]. To maximize coverage, we focus on WholeLife for the remainder of our analysis. Fig. 3 shows the ROC curve for the WholeLife period of evaluation.

3.2. Comparison with FAERS

Table 2 displays the results of a similar analysis based on spontaneous reports from FDA FAERS. The analysis was based on the same 74 test cases employed in the evaluation of the search logs. The aim of this comparative analysis is to obtain a baseline benchmark for log performance and understand the difference between our log-based method and FAERS. FAERS has a distinct advantage for this analysis: it is one of the main sources of evidence used by the FDA in determining label changes.

Each row of Table 2 provides the AUROC of signals generated from FAERS by backdating the analysis to the specified year ends (2010–2013). Each row also displays the total number of spontaneous reports used for the analysis at each time point. In addition, Table 2 displays a backdated analysis using FAERS aligned to the dates used for the WholeLife analysis of search logs (Table 1), i.e., label change dates revised by the Wayback Machine. The signals were generated using FDA's primary signal-detection algorithm: the Multi-Item Gamma Poisson Shrinker [45,46].

A strictly fair comparison with FAERS for the task of early detection is not possible given the causal connection between FAERS reports and label changes that is not present for the logs. For this reason, we did not perform tests on the significance of differences between FAERS and the logs. However, Table 2 is still informative. The performance of search logs (AUROC = 0.82) is similar to FAERS for detecting ADRs at least one year ahead of their publication (i.e., the 2010–2011 timeframe, AUROC = 0.79–0.80). The signals from the two data sources are most comparable at this time, where signals derived from FAERS are least likely to be related to the FDA's decision making processes regarding label changes. We included 2010 in this analysis because the data were available in FAERS and it provided an additional data point on the performance of FAERS as an early detection method well before the label changes occurred.

3.3. Varying lead times

The analysis in the previous section reported results using logs right up until the index date. Forecasting future events well in advance of their occurrence is an important aspect of prospective analysis. There is value in understanding the effectiveness of our method given different increasing lead times in advance of the index date. We note that given the fixed duration of the log data that we use, less log data were available as we move back in time, potentially impacting the predictive power of the methods. Table 3 reports the performance of the search logs in predicting ADRs at different lead times (in months) before the index date in six month increments. There are fewer positive and negative cases as we regress further back in time. For consistency we report the AUROC

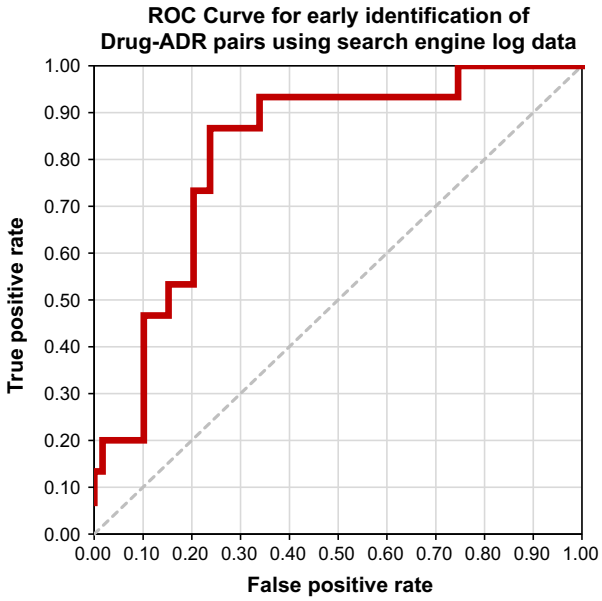

Fig. 3. ROC curve illustrating predictive performance using behavioral search log data collected during the WholeLife time interval (AUROC = 0.817).

Table 2
Prospective signal detection performance of FAERS.

Year	AUROC	Number of reports
2013	0.84	6,268,207
2012	0.88	5,494,345
2011	0.79	4,857,012
2010	0.80	4,334,379
Aligned	0.86	N/A

at each lead time over the 63 pairs (9 positives and 54 negatives) (85% of the original set of pairs) for which we had behavioral evidence in the search logs in the 24 months before the index date. For completeness we also report the AUROC across all available pairs at each lead time interval. The available pairs change (increase) as the lead time drops, given the appearance of additional behavioral evidence in the log data over time.

The results show that, while performance generally decreases with increased lead time, we can still accurately forecast ADR well in advance of the label changes. When focusing on the 63 pairs for which we had evidence 24 months in advance of the label change date, the AUROCs are often higher than that reported in Table 1. This may be because these pairs are popular in the logs, and for WholeLife we can make more reliable estimates of their performance historically and even better estimates as we obtain more data as the index date approaches. When considering all available pairs at each lead time, there is a dip in AUROC at 18 months (i.e., from 0.637 at across the 63 drug–ADR pairs at 24 months to 0.568 across the 69 pairs at 18 months). During that time, six new pairs are added for which there may be little behavioral log data. This can lead to less reliable signal detection across the 69 pairs.

4. Discussion

We describe methods that can be used in for prospective ADR detection scenarios where effective drug surveillance is most necessary. We showed that we can reliably predict unknown adverse drug reactions using search log data. We compared the performance with that of FAERS and at different timeframes and the results show that the performance of the two methods is

Table 3
Performance of the models given different lead times (t) in months before the index date. Pairs are ranked based on signal scores using the WholeLife time interval.

Lead time (t)	Same pairs as $t = 24$	All available pairs at each t				Total number of users
	AUROC	AUROC	#Positives	#Negatives	#Total	
None	0.883	0.817	15	59	74	57,101,343
6 months	0.858	0.775	15	59	74	52,420,889
12 months	0.759	0.635	14	57	71	43,099,412
18 months	0.669	0.568	13	56	69	33,920,410
24 months	0.637	0.637	9	54	63	24,538,372

comparable. This is promising, especially in light of the strong baseline provided by the FAERS-based prediction. The findings in Table 1 suggest that longer time intervals (TwoWeeks or more) result in more accurate log-based predictions. Reasons for this include (a) more opportunity to observe the ADR, leading to larger user counts and more reliable statistical estimates, and (b) the time between the drug and the ADR searches may be more typical of an individual experiencing the ADR, while shorter time intervals (OneWeek or less) may be more likely to reflect user interests in possible connections between the drug and the ADR.

In summary, we have demonstrated that candidate ADRs can be detected in advance of public knowledge about the adverse effects via a log-based signal-detection methodology. These findings resonate with other successes of using log-based signals in pharmacovigilance [21,22]. However, we now show the successful detection of ADRs that were unknown during the timeframe studied (vs. only a single drug pair in previous work [21]). This brings the solution closer to the application scenario in which the search logs could have strong utility: early detection of unknown ADRs, so as to inform follow-up studies by the FDA and others. Although the label change date reflected the time at which the FDA announced the change publically, suspicions and concerns about the ADR that are reported on Websites, based on signs and symptoms experienced by users or scientific studies, could still be surfaced by search engines. This information could influence search behavior well in advance of the label change date. The use of the Internet Archive to estimate when signals appeared in publically facing resources retrospectively was valuable in our context and has utility for other applications. For example, this technique could be used to discount the effect of the news media and other influencing factors on the aggregated behavioral signals that are observed in the search logs. Although the analysis is performed using logs from the Microsoft Bing search engine, nothing in our approach is specific to Bing. The same methods could be applied in other search engines given longitudinal search data from their users. Only three years of logs were available for experimentation in this study. Longer-term data are required to make generalizable claims about the performance of our methods, including whether the accuracy of the predictions increases over time as more than three years of data become available.

We note that our findings demonstrate the value of using logs to accurately distinguish between a given reference set of positives and negatives. Thus, the methods and results were framed by a predefined set of drug–ADR pairs of interest. Such a fixed set of conditions and associated symptoms reduces the scope of the surveillance challenge and thus makes the analysis computationally tractable. The challenge of discovering ADRs from large numbers of potential drug–event pairs remains an open question. Directions of future research include the application of the methods to new medications or to search for adverse interactions in frequently-occurring combinations of medications. Polypharmacy is common, especially among the elderly [47]. Applying search log data to better understand interactions between combinations of drugs—pairs, triples, and beyond—could help discover unknown side effects.

The risk of false positives in operational surveillance systems run by governmental agencies and pharmaceutical companies represents a significant challenge. Each drug–ADR pair that is identified as a potential hazard poses a potentially costly challenge for deeper examination. As such, false positives can lead to alert fatigue, wasted resources, and render signal detection systems impractical. The ROC curve in Fig. 3 suggests that we can attain reasonable recall of true positives while preserving a low false-positive rate. For example, in fixing the false-positive rate to 0.3 (specificity: $\geqslant 0.7$, a suggested value for clinical significance for signal-detection assessment [48]) we can recall 86.7% of the true positive cases in our dataset. The discrimination threshold used in practice should be determined based on desired performance profile of the signal detection model. The threshold could be reduced to improve recall or increased to improve precision. Further experiments are needed to better understand the costs that could be incurred from mistakenly identifying drug–ADR pairs at different rates in practice. As part of this, we need to establish thresholds that maintain reasonable levels of tolerance with regard to false positives while still recalling sufficient numbers of true positives for a surveillance system based on these methods to have practical utility.

We also note limitations in the study introduced by the data and reference standard that we are using. For example, we only had access to three years of search logs in total and 18 months for any particular user; insufficient data to perform backdated types of analysis to earlier time points. Our results are also limited to the particular reference set used and the number of valid test cases in that dataset.

Since we do not have ground truth about the intentions behind people's observed activity, we cannot be certain that they are truly experiencing the side effects of interest. The use of ADR symptoms reduced predictive performance, likely because symptoms can be associated with a variety of medical conditions, only some of which are the ADRs of interest in this study. Variations of the methods presented in this article where we target users who focus only a set of sentinel symptoms for the duration of our observations of them in the log data (vs. broad explorations of many symptoms) may be valuable. We can train classifiers to identify whether a search was experiential based on aspects of the user's search behavior including the recurrence of related queries over time and the content of visited Web pages [49]. The use of specific query terminology such as first-person statements (e.g., queries containing "I was prescribed") or those pertaining to starting a course of medications (e.g., queries containing "first week on" or "just started") may also be valuable in detecting searchers who are likely to be experiencing associated side effects.

Author contributions

All authors contributed to the design of the methods and experiments. RW, EH, and RH wrote the manuscript. RW, SW, and AP analyzed the log data. RH analyzed the FAERS data. WD provided guidance on statistical methods. All authors reviewed the manuscript.

Competing interests

RW, EH, AP, PS, and WS are employed by Microsoft. RH and WD are employed by Oracle. SW is a research student at University of Illinois at Urbana-Champaign and was an intern at Microsoft when performing this research.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.11.005.

References

[1] J. Lazarou, B.H. Pomeranz, P.N. Corey, Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies, J. Am. Med. Assoc. 279 (15) (1998) 1200–1205.
[2] D.C. Classen, S.L. Pestotnik, R.S. Evans, J.F. Lloyd, J.P. Burke, Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality, J. Am. Med. Assoc. 277 (4) (1997) 301–306.
[3] S.R. Ahmad, Adverse drug event monitoring at the food and drug administration, J. Gen. Intern. Med. 18 (1) (2003) 57–60.
[4] R. Harpaz, W. DuMouchel, N.H. Shah, D. Madigan, P. Ryan, C. Friedman, Novel data-mining methodologies for adverse drug event discovery and analysis, Nat. Clin. Pharmacol. Therap. 91 (6) (2012) 1010–1021.
[5] A. Szarfman, S.G. Machado, R.T. O'Neill, Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database, Drug Saf. 25 (6) (2002) 381–392.
[6] R. Harpaz, W. DuMouchel, P. LePendu, A. Bauer-Mehren, P. Ryan, N.H. Shah, Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system, Nat. Clin. Pharmacol. Therap. 93 (6) (2013) 539–546.
[7] R. Platt, M. Wilson, K.A. Chan, J.S. Benner, J. Marchibroda, M. McClellan, The new sentinel network: improving the evidence of medical-product safety, N. Engl. J. Med. 361 (7) (2009) 645–647.
[8] P.M. Coloma, M.J. Schuemie, G. Trifirò, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, G. Corrao, L. Pedersen, J. van der Lei, M. SturkenboomEU-ADR Consortium, Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EUADR Project, Pharmacoepidemiol. Drug Saf. 20 (1) (2011) 1–11.
[9] P.E. Stang, P.B. Ryan, J.A. Racoosin, J.M. Overhage, A.G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, J. Woodcock, Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership, Ann. Intern. Med. 153 (9) (2010) 600–606.
[10] K.D. Shetty, S.R. Dalal, Using information mining of the medical literature to improve drug safety, J. Am. Med. Inform. Assoc. 18 (5) (2011) 668–674.
[11] P. Avillach, J.C. Dufour, G. Diallo, F. Salvo, M. Joubert, F. Thiessard, F. Mougin, G. Trifirò, A. Fourrier-Réglat, A. Pariente, M. Fieschi, Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project, J. Am. Med. Inform. Assoc. 20 (3) (2013) 446–452.
[12] H. Pontes, M. Clement, V. Rollason, Safety signal detection: the relevance of literature review, Drug Saf. 37 (7) (2014) 471–479.
[13] S. Vilar, R. Harpaz, H.S. Chase, S. Costanzi, R. Rabadan, C. Friedman, Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis, J. Am. Med. Inform. Assoc. 18 (1) (2011) 73–80.
[14] A.P. Chiang, A.J. Butte, Data-driven methods to discover molecular determinants of serious adverse drug events, Nat. Clin. Pharmacol. Therap. 85 (3) (2009) 259–268.
[15] M. De Choudhury, M.R. Morris, R.W. White, Seeking and sharing health information online: comparing search engines and social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014, pp. 1365–1376.
[16] R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu, N.H. Shah, Text mining for adverse drug events: the promise, challenges, and state of the art, Drug Saf. 37 (10) (2014) 777–790.
[17] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts in health-related social networks, in: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, 2010, pp. 117–125.
[18] P. Wicks, T.E. Vaughan, M.P. Massagli, J. Heywood, Accelerated clinical discovery using self-reported patient data collected online and a patient matching algorithm, Nat. Biotechnol. 29 (5) (2011) 411–414.
[19] Pew Research Center. Pew Internet & American Life Project: Health Online 2013. <http://www.pewinternet.org/2013/01/15/health-online-2013> (Accessed July 2015).
[20] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (2009) 1012–1014.
[21] R.W. White, N.P. Tatonetti, N.H. Shah, R.B. Altman, E. Horvitz, Web-scale pharmacovigilance: listening to signals from the crowd, J. Am. Med. Inform. Assoc. 20 (3) (2013) 404–408.
[22] R.W. White, R. Harpaz, N.H. Shah, W. DuMouchel, E. Horvitz, Toward enhanced pharmacovigilance using patient-generated data on the internet, Nat. Clin. Pharmacol. Therap. 96 (2) (2014) 239–246.
[23] E. Yom-Tov, E. Gabrilovich, Post-market drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries, J. Med. Internet Res. 15 (6) (2013) e124.
[24] D. Butler, When Google got flu wrong, Nature 494 (7436) (2013) 155–156.
[25] S. Cook, C. Conrad, A.L. Fowlkes, M.H. Mohebbi, Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic, PLoS ONE 6 (8) (2011) e23610.
[26] D. Lazer, R. Kennedy, G. King, A. Vespigiani, The parable of Google flu: traps in big data analysis, Science 343 (6176) (2014) 1203–1205.
[27] D. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. 99 (9) (2002) 5766.
[28] E. Bakshy, B. Karrer, L.A. Adamic, Social influence and the diffusion of user-created content, in: Proceedings of the ACM Conference on Electronic Commerce, 2009, pp. 325–334.
[29] P.B. Ryan, P.E. Stang, J.M. Overhage, M.A. Suchard, A.G. Hartzema, W. DuMouchel, C.G. Reich, M.J. Schuemie, D. Madigan, A comparison of the empirical performance of methods for a risk identification system, Drug Saf. 36 (1) (2013) 143–158.
[30] G.N. Noren, O. Caster, K. Juhlin, M. Lindquist, Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance, Drug Saf. 37 (9) (2014) 655–659.
[31] R. Harpaz, W. DuMouchel, N.H. Shah, Comment on: "Zoo or Savannah? Choice of training ground for evidence-based pharmacovigilance", Drug Saf. 38 (1) (2015) 113–114.
[32] R. Harpaz, D. Odgers, G. Gaskin, W. DuMouchel, R. Winnenburg, O. Bodenreider, A. Ripple, A. Szarfman, A. Sorbello, E. Horvitz, R.W. White, N.H. Shah, A time-indexed reference standard of adverse drug reactions, Nat. Sci. Data 1 (2014).
[33] D. Downey, S.T. Dumais, E. Horvitz, Models of searching and browsing: Languages, studies, and application, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2007, pp. 2740–2747.
[34] R.W. White, S.M. Drucker, Investigating behavioral variability in web search, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 21–30.
[35] D.J. Odgers, R. Harpaz, A. Callahan, G. Stiglic, N.H. Shah, Analyzing search behavior of healthcare professionals for drug safety surveillance, in: Proceedings of Pacific Symposium on Biocomputing, 2014, pp. 306–317.
[36] MedWatch, <http://www.fda.gov/Safety/MedWatch> (Accessed July 2015).
[37] E. Adar, J. Teevan, S.T. Dumais, Large scale analysis of web revisitation patterns, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008, pp. 1197–1206.
[38] D.J. Liebling, P.N. Bennett, R.W. White, Anticipatory search: using context to initiate search, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, pp. 1035–1036.
[39] Y. Ustinovskiy, P. Serdyukov, Personalization of web-search using short-term browsing context, in: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, 2013, pp. 1979–1988.
[40] D. Beeferman, A. Berger, Agglomerative clustering of a search engine query log, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 407–416.
[41] M. Richardson, Learning about the world through long-term query logs, ACM Trans. Web 2 (4) (2008) 21.
[42] A. Bate, S.J.W. Evans, Quantitative signal detection using spontaneous ADR reporting, Pharmacoepidemiol. Drug Saf. 18 (6) (2009) 427–436.
[43] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (1988) 837–845.
[44] R. Harpaz, W. DuMouchel, P. LePendu, N.H. Shah, Empirical Bayes model to combine signals of adverse drug reactions, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1339–1347.
[45] W. DuMouchel, Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system, Am. Stat. 53 (3) (1999) 177–190.
[46] W. DuMouchel, D. Pregibon, Empirical Bayes screening for multi-item associations, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, pp. 67–76.
[47] S.I. Haider, K. Johnell, M. Thorslund, J. Fastbom, Trends in polypharmacy and potential drug–drug interactions across educational groups in elderly patients in Sweden for the period 1992–2002, Int. J. Clin. Pharmacol. Ther. 45 (12) (2007) 643–665.
[48] P.B. Ryan, D. Madigan, P.E. Stang, O.J. Marc, J.A. Racoosin, A.G. Hartzema, Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership, Stat. Med. 31 (30) (2012) 4401–4415.
[49] M.J. Paul, R.W. White, E. Horvitz, Diagnoses, decisions, and outcomes: Web search as decision support for cancer, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 831–841.