Web-Scale Pharmacovigilance: Listening to Signals from the Crowd

Ryen W. White Ph.D.[1]*, Nicholas P. Tatonetti Ph.D.[2], Nigam H. Shah M.B.B.S. Ph.D.[3], Russ B. Altman M.D. Ph.D.[4] and Eric Horvitz M.D. Ph.D.[1]

[1] Microsoft Research, Redmond, WA 98052, USA

[2] Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA

[3] Department of Medicine, Stanford University, Stanford, CA 94028, USA

[4] Departments of Bioengineering and Genetics, Stanford University, Stanford, CA 94028, USA

* Correspondence to: ryenw@microsoft.com

Word count: 2628

ABSTRACT

Context: Adverse drug events cause substantial morbidity and mortality and are often discovered after a drug comes to market.[1] In the US alone, adverse drug events cause thousands of deaths annually and their associated medical treatment costs billions of dollars.[2,3]

Objective: Given that a significant use of the Internet is for health searches,[4] we hypothesized that Internet users may provide early clues about adverse drug events via their online information-seeking activities.

Design: We conducted a large-scale study of Web search log data gathered during 2010. We pay particular attention to the specific drug pairing of paroxetine and pravastatin, whose interaction to cause hyperglycemia was reported *after* the time period of the online logs used in the analysis. We also examine sets of drug pairs known to be associated with hyperglycemia and those that have not been associated with hyperglycemia.

Results: Our study shows that signals concerning drug interactions can be mined directly from search logs and confirms the findings of laboratory studies as well as prior known associations.

Conclusions: This is the first study to extract evidence of drug interactions from search log data. Compared to analysis of other sources such as electronic health records (EHR), logs are inexpensive to collect and mine, are not dependent on healthcare utilization, and are not subject to the same latencies. The results demonstrate that logged search activities by populations of computer users captured by Internet services can contribute to drug safety surveillance.

BODY

Background

The Food and Drug Administration (FDA) and other organizations collect reports on drug side effects from physicians, pharmacists, patients, and drug companies. These reports provide valuable clues about drug-related adverse events, but are incomplete and biased.[5,6,7] As a result, adverse event alerts for single drugs are often delayed as evidence accumulates.[8,9] These challenges are compounded in the setting of adverse events resulting from multiple drugs that interact in unexpected ways.

Previous research on tracking seasonal influenza has demonstrated that search logs can form an implicit sensor network for health monitoring.[10,11] In that work, search logs accurately estimated the weekly levels of influenza activity in different regions of the United States, with a reporting delay of approximately one day. The authors showed that health-seeking activity captured in queries to online Web search services mirrors trends in data gathered by traditional surveillance systems based on virological and clinical data.

We employ search log data for a different purpose: we seek to harness people's online health-seeking search activity in the aggregate to identify adverse drug events associated with drug interactions. Patients may seek information on the Web about the drugs prescribed to them or to close family members, and to explore the potential explanations of new symptoms.[12] We consider as a test case an interaction between paroxetine (an antidepressant) and pravastatin (a cholesterol-lowering drug) which was recently reported to create hyperglycemia.[13,14] This association was extracted from the FDA Adverse Event Reporting System (AERS) using a data-mining algorithm that aggregates reports to identify drug-drug interactions.[13] The finding was confirmed in a retrospective analysis of the electronic health records of three regionally distinct medical institutions and confirmed in a mouse model.[14] We hypothesize that patients taking these two drugs might experience symptoms of hyperglycemia and may have conducted Internet searches on these symptoms and concerns related to hyperglycemia *before* the association was reported in 2011.

Method

We analyzed the search logs of millions of consenting Web users who opted to share search activities with Microsoft via the installation of a browser add-on, spanning a 12-month period of all of 2010 and comprising searches on Google, Bing, and Yahoo!. An anonymous identifier tied to the instance of the browser add-on was used to track the drugs and symptom queries that each user performed over time (note that we were unable to distinguish between multiple users of the same machine). Searches for information on prescription drugs are common. We found that over 1 in 250 of people (0.43%) pursued information on at least one of the top-100 best-selling drugs in the United States, including paroxetine and pravastatin, the medications that we focus on here.[15]

By examining words used in user queries, we sought evidence that searches from people exploring pravastatin and paroxetine over time (using logs from 2010) would have a higher rate of including hyperglycemia-associated words than people searching for only one of the drugs. The list of hyperglycemia-related terminology that was used is included in the supplementary materials (Table S1). We generated the list based on a review of medical literature. The list is broad to ensure that we covered a majority of related symptoms. Although there are many possible causes for the symptoms listed, each can be associated with hyperglycemia. We seek to detect increases in the use of terms from the list in exploratory Web searches by holding the list constant and noting the presence or absence in user logs of queries for the medications which have been found to cause hyperglycemia when taken together.

We first mined the 12 months of search logs to identify users who had searched for hyperglycemia-related symptoms or terms. We then identified users in each of the following groups: (i) *Both (paroxetine & pravastatin) searchers*, comprising those who searched on paroxetine (or one of its trade name variants: Aropax, Paxil, Seroxat, and Sereupin) and pravastatin (or its trade name Pravachol); (ii) *Pravastatin, independent of paroxetine, searchers*, comprising those users who searched for pravastatin regardless of whether they also searched for paroxetine; and (iii) *Paroxetine, independent of pravastatin, searchers*, comprising those users who searched for paroxetine irrespective of whether they also searched for pravastatin.

We count the number of users in each of the three user groups, and the number of users in each group who searched for at least one of the terms associated with hyperglycemia (i.e., the intersection with the set of hyperglycemia searchers). These populations can be visualized with a Venn diagram, as shown in Figure 1.

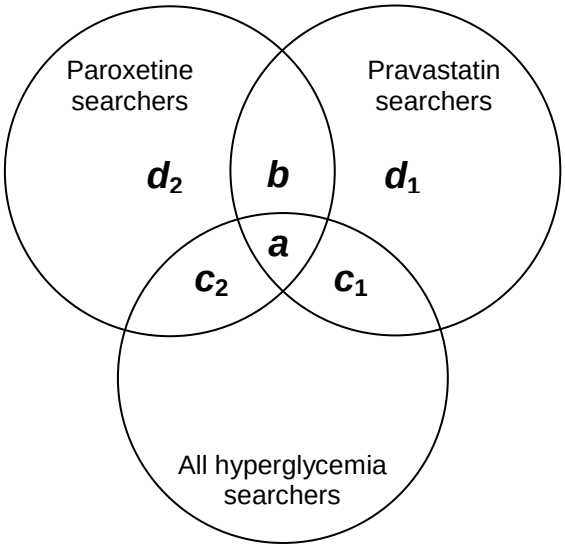

Fig 1. Venn diagram showing the different user groups in our analysis (not drawn to scale). Letters denote different subsets of searchers, with *a* referring to those who search on both paroxetine and pravastatin and also search on hyperglycemia-related terminology, and *b* to those who search on both drugs. Subsets d_1 and d_2 refer to those who search on pravastatin and on paroxetine, respectively. Subset c_1 denotes those who search for pravastatin and hyperglycemia-related terms and c_2 those who search o paroxetine and hyperglycemia-related terms. We perform two disproportionality analyses, with expected (background) based on pravastatin using c_1 and d_1 and with expected based on paroxetine using c_2 and d_2.

We use disproportionality analysis[6] to assess the increased chance of a user searching for hyperglycemia-related terms given that they search for both pravastatin and paroxetine. Reporting ratios (RR) are computed based on observed versus expected adverse reports.[16] Given the broad spectrum of information goals on the Web, for the search logs, we use a conditional disproportionality analysis that introduces a contextual focus to minimize false positives. In this case, we seek evidence for increased searches for hyperglycemia related terms within the specific context of searches on a drug or drugs of interest. In exploring the potential influence of the two

drugs together, we consider people who have searched for each of the drugs individually over the same period as controls.

Given the subsets of users defined above, disproportionality analysis was used to identify drug pairs that occur at higher than expected frequencies with hyperglycemia related terms. RR is defined as *observed/expected* or (a/b)/(c/d). *Observed* is defined as the fraction of users who search for both pravastatin and paroxetine (b) who also query for hyperglycemia symptoms (a), and *expected* is defined as the fraction of users who search for pravastatin (d_1) who also search for hyperglycemia symptoms (c_1), or (symmetrically) the fraction of users who search for paroxetine (d_2) who also search for hyperglycemia symptoms (c_2).

When RR is based on expected for pravastatin as background and search logs, a = number of users in the paroxetine & pravastatin set who searched for hyperglycemia-related terminology; b = number of users in the paroxetine & pravastatin set; c_1 = number of users in the pravastatin-only set who searched for hyperglycemia-related terminology, and d_1 = number of users in the pravastatin-only set. Figure 1 shows how each of these variables (a-d) relates to the three user groups defined earlier and their intersection with each other and all hyperglycemia searchers. We similarly compute RR with expected conditioned on paroxetine as background.

Findings

User Groups and Prevalence

To perform the analysis described in the remainder of this article, we analyzed 82 million drug, symptom, and condition queries from 6 million Web searchers. To ensure coverage, we looked for co-occurrences of the two medications for each user within the 12-month timeframe. For the group of users showing these co-occurrences, paroxetine and pravastatin did not co-occur within the same query; 29.61% of the observed drug pairs occurred in searches within the same day, 41.90% within the same week, and 60.89% within the same month. Figure 2 shows the fraction of users in each of the groups who query for any of the hyperglycemia-related terms in Table S1. The value for *Background* in the figure is the fraction of all users who query for the hyperglycemia-linked terms independent of the presence of pravastatin and paroxetine in any of

their queries. The figure shows that people who search for both paroxetine and pravastatin over the 12-month period are more likely to perform searches on the terms associated with hyperglycemia (around 10% of users who search for the drug pair) than those who search on only one of the drugs (around 5% of paroxetine users, around 4% of pravastatin users). Around 0.3% of all users search for one or more terms from the list (shown as *Background* in the figure). The figure also shows that the difference between the groups is consistent over the 12-month period and that there are no temporal variations such as seasonal affects.

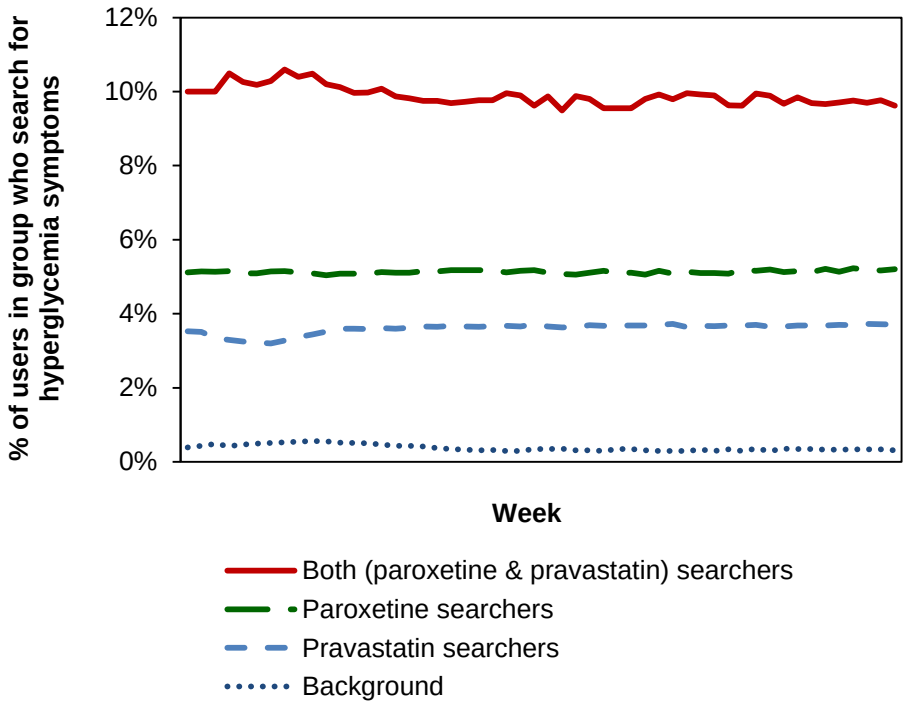

Fig. 2. Percentage of users in each of the three user groups searching for hyperglycemia-related terms. Percentage is computed per week over 12 months of search log data. *Background* refers to the fraction of all searchers who search for hyperglycemia-related symptoms or terminology independent of the presence of the drugs in the users' search histories.

Disproportionality Analysis

Table 1 shows the results of the conditional disproportionality analysis for RR computed using expected for pravastatin and expected for paroxetine.

Table 1. Results of disproportionality analysis for Expected (pravastatin), Expected (paroxetine).

	a	***b***	***c***	***d***	***RR***	***95% CI (Lower, Upper)***	***p-value (one-tailed)***
Expected (pravastatin)	342	2716	2581	56302	2.747	2.438, 3.094	< 0.0001
Expected (paroxetine)	342	2716	3645	71243	2.461	2.189, 2.767	< 0.0001

The results in Table 1 show that searching with terms that capture hyperglycemia symptomatology is observed more frequently in users searching for both drugs than in those searching for each drug separately. This result based on data from a non-clinical source resonates with findings from AERS and laboratory analysis described earlier.[13,15] As we know the date that the discovery of the interaction was made public, we can examine prior log data prior with confidence that the logged activities are not influenced by information about known interactions published later. However, since this is only a single drug pair, it is possible that the results are explained by an un-modeled mechanism or by chance.

Disproportionality Analysis for Known Drug-Drug Interactions

To address the concern associated with focusing on a single pair, we tested 31 other drug pairs that are known to interact and cause hyperglycemia (true positives, TP). Known drug-drug interactions are extracted (and manually validated) from textual monographs in DrugBank and the Medi-Span drug therapy monitoring system®. These sources are highly technical in nature or require paid access, making it less likely that ordinary health consumers would visit them and have the information bias their searches. Note that this is a less strict criterion than the pravastatin-paroxetine interaction, where we can guarantee that knowledge had not been available before the public release of the information. In order to compile a set of drug-pairs that are not associated with hyperglycemia, we create a negative set of 31 other drug pairs (TN) by associating drug-pairs with a randomly chosen adverse event, and removing any drug-drug-event pairings that are known to be associated based on external knowledge (DrugBank, Medi-Span, Drugs.com, UMLS or SIDER). We mapped the generic names for the drugs to their brand names, as we did with paroxetine and pravastatin, and searched for the presence of both drugs in the log data described above. We then performed the same type of log-based disproportionality analysis, including computing RR based on the expected counts from each drug in the pair.

Table S2 presents the results of this additional disproportionality analysis. The drug pairs are ranked in descending order by the average RR for the pair. We preserve the TP/TN label to show where in the list the TPs appear. If the log-based method performed perfectly, then all TPs would be ranked above all TNs. The results show that the majority of the drug-pairs identified as having a strong relationship with hyperglycemia are TP (i.e., 74% of the top half of the table is TP; two proportion *Z*-test (*Z*=-2.086, *p*=0.019)) and consequentially, the TN are least strongly related to hyperglycemia. In addition, if we assume that the pairings where the average RR values > 2 predict a TP (an RR value of two has been shown to be a meaningful threshold in previous work [17,18]), we estimate a false positive rate of 12.5% from the 62 pairings we examined. To further study performance across the range of threshold values, we construct a receiver operating characteristic (ROC) curve, shown in Figure 3. The area under the curve (AUC_{All}) is 0.8189, signifying strong performance in distinguishing TP from TN using the log data.

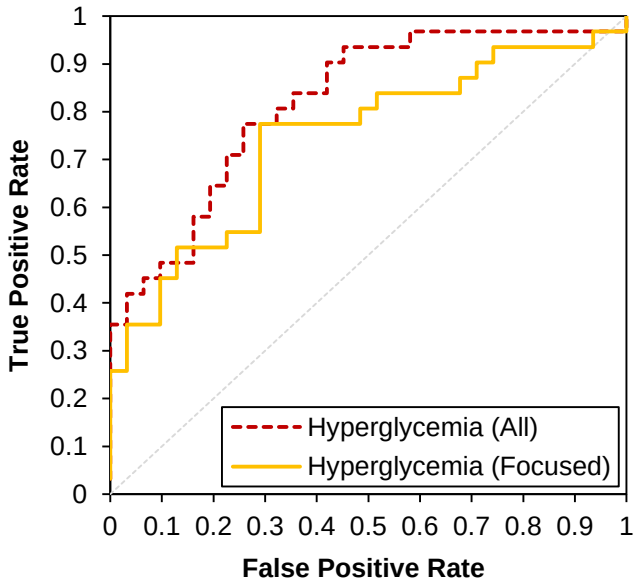

Fig 3. ROC curve for the identification of drug pairs known to be associated with hyperglycemia using search log data. Red (dashed) line denotes the performance when using all hyperglycemia-related terminology in our set. Yellow (solid) line denotes the performance a more narrowly focused set of symptoms strongly connected to hyperglycemia.

As the behavioral data for a large population used in the analyses are noisy we sought in our first phase of study to be inclusive with the use of a broad term list. We probed the sensitivity of the

results to reducing the set of terms to a more focused subset of terms restricted to synonyms of hyperglycemia and three primary hyperglycemic symptoms: polyphagia, polydipsia, and polyuria (and their related synonyms). The focused list appears in Table S3. The ROC curve for the more focused subset is shown in Figure 3. The value of $AUC_{Focused}$ is 0.7429, showing good performance in distinguishing TP from TN (i.e., 71% of the top-half of the ranking is TP; two proportion *Z*-test (Z=-1.815, p=0.035)). The performance with the focused subset of terms is lower than for the full set of hyperglycemia-related terminology, but not significantly so (Z=0.914, p=0.180)[19].

To understand which of the terms yielded the most benefit, we performed an ablation analysis of the symptoms/conditions. We iterate through sets of terms for each of the conditions/symptoms considered, starting with all terms, and remove successively sets of terms whose deletion leads to the largest decrement in the area under the ROC curve. Figure 4 shows the list of symptoms and conditions and the influence on AUC of removing each of them with this greedy procedure.

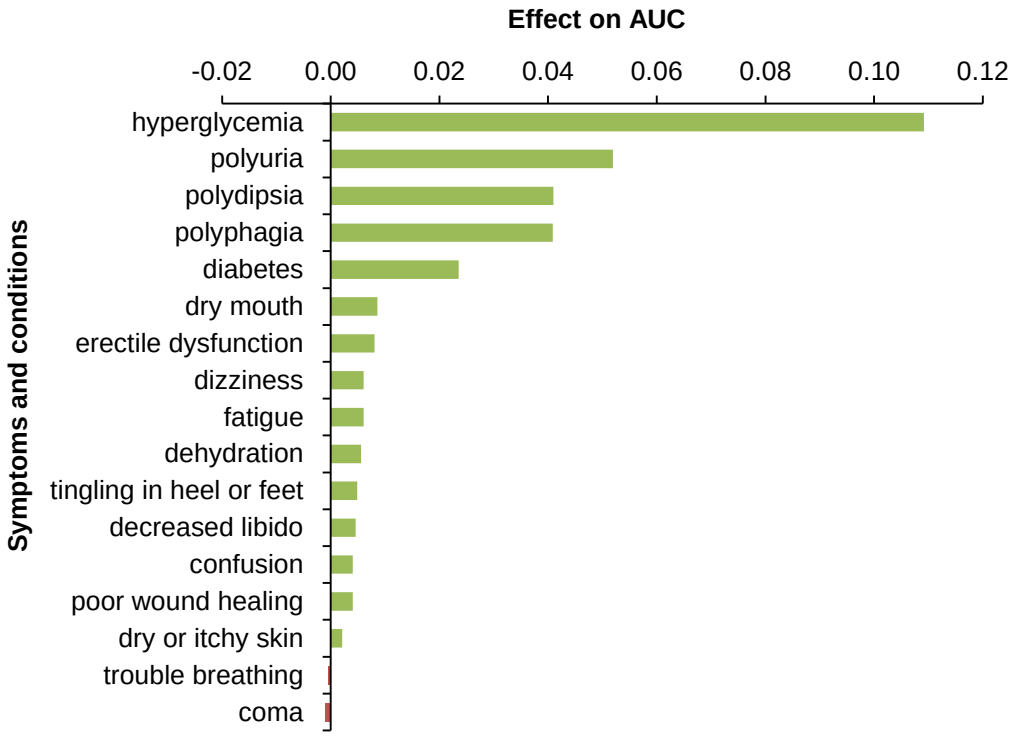

Fig 4. Influence of removing symptoms and conditions on the classification performance as measured by change in AUC_{All}.

Figure 4 shows that hyperglycemia (and its synonyms such as "high blood sugar") has the largest effect on AUC_{All}, followed by each of the three core hyperglycemic symptoms in the order polyuria, polydipsia, and polyphagia. The AUC remains high even when direct references to hyperglycemia (first bar in Figure 4) are removed ($AUC_{All-Hyperglycemia} = 0.7097$), illustrating the value of employing the pooled related symptoms and conditions for this classification task. The most influential additional terms outside of the core hyperglycemic symptoms (diabetes, dry mouth, etc.) are also known to be related to hyperglycemia. The terms become less strongly related as we move down the list. Note that removing "trouble breathing" and "coma" improves performance, signaling that these terms may add noise to the classifier.

Discussion and Conclusions

Overall, these findings demonstrate the potential value of the log analysis for identifying drug pairs linked to hyperglycemia and illustrates the generalizability of the method beyond just the pravastatin-paroxetine pairing. Given that the majority of the TPs can be identified from logs of search activity also provides validation for the set of terms used to identify hyperglycemia related searches (Table S1). Given the many pairs with little or no effect from the interaction also shows that the act of searching for multiple drugs is insufficient on its own to explain the heightened interest in hyperglycemia-related material.

The prolific use of Web search to pursue information can be likened to a large-scale distributed network of sensors for identifying potential side effects of drugs. There is a potential public health benefit in listening to such signals, and integrating them with other sources of information. We see a potentially valuable signal even though search logs are unstructured, not necessarily related to health, and can include any words entered by users. More in-depth analysis is needed to better understand biases and sources of noise in Web search logs. We particularly seek to understand potential non-pharmacological explanations for the trends observed in the log data. For example, confounding or hidden variables may play a role in boosting searches for terms associated with symptoms of hyperglycemia for the joint cohort. For example, demographic factors such as age and gender (not directly observable via log data) may contribute to the observed interactions. Psychological influences on health-seeking behavior may also play a role. For example, people prescribed paroxetine for anxiety may be more likely to focus on and

inquire about their symptomatology online than others, and this anxiety may rise more than others with the growing list of prescribed medications. We note that the data does not support this potential explanation; Figure 2 shows that there is less of an effect for those who search for paroxetine alone.

The pravastatin-paroxetine interaction was not known at the time the logs gathered (in 2010). Thus, the analysis we performed was similar to a prediction task. While further work is needed to explore the predictive value of signals from search logs, the methods and findings highlight the potential value of harnessing anonymized search logs captured by Internet services as complements to other signals for pharmacovigilance.[20] We believe that patient search behavior directly captures aspects of patients' concerns about sensed symptomatology and can complement more traditional sources of data for pharmacovigilance, including AERS and EHR data. We anticipate more sophisticated log-based detection of adverse events associated with medications, and that these will contribute to the faster identification of drug safety information.

References

1. D. C. Classen, R. Resar, F. Griffin, F. Federico, T. Frankel, N. Kimmel, J. C. Whittington, A. Frankel, A. Seger, B. C. James, 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff.* **30**(4), 581-589 (2011).
2. J. A. Johnson, J. L. Bootman. Drug-related morbidity and mortality: a cost of illness model. *Arch. Intern. Med*. **155**(18), 1949-1945 (1995).
3. J. Lazarou, B. Pomeranz, P. N. Corey. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *JAMA*, **279**, 1200-1205 (1998).
4. S. Fox, Health topics: 80% of internet users look for health information online. *Pew Internet and American Life Project* (2011).
   [online at http://pewinternet.org/~/media/Files/Reports/2011/PIP_HealthTopics.pdf, accessed 25 June 2012].
5. S. Schneeweiss, J. Avorn, A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J. Clin. Epidemiol.* **58**(4), 323-337 (2005).
6. A. Bate, S. J. W. Evans, Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol. Drug Saf.* **18**(6), 427-436 (2009).

7. P. M. Coloma, G. Trifiro, M. J. Schuemie, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, G. Picelli, G. Corrao, L. Pedersen, J. van der Lei, M. Sturkenboom. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? *Pharmacoepidemiol. Drug Saf.* **21**(6), 611-621 (2012).
8. J. Avorn, S. Schneeweiss, Managing drug-risk information: what to do with all those new numbers. *N. Engl. J. Med.* **361**(7), 647-649 (2009).
9. M. Hauben, A. Bate, Decision support methods for the detection of adverse events in post-marketing data. *Drug Discov. Today* **14**(7-8), 343-357 (2009).
10. P. M. Polgreen, Y., Chen, D. M. Pennock, N. D. Forrest. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* **47**, 1443-1448 (2008).
11. 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**, 1012-1014 (2009).
12. R. W. White, E. Horvitz, Cyberchondria: studies on the escalation of medical concerns in Web search. *Trans. Inf. Sys.* **27**(4), 23 (2009).
13. N. P. Tatonetti, G. H. Fernald, R. B. Altman, A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *J. Am. Med. Inform. Assoc.* **19**(1), 79-85 (2011).
14. N. P. Tatonetti, J. C. Denny, S. N. Murphy, G. H. Fernald, G. Krishnan, V. Castro, P. Yue, P. S. Tsao, I. Kohane, D. M. Roden, R. B. Altman. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin. Pharmacol. Ther.* **90**(1), 133-142 (2011).
15. A. Humphreys, MedAdNews 200 - World's Best-Selling Medicines. *MedAdNews* (2007). [online at http://en.wikipedia.org/wiki/List_of_bestselling_drugs, accessed 25 June 2012].
16. W. DuMouchel, Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Amer. Stat.* **53**, 177e90 (1999).
17. 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 Safety*, **25**(6): 381-392 (2002).

18. G. Deshpande, V. Gogolak, W. S. Sheila. Data mining in drug safety: review of published threshold criteria for defining signals of disproportionate reporting. *Pharmaceutical Medicine*, **24**(1): 37-43 (2010).
19. J. A. Hanley, B. J. McNeil. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, **143**, 29-36 (1982).
20. R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, C. Friedman. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin. Pharmacol. Ther.*, 91(6):1010-1021 (2012).

CONTRIBUTORSHIP STATEMENT

All authors planned the study and drafted and revised the paper. RW mined and analyzed the log data, and developed and evaluated the classifier. NT, NS, RA, EH advised on analysis and modeling strategies. NS provided data on known drug-drug interactions.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

COMPETING INTERESTS STATEMENT

There are no competing interests.

Table S1. List of hyperglycemia symptoms and conditions used in automated analysis.

appetite increase	feet tingling	poor wound healing
blood glucose high	frequent urinating	short of breath
blood sugar high	frequent urination	shortness of breath
blood sugar increase	glucose high	skin tingling
blurred vision	heel tingling	sleepiness
blurry vision	high glucose	sleepy
breathing difficulty	high blood glucose	slow healing
breathing trouble	high blood sugar	slow wound healing
breathless	hunger	thirst
breathlessness	hungry	thirstiness
coma	hyperglycemia	thirsty
confused	hyperglycaemia	tingling feet
confusion	impotence	tingling heel
decreased libido	impotent	tingling skin
decreased sex drive	increase blood sugar	tired
decreased sexual desire	increased appetite	tiredness
dehydrated	increased urination	trouble breathing
dehydration	itchy skin	xerostomia
diabetes	labored breathing	
diabetic	light headed	
difficulty breathing	lightheaded	
dizziness	light-headed	
dizzy	lightheadedness	
drowsiness	loss in weight	
drowsy	loss of weight	
dry mouth	low sex drive	
dry skin	polydipsia	
erectile dysfunction	polyphagia	
fatigue	polyuria	
fatigued	poor healing	

Table S2. Disproportionality analysis of true positive (TP) and true negative (TN) drug pairs with known association or dissociation with hyperglycemia. We include the analysis using both Expected (Drug 1) and Expected (Drug 2). **The pairs are ranked in descending order by the Average RR for Drug 1 and Drug 2.** Statistical significance for a one-tailed test performed using a Taylor series is denoted as: *** p < 0.001, ** p < 0.01, * p < 0.05. Column headers *a-d* have the same meaning as elsewhere in the article.

Label	***Drug 1***	***Drug 2***	***a***	***b***	***Expected (Drug 1)***				***Expected (Drug 2)***				***Avg. RR*** ↓
					c	***d***	***RR***	***95% CI***	***c***	***d***	***RR***	***95% CI***	
TP	dobutamine	hydrocortisone	43	150	645	5646	2.509***	1.755, 3.534	2595	36420	4.026***	2.836, 5.628	3.266
TP	dobutamine	triamcinolone	39	207	645	5646	1.649**	1.148, 2.325	3111	62047	3.758***	2.634, 5.256	2.703
TP	dobutamine	prednisolone	29	132	645	5646	1.923**	1.257, 2.868	1494	23115	3.399***	2.232, 5.044	2.661
TP	betamethasone	dobutamine	31	156	1515	26607	3.490***	2.332, 5.096	645	5646	1.739**	1.257, 2.868	2.615
TP	glipizide	phenytoin	273	1559	4817	25935	0.943	0.825, 1.075	6993	155357	3.890***	3.413, 4.434	2.417
TP	dobutamine	methylprednisolone	61	323	645	5646	1.653***	1.234, 2.188	3553	55652	2.958***	2.230, 3.876	2.306
TP	prednisolone	salmeterol	40	210	1494	23115	2.947***	2.070, 4.115	550	4656	1.612**	1.125, 2.269	2.280
TP	salmeterol	triamcinolone	69	437	550	4656	1.337*	1.016, 1.741	3111	62047	3.149***	2.421, 4.049	2.243
TP	betamethasone	terbutaline	58	328	1515	26607	3.106***	2.321, 4.101	721	5468	1.341*	0.997, 1.782	2.223
TP	dexamethasone	dobutamine	88	418	3424	42216	2.596***	2.048, 3.263	645	5646	1.843***	1.438, 2.344	2.219
TP	betamethasone	salmeterol	36	213	1515	26607	2.968***	2.908, 4.662	550	4656	1.431***	1.387, 2.272	2.200
TN	celecoxib	salmeterol	63	377	3943	69221	2.934***	2.227, 3.817	550	4656	1.415**	1.061, 1.864	2.174
TP	brimonidine	methylprednisolone	85	543	645	7496	1.819***	1.421, 2.309	3553	55652	2.452***	1.946, 3.089	2.136
TP	dobutamine	prednisone	103	497	645	5646	1.814***	1.440, 2.271	16191	191622	2.453***	1.975, 3.024	2.133
TN	heparin	lamivudine	45	195	4143	33666	1.875***	1.341, 2.579	229	2180	2.197***	1.533, 3.104	2.036
TP	hydrocortisone	salmeterol	44	244	2595	36420	2.531***	1.813, 3.471	550	4656	1.527**	1.084, 2.114	2.029
TN	ampicillin	tazobactam	20	105	923	12092	2.495***	1.504, 3.982	138	1088	1.501	0.883, 2.468	1.999
TP	prednisone	terbutaline	108	589	14191	191622	2.476***	2.015, 3.042	721	5468	1.391**	1.113, 1.727	1.933
TN	clopidogrel	famotidine	232	3974	71083	3653912	3.001***	2.628, 3.426	3431	48090	0.818	0.712, 0.937	1.910
TN	meropenem	methylprednisolone	37	276	274	3496	1.710**	1.174, 2.442	3553	55652	2.100***	1.469, 2.936	1.905
TP	formoterol	methylprednisolone	70	503	578	6372	1.534**	1.171, 1.989	3553	55652	2.180***	1.683, 2.791	1.857
TP	glucosamine	metformin	335	2378	4287	71126	2.337***	2.076, 2.631	14819	143737	1.366***	1.216, 1.348	1.852
TP	formoterol	triamcinolone	61	512	578	6372	1.313*	0.987, 1.726	3111	62047	2.376***	1.804, 3.087	1.845

TN	donepezil	sodium bicarbonate	58	429	3105	42237	1.839***	1.384, 2.410	1817	24120	1.795***	1.348, 2.356	1.817
TP	dexamethasone	salmeterol	72	419	3424	42216	2.119***	1.636, 2.714	550	4656	1.455**	1.110, 1.888	1.787
TP	methylprednisolone	salmeterol	59	413	3553	55652	2.238***	1.687, 2.927	550	4656	1.209	0.902, 1.602	1.723
TN	amitriptyline	bacitracin	68	576	4863	68827	1.671***	1.297, 2.152	776	11630	1.769***	1.362, 2.298	1.720
TP	brimonidine	dexamethasone	78	546	645	7496	1.660***	1.286, 2.124	3424	42216	1.761***	1.378, 2.228	1.711
TP	budesonide	dobutamine	26	147	1107	11609	1.855**	1.196, 2.792	645	5646	1.548*	0.995, 2.339	1.702
TN	oxcarbazepine	trimethoprim	34	331	1423	20977	1.514*	1.045, 2.140	919	16427	1.836**	1.264, 2.601	1.675
TP	prednisone	salmeterol	114	757	14191	191622	2.033***	1.669, 2.478	550	4656	1.275*	1.024, 1.578	1.654
TP	methylprednisolone	terbutaline	46	331	3553	55652	2.177***	1.58, 2.945	721	5468	1.054	0.760, 1.438	1.615
TN	metoprolol	piperacillin	39	266	7492	103085	2.017***	1.424, 2.800	222	1832	1.210	0.832, 1.728	1.614
TN	sulfamethoxazole	valproic acid	56	547	1202	21967	1.871***	1.413, 2.478	3673	47637	1.328*	0.998, 1.740	1.599
TP	brimonidine	prednisone	116	858	645	7496	1.571***	1.274, 1.938	16191	191622	1.600***	1.317, 1.944	1.586
TN	ketorolac	sucralfate	118	1084	2689	42126	1.705***	1.405, 2.070	1961	25530	1.417***	1.161, 1.718	1.561
TP	glucosamine	pioglitazone	107	876	4287	71126	2.027***	1.654, 2.483	5036	44450	1.078	0.877, 1.316	1.552
TN	metronidazole	ranitidine	319	3358	5886	103201	1.666***	1.481, 1.874	5076	74550	1.395***	1.240, 1.570	1.530
TN	clindamycin	diltiazem	224	1978	5028	70841	1.596***	1.386, 1.837	3677	43125	1.328***	1.150, 1.432	1.462
TP	formoterol	prednisone	104	826	578	6372	1.388**	1.109, 1.726	16191	191622	1.490***	1.209, 1.822	1.439
TP	epinephrine	prednisone	250	1739	3628	27910	1.106	0.963, 1.267	16191	191622	1.701***	1.489, 1.944	1.404
TN	dipyridamole	prednisone	83	661	622	6198	1.251*	0.977, 1.589	16191	191622	1.486***	1.176, 1.860	1.369
TP	budesonide	salmeterol	45	331	1107	11609	1.426*	1.028, 1.943	550	4656	1.151	0.825, 1.579	1.288
TN	hydrochlorothiazide	tazobactam	20	139	6254	60533	1.393	0.850, 2.189	138	1088	1.134	0.672, 1.846	1.264
TN	clindamycin	montelukast	141	1560	5028	70841	1.273**	1.066, 1.512	3729	50425	1.222*	1.022, 1.453	1.248
TN	lamotrigine	nystatin	116	1167	4523	58207	1.279**	1.050, 1.547	3257	36945	1.128	0.925, 1.365	1.203
TN	methylprednisolone	rosuvastatin	158	1774	3553	55652	1.395***	1.181, 1.647	6559	68810	0.934	0.790, 1.099	1.165
TP	budesonide	formoterol	115	1072	1107	11609	1.125	0.916, 1.373	578	6372	1.183	0.952, 1.456	1.154
TN	loratadine	nystatin	156	1635	7929	99985	1.203*	1.016, 1.417	3257	36945	1.082	0.912, 1.277	1.143
TN	hydroxychloroquine	prochlorperazine	86	743	2435	23664	1.125	0.892, 1.406	2028	19705	1.125	0.891, 1.406	1.125
TN	labetalol	sertraline	150	1429	1153	11795	1.074	0.896, 1.281	14153	157269	1.166*	0.982, 1.378	1.120
TN	ciprofloxacin	vecuronium	17	147	12497	126896	1.174	0.670, 1.900	187	1708	1.056	0.608, 1.752	1.115

TN	asparaginase	promethazine	7	61	110	911	0.950	0.392, 2.040	7386	77771	1.208	0.509, 2.520	1.079
TN	doxycycline	lovastatin	112	1375	7748	104419	1.098	0.901, 1.328	1903	24672	1.056	0.863, 1.283	1.077
TN	bumetanide	ondansetron	113	1023	944	9304	1.089	0.883, 1.333	5932	56427	1.051	0.860, 1.274	1.070
TN	amlodipine	amoxicillin	347	4261	9219	121885	1.077	0.963, 1.204	9062	117320	1.054	0.943, 1.179	1.065
TN	dextromethorphan	diazepam	49	564	1100	12943	1.022	0.752, 1.367	7589	96112	1.100	0.813, 1.463	1.061
TN	ibuprofen	ketamine	76	1055	16324	246924	1.090	0.858, 1.369	1801	25574	1.023	0.802, 1.291	1.056
TN	pantoprazole	promethazine	361	3838	5358	62571	1.098	0.983, 1.228	7386	77771	0.990	0.886, 1.105	1.044
TN	promethazine	sertraline	411	4751	6886	77771	0.977	0.880, 1.083	14153	157269	0.961	0.870, 1.064	0.969
TN	sitagliptin	tazobactam	17	103	5172	30603	0.977	0.568, 1.602	188	1088	0.955	0.544, 1.605	0.966
TP	budesonide	epinephrine	39	369	1107	11609	1.108	0.783, 1.536	3628	27910	0.813	0.576, 1.122	0.961

Table S3. Focused list hyperglycemia symptoms and conditions used in automated analysis

polydipsia
thirst
thirstiness
thirsty
polyphagia
appetite increase
increased appetite
hunger
hungry
polyuria
frequent urinating
frequent urination
increased urination
hyperglycemia
hyperglycaemia
high glucose
glucose high
high blood glucose
blood glucose high
high blood sugar
blood sugar high
increase blood sugar
blood sugar increase