PHARMACOEPIDEMIOLOGY AND DRUG SAFETY (2014)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/pds.3561

ORIGINAL REPORT

AERS spider: an online interactive tool to mine statistical associations in Adverse Event Reporting System

Igor Grigoriev[1], Wolfgang zu Castell[1], Philipp Tsvetkov[2] and Alexey V. Antonov[3]*

[1] *Helmholtz Zentrum München GmbH, Department of Scientific Computing, Neuherberg, Germany*
[2] *Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia*
[3] *MRC Toxicology Unit, Hodgkin Building, Leicester, UK*

ABSTRACT

Background Exploration of the Adverse Event Reporting System (AERS) data by a wide scientific community is limited due to several factors. First, AERS data must be intensively preprocessed to be converted into analyzable format. Second, application of the currently accepted disproportional reporting measures results in false positive signals.
Methods We proposed a data mining strategy to improve hypothesis generation with respect to potential associations.
Results By numerous examples, we illustrate that our strategy controls the false positive signals. We implemented a free online tool, AERS spider (www.chemoprofiling.org/AERS).
Conclusions We believe that AERS spider would be a valuable tool for drug safety experts. Copyright © 2014 John Wiley & Sons, Ltd.

KEY WORDS—pharmacovigilance; data mining; AERS; FDA; pharmacoepidemiology

Received 7 June 2013; Revised 28 October 2013; Accepted 3 December 2013

INTRODUCTION

The Adverse Event Reporting System (AERS) is a computerized information database designed to support the FDA's post-marketing safety surveillance program for all approved drug. Currently, the publicly available data covers more than one million reports submitted mainly from 1999 to 2011. The publicly available quarterly data files (http://www.fda.gov/) include patient demographic and administrative information, patient indications, drug information, reaction information, patient outcome information as well as some additional much less regular information.[1]

A variety of data mining procedures have been proposed recently to screen pharmacovigilance databases for potential "drug-to-Adverse Event" associations.[2–12] A commonly accepted principle to detect signals is to use different forms of disproportional reporting rates.[13,14] For example, proportional reporting ratio (PRR) is the ratio of the frequency of an Adverse Event (AE) in subpopulation of reports exposed to a drug and the frequency of an AE in the background population. Disproportional reporting principles are widely used for the detection of signals in pharmacovigilance databases at regulatory agencies and pharmaceutical companies in many countries.[7]

The analysis of AERS data is complicated by the need to extensively preprocess the data. In its publicly available form, AERS data is rather a collection of reports than a database. Drugs are usually reported by various brand and generic names. Several other issues, e.g. multiple report duplication, must also be resolved. Data preparation is absolutely required before any analytical steps are attempted and represents a serious technical challenge. These factors make it extremely complicated for the broad scientific community to obtain analytical results from AERS data on a particular issue. It is imperative to develop tools which could provide easy-to-use data mining access to AERS data.[15–19]

Several online tools are currently available providing analytical access to AERS data. FDAble seems to be the first web tool which provides analytical services

*Correspondence to: A. V. Antonov, MRC Toxicology Unit, Hodgkin Building, Lancaster Road, Leicester LE1 9HN, UK. E-mail: aa668@leicester.ac.uk

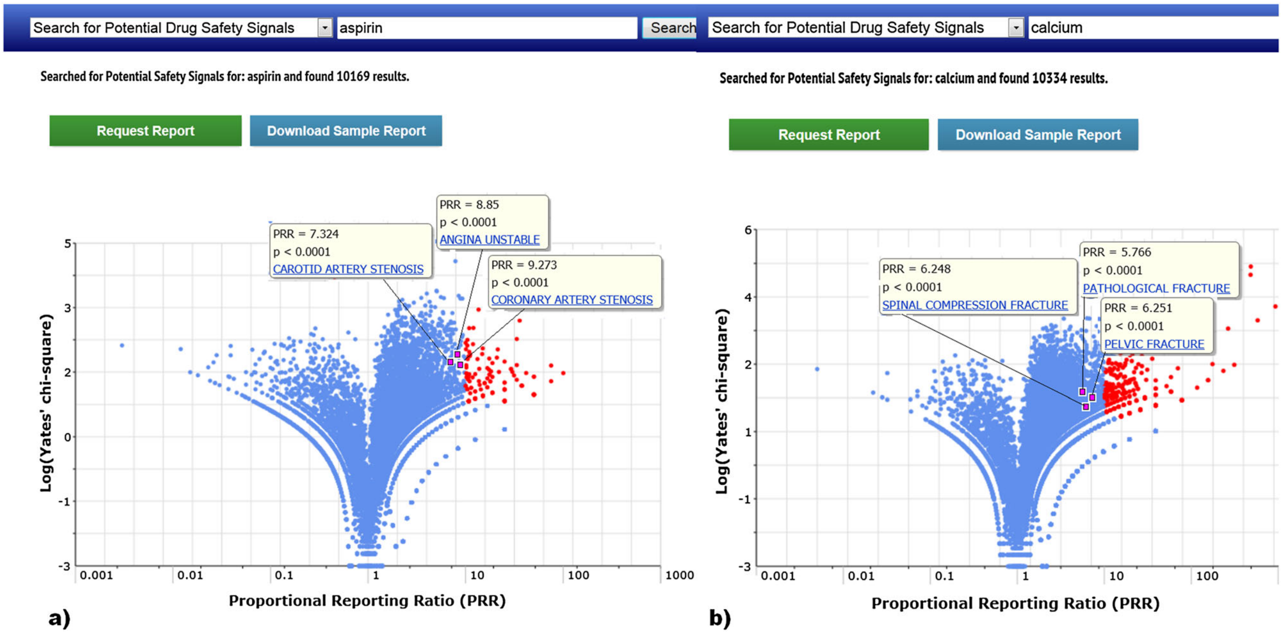
by computing PRRs.[15] FDAble provides an option to search for potential drug safety signals. FDAble reports PRR values for all AEs. Similar to FDAble, OpenVigil provides several options to query a drug-to-AE association providing PRR value.[17] However, both tools do not account for the intrinsic complexity of AERS data that practically results in reporting of false positive signals. Let us consider several illustrative examples related to "aspirin" and "calcium".

Brief inspection of results provided by FDAble for aspirin reveals the issue with false positive signals (see Figure 1a). Multiple cardio related and obviously false positive associated adverse reactions are reported with PRR > 5. OpenVigil provides similar results. For example, PRR reported for "aspirin-to-ANGINA UNSTABLE" association is 4.21 with the general conclusion that this drug-event-association is significant. For "calcium", FDAble reports all possible "fracture" outcomes with PRR from 3 up to 15. These simple examples demonstrate that practical application of disproportional reporting principles to detect signals from AERS data often leads to identification of AEs for which the drug is actually being used.

In our estimate, the share of false positive signals (with PRR > 2) can reach up to 50% for most of the drugs. The authors of FDAble are probably aware of the issue and impose very strict threshold (PRR > 10, red points in Figure 1) to consider the signal to be significant. However, it is clear that even such a strict threshold does not guarantee the absence of false positive signals while filtering many true positive associations (for example, for "calcium" the obvious false positive outcome "OPEN REDUCTION OF FRACTURE" has PRR > 15, and present among red points which are not freely available, FDAble is commercial tool). We used such obvious false positive cases to illustrate the issue. In general, there is no reliable way to figure out the false positive nature of the signal based only on PRR value, even imposing strict thresholds.

We would like to stress that in the majority of cases the false positive signals have nothing to do with well-known concerns regarding the quality of AERS data, such as duplication of reports, missing data, typographical errors, inaccuracy of reporting, variation in granularity and terminology used, underreporting and media influences.[2,4,7,18,19] The reason lies in the nature of AERS data and the need to consider multiple risk factors present in the data which are not accounted for by simple PRR measure.

The prior risk (the risk before drug usage) of certain adverse outcomes is extremely unequally distributed across various report groups in AERS data. Patients taking a particular drug, as a rule, have an incomparably high risk of certain adverse outcomes in comparison to a general population (a considerable proportion of people taking aspirin has a high risk of cardio adverse outcomes). In


Search for Potential Drug Safety Signals: aspirin

Searched for Potential Safety Signals for: aspirin and found 10169 results.

Search for Potential Drug Safety Signals: calcium

Searched for Potential Safety Signals for: calcium and found 10334 results.

Figure 1. False positive signal issue: a) FDAble search for potential "aspirin" safety signals. The figure lists a few obviously false positive signals with PRR > 5. In total, there are more than 50 cardio related signals with PRR > 5. b) FDAble search for potential "calcium" safety signals. The figure shows several "fracture" signals with PRR > 5. Almost all "fracture" related outcomes are reported with PRR > 3

many cases, the PRR merely indicates this fact, rather than the risk which is directly associated to the drug administration. To account for this, we propose a semi-automated strategy to correct prior-to-drug usage risk of the adverse outcome between drug and background subpopulations, by removing report subgroups marked by factors with apparently high risk of the adverse outcome. We refer to such factors (indications, age groups, gender, others drugs) as "mask" factors to point out that they can disguise the real strength of a drug-to-AE association.

In this paper, we present AERS Spider (www.chemoprofiling.org/AERS), a web tool which implements an iterative procedure to explore "drug-to-AE" association pattern. AERS Spider computes not only the PRR for a given drug and a given AE but also derives all other potential "mask" factors which are considerably associated (based on computed PRR value) to the AE. The user can selectively (based on his/her expert opinion) remove these factors from consideration and, thus, account for all (or at least available in the AERS data) risk factors which might disguise the real value of association between the drug and the AE. We provided examples supporting that the iterative procedure implemented in AERS spider improves hypothesis generation with respect to potential associations.

METHODS

Data pre-processing

Information from the AERS database is reduced to unified data model (Figure 2). The following attributes of each report are considered: date of the report, patent gender, age, indication(s), drug(s) and reaction(s). There are many reports with one or several missing attributes, i.e. no indication or gender is specified. Only reports which have all attributes available (at least one value specified) are considered. The one widely recognized issue with the AERS data is multiple duplication entries. To address this issue, reports from the same date, gender and age, which had the same indication(s) and reaction (s), the same or at least three common drugs were

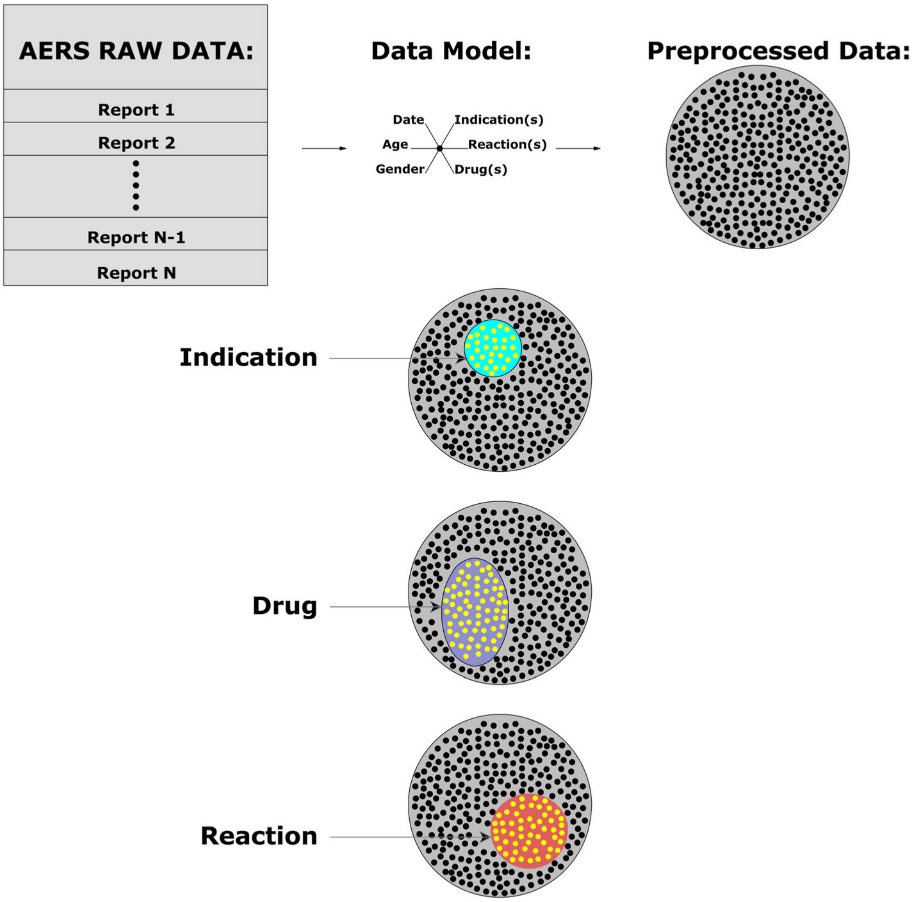


Figure 2. Each report from AERS raw data is converted into object with unified list of attributes (Date, Age, Gender, Indication(s), Reaction(s), Drug(s)). The name space for each attribute is unified. Each attribute (a drug, an indication, a reaction) defines a subpopulation of objects

DOI: 10.1002/pds

considered to be multiple entries and only one report is considered for the purpose of our analyses.

The key issue in data preparation is mapping of drug names from original reports to unified drug name space. In original AERS data, drugs are reported by multiple brand or multiple generic names including multiple mixtures. In addition, the same name can be written in multiple ways, which would be treated by the computer as different. We implemented a semi-automatic mapping procedure using files from DrugBank.[20] In the majority of cases, the DrugBank provides, for most FDA-approved drugs, a comprehensive list of brand and generic names, including names of drug mixtures where the drug (active compound) is present. Thus, original drug names from AERS data were remapped to unified name space of DrugBank IDs. In a few very specific cases, several DrugBank IDs were grouped into one. For example, one of such cases is related to "Insulin". DrugBank has several IDs which are related to slightly different (on molecular level) forms of Insulin. In most cases, from AERS report, it is not clear which one was actually used. For the purpose of our analyses, all such cases are mapped to a general ID "Insulin" and are considered as one drug.

Adverse reactions (as well as indications) in AERS are coded to terms in the Medical Dictionary for Regulatory Activities terminology (MedDRA). We created aggregate Adverse Reactions attributes which are merely a union of several Adverse Reactions. The name of aggregate Adverse Reaction corresponds to the most representative Adverse Reaction with extension "Sub Group" or "Group" at the end. "Sub Group" means more specific class while "Group" means a very non-specific set of adverse outcomes.

AERS spider

AERS spider is a free online tool (http://www.chemoprofiling.org/AERS/). Currently, AERS spider provides two query options, General Drug Query and Specific Drug Query. For the General Drug Query, the user needs to specify a drug and a background set from available options. As output, the list of putative signals ($PRR > 2$) is provided. The user can explore each reported "drug to adverse reaction" association using Specific Drug Query.

Specific Drug Query is an online interactive form which allows the user to remove/restore "mask" factors while exploring "drug to adverse reaction" association. As output, the user gets two tables. The first table provides the list of factors which were removed from consideration at previous steps and the option to restore each of them. The second table provides the list of current "mask" factors. By definition, all factors with PRR value higher than 3 are provided, i.e. all indications, age groups and other drugs which according to AERS data have strong ($PRR > 3$) association to adverse reaction are computed and provided. In some cases, the list may consist of up to several hundred factors. The user can either mark them for removal manually or remove them in a semi-automated way by setting the minimal PRR threshold value. After marking the "mask" factors, the user can update the signal. The list of "mask" factors is also updated. The procedure can to be repeated iteratively.

Figure 3 illustrates a general principle of our data mining strategy. In the first step, similar to other available tools, we compute PRR value for the drug and AE of interest. However, additionally we provide a possibility for the user to see all other potential "mask" factors (other drugs, indications, age groups, gender) which have high PRR value of association to the AE. We provide interactive possibility for the user to remove these factors from consideration. The new updated PRR is computed accounting for removed factors. The list of potential "mask" factors is updated as well. The procedure can be repeated iteratively unless all suspicious factors are removed.

RESULTS

Examples

To illustrate that our strategy resolves the false positive signal issue, we explore several obvious false positive associations which were used in the introduction. Table 1 summarizes PRR values for several "calcium-to-fracture" and "aspirin-to-cardio" cases computed based on AERS data. Rows 1a, 2a and 3a report raw PRR values without accounting for multiple "mask" factors. Rows 1b, 2b and 3b provide the PRR values for the same associations after removal of apparent risk "mask" factors.

The raw PRR (without accounting for multiple risk factors) for "Calcium-to-OPEN REDUCTION OF FRACTURE" association is 17.32 with 20 incidents of AE for patients administrating calcium (see Table 1, row 1a). The list of potential "Mask" factors proposed by the AERS spider (the list of potential "Mask" factors includes all "indications", "drugs" and "age groups" which are associated to "Reaction" with $PRR > 3$) includes multiple factors which point out to the group of patients with "BONE DENSITY" problems ("OSTEOPOROSIS", "Bone Density Conservation Agents"). Most of these patients have incomparably higher risk of various "fracture" outcomes. After removal a subgroup of reports marked with "Bone Density Conservation Agents" there is no one incidence of

DOI: 10.1002/pds

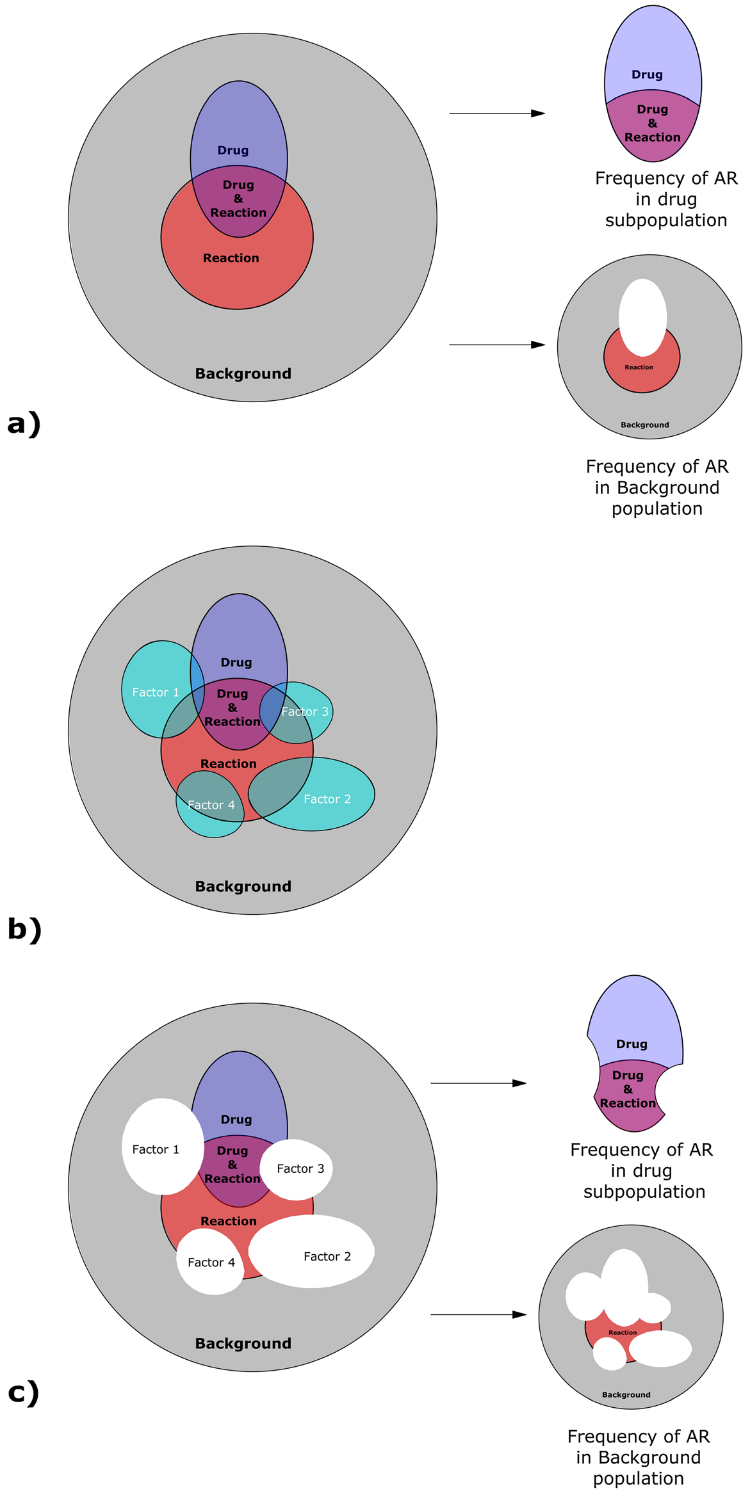


Figure 3. Principles of AERS spider. a) PRR: the frequency of the adverse reaction (AR) in the drug subpopulation is defined as a proportion of reports with AR. Proportional reporting ratio (PRR) compares the frequencies of AR in the drug and background populations. b) "Mask" factors (shown as Factor1, Factor2, Factor3 and Factor4) are report subgroups (indications, other drugs, age groups) with incomparable high prior risk of AR ("mask" factors are inferred by AERS spider by computing PRR of association between the factor and the AR, by definition all factors with PRR > 3 are reported by AERS spider as "mask" factors). The "mask" factors may substantially affect the estimate of frequency of the AR either in drug subpopulation (Factor1, Factor3) or in the background population (Factor2, Factor4). c) Data mining strategy: "Mask" factors are removed from consideration (based on high PRR value and the user expert opinion). Both frequencies of AR in the drug and the background populations are recomputed accounting for removed factors (shown as blank space). The updated value of PRR is supposed to reflect the risk of AR more objectively in relation to the drug administration. The procedure may be repeated iteratively

DOI: 10.1002/pds

Table 1. False positive signals used as negative control examples. The table reports raw PRR values (1a, 2a, 3a rows) without accounting for multiple risk "mask" factors. Rows 1b, 2b and 3b provide corrected signal values after removal of several obvious "mask" risk factors

#	Drug	Reaction	PRR	Number of reports with drug and reaction	Number of reports with drug	Number of reports with reaction	Number of reports background
1a	Calcium	OPEN REDUCTION OF FRACTURE	**17.32**	20	36 482	54	1 110 704
1b	Calcium	OPEN REDUCTION OF FRACTURE	**0.00**	0	18 751	9	1 023 140
2a	Calcium	COMPRESSION FRACTURE	**7.61**	128	36 482	623	1 110 704
2b	Calcium	COMPRESSION FRACTURE	**0.86**	1	12 050	80	833 266
3a	Aspirin	ANGINA UNSTABLE	**7.99**	300	79 845	770	1 078 929
3b	Aspirin	ANGINA UNSTABLE	**0.99**	4	20 329	153	771 553

"OPEN REDUCTION OF FRACTURE" in calcium report subpopulation (see Table 1, row 1b).

In the second "calcium-to-fracture" case (see Table 1, row 2a), in addition to apparent "BONE DENSITY" factors, one needs to remove less obvious ones. "CEREBROVASCULAR ACCIDENT" patients are known to have high risk of fractures due to the frequent falls at the start of the incident. "CANCER" patients are also known to have a higher risk of bone fractures.[21] 3After removal of above mentioned factors (for the full list, please, see supplementary table 1.2), the new recomputed PRR value for "Calcium-to-COMPRESSION FRACTURE" is 0.86 with the only one registered incident in "Calcium" subpopulation (see Table 1, row 2b).

For "Aspirin-to-ANGINA UNSTABLE" case, the list of potential "mask" factors proposed by AERS spider points out mainly to the group of patients with obvious cardio related problems, like, indications "MYOCARDIAL INFARCTION", "ANGINA PECTORIS" or patients who administered the drugs which obviously point out to the patient cardiovascular problems ("Antihypertensive Agents", "Antithrombotic Agents", "Nitrates and Nitrites"). After removal of these factors (for the full list see supplementary table 1.3), the new recomputed PRR value for "Aspirin-to-ANGINA UNSTABLE" is 0.99 with the only four registered incidents out of 20 000 reports in "Aspirin" subpopulation.

The false positive nature of the signals presented in Table 1 is obvious even for the non-expert. However, there are many cases which are not so transparent. Let us consider one of such examples. Glyburide is an oral antihyperglycemic agent used for the treatment of non-insulin-dependent diabetes. PRR value for "Glyburide-to-LACTIC ACIDOSIS" association is 3.91 with 52 incidences of "LACTIC ACIDOSIS" followed by "Glyburide" administration. The false positive nature of association is not obvious taking into account that some of the antihyperglycemic agents, like, "Metformin" are known to be associated with "LACTIC ACIDOSIS".[22] The fact actually could be used as an argument to consider "Glyburide" as a risk factor (because both drugs are used to treat the same indication and could potentially share a common mechanism which leads to a common side effect). In fact, "Metformin" is a mask factor. After removal of "Metformin" reports from consideration, the updated PRR value for "Glyburide-to-LACTIC ACIDOSIS" association is only 1.19 and indicates of no causative relation between administration of "Glyburide" and "LACTIC ACIDOSIS".

Examples presented above illustrate that false positive associations can be filtered out by removal of several related "mask" factors. This supports that our data mining strategy accounts for multiple risk factors and thereby reduces the number of PRR signals to more informative ones.

DISCUSSION

AERS data represent a large and extremely valuable resource to explore relations of multiple factors in respect to drug safety issues. The database is growing constantly by approximately 250 thousands new reports each year, covering new and established drugs on the market. It is imperative to develop tools which could provide easy-to-use data mining access to AERS data for drug safety experts.

The aim of this paper is mainly to address the issue with false positive signals resulting from practical application of disproportional reporting principles to AERS data without accounting for multiple risk factors. In the majority of cases, the phenomenon has nothing to do with widely recognized issues related to quality of AERS data and thus can be overcome by appropriate data mining strategy. We propose the data mining strategy implemented as online interactive web tool and illustrate that our tool is an efficient instrument to control false positive signals. We believe that AERS spider would be a valuable tool for drug safety experts for fast validation of various hypotheses related to drug safety issues. On the other hand, we would like to make crystal clear that provided hypotheses need further investigation.

CONFLICT OF INTEREST

The authors certify that there are no personal, commercial or academic conflicts of interest with any financial organization regarding the material discussed in the manuscript.

ETHICS STATEMENT

This study did not require ethics approval, as no human or animal subjects were involved.

KEY POINTS

- The analysis of AERS data for a wide public is complicated by the need to extensively preprocess the data.
- Currently available tools do not account for the intrinsic complexity of AERS data that practically results in reporting of false positive signals.
- We present AERS, a web tool which implements an iterative strategy to test false positive nature of detected signals.

REFERENCES

1. US Food and Drug Administration. Adverse Event Reporting System <http://www.fda.gov/cder/aers/default.htm>. FDA 2012.
2. Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N. Novel statistical tools for monitoring the safety of marketed drugs 1. *Clin Pharmacol Ther* 2007; **82**(2): 157–166.
3. Bate A, Evans SJ. Quantitative signal detection using spontaneous ADR reporting 1. *Pharmacoepidemiol Drug Saf* 2009; **18**(6): 427–436.
4. Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, Friedman C. Biclustering of adverse drug events in the FDA's spontaneous reporting system 1. *Clin Pharmacol Ther* 2011; **89**(2): 243–250.
5. Hauben M, Madigan D, Patadia V, Sakaguchi M, van Puijenbroek E. Quantitative signal detection for vaccines 1. *Hum Vaccin* 2010; **6**(8): 681.
6. Hochberg AM, Hauben M. Time-to-signal comparison for drug safety data-mining algorithms vs. traditional signaling criteria 4. *Clin Pharmacol Ther* 2009; **85**(6): 600–606.
7. Moore N, Hall G, Sturkenboom M, Mann R, Lagnaoui R, Begaud B. Biases affecting the proportional reporting ratio (PPR) in spontaneous reports pharmacovigilance databases: the example of sertindole. *Pharmacoepidemiol Drug Saf* 2003; **12**(4): 271–281.
8. Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: proceed with caution 1. *Pharmacoepidemiol Drug Saf* 2007; **16**(4): 359–365.
9. Trifiro G, Pariente A, Coloma PM, *et al*. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? 5. *Pharmacoepidemiol Drug Saf* 2009; **18**(12): 1176–1184.
10. Waller P, Beard K, Egberts T, *et al*. European commission consultation on pharmacovigilance 6. *Pharmacoepidemiol Drug Saf* 2008; **17**(2): 108–9.
11. Wang HW, Hochberg AM, Pearson RK, Hauben M. An experimental investigation of masking in the US FDA adverse event reporting system database 1. *Drug Saf* 2010; **33**(12): 1117–1133.
12. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions 1. *Sci Transl Med* 2012; **4**(125): 125ra31.
13. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001; **10**(6): 483–486.
14. Levine JG, Tonning JM, Szarfman A. Reply: The evaluation of data mining methods for the simultaneous and systematic detection of safety signals in large databases: lessons to be learned 4. *Br J Clin Pharmacol* 2006; **61**(1): 105–113.
15. Pratt LA, Danese PN. More eyeballs on AERS 1. *Nat Biotechnol* 2009; **27**(7): 601–2.
16. Anonymous. Making a difference. *Nat Biotechnol* 2009; **27**(297).
17. Bohm R, Hocker J, Cascorbi I, Herdegen T. OpenVigil-free eyeballs on AERS pharmacovigilance data 1. *Nat Biotechnol* 2012; **30**(2): 137–138.
18. Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP. The role of data mining in pharmacovigilance 3. *Expert Opin Drug Saf* 2005; **4**(5): 929–948.
19. Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance 4. *Br J Clin Pharmacol* 2004; **57**(2): 127–134.
20. Knox C, Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for "omics" research on drugs 1. *Nucleic Acids Res* 2011; **39**(Database issue): D1035–D1041.
21. Guise TA. Bone loss and fracture risk associated with cancer therapy. *Oncologist* 2006; **11**(10): 1121–1131.
22. Misbin RI. The phantom of lactic acidosis due to metformin in patients with diabetes 3. *Diabetes Care* 2004; **27**(7): 1791–1803.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version at the publisher's web-site.