Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU–ADR project

Paul Avillach,[1,2,3] Jean-Charles Dufour,[3] Gayo Diallo,[1] Francesco Salvo,[4,5] Michel Joubert,[3] Frantz Thiessard,[1,2] Fleur Mougin,[1] Gianluca Trifirò,[6] Annie Fourrier-Réglat,[2,4,5] Antoine Pariente,[2,4,5] Marius Fieschi[3]

▸ Additional supplementary files are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001083).

[1]LESIM, ISPED, University of Bordeaux, Bordeaux, France
[2]Pole de sante publique, CHU Bordeaux, Bordeaux, France
[3]LERTIM, EA 3283, Faculté de Médecine, University Aix Marseille, Marseille, France
[4]Department of Pharmacology, University Bordeaux, Bordeaux, France
[5]INSERM U657, University Bordeaux Segalen, Bordeaux, France
[6]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Correspondence to
Dr Paul Avillach, LESIM, ISPED, Université Bordeaux 2, 146 rue Léo-Saignat, Bordeaux cedex F-33076, France; avillach@mac.com

Received 5 May 2012
Accepted 1 November 2012
Published Online First 29 November 2012

ABSTRACT

Objectives The aim of this research was to automate the search of publications concerning adverse drug reactions (ADR) by defining the queries used to search MEDLINE and by determining the required threshold for the number of extracted publications to confirm the drug/event association in the literature.

Methods We defined an approach based on the medical subject headings (MeSH) 'descriptor records' and 'supplementary concept records' thesaurus, using the subheadings 'chemically induced' and 'adverse effects' with the 'pharmacological action' knowledge. An expert-built validation set of true positive and true negative drug/adverse event associations (n=61) was used to validate our method.

Results Using a threshold of three of more extracted publications, the automated search method presented a sensitivity of 90% and a specificity of 100%. For nine different drug/event pairs selected, the recall of the automated search ranged from 24% to 64% and the precision from 93% to 48%.

Conclusions This work presents a method to find previously established relationships between drugs and adverse events in the literature. Using MEDLINE, following a MeSH approach to filter the signals, is a valid option. Our contribution is available as a web service that will be integrated in the final European EU–ADR project (Exploring and Understanding Adverse Drug Reactions by integrative mining of clinical records and biomedical knowledge) automated system.

INTRODUCTION

In the area of drug safety, information sharing could enhance the current spontaneously reported information on adverse drug reactions (ADR), as reporting is far from optimal. In pharmacovigilance, a drug safety signal is defined as an unexpected association between a given event and a given drug.[1] According to the WHO definition, it refers to reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented. Usually more than a single case report of ADR is required to generate a drug safety signal, depending on the seriousness of the event and the quality of the information. Sources of drug safety signals are various: from safety data from clinical trials, to spontaneous reports of suspected ADR in pharmacovigilance systems, publication of case reports, case series and results of post-marketing observational studies. It is estimated that only 4% of ADR are reported through the current spontaneous reporting channels.[2] Drug safety signals may be detected too late, as was recently debated after the Vioxx withdrawal due to cardiovascular toxicity.[3] It has been recognized that additional complementary systems for signal detection are necessary. This could profit from the wide availability of healthcare databases throughout Europe.[4] Applying data mining techniques on these databases for signal detection could overcome the underreporting limitation of spontaneous reporting systems and may detect signals earlier. From this rationale, the EU–ADR project (Exploring and Understanding Adverse Drug Reactions by integrative mining of clinical records and biomedical knowledge, available at: http://www.euadr-project.org) was funded by the European Commission and started in February 2008. The aim of this project is to design, develop and validate a computerized system to process data from eight electronic health record (EHR) databases and biomedical knowledge databases for the early detection of safety signals.

The EU–ADR project platform incorporates a network of EHR databases from different European countries. The eight databases involved in the EU–ADR project contain information stemming from the medical files of more than 20 million European citizens.[1] Those databases are heterogeneous in structure (general practice vs claims databases) and available information (ie, EHR, hospital discharge diagnoses, death registries, laboratory values).[5] Automated signal generation has currently been implemented by applying data mining techniques on medical data from these eight databases.[6] All the identified drug–event pairs represent potential signals that will be thereafter substantiated by a computer-assisted exploration of biological plausibility in the context of current biomedical knowledge to reduce the false positive signals. The list of potential signals will be assessed by automatically investigating feasible biological paths connecting the drug and the adverse reaction involved in the proposed signal. In other words, there is a connection between a drug (or drug metabolites) and an event through common genes/proteins.[7] The ultimate goal of this task is to integrate all the developed tools through web services to create a unique environment where the end user may find the most significant scientific evidence for a set of signals.

To cite: Avillach P, Dufour J-C, Diallo G, *et al*. *J Am Med Inform Assoc* 2013;**20**:446–452.

The potential generation of a very high number of signals constitutes a major concern in the EU–ADR project as this could hamper the rapid identification of those of most importance in a public health matter. Therefore, various additional sources of information will be used to reduce the number of spurious signals and to identify emerging signals among those generated through the exploitation of the EHR databases. The preexisting level of knowledge of each identified drug/adverse event association will be evaluated by studying the reporting of corresponding ADR in the biomedical literature. Various techniques have been developed to automate knowledge extraction for providing appropriate information.[8] The MEDLINE database from the National Library of Medicine (NLM) is a leading source of scientific information. Extracting articles related to ADR from MEDLINE using a medical subject headings (MeSH) approach has been described previously.[9–11]

Objective

The aim of this research was to determinate automatically if a given drug/adverse event association was already known as an ADR reported in the literature.

METHODS

To determine if an ADR has already been described in the literature, we first defined a pattern for the queries used to search in MEDLINE, and then determined the threshold number of extracted publications needed to confirm or reject that ADR (ie, causal drug–event association). Our contribution is available as a web service that will be integrated in the final EU–ADR automated system.

We used MEDLINE as a knowledge source. The relevant publications were defined as those in which the drug and the adverse event of interest were co-occurring in the same relevant citations. The MeSH thesaurus is a controlled vocabulary produced by the NLM and used for indexing, cataloging, and searching for biomedical and health-related information and documents. NLM indexers select the most appropriate MeSH descriptors and subheadings (or qualifiers) to resume the full content of an article after reading the full text. This professional indexation enhances the quality of information retrieval. To automate the search of publications concerning ADR, we used the following MeSH-based approach: (1) map the events to MeSH; (2) map the drugs to MeSH; (3) construct the query with mapped MeSH terms and filter the results by publication type; and (4) determine a threshold number of publications to confirm or inform the knowledge of the drug/event association in the literature. This last step was performed by testing the method on an expert-built validation set including true positive (ie, known signals) and true negative drug/adverse event associations.

Resources

We downloaded (via PubMed) and imported in a database a subset of MEDLINE including all the citations with the 'adverse effects (AE)' MeSH subheading. For each citation, we gathered the following information: PMID, MeSH descriptors, subheadings, substances, and date of creation of the citation. We used the 2009AA version of the unified medical language system (UMLS), a biomedical terminology integration system handling more than 150 terminologies,[12] including MeSH.

Which events to monitor?

When using data mining to detect signals in EHR databases, either a drug or an event-based approach can be adopted. The EU–ADR project used an event-based approach in which a limited set of specific events are inspected for their association with all available drugs in the EHR databases participating in the project. One of the challenges in the event-based approach for signal detection through data mining on EHR databases is the identification of events that are most important in pharmacovigilance and thus warrant priority for monitoring.[13] This ranked list comprised 23 adverse events. The top-ranking events were: cutaneous bullous eruption (BE), acute renal failure, acute myocardial infarction (AMI), anaphylactic shock (AS), and rhabdomyolysis (RHABD). Because of its complexity, an additional event, upper gastrointestinal bleeding (UGIB), was selected to test the method. A definition of each event was written by medical specialists to facilitate the identification of the medical concepts defining it.

Concept selection

To make explicit and to harmonize the definition of the events among the different databases, a shared semantic foundation for the eight databases was built.[14] Its constituents are the UMLS concepts grouping together terms from different terminologies with the same medical meaning. The aim was to provide researchers with a formalized and standardized list of medical concepts concept unique identifier (CUI) and associated terms to be used for identifying the events investigated in their respective EHR databases, and providing the same definition of the event to filter and substantiate the generated signal.

Mapping of events

For each of the six events initially investigated, the corresponding UMLS concepts were listed with their CUI. We used the metathesaurus of the UMLS to get MeSH codes and the preferred terms in English. If the concept had no direct mapping in MeSH, we used the 'restrict to MeSH' algorithm[15] to get the nearest MeSH codes.

Events knowledge in MEDLINE citations

The MeSH heading field is used in a MEDLINE citation to describe the event focus by the corresponding article. Topical subheadings (or qualifiers) are used to narrow the specific focus of a main MeSH heading to a particular aspect of the subject. The subheading 'chemically induced' is used to qualify the adverse events in drug safety articles.

Mapping of drugs

In the EU–ADR project, all the databases code their drugs using the anatomical therapeutic chemical (ATC) classification except for one (Qresearch uses BNF codes, which they have mapped to ATC). Those ATC codes also need to be mapped into MeSH to query MEDLINE. As the ATC classification is not included in the UMLS, we used a mapping of ATC to UMLS CUI concepts (see Acknowledgments) that allowed the mapping from the ATC codes to the MeSH terms. We used the UMLS to find synonyms if a CUI concept (of an ATC code) did not have a direct mapping to a MeSH term.

Drugs knowledge in MEDLINE citations

Despite the richness of concepts contained within the many main headings (descriptors, MeSH terms) that comprise the cemicals and drugs category of MeSH, the rapid expansion of knowledge about chemicals and proteins requires a correspondingly rapid-changing supplementary vocabulary. Main headings are updated annually; supplementary concept records (SCR), discussed below, are updated on a daily basis.[16]

The many substances that are described in the literature need to be named with a controlled vocabulary. Some of them, like aspirin or lisinopril, are MeSH terms, but most substances are not part of the MeSH trees of interrelated MeSH terms. Instead, these substances are part of a separate controlled vocabulary called 'supplementary concepts', which are mainly chemical and protein concepts. Each of these 'supplementary concepts' or 'substance names' is described in a 'SCR', of which there are more than 150 000. Following a procedure that began in 1996, articles about the action of a drug or chemical are indexed both under the MeSH term for the drug or chemical and that for the pharmacological action being studied[16] (see example below).

In MEDLINE citations, the subheading 'AE' is used to qualify the drugs in drug safety articles and this can be used only to qualify drugs mentioned using MeSH headings (in the MeSH terms field of the citations). SCR terms do not have any subheading. However, citations with drugs SCR in the 'substances' field can have, in the MeSH terms field, MeSH headings for the drug's pharmacological actions, with the appropriate subheadings.

This situation is illustrated in the following example: moxifloxacin is a SCR with the MeSH heading 'anti-infective agents' as pharmacological action. Aspirin is a MeSH heading with four pharmacological actions: 'anti-inflammatory agents, non-steroidal', 'platelet aggregation', […]. Both drugs were mentioned in the case report entitled: 'Drug points: tachycardia associated with moxifloxacin' (PMID:11141146). In this case report, a 49-year-old man was prescribed moxifloxacin because of sinusitis and developed tachycardia as an adverse reaction to moxifloxacin. It is also reported that the patient took aspirin for the treatment of headache (with no adverse effect). In the citation, two substances are indexed: moxifloxacin and aspirin. MeSH terms are:

- Tachycardia/chemically induced*
- Anti-infective agents/AE*
- Sinusitis/drug therapy
- Anti-inflammatory agents
- Non-steroidal/therapeutic use
- Aspirin/therapeutic use
- Headache/drug therapy, […]

Moxifloxacin can only be indexed in the 'substances' field and not in the MeSH terms field (it is a SCR), differently from aspirin, which is a MeSH heading with the 'therapeutic use' subheading. The pharmacological action of moxifloxacin, 'anti-infective agents', has the subheading 'AE' so the AE knowledge can be linked to the appropriate drug (moxifloxacin and NOT aspirin).

Query construction

To retrieve the appropriate publications, we used the co-occurrence of four elements in a citation: the drug (from 'substances' OR 'MeSH heading' fields), the adverse effect and the two subheadings, AE and chemically induced. We only took into account drugs from the 'substances' field if their pharmacological action was qualified by the subheading 'AE' (see the previous example). In this case, the pharmacological action was an additional co-occurring element (the fifth one). This was a key point of our method: always having a link between an adverse event and a drug in the context of drug safety and not just a co-occurrence in a MEDLINE citation.

Filtering the results by publication type

We considered all the following types of publication as non-contributive to describe ADR: addresses, bibliography, biography, comment, dictionary, directory, duplicate publication, editorial, Festschrift, government publications, historical article, in vitro, interactive tutorial, interview, introductory journal article, lectures, legislation, patient education handout, periodical index, published erratum and retracted publication.

Evaluation

An evaluation with two phases was conducted. The first phase consisted of using validations sets of true positive and true negative signals to evaluate the sensibility and specificity of the system. Second, a manual search was performed in MEDLINE by using PubMed by two pharmacovigilance experts on four true positive signals, constituting the gold standard, to evaluate the recall and precision of the method.

Constituting the validation sets

The first drug–event set consisted of true positive associations. These associations constitute well-recognized safety signals. True positive signals consist of drug–event combinations for which a signal was generated and confirmed in the past. The combinations corresponding to this definition were identified through the following procedure:

- First, a search was performed through a spontaneous reporting database (AFSSAPS, the French pharmacovigilance database) for drugs associated with the selected events using the reporting OR in the case–non-case method. This step was performed to provide an orientation for the search in the literature and the websites of the regulatory agencies.
- In a second step, we searched evidence in the literature (MEDLINE and EMBASE) and the website of the regulatory agencies (US Food and Drug Administration (FDA), AFSSAPS). The confidence in the status of a signal increases with the addition of evidence provided there is no study questioning its reliability.

Therefore, the signals selected for the constitution of the true positive set are mostly referring to historical and very well known associations for which no doubts remain.

The second set consists of true negative signals, which are defined as drug–event combinations for which no evidence has ever been generated until the time of the study. The combinations corresponding to this definition were identified through the following procedure:

- We first searched in the French pharmacovigilance database for drugs that are not associated with the selected events using the reporting OR in the case–non-case method. This step was performed to provide an orientation for the search in the literature and the databases of the regulatory agencies.
- In a second step, we searched in the literature, the Thomson Reuters Micromedex database, the websites of the regulatory agencies, the FDA adverse event reporting system spontaneous reporting database. The confidence in the status of a signal increases with the addition of evidence provided there is no study questioning its reliability. Therefore, the signals selected for the constitution of the true negative set are mostly referring to drugs that were marketed for a long time and for which no question remains about a potential association to an event of interest. No data providing more evidence than a single case report had to be found for these combinations in the literature. If a signal selected as true negative was the focus of a report in the Thomson Reuters Micromedex database (Micromedex Healthcare Series, Greenwood Village,

Colo: Thomson Reuters (Healthcare) Inc., updated periodically) it was definitely rejected from the set.

Constitution of the gold standard by pharmacological experts

To study the performances of our automated search system, we considered as a gold standard a method that had two components: (1) a traditional PubMed search in MEDLINE associating drug names and event names, with a restriction on the publication type (the same as described before) and upper date limit identical (15/02/2010) with the date of extraction used by our method; (2) an expert assessment of the relevancy of the retrieved references. After the literature search was performed (25/09/2010) for each of the selected drug/event pairs described above, the articles retrieved were evaluated independently by two pharmacovigilance experts to determine if the responsibility of the drug mentioned in the occurrence of the event of interest appeared clearly in the title and/or abstract. The retrieved articles judged relevant by the two experts constituted the true positives identified by this gold standard method. Four drug event pairs from the true positive validation set were analysed: (1) bezafibrate and RHABD; (2) ceftriaxone and AS; (3) lamotrigine and BE; (4) sildenafil and AMI. The PubMed interface was used as they usually do to search if there is an association between a drug and an event.

RESULTS

Validation sets

A list of five drugs for the true positive set and of five drugs for the true negative set was constituted for each of the six events. Overall, a list of 61 pairs of drug/events was available to test our method.

Mapping of the drugs

All the ATC codes were mapped successfully to the MeSH (MeSH descriptors or SCR) (precision 100%) with the 'ATC to CUI' (see Acknowledgments) and the 'CUI to MeSH' mapping (UMLS); 83% of drugs were MeSH headings and 17% were SCR.

Retrieved publications

The number of retrieved publications for each of the 61 pairs of the validations sets is available for consultation in supplementary appendix 1 (available online only). The specificity and sensitivity for each threshold are given in table 1. The receiver operating characteristic performance of the model is graphically presented in figure 1.

Evaluation by the pharmacologist experts

For the different drug/event pairs selected, the precision of the automated search ranged from 93% to 48% and the recall ranged from 64% to 24% (see table 2).

Table 1 Overall sensibility and specificity for each threshold in the validation sets

Threshold	≥1	≥2	≥3	≥4	≥5	≥10
Specificity (%)	93	93	100	100	100	100
Sensitivity (%)	97	94	90	81	81	65

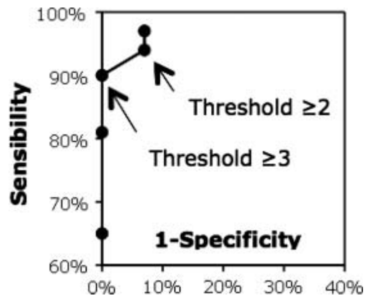

Figure 1 Receiver operating characteristic curve of the model.

Web service

A web service implementing our method has been launched. Its endpoint is accessible at http://lesim.isped.u-bordeaux2.fr/axis2/services/UB2_EUADR?wsdl.

The input is the ATC code of the drug and the coded name of the event (eg, UGIB for UGIB). The web service returns an XLM file, which conforms to the XSD schema of the EU–ADR project (http://bioinformatics.ua.pt/euadr/euadr_types.xsd) and includes: (1) a list of publications (PMID), in which the drug–event pair has been reported, classified by publication type; (2) A URL to build a chart to visualize the number of times the adverse event and the drug are seen together in MEDLINE in the context of ADR by year (see examples in figures 2 and 3); (3) the total number of citations retrieved; (4) a global score of the system to confirm (1) or inform (0) the knowledge of this association in the literature. A workflow calling the web service can be implemented within the Taverna workbench.[17] Figure 4 is a screenshot of the web service workflow in Taverna. The EU–ADR project is developing an end user platform that will integrate all web services and workflows. We have embedded this web service in a complete Taverva workflow available online at: http://www.myexperiment.org/workflows/2280.html.

DISCUSSION

Findings

The sensitivity and specificity measures of publication retrieval in the automated search show a good performance. When using a threshold of three or more extracted publications, the method presented a sensitivity of 90% and a specificity of 100%. This result is comparable with a previous work with the MeSH and subheading approach.[9] In Garcelon *et al*,[9] the most relevant threshold was three or more with a sensitivity of 65% and

Table 2 Precision and recall for nine true positive signals

Signals	Precision (%)	Total test positive	Recall (%)	Total gold standard positive
Lamotrigine and BE	93	15	64	22
Furosemide and BE	86	21	58	31
Bezafibrate and RHABD	82	17	54	26
Atorvastatine and RHABD	81	21	31	54
Valproic acid and BE	80	10	57	14
Ceftriaxone and AS	75	4	33	9
Diclofenac and AS	71	14	34	29
Sildenafil and AMI	65	23	52	29
Pravastatine and RHABD	48	21	24	42

AMI, acute myocardial infarction; ARF, acute renal failure; AS, anaphylactic shock; BE, cutaneous bullous eruption; RHABD, rhabdomyolysis.

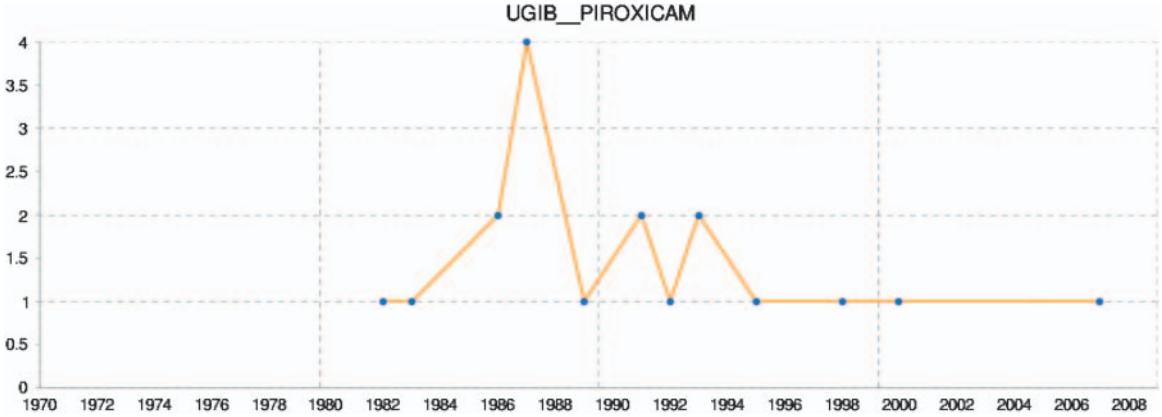

Figure 2 Number of times upper gastrointestinal bleeding (UGIB) and piroxicam are seen together in MEDLINE in the context of adverse drug reactions by year. This figure is only reproduced in colour in the online version.

specificity of 97%, but the pharmacological action knowledge to substance MeSH terms was not considered in their model. Their research work, conducted within the French VigiTermes project, aimed at developing a platform to improve the documentation of pharmacovigilance case reports for the pharmaceutical industry and regulatory authorities.[11] Another approach[18] in the VigiTermes project, using MeSH terms with extensions, had a recall of 67% and a lower precision of 29%. Zeng and Cimino carried out an automated disease–chemical knowledge extraction based on the co-occurrence of UMLS concepts.[8 19] The results estimated a sensitivity of 93%. In order to determine drug–adverse event (ADE) relationships, Wang and colleagues[20] developed a machine learning approach to extract knowledge from PubMed for the purpose of supporting pharmacovigilance and decision support. The approach includes a manual step for term synonyms searching and a classification algorithm for discarding articles denoting drug–ADE relationship from others by exploiting the text of the article in PubMed. They obtained a sensitivity of 90% and a specificity of 78% when testing myocardial infarction with 38 drugs. This method involves determining a new threshold for each adverse event. Our approach has been tested on a wider set of adverse events with the same threshold, to enable scaling with new adverse events. However, we do not perform any text natural language processing as this is done by another task[21] within the EU–ADR framework. Another study from Shetty and Dalal[22] developed a lasso-based statistical document classifier using MeSH terms identifying relevant articles with 71.4% sensitivity and 40.7% positive predictive value.

Limitations

Our approach offers the opportunity to determine automatically if an ADR has already been described in MEDLINE. However, the causality relationship between the drug and an event can be confirmed only by an expert reading the full text article. Because specific subheadings and keywords are used in the queries that are automatically built, the automated search may be more specific than a manual query. Only publications already indexed with MeSH could be detected by our method.

Despite the interesting results, the semantics is partly limited because it is based on librarian indexing and deals only with the co-occurrence of MeSH terms and subheading usage. Co-occurrence of 'a MeSH given drug' (or its pharmacological action)/AE and 'a MeSH given event'/chemically induced is always considered as a 'causes' semantic relationship by our method. Therefore, a semantic relationship between concepts is not fully refined, for example, it does not treat negation between concepts. Multiple co-occurrence situations (eg, the same citation with several events and several drugs or pharmacological actions mentioned) can also be problematical because our method considers the combinatorial relationships between all drugs and events of the citation. As an example, both

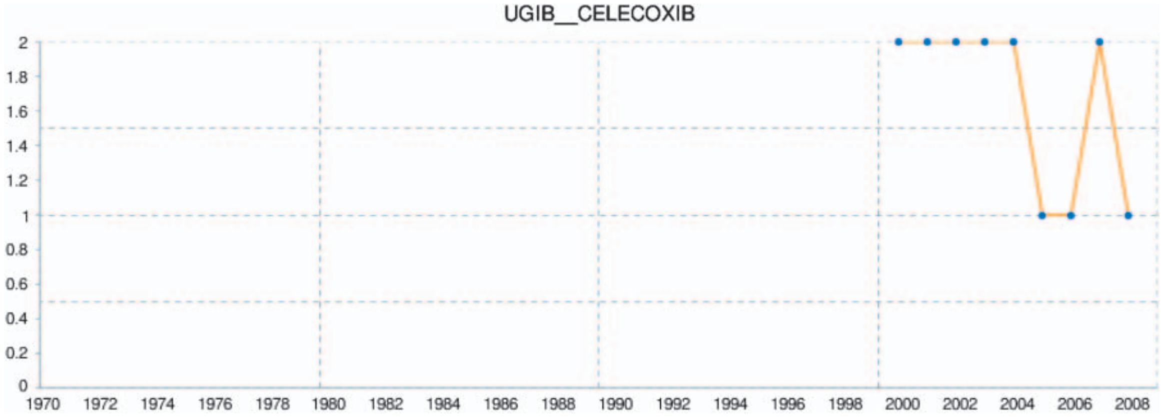

Figure 3 Number of times upper gastrointestinal bleeding (UGIB) and celecoxib are seen together in MEDLINE in the context of adverse drug reactions by year. This figure is only reproduced in colour in the online version.

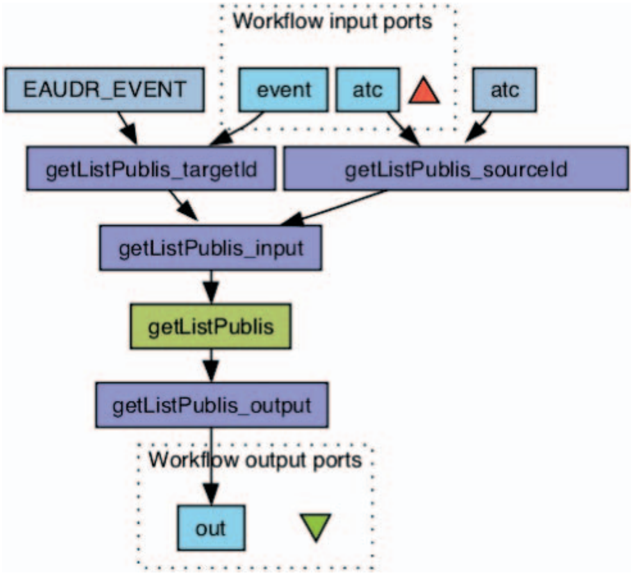

Figure 4 Web service workflow in Taverna. This figure is only reproduced in colour in the online version.

relations of the MEDLINE citation with two drugs: atorvastatin/ 'adverse effect', acetylsalicylic acid/'adverse effect' and one adverse effect: gastrointestinal bleeding/'chemically induced'.

Validation set

As the definition of true negative signals is based on the existing knowledge at the time the signal is investigated, it has to be understood that signals that are currently considered as true negatives could become positive signals in the future based on new evidence on drug safety. The set of true negative signals can thus only be considered as such at the time and date of its constitution.

Gold standard

The evaluation of the performances of the automated method for literature selection that we proposed was performed using nine drug/event pairs that presented various characteristics with regard to the type of the drug, the nature of the adverse event and the amount of papers identified through a traditional literature search. In three of these, the performances of the automated methods appeared interesting. Literature search and knowledge identification on the potential responsibility of a given drug in the occurrence of a given event is a mainstay of everyday pharmacovigilance work. To be performed rigorously, this identification process happens to be very time-consuming using traditional methods, mostly because a potentially very important proportion of irrelevant papers are among the retrieved references. In her/his everyday work, the objective of the pharmacovigilance expert is not to perform an exhaustive review of the existing knowledge concerning a drug/event association, but to identify rapidly relevant information that will allow her/him to give to a clinician an answer concerning the therapeutic management of a patient who presents with an adverse event for which drugs are suspects. Therefore, its objective when performing a literature search is to identify some, but not necessarily all, relevant papers on the topic, if possible from a limited number of references to be as time-efficient as possible. In this sense, the automated method we propose seems to be interesting: it succeeded in identifying relevant papers in all the studied situations and provided reference lists that were shorter that those obtained using a traditional search method (considered as a gold standard).

Further work

We want to analyze the time period between the publications: a short and productive period is more significant than a long period with only a few publications. The overall number of publications can be the same in those two cases. We also plan to analyze the different publication types and the major/minor character of the MeSH descriptors and subheadings to evaluate if they could enhance the filtering process. Another research group of the EU–ADR project from the Erasmus University MC is also using MEDLINE as a knowledge source to filter signals following a natural language processing approach. The results of the two approaches will be compared.

CONCLUSION

This work presents a method to find previously established relationships between drugs and adverse events in the literature. Using MEDLINE, following a MeSH approach to filter the signals, is a valid option. Using a threshold or three or more publications containing adverse event and drug co-occurrences, the extracting method shows an enthusiastic result on the studied couple of drug/adverse drug associations with a sensitivity of 90%, a specificity of 100% and moreover a precision of up to 93%.

Acknowledgements The authors wish to thank the NLM for making UMLS available free of charge. They also wish to thank Erik van Mulligen from the Department of Medical Informatics, Erasmus University, Medical Center, Rotterdam, The Netherlands for providing the ATC to CUI mapping table.

Contributors PA, JCD, GD, MJ, FT, FM and MF participated in designing the method and its evaluation. GT, AFR and AP participated in the design and elaboration of the validation sets. All authors participated in the elaboration and correction of the manuscript.

Funding This work was supported by the European Commission Seventh Framework Programme (FP7/2007-2013) under grant no. 215847 – the EU–ADR Project.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Coloma PM, Schuemie MJ, Trifiro G, *et al*. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU–ADR Project. *Pharmacoepidemiol Drug Saf* 2011;20:1–11.
2 Begaud B, Martin K, Haramburu F, *et al*. Rates of spontaneous reporting of adverse drug reactions in France. *JAMA* 2002;288:1588.
3 Singh D. Merck withdraws arthritis drug worldwide. *BMJ* 2004;329:816.
4 Bates DW, Evans RS, Murff H, *et al*. Detecting adverse events using information technology. *J Am Med Inform Assoc* 2003;10:115–28.
5 Avillach P, Joubert M, Thiessard F, *et al*. Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU–ADR project. *Stud Health Technol Inform* 2010;160:1085–9.
6 Trifiro G, Fourrier-Reglat A, Sturkenboom MC, *et al*. The EU–ADR project: preliminary results and perspective. *Stud Health Technol Inform* 2009;148:43–9.
7 Bauer-Mehren A, Furlong LI, Rautschka M, *et al*. From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways. *BMC Bioinformatics* 2009;10(Suppl. 8):S6.
8 Mendonca EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. *Proc/AMIA Ann Symp* 2000:575–9.
9 Garcelon N, Mougin F, Bousquet C, *et al*. Evidence in pharmacovigilance: extracting adverse drug reactions articles from MEDLINE to link them to case databases. *Stud Health Technol Inform* 2006;124:528–33.
10 Duda S, Aliferis C, Miller R, *et al*. Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases. *AMIA Annu Symp Proc* 2005:216–20.
11 Amardeilh F, Bousquet C, Guillemin-Lanne S, *et al*. A knowledge management platform for documentation of case reports in pharmacovigilance. *Stud Health Technol Inform* 2009;150:517–21.
12 Humphreys BL. The 1994 unified medical language system knowledge sources. *Health Libr Rev*. 1994;11:200–3.

13 Trifiro G, Pariente A, Coloma PM, *et al*. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Safe* 2009;18:1176–84.

14 **Avillach P**, Coloma P, Gini R, *et al*. Harmonization process for the identification of clinical events in eight European healthcare databases: the experience from the EU-ADR project. *J Am Med Inform Assoc* 2012. Published Online First 6 Sept 2012. doi:10.1136/amiajnl-2012-000933

15 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32:D267–70.

16 Katcher BS. *MEDLINE: a guide to effective searching in PubMed and other interfaces*. 2nd edn. San Francisco: Ashbury Press, 2006.

17 Oinn T, Addis M, Ferris J, *et al*. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 2004;20:3045–54.

18 Delamarre D, Lillo-Le , Louet A, *et al*. Documentation in pharmacovigilance: using an ontology to extend and normalize Pubmed queries. *Stud Health Technol Inform* 2010;160:518–22.

19 Zeng Q, Cimino JJ. Automated knowledge extraction from the UMLS. *ProcAMIA Ann Symp* 1998;568–72.

20 Wang W, Haerian K, Salmasian H, *et al*. A drug–adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations. *AMIA Annu Symp Proc* 2011;2011:1464–70.

21 van Mulligen EM, Fourrier-Reglat A, Gurwitz D, *et al*. The EU–ADR corpus: annotated drugs, diseases, targets, and their relationships. *J Biomed Inform* 2012;45:879–84.

22 Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. *J Am Med Inform Assoc* 2011;18:668–74.