Journal of Biomedical Informatics 66 (2017) 72–81

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Accuracy of an automated knowledge base for identifying drug adverse reactions

E.A. Voss [a,b,c,*], R.D. Boyce [d,c], P.B. Ryan [a,e,c], J. van der Lei [b,c], P.R. Rijnbeek [b,c], M.J. Schuemie [a,c]

[a] Epidemiology Analytics, Janssen Research & Development, LLC, Raritan, NJ, United States
[b] Erasmus University Medical Center, Rotterdam, Netherlands
[c] Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States
[d] University of Pittsburgh, Pittsburgh, PA, United States
[e] Columbia University, New York, NY, United States

ARTICLE INFO

Article history:
Received 21 July 2016
Revised 8 December 2016
Accepted 10 December 2016
Available online 16 December 2016

Keywords:
Pharmacovigilance
Adverse drug reaction
Machine-learning experiment
Knowledge base
Health outcome

ABSTRACT

Introduction: Drug safety researchers seek to know the degree of certainty with which a particular drug is associated with an adverse drug reaction. There are different sources of information used in pharmacovigilance to identify, evaluate, and disseminate medical product safety evidence including spontaneous reports, published peer-reviewed literature, and product labels. Automated data processing and classification using these evidence sources can greatly reduce the manual curation currently required to develop reference sets of positive and negative controls (i.e. drugs that cause adverse drug events and those that do not) to be used in drug safety research.

Methods: In this paper we explore a method for automatically aggregating disparate sources of information together into a single repository, developing a predictive model to classify drug-adverse event relationships, and applying those predictions to a real world problem of identifying negative controls for statistical method calibration.

Results: Our results showed high predictive accuracy for the models combining all available evidence, with an area under the receiver-operator curve of $\geqslant$0.92 when tested on three manually generated lists of drugs and conditions that are known to either have or not have an association with an adverse drug event.

Conclusions: Results from a pilot implementation of the method suggests that it is feasible to develop a scalable alternative to the time-and-resource-intensive, manual curation exercise previously applied to develop reference sets of positive and negative controls to be used in drug safety research.

1. Introduction

An adverse drug reaction (ADR) is a response to a drug which is noxious and unintended at a dose that is normally used in humans [1]. ADRs may be distinguished from "adverse events" by the identification of a causal relationship to a drug [2]. Approximately 3.6% of all hospital admissions are caused by ADRs [3] and 16.88% of patients experience an ADR during hospitalization [4]. An older, well cited, publication by Lazarou et al. found that of hospitalized patients 6.7% had a serious ADR, 0.32% of which were fatal [5]. Medical decision making could be better informed if the level of certainty regarding potential ADRs were known.

The product label is a primary method for the drug manufacturer to communicate with health providers about the potential effects of drug exposure. However product labels can be difficult to read; one label can list many potential adverse events of which not all have the same probability of occurring [6], and one active ingredient can be included on multiple labels which can provide inconsistent information [7]. In addition to product labels, researchers can find ADR information from spontaneous adverse event reporting or published peer-reviewed literature (e.g., case reports, summaries from randomized clinical trials, and non-interventional observational studies).

Abbreviations: OHDSI, Observational Health Data Sciences and Informatics; OMOP, Observational Medical Outcomes Partnership; LAERTES, Largescale Adverse Effects Related to Treatment Evidence Standardization; HOI, health outcomes of interest; EU-ADR, Exploring and Understanding Adverse Drug Reactions; AZCERT, Arizona Center for Education and Research on Therapeutics.

* Corresponding author at: 920 Route 202, Raritan, NJ 08869, United States.
E-mail address: evoss3@its.jnj.com (E.A. Voss).

http://dx.doi.org/10.1016/j.jbi.2016.12.005
1532-0464/

Spontaneous reports have been successfully used to detect rare events and to stimulate hypotheses about potential associations that warrant further evaluation, but underreporting and other biases can limit their utility [8]. While generally providing more detailed information than individual spontaneous reports, published case reports also tend to be a very limited representation of an unknown fraction of similar events that occur. Clinical studies conducted prior to regulatory approval can identify important safety information about a drug [9] but are generally focused on drug efficacy in a very well-defined but restricted subset of subjects who will eventually have access to the treatment.

There is increasing attention to the application of non-interventional observational studies – using retrospective data from administrative claims and electronic health records or prospective data collected in clinical registries – for post-approval medical product safety surveillance [10]; however the quality of evidence from these studies is still an active area of methodological research [11–14].

Given the wide range of data types, populations, and quality issues, great effort is needed in order to summarize the relationship between a drug and condition. In this paper we explore automating of assembling evidence from disparate sources into a single repository, summarizing it, and applying that knowledge to a real world problem of identifying negative controls for statistical method calibration.

In 2014, members of the Observational Health Data Sciences and Informatics (OHDSI, http://ohdsi.org) [15] community published a proposal to develop an open-source framework to store all relevant sources of pharmacovigilance evidence in a single system [16]. The method would merge the evidence sources into a single evidence database and standardize the terminology leveraging the Observational Medical Outcomes Partnership (OMOP) Vocabularies [17]. One specific purpose for stitching this information together is to use the information as a 'gold standard' in evaluation of study designs and produce their operating characteristics. A pilot version of this method has since been implemented into a system named the Largescale Adverse Effects Related to Treatment Evidence Standardization (LAERTES) [18–20].

The evidence base integrates evidence about the potential relationship between drugs and health outcomes of interest (HOIs) from spontaneous reports, scientific literature, and both American and European product labeling. Spontaneous reporting evidence was from the US Food and Drug Administration's (FDA) Adverse Event Reporting System (FAERS) and included counts of reports and proportional reporting ratio (PRR) scores [21,22]. Evidence from the scientific literature was processed in two ways: the first leveraged Medical Subject Headings (MeSH) tags in a method described by Avillach et al. [23] and the second used relationships semantically tagged Medline abstracts via natural language processing from SemMedDB [24]. These two methods are additionally stratified by Medline publication types: clinical trials, case report, and all other abstracts (i.e., of type Meta-Analysis, Comparative Study, Multicenter Study, or Journal Article). Finally, American product labels are parsed by a method developed by Duke et al. [7] and ADRs mentioned in European labels are provided by the PROTECT project [25]. Table 1 provides additional details on the evidence sources currently included in LAERTES.

Using evidence available in LAERTES, we wanted to quantify the relationship between a drug and an HOI. We performed a quantitative assessment of the predictive accuracy of the evidence base for discriminating between known positive drug-condition causal relationships and drugs known to be unassociated with a condition. This machine-learning experiment has direct application for the research community; information on the relationship between a drug and HOI (particularly ones that have no association) can be used to evaluate pharmacovigilance research study designs and produce their operating characteristics. Measuring a study design's operating characteristics through drug-HOI pairs that should have no association, and using those characteristics to calibrate statistics produced during the study is a recommended practice [12]. Currently, this is a manual process. Specifically, the calibration requires identifying drug-condition pairs that are known not to have a relationship. An automated method for bringing together and quantifying ADR evidence would greatly accelerate the generation of drug-HOI pairs needed for calibration purposes. In this paper, we propose that the automation of data processing and classification of its evidence can greatly reduce the manual curation currently required to develop reference sets of positive and negative controls to be used in drug safety research.

2. Materials and methods

2.1. Loading evidence items into LAERTES

Each evidence item was identified by a drug-HOI pair and loaded into LAERTES with a label indicating the source and its "type" (Table 1). For example, "MEDLINE MeSH ClinTrial" represents a clinical trial report from MEDLINE that uses the Avillach et al. [23] method to identify ADRs. Each piece of evidence was also tagged with a label indicated "modality", i.e., whether the evidence supported a positive or negative association. A piece of evidence was also quantified by one of two possible "statistic" values – (1) count (e.g. the number of Medline abstracts that support the drug-HOI association) or (2) the proportional reporting ratio of the drug-HOI (used only for spontaneous reports). All this information is stored in LAERTES in a single table that contains a key for the drug-HOI pair, the type of evidence (e.g. FAERS Report Count), the modality, the evidence figure (e.g. count of reports), and a URL that can be used to see more details about the underlying evidence items. This overall generic structure within LAERTES allows for disparate sources to be included and then queried once the data was extracted, transformed, and loaded.

2.2. Terminology mapping

Table 1 discusses the differences in how evidence on drugs and HOIs are communicated. Incoming evidence from source databases to LAERTES ranges from free text (e.g. "ALLOPURINOL") to coding vocabularies (e.g. MeSH Unique ID: D000493 for Allopurinol). However, working across evidence sources in this manner is difficult. Standardizing to specific terminologies is critical to enabling evidence from disparate sources to be meaningfully comparable by using a common language to relate drugs and HOIs. For this purpose we depend on the OMOP Vocabulary which contains a library of terminologies (e.g., RxNorm, National Drug Code [NDC], Systematized Nomenclature of Medicine-Clinical Terms [SNOMED-CT], etc.) and provides the relationships between them [17]. The OHDSI collaborative does not generate any of the terminology standards, but instead leverages existing sources and aggregates the terminology concepts and relationships into one common vocabulary model as part of the OMOP Common Data Model. LAERTES relies on content within the OMOP Vocabulary to standardize drugs to RxNorm and standardize conditions to SNOMED-CT. The LAERTES record for each evidence item included a single drug and HOI concept pair (e.g. "omeprazole – anaphylaxis").

In translating evidence from the sources, we found data at varying levels of granularity; for example, one source might provide evidence at the ingredient level (e.g. "omeprazole"), while another might provide similar evidence at the clinical drug level (e.g. "omeprazole 10 MG"). When a drug concept in an evidence source was mentioned at the ingredient or clinical drug level, it was trans-

Table 1
Description of LAERTES sources.

Data source	Description
FAERS Proportional Reporting Ratio (FAERS PRR)	Data files from the FDA Adverse Event Reporting System (FAERS) Latest Quarterly Data Files website [44] were used to generate evidence. The FAERS drug/outcome pairs were standardized from free text drug names and outcomes in MedDRA Preferred Terms to RxNorm OMOP concepts and MedDRA condition OMOP concepts. In addition, the MedDRA condition concepts were mapped to SNOMED-CT concepts based on the OMOP mappings available in the OMOP Vocabulary. The ETL process also included logic to remove duplicate adverse drug event reports [22]. The PRR metric generated by work by Van Puijenbroek et al. [21]. The FAERS data currently available in LAERTES covers Q4 2004 through Q4 2014
FAERS Report Count (FAERS Report Count)	Similar to FAERS PRR except a count of reports is provided for each drug-condition pair
Medline MeSH Clinical Trials (MEDLINE MeSH ClinTrial)	Looking for ADRs in MeSH terms for clinical trials in Medline. The process to identify ADRs was leveraged from Avillach et al. [23]. The Avillach method using MeSH tagged publications from Medline looked for adverse drug reactions based on the co-occurrence of a drug and an adverse event on the same citation. The source of the data used was directly from the National Library of Medicine and gathered from 1946 until September 2015
Medline MeSH Case Reports (MEDLINE MeSH CR)	Similar to MEDLINE_MeSH_ClinTrial except for case reports
Medline Mesh Other (MEDLINE MeSH Other)	Similar to MEDLINE_MeSH_ClinTrial except for it reports on things other than clinical trials or case reports in Medline (i.e. Meta-Analysis, Comparative Study, Multicenter Study, or Journal Article)
Medline SemMedDB Clinical Trials (MEDLINE SemMedDB ClinTrial)	For clinical trials, provides MeSH tagged drug-HOI clinical trial abstracts from PubMed that look for associations such as: causes, affects, associated with, complicates, or disrupts [24]. All of these associations also have a negative modality, meaning SemMedDB provides both positive and negative associations. The data was last mined June 30, 2015
Medline SemMedDB Case Reports (MEDLINE SemMedDB CT)	Similar to MEDLINE_SemMedDB_ClinTrial except for case reports
Medline SemMedDB Other (MEDLINE SemMedDB Other)	Similar to MEDLINE_SemMedDB_ClinTrial except for it reports on things other than clinical trials or case reports in Medline
Structured Product Label Adverse Drug Reactions from SPLICER (SPL SPLICER ADR)	SPLICER, a tool that reads and parses United States Structured Product Labels (SPLs) for drugs and HOIs in the sections "Adverse Drug Reactions" or "Postmarketing" [7]. SPLICER already utilizes the OMOP Vocabulary and maps drugs to RxNorm and HOIs to MedDRA terms. The SPLICER data was up-to-date as of September 2015
European Product Label Adverse Drug Reactions (SPL EU SPC)	From the PROTECT ADR database, this provided a list of ADRS on Summary of Product Characteristics (SPC) of products authorized in the European Union [25]. The drugs come across as free text and the HOIs come across as descriptions of MedDRA Preferred Terms. It was last updated on December 31, 2013

lated through the OMOP Vocabulary to the RxNorm concepts at the same level. Since evidence for drug-HOI pairs could be at either the ingredient or clinical drug level, we aggregated evidence to individual ingredients for further analysis. This aggregation was straightforward using the clinical-drug-to-ingredient relationships from RxNorm, as available in the OMOP Vocabulary.

The various evidence sources provided HOI concepts using three different terminologies. The Medline source used the MeSH terminology. SemMedDB used UMLS concepts that represented concepts in MeSH, MedDRA, or SNOMED. Both strategies for parsing American and European product labeling used MedDRA. The mapping process used relationships in the OMOP Vocabulary to map from the source terminologies to SNOMED. Sometimes this resulted in evidence for very similar HOI concepts that resided at different levels of the SNOMED "is a" hierarchy. For example, if a source used MeSH to represent the concept "Myocardial Infarction" it might be mapped to the SNOMED concept "Myocardial Infarction", while the same concept from a source that used MedDRA might be mapped to the highly similar child concept "Acute Myocardial Infarction". In using the evidence base for this study, each HOI was defined as the aggregate evidence from the concept itself and all its descendant concepts. For example, the concept of "Myocardial Infarction" was considered to be representative of all of its children concepts including "Acute Myocardial Infarction" and "Acute Subendocardinal Infarction".

2.3. The reference sets

Two existing manually-created reference sets were used to train an automated classifier for ADR signal detection and estimate its accuracy using cross-validation: the OMOP Reference Set [26] and the Exploring and Understanding Adverse Drug Reactions (EU-ADR) Reference Set [27–29]. These were chosen because their drugs and HOIs were already translated in OMOP Vocabulary concepts. They provided the ground truth that served as the basis for understanding LAERTES performance and defining an algorithm for prediction.

The OMOP Reference Set, initially developed in 2010, contains 4 HOIs each with its own positive and negative control drug set. The 4 HOIs are acute kidney injury, acute liver injury, acute myocardial infarction, and gastrointestinal bleed, which were all originally chosen to provide a spectrum of adverse events or likelihood of being the focus of ongoing drug safety surveillance [30]. The OMOP Reference Set positive controls list started from product labels that listed the HOIs of interest in the "Black Box Warning" section, the "Warnings and Precautions", or "Adverse Reactions" sections [26]. The list was further defined using an independent literature review of randomized trials or observational studies, as well as systematic literature review provided by Tisdale and Miller [31]. These same information sources were also used to define the negative controls. In this case, by assuming that a lack of evidence across all of the sources for a given drug-HOI association indicated that no association exists. Across the 4 HOIs in the OMOP Reference Set there are 165 positive controls (drugs that are known to cause an HOI, ground truth is 1) and 234 are negative controls (drugs that are known to not cause the HOIs, ground truth is 0).

The EU-ADR Reference Set, initially developed in 2012, has 10 HOIs (liver disorder, acute myocardial infarction, renal failure acute, anaphylactic shock, erythema multiform, mitral valve disease, neutropenia, aplastic anemia, rhabdomyolysis, and gastrointestinal hemorrhage) each with its own positive and negative control set. The HOIs chosen for this reference set were considered important from a pharmacovigilance and public health perspective [27]. The process of generating the positive and negative controls for this reference set started slightly differently than for the OMOP Reference Set. The originating team started by looking for drugs with enough exposure in the EU-ADR database network [27,32]. Following this, a literature review was conducted to find associations between the initial drug list and the HOIs of interest. If more than 3 citations could be found in MEDLINE for a drug-ADR association then, it was considered for a positive control. If there were no

literature citations and no World Health Organization (WHO) Vigibase® mentions of the drug-HOI pair then, it was considered as a negative pair. All final drug-ADR pairs for the EU-ADR Reference Set went through a manual review. In total, there are 93 positive and negative controls across the 10 HOIs in the EU-ADR Reference Set; 43 are positive controls and 50 are negative controls. EU-ADR also has one HOI that only has negative controls, mitral valve disorder.

Both the OMOP and EU-ADR Reference Sets were translated to standard OMOP Vocabulary concepts; RxNorm for drugs and SNOMED for conditions. The definitions for the drug-HOI pairs can be found in Supplementary data 1. Our hypothesis was that a composite summary of evidence from LAERTES can be predictive of negative and positive controls in OMOP and EU-ADR Reference Sets that were manually generated. In addition, we hypothesize that using all the evidence sources to predict the reference sets will have improved performance over the individual sources independently.

While the OMOP and EU-ADR Reference Sets were used for training and testing the pilot method, an additional reference set called the Arizona Center for Education and Research on Therapeutics (AZCERT) dataset [33,34] was utilized in a validation manner. The CredibleMeds group, which manages the AZCERT list, focuses on programs to reduce preventable harm from medication. That AZCERT dataset used for this study focused on two HOIs: Torsade De Pointes (TDP) and QT Prolongation. We used the OMOP Vocabulary to define what condition concepts in LAERTES would relate to these HOIs. Specifically, we used SNOMED concepts 314664-"Long QT syndrome" and 4135823-"Torsades de pointes" and their children to define this HOI. 4008859-"Prolonged QT interval" was also associated to "Long QT syndrome" for this work; "Prolonged QT interval" is considered a measurement in the OMOP Vocabulary but is relevant for this reference set. The AZCERT drugs considered associated to QT prolongation and TDP were downloaded from CredibleMeds® (https://www.crediblemeds.org/) in July 2015 and the OHDSI USAGI concept mapping program [35] was used to associate the ingredients to OMOP Vocabulary RxNorm ingredients. This mapping can be found in Appendix B. Of the 182 drugs, only three did not have mappings in the OMOP Vocabulary, most likely due to no or recent approval in the US: dihydroartemisinin & piperaquine, ivabradine, and panobinostat.

2.4. The drug and HOI "Universe"

A subset of drugs and HOIs were chosen to be the "universe" for this analysis. The term "universe" in this paper refers to the ingredients and HOIs for which sufficient evidence exists to suggest that medication safety issues would have been known and reported at the time of the current experiment. Operationally, we included drugs for which the ingredient or HOI had at least one FAERS evidence item, one Medline evidence item (either from the Avillach method or SemMedDB), and one evidence item from either EU or US product labels. Another way to say this is to remove evidence that is lacking in at least one of FAERS, Medline, or product labels (e.g. FAERS has the ingredient "bees wax" however this does not appear in Medline so it cannot exist in the universe). This step was taken because an HOI or drug that showed up in only one source might indicate a lack of clinical experience. For example, a novel topic may appear in FAERS but take a while to start showing up in literature. Also, a novel drug may show up in product labels fairly quickly and not the other sources. In either case, the limited available evidence was not considered in the experiment because we thought that the lack of evidence available in the other sources would likely be due to novelty.

The evidence for drugs was only reviewed at the ingredient level. For example, SPLs are at the clinical drug level however that is too specific for the other sources therefore we translate all drug mentions to ingredients. Only ingredients across the three sources are part of the universe.

For HOIs, all concepts tagged from incoming sources were considered as well as their ancestor concepts as identified through the OMOP Vocabulary relationships. This would give a higher level term within the Vocabulary more potential opportunity to contain enough evidence assuming it will have more descendants than its children concept). Aggregation of conditions concepts in this manner allows evidence at a low-level concept to be rolled-up into more general concepts based on the SNOMED hierarchy. This is important because different source's HOIs will be mapped at different levels of granularity and the OMOP Vocabulary hierarchy enables synthesis of the evidence at each level of detail.

2.5. Statistics

Logistic regression was used to build multivariate models on the LAERTES data that could discriminate between positive and negative controls. Regularization with a Laplace prior on the regression coefficients was used to allow the model to perform parameter selection.

2.6. Building classification models

Models that predicted drug-HOI associations were built using the evidence provided by each source in LAERTES as predictors. Referring to sources listed in Table 1, the following is a discussion on how features (predictor variables) were constructed for each drug-HOI pair. Each predictor was finally scaled by dividing by its standard deviation.

- FAERS:
  - The FAERS proportional reporting ratio (PRR) for the drug-HOI pair was the geometric mean of all PRR values assigned to the drug-HOI pair (which included all children condition concepts to that HOI).
  - FAERS report counts were summed across the counts for the drug-HOI pair (which included all children condition concepts to that HOI). The natural log of the total was taken to address the positive skew found in the count data.
- Medline:
  - Avillach Method [23]: Evidence for the drug-HOI pair from the three Medline MeSH sources was treated as discrete count data (the number of articles).
  - SemMedDB Method [24]: Evidence from the drug-HOI pair from the three Medline SemMedDB sources was summarized as a categorical variable (1 if present else 0 if no evidence existed). SemMedDB was treated as categorical because testing found some of the drug-HOI counts were erroneously inflated because of a terminology mapping issue from this source to the Vocabulary (this programming bug only affected this source). In addition, Medline SemMedDB sources were also broken out into positive and negative modality, making a total of six pieces of evidence from SemMedDB. SemMedDB is the only data source currently to make use of the negative modality.
- Product Labels:
  - There can be many product labels per ingredient in both the US (parsed using Duke et al. method [7]) and EU (parsed using the and Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT) method [25]) that for the most part all say the same thing; therefore an evidence count here is not appropriate. Instead a categorical variable was used to indicate that evidence existed for the drug-HOI pair.

2.7. Evaluation

As a first step, we evaluated the performance of our approach within the OMOP and EU-ADR Reference Sets separately. To prevent overly optimistic performance metrics due to overfitting, we utilized leave-pair-out cross-validation [36]. In leave-pair-out cross validation, for every combination of a positive and a negative control in a reference set, a model is fit using all data except the left-out pair and then evaluated on its ability to rank the left-out pair. We then compute an overall predictive accuracy across all folds (all pairs of negative and positive controls) as the area under the receiver-operator curve (AUC); 95% confidence intervals for the AUC were computed to account for uncertainty due to random error [37].

The second step consisted of evaluating the generalizability of the combined model by fitting the model on the combination of the OMOP and EU-ADR Reference Sets, and using the AZCERT for evaluation. In AZCERT, only drugs that were labeled as "Risk of TdP" or "Conditional Risk of TdP" were considered confident ADR associations while other drugs under "Avoid in congenital long QT" and "Possible Risk of TDP" seemed to reflect less confident associations, or ones that were likely only under certain conditions, and therefore were eliminated for consideration in this analysis. When the list was narrowed there were 77 drugs left of which only ivabradine was unmapped to a Vocabulary concept. When comparing these 76 mapped drugs 55 were in the LAERTES universe. These 55 served as the positive controls and all other drugs in the LAERTES universal set, excluding the "Avoid in congenital long QT" and "Possible Risk of TDP" drugs from AZCERT, were used to see if the model's predicted probabilities could discern between them and the AZCERT-identified positive controls. Because some of these negative controls could in reality be positive controls that were not catalogued in AZCERT, we also assessed AUC and the raw counts assuming 1% of the negative controls were misclassified in both the best-case scenario (the 1% highest ranked negative controls according to our algorithm were misclassified) and the worst-case scenario (the 1% lowest ranked negative controls according to our algorithm were misclassified). The full AZCERT Reference Set can be found in Appendix C. Additionally, Fig. 1 depicts both the first step and second step of evaluation of our models built.

2.8. Software and tools used

The LAERTES data was stored in a PostgreSQL database v9.3 however the SemMedDB and Medline of the extract, transform, and load (ETL) processes used MySQL v5.5. A Virtuoso server v6.1 was used for Resource Description Framework (RDF) graphs that provided the link out data to source content. All LAERTES content was stored in an Amazon AWS cloud server running Ubuntu 14.04 in order to provide scalability as well ease team participate across geographies and organizations (including both public and private sector participants). All analyses were conducted in R version 3.2.1 [38] using the Cyclops package [39].

3. Results

The summary statistics of the state of LAERTES at the time of the experiment can be found in Table 2. For each piece of evidence and modality, Table 2 reports across its columns the number of distinct rows, distinct drugs, and distinct conditions. FAERS accounted for a 34.4% of the summary statistics in LAERTES. There were 2.7 million FAERS records within LAERTES, each with a count of reports and a proportional reporting ratio (PRR). Medline made up over 2 million rows or 27% of the records. Product labels made up 220,809 rows or about 3% of the LAERTES records. Across all the incoming evidence there were 3797 distinct ingredients and 9403 distinct conditions. Focusing on the aforementioned "universe" of drugs and HOIs for this analysis, LAERTES provided evidence for 992 distinct ingredients and 3488 distinct HOIs. The reduction was caused by the "drug universe" membership requirement that the ingredient or HOI was in FAERS, Medline, and product labels at least once. After review of the number of rows per data source, we decided not to use the negative modality evidence (Medline SemMedDB Clinical Trial, Medline SemMedDB Case Report, Medline SemMedDB Other) as well as positive modality from 'Medline SemMedDB other' as these data elements were found to be too sparse for building predictive models.

Table 3 shows the number of ingredients, HOIs, and ingredient-HOIs pairs in the OMOP and EU-ADR Reference Sets and in the entire LAERTES database, both before and after restricting to those items that meet our criterion of minimum amount of data. As discussed above, LAERTES drug evidence was included at the ingredient level while HOIs evidence was included at the level of a given concept and all of its children concepts in the OMOP Vocabulary. For example, product label evidence for 19019271-"Naproxen 375 MG Oral Tablet" and 197925-"Hemorrhage of rectum and anus" was associated to the OMOP Reference Set item 1115008-"Naproxen" and 192671-"Gastrointestinal hemorrhage". In the OMOP Vocabulary the SNOMED term 192671-"Gastrointestinal hemorrhage" was inclusive of 197925-"Hemorrhage of rectum and anus" and 1115008-"Naproxen" was the ingredient associated to 19019271-"Naproxen 375 MG Oral Tablet".

LAERTES contained evidence on 3797 distinct ingredients with 992 having enough evidence to be considered for analysis. There were a potential of 9403 HOIs and 3488 were considered to have enough evidence for analysis. We refer to the included drugs and HOIs as the "LAERTES universe". Because LAERTES was not designed for specific drug-HOI pairs, every permutation of ingredients and HOIs is available for evidence which provides over

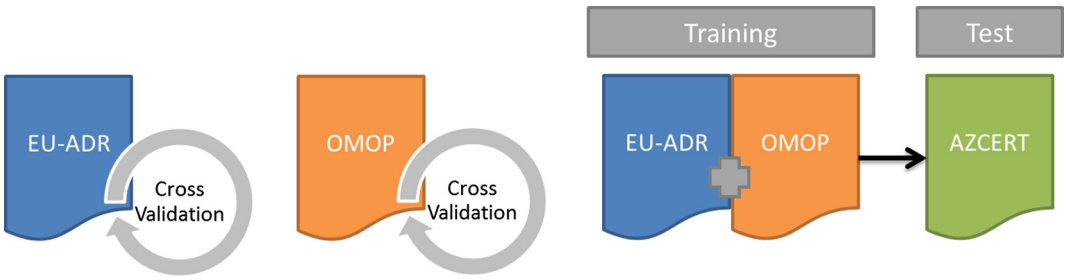

Fig. 1. Graphical depiction of how the reference sets are used and how the models were training and tested. Leave-pair-out cross validation was used to evaluate the models built independently on EU-ADR and OMOP reference sets. The third model was trained on the combination of both the EU-ADR and OMOP reference sets and then tested using the AZCERT as the test set.

Table 2
Dimensions of LAERTES.

	Modality	Rows	Distinct ingredients	Distinct HOI
FAERS PRR	Positive	2,742,314 (34.4%)	3534 (93.1%)	8463 (90.0%)
FAERS	Positive	2,742,314 (34.4%)	3534 (93.1%)	8463 (90.0%)
Medline Clinical Trial	Positive	207,018 (2.6%)	2292 (60.4%)	970 (10.3%)
Medline Case Reports	Positive	825,635 (10.4%)	2358 (62.1%)	2248 (23.9%)
Medline Other	Positive	1,122,493 (14.1%)	2389 (62.9%)	2458 (26.1%)
Medline SemMedDB Clinical Trial	Positive	11,595 (0.2%)	1353 (35.6%)	202 (2.2%)
Medline SemMedDB Clinical Trial	Negative	906 (0.0%)	302 (8.0%)	48 (0.5%)
Medline SemMedDB Case Report	Positive	17,285 (0.2%)	1668 (43.9%)	227 (2.4%)
Medline SemMedDB Case Report	Negative	550 (0.0%)	330 (8.7%)	25 (0.3%)
Medline SemMedDB Other	Positive	79,725 (1.0%)	2049 (54.0%)	366 (3.9%)
Medline SemMedDB Other	Negative	4498 (0.1%)	1083 (28.5%)	125 (1.3%)
EU Product Labels	Positive	24,626 (0.3%)	315 (8.3%)	2052 (21.8%)
US Product Labels	Positive	196,183 (2.5%)	1085 (28.6%)	2645 (28.1%)
Total	Positive	7,975,142 (100.0%)	3797 (100.0%)	9403 (100.0%)
LAERTES Universe Set Evidence Across FAERS, Medline, and product labels	–	–	992 (26.1%)	3488 (37.1%)

FAERS: FDA Adverse Event Reporting System, PRR: proportional reporting ratio, HOI: health outcome of interest.
This table communicates the size of LAERTES in terms of distinct number of rows, distinct ingredients, and distinct HOIs. Utilizing this data we are able to find the LAERTES universe set which is the ingredients and HOIs that contain at least one piece of evidence in each of the following: Medline, product labels, and spontaneous reports.

Table 3
Reference set sizes narrowed to available evidence in LAERTES Universe.

Reference set	Before restricting to LAERTES universe			Restricting to LAERTES universe		
	Distinct ingredients	Distinct HOIs	Distinct ingredient-HOI	Distinct ingredients	Distinct HOI	Distinct ingredients-HOI
OMOP	182	4	399	151	4	329
EU-ADR	65	10	93	59	9	77
LAERTES	3797	9403	35,703,191	992	3488	3,460,096

OMOP: Observational Medical Outcomes Partnership, EU-ADR: Exploring and Understanding Adverse Drug Reactions, HOI: Health Outcomes of Interest.
The number of ingredients, HOIs, and ingredient-HOIs pairs in the OMOP and EU-ADR Reference Sets and in the entire LAERTES database, both before and after restricting to those items that meet our criterion of minimum amount of data.

3 million potential permutations. The OMOP Reference Set had 182 distinct ingredients, 151 (83%) of which were included in the LAERTES universe. There were 4 HOIs listed in the OMOP Reference Set and all 4 (100%) had enough evidence within the LAERTES universe. The OMOP Reference Set lists 399 drug-HOI pairs, of which 329 exist in the LAERTES universe. For EU-ADR, there were 65 ingredients, 59 (91%) of which were in the LAERTES universe. Additionally, there were 10 HOIs in the EU-ADR Reference Set. However, 1 HOI (rhabdomyolysis) was missing because it lacked enough evidence due to a Vocabulary mapping inconsistency discussed in the Limitations section. The EU-ADR Reference Set lists 93 drug-HOI pairs, of which 77 are within the LAERTES universe.

Table 4
AUC (Area Under the Curve) and 95% confidence interval for individual predictors and a regularized logistic regression model using all predictors, using leave-pair-out cross-validation.

Column(s) in model	OMOP AUC	EU-ADR AUC
Medline Clinical Trial	0.74 (0.69–0.79)	0.73 (0.63–0.83)
Medline Case Reports	0.85 (0.81–0.89)	0.88 (0.81–0.96)
Medline Other	0.85 (0.80–0.89)	0.87 (0.79–0.95)
Medline SemMedDB Clinical Trial	0.58 (0.55–0.61)	0.57 (0.51–0.63)
Medline SemMedDB Case Reports	0.58 (0.55–0.61)	0.59 (0.52–0.65)
EU Product Labels	0.57 (0.54–0.60)	0.53 (0.49–0.57)
US Product Labels	0.87 (0.84–0.91)	0.80 (0.71–0.89)
FAERS[a]	0.73 (0.67–0.78)	0.70 (0.57–0.82)
FAERS PRR[b]	0.64 (0.58–0.70)	0.75 (0.63–0.86)
All Predictors	0.94 (0.91–0.97)	0.92 (0.86–0.98)

OMOP: Observational Medical Outcomes Partnership, EU-ADR: Exploring and Understanding Adverse Drug Reactions, AUC: area under the curve, LBCI: lower bound 95% confidence interval, UPCI: upper bound 95% confidence interval, FAERS: FDA Adverse Event Reporting System, PRR: proportional reporting ratio.
[a] Natural logs were taken to scale predictor.
[b] Geometric mean was used to scale predictor.

Table 4 provides the predictive accuracy of each evidence type alone as well as the full model with all evidence types. However, due to regularization, not all evidence may play a role in the model. For example, in the EU-ADR model "Medline Other", "SemMedDB Clinical Trial", "SemMedDB Case Reports" model coefficients were shrunk to 0 as they did not provide additional information to the model). Looking at the predictive accuracy of individual pieces of information in the OMOP Reference Set, US product labels were the most predictive (AUC = 0.87 [95% CI: 0.84–0.91]). For the EU-ADR Reference Set, the Medline case reports were most predictive (AUC = 0.88 [95% CI 0.81–0.96). For both the OMOP and EU-ADR Reference Sets, the model with all the predictors performed better than any one single predictor, AUC = 0.94 [95% CI: 0.91–0.97] and AUC = 0.92 [95% CI: 0.86–0.98] respectively.

With the results of Table 4 and the EU-ADR demonstrating large predictive power for Medline Case Reports, Medline Other, and US product labels, we performed additional investigation to validate these findings. Across both cases, out of 36 positive controls, 34 had some type of case report, and of the 41 negative controls, 21 had a Case Report and 24 had an "Other" type of report. In addition, the large coefficient for US product labels scoring was a bit of surprise for EU-ADR since it was not information used in the original heuristic; however, the US product labels data processed by SPLICER has more evidence compared to the EU label data processed in PROTECT which was only a subset of drugs. Therefore, US labels had more opportunity to be predictive of positive and negative controls. When there was at least a label for a drug/HOI combination within the OMOP and EU-ADR Reference Sets 83% of the time we had only a US label, 13% of the time there was both a US and EU label, and 4% of the time there was an EU but not a US label.

To understand how the reported performance was achieved, we fitted models using the reference sets individually and then on both reference sets. Table 5 shows the regression coefficients of

Table 5
Coefficients in models built on the full datasets (not using cross-validation) using regularized logistic regression.

Ref. Set	Intercept	Medline clinical trial	Medline case reports	Medline other	SemMedDB clinical trial	SemMedDB case reports	EU product labels	US product labels	FAERS[a]	FAERS PRR[b]
OMOP	−3.67	−5.55	0.51	5.43	0.47	0.02	0.46	1.81	0.59	−0.07
EU-ADR	−2.65	0.75	3.89	-	-	-	0.18	0.85	0.09	0.61
OMOP/EU-ADR	−2.86	-	1.54	-	0.04	0.05	0.36	1.56	0.38	-

- (hyphen): indicates a coefficient was shrunken to exactly zero.
OMOP: Observational Medical Outcomes Partnership, EU-ADR: Exploring and Understanding Adverse Drug Reactions, FAERS: FDA Adverse Event Reporting System, PRR: proportional reporting ratio.
All predictors were scaled by dividing by the variable's standard deviation prior to fitting the model.
[a] Natural logs were taken to scale predictor.
[b] Geometric mean was used to scale predictor.

these overall models. The largest coefficients in the OMOP model were for "Medline Other". The largest coefficient in the EU-ADR model was for Medline Case Reports. Table 5 also shows a model built off both the OMOP and EU-ADR set together (the last row). The largest coefficients were the Medline Case Reports and US Product Labels.

Fig. 2 represents the distribution of predicted probability of drugs being positive or negative controls using the model built on the combined OMOP and EU-ADR Reference Sets. The plots suggest that the predicted probabilities produced by the algorithm were useful for segregating positive and negative controls. For example, the model predicted a probability of 1.00 for ketoprofen associated to the HOI of gastrointestinal hemorrhage (see Appendix D, OMOP Reference Set) and there is a black box warning for stomach bleed with NSAIDs like ketoprofen. For negative controls in the OMOP Reference Set, the model indicates a probability of 0.05 associating almotriptan to acute renal failure, which is in agreement with the OMOP Reference Set's choice with this ingredient-HOI pair (both LAERTES and the OMOP work found no evidence of published papers, no elevated PRR, and nothing on the product label).

Using the model built off both the OMOP and EU-ADR Reference Set to predict AZCERT drug-HOI associations, we found that the model was able to separate the ingredients that were positive controls from those that were not positive controls. Drugs that are "not positive controls" are referred to as negative controls in Fig. 2, however technically the AZCERT does not provide negative controls and they were inferred from LAERTES. Due to the non-positive controls being such a large portion of the reference set, there is a bit of imbalance between the two classes. Appendix D provides the predicted probabilities. Some positive controls do have low predictive probabilities, however most of the negative controls are substantially lower. The 865 non-positive controls predictive probabilities range from 0.05 to 1.00; however, 75% of the non-positive controls are 0.25 or below, while the 55 positive controls range from 0.17 to 1.00 and 89% of the positive controls are greater than 0.25. Fig. 2 shows a fairly high AUC = 0.92 (CI [assuming no misclassification]: 0.89–0.95, assuming 1% worst case misclassification [about 9 drugs]: 0.79, assuming 1% best case misclassification: 0.94) for AZCERT however the AUC has wide bounds for the point estimate suggesting some uncertainty in the model. We note that the point estimate bounds were calculated differently for AZCERT than the confidence intervals developed for OMOP and EU-ADR, however all give a sense of the model AUC range.

4. Discussion

LAERTES brings together adverse event evidence from spontaneous reports, medical literature, and product labels into one database and standardizes the source's terminologies to standard condition and drug vocabularies. In addition to standard terminologies, the different types of LAERTES evidence are stored in a standardized structure – a structure that is flexible to accept additional future evidence sources (e.g. observational evidence). This allowed us to see if a classifier built using evidence from disparate sources can identify drugs that cause certain outcomes (positive controls) and drugs that lack evidence for certain negative outcomes (negative controls).

The prediction results for the full model suggested that evidence used in aggregate is more predictive than the univariate models. This may be suggestive that individual pieces of evidence only bring certain amounts of information to the researcher, and if that researcher stops at only reviewing one data source they are most likely not getting a complete picture of the potential ADR issues. This work found that individual US product labels, Medline data, and FAERS counts to be informative for predicting drug-HOI associations in the OMOP and EU-ADR Reference Sets. Our results showed fairly high predictive accuracy for the models combining all available evidence, an AUC of 0.93 for the OMOP Reference Set, 0.92 for EU-ADR, and 0.92 for AZCERT. Another way to interpret this last AUC of 0.92 is if we picked a sensitivity of 50% we would achieve a specificity of 97% and a positive predictive value of 48% on the AZCERT Reference Set.

Reviewing the predictors found in Table 5 it may seem hard to draw generalizable conclusions from coefficients that vary so much between models. However, the reference sets themselves were made by different processes and selection of the model coefficients reflect this; two different reference sets (OMOP and EU-ADR) made by two different groups using two different processes. Additionally, some predictors seem to be more predictive than others, for example the US product labels have larger model coefficients than the EU product labels. Again the model is reflecting reality; first there are more US labels parsed by SPLICER in LAERTES than EU labels from PROTECT and additionally the OMOP Vocabulary at the time of analysis had a slight bias towards US centric drugs thus making some of the European labels not able to participate.

It is important to point out that there may seem to be a bit of a recursive argument here in that the reference sets were built using the same data that LAERTES is using. However, the OMOP and EU-ADR Reference Sets were generated manually and while the final model generated drug-HOI associations in an automated fashion. One of the main reasons AZCERT was included was to help understand if circular argument was an issue for LAERTES, however the data was still predictive on this reference set that was not necessarily generated in the same manner as the OMOP and EU-ADR Reference Sets.

AZCERT provided an independent reference set, a reference set not used in the development of the model. We used other drugs in LAERTES not part of the AZCERT to represent the negative controls. Assuming all the drugs in LAERTES not in AZCERT are negative controls may not be a correct assumption; AZCERT does not

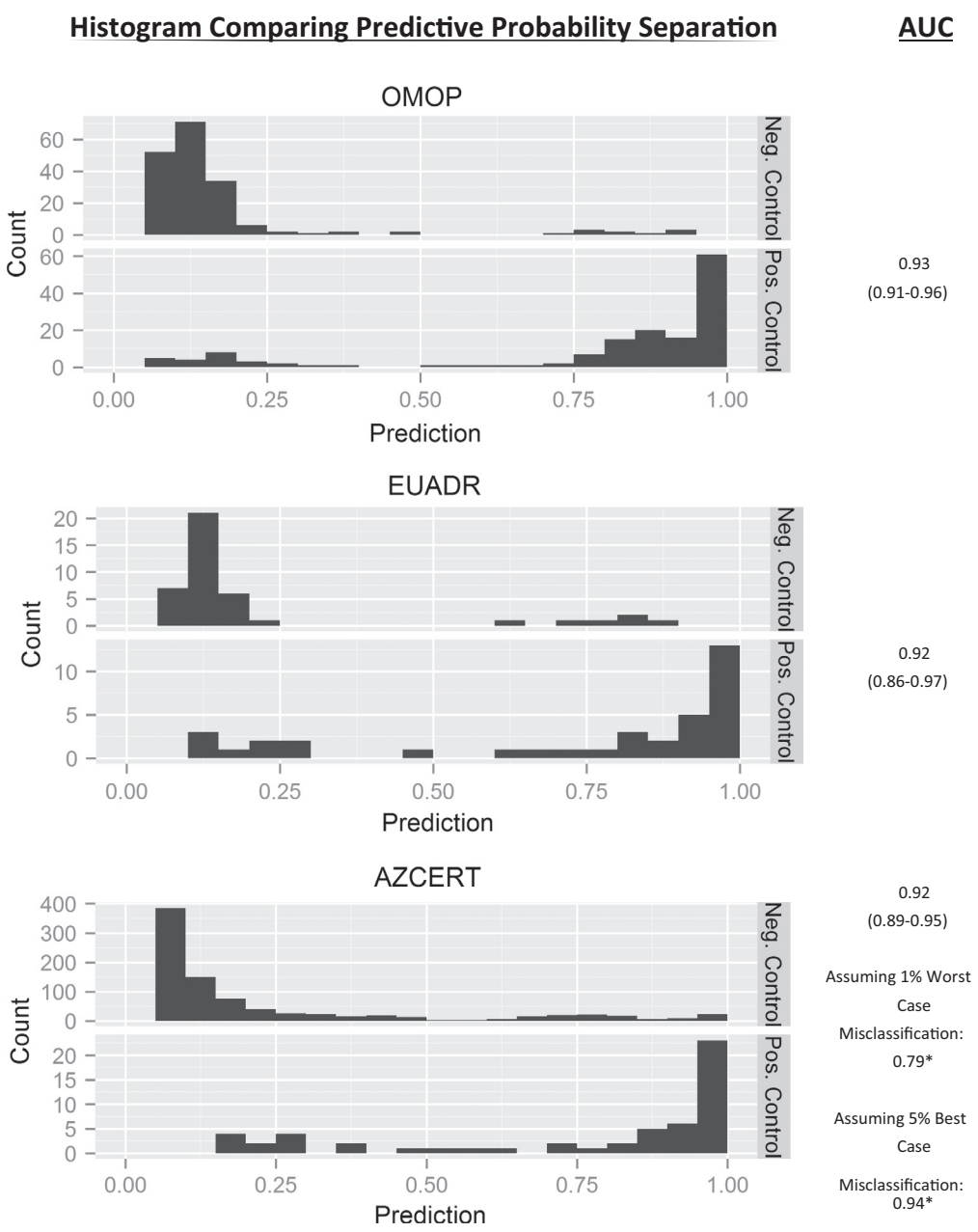

Fig. 2. Histograms of predicted probabilities with AUCs for positive and negative controls in the various reference sets, using the model trained on both OMOP and EU-ADR Reference Set. Values: AUC (Lower Bound AUC-Upper Bound AUC). *Calculated by assuming a 1% of the negative controls were misclassified. OMOP: Observational Medical Outcomes Partnership, EU-ADR: Exploring and Understanding Adverse Drug Reactions, AUC: area under the curve.

necessarily contain all drugs that cause TDP/QT prolongation meaning that some of the negative controls may actually be positive controls. We assume that the list is more or less inclusive of all drugs but still calculated confidence interval for the AUC guessing that 1% of the non-positive drugs were in fact positive controls (this gives you the best and worst cases if some drugs are misclassified). With the model predicting positive controls, even though some of the predictive probabilities were low, compared to the non-positive control set, they had higher values. The lowest positive control was sevoflurane with a predictive probability of 0.17. The lowest performing drug from AZCERT was sevoflurane. Sevoflurane had evidence from Medline and FAERS but no information from SemMedDB or product labels. The resulting model weighted heavily having a product label, since suvoflurane does not seem to make mention of QT prolongation as an issue in either the ‘‘Adverse Drug Reactions” or ‘‘Postmarketing” section of its labels than its predicted probability is lower. The smallest 18 predicted probabilities from AZCERT do not have an US product label which is deemed important in our prediction model.

This work demonstrates the feasibility of using the evidence from disparate sources to select positive and negative controls using the automated machine learning process. In its current form, the predictive probabilities generated from the final model have been used to find suitable negative controls for model calibration (e.g. for a certain condition, asking what drugs have a low probability of being associated to it based on the evidence in LAERTES, or for a given drug, identifying candidate conditions that are not observed to be associated). The candidate negative controls identified through this process still require clinical adjudication prior to use, but using the automated procedure to identify the set of candidates greatly reduces the required resource without substantially decreasing specificity of the controls selected.

Additional future work could include advancing the model built so that it could easily be used for drug-HOI prediction. This may need to include weighting of the evidence; we may not want to treat all LAERTES parameters as equal. Weighting could add additional information about the quality of data received from LAERTES or include our belief on the quality of the suggested association. Additionally, a more ambitious goal is to determine if an automated ADR prediction method using LAERTES could achieve sufficient performance to supplement, or even be a replacement for, the expensive manual evidence synthesis effort currently required to investigate pharmacovigilance signals. For example, LAERTES holds promise of being a tool to be used directly for signal detection; instead of having many places to review evidence each using their own terminology LAERTES could provide the "one stop shop" for reviewing the evidence, the accessibility of evidence would improve ease of signal review.

5. Limitations

One challenge in the application of the LAERTES data was suboptimal mappings between source vocabularies within the OMOP Vocabulary. Specifically mapping to SNOMED conditions from different starting source codes can be difficult. Earlier it was described that rhabdomyolysis was not in the LAERTES universal set was due specifically to a mapping issue (MedDRA concept was mapping to 'Muscle, ligament and fascia disorders' instead of 'Rhabdomyoly sis'). Mapping within conditions is not straightforward and this area will take continued data investigation to uncover where our mapping to the standard terminologies could be improved.

In order to improve comparability between drugs, when analyzing LAERTES evidence we translated all drugs to their ingredients. One consequence of this decisions is that combination drugs will provide evidence all associated ingredients; 'dapagliflozin 10 MG/metformin hydrochloride 500 MG Extended Release Oral Tablet' would be associated to both dapagliflozin and metformin and the evidence associated to that clinical drug will appear both for dapagliflozin and metformin. We felt that this was appropriate because even though we may have a strong feeling about which ingredient caused the ADE we should not assume this. If there is a true relationship, the ADE will show up multiple times in multiple evidence sources. In future releases, however, we would like to explore adding information on the certainty of the association (i.e. this drug had a direct map to an ingredient or this drug was mapped to multiple ingredients) which should in turn help them model make better predictions.

As highlighted earlier, future work should include weighting of the evidence. Currently all evidence is treated as equal in the model and this is most likely not representative of truth; it is intuitive to believe that information from a product label is more likely to represent an ADR over a spontaneous report. Additionally, we require the existence of a condition and separately of a drug in multiple sources of evidence to be considered to participate in drug-condition pairs. This forces us only to review drugs or conditions that we are confident there is evidence for. To improve the model so that it more closely represents real world scenarios we should evaluate our confidence in the evidence sources, allowing that confidence level to participate in the model, and exploring consideration of drugs or conditions that may have evidence only found in one or two types of evidence sources.

This experiment also heavily relies on the OMOP and EU-ADR Reference Sets. Despite extensive efforts by the teams that developed both reference sets to get what they determined to be a high quality reference set, it is still possible that not all controls are correctly classified [40]. However, both reference sets have been used in multiple studies [41–43]. Additionally, it can be argued that the reference sets do not include a diverse enough set of conditions reviewed. The HOIs tend to be acute conditions rather than chronic ones like diabetes. This problem was also found with the AZCERT reference set where only one condition was used for testing the model built on the combination of OMOP and EU-ADR. The lack of diversity in the reference sets may limit the generalizability of the results to all conditions and this should be taken into account when utilizing the model for prediction.

In preparing for this publication a few limitations became evident with the implementation of LAERTES. The knowledge base has the data it needs to allow users to search on standardized terminologies and retrieve link outs to evidence sources. However, access to this "drill down" data is currently difficult as full access requires programming against the OMOP Vocabulary and LAERTES using SQL. There is an experimental LAERTES evidence explorer (http://www.ohdsi.org/web/knowledgebaseweb) but it is currently a prototype user interface.

As discussed earlier, there is some evidence in LAERTES that is sparse (e.g. negative modality for Medline SemMedDB Clinical Trial) and was not included in our models; there may be future opportunity to improve in this area. Also, the team needs to develop appropriate views into the data (e.g. a web interface accessing the data in a certain manner); however efforts such as publication and application will highlight what views make the most sense. Additionally, with some evidence sources there are some Vocabulary mapping that should be reviewed. For example, with Medline publications, some of the MeSH terms map out into the Vocabulary to many board concepts which then seems to associate tangentially associated abstracts to your drug-HOIs of interest. Finally, we do not know how the model will perform as the evidence in LAERTES evolves (e.g. additional evidence within a data source, changes in the Vocabulary mappings, more data sources added). This paper outlines the process for learning a model based on the evidence and the model coefficients most likely will change as LAERTES advances. But without applying LAERTES to the "real world" it would be impossible to understand where and how to improve upon the tool.

6. Conclusions

The goal of this paper was to use the designed method for gathering evidence implemented in LAERTES to explore the relationship between drugs and HOIs. We demonstrated that using LAERTES its evidence was predictive of the reference sets, particularly when using all the predictors with sufficient data. The model classifier also performed well on AZCERT. The method to pull disparate data sources together will only continue to improve as new evidence sources are added. As this process implemented to generate LAERTES provides a scalable alternative to the time-and-resource-intensive, manual curation exercise previously applied to develop reference sets of positive and negative controls used in drug safety research.

Funding

This research was funded in part by the National Institute on Aging (K01AG044433), and the National Library of Medicine (R01LM011838).

Acknowledgements

The authors would like to acknowledge the support of Nick Tatonetti, Lee Evans, and Nigam Shah for their expertise surrounding FDA Adverse Event Reporting System data. We would also additionally like to thank Lee Evans for his support of the underlying infrastructure that LAERTES currently sits.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.12.005.

References

[1] Safety of Medicine, A guide to detecting and reporting adverse drug reactions, [PDF] 2002 03/24/2012], Available from: <http://whqlibdoc.who.int/hq/2002/WHO_EDM_QSM_2002.2.pdf>.

[2] J.R. Nebeker, P. Barach, M.H. Samore, Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting, Ann. Intern. Med. 140 (10) (2004) 795–801.

[3] J.C. Bouvy, M.L. De Bruin, M.A. Koopmanschap, Epidemiology of adverse drug reactions in Europe: a review of recent observational studies, Drug Saf. 38 (5) (2015) 437–453.

[4] A. Miguel et al., Frequency of adverse drug reactions in hospitalized patients: a systematic review and meta-analysis, Pharmacoepidemiol. Drug Saf. 21 (11) (2012) 1139–1154.

[5] J. Lazarou, B.H. Pomeranz, P.N. Corey, Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies, JAMA 279 (15) (1998) 1200–1205.

[6] J. Duke, J. Friedlin, P. Ryan, A quantitative analysis of adverse events and "overwarning" in drug labeling, Arch. Intern. Med. 171 (10) (2011) 944–946.

[7] J. Duke, J. Friedlin, X. Li, Consistency in the safety labeling of bioequivalent medications, Pharmacoepidemiol. Drug Saf. 22 (3) (2013) 294–301.

[8] Guidance for Industry, Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment, [PDF] 2005.03.22, Available from: <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071696.pdf>.

[9] J.A. Berlin, S.C. Glasser, S.S. Ellenberg, Adverse event detection in drug development: recommendations and obligations beyond phase 3, Am. J. Public Health 98 (8) (2008) 1366–1371.

[10] R.E. Behrman et al., Developing the sentinel system—a national resource for evidence development, N. Engl. J. Med. 364 (6) (2011) 498–499.

[11] P.B. Ryan et al., A comparison of the empirical performance of methods for a risk identification system, Drug Saf. 36 (Suppl. 1) (2013) S143–S158.

[12] M.J. Schuemie et al., Interpreting observational studies: why empirical calibration is needed to correct p-values, Stat. Med. 33 (2) (2014) 209–218.

[13] P.B. Ryan et al., Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership, Stat. Med. 31 (30) (2012) 4401–4415.

[14] M.J. Schuemie et al., Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods, Med. Care 50 (10) (2012) 890–897.

[15] G. Hripcsak et al., Observational health data sciences and informatics (OHDSI): opportunities for observational researchers, Stud. Health Technol. Inform. 216 (2015) 574–578.

[16] R.D. Boyce et al., Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest, Drug Saf. 37 (8) (2014) 557–567.

[17] Observational Health Data Sciences and Informatics (OHDSI) Vocabulary Resources, [web page] 2015 (2015.06.03), Available from: <http://www.ohdsi.org/data-standardization/vocabulary-resources/>.

[18] V. Huser, et al., Piloting a Comprehensive Knowledge Base for Pharmacovigilance Using Standardized Vocabularies, [Web Page] 03/26/2015, Available from: <http://www.slideshare.net/boycer/piloting-acomprehensivepharmacovigilanceknowledgebase2015v2>.

[19] KnowledgeBase GitHub, [Web Page], Available from: <https://github.com/OHDSI/KnowledgeBase/>.

[20] R. Boyce, et al., LAERTES: An open scalable architecture for linking pharmacovigilance evidence sources with clinical data, 2016 1-AUG-2016, Available from: <http://icbo.cgrb.oregonstate.edu/node/354>.

[21] E.P. van Puijenbroek et al., A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions, Pharmacoepidemiol. Drug Saf. 11 (1) (2002) 3–10.

[22] J.M. Banda et al., A curated and standardized adverse drug event resource to accelerate drug safety research, Sci. Data 3 (2016) 160026.

[23] P. Avillach et al., Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project, J. Am. Med. Inform. Assoc. 20 (3) (2013) 446–452.

[24] H. Kilicoglu et al., Constructing a semantic predication gold standard from the biomedical literature, BMC Bioinformatics 12 (2011) 486.

[25] Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT), Adverse Drug Reactions Database, [web page] (2015.05.07), Available from: <http://www.imi-protect.eu/adverseDrugReactions.shtml>.

[26] P.B. Ryan et al., Defining a reference set to support methodological research in drug safety, Drug Saf. 36 (Suppl 1) (2013) S33–S47.

[27] P.M. Coloma et al., A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases, Drug Saf. 36 (1) (2013) 13–23.

[28] M.J. Schuemie et al., Detecting adverse drug reactions following long-term exposure in longitudinal observational data: the exposure-adjusted self-controlled case series, Stat. Meth. Med. Res. (2014).

[29] M.J. Schuemie et al., Replication of the OMOP experiment in Europe: evaluating methods for risk identification in electronic health record databases, Drug Saf. 36 (Suppl 1) (2013) S159–S169.

[30] P.E. Stang et al., Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership, Ann. Intern. Med. 153 (9) (2010) 600–606.

[31] J. Tisdale, D. Miller, Drug-Induced Diseases: Prevention, Detection and Management, second ed., American Society of Health-System Pharmacists (ASHP), 2010.

[32] G. Trifiro et al., The EU-ADR project: preliminary results and perspective, Stud. Health Technol. Inform. 148 (2009) 43–49.

[33] CredibleMeds(R), [web page] (2015.06.18), Available from: <https://www.crediblemeds.org/>.

[34] R.L. Woosley, K. Romero, Assessing cardiovascular drug safety for clinical decision-making, Nat. Rev. Cardiol. 10 (6) (2013) 330–337.

[35] Usagi, [web page] (2015.05.21 2015.06.17), Available from: <http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi>.

[36] A. Airola et al., An experimental comparison of cross-validation techniques for estimating the area under the ROC curve, Comput. Stat. Data Anal. 55 (4) (2011) 1828–1844.

[37] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (3) (1988) 837–845.

[38] R: A language and environment for statistical computing, R Foundation for Statistical Computing, 2015, R Core Team, Vienna, Austria.

[39] M.A. Suchard et al., Massive parallelization of serial inference algorithms for a complex generalized linear model, ACM Trans. Model. Comput. Simul. 23 (1) (2013) 1–17.

[40] M. Hauben, J.K. Aronson, R.E. Ferner, Evidence of misclassification of drug-event associations classified as gold standard 'negative controls' by the observational medical outcomes partnership (OMOP), Drug Saf. 39 (5) (2016) 421–432.

[41] P. Ryan, M. Schuemie, Evaluating performance of risk identification methods through a large-scale simulation of observational data, Drug Saf. 36 (1) (2013) 171–180.

[42] R. Harpaz et al., Text mining for adverse drug events: the promise, challenges, and state of the art, Drug Saf. 37 (10) (2014) 777–790.

[43] J.M. Reps et al., A supervised adverse drug reaction signalling framework imitating Bradford Hill's causality considerations, J. Biomed. Inform. 56 (2015) 356–368.

[44] FDA Adverse Event Reporting System (FAERS), Latest Quarterly Data Files, [web page] 2015.09.29 2015.11.13]; Available from: <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm082193.htm>.