HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2018 September 07.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2017 November ; 2017: 1809–1814. doi:10.1109/BIBM.2017.8217935.

Mining FDA resources to compute population-specific frequencies of adverse drug reactions

Aleksandar Poleksic,
Department of Computer Science, University of Northern Iowa, Cedar Falls, USA

Carson Turner,
Department of Computer Science, University of Northern Iowa, Cedar Falls, USA

Rishabh Dalal,
Department of Computer Science, University of Northern Iowa, Cedar Falls, USA

Paul Gray, and
Department of Computer Science, University of Northern Iowa, Cedar Falls, USA

Lei Xie
Department of Computer Science, Hunter College and CUNY Graduate Center, New York, USA

Abstract

Adverse drug reactions (ADRs) represent one of the main health and economic problems in the world. With increasing data on ADRs, there is an increased need for software tools capable of organizing and storing the information on drug-ADR associations in a form that is easy to use and understand. Here we present a step by step computational procedure capable of extracting drug-ADR frequency data from the large collection of patient safety reports stored in the Federal Drug Administration database. Our procedure is the first of its type capable of generating population specific drug-ADR frequencies. The drug-ADR data generated by our method can be made specific to a single patient population group (such as gender or age) or a single therapy characteristic (such as drug dosage, duration of therapy) or any combination of such.

Keywords

adverse drug reaction; ADR; side-effect; FDA; case report

I. INTRODUCTION

Adverse drug reactions (ADRs) are unwanted outcomes of drug treatments. Due to their frequencies and severity of their presentations, ADRs represent one of the main public health problems [1,2]. The clinical impact of ADRs is manifested by frequent visits to emergency departments and extended hospital stays [3]. ADRs are the major cause of failed drug-discovery pipelines and, in turn, loss of revenue by the drug manufacturers. The

poleksic@cs.uni.edu, turnecak@uni.edu, dalalr@uni.edu, paul.gray@uni.edu, lei.xie@hunter.cuny.ed.

accompanying burden on world economy is large and is measured by tens of billions of dollars annually [4].

Understanding side-effects and their associations to drugs is of utmost importance in medical research. Databases of drugs, side-effects and their associations are used to predict other side-effects and ensure the safe use of drugs. Drug-ADR databases can be used to prescribe the correct drugs to patients or to predict the likely side effects of a yet-untested drug. As such, drug-ADR association databases are vital in medicine, biomedical research, and drug discovery.

A vast amount of raw data on drugs and their observed side effects is already made available as part of the Food and Drug Administration database openFDA (https://open.fda.gov/) [5]. OpenFDA was launched in 2014 with the goal of providing easy access to data which, in turn, is expected to educate the public and save lives. The public data is also intended to support biomedical research and drug discovery.

The openFDA data is available for direct download or can be accessed through the Application Programming Interface (API). The data is massive is size (~100 GB) as it contains millions of reports on drugs, medical devices, and food. These categories are further divided into specific types of reports. For example, drug reports are divided into drug adverse events, drug product labeling, and drug recall enforcement reports. There are over 7.4 million drug adverse event reports which can be queried using the API.

The openFDA API provides a large number of search parameters as well as the ability to filter data by the parameters, such as the drug dosage, patient gender, age and weight, to name a few. While adverse event reports contain varying levels of details, they all follow the same general structure. At the top is the information about the report itself, such as the report ID and receive date. Each report contains the information about the patient's statistics (age, weight, sex, etc.). Finally, there is a list of medicinal products the patient is taking and a list of reactions the patient suffered from.

The ultimate goal of our project is to compile and organize the openFDA reports into a collection of drug-ADR databases that will provide the frequency (probability) that any given drug will give rise to any particular ADR. Most importantly, by taking advantage of the openFDA data categorized by patient characteristics and drug treatment, we can compute the relative frequency of an adverse reaction for each drug each given patient population. Availability of patient and treatment specific ADR frequencies is very important since the probability of a drug giving rise to a particular ADR vary among groups. For instance, unlike the pediatric population, the geriatric population has several known issues with antibiotics [6], such as the acute *renal insufficiency*. Other frequently seen ADRs in elderly are *hypokalaemia* and *symptomatic hypotension* [7]. The pediatric population is prone to skin reactions (e.g. *urticaria*) and adverse events affecting the gastrointestinal system, although serious ADRs, such as those related to the central nervous system, are also frequently observed [8].

While the importance of accurate population- and therapy-specific drug-ADR databases cannot be underestimated in biomedical research and drug discovery in general, the field that

can immediately take advantage of our data and programs is the computational prediction of drug side-effects [9,10,11]. Currently, all computational tools for predicting ADRs for a given drug use a single, one-fits-all binary matrix of known drug-ADR associations [12]. Our project is expected to enhance the computational ADR prediction in two directions. First, instead of using a binary (0/1) matrix of known drug-ADR associations (where 1 represents a known association and 0 represents no known association), future computational methods will be able to employ the probabilities of drug-ADR associations (ranging from 0 to 1). Furthermore, the ADR predictions can be made more accurate by utilizing drug-ADR frequencies specific to the population group of interest (gender, age, etc.) or the drug treatment characteristics (such as drug dosage, duration of treatment, etc.) or any combination of such.

II. METHODS

Our research was carried out in three phases. Section A below describes how we identify a representative set of FDA approved drugs according to some well-defined criteria. In section B we describe how we modify and use the FDA API to extract all drugs associated with a given ADR. The work-in-progress procedure for normalizing the output set of drugs and the procedure for generating drug-ADR counts (frequencies) is described in section C.

A. Generating a representative drug set

We were primarily interested in drugs approved for use in the United States. Thus, we began compiling drugs from the Food and Drug Administration (FDA). The FDA has two resources for approved drugs, Drugs@FDA [13] and Orange Book [14]. Both of these databases contain prescription as well as over-the-counter drugs approved for human use in the United States. Both lists also contain approved drugs which have been labeled “discontinued” due to never being marketed, discontinued from marketing, drugs intended for military use, for export only, or drugs having approval withdrawn. These sets of drugs do not contain dietary supplements, animal drugs, or drugs approved outside the United States that have not been approved for marketing inside the country. Of these two drug sets, Orange Book is the smaller one and is a strict subset of Drugs@FDA. This is because Drug@FDA also includes tentatively approved drugs and therapeutic biological products. However, we were only concerned with approved drugs and so we used Orange Book as a starting point to compile our initial drug list.

Orange Book is available for download from the FDA’s website and includes both the ingredients and trade names of each drug. The same ingredient may appear many times under different trade names, and the ingredient information (column) for a drug may actually contain a mixture of multiple ingredients. To obtain a set of unique ingredients, we extracted the ingredient column of the Orange Book’s *products.txt* file and used it as our initial (yet unrefined) set of drugs.

However, when compared to similar work, such as SIDER [15], our set of drugs turned out to be significantly larger. We found this difference was due to three main factors: our set contained mixtures of ingredients, discontinued drugs, and ingredients in pharmaceutical salt forms. The differences due to mixtures and discontinued drugs were intentional and required

no action, but several ingredients appeared in the set multiple times under different salt forms.

Our solution to merge the salt forms of the same ingredients together was to use the RxNorm [16]. RxNorm is a publicly available suite of designated standards that provides normalized names for clinical drugs. Each concept within RxNorm is assigned a term type (TTY) and links to other related concepts. For example, a concept under the Ingredient (IN) TTY may have related salt forms under the Precise Ingredient (PIN) TTY and brand name drugs which contain that ingredient may appear under the Brand Name (BN) TTY. We used RxNorm to check our set of drugs (each represented by the corresponding ingredient specified in the Orange Book), and if any were found to be PINs, we would replace it with the related Ingredients (INs).

Specifically, for each drug in our set, we used the RxNorm Web API to determine the TTY. The API accepts an ingredient (represented as a string in Orange Book) as input and returns the concept in RxNorm including a unique identifying number and the TTY (such as IN or PIN). If the search returned a TTY other than IN, the Web API was used again to find the related IN concept. In rare cases when there were more than one related IN, we chose the best matching IN manually. An illustrative example is provided in Fig. 1.

However, we encountered a problem when it came to finding mixtures from our set in RxNorm. Namely, while RxNorm also contains Mixtures (MINs), the TTYs of IN and MIN are grouped together. In other words, RxNorm's MIN is a combination of INs, and contains no PINs. This created a problem because many of the mixtures from the Orange Book contained at least one PIN which led to the search API failing to find the mixtures in RxNorm.

An attempted solution was to use the approximate search also available in the API. Approximate search would return a large number of results and rank them with a score indicating the best match. Ultimately this proved unreliable in converting PINs to INs. More specifically, the search returned many different candidates tied for the highest score, and these candidates were often MINs with an ingredient from the search string missing, or an unrelated ingredient included.

Since approximate search was not working, we sought a solution to converting salt forms within a mixture to INs using the original API search. While the original API search would not work with the entire mixture string from the Orange Book at once, we could make it work by searching one ingredient at a time. Each mixture was split into its component ingredients, the TTY checking procedure was performed on each component, and then the components were recombined back into a mixture.

As we converted PINs to INs, we saved the salt form (PIN) to ingredient (IN) associations for future use. This was done in the hope of reducing the future run time by loading the associations into a dictionary for constant time lookup and no longer needing to query the RxNorm API.

However we also discovered the need for another dictionary unrelated to PINs. Some INs have multiple synonymous names and the name preferred by Orange Book is different than that preferred by RxNorm. The RxNorm API will correctly direct the Orange Book name to the RxNorm name, but a direct string comparison would fail. As a remedy, we created a second dictionary, this one for non-salt form synonyms to also reduce the number of queries to the RxNorm API. In the end, we had our set of drugs from the Orange Book with the pharmaceutical salt forms removed, along with two lists of synonymous terms to aid in future work.

B. Querying openFDA to generate a list of drugs for a given ADR

The openFDA database provides drugs' side-effect data through its FDA Adverse Event Reporting System (FAERS, https://open.fda.gov/data/faers/). FAERS contains ADR information specific to not only drug indication and dosage, but also to patient sex, onset age, demographics, weight, start and end date of a reaction, etc. Although FAERS mostly contains ADRs for combination drugs, the resource is large enough to provide meaningful single-drug-to-single-ADR relationships. For instance, the latest quarterly release (Oct–Dec 2017) (among 17 published) contains over 300,000 case reports.

Since FAERS case reports are given in the form of .xml files with well-defined tag hierarchy, the first approach we considered was to create a MySQL database containing two distinct tables. One table would specify the relationship between medicinal products and the other would specify the relationship between medicinal products and active substance names. These would later be mapped to the normalized set of drugs found earlier (as explained in section A). We abandoned that approach due to the concern that MySQL databases would not perform well because of the large amount of data (although we could have tried creating multiple MySQL databases and query them in parallel).

Instead of directly downloading and dumping FAERS .xml files into MySQL databases, we decided to take advantage of the recently made available API from FDA. While this approach proved feasible, it was also faced with difficulties. First, an FDA API is not configured to return all safety reports for each query ADR. Specifically, upon getting the first 100 safety reports related to the query, the API query parameter "skip" had to be set in order to access the next 100 safety reports. Since the maximum limit on the "skip" is 50 we could access only (100) + (50*100) = 5100 safety reports. FDA has recently increased the limit on the total number of safety reports to 25,000. However, even the new limit is not large enough, since some query ADRs are found in hundreds of thousands of safety reports.

We thought of two potential ways to address the issue above. The first was to create our own API (using the source code provided by FDA at *github*) and modify it to fit our needs. However, we found that it was simpler to do some adjustments in the processing of data and queries. We could eliminate the upper limit on the number of safety reports returned for a query by decomposing the search. More specifically, we divided original query into several smaller queries by placing the limits on safety report "receive dates". In some cases the increments in safety report's "receive dates" were as small as 10 days (an example case being the drug "aspirin" which shows in 404,389 safety reports). The procedure of decomposing the search is illustrated in Fig. 2.

The other difficulties were encountered when trying to parse the output data for drugs related to query ADRs. In some cases, API queries provided false results. Occasionally, the query ADR is not associated to any drug listed in the output but only appears in the reference section of the safety report. We also found that the metadata section of the output *json* safety report file often had improperly nested parentheses, which created further difficulties. Both of these issues required post-processing of the output files to ensure the accuracy of the drug-ADR association data.

C. Normalizing the drug names in the output files and computing drug-ADR counts

The drugs from the openFDA safety reports contained many spelling errors and random punctuation, making them difficult to compare with our set of drugs. In order to identify these drugs we used the natural-language processing software MetaMap [17], provided by the U. S. National Library of Medicine. MetaMap dramatically improved our ability to identify drugs in the safety reports. One case went from 28.2% of drugs identified before MetaMap and rose to 59.7% identified after using MetaMap.

In order to run MetaMap on the drugs from the openFDA safety reports associated to query ADR, we first output the drugs from the safety report to a file containing one drug per line. This is the format that MetaMap can read. We configured MetaMap to identify terms (rather than complete sentences) by ignoring the order of words in the input text. The output was restricted to only concepts of the specified semantic types. In our case this was Pharmacologic Substance (phsu), Antibiotic (antb), Organic Chemical (orch), and Vitamin (vita). Finally, we configured MetaMap to generate an unformatted XML file with the results. This XML file was then parsed by a Python script to find the results for each drug, if any exist.

In the end, each drug from the safety reports had a pair of names: the original name as it appeared in the safety report, and the name returned by MetaMap (if any). Both of these names were compared to our set of drugs, and if at least one matched, then that drug was identified to be part of our set of drugs.

III. AVAILABLE DATA AND PROGRAMS

The Python scrips for computing the representative set of drugs and querying openFDA safety reports can be downloaded from http://bioinfo.cs.uni.edu/Drug-Adr/. At the same URL, we provide the script getDrugs.py for creating a compact set of FDA approved drugs using FDA's Orange Book and RxNorm as well as two dictionaries: one for replacing salt forms (or other PINs) with their associated INs and another one for translating between vocabularies that use different names for the same ingredient. We also provide the script query_hard.py for compiling the set of safety reports containing the query ADR of interest.

IV. DISCUSSION AND CONCLUSION

Adverse drug reactions (ADRs) represent a major burden on public health and world economy [18,19]. While most ADRs for individual drugs are reported and documented on drug package inserts, limited resources currently exist that provide explicit statistics on drug-

ADR associations. And while current databases of drug-ADR associations, such as SIDER [15], provide comprehensive binary drug-ADR associations, the actual probabilities (frequencies) of those associations are only reported for less than 50% of drug-ADR pairs. Furthermore, the drug-ADR associations in SIDER are only cumulative associations, computed for all patient and treatment categories combined. To the best of our knowledge, patient-specific or treatment-specific drug-ADR frequencies are not currently available in a form that is easy to use and understand.

The method we present here is able to explore large openFDA resources using the FDA provided API and the natural language processing and text mining tools to extract comprehensive data on drug-ADR association counts. Our method is able to generate drug-ADR counts for different population groups such as gender-, age- and demographics-specific drug-ADR associations, drug-dose specific associations, and side-effects arising from combination drugs.

We believe that population specific drug-ADR databases, such as those developed using our methodology, will fuel progress in biomedical research. For instance, current methods for computational prediction of ADRs associated with novel chemicals can immediately take advantage of population specific data to improve accuracy of ADR predictions. At present, methods for computational prediction of ADRs, such as ML (multi-label learning) or CCA (canonical correlation analysis) can predict ADRs for a given drug based upon the knowledge of ADRs associated to other drugs. This knowledge is stored in a binary $m \times n$ matrix R, where each row represents a drug and each column represents an ADR. The entries of R are defined as

$$r_{ij} = \begin{cases} 1 \text{ if drug } i \text{ causes ADR } j \\ 0 \qquad \text{otherwise} \end{cases}$$

Techniques such as collaborative filtering or multi-label learning can also take advantage of known drug-ADR associations to identify ADRs for novel chemicals. Currently, the matrices of drug-ADR associations represent cumulative knowledge, computed for all populations combined. It is reasonable to expect that replacing current, one-fits-all matrices with population specific matrices will increase the accuracy of computational ADR prediction.

Our method for extracting drug-ADR associations can be viewed as a three-step procedure. In the first step we generate a representative set of FDA approved drugs according to some well-defined criteria. We require that our set contains neither animal drugs nor dietary supplements. We also exclude drugs approved outside the United States that have not been approved for marketing inside the country. However, since a larger data sets is still desirable (e.g. for use in computational ADR prediction methods) we do allow into our set the previously approved but discontinued drugs. Drug duplicates are removed using the RxNorm resource that provides links between drug terminologies and pharmacy knowledge systems [16].

To compile the set of drugs responsible for each given ADR, we mine openFDA adverse reaction database (FAERS) using the API provided by FDA. This API is modified to ensure

that each search with a query ADR returns all safety reports containing that ADR. A straightforward post-processing of the results is needed to ensure the accuracy of extracted drug-ADR associations.

Finally, we use a natural language processing tool (namely MetaMap [17]) to map the set of drugs returned in the safety reports to the set of FDA approved drugs compiled in the first step. This ensures better drug coverage and results in more accurate drug-ADR association data.

Our method for computing drug-ADR counts is different and complementary to all existing methods for the same problem. For instance, in contrast to drug-ADR associations extracted from the drug package inserts (as done in SIDER) our associations are derived from patient safety reports deposited in openFDA. In general, there may be multiple drugs and reactions in a single report, which makes it difficult to determine which drug is responsible for which reaction. This uncertainty can be accounted for by taking advantage of large and increasing number of safety reports available in openFDA. Moreover, a simple statistical model can be developed for computing accurate association frequencies that takes account of the overall occurrence of ADRs across all openFDA safety reports.

To increase the database accuracy further, our future research will focus on the openFDA records that pertain to the reported role of each drug in the adverse event. We will also take advantage of the “Potential Signals of Serious Risks/New Safety Information” resource, which provides a listing of possible safety concerns for selected drugs from FAERS. For instance, the July-Sept 2016 quarterly report lists the use of *sensipar* tablets as a potential risk for gastrointestinal bleeding, which is invaluable information in predicting *gastrointestinal bleeding* (or a similar ADR) as an adverse event of drugs similar to *sensipar*.

ACKNOWLEDGMENT

We thank Ms. Nadhar Alsabban, the Customer Success Manager at DrugBank as well as the Drug Information Specialists at the Center for Drug Evaluation and Research, Food and Drug Administration, for their help and assistance during this project. We also thank Mr. Noel Southall at NIH/NCATS for helping us use their G-SRS browser to search for compound structures.

REFERENCES

[1]. Impicciatore P, Choonara I, Clarkson A, Provasi D, Pandolfini C, & Bonati M (2001). Incidence of adverse drug reactions in paediatric in/out patients: a systematic review and meta-analysis of prospective studies. British journal of clinical pharmacology, 52(1), 77–83. [PubMed: 11453893]

[2]. Lazarou J, Pomeranz BH, & Corey PN (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama, 279(15), 1200–1205. [PubMed: 9555760]

[3]. Budnitz DS, Shehab N, Kegler SR, & Richards CL (2007). Medication use leading to emergency department visits for adverse drug events in older adults. Annals of internal medicine, 147(11), 755–765. [PubMed: 18056659]

[4]. Sultana J, Cutroneo P, & Trifirò G (2013). Clinical and economic burden of adverse drug reactions. Journal of Pharmacology and Pharmacotherapeutics, 4(5), 73.

[5]. Kass-Hout TA, Xu Z, Mohebbi M, Nelsen H, Baker A, Levine J, … & Bright RA (2015). OpenFDA: an innovative platform providing access to a wealth of FDA’s publicly available data. Journal of the American Medical Informatics Association, 23(3), 596–600. [PubMed: 26644398]

[6]. Veehof LJG, Stewart RE, Meyboom-de Jong B, & Haaijer-Ruskamp FM (1999). Adverse drug reactions and polypharmacy in the elderly in general practice. European journal of clinical pharmacology, 55(7), 533–536. [PubMed: 10501824]

[7]. Passarelli MCG, Jacob-Filho W, & Figueras A (2005). Adverse drug reactions in an elderly hospitalised population. Drugs & aging, 22(9), 767–777. [PubMed: 16156680]

[8]. Napoleone E (2010). Children and ADRs (adverse drug reactions). Italian journal of pediatrics, 36(1), 4. [PubMed: 20180963]

[9]. Zhang W, Liu F, Luo L, & Zhang J (2015). Predicting drug side effects by multi-label learning and ensemble learning. BMC bioinformatics, 16(1), 365. [PubMed: 26537615]

[10]. Mizutani S, Pauwels E, Stoven V, Goto S, & Yamanishi Y (2012). Relating drug–protein interaction network with drug sideeffects. Bioinformatics, 28(18), i522–i528. [PubMed: 22962476]

[11]. Poleksic A, & Xie L (2017). Predicting serious rare adverse reactions of novel chemicals. bioRxiv, 160473.

[12]. Kuhn M, Campillos M, Letunic I, Jensen LJ, & Bork P (2010). A side effect resource to capture phenotypic effects of drugs. Molecular systems biology, 6(1), 343. [PubMed: 20087340]

[13]. US Food and Drug Administration. (2013). Drugs@ FDA: FDA approved drug products https://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm

[14]. US Food and Drug Administration. (1985). Approved drug products with therapeutic equivalence evaluations. In Approved drug products with therapeutic equivalence evaluations FDA. https://www.fda.gov/Drugs/InformationOnDrugs/ucm129662.htm

[15]. Kuhn M, Letunic I, Jensen LJ, & Bork P (2015). The SIDER database of drugs and side effects. Nucleic acids research, 44(D1), D1075–D1079. [PubMed: 26481350]

[16]. Nelson SJ, Zeng K, Kilbourne J, Powell T, & Moore R (2011). Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association, 18(4), 441–448. [PubMed: 21515544]

[17]. Aronson AR, & Lang FM (2010). An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3), 229–236. [PubMed: 20442139]

[18]. Bouvy JC, De Bruin ML, & Koopmanschap MA (2015). Epidemiology of adverse drug reactions in Europe: a review of recent observational studies. Drug safety, 38(5), 437–453. [PubMed: 25822400]

[19]. Patel KJ, Kedia MS, Bajpai D, Mehta SS, Kshirsagar NA, & Gogtay NJ (2007). Evaluation of the prevalence and economic burden of adverse drug reactions presenting to the medical emergency department of a tertiary referral centre: a prospective study. BMC clinical pharmacology, 7(1), 8. [PubMed: 17662147]

(A)

Mkt. Status	Active Ingredient	Proprietary Name	Appl No	Dosage Form	Route	Strength	TE Code	RLD	RS	Applicant Holder
RX	ABIRATERONE ACETATE	ZYTIGA	N202379	TABLET	ORAL	250MG		RLD	RS	JANSSEN BIOTECH INC
RX	ABIRATERONE ACETATE	ZYTIGA	N202379	TABLET	ORAL	500MG		RLD		JANSSEN BIOTECH INC
Mkt. Status	Active Ingredient	Proprietary Name	Appl No	Dosage Form	Route	Strength	TE Code	RLD	RS	Applicant Holder

Abiraterone Acetate as it appears in Orange Book

(B)

Abiraterone acetate appearing as a PIN in RxNorm, with related IN abiraterone.

(C)

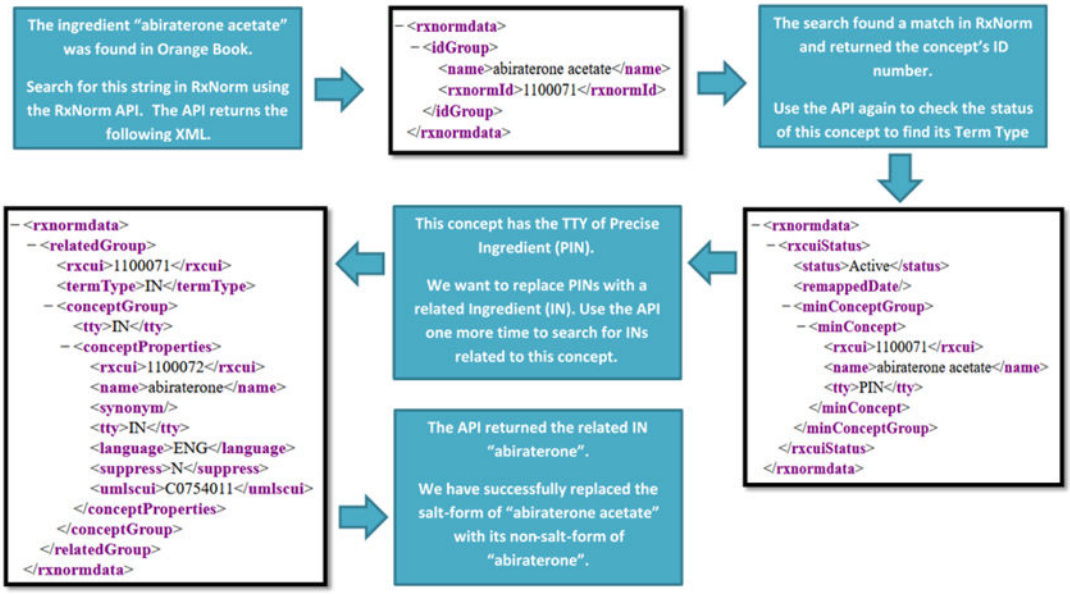

Figure 1.

Normalizing the active ingredient *abiraterone acetate*. Figure A shows the entry of this ingredient in the Orange Book. Figure B shows the mapping of the same ingredient in "salt" form (PIN) to the corresponding (IN) TTY form given in the RxNorm. The detailed mapping procedure is presented in figure C.

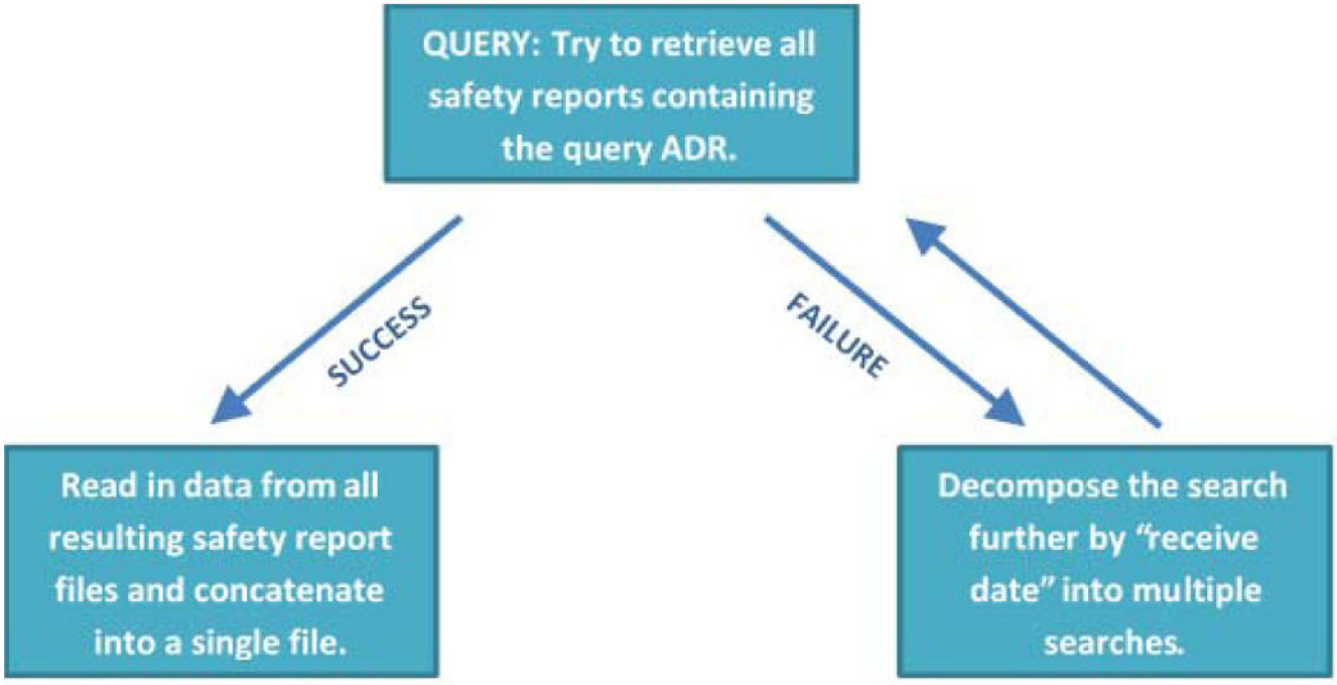

Figure 2.
Decomposing the query search to circumvent the limit on the number of returned safety reports.