HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2015 June 02.

Published in final edited form as:

J Biomed Inform. 2014 February ; 47: 171–177. doi:10.1016/j.jbi.2013.10.008.

Automatic signal extraction, prioritizing and filtering approaches in detecting post-marketing cardiovascular events associated with targeted cancer drugs from the FDA Adverse Event Reporting System (FAERS)

Rong Xu, PhD[1] and **QuanQiu Wang, MS**[2]

[1]Medical Informatics Program, Center for Clinical Investigation, Case Western Reserve University

[2]ThinTek, LLC, Palo Alto, CA

Abstract

Objective—Targeted drugs dramatically improve the treatment outcomes in cancer patients; however, these innovative drugs are often associated with unexpectedly high cardiovascular toxicity. Currently, cardiovascular safety represents both a challenging issue for drug developers, regulators, researchers, and clinicians and a concern for patients. While FDA drug labels have captured many of these events, spontaneous reporting systems are a main source for post-marketing drug safety surveillance in 'real-world' (outside of clinical trials) cancer patients. In this study, we present approaches to extracting, prioritizing, filtering, and confirming cardiovascular events associated with targeted cancer drugs from the FDA Adverse Event Reporting System (FAERS).

Data and Methods—The dataset includes records of 4,285,097 patients from FAERS. We first extracted drug-cardiovascular event (drug-CV) pairs from FAERS through named entity recognition and mapping processes. We then compared six ranking algorithms in prioritizing true positive signals among extracted pairs using known drug-CV pairs derived from FDA drug labels. We also developed three filtering algorithms to further improve precision. Finally, we manually validated extracted drug-CV pairs using 21 million published MEDLINE records.

Results—We extracted a total of 11,173 drug-CV pairs from FAERS. We showed that ranking by frequency is significantly more effective than by the five standard signal detection methods

Correspondence: Medical Informatics Program, Center for Clinical Investigation, Case Western Reserve University, 2103 Cornell Road, Room 6145, Cleveland, OH 44106-3860, Fax: 216-368-0203, rxx@case.edu.

Data Availability

http://nlp.case.edu/public/data/TargetedCancerDrug_Cardiotoxicity_AERS

Author's contributions

Xu and Wang have jointly conceived the idea, designed and implemented the algorithms, and prepared the manuscript.

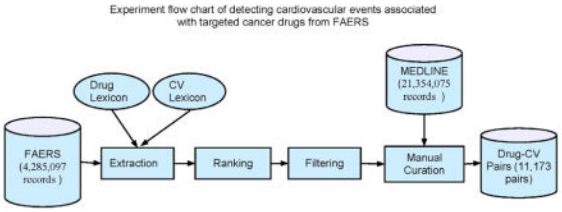
Competing Interests

None

(246% improvement in precision for top-ranked pairs). The filtering algorithm we developed further improved overall precision by 91.3%. By manual curation using literature evidence, we show that about 51.9% of the 617 drug-CV pairs that appeared in both FAERS and MEDLINE sentences are true positives. In addition, 80.6% of these positive pairs have not been captured by FDA drug labeling.

Conclusions—The unique drug-CV association dataset that we created based on FAERS could facilitate our understanding and prediction of cardiotoxic events associated with targeted cancer drugs.

Graphical Abstract



Keywords

targeted cancer therapy; cardiotoxicity; data mining; post-market drug safety surveillance; Personalized Medicine

1. Introduction

Treatment outcomes in cancer patients have dramatically improved since the introduction of targeted drugs. However, targeted drugs are often associated with unexpectedly high cardiovascular toxicity in cancer patients [1–2]. The mechanisms by which targeted drugs become cardiotoxic in cancer patients are not well-understood [3–5]. To ensure personalized cancer treatment, research efforts are needed to understand cardiovascular toxicities associated with targeted drugs. Toxic effects on the heart and vascular system caused by targeted cancer drugs can reveal insights into the biology of cardiovascular disease in humans [6]. Systematic and integrated approaches to studying cardiovascular events associated with targeted drugs have high potential for elucidating the complex pathways of drug-induced toxicities, identifying the on- and off-targets of undesirable cardiovascular events, and predicting unknown cardiotoxicities [7–9]. However, systematic study of targeted drug-induced cardiovascular events has been hampered by the lack of a comprehensive and machine-understandable knowledge base of drug-associated cardiotoxicity in cancer patients. This knowledge is often buried throughout multiple disparate and complementary information sources in various formats, including the U.S. Food and Drug Administration (FDA) (commercial drugs), the vast amount of published biomedical literature (commercial drugs, drugs under preclinical or clinical development, and even failed drugs), the FDA Adverse Event Reporting System (FAERS) (commercial drugs), and patient electronic health records (EHRs) (commercial drugs). Our long-term research goal is to extract drug-side effect (drug-SE) relationships for cancer drugs from all these information sources and build a comprehensive cancer toxicity knowledge base. In this

study, we focus on mining cardiovascular events associated with targeted cancer drugs (drug-CV) from the FDA post-marketing FAERS database, which holds data on more than 4.2 million patients, including 443,226 cancer patients. The unique dataset of drug-CV relationship we have created could facilitate our understanding of the underlying biological mechanisms of the toxic effects under consideration, the development of computational models predictive of adverse events related to targeted cancer therapies, the identification of cancer patients at risk for specific cardiovascular toxic events, and understanding the biology of cardiovascular disease in humans. Ultimately, our database can potentially capacitate the achievement of more effective, safer, and more personalized cancer care. As far as we know, this is the first study to systematically extract, rank, filter, and confirm cardiovascular events associated with targeted anticancer drugs using large amount of published biomedical literature.

Targeted cancer drugs control cancer cell proliferation and spread by interfering with specific molecular targets involved in tumor growth and progression [13]. Targeted cancer therapies have significantly (positively) impacted the survival and quality of life of cancer patients [2]. Currently, approximately 500 novel targeted anti-cancer agents are under preclinical or clinical development or have been approved by the FDA for the treatment of specific types of cancers [13–14]. Targeted cancer drugs promise new ways to personalize cancer treatments based on the unique molecular targets expressed by tumor cells. However, a major challenge posed by these targeted agents lies in maintaining the balance between tumor control and drug-induced toxicity [2], especially cardiotoxicity, which represents one of the most significant complications of targeted drug use in cancer patients [15]. The mechanisms by which targeted agents manifest their cardiotoxicity remain unclear [3–5]. Unlike the side effects induced by non-specific cytotoxic chemotherapeutics, which are often similar, cardiotoxicity associated with targeted cancer drugs often differ among even drugs of the same class such as erlotinib and gefitinib [16]. Currently available approaches in predicting the cardiotoxicity of targeted drugs have had limited success [17], demonstrating the need for researchers to gain mechanistic insight into cardiotoxicity associated with targeted drug use in cancer patients.

Side effects, including drug-induced cardiotoxicity, are observable phenotypes of drugs at the level of the whole body system and are mediated by a drug interacting with its targets through a cascade of downstream pathway perturbations [18]. Systems biology creates the capability to elucidate complex and highly interconnected pathways of toxicities and to develop computational models predictive of unknown toxicities [8–10]. It has been increasingly recognized that similar side effects of seemingly unrelated drugs can be caused by their common off-targets and that drugs with similar side effects are likely to share molecular targets [19]. Therefore, systems approaches to studying the phenotypic relationships among targeted cancer drugs and integration of this drug phenotypic data with genetic and other ‘omics’ data, such as protein-target interactions, chemical structure, and gene co-expression, will allow for a better understanding of drug toxicities. For phenotype-driven systems approaches to understanding targeted drug-induced cardiotoxicity in cancer patients, the availability of a comprehensive and machine-understandable drug-cardiovascular events (drug-CV) relationship knowledge base is critical. At present, no such

data source exists and the published literature on automatically extracting drug-SE relationships for cancer drugs is scant.

It was recently demonstrated that 39% of serious events associated with targeted cancer drugs are not reported in clinical trials and 49% are not described in FDA drug labels [20]. In addition, the accelerated approval of anticancer drugs initiated by the FDA in 1992 may cause the early release of unsafe drugs [21]. Therefore, to build a comprehensive knowledge base of drug-SE relationships for cancer drugs, it is important to extract knowledge from multiple sources, including FDA drug labels, the FDA post-market drug safety surveillance system FAERS, patient EHRs, and the large body of published biomedical literature. Recently, we developed a simple but effective method for extracting a large number of drug-SE pairs for cancer drugs from published biomedical literature abstracts [11]. In this study, we focus on mining targeted cancer drug-related cardiovascular events from FAERS, which includes data for more than 4.4 million patients.

Spontaneous reporting systems are the main resources for post-marketing drug safety surveillance. Mining drug-SE relationships from FAERS is a highly active area. Data mining algorithms such as disproportionality analysis, correlation analysis, and multivariate regression have been developed to detect adverse drug signals from FAERS [22–26]. Current signal detection methods often suffer from a range of limitations including biased reporting and misattribution of causality in drug-SE combinations [27]. Therefore, it is important to develop robust signal detection methods to identify anticancer drug-related adverse events from FAERS. In this study, we developed a simple signal-ranking method that is more effective than standard signal detection approaches. We also developed three signal-filtering methods to further improve precisions. After these initial signal detection and strengthening steps, we confirmed extracted CV pairs by manually looking for evidence in the 21 million published MEDLINE records.

2. Data and Methods

The processing of detecting cardiovascular events associated with targeted cancer drugs is depicted in Figure 1 and consisted of the following steps: (1) Extracting drug-CV pairs from FAERS through named entity recognition and drug mapping processes; (2) Ranking extracted drug-CV pairs by comparing six different ranking approaches; (3) Filtering extracted drug-CV pairs by removing false positives (drug-disease treatment pairs and disease-manifestation pairs) and by extracting drug-CV pairs from patients taking single drug; and (4) Manual curation of all extracted drug-CV pairs that appeared in MEDLINE sentences.

2.1 Data

2.1.1 FDA Adverse Event Reporting System (FAERS)—The FDA FAERS data files for the time period from year 2004 through 2012 were downloaded [28]. A total of 4,285,097 records were extracted from the downloaded datasets. Among the downloaded files, files DRUGyyQq.TXT contain drug information associated with reported adverse event. Files REACyyQq.TXT contain all “Medical Dictionary for Regulatory Activities” (MedDRA) terms coded for an adverse event. Files DRUGyyQq.TXT and REACyyQq.TXT

are the sources of drug-CV association extraction. Files INDIyyQq.TXT contain patients' disease information, which was used in our study to stratify patients into cancer and non-cancer populations. Files DEMOyyQq.TXT contain patients' demographic information, such as age and gender, which was used in our study to stratify patients into different demographic groups.

2.1.2 Lexicon of targeted cancer drugs—A list of 45 targeted cancer drugs (including both generic names and trade names) that are FDA-approved or being studied in ongoing clinical trials was obtained from the National Cancer Institute (NCI) [29].

2.1.3 Lexicon of cardiovascular event (CV) terms—We built a lexicon of CV terms based on MedDRA, which is a widely-used medical terminology to classify adverse event information associated with the use of biopharmaceuticals and other medical products [30]. The adverse events in FAERS are also coded with MedDRA terms. The lexicon was created by finding all leaf nodes with the ancestor "vascular disorders" or "cardiac disorders." This lexicon consisted of a total of 1,712 CV terms, including 1,269 vascular disorders and 527 cardiac disorders.

2.2 Methods

2.2.1 Extraction of drug-CV pairs from FAERS—The drug-CV pairs were extracted by linking DRUGyyQq.TXT with REACyyQq.TXT using patient report ID numbers. We then cleaned up the extracted pairs as following: (1) drug entity recognition and mapping: drug names used in DRUGyyQq.TXT often consisted of drug trade name, generic name, or both. In addition, many drug strings are in free text form. We performed drug entity recognition and mapped trade names to generic names. We first built a drug lexicon consisting of both generic names (i.e. Trastuzumab) and trade names (i.e. Herceptin) for the 45 targeted drugs. The trade names We then recognized drug entities (both trade names and generic names) from drug strings through a simple dictionary-based exact string match process. After named entity recognition, we then mapped all trade names to their corresponding generic names. For example, from the drug string "herceptin (trastuzumab) pwdr + solvent, infusion soln, 440mg", we first recognized the two drug terms "Herceptin" and "trastuzumab" using the drug lexicon that consisted of both drug generic names and trade names. We then mapped the trade name "Herceptin" to its generic name "trastuzumab." Therefore, from the above free-text drug string, we extracted one drug entity "trastuzumab." (2) CV entity recognition: CV entities were recognized from adverse event strings in the REACYYQq.TXT files using a dictionary-based exact string matching approach. The dictionary of CV terms was created as above and consisted of a total of 1,712 CV terms, including 1,269 vascular disorders and 527 cardiac disorders.

After these two steps, we obtained a total of 11,173 drug-CV pairs, representing 39 (out of 45) targeted drugs and 1,097 (out of 1,712) CV events. Six targeted drugs (bosutinib, cabozantinib, carfilzomib, crizotinib, regorafenib, and ziv-aflibercept) were not in the database since the list of targeted cancer drugs included both FDA-approved drugs and drugs still in ongoing clinical trials. These six drugs are presumably still in clinical trial stages.

2.2.1 Prioritizing extracted drug-CV pairs—We ranked the extracted drug-CV pairs using six measures: frequency, relative reporting ration (RRR), proportional reporting ratio (PRR), reporting odds ratio (ROR), phi coefficient (PhiCorr), and information component (IC). Frequency is the co-occurrence count of a drug-CV pair in the database. The RRR, PRR, ROR, PhiCorr, and IC are popular signal detection methods based on frequency analysis of 2 × 2 contingency tables to estimate statistical association of a drug-adverse event pair in the databases [22–27].

In order to compare different ranking methods, we used *11-point interpolated average precision*, which is commonly used to evaluate retrieved ranked lists for search engines [31]. For each ranked list, the interpolated precision was measured at the 11 recall levels of 0.0, 0.1, 0.2, …, 1.0. At each recall level, we calculated the arithmetic mean of the interpolated precision. A composite precision-recall curve showing 11 points was then graphed.

It is often difficult to evaluate the performance of signal detection from FAERS since we don't have a gold standard that accurately represents the space of all true positive signals (including both known true positives and unknown true positives) reported in FAERS. In order to compare these six ranking approaches in prioritizing true signals, we used a total of 259 drug-CV pairs extracted from FDA drug labels as the evaluation dataset. These pairs contain not only true signals reported in pre-marketing controlled clinical trials (major portion) but also many confirmed post-marketing signals. Note this evaluation dataset was not intended to calculate the true precisions and recalls, but to compare the six ranking approaches in prioritizing true signals. The advantages in this evaluation dataset are that it is mainly comprised true positives and it is not biased towards specific drugs or CV events. The limitation is that we cannot measure the true precision and recall of drug-CV pair extraction from FAERS.

For creating the gold standard, we downloaded a total of 44,979 drug labels, including 21,610 Human prescription labels, and 23,369 Human OTC labels, from DailyMed [32]. DailyMed is maintained by the National Library of Medicine (NLM) and provides high quality information about marketed drugs. Drug labeling on DailyMed is the most recent FDA labels (package inserts). Majority of the drug side effect information on FDA drug labels was obtained from clinical trials and some was from post-marketing surveillance. We used the publically available information retrieval library Lucene (http://lucene.apache.org) to create a local FDA drug label search engine with indices created on both drugs, section headers in the labels such as "Indications," "Contraindications," and "Adverse Reactions" and sentences. Each sentence was associated with a drug and a subsection header name. We used each of the 45 * 1,712 drug-CV combinations (45 targeted drugs and 1,712 CV terms) as search queries to the local FDA drug label search engine. Drug-CV pairs that appeared in sentences with header "Adverse Reactions" were retrieved. We extracted a total of 259 drug-CV pairs from FDA drug labels, representing 15 targeted drugs and 75 CV events. These pairs are of high quality and represent known cardiovascular events associated with targeted cancer drugs.

2.2.3 Filtering extracted drug-CV pairs—We investigated three filtering schemes to improve the precision of the extracted drug-CV pairs. Normally, for a patient taking *n* drugs

and reporting *m* CV events, a total of *n* * *m* drug-CV pairs are possible. At least three factors can contribute to false positives: (1) misattribution among drugs and CVs; (2) some of the reported side effects are in fact indications of some of the drugs a patient is taking; and (3) the reported side effects are in fact manifestations of the diseases. We developed three different filtering algorithms to deal with each of the above-mentioned scenarios. The filtered drug-CV pairs were then ranked. Ranked performance of the filtered pairs was compared to that of unfiltered pairs.

Filter 1: Extracting drug-CV pairs from patients taking a single drug: As is later shown, cancer patients in FAERS, on average, took 4.62 drugs at the same time. Therefore, misattribution between drugs and CV events can be a significant problem contributing to false positives. The first filtering approach was to extract drug-CV pairs from patients who only took one drug, which is a targeted drug, and also reported at least one CV event.

Filter 2: removing known drug-disease treatment pairs from extracted drug-CV pairs: As our Results section indicates, about 25% of drug-CV pairs that appeared in both FAERS and in biomedical literature were in fact drug-disease treatment pairs. Our second filtering approach was to systematically remove all known drug-disease treatment pairs from extracted drug-CV pairs. We compiled a large dataset consisting of 184,442 drug-disease treatment pairs by combining information from FAERS (52,066 pairs) and clinicaltrials.gov (139,669 pairs). Pairs from FAERS were extracted by linking DRUGyyQq.TXT to INDIyyQq.TXT (with named entity recognition and mapping for both drugs and diseases). Drug-disease treatment pairs from clinicaltrials.gov were generated in one of our recent studies [11]. For each patient, we filtered out known drug-disease treatment pairs from the drug-CV pairs.

Filter 3: removing known disease-CV manifestation associations from patient records: Cardiovascular diseases often co-occur in cancer patients since the incidence of both increases with age. Therefore it is likely that the reported cardiotoxicities are in fact the clinical manifestations of co-morbid cardiovascular events in cancer patients. We extracted a total of 50,551 disease-manifestation pairs from the Unified Medical Language System (UMLS) (2011 version) file MRREL.RRF [33]. We then expanded the terms in the pairs to include all the synonyms in order to capture disease term usage variations in FAERS. After expansion, we obtained a total of 3,499,87 pairs, which were then used to filter out side effects that are known manifestations (symptoms) of diseases being treated. For each patient, we simply removed all side effects that are known clinical manifestations of the patient's disease. Then, drug-CV pairs were extracted from the filtered patient records.

2.2.4 Manual confirmation of drug-CV pairs using supporting evidence from MEDLINE—

In one of our previous studies [11], we built a local MEDLINE search engine with indices on a total of 21,354,075 MEDLINE records (119,085,682 sentences) published between 1965 and 2012. For each targeted drug-CV pair extracted from FAERS, we retrieved all of its associated MEDLINE sentences using the local search engine. In total, we retrieved 3,628 sentences from MEDLINE. We then manually classified these pairs into three classes (CAUSE, TREAT and NONE) using the sentences (and abstracts when

necessary) as evidence. Three curators with graduate degrees in clinical or biomedical science performed the curation task. Each curator independently curated all the sentences (a total of 25 hours/each curator). For each drug-CV pairs, if there is one sentence that provides supporting evidence for causal relationship, then it is determined as positive (even though other sentences may show contradictory/inconclusive evidence). Majority vote was used to decide the final classification of each drug-CV pair. We found the inter-annotator agreement is as high as 86%.

3 Results

3.1 Many cancer patients are under targeted therapy and reported significantly more CV events than non-cancer patients

There are a total of 4,285,097 patient records in FAERS, among which 443,226 (10.34%) are of cancer patients and 3,841,871 (89.66%) are of non-cancer patients. About half of cancer patients (47.7%) took at least one targeted drug (Table 1). Only a small percentage of non-cancer patients (2.56%) took targeted drugs. About 39.0% cancer patients taking targeted drugs also reported at least one cardiovascular adverse event, indicating that cardiovascular safety is indeed a significant issue of targeted therapy in cancer patients. Note that even though these 45 target drugs were used in treating cancers, they are also used in treating non-cancer diseases. For example, bevacizumab is FDA-approved drug for the treatment of colorectal cancer and lung cancer, it is also used in treating eye diseases such as age-related macular degeneration (AMD) and diabetic retinopathy. The prevalence of targeted drug-associated CV events in cancer patients is significantly higher than that in non-cancer patients (39.0% vs. 32.7%, Z-score is 12.538 and p-value = 0), indicating that risk factors other than the drug itself, such as drug combination, disease characteristics, and co-morbidities, may contribute to this increased prevalence. Cancer patients also tended to use significantly more drugs (average of 4.62) than non-cancer patients (average of 3.09). The average number of co-morbidities in cancer patients is 1.45, which is also significantly higher than 1.25 in non-cancer patients. Since cancer patients are often under the treatment of multiple drugs and also have co-morbidities, accurately extracting drug-SE pairs, including drug-CV pairs, is a challenging task. In the following sections, we present methods to rank lowly and filter out false positives.

3.2 Ranking by co-occurrence frequency is more effective than standard signal detection methods

We compared six different ranking measures: frequency (Freq), relative reporting ratio (RRR), proportional reporting ratio (PRR), reporting odds ratio (ROR), phi coefficient (PhiCorr), and information component (IC). We extracted a total of 11,173 drug-CV pairs, representing 39 targeted cancer drugs and 1,095 CVs. We used the 259 drug-CV pairs from FDA drug labels as gold standard to compare the precisions at 11 recall values. Note that the precisions measured using this gold standard did not necessarily represent the actual precisions since many true signals may not be captured in FDA drug labels. As shown in Figure 2, ranking by co-occurrence significantly improved the precisions of top-ranked pairs; the precision of top-ranked pairs (recall =0.1) was 0.208, a more than 800% elevation in precision compared to 0.023 at recall of 1.0 (the whole list). In addition, ranking by co-

occurrence is more effective than all five other methods. At a recall of 0.1, the precision was 0.208 for frequency-based ranking, which is 246% increase compared to the 0.06 for ranking by PRR. Ranking by all other five methods had no effects on ranking known drug-CV pairs highly (data for RRR, ROR and IC were similar to that for PRR and is not shown). The ineffectiveness of the widely adopted signal detection methods such as PRR in ranking targeted drugs-related CVs may be partly due to inherent data biases. In fact, many known drug-CV pairs from FDA drug labels had low, even negative, PRR values.

Even though the top pairs ranked by frequency were significantly enriched with true positives compared to the whole list, the precision was still very low (0.208 at a recall of 0.1 and 0.180 at a recall of 0.2). There are at least two explanations for these low precisions. First, not all signals from FAERS were included in FDA drug labels, resulting in false negatives. The information from FDA drug labels was mainly sourced from controlled clinical trials, which do not necessarily reflect the 'real-world' clinical practice. The information from FAERS reflects the actual clinical situation, where the presence of co-morbidities, age, and other risk factors may contribute to cardiovascular events in cancer patients. As shown later in this study, 80.9% true positive drug-CV pairs that appeared in both FAERS and published literature were not captured by the FDA drug labels, clearly demonstrating that using drug-CV pairs included in FDA drug labels as gold standard will greatly under-estimate the actual precision of drug-CV pairs extracted from FAERS. The second cause of the low precision of top-ranked pairs may be true negatives, since not all top-ranked pairs are necessarily true signals. Even if a drug and a CV event appear together many times, it is still possible that the underlying diseases or other drugs in a combination therapy may contribute to the CV event. On the other hand, even though a cardiovascular event appeared with a targeted drug only once in the database, it is still likely to be a true association, which may only appear in some specific patients with certain demographic and disease characteristics. Another caveat in using well-known drug-SE pairs from FDA drug labels as gold standard is that it will particularly underestimate the precision of low-incidence or idiosyncratic events in cancer patients. In fact, to identify, understand, and predict these rare events in cancer patients is perhaps the most important and challenging task.

3.3 Drug-CV pairs extracted from patients taking only one drug have the most improved precision of all pairs

At least three possible factors may contribute to the low precision of extracted drug-CV pairs: (1) inclusion of CV events caused by drugs other than the targeted cancer drugs; (2) inclusion of drug-disease treatment pairs; and (3) the possibility that reported adverse events are clinical manifestations of a patient's existing diseases.

We extracted a total of 4,549 drug-CV pairs from patients who took only one drug and reported at least one CV event, and ranked the pairs by frequency. As shown in Figure 3, these pairs had higher precisions than unfiltered pairs at 9 recalls. The overall precision of drug-CV pairs extracted from patients taking single drug is 0.044, which represents a 91.3% increase from the precision of 0.023 for unfiltered pairs.

However, the other two filtering approaches, filtering out known drug-disease treatment pairs from extracted drug-CV pairs and removing known clinical manifestations from reported side effects, had no effect on precisions (Figure 3). As our results show, about 25% of drug-CV pairs that appeared in both FAERS and biomedical literature were in fact drug-disease treatment pairs. The ineffectiveness in improving precision by filtering out known drug-disease treatment pairs may be due to the limited coverage of our compiled drug-disease treatment pairs, even though it consisted of 184,442 pairs. The same explanation can apply to the filtering approach by removing known disease manifestations.

The higher precision of pairs extracted from patients taking a single drug when evaluated with pairs from FDA drug labels reflects the fact that these patients may be more similar to patients in controlled clinical trials (less co-comorbities, taking less multiple drugs simultaneously). Some CV events may manifest only when targeted drugs are used in combination with each other or with other cancer therapies such as chemotherapy or radiotherapy.

3.4 Systematically curation of drug-CV pairs that appeared in MEDLINE sentences

We used evidence from the vast amount of published literature to systematically confirm extracted drug-CV pairs. Among 11,173 drug-CV pairs extracted from FAERS, only 617 pairs also appeared in MEDLINE sentences. We retrieved a total of 3628 sentences that these 617 pairs appeared in and manually curated these sentences. Among the 617 pairs that also appeared in MEDLINE sentences, 320 pairs (51.9%) were true positive (CAUSE) pairs, demonstrating that if a pair appears in both FAERS and in the literature, it is highly likely to be a true signal. 154 (25.0%) out of the 617 pairs were in fact drug-disease treatment pairs, demonstrating that the inclusion of drug-disease treatment pairs can adversely affect precision. The reason that our strategy of filtering out known drug-disease treatment pairs was not effective in improving precision might be due to the limited coverage of the drug-disease treatment dataset. The rest of the 143 pairs (23.1%) have no obvious semantic relationships based on the evidence sentences. More significantly, among the 320 true positive pairs, 258 pairs (80.6%) are not included in the FDA drug labels even though there exists evidence from both FAERS and published literature. These pairs along with their associated MEDLINE sentences and FAERS counts are available at: http://nlp.case.edu/public/data/TargetedCancerDrug_Cardiotoxicity_AERS/3_manual_curation.

In summary, even though the presence of FAERS reported drug-CV pairs in MEDLINE sentences is low, among the 617 pairs that did appear in MEDLINE sentences, 51.9% are true positives. 80.6% of these true positives are not included in current FDA drug labels, demonstrating the importance of active monitoring of post-marketing cardiovascular events reportedly associated with targeted drugs in cancer patients. As we mentioned before, the knowledge in FAERS and in published literature is largely complementary. In this study, we only checked 618 drug-CV pairs that appeared in both FAERS and in the literature. However, we don't know how many drug-CV pairs appeared in literature, but not in FAERS or FDA drug labels.

4. Discussion

In this study, we systematically mined and manually confirmed post-marketing cardiovascular events associated with targeted anticancer drugs. To the best of our knowledge, this is the first attempt at creating a knowledge base of drug-CV associations for targeted cancer drugs. However, several limitations must be considered. First, the true precision of the extracted drug-CV pairs is unknown. Evaluations using drug-CV pairs extracted from FDA drug labels may have significantly underestimated the true precision. The cardiovascular adverse events for targeted cancer drugs from FDA drug labels were mainly from clinical trials and the ones in FAERS are from 'real-world' patients (older patients or patients with significant co-morbidities). Manual evaluation with evidence from MEDLINE also has its own limitations and biases since we can only calculate the precision of the pairs that appeared in both FAERS and MEDLINE. Second, targeted cancer therapies are often used in combination with other cancer treatments, including other targeted therapies, chemotherapy, and radiotherapy. The risk of certain cardiovascular events may increase when the targeted drugs are used in combination. At present, the drug-CV pairs we extracted from FAERS are only for single drugs. It will be interesting to investigate whether combinations of targeted cancer drugs may lead to specific cardiotoxicities due to synergistic actions. Finally, it will be important to identify potential risk factors for drug-induced cardiovascular events. These risk factors may include patient demographics such as age and gender, patient characteristics such as smoking status and physical stress, disease characteristics such as cancer types and co-morbidities, and prior cancer treatments. All these factors can potentially affect the presence and severity of cardiovascular events associated with targeted drug therapy in cancer patients. Ideally, results from this study will guide similar future studies of toxic effects of cancer treatments.

In this study, we mined targeted drug-associated cardiovascular events from FAERS. To make the knowledge base more complete, other data sources are necessary, including the vast amount of published biomedical literature [11], the patient electronic medical records [36] and even the Web [37–38]. For example, a recent study shows that mining consumers' web search history can reveal unreported side effects of drugs or drug combinations [37]. Our study in building a targeted cancer drug-associated cardiotoxicity knowledge base from FAERS is the first step towards mechanistic understanding of these adverse events. In-depth meta-analysis or network-based systems approach can be performed by combining this knowledge base with other data, such as patients' age, gender, disease characteristics, drug-drug combinations, drug targets, and drug metabolism.

5. Conclusions

In this study, we present automatic approaches in mining cardiovascular events associated with targeted cancer drugs from FDA post-marketing FAERS database. We have built a database consisted of 11,173 post-marketing cardiovascular events associated with targeted cancer drugs. We have developed signal extraction, ranking, filtering, and confirming approaches to improve the precision of the dataset. The unique drug-CV relationship dataset we created could capacitate the development of computational models that would in turn

facilitate our understanding and prediction of cardiotoxic effects associated with targeted drugs, and help us to achieve more effective, safer, and more personalized cancer care.

Acknowledgments

Funding statement

This work by RX was supported by Case Western Reserve University/Cleveland Clinic CTSA Grant (UL1 RR024989), and the work by QW was supported by ThinTek LLC.

References

1. Cleeland CS, Allen JD, Roberts SA, Brell JM, Giralt SA, Khakoo AY, Skillings J. Reducing the toxicity of cancer therapy: recognizing needs, taking action. Nature Reviews Clinical Oncology. 2012
2. Keefe DM, Bateman EH. Tumor control versus adverse events with targeted anticancer therapies. Nature Reviews Clinical Oncology. 2011; 9(2):98–109.
3. Ewer MS, Ewer SM. Cardiotoxicity of anticancer treatments: what the cardiologist needs to know. Nature Reviews Cardiology. 2010; 7(10):564–575.
4. Schmidinger M, Zielinski CC, Vogl UM, Bojic A, Bojic M, Schukro C, et al. Force T, Ewer MS, de Keulenaer GW, Suter TM, Anker SD, Shah AM. Cardiovascular side effects of cancer therapies: a position statement from the Heart Failure Association of the European Society of Cardiology. European journal of heart failure. 2011; 13(1):1–10. [PubMed: 21169385]
5. Mellor HR, Bell AR, Valentin JP, Roberts RR. Cardiotoxicity associated with targeting kinase pathways in cancer. Toxicological Sciences. 2011; 120(1):14–32. [PubMed: 21177772]
6. Moslehi J, Cheng S. Cardio-Oncology: It Takes Two to Translate. Sci Transl Med. May 29.2013 5(187):187fs20.
7. Gibb S. Toxicity testing in the 21st Century. A vision and a strategy. Reprod Toxicol. 2008; 25:136–8. [PubMed: 18093799]
8. Raschi E, De Ponti F. Cardiovascular toxicity of anticancer-targeted therapy: emerging issues in the era of cardio-oncology. Internal and emergency medicine. 2012; 7(2):113–131. [PubMed: 22161318]
9. Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends in pharmacological sciences. 2010; 31(3):115–123. [PubMed: 20117850]
10. Turteltaub KW, Davis MA, Burns-Naas LA, Lawton MP, Clark AM, Reynolds JA. Identification and elucidation of the biology of adverse events: the challenges of safety assessment and translational medicine. Clinical Cancer Research. 2011; 17(21):6641–6645. [PubMed: 22046025]
11. Xu R, Wang Q. Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug side effect relationships from literature. J Am Med Inform Assoc. 2013 Published Online First: 18 May 2013. 10.1136/amiajnl-2012-001584
12. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm
13. http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted
14. Chen MH, Kerkelä R, Force T. Mechanisms of cardiac dysfunction associated with tyrosine kinase inhibitor cancer therapeutics. Circulation. 2008; 118:84–95. [PubMed: 18591451]
15. Ravaud A. How to optimise treatment compliance in metastatic renal cell carcinoma with targeted agents. Annals of oncology. 2009; 20(suppl 1):i7–i12. [PubMed: 19430007]
16. Bonura F, Di Lisi D, Novo S, D'Alessandro N. Timely Recognition of Cardiovascular Toxicity by Anticancer Agents: A Common Objective of the Pharmacologist, Oncologist and Cardiologist. Cardiovascular toxicology. 2012; 12(2):93–107. [PubMed: 21894547]
17. Force T, Kolaja KL. Cardiotoxicity of kinase inhibitors: the prediction and translation of preclinical models to clinical outcomes. Nature Reviews Drug Discovery. 2011; 10(2):111–126.

18. Liebler DC, Guengerich FP. Elucidating mechanisms of drug-induced toxicity. Nature reviews Drug discovery. 2005; 4(5):410–420.
19. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008; 321(5886):263–266. [PubMed: 18621671]
20. Seruga B, Sterling L, Wang L, Tannock IF. Reporting of serious adverse drug reactions of targeted anticancer agents in pivotal phase III clinical trials. J Clin Oncol. 2011; 29:174–185. [PubMed: 21135271]
21. Richey EA, Lyons EA, Nebeker JR, Shankaran V, McKoy JM, Luu TH, et al. Accelerated approval of cancer drugs: improved access to therapeutic breakthroughs or early release of unsafe and ineffective drugs? J Clin Oncol. 2009; 27:4398–4405. [PubMed: 19636013]
22. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics. 2012; 91(6):1010–1021. [PubMed: 22549283]
23. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiology and drug safety. 2001; 10(6):483–486. [PubMed: 11828828]
24. Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, Friedman C. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. Journal of the American Medical Informatics Association. 2013; 20(3):413–419. [PubMed: 23118093]
25. Wilson AM, Thabane L, Holbrook A. Application of data mining techniques in pharmacovigilance. British journal of clinical pharmacology. 2004; 57(2):127–134. [PubMed: 14748811]
26. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiology and drug safety. 2009; 18(6):427–436. [PubMed: 19358225]
27. Stephenson WP, Hauben M. Data mining for signals in spontaneous reporting databases: proceed with caution. Pharmacoepidemiology and drug safety. 2007; 16(4):359–365. [PubMed: 17019675]
28. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm083765.htm
29. http://www.cancer.gov/cancertopics/factsheet/Therapy/targeted
30. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Safety. 1999; 20(2):109–117. [PubMed: 10082069]
31. Manning, CD.; Raghavan, P.; Schütze, H. Introduction to information retrieval. Vol. 1. Cambridge: Cambridge University Press; 2008.
32. http://dailymed.nlm.nih.gov/dailymed/downloadLabels.cfm
33. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004; 32(suppl 1):D267–D270. [PubMed: 14681409]
34. Friedman, C. Artificial Intelligence in Medicine. Springer; Berlin Heidelberg: 2009. Discovering novel adverse drug events using natural language processing and mining of the electronic health record; p. 1-5.
35. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. Journal of the American Medical Informatics Association. 2013; 20(3):404–408. [PubMed: 23467469]
36. Leaman, R.; Wojtulewicz, L.; Sullivan, R.; Skariah, A.; Yang, J.; Gonzalez, G. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the 2010 workshop on biomedical natural language processing; Association for Computational Linguistics; 2010 Jul. p. 117-125.

Highlights

- Targeted cancer drugs are often associated with unexpectedly high cardiotoxicity.
- The Mechanisms underlying these drug-cardiovascular (CV) relationships are unclear
- Systems approaches in studying targeted cancer drug-CV relationships are important.
- There exists no knowledge base of targeted drug-CV relationships.
- We extract, rank, filter, and curate post-marketing drug-CV pairs from FAERS.

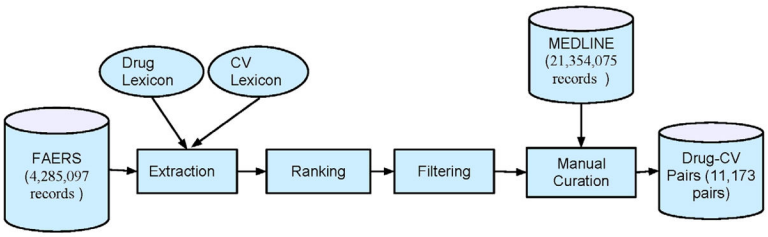


Figure 1.
Experimental flow chart.

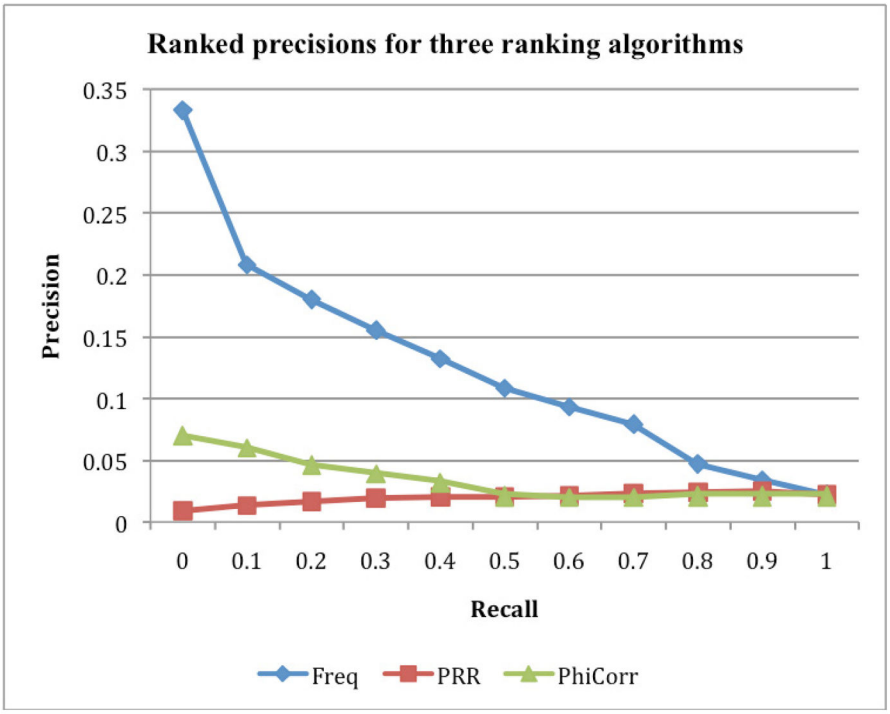


Figure 2.

Ranked precisions at 11 recalls for three ranking measures: frequency (Freq), proportional reporting ratio (PRR) and phi coefficient (PhiCorr).

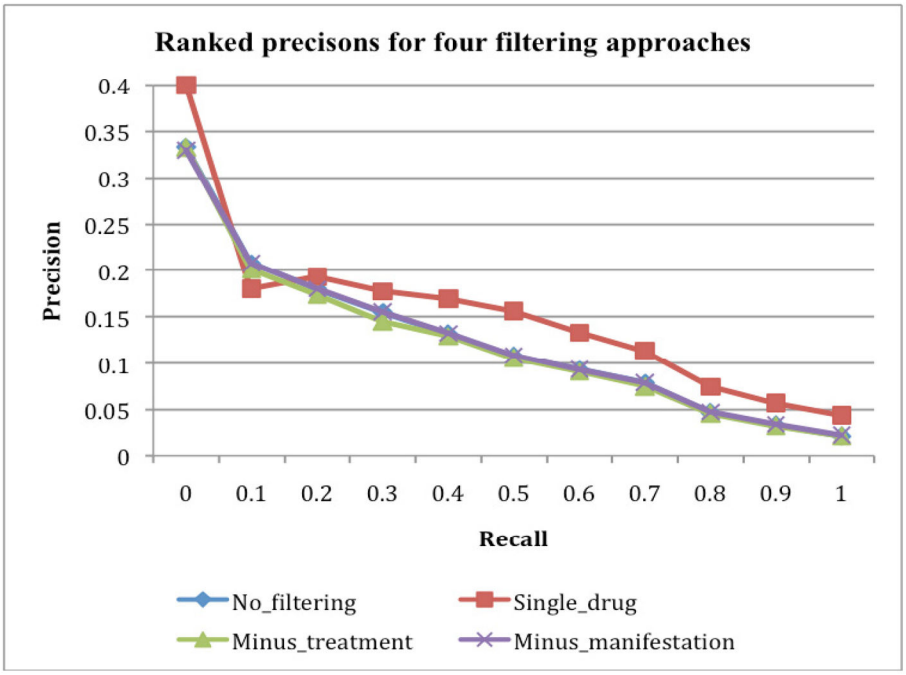


Figure 3.

Ranked precisions at 11 recalls for three different filtering schemes: “Single_Drug_Patients” (pairs from patients taking one drug), “Minus_Treatement_Pairs” (filtered out drug-disease treatment pairs), “Minus_manifestation” (filtered out disease manifestations).

Table 1

Targeted drug usage and prevalence of CV events in cancer versus non-cancer patients

Patients	Targeted drugs	CVs_targeted	Polypharmacy	Comorbidity
Cancer patients N =443,226 (10.34%)	47.7%	39.0%	4.62	1.45
Non-Cancer patients N = 3,841,871 (89.66%)	2.56%	32.7%	3.09	1.25

Table 2

Manual curation of all drug-CV pairs that appeared in both FAERS and MEDLINE sentences.

Pairs	CAUSE	TREAT	NONE	CAUSE pairs not included in FDA drug labels
617	320	154	143	258
	51.9%	25.0%	23.1%	**80.6%**