Middle-East Journal of Scientific Research 24 (S2): 163-168, 2016
ISSN 1990-9233
© IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.S2.138

Classification of Side Effects Associated with Anticancer Drugs Using Oncological Articles

G. Aruna, G. Priyanka and N. Sivaram

Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India

Abstract: Anticancer drugs that are used for the treatment of cancer generally improve the patients health. However they create number of side effects. These drugs may create serious problem which may even cause death. To avoid this problem it is necessary to have a mechanism that identifies the side effect associated with the drug. To perform this, the articles from Journal of Clinical Oncology (JCO) is collected as the dataset. This articles are given by doctors which includes the treatment given to patients along with the result that were developed for the treatment. The reason to choose the JCO article is that, it is published by the doctors so it may provide data correctness. Once the JCO articles are collected, the articles are classified as side effect related or not using Support Vector Machine(SVM). Only side effect related papers are used. From the SE related article the drug-SE pairs are extracted. The drug-SE pairs are then ranked using term frequency and document frequency. Finally the drugs along with its side effects are obtained.

Key words: JCO · SVM · Drug-SE pair · Term frequency · Document frequency

INTRODUCTION

Cancer drugs can have potentially severe or even toxic side effects (SEs). The effectiveness of a lot of cancer drugs is greatly restricted by their level of toxicity. The majority of cytotoxic cancer drugs are not cancer cell-specific, which leads to various familiar severe SEs such as thrombocytopenia, neutropenia and anemia etc., The biological cancer medicine cetuximab for the treatment of metastatic colorectal cancer produces an acne-like rash, which may be acceptable for many patients. In contrast, the biological agent rituximab (for treating chemotherapy-refractory B-cell non-Hodgkin lymphomas) can cause death, acute kidney failure and cardiac arrest, in fact that would make this treatment unappealing to all but the most sophisticated cancer patients [1]. Anticancer drugs manage cancer cell improvement by interfering with precise molecular targets involved in tumor enlargement and progression. Targeted cancer therapies have considerably (positively) impacted the survival and quality of life of cancer patients [2]. Systematic and integrated approaches to study drug-associated side effects is possible to illuminate the composite pathways of drug-induced toxicities, allowing the recognition of narrative drug targets, prophecy of unknown drug toxicities and relocation of existing drugs for innovative disease indications. System approaches would greatly benefit from the enormous amount of higher-level clinical phenotype data such as observed drug-related side effects [3] and [4]. It has been gradually more recognized that similar side effects of seemingly dissimilar drugs. Anticancer drugs promised innovative ways to personalize cancer treatments based on single molecular targets expressed by tumor cells. Yet, current studies have shown that these innovative drugs are frequently associated with unanticipated high toxicities [5]. Recent meta-analysis studies show that the majority of newly-approved anticancer drugs are extra toxic than regular treatments and are associated with enlarged rates of toxic death, severe adverse events, treatment discontinuation [6] and [7].

Besides the overall toxicity levels, more than a few targeted anticancer drugs are associated with unexpected toxicities, such as cardiovascular events, that are idiosyncratic and their underlying molecular mechanisms remain largely unknown [8].Unlike side effects induced by cytotoxic chemotherapeutics, which are like among drugs, side effects associated with anticancer drugs often differ among drugs of the similar class such as erlotinib and gefitinib [9]. These toxicities may be caused by the receptor cross-reactivity, the occurrence of receptors on normal cells [10] or the multiplicity of affected off-target proteins. In order to maintain the balance among tumor control and drug-induced toxicities, research is needed to improve our understanding of the molecular mechanisms of anticancer drug-related toxicities [11].

Corresponding Author: G. Aruna, Department of Computer Science and Engineering,
Mepco Schlenk Engineering College, Sivakasi, India.

The accessibility of a comprehensive side effect knowledge base for targeted drugs and innovative computational methodologies to predicting unanticipated toxicities are significant for the successful enlargement of anticancer agents. The FDA Adverse Event Reporting System (FAERS) is the spontaneous reporting system overseen by the U.S. FDA and the main resources for post-marketing drug safety surveillance. Mining drug–side effect (drug–SE) relationships from FAERS is a highly active research area [12] and [13] Data mining methods such as disproportionality analysis, multivariate regression, correlation analysis and signal ranking and filtering leveraging external knowledge have been developed to detect adverse drug signals from FAERS. Another important information source of drug–SE relations is the vast amount of published biomedical literature. Now, more than 22 million biomedical abstracts are publicly accessible on MEDLINE, making it a source of side effect information for drugs at all clinical phases, including drugs in pre-marketing clinical trials, post-marketing clinical case reports and clinical trials. Statistical, machine learning and signal ranking methodologies have been established in extracting drug–SE pairs from free-text MEDLINE abstracts [14][15] and[16]. In summary, Anticancer drugs are used to improve the treatment outcomes in cancer patients. However, these innovative agents are often associated with unexpected side effects. These side effects are not well understood. The availability of comprehensive knowledge base of side effects associated with targeted anticancer drugs has the potential to illuminate complex pathways underlying toxicities induced by these innovative drugs.

Survey: Literature reports on the corpora that are annotated to facilitate information extraction from biomedical literature, electronic health records and other textual data. In their annotations, Gurulingappa et al. [17] excluded names of medical devices or hospital chemicals. Also, they only annotate drug name mentioned in relation to an adverse event. The final corpus includes those sentences from the abstracts that had at least individual reference of an adverse effect. VanMulligen et al. [18] reports a widely available annotated corpus of biomedical literature where instances of drugs, disorders, genes and the relationships between the well-known entities are annotated. Deleger et al. [19] created an annotated corpus of clinical records. The corpus was created with two different purposes:

(1) a de-identification mission for which the data was annotated for individual health information such as patient‘s age or email address; and (2) annotation of entities associated to the medication, such as medication name and type, as well as disease and symptoms. In most literature-based drug–SE relationship extraction tasks used only the abstracts of biomedical research articles, while we used full-text articles. While full-text articles contain richer drug–SE association knowledge compared to abstracts, they also encompass much noise, which renders the extraction task more challenging.

Related Work: In this study, we developed automatic methodologies to extract anticancer drug–SE pairs from the Journal of Clinical Oncology (JCO).JCO was established in 1983 and is the official journal of the American Society of Clinical Oncology and the foremost journal in oncology. JCO articles comprise a variety of cancer-related investigation articles, including clinical trials reporting drug effectiveness and toxicity in cancer patients, trial reports estimating the effectiveness of biomarkers, clinical case reports and meta-analysis studies, among other article types. JCO articles not only comprise pivotal clinical trials that have led to drug authorization, but also trials that are still in investigational phases and even failed trials. Side effect knowledge of drugs at dissimilar clinical stages is crucial to our understanding of the molecular mechanisms underlying the observed toxicities. First developed a support vector machine (SVM) classifier to classify downloaded articles into drug SE-related and SE-unrelated. Then extracted drug–SE co-occurrence pairs from articles that were classified as SE-related. Then developed ranking algorithms to further prioritize extracted drug–SE pairs based on their term frequencies and document frequencies.

Methods: The overall experiment consists of the following steps: (1)download JCO full-text articles;(2) Calculate Term Frequency and Document Frequency; (3) Classify JCO articles into drug SE-related and unrelated using support vector machine; (4) Extract drug–SE pairs from articles classified as SE-related; (5) Rank drug-SE pair;

Preprocessing: JCO is the foremost peer-reviewed journal focusing on clinical cancer research and the authoritative source for current information on the diagnosis and treatment of patients with cancer. Readers are physicians, researchers, oncology nurses and other health care practitioners with a predominant interest in oncology.

The reader is advised to review the appropriate medical literature and the product information currently provided by the manufacturer of each drug to be administered to verify the dosage, the method and duration of administration, or contraindications.

It is the responsibility of the treating physician or other health care professional, relying on independent experience and knowledge of the patient, to determine drug dosages and the best treatment for the patient. We downloaded 500 articles from full-text JCO articles. To remove stemming, stopwords, numbers etc.,

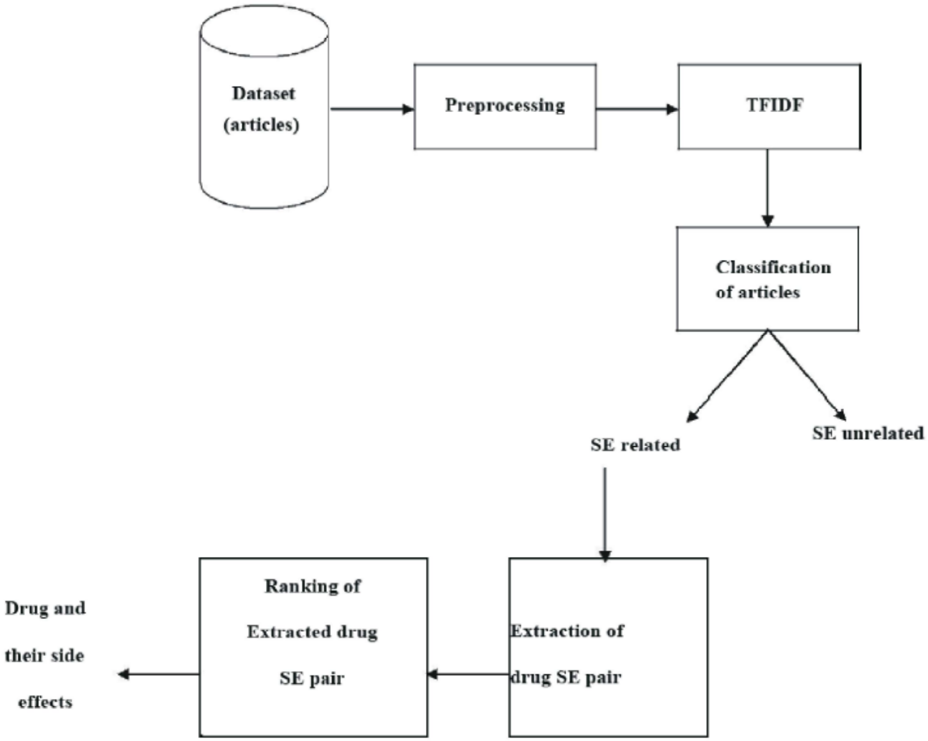

Fig. 1: Overall system design

Proposed System: This section provides the detailed explanation of the proposed system. The system design is shown in Fig. 1.

Term Frequency and Document Frequency: The Term Frequency and Inverse Document Frequency is an arithmetic statistic that is intended to reflect how significant a word is to a document in a collected works or corpus. It is often used as a weighting factor in information retrieval. The tf-idf value enlarge proportionally to the number of times a word appears in the document, but it is offset by the occurrence of the word in the corpus, which helps to alter for the fact that a few words appear more frequently in general.

One of the simplest ranking functionsis compute by calculating the tfidf for all query term; many more sophisticated ranking functions are variants of this simple model.This weight criterion is a arithmetical measure used to evaluate the importance of a term from a text in a corpus. The importance growths proportionally to the number of times aterm appears in the document but is offset by the occurrence of the word in the respective collection [20].For a term ti from the document dj, its term frequency (TF) is defined as follows:

$$f_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

The inverse document frequency is a measure of how much information the word provides, that is, whether the term is mutual or rare across all documents. It is the logarithmically scaled fraction of the documents that contain the word, obtained by dividing the entire number of documents by the number of documents containing the term and then taking the logarithm of that quotient.

$$idf_i \ \square \ \log_2 \frac{|D|}{|d_j : t_i \square d_j|}$$

And, finally, the TF-IDF weight of a term ti is the product TF and IDF

$$T\bar{F} \ IDF_{i,j} \ \square TF_{i,j} \ \square \ IDF_i$$

Classification: The SVM-based classifier used polynomial kernel, bag-of-words feature, TF-IDF weighting, stemming and stopwords removal. The bag-of-words feature was used since it is often the case that the appearance of one specific word such as toxicity‘, ‘adverse‘, ‘toxicity‘, ‘adversely‘, ‘toxicities‘ can be used to determine whether a sentence is drug–SE-related.

Support Vector Machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one target values and several attributes. The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.Classification in SVM is an example of Supervised Learning Known labels help indicate whether the system is performing in a right way or not. The labels are side effect related and side effect unrelated. This information points to a desired response, validating the accuracy of the system, or be used to help the system learn to act correctly. A step in SVM classification involves identification as which are intimately connected to the known classes.SVM are based on statistical learning theory. They can be used for learning to predict future data.

Extract Drug–SE Pair: The inputs to the drug–SE pair extraction algorithm were a list of targeted anticancer drugs, a list of SE terms and JCO articles that were automatically classified as SE-related.

Targeted Drug Lexicon: A list of targeted cancer drugs was obtained from the National Cancer Institute (NCI). The example targeted drugs are: Abitrateroneacetate, Abitrexate, Alemtuzumab, Anastrozole, Belinostat, Bevacizumab Cabazitaxel, Capecitabine, Cetuximab, Dabrafenib, Dactinimycin, Denosumab, Elotuzumab, Enazalutamide, Filgrastim, Fludarabine, Phosphate, Gefitinib, Ibrutinib, Idamycin, Interleukin2, Ianreotide, Acetate, Lenalidomide, Letrozole Lomustine, Methazolastone, Necitumumab, Nilotinib, Nivolumab, Obinutuzumab, Ondansetronhydrochloride, Ocaliplatin, Paclitaxel, Palbociclib, Pamidronatedisodium, Panitumumab, Ramucirumab, Trastuzumab, Vemurafenib, Vinorelbine tartrate.

Targeted Side Effect Lexicon: A list of targeted cancer side effects was obtained from the Medical Dictionary for Regulatory Activities (MedDRA) and Unified Medical Language System(UMLS). The example cancer side effects are: Anemia, Alopecia, Fatigue, Rapid, Chestpain, Shortness of breath, Irregular heart beat, Dizziness, Headache, Insomnia, Blockedintestine, Gastrointestinal obstruction, Dental, Oralhealth, Ascites, Peripheral neuropathy, Hairloss, Diarrhea, Lymphedemia, Vomiting, Neutropenia, Edema, Hypercalcemia, Diagnosis, Nailchanges, Weightgain, Jaundice, Nerve problems. If a drug term and side effect term appeared in the title or text of an SE-related article, then extract the drug-SE pair.

Rank: Two ranking algorithms to rank the extracted drug–SE pairs. The first one is to rank drug–SE pairs according to their total occurrences in the entire corpus, which is equivalent to the term frequency used in information retrieval. The second one is to rank drug–SE pairs according to their document frequencies (the number of documents where a pair appeared).

Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation.

Similarity.co.occurrence(drug,side)freq(drug) freq(side)*

Anticancer Drugs	Side effects
Antiemetic	Diarrhea
Bicalutamide	Edema
Carmustine	Vomiting
Docetaxel	Fatigue
Epirubicin	Fever
Fluorouracil	Anemia
Trastuzumab	Alopecia
Prednisone	Rapid

Fig. 2: Drugs along with its side effects

RESULT

The drug-SE pairs are then ranked using term frequency and document frequency. Finally the drugs along with its side effects are obtained.

CONCLUSION

The method proposed in this work is to extract targeted anticancer drug associated side effects from a large number of high profile full text oncological articles. The Five hundred articles are downloaded from Journal of Clinical Oncology (JCO). We maintain a bag-of-words which contains terms like toxicity', 'adverse' which is used to determine whether a sentence is drug-SE-related or drug-SE-nonrelated. To classify the articles into drug-side effect(SE) related and drug non SE related.

Support Vector Machine(SVM) classifier is used to classify downloaded articles into drug-side effect(SE) related and non SE related. Then drug-SE co-occurrence pairs are extracted from articles that were classified as SE related. For the drug term and side effect term appeared in the text of an SE related article, then extract the drug-SE pair. To rank the extracted drug-SE pairs two techniques are used. The first one is to rank drug-SE pairs according to their total occurrences in the entire corpus, which is equivalent to the term frequency used in information retrieval. The second one is to rank drug-SE pairs according to their document frequencies. Finally to detect the side effect along with their anticancer drugs.

ACKNOWLEDGEMENT

We thankful to Management and Principal of the Mepco Schlenk Engineering College for their financial and logistical support and for providing necessary guidance concerning projects implementation.

REFERENCES

1. Xu, R. and Q. Wang, 2014. Toward creation of a cancer drug toxicity knowledge base: automatically extracting cancer drug side effect relationships from literature, J. Am. Med. Inf. Assoc., 21(1): 90-96.
2. Rong Xu and Quan Qiu Wang 2015. Large-scale automatic extraction of side effects associated with targeted anticancer drugs from full-text oncological articles, Journal of Biomedical Informatics, 55: 64-72. ISSN 1532-0464
3. Cami, A., A. Arnold, S. Manzi and B. Reis, 2011. Predicting adverse drug events using pharmacological network models. Sci. Transl. Med. 3(114): 114ra27.
4. Campillos, M., M. Kuhn, A.C. Gavin, L.J. Jensen and P. Bork, 2008. Drug target identification using side-effect similarity. Science, 321(5886): 263-6.
5. Cleeland, C.S., J.D. Allen, S.A. Roberts, J.M. Brell, S.A. Giralt, A.Y. Khakoo, R.A. Kirch, V.E. Kwitkowski, Z. Liao and J. Skillings, 2012. Reducing the toxicity of cancer therapy: recognizing needs, taking action, Nat. Rev. Clin. Oncol., 9: 471-478.
6. Kirk, R., 2012. Targeted therapies: the toxic reality of new drugs, Nat. Rev. Clin. Oncol., 9(9): 488.
7. Niraula, S., B. Seruga, A. Ocana, T. Shao, R. Goldstein, I.F. Tannock and E. Amir, 2012. The price we pay for progress: a meta-analysis of harms of newly approved anticancer drugs, J. Clin. Oncol., 30(24): 3012-3019.
8. Eschenhagen, T., T. Force, M.S. Ewer, G.W. de Keulenaer, T.M. Suter, S.D. Anker, M. Avkiran, E. de Azambuja, J. Balligand, D.L. Brutsa, G. Condorelli, A. Hansen, S. Heymans, J.A. Hill, E. Hirsch, D. Hilfiker-Kleiner, S. Janssens, S. de Jong, G. Neubauer, B. Pieske, P. Ponikowski, M. Pirmohamed, M. Rauchhaus, D. Sawyer, P.H. Sugden, J. Wojta, F. Zannad and A.M. Shah, 2011. Cardiovascular side effects of cancer therapies: a position statement from the Heart Failure Association of the European Society of Cardiology, Eur. J. Heart Fail., 13(1): 1-10.
9. Bonura, F., D. Di Lisi, S. Novo and N. D'Alessandro, 2012. Timely recognition of cardiovascular toxicity by anticancer agents: a common objective of the pharmacologis oncologist and cardiologist, Cardiovasc. Toxicol., 12(2): 93-107.
10. Ravaud, A., 2009. How to optimise treatment compliance in metastatic renal cell carcinoma with targeted agents, Ann. Oncol., 20(suppl. 1): i7-i12.
11. Keefe, D.M. and E.H. Bateman, 0000. Tumor control versus adverse, events with targeted anticancer therapies, Nat. Rev. Clin. Oncol., 9(2): 98-109.
12. Weitman, S.D.; Glatstein, E. Kamen, B.A. Back to the basics: the importance of concentration x time in oncology. J. Clin. Oncol., 1993, 11(5), 820-821.
13. Harpaz, R., W. DuMouchel, N.H. Shah, D. Madigan, P. Ryan and C. Friedman, 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Therapeut, 91(6): 1010-21.
14. Tatonetti, N.P., P.Y. Patrick, R. Daneshjou and R.B. Altman, 2012. Data-driven prediction of drug effects and interactions. Sci Transl Med., 4(125): 125ra31
15. Gurulingappa H., A. Mateen Rajput and L. Toldo, 2012. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics, 3(1): 15.

16. Shetty, K.D. and S.R. Dalal, 2011. Using information mining of the medical literature to improve drug safety. J Am Med Inform Assoc., 18(5): 668-74.
17. Xu, R. and Q. Wang, 2013. Automatic construction and integrated analysis of a cancer drug side effect knowledge base. J Am Med Inform Assoc., http://dx.doi.org/10.1136/amiajnl-2012-001584.
18. Gurulingappa, H., A. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius and L. Toldo, 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, Biomed. Inform., 45(5): 885-892.
18. Van Mulligen, E., A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. Kors and L. Furlong, 2012. The EU-ADR corpus: annotated drugs, diseases, targets and their relationships, Biomed. Inform., 45(5): 879-884.
19. Deleger, L., Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo and I. Solti, 2012. Building gold standard corpora for medical natural language processing tasks, in: AMIA Annual Symposium, Washington, DC, pp: 144-153.
20. Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing,? Commun. ACM, 18(11): 613-620.