Mining Biomedical Literature to Explore Interactions between Cancer Drugs and Dietary Supplements

Rui Zhang, PhD[1, 2], Terrance J. Adam, RPh, MD, PhD[1, 3], Gyorgy Simon, PhD[1], Michael J. Cairelli, DO, MS[4], Thomas Rindflesch, PhD[4], Serguei Pakhomov, PhD[1, 3], Genevieve B. Melton, MD, MA[1, 2]
[1]Institute for Health Informatics; [2]Department of Surgery; [3]College of Pharmacy, University of Minnesota, Minneapolis, MN; [4]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health

Abstract

Interactions between cancer drugs and dietary supplements are clinically important and have not been extensively investigated through mining of the biomedical literature. We report on a previously introduced method now enhanced by machine learning-based filtering. Potential interactions are extracted by using relationships in the form of semantic predications. Semantic predications stored in SemMedDB, a database of structured knowledge generated from MEDLINE, were filtered and connected by two interaction pathways to explore potential drug-supplement interactions (DSIs). The lasso regression filter was trained by using SemRep output features in an expert annotated corpus and used to rank retrieved predications by predicted precision. We found not only known interactions but also inferred several unknown potential DSIs by appropriate filtering and linking of semantic predications.

Introduction

The use of natural supplements in the US has increased dramatically in recent years. According to the results of a National Health Interview Survey in 2012, 17.9% of American adults had used dietary supplements (not including vitamins and minerals).[1] National Health and Nutrition Examination Survey (NHANEs) data also indicated that 53% of American adults took at least one dietary supplement during 2003-2006, mostly multivitamin and multimineral supplements.[2] When taking occasional and seasonal use into account, the prevalence of supplement use was 69% in 2011. Additionally, supplement use in women is higher than in men according to consumer surveys by the Council for Responsible Nutrition.[3] Those who use the supplements often take them in combination with conventional drugs. About 30% of the elderly population (age >65), the largest group consuming prescription drugs, use at least one daily supplement, thus placing the patient at risk for potential drug-supplement interactions (DSIs). Supplements are also increasingly used in the US by patients diagnosed with cancer to help strengthen their immune system and ease the side effects of treatments.

The growing popularity of supplements has focused attention on DSIs.[4] One study suggested that patients on medications with a narrow therapeutic index (e.g., cyclosporine, phenytoin, warfarin) should avoid the use of herbal products, as those drugs may either have adverse effects or be less effective when combined with such products.[5] Gurley et al. reported that the concomitant administration of botanical supplements with P-glycoprotein (P-gp) substrates can lead to clinically significant interactions. The study was based on evaluating the effects St. John's wort (SJW) and Echinacea on the pharmacokinetics of a P-gp substrate, digoxin.[6] It was suggested that the concomitant use of docetaxel and SJW should be avoided in cancer patients, as the hyperforin component in SJW can induce cytochrome P450 3A4, thus leading to changes in drug metabolism for a number of chemotherapeutic and other conventional drugs.

Many of these studies have focused on limited sets of supplements and drugs. Most supplements have not been studied extensively in clinical trials. Some serious DSIs are not found until a new drug is already on the market, since clinical trials for new drugs do not typically consider DSIs. Therefore, many DSIs are unknown to both health care providers and patients themselves. Current DSI documentation is limited, as it is only based on pharmacological, in vitro, or animal model data. Moreover, because of the less rigorous regulatory rules regarding dietary supplements, formulations may vary significantly by manufacturer, and similar products may be derived from a variety of sources.

In addition, potential DSIs may result from undefined pathways that have yet to be discovered. Such interactions often can only be derived with an indirect approach, such as mining the scientific literature. This resource contains a large amount of pharmacokinetic and pharmacodynamic knowledge in free text and expands the range of drugs, supplements and genes. Compared to traditional drug-drug interaction work, the use of literature-based discovery for DSI identification has not been adequately investigated. We hypothesize that a powerful literature-based information discovery system could significantly enhance DSI knowledge bases and further translate to clinical practice for

increased quality of patient care. In this study, we investigated the use of the structured knowledge extracted from biomedical literature for exploration of DSIs.

Background

Literature-based discovery (LBD) is an automatic method to generate hypotheses by connecting findings in the literature. In general, if a concept Y is related to both concept X and concept Z, there exists a potential association between X and Z. For example, Swanson et al. applied this approach to propose fish oil as a potential treatment for Raynaud's disease.[7] Hristovski et al. proposed to enhance LBD by using semantic predications instead of depending on co-occurrence of words or concepts.[8] They found that linking semantic predications could generate a larger number of useful findings. In this study, we will treat a gene as the concept Y linking a cancer drug (X) and a dietary supplement (Z).

SemRep is a semantic interpreter that extracts structured knowledge in the form of semantic predications from MEDLINE citations[9]. Each predication consists of two UMLS Metathesaurus concepts as subject and object, and a semantic relationship (from the UMLS Semantic Network) as a predicate. SemMedDB[10] is a database containing semantic predications generated by SemRep from over 23.6 million MEDLINE citations. The database for this study contains 69 million semantic predications from processed citations published as of March 31, 2014. SemRep and SemMedDB have been used in pharmacogenomic information extraction[11, 12] and information extraction for clinicians' needs[13].

Methods

The overall methodology includes 1) drug and supplement concept mapping, 2) related semantic predication extraction, 3) machine-learning based filtering, 4) extraction of potential interaction pathways, and 5) expert judgment and verification of known interactions.

Supplements and cancer drugs

In this study, we focused on a limited subset of dietary supplements and cancer drugs. A supplement list was obtained from the Office of Dietary Supplements website (http://ods.od.nih.gov/factsheets/list-all/ - accessed August 1, 2014). Drugs approved by the FDA to treat breast and prostate cancer were also collected from the National Cancer Institute website (http://www.cancer.gov/cancertopics/druginfo/drug-page-index - accessed August 1, 2014). The corresponding CUIs for supplements and cancer drugs were retrieved by mapping the supplements and drugs to the UMLS Metathesaurus via MetaMap.

Extraction of semantic predications

We extracted from SemMedDB predications with various semantic types: supplement-gene (i.e., predications with a dietary supplement as the subject and a gene as the object), breast_cancer_drug-gene, prostate_cancer_drug-gene, gene-supplement, gene-breast_cancer_drug, and gene-prostate_cancer_drug. We restricted our analysis to three predicate types: STIMULATES, INHIBITS, and INTERACTS_WITH.

Machine learning-based filter

Considering the limited precision of SemRep that we have seen previously[11], we developed a machine learning (ML)-based filter to rank the generated semantic predications by probability of being correct. We used the reference standard from a previous study containing annotations for 300 randomly selected sentences that involve drug-gene, gene-drug and gene-biological function predications (dataset details described in the paper)[11] as our training set. SemRep extracted 524 total semantic predications from these sentences, 304 of which were evaluated to be correct. In this study, we only used predications generated by SemRep that were judged by experts as either true positive (TP) or false positive (FP). The supervised machine-learning algorithm, lasso regression (LR), was used to classify SemRep output as either TP or FP. We used SemRep output features as predictors, including UMLS biomedical concepts, argument distance, indicator types (e.g., verb), predicate (e.g., TREATS), and UMLS semantic types.

To evaluate the effectiveness of the filter, we ranked the generated semantic predications based on the score the model assigned to each semantic predication. A physician (MJC) was asked to judge the top 20 for each type of predication (including supplement-gene, gene-supplement, breast cancer drugs-gene, gene-breast cancer drugs, prostate cancer drugs-gene, gene-prostate cancer drugs). The precision was calculated as *True Positives/(True Positives+False Positives)*.

DSI discovery

We further filtered out some nonspecific concepts such as "gene", "receptor" and "proteins" before discovery. We ranked the potential DSIs based on two pathways modified from earlier work[11]: Drug→Gene→Supplement and Supplement→Gene→Drug schemas. For the first pathway, when semantic predications Drug→Gene and

Gene→Supplement share the same gene, an interaction may be indicated. For example, the predications Echinacea INHIBITS CYP450, and CYP450 INTERACTS_WITH Docetaxel generate the potential interaction Echinacea→CYP450→Docetaxel. A score was then assigned to each potential interaction (e.g., Supplement→Gene→Drug) by adding two rank scores of Supplement→Gene and Gene→Drug, which were obtained from the ML-based filter. These interactions were ranked based on this score and then evaluated.

Expert selection and database checking

A drug interaction expert that is a pharmacist and physician (TJA) manually reviewed the top 100 ranked interactions. Priority for review of potential interactions was first given to gene specificity, with priority for specific gene paths (e.g. CYP 450 3A4) preferred over more general gene categories (CYP p450). Second priority was given to highly plausible combinations such as those including supplements to enhance the immune system or help with chemotherapy side effects, where the supplements would be likely to be used in combination in a clinical setting. The next priority was the level of evidence in describing potential interactions and was assessed based on the pharmacologic data, with priority given to interaction data providing evidence of substance-gene activity.

To compare potential DSIs found by our method with established interactions, a drug profile for drugs for breast and prostate cancer, as well as a number of targeted supplements was entered into a well-known drug-drug interaction (DDI) website (http://cpref.goldstandard.com/interreport.asp) with a theoretical multiple drug profile including pertinent supplements and chemotherapy drugs. Another website, at Case Western Reserve University Hospital, was used to provide additional support: http://www.uhhospitals.org/health-and-wellness/drug-information-center/drug-interaction-tool.

Results

We extracted 10,500 supplement→gene predications, 270 prostate_cancer_drug→gene predications, 991 breast_cancer_drug→gene predications, 7732 gene→supplement predications, 280 gene→breast_cancer_drugs predications, and 217 gene→prostate_cancer_drugs predications. The precision of the top ranked predications after filtering was 69%, a significant increase over the 58% previously reported for a randomly selected set of similar predications[10]. After combining the top ranking 500 predications from each side of the pathway (supplement-gene and gene-drug), 1095 combinations were formed and ranked by score. After expert review, we examined some of these DSIs focusing on the pharmacologically active CYP450 gene family, which has potential effects on multiple therapeutic classes. We found both known and unknown DSIs, and we found five more interactions with filtered predications than with unfiltered predications. Some of the potential DSIs identified shared the same pathway, such as the interactions between Echinacea and chemotherapeutic medications cyclophosphamide, docetaxcel, everolimus, fluorouracil and toremifine (Table 1). Table 2 lists selected semantic predications and citations.

Table 1. Selected DSI examples and pathways. INH, INHIBITS; STI, STIMULATES; INT, INTERACTS_WITH.

Drug/Supplement	Predicate	Gene/Gene Class	Predicate	Supplement/Drug	Known	Filter/ Unfilter
Echinacea	INH	CYP450	INT	Cyclophosphamide	Y	Both
Echinacea	INH	CYP450	INT	Docetaxel	Y	Both
Echinacea	INH	CYP450	INT	Everolimus	Y	Both
Echinacea	INH	CYP450	INT	Fluorouracil	Y	Both
Echinacea	INH	CYP450	INT	Toremifine	N	Both
Echinacea	STI	CYP1A1	INT	Exemestane	N	Both
Grape seed extract	INH	CYP3A4	INT	Docetaxel	N	Both
Kava preparation	STI	CYP3A4	INT	Docetaxel	Y	Filter
Ginseng	INH	CYP3A	INT	Ginkgo bilob extract	Y	Unfilter
Ginseng	INH	CYP3A	INT	Docetaxel	N	Unfilter
Prednisone	INT	P-glycoprotein	STI	Vitamin E	N	Filter
Cyclophosphamide	INT	P-glycoprotein	STI	Vitamin E	N	Filter
Glucosamine	INH	COX2	STI	Docetaxel	Y	Filter
Melatonin	INH	COX2	STI	Docetaxel	N	Filter

Table 2. Selected semantic predications and citations.

Semantic Predications	Citations (PMID)
Echinacea STIMULATES CYP1A1	Our in vivo data indicate that the Echinacea ethanolic extract can potently inhibit the expression of CYP3A1/2 and can also induce of CYP1A1, CYP2D1. (20374973)
Grape seed extract	Four brands of GSE had no effect, while another five produced mild to moderate but variable

INHIBITS CYP3A4	inhibition of CYP3A4, ranging from 6.4% by Country Life GSE to 26.8% by Loma Linda Market brand. (19353999)
Melatonin INHIBITS Cyclooxygenase-2	Moreover, Western blot analysis showed that melatonin inhibited LPS/IFN-gamma-induced expression of COX-2 protein, but not that of constitutive cyclooxygenase. (18078452).
Prednisone INTERACTS_WITH P-glycoprotein	PRED is also a substrate of P-gp and is a weak inducer of CYP3A, and drug-drug interactions within this combination therapy might occur. (23267661)
Cyclophosphamide INTERACTS_WITH P-glycoprotein	These findings suggest that active cyclophosphamide metabolite can be a substrate for P-glycoprotein. (22803083)
CYP450 INTERACTS_WITH Toremifene	Tamoxifen and toremifene are metabolised by the cytochrome p450 enzyme system, and raloxifene is metabolised by glucuronide conjugation. (12648026)
CYP3A INHIBITS Docetaxel	Because docetaxel is inactivated by CYP3A, we studied the effects of the St. John's wort constituent hyperforin on docetaxel metabolism in a human hepatocyte model. (16203790)
CYP1A1 INTERACTS_WITH Exemestane	Recombinant CYP1A1 metabolized exemestane to MI with a catalytic efficiency (Cl(int)) of 150 nl/pmol P450 x min that was at least 3.5-fold higher than those of other P450s investigated. (20876785)
Cyclooxygenase 2 STIMULATES Docetaxel	We investigated whether prostate tumor-associated stromal cells, marrow-derived osteoblasts, affect cytotoxicity of 2 antitumor drugs, COL-3 and docetaxel (TXTR), and whether it is dependent on COX-2 activity. (15688368)
P-glycoprotein STIMULATES Vitamin E	Expression of multiple drug resistant (MDR) phenotype and over-expression of P-glycoprotein (P-gp) in the human hepatocellular carcinoma (HCC) cell clone P1(0.5), derived from the PLC/PRF/5 cell line (P5), are associated with strong resistance to oxidative stress and a significant ($p < 0.01$) increase in intracellular vitamin E content as compared with the parental cell line. (15453640)

Discussion

DSI is an important topic and deserving of additional investigation with informatics methods to explore potential interactions extracted from the biomedical literature that may have a significant effect on medication therapy. In this study, we investigated the use of semantic knowledge provided by SemRep for DSI extraction. However, due to its limited recall and precision, human review is typically required for maximally effective use of this resource. This intensive manual process of filtering has limited the general use of SemRep in larger scale applications in biomedical and health informatics. Although argument-predicate distance has been used to enhance the precision of extracting semantic predications[14], the use of machine-learning for automatic filtering of semantic predications has not been investigated. The ranking score for each potential interaction was used as an additional means lower the number of interactions subjected to human review. It was found that the filter helped to discover not only those found by using unfiltered results but also found several additional DSIs. Two DSIs that were found without filtering were not found in the filtered interactions.

Both known and unknown interactions were found. The first DSI candidate area of interest is Echninacea which has been known to affect chemotherapy drugs. Cyclophosphamide, docetaxel, fluorouracil are all standard therapeutics for breast and prostate cancer patients and all were noted in our data, as well as in the DDI medication site used for confirmation. We also identified a potentially novel DDI with Echinacea. Exemestane was noted to have an interaction with Echinacea in the test data, with specific activity identified at the CYP1A1 gene. This was confirmed after expert review. This interaction did not show up on the DDI sites consulted. This may be worth additional exploration, especially since metabolites of Exemestane result in reduced activity or non-activity of this medication. The potential implication of therapeutic failure may have an impact on patient survival. In Another example involves P-glycoprotein, which can affect many drugs. Prednisone and cyclophosphamide were identified to be substrates for P-glycoprotein with both having the potential for interactions with the other. In addition they both may interact with the supplement Vitamin E. Melatonin inhibits COX2 via suppression of protein expression potentially creating an interaction with docetaxel which also has activity via the COX2 pathway. Ginseng was also explored for possible interaction in our theoretical patient profile. It was noted to interact with ginkgo on the DDI checker website. This is confirmed in our test data set via CYP450 pathway. In the case of prostate cancer, the potential effect of ginseng on multiple CYP450 pathways was noted, which may result in a DSI with docetaxel through the CYP450 3A pathway. This interaction is not noted at either of the DDI websites we consulted and is an area for future exploration.

In this pilot study, we did not define the specific interaction relations between cancer drugs and dietary supplements,

although we found their potential pathways. In previous SemRep error analysis11, false positives of semantic predications were mainly due to knowledge source shortcomings (e.g., incorrect mapping of gene or protein mentions to UMLS concepts or Entrez Gene terms) and SemRep processing shortcomings with linguistic phenomena (e.g., negation, serial coordination). Future SemRep development efforts include addressing these shortcomings.

Although known interactions were discovered using this methodology, the goal was to identify unknown interactions. Verifying such interactions is a significant challenge for this type of methodological development since, by definition, there is no reference standard. The best validation method would be to translate the findings into a clinical trial for the suggested combination. While our approach of identifying literature support for the proposed interaction is significantly more expedient, it is less robust and further clinical evidence is required to fully validate these findings.

In conclusion, we found both known and unknown DSIs by using combining a supplement→gene→drug schema data with LR to filter and rank the semantic predications before expert manually review. We found that, although further development could improve the performance of the system, this filter provides a foundation to facilitate DSI discovery by avoid experts screen a larger reducing and enriching the set of semantic predications for expert review.

Acknowledgments

This research was supported by the University of Minnesota Informatics Institute On the Horizon Grant (RZ), Agency for Healthcare Research & Quality Grant (#R01HS022085-01) (GM), and University of Minnesota clinical and Translational Science Award (#8UL1TR000114-02) (Blazer). This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. This research was supported in part by an appointment to the NLM Research Participation Program, which is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

Reference

1. Peregoy JA, Clarke TC, Jones LI, Stussman BJ, Nahin RL. Regional variation in use of complementary health approaches by U.S. adults. NCHS data brief, no 146 Hyattsville, MD: National Center for Health Statistics. 2014.
2. National Center for Health Statistics About the National Health and Nutrition Examination Survey. Hyattsville (MD). 2009.
3. Dickinson A, Blatman J, El-Dash N, Franco JC. Consumer usage and reasons for using dietary supplements: report of a series of surveys. Journal of Americian College of Nutrition. 2014;33(2):176-82.
4. Kennedy DA, Seely D. Clinically based evidence of drug-herb interactions: a systematic review. Expert opinion on drug safety. 2010 Jan;9(1):79-124.
5. Kuhn MA. Herbal remedies: drug-herb interactions. Critical care nurse. 2002 Apr;22(2):22-8, 30, 2; quiz 4-5.
6. Gurley BJ, Swain A, Williams DK, Barone G, Battu SK. Gauging the clinical significance of P-glycoprotein-mediated herb-drug interactions: comparative effects of St. John's wort, Echinacea, clarithromycin, and rifampin on digoxin pharmacokinetics. Molecular nutrition & food research. 2008 Jul;52(7):772-9.
7. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine. 1986 Autumn;30(1):7-18.
8. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc. 2006:349-53.
9. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.
10. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012 Dec 1;28(23):3158-60.
11. Zhang R, Cairelli MJ, Fiszman M, Rosemblat G, Kilicoglu H, Rindflesch TC, et al. Using semantic predications to uncover drug-drug interactions in clinical data. Journal of Biomedical Informatics. 2014;49:134-47.
12. Zhang R, Cairelli MJ, Fiszman M, Kilicoglu H, Rindflesch TC, Pakhomov SV, et al. Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs. Cancer informatics. 2014.
13. Jonnalagadda SR, Del Fiol G, Medlin R, Weir C, Fiszman M, Mostafa J, et al. Automatically extracting sentences from Medline citations to support clinicians' information needs. Journal of the American Medical Informatics Association. 2013 Sep-Oct;20(5):995-1000.
14. Masseroli M, Kilicoglu H, Lang FM, Rindflesch TC. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. BMC bioinformatics. 2006 Jun 8;7:291.