POMELO: Medline corpus with manually annotated food-drug interactions

Thierry Hamon[1,2]**, Vincent Tabanou**[3]**, Fleur Mougin**[4]**, Natalia Grabar**[5]**, Frantz Thiessard**[3,4]

[1] LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[2] Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France
[3] CHU de Bordeaux, Pole de sante publique, Service d'information médicale, Bordeaux, France
[4] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center,
team ERIAS, UMR 1219, F-33000 Bordeaux, France
[5] CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
hamon@limsi.fr, vincenttabanou@aol.com, fleur.mougin@u-bordeaux.fr,
natalia.grabar@univ-lille3.fr, frantz.thiessard@u-bordeaux.fr

Abstract

When patients take more than one medication, they may be at risk of drug interactions, which means that a given drug can cause unexpected effects when taken in combination with other drugs. Similar effects may occur when drugs are taken together with some food or beverages. For instance, grapefruit has interactions with several drugs, because its active ingredients inhibit enzymes involved in the drugs metabolism and can then cause an excessive dosage of these drugs. Yet, information on food/drug interactions is poorly researched. The current research is mainly provided by the medical domain and a very tentative work is provided by computer sciences and NLP domains. One factor that motivates the research is related to the availability of the annotated corpora and the reference data. The purpose of our work is to describe the rationale and approach for creation and annotation of scientific corpus with information on food/drug interactions. This corpus contains 639 MEDLINE citations (titles and abstracts), corresponding to 5,752 sentences. It is manually annotated by two experts. The corpus is named *POMELO*. This annotated corpus will be made available for the research purposes.

1 Introduction

Prescribed medicines depend on initial marketing authorization to guarantee the security of patients. Nevertheless, medicines can cause adverse drug reactions (ADRs) discovered during clinical trials, but usually later, in a pharmacovigilance context, while drugs are administered to patients (Aagaard and Hansen, 2013; Brahma et al., 2013; Cote and Choy, 2013; Yom-Tov and Gabrilovich, 2013; Wei et al., 2013). For this reason, prescription and intake of drugs is controlled all over their marketing and use by patients.

When patients take more than one medication, they may be at risk of drug interactions, which means that a given drug can cause unexpected effects when taken in combination with other drugs. For instance, sedative and pain medication cause an important drowsiness in patients, while other drugs (*benzylpenicillin* and *heparin*) interact between them and cannot be placed in the same syringe. Most drugs are not concerned by such effects. Yet, when such adverse events happen, they may have negative and serious effects on patients and their health. Hence, it is important to know the possible interactions between drugs and to clearly indicate them to patients and to medical staff. For this reason, for several years now, automatic extraction of drug-drug interactions is heavily researched in order to provide an updated and timely information on known interactions between drugs or between their active principles. In the same way, it becomes possible to discover potential adverse effects and adverse reactions (Aronson and Ferner, 2005) of these drugs.

A different but yet related situation occurs when drugs are taken together with certain food or beverages. Food and drug interaction may also lead to negative effects on health and well-being of patients. For instance, grapefruit has interactions with several drugs, because its active ingredients inhibit enzymes involved in the drugs metabolism and can then cause an excessive dosage of these drugs (Duke Med Health News, 2013; Greenblatt and Derendorf, 2013). Due to the difficulty to detect them, because patients usually do not remem-

ber the food they have taken, and to their complexity, such situations are studied less frequently, even if they are important for patients and for the medical care process. Our main interest is to study food-drug interactions and to automatically detect them in scientific literature.

Most information on food/drug interactions is recorded in unstructured sources, such as scientific articles and some knowledge bases, like Drug-Bank[1] (Wishart et al., 2006), or possibly in discussion fora which provide patient point of view of adverse events. Yet, this information remains poor. For instance, DrugBank records textual information about food/drug interactions for less than 10% of drugs, and it mainly provides information on optimal drug intake time.

Regarding these observations, our objective is to use and mine scientific bibliographical data in order to describe interactions that exist between drugs and food, and that may lead to adverse effects. To achieve this goal, our first step is to design and to annotate a dataset of MEDLINE abstracts. This is the purpose of the work presented in this paper.

In what follows, we first present some related work (section 2). We then present the annotation scheme (section 3), and describe the corpus definition and the annotation process (sections 4 and 5). Finally, we present our results (section 6), and discuss our work and conclude with future orientations (section 7).

2 Related work

We present two kinds of works: those performed by pharmacists and pharmacovigilance experts on drug/drug interactions (DDI) and food/drug interactions (FDI) in medical domain, and those performed by computer scientists on information extraction.

2.1 Medical domain

In has been defined that interactions can occur in different ways. The interactions presented here have been defined for the DDI cases, but they show very similar effects when the FDIs occur. Hence, two drugs given together may act at the same or similar receptor, which can lead to a greater or to a decreased effect of either drug. Another situation is when one drug is affected by action of

another drug. In this case, their absorption, distribution, metabolism or excretion (commonly called ADME) are involved (Doogue and Polasek, 2013). We give here some examples of DDIs found online[2]:

- *Absorption.* Some drugs can alter the absorption of another drug. For example, *calcium* can block absorption of some medications. Hence, the HIV treatment *dolutegravir (Tivicay)* should not be taken at the same time as *e.g. calcium carbonate (Tums, Maalox)*, because it can lower the amount of *dolutegravir* absorbed and reduce its effectiveness in treating HIV infection. For the same reason, many drugs cannot be taken with milk or dairy products because they will bind with the calcium;

- *Distribution.* Protein-binding interactions can occur when highly protein-bound drugs compete for a limited number of binding sites. One example is between *fenofibric acid (Trilipix)*, used to lower cholesterol in the blood, and *warfarin*, a common blood thinner to help prevent clots. *Fenofibric acid* can increase the effects of *warfarin* and cause the bleeding in patient;

- *Metabolism.* Drugs are usually eliminated from the body further to their changes through metabolism. Enzymes in the liver, usually the CYP450 enzymes, are often responsible for breaking down drugs and for their elimination from the body. However, enzyme levels may go up or down and affect how drugs are broken down. For example, using *diltiazem*, a blood pressure medication, with *simvastatin*, a medicine to lower cholesterol, may elevate the blood levels and cause side effects due to *simvastatin*. Indeed, *diltiazem* can block the CYP450 3A4 enzymes needed for the breakdown of *simvastatin*, in which case, high blood levels of *simvastatin* can lead to serious liver and muscle side effects. Another example is when grapefruit affects the action of the CYP3A4 enzyme, thus also affecting the intake of several drugs (Duke Med Health News, 2013; Greenblatt and Derendorf, 2013);

---

[1] http://www.drugbank.ca/

[2] https://www.drugs.com/drug_interactions.html

- *Excretion.* Some *nonsteroidal antiinflammatory drugs (NSAIDs)* (e.g. *indomethacin*), may lower kidney function and affect the excretion of *lithium*, a drug used for bipolar disorder, in which case its action can be increased.

From these examples, we can highlight several points related to the intake of drugs:

- when several medications are taken together, patients should define the best way to take them (*e.g.* time, dose) with their doctor;

- DDIs and FDIs may show similar action patterns because they may contains same or similar active principles, like shown above with the *calcium* intake, present in drugs and in diary products, or with grapefruit and *diltiazem* changing the behavior of some enzymes;

- the interaction can follow several patterns: action of a given medication can be decreased, increased or cause side effects which are usually not observed.

2.2 Computer Sciences

As noticed above, it is important to research the issues related to DDIs and FDIs. Concerning the DDIs, and globally the ADRs (Adverse Drug Reactions), their reporting is extremely low. For instance, in France, 96% of the ADRs are simply not reported (Moride et al., 1997; Lacoste-Roussillon et al., 2001). As for the FDIs, they are even more difficult to identify for several reasons: the large number of possible interactions, the difficulty of describing meals in a standard ADR reporting form, the difficulty to remember exactly which food has been taken at a given moment, and probably also for sociological reasons because drugs and pathology are connected unconsciously and associated with negative feelings, whereas food is rather an indication of good health. For these reasons, it is important to provide automatic NLP (Natural Language Processing) methods for mining available sources of information, like MED-LINE bibliographical database[3].

Several works have been done on automatic extraction of DDIs (Duda et al., 2005; Björne et al., 2013; Ayvaz et al., 2015; Kim et al., 2015; Kolchinsky et al., 2015; Liu et al., 2016; Schneider and Boyce, 2016). In most cases, supervised

categorization methods are exploited for the detection of entities and of their interactions. This explains why the majority of these works are part of the NLP challenges, like *SEM (Segura-Bedmar et al., 2013) for DDI extraction and BIOCREATIVE (Krallinger et al., 2009) for PPI (Protein–Protein Interaction) extraction. Indeed, the *SEM challenge proposes task dedicated to DDI extraction and provides annotated corpora. The main contribution of our work is related to the creation of biomedical corpora and their annotation with information on food/drug interactions.

Notice that there is very little work on automatic FDI extraction. Currently, several knowledge bases, semantic resources and repositories concerning the involved entities (i.e., drugs, food and diseases) are available (Brown et al., 1999; NLM; RxNorm; Kuhn et al., 2010). Yet, the FDI information is fragmented and scattered across these bases and repositories. As consequence, there is no explicit relations between these entities. Linked Data projects, such as Linked Open Drug Data (LODD[4]), attempt to create fine-grained links between such knowledge bases. In addition, the already mentioned DrugBank (Wishart et al., 2006) knowledge base contains various kinds of information on drugs, although their relations with food is provided as free-text fields and is mainly concerned with the optimal drug intake time. The effort done for the formalization of the FDI information from DrugBank (Jovanovik et al., 2014) has been performed manually.

In the next sections, we propose the description of the methodology for the creation of annotated corpus with FDIs.

3 Annotation schema: Representation of drug related information and of FDIs

In Figure 1, we present the model of the drug-related information. The model is instantiated for the *solumedrol* medication. Hence, a given drug has an international name (*DCI*) and a therapeutic class (or `is-a` relation). It has a composition and is prescribed for specific indications and with specific features (*e.g.* dosage, mode, frequency and duration of administration). Then, a drug may have adverse effects, including those due to action of food (FDIs) or of other drugs (DDIs).

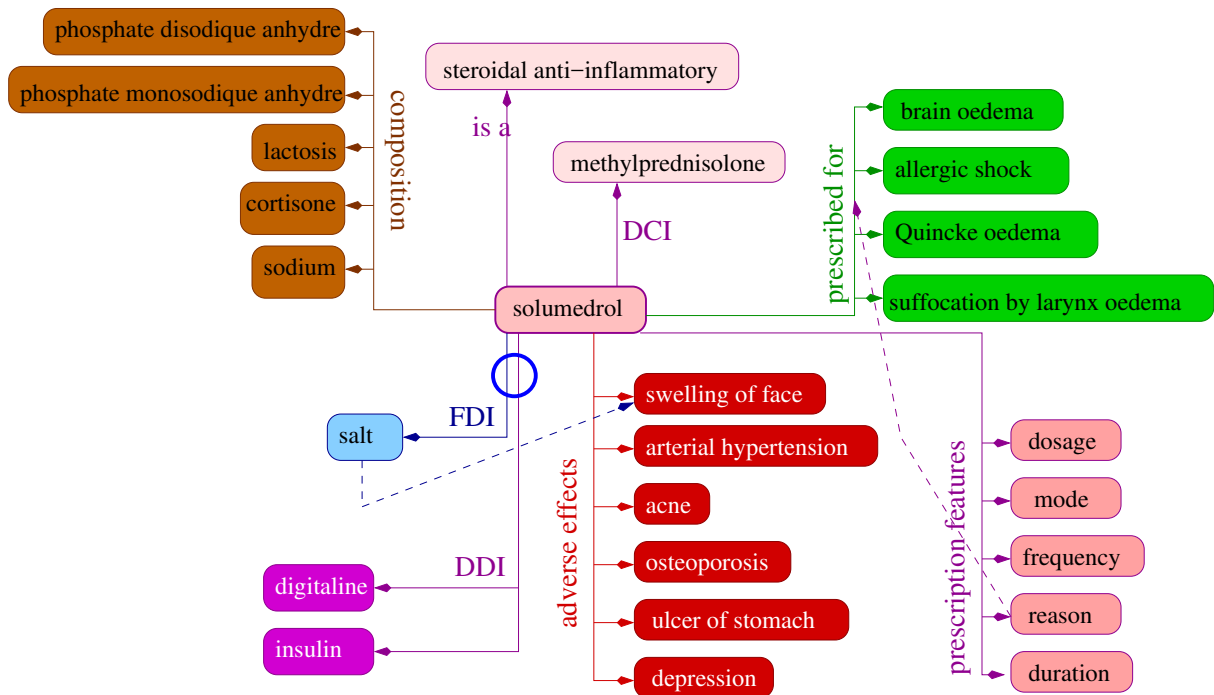The annotated entities must belong to one of these categories:

Figure 1: Model of the drug-related information

1. food names and their different types (ingredients, cooked meals, food supplements...),

2. meal time (before, during, after),

3. drug names and related information (dosage, frequency, duration, mode),

4. disorders for which a given drug is indicated,

5. side effects (including FDIs) of drugs.

According to types of actions of drugs on patients and to the ADME model (section 2.1), we propose to investigate the following types of interactions between these entities:

- decrease, reduce, slow down or make disappear drug effect (absorption, elimination...) due to food,

- increase or speed up drug effect due to food,

- make appear, have (new) side effects, make appear negative effect, or worsen drug effect due to food,

- increase side effect of drugs due to food,

- improve drug effect, have positive effect on drug, or reduce side effect of drug,

- treat a disorder,

- have no effect on drug,

- drug must be taken without food,

- have effect on drug or general relation with drug. These relations are under-defined: relation or effect exist but it is not possible to decide what kind of relation or effect it is.

4 Corpus design

The MEDLINE bibliographical base has been queried in order to extract citations related to food-drug interactions with the following query:

("FOOD DRUG INTERACTIONS"[MH] OR "FOOD DRUG INTERACTIONS*") AND ("adverse effects*")

In December 2013, it permitted to obtain a set of 639 citations, of which we exploit titles and abstracts. This corpus is called *POMELO*, namely *grapefruit* in French, because has been built and annotated during the French MESHS-funded project POMELO.

5 Corpus annotation

The POMELO corpus (639 titles and abstracts corresponding to 5,752 sentences) has been manually annotated in order to make explicit the information on food/drug interactions. Two experts have

been involved in the annotation process: one resident and one medical doctor. The annotation has been done mainly by the resident, who was helped by the medical doctor when facing difficult situations. The annotation has been performed with the BRAT software (Stenetorp et al., 2012).

To prepare the annotation, we exploited some existing resources in English and French, which have been automatically projected on corpus. For instance, the food has been pre-annotated using:

- the USDA National Nutrient Database[5]

- the Codex Alimentarius of the WHO (World Health, Organization)[6],

- and resources built from some recipes.

Other entities have been pre-annotated with existing terminologies: disorders and side effects (Brown et al., 1999; NLM; Kuhn et al., 2010), and drugs (RxNorm; Wishart et al., 2006). Besides, specific resources have been built for the annotation of dosage, frequency, duration and mode of drug administration.

Then, the POMELO corpus has been checked out for the correctness of entities and further annotated with relations by the annotators.

6 Results

In Figure 2, we give an example of an annotated citation. Drugs are in blue, food in green and adverse effects in cyan. Other entities are related to dosage, frequency and duration. Then, relations between these entities are marked up.

---

[5] http://ndb.nal.usda.gov/ndb/search/list
[6] http://www.codexalimentarius.org

Entities	Nb
drug	4,953
food	2,783
treated disease	645
drug effect	558
side effect	1,985
meal time	1,027
mode	539
dosage	767
duration	86
frequency	282

Table 1: Types and number of the annotated entities

Relation	Nb
decrease absorption	64
slow absorption	21
slow elimination	18
increase absorption	52
speed up absorption	4
new side effect	4
negative effect on drug	91
worsen drug effect	16
has side effect	434
increase side effect	239
positive effect on drug	23
reduce side effect	23
improve drug effect	14
treat	350
no effect on drug	145
without food	23
has effect	233
relation	706

Table 2: Types and number of the annotated relations

In Tables 1 and 2, we indicate the types and numbers of entities and relations manually annotated by experts. We can see for instance that among the most frequent relations we can find:

- drugs `have` side effects (n=434):
  Atovaquone suspension was well tolerated; diarrhea, nausea, fatigue, and rash were the most common adverse events.

- drugs `treat` disorders (n=350):
  Metrifonate is an inhibitor of cholinesterase effective in the treatment of Alzheimer's disease.

- food `increases` side effects (n=239):
  Animals infused ethanol-containing diets adequate in carbohydrate developed steatosis, but had no other signs of hepatic pathology.

- food `has no effect` on drugs (n=145):
  Azimilide dihydrochloride may be orally administered to patients without regard to the prandial state.

- food `has undefined effect` on drugs (n=233):
  In both studies the equivalence in AUC of DDVP was paralleled by equivalent effects on BChE inhibition.

Figure 2: Sample of annotated document (title and abstract) issued from a Medline citation
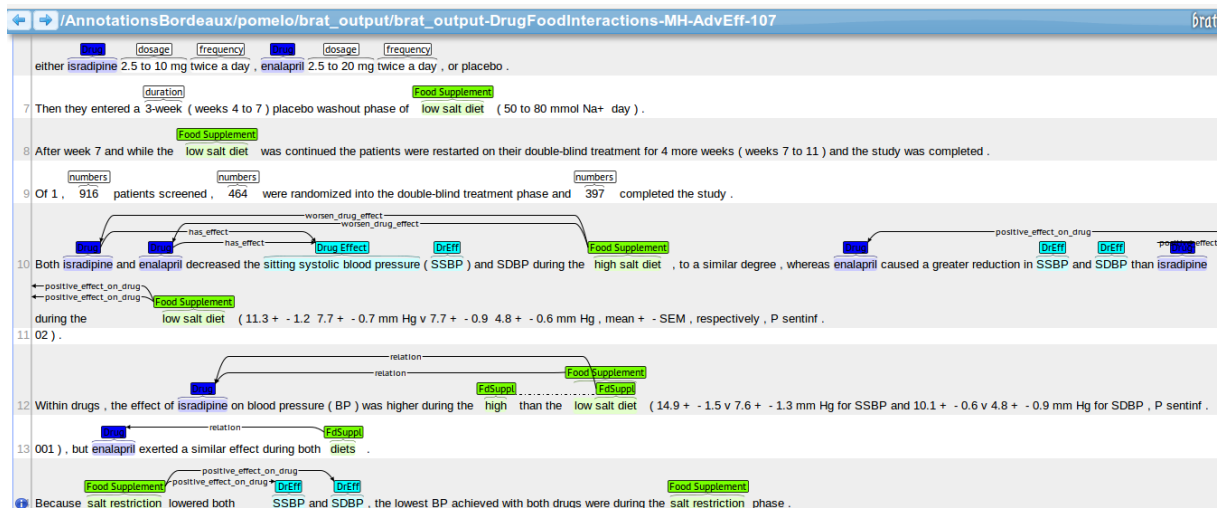
- food `has undefined relation` with drugs (n=706):
  We know that changing a customary diet to one high in protein and low in carbohydrate increases the rates of metabolism of antipyrine and theophylline, and shifting to an isocaloric diet of low protein- protein-high carbohydrate slows the rates of metabolism of these drugs.

This corpus will be made available for the research purposes. In this way, we expect to encourage the research on food/drug interactions.

7 Conclusion

We described our work done in order to create a corpus of biomedical literature (titles and abstracts) annotated with information on food/drug interactions. The corpus is called *POMELO*, namely *grapefruit* in French. Titles and abstracts are obtained from MEDLINE bibliographical base. We first propose a model of the food/drug-related information, which takes into account various aspects going from composition of drugs to their intake and possible adverse effects. Then, the annotation is performed by two experts: one resident and one medical doctor. Several entities are annotated, such as drugs, food, diseases and drug effects, meal time, and drug-related information (mode, dosage, duration and frequency). These entities are pre-annotated automatically and then checked out manually by the annotators. Then, relations between these entities are manually annotated. Among the most frequent relations, we can observe for instance: drugs `have` side effects, drugs `treat` disorders, food `increases` side effects, food `has no effect` on drugs, food `has undefined effect` on drugs, food `has undefined relation` with drugs.

We have several orientations for the future work on this research. One of the orientations is concerned by increasing the size and quality of the annotated corpus: (1) an updated MEDLINE query indicates that currently there are more citations indexed with the queried keywords; (2) another related MESH keyword (*herb/drug interactions*) can be exploited to enrich the corpus; (3) even if two experts have been involved in the annotation, the annotation can be done by other independent annotators; (4) finally, this English-language annotated corpus can be enriched with French-language citations and documents. Another orientation is related to the exploitation of this annotated corpus: (1) use of the annotations for creating the model for automatic extraction of food/drug interactions; (2) exploit this model for a systematic extraction of FDIs and their recording together with their evidence level; (3) creation of a knowledge base with food/drug interactions and their use by medical professionals and patients.

The *POMELO* corpus will be made available for the research purposes on the web site of the MIAM project (`https://miam.limsi.fr/`).

Acknowledgments

References

L Aagaard and EH Hansen. 2013. Adverse drug reactions reported by consumers for nervous system medications in europe 2007 to 2011. *BMC Pharmacol Toxicol* 14:30.

JK Aronson and RE Ferner. 2005. Clarification of terminology in drug safety. *Drug Saf* 28(10):851–70.

S Ayvaz, J Horn, O Hassanzadeh, Q Zhu, J Stan, NP Tatonetti, S Vilar, M Brochhausen, M Samwald, M Rastegar-Mojarad, M Dumontier, and RD Boyce. 2015. Toward a complete dataset of drug-drug interaction information from publicly available sources. *J Biomed Inform* 55:206–17.

J Björne, S Kaewphan, and T Salakoski. 2013. UTurku: Drug named entity detection and drug-drug interaction extraction using svm classification and domain knowledge. In *International Workshop on Semantic Evaluation (SemEval 2013)*. pages 1–9.

DK Brahma, JB Wahlang, MD Marak, and MCh Sangma. 2013. Adverse drug reactions in the elderly. *J Pharmacol Pharmacother* 4(2):91–4.

EG Brown, L Wood, and S Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.* 20(2):109–117.

GM Cote and E Choy. 2013. Role of epigenetic modulation for the treatment of sarcoma. *Curr Treat Options Oncol* .

MP Doogue and TM Polasek. 2013. The abcd of clinical pharmacokinetics. *Ther Adv Drug Saf* 4(1):5–7.

Stephany Duda, Constantin Aliferis, Randolph Miller, Alexander Slatnikov, and Kevin Johnson. 2005. Extracting drug-drug interaction articles from Medline to improve the content of drug databases. In *AMIA Symp*. pages 216–20.

Duke Med Health News. 2013. Grapefruit: enemy of many medications. in some patients, the interaction of fruit and drug may put their life and health at risk. *Duke Med Health News* 19(2):1–2.

DJ Greenblatt and H Derendorf. 2013. Grapefruit-medication interactions. *CMAJ* 185(6):507.

Milos Jovanovik, Aleksandra Bogojeska, Dimitar Trajanov, and Ljupco Kocarev. 2014. Inferring cuisine - drug interactions using the linked data approach. *Nature* 5(9346):1–7.

Sun Kim, Haibin Liu, Lana Yeganova, and W. John Wilbur. 2015. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of Biomedical Informatics* 55:23–30.

Artemy Kolchinsky, Anália Lourenço, Heng-Yi Wu, Lang Li, and Luis M. Rocha. 2015. Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. *PLoS ONE* 10(5):1–24.

M Krallinger, F Leitner, and A Valencia. 2009. The biocreative ii.5 challenge overview. In *BioCreative II 5 Workshop 2009 on Digital Annotations*.

Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2010. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology* 6(1).

C. Lacoste-Roussillon, P. Pouyanne, F. Haramburu, G. Miremont, and B. Bégaud. 2001. Incidence of serious adverse drug reactions in general practice: a prospective study. *Clin Pharmacol Ther* 69(6):458–462.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2016. Drug-drug interaction extraction via convolutional neural networks. *Comput Math Methods Med. 2016; 2016: 6918381* 2016:1–8.

Y. Moride, F. Haramburu, A. Requejo Alvarez, and B. Bégaud. 1997. Under-reporting of adverse drug reactions in general practice. *Br J Clin Pharmacol* 43(2):177–181.

NLM. 2001. *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. *www.nlm.nih.gov/mesh/meshhome.html*.

RxNorm. 2009. RxNorm, a standardized nomenclature for clinical drugs. Technical report, National Library of Medicine, Bethesda, Maryland. Available at www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.

Jodi Schneider and Richard D. Boyce. 2016. Acquiring and representing drug-drug interaction knowledge as claims and evidence. In *NLM Informatics Training Conference*.

Isabel Segura-Bedmar, Paloma Martinez, and Maria Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Lexical and Computational Semantics (*SEM)*. pages 341–350.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, pages 102–107. http://www.aclweb.org/anthology/E12-2021.

Z Wei, C Doria, and Y Liu. 2013. Targeted therapies in the treatment of advanced hepatocellular carcinoma. *Clin Med Oncol* 7:87–102.

David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 34:668–672. Database issue.

E Yom-Tov and E Gabrilovich. 2013. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *J Med Internet Res* 15(6):124–125.