

*Bachelor’s Degree in Bioinformatics (UPF-UPC-UB)*

*Final Grade Project*

|  |
| --- |
| **Elucidating drug-drug interactions underlying drug polypharmacy profiles**  Jorge Hernansanz1\*, Cristina Leal1, Gianluca Mazzoni1, Søren Brunak1  1 Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark.  \*Correspondence: jorge.hernansanz@cpr.ku.dk  Abstract  **Background:** Drug-drug interactions (DDI) are a common characteristic of adverse drug reactions (ADR) and suppose an emerging threat to public health. DDI data remains disperse and its integration into a unique source can benefit other research for an easier access to data.  **Methods:** We generated a complete compendium of DDI data through the integration of 15 publicly available drug sources. Then, we characterize each source by looking to its drug and DDI distribution. Finally, we performed a network cluster analysis in order to find significant DDI in clinical data.  **Results:** Across our databases, there are 4.288 unique drug names and 2.560.648 unique DDIs. We extracted 5 different DDI features; clinical significance, management options, evidence level, mechanism of action and side effects, that can be factorized and be used in posterior exploratory analysis. We observed in specific disease databases a bias for some of the ATC class. We found two significant clusters from the network analysis performed on the drug combination clinical data.  **Conclusion:** There is little overlap between the different databases, indicating thus the sparsity information across DDI interaction databases. Often being disease-specific and containing limited DDI information of certain drug classes. A network analysis approach was used for elucidating clinically significant DDI (major, moderate and minor) in hospital clinical data from Denmark.  **Keywords:**  **Supplementary information:** … |

1. **Background**

A DDI occurs when one drug modifies the pharmacological activity of other drug. It is a common characteristic of adverse drug reactions (ADR) [1], which is the focus of pharmacovigilance [2]. DDIs can be divided into two major groups, pharmacokinetic interactions, which involves changes in the absorption, distribution, metabolism and excretion of the drugs [3], and pharmacodynamics interactions; in which directly one drug changes the pharmacological effect of another one being additive or antagonist [4].

DDI data can be found in both public website sources (Drugbank, Crediblemeds…) and private commercial sources (Lexicomp, Stockley’s drug interactions…). Current DDI can be used to train statistical models that predicts new DDIs. Some DDI extraction challenges were hold in the past with the goal of covering different approaches when extracting DDI data from biomedical texts [5] [6]. Also, other project evaluated the performance of unsupervised and supervised machine learning methods for predicting potential DDIs [7].

Another challenge met is the integration and common consensus of DDI data. Previous studies have already tried to integrate multiple DDI databases [8] [9]. In a society where co-administration of drugs is becoming more and more frequent, a synthesis of this knowledge would help clinicians guide patients to effective and less noxious medication therapies [10].

In this paper, we extracted all DDI data from 15 publicly available databases, and gather all of them by mapping each drug to a common identifier; the ATC code. We provide a complete DDI dataset with a total of 3.049.224 DDIs. We also characterized drugs and factorized relevant clinical information so that it can be used as a tool for exploratory analysis of DDIs. Finally, we used our compendia to characterize and elucidate common patterns of DDI from a clinical setting.

This work provides a useful and important resource for the pharmacovigilance field, where it is trivial to harmonize and integrate DDI data in a unique compendium.

1. **Methods**
   1. **Anatomical therapeutical classification**

Anatomical Therapeutic Chemical (ATC) Classification System is a drug-related ontology that classifies within different levels the chemical group of each drug. Each drug may have several ATC codes but there is only one drug for each ATC code. It consists of 14 different anatomical groups, each of them reaching 5 different levels of differentiation.

The ATC classification system is developed and maintained by the World Health Organization Collaborating Center (WHOCC) [] for drug statistics methodology.

* 1. **Overview of DDI sources**

We created a drug- drug interaction dataset using content from 15 openly available database resources; Drugbank, KEGG, NDF-RT, Twosides, Interaktion Databasen, Crediblemeds, ONC-High, ONC-Non interruptive, HIV, HEP, CANCER, DDI Corpus 2011, DDI Corpus 2013, PK Corpus and NLM CV Corpus. Many of them are continuously updating, but others are just fixed projects just one time updated. The method each database was extracted is mentioned below. Several kind of sources were met such as regular XML format files, REST APIs, or information on website that was downloaded directly from there.

Each database contains specific information and features related to DDI pairs (Table 2). From a total of 15 public databases, we extracted five specific DDI features: clinical significance, evidence level, management, mechanism of action and side effects . Clinical significance feature associates to the DDI the level of change in the physiological effect of the drugs affected. Evidence level gives an estimation of how well documented is that DDI. Management options refers to how to proceed with the administration of the drug combination. Mechanism of action reflects the target/s involved due to the DDI. Side effect states the adverse reaction produced by the DDI.

Clinical significance, evidence level, and management features were transformed to digits in order to characterize in a numerical scale the magnitude of that information for each DDI.

Mechanism of action and Side effect features were also treated to be used in exploratory analyses. Mechanism of action output was formatted to a general vocabulary of actions easy to deal with. Side effect information was coded with the Medical Dictionary for Regulatory Activities (MedDRA) [] identifier. MedDRA identifier is the international medical terminology developed under the auspices of the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human use (ICH).

* + 1. **DrugBank**

The DRUGBANK database [11] is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information. Information retrieval was performed with the R package “DBparsed” [12], obtaining a brief description for each DDI. The version used in this study (5.0) was downloaded from the Drugbank website on January 2020.

* + 1. **KEGG**

The KEGG DRUG database [13] is a comprehensive drug information resource for approved drugs in Japan, USA, and Europe. Information is unified based on the chemical structure and/or the chemical component of active ingredients. Information retrieval was performed with its REST API [14], such as drugs’ mechanism of action and management DDI (“Contraindicated”, “precaution”). The version used in this study (93.0) was downloaded from the KEGG API in February 2020.

* + 1. **NDF-RT**

NDF-RT [16] is the National Drug File – Reference Terminology from the U.S. Department of Veteran Affairs, Veterans Health Administration (VHA). It is an extension of the VHA National Drug File (NDF) that combines its hierarchical drug classification with a multi-category reference model. “VA Drug interactions” was the category extracted, which contains clinical significance concept for each of the DDIs. The version used in this study dates from July 2014 as this was the last release VA was maintaining NDF-RT interactions, resulting in its removal from their posterior updates. Data was downloaded from the National Institutes of Health (NIH) repository in March 2020.

* + 1. **Twosides**

The Twosides project [15] consists of a comprehensive database of DDI side effects elaborated by Tatonetti’s laboratory. Twosides reports a total of 40 million DDI-related side effects, with their corresponding propensity scores for the evidence level. We downloaded the dataset from their lab repository in February 2020.

* + 1. **Crediblemeds**

Crediblemeds [17] is a clinically oriented information source that is used to guide clinical decision-making and safe use of drugs with risk of QT prolongation and/or torsades de pointes (TdP). Crediblemeds comprises a small DDI dataset and contains information such as Precipitant and Object drugs, as well as mechanism of action and management options features for each of the DDIs. The newest version of it dates from May 2010 with no identifiers, and it was extracted from [] as they already had parsed the drugs to their Drugbank identifiers.

* + 1. **Interaktion Databasen**

The Danish DDI database [18] is an electronic search tool that describes evidence-based interactions documented by clinical/case studies. It provides a common frame of reference for the on the handling of drug interactions in the Danish healthcare system. Data was retrieved from a public document provided by the Danish Medicines Agency. Information provided includes clinical significance, evidence level, and management administration features for each DDI. The version used in this study (XML-dato-3 release 2.7) was downloaded in March 2020.

* + 1. **ONC High-priority/ Non-interruptive**

ONC High-Priority (ONC-High) [19] consists of a set of high-severity DDIs for use in electronic health records (EHR). This dataset is characterized by contraindicated and highly clinicalclinically significant DDIs. Nevertheless, most of the DDIs included in this resource lack of primary literature supporting their evidence. We retrieve information regarding clinical significance, evidence level, and management options features. It is a one-time updated project published in April 2012.

ONC Non-Interruptive (ONC-NI) [20] consists of a set of low priority DDIs feasible for non-interruptive alerts that aims to reduce alert fatigue for the provider’s workflow of EHRs. We retrieved information regarding the clinical significance and evidence level features. Like ONC-High, this is also a one-time updated project published in September 2012.

These two projects comes from research organized by the Office of The National Coordinator for Health Information Technology (ONC) where in both cases the set of DDIs was a consensus between the different commercial drug providers that participated. The projects were already treated by [8], parsing their drug names to Drugbank identifiers. DDI datasets were extracted from that source in February 2020.

* + 1. **DDI Corpus 2011/2013**

DDI Corpus 2011 [5] relates to a project included in the DDI Extraction 2011 workshop focused on natural language processing (NLP) techniques for drug-drug extraction from text selected from the Drugbank database. This project dates from September 2011.

DDI Corpus 2013 [6] relates to a project included in the SemEval 2013 DDI Extraction challenge for the evaluation of NLP techniques applied to recognition of pharmacological substances and drug-drug extraction from Drugbank and Medline databases. This project dates from July 2013.

Isabel Segura’s Lab carried out both projects and its raw data can be accessed from her own GitHub. Those projects were already treated by [8], parsing their drug names to Drugbank identifiers. DDI datasets were extracted from that source in February 2020.

* + 1. **NLM CV Corpus**

NLM CV Corpus [21] is a DDI corpus used for training NLP techniques that identifies possible DDIs where cardiovascular drugs are involved. This project was already treated by [8], parsing their drug names to Drugbank identifiers. DDI datasets were extracted from that source in February 2020.

* + 1. **PK Corpus**

PK Corpus [22] is a specific-pharmacokinetic DDI corpus from a pharmacokinetic ontology used for DDI text mining analysis in drug product labels. This project was already treated by [8], parsing their drug names to Drugbank identifiers. DDI datasets were extracted from that source in February 2020.

* + 1. **HIV/HEP/CANCER Interactions**

HIV/HEP/CANCER drug interactions [23] [24] [25] correspond to three DDI resources from the Liverpool university. They report the clinical significance between a selected small amount of disease-specific drugs and a bigger set of standardized drugs, independently of if there is interaction or not. We extracted information related to the clinical significance and the evidence level for each DDI. The three of them were downloaded in April 2020.

* 1. **Parsing DDI pairs**

Here we describe a summary of the mapping procedure and its results for the different databases parsed. Table 1 provides a summary statistic about the unique number of drugs and DDI pairs, as well as the owner of the source and the version used when information was retrieved.

DrugBank and KEGG codify drugs with their own type of identifier, and provide a complementary section of data that links to ATC codes. This complementary section of DrugBank was also used for the four corpus: DDI Corpus 2011, DDI Corpus 2013, PK Corpus, and NLM CV Corpus, Crediblemeds dataset, and the ONC projects, that were already encoded with Drugbank identifiers.

NDF-RT uses its own type of identifier, but does not provide in a direct way a link of its drugs with the ATC code. For it, we used a comprehensive repository of biomedical ontologies called BioPortal [26] that provided us for indexes to parse the NDF-RT identifier to the Concept Unique Identifier (CUI), and from CUI to ATC code.

Interaktion Databasen encodes its drugs and every other component with its own identifiers. For the mapping procedure it was used a CSV file retrieved from the Danish Medicine Agency that accounts for the ATC code for each drug identifier stored in the database.

Twosides uses the RXNORM identifier for its drugs, which is the normalized clinical drug dictionary of the Unified Medical Language System (UMLS). As it didn’t provided us with a direct link to the ATC code, we retrieve form BioPortal an index to parse RXnorm to CUI identifier, and used the previous index retrieved of CUI – ATC identifier.

HIV, HEP and CANCER interactions are just online websites where it is only registered the regular drug name. Therefore, drugs were name-linked to Drugbank identifiers so as to be linked to the ATC code (From the index obtained from Drugbank), and those that were not mapped to DrugBank ones were manually annotated their ATC code using as index the WHO ATC list of 2017.

* 1. **Clinical data**

In order to elucidate DDI on clinical data, we collected drugpairs from actual prescription data from electronic health records (EHR) covering the period January 2008 to June 2016 from Denmark. The clinical data contained drug pairs information including the number of patients and the age distribution. All possible pairwise combinations were compared for finding the coverage of DDIs in our compendia.

* 1. **Visualization tools**

Overlap analysis between our sources was carried out with upserR [], an R package that visualize the different interaction intersections between sources, showing a range from the interactions that are unique for each database, to those ones that are covered by most databases.

For the visualization of the interaction distribution for each database, we used an R package called “Circlize” [] that gives the distribution desired in a circular layout.

* 1. **Network analysis**

We performed a network analysis using Cytoscape [27], an open source software platform used for visualizing and integrating networks. We used the MCODE algorithm [25], which is a graph theoretic-based clustering algorithm with three stages: network weighting, complex detection and optional post-processing. This cluster algorithm detects densely connected regions in large interaction networks. Although it has been primarily used for protein-protein interactions, its use for DDI can help us identify significant drug clusters.

Parameters were personalized for our analysis. “Degree cutoff” was set to 2 in order to prevent single-connected nodes in clusters. We represented the clinical drug pairs as a directed network and as 2 drugs may share 2 edges, the “K-Core” parameter was set to 3 to force each node to have 3 connections per node and to connect at least with 2 drugs. Following the FAQ section of MCODE [28] it is stated to apply the “MCODE level 3” parameters in case of seeking for dense interconnected clusters. This changed parameter “Node Score Cutoff” to 0.1, “fluff” to zero, and disabling “haircut” option.

1. **Results**
   1. **Characterizing DDIs**

An overlapping analysis was performed involving the 15 databases parsed, ranging from DDIs that are unique for a database to the DDIs that can be found in most of them (supplementary file 1). We found our DDI data to have a significant limited overlap where we can observe there is an enormous quantity of DDIs that just appear in one, two, or three databases, whereas very few DDIs appear in higher degree intersections (Figure 1 a, b). The highest degree of overlapping achieved by our compendia is of nine databases with only five DDIs in that section.

An analysis for the distribution of DDIs was run inside each database. For it, we categorized drugs according to their first ATC level. We checked the distribution of each kind of drug across databases, in figure 2, and the intersection of interaction distribution for each database, in figure 3.

Attending the distribution of drugs, we observed that the databases follow a similar one where the level 1 ATC class most relevant corresponds to the Cardiovascular System drugs (C). In addition, Nervous System drugs (N) and Antiinfectives for Systemic Use drugs (J) have a major overall prevalence among most of classes. Some databases account for a more detailed distribution of drugs. For example, the NLM CV Corpus and ONC NI have a bigger proportion of Cardiovascular System drugs (C). This is expected from the first one as it is used to identify DDIs in drug product labeling affecting cardiovascular drugs [8]. In contrast, Crediblemeds contains a higher degree of Musculo-skeletal System (M) drug-related interactions compared to the rest of databases.

Looking at the distribution of interactions inside each database, we observe again that the most predominant ATC class correspond to the one of Cardiovascular System. Both NLM CV Corpus and ONC-NI display a big set of ATC class C receiving most of interactions from the rest of classes. It is significant to observe Crediblemeds that most drugs from the ATC class J are interacting with the class C and that ATC class M, although having a good percentage of drugs inside the database, plays an insignificant role at the interactions.

* 1. **Cluster analysis**

We represented in a network how clinical DDIs are distributed according to their clinical significance: minor, moderate and major interactions. DDIs were filtered to actual drug pairs prescriptions, allowing the analysis to have a real implementation in a clinical environment. We considered the anatomical level of the ATC classification and ranked the clinical significance of interactions in four levels (0,1,2,3), ranging from 0 when a DDI is not expected to present pharmacokinetic changes to 3 when a DDI represents a major change in the pharmacokinetics one of the drugs (Table 2). For the purpose of our analysis, DDIs with a score of zero were removed.

We found two relevant clinical DDI clusters (Figure 4). The first cluster consisted of 10 nodes and 90 edges, all of them with a clinical significance level 3, indicating a total major level of interactions between them. This cluster is characterized predominantly by nervous system drugs (N).

The second cluster we identified is less dense and more heterogeneous in terms of clinical significant levels (indicate number of edges for each level……), with 21 nodes and 78 edges . The great majority of interactions correspond to moderate interactions (level 2). The ATC level N remains again as the class that groups more nodes in the cluster.

1. **Discussion**

Up to day there is still a limited overlap between the different publicly DDI sources available on the internet. Moreover, this project did not cover all existing

databases and DDI knowledge continues appearing, indicating the necessity of a continuous work to have a well-updated Compendia.

Network analysis lead us to a small cluster of mayor type interaction between its drugs. We used SUPERCYP [], a web tool that checks interactions between drugs and P450 cytochromes, to check for the mechanism of action of these drugs and find similarities. We observed nine of them are highly interacting with cytochromes 2D6 and 3A4, and the other one des not present any cytochrome interaction, even though it is included in the cluster. Althogh needed more exhaustive analysis, this sets a precedent that DDI information features recollected can be used in order to detect potential significant patterns.

As future work, this project may continue with the assignment of extracting more DDI sources, and be used as a complement for other research papers.

1. **Conclusion**

In this project, we integrated a wide range of publicly available sources of DDI information. We analyzed the overlap across the different databases and we found that there is actually little overlap between DDI pairs and high heterogeneity between the features extracted from the sources. Interestingly, Drugbank, KEGG, and Twosides appeared in nearly all intersections for the most overlapped DDIs, justifying that those databases are the richest sources of DDIs.

Then, we looked to the distribution of the interactions for each of the databases and checked for significant patterns. We observed that specific-related DDI source are more likely to have a predominant ATC class being the hotspot for the majority of interactions, whereas non-disease specific ones tend to have a greater interaction diversity between their classes.

Finally, we performed a cluster analysis on the elucidated clinical DDI using clinical significance information. Results from this analysis revealed a highly dense cluster of 10 drugs that were to interact between each of them.