Universidad Autónoma de Madrid
Escuela Politécnica Superior

Master's Degree in Bioinformatics and Computational Biology

Master's Degree Final Project

INTEGRATION OF NEW PHARMACOGENOMICS FUNCTIONALITIES IN PANDRUGS

Author: María José Jiménez Santos
Tutor: Fátima Al-Shahrour Núñez
Tutor: Elena Piñeiro Yáñez
Rapporteur: Enrique Carrillo de Santa Pau

FEBRUARY 2019

INTEGRATION OF NEW PHARMACOGENOMICS FUNCTIONALITIES IN PANDRUGS

Author: María José Jiménez Santos
Tutor: Fátima Al-Shahrour Núñez
Tutor: Elena Piñeiro Yáñez
Rapporteur: Enrique Carrillo de Santa Pau

Unidad de Bioinformática (BU)
Centro Nacional de Investigaciones Oncológicas (CNIO)

Escuela Politécnica Superior (EPS)
Universidad Autónoma de Madrid (UAM)

FEBRUARY 2019

Abstract

Abstract

Pandrugs is a bioinformatics platform for providing personalized drug prescription to cancer patients based on tumor genetics and drug efficacy. This resource computes a GScore (Gene Score), which measures the relevance of tumor genetic variants in cancer initiation and progression, and a DScore (Drug Score) of drug efficacy against different targets. Then, Pandrugs outputs a ranking of the best therapeutic candidates for a particular patient. Nowadays, PanDrugs does not take into account patient's germinal variants in order to prioritize some drugs over others, although there is evidence of the influence of these variants in drug responses. Pharmacogenomics studies how the same drug administered to people with different germinal variants generates different responses that may affect drug efficacy and toxicity. Drugs used in cancer therapy are very aggressive and their toxicity may be increased due to pharmacogenomics interactions. For this reason, we consider important to incorporate this kind information in PanDrugs. We have developed a new score, which we have called ToxScore, that measures the noxiousness of a drug for patients with different genetic variants. Moreover, we have proven that ToxScores vary between drugs with the same therapeutic use. This result supports our decision of incorporating pharmacogenomics information about toxicity in PanDrugs to suggest, for each patient, the best and least noxious drug that targets a specific gene product. Finally, we used sequencing data from a paraganglioma patient in order to test whether or not PanDrugs ranking varied when using this new ToxScore. We were able to identify germinal variants that were associated to a higher risk of adverse drug reactions in response to some of the top ranked drugs. Consequently, PanDrugs output was reordered and therapeutic candidates with the same effectiveness but higher toxicity descended in the ranking of therapeutic candidates. Thus, our new ToxScore is able to integrate pharmacogenomics data in order to reorder PanDrugs ranking and penalize drugs with an increased risk of causing adverse drug reactions to the patient.

Key words

Bioinformatics, personalized medicine, cancer, genetic variants, drug targets, pharmacogenomics, drug response, drug toxicity, adverse drug reactions.

Acknowledgments

I want to thank all members of CNIO's Bioinformatics Unit, and specially my tutors Fátima and Elena. Thank you for letting me be part of this project, for your support and help. It has been a fulfilling experience.

Contents

Figure Index

Table Index

1 Introduction

1.1 Motivation

PanDrugs is a bioinformatics platform for providing personalized drug prescription that can be accessed at `www.pandrugs.org`. This resource was developed by members of CNIO' s BU (Spanish National Research Center's Bioinformatics Unit) in collaboration with SING (Next Generation Computer Systems Group) from the University of Vigo. The objective of this project was to apply precision medicine to cancer patients.

Personalized or precision medicine is an emergent field whose aim is to make improved medical decisions for each patient according to thousands of people's epidemiological, clinical and genomics data [1]. Precision medicine is particularly interesting in cancer context, due to the high mutational heterogeneity that characterizes this illness. In fact, patients with the same type of cancer can have distinct mutations that affect cancer development and progression. Moreover, tumors within the same patient or even cells within the same tumor can contain different mutations [2]. Thus, personalized medicine, as it takes into account genetic variability and tries to give the best treatment in each case, could be key to improve the outcome of current cancer therapies.

Some of the most frequent cancer treatments include surgical removal of the tumor, chemotherapy, radiotherapy, immunotherapy or drugs that target mutant genes that contribute to tumor initiation and progression [3]. PanDrugs centres in the later, suggesting the best therapeutic candidates for each cancer patient. This platform mines PanDrugsdb, a database with information about the implication of genes in different types of cancer and the effectiveness of drugs that target those mutated genes. When the user inputs the type of tumor and its mutated genes or somatic variants, PanDrugs computes a GScore (Gene Score) and a DScore (Drug Score) using this information. GScores range from 0 to 1 and indicate the biological relevance and level of evidence of the implication of a particular gene or genetic variant in cancer. DScores range between -1 and 1 and estimate target's resistance (negative values) or sensitivity (positive values) to the drug. Then, PanDrugs ranks all the drugs in PanDrugsdb according to these two scores. The drugs with higher DScores that target genes with GScores closer to 1 are suggested as the best therapeutic candidates for that particular patient [4].

Nowadays, PanDrugs does not take into account patient's germinal variants in order to compute this ranking. Nevertheless, many pharmacogenomics (PGx) studies highlight the influence

of genetic polymorphisms in drug response. Genetic variants are changes in the DNA sequence with respect to the reference genome. Those variants with a frequency $> 1\%$ in the population are called polymorphisms. Moreover, genetic variants can be classified as germinal or somatic depending on the affected tissue. Germinal variants are present in all cells of a person's body, including germinal cells, and can be transmitted to progeny. On the contrary, somatic variants appear spontaneously in a certain tissue [5]. Somatic variants include tumor mutations and are the ones used as input by PanDrugs. Polymorphisms can occur in intergenic regions but also in coding sequences. In the later case, they can provoke a non-synonym change of the amino acid sequence of the resulting protein. This change may alter protein stability or function and, if it affects drug metabolizing enzymes, the polymorphism can increase the risk of side effects. This is particularly concerning in cancer treatment, since therapeutic drugs are already very aggressive and their toxicity may be increased due to the presence of certain polymorphisms in patient's genome [6]. Thus, we consider important to take into account germinal variability when giving personalized drug prescription, in order to select the best non-toxic treatment for each patient.

Side effects are defined as phenotypical changes caused by the interaction of a drug with a molecule different from its target [7]. In this project, we are interested in adverse side effects, referred hereafter as ADRs (Adverse Drug Reactions). ADRs are important causes of morbidity and mortality, and occasion significant financial expenses to the healthcare system [8]. There are several resources that contain drug ADR data obtained from different electronic records. Moreover, in the recent years, pharmacogenomics databases with information about variants that increase the risk of drug ADRs have been created.

In this project, we aim to integrate pharmacogenomics information, as well as ADR severity and frequency data, in order to compute a score of drug noxiousness due to genetic variation. This toxicity score could be used to penalize those drugs in PanDrugs ranking that have been reported to have toxic effects in people with the same germinal variants as the patient. That way, we expect to improve PanDrugs personalized drug prescription for cancer patients.

1.2 Objectives

1) **To obtain drug-variant and drug-ADR associations, as well as ADR severities and frequencies from the source databases.**

   We downloaded four databases that contained drug-ADR associations: SIDER (Side Effect Resource), IntSide, SPL-ADR-200db and PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium). Moreover, we downloaded data from PharmGKB (Pharmacogenomics Knowledgebase), a database with information about genetic variants that contribute to a higher susceptibility to suffer ADRs in response to drug treatments.

   The drug terms were standardized using PanDrugsdb, DrugBank, DrugCentral, PubChem and EMA (European Medicines Agency) Medicines database. The ADRs were mapped to MedDRA (Medical Dictionary for Drug Regulatory Activities) PTs (Preferred Terms) and LLTs (Low Level Terms), unifying LLTs that were synonyms or lexical variants.

2) **To compute a metric to measure the noxiousness of each drug in response to germinal variants.**

   For each ADR of each drug, we retrieved all the available severity and frequency data. Moreover, we measured the proportion of databases where this association was reported.

In addition, a level of evidence of the drug-variant-ADR association was obtained from PharmGKB. Then, we computed a ToxScore ranged from 0.02 to 1 for each compound and variant. In order to do so, we used the evidence of the pharmacogenomics association, the overlapping between databases and the severity and frequency of each ADR of a particular drug. As there were many missing data, we also computed a Reliability Level for each ToxScore. This Reliability ranked from A to C and represented the amount of missing data. ToxScores in A group were the most reliable ones and scores in group C were considered unreliable. This way, for a particular patient, drugs with a higher ToxScore would be considered more noxious than drugs with a lower ToxScore within the same Reliability Level.

3) To verify that drugs that target the same gene or that belong to the same family have different ToxScores.

We retrieved all PanDrugsdb compounds with a computed ToxScore that targeted *EGFR* (Epidermal Growth Factor Receptor) gene or belonged to tyrosine kinase receptor inhibitors family. Then, we verified that drugs with the same target or that belonged to the same family of compounds had different toxicities.

4) To confirm that ToxScores modify the ranking of therapeutic candidates returned by PanDrugs.

Finally, we used real patient data to corroborate that PanDrugs rankings varied when we took into account pharmacogenomics information about toxicity. We input the somatic variants of the primary tumor and the metastasis of a paraganglioma patient to PanDrugs. Then, we reordered the list of best therapeutic candidates using the ToxScores for the drugs reported in the ranking and the germinal variants of the patient. ToxScores were used to penalize drugs with an increased risk of causing adverse drug reactions to that patient.

1.3 Metodology and work plan

In the table below, we have broken down the tasks that we completed during this project.

Table 1.1: Work plan

Tasks / Subtasks	Hours
T1. Project proposal	**35**
T1.1 Read bibliography	15
T1.2 Redact the project proposal	20
T2. Download the source databases to obtain drug-variant and drug-ADR associations, as well as ADR severity and frequency data	**3**
T3. Drug and ADR standardization	**125**
T3.1 ADR standardization	75
T3.1.1 Map all ADRs to MedDRA PT and LLT terms	*5*
T3.1.2 Manual standardization of all LLTs within the same PT that are synonyms or lexical variants	*70*
T3.2 Drug term standardization	50
T3.2.1 Automatic standardization of all drug terms included in PanDrugsdb	*5*
T3.2.2 Manual standardization of the rest of the drug terms using DrugBank, DrugCentral, PubChem and EMA Medicines databases	*45*
T4. Data merging	**12**
T4.1 Merge all data using as common fields the standardized drug and ADR terms	6
T.4.2 Subset the entries with PanDrugsdb compounds and remove duplicates	6
T5. ToxScore computation	**40**
T5.1 Determine the overlapping of each drug-ADR pair in the source databases	5
T5.2 Assign a numeric value to each level of pharmacogenomics evidence	5
T5.3 Unify the severity data and assign a numeric value to each category	10
T5.4 Unify the frequency data and assign a numeric value to each category	10
T5.5 Compute a ToxScore for each drug-variant association using the pharmacogenomics evidence, drug-ADR overlapping information, ADR severity and frequency	5
T5.6 Compute a Reliability Level for each ToxScore based on the amount of missing data	5
T6. Verify that drugs that target the same gene or are from the same family have different ToxScores	**15**
T6.1 Select a gene that plays an important biological role in cancer and that is targeted by a considerable number of drugs with computed ToxScores	3
T6.2 Select a drug family with a large number of members with known ToxScores	2
T6.3 Study the differences between the ToxScores of drugs that target the same gene or belong to the same compound family	10
T7. Confirm that the ToxScore modifies the ranking of therapeutic candidates returned by PanDrugs	**15**
T7.1 Execute PanDrugs using the somatic variants of the primary tumor and the metastasis of a paraganglioma patient	1
T7.2 Add to PanDrugs output the ToxScores corresponding to the ranked drugs and patient's germinal variants	10
T7.3 Re-rank the PanDrugs output taking into account the ToxScores	3
T7.4 Check that the two rankings differ	1
T8. Project defense	**55**
T8.1 Redact the dissertation	40
T8.2 Create the presentation	14
T8.3 Present the dissertation to the laboratory	0.5
T8.4 Present the dissertation to the committee	0.5
TOTAL HOURS	**300**

2

State of the Art

The aim of this chapter is to clarify key concepts that will be further explained in the rest of the dissertation.

2.1 Cancer

The term cancer refers to a group of diseases characterized by the uncontrolled proliferation of cells. As a consequence, a primary malignant tumour is formed. There are more than 100 types of cancer depending on the tissue where the malignant tumor arises. In late stages of this disease, the cells of the primary tumor can become motile, disseminate to other body parts through the blood or the lymph and generate a secondary neoplasm. The process of migration and reproduction of a malignant tumor in another tissue is called metastasis [9].

Cancer is one of the major health problems worldwide. In our country, cancer is the second principal cause of death after cardiac diseases [10]. Some of the most frequent cancer treatments include surgical removal of the tumor, chemotherapy, radiotherapy, immunotherapy or drugs that target mutant genes that contribute to tumor apparition and progression [3]. However, many cancer patients die, specially those in late stages of the illness. In fact, metastasis is responsible of about 90% of cancer deaths [11]. For this reason, a big part of the scientific community is trying to understand the molecular mechanisms that underlie cancer to ultimately find a cure.

Cancer initiation and progression is the result of a Darwinian evolutionary process in which the acquisition of mutations in driver genes confers a selective advantage to cancer cells. Genes whose mutation do not increase cancer cell fitness are called passengers [12]. Mutations in driver genes can activate proto-oncogenes or inactivate tumor suppressor genes [13]:

- **Proto-oncogenes:** Genes that favor cell division. When they mutate, they become constitutively active and are called oncogenes. These genes suffer gain-of-function and dominant mutations, so only one of the two alleles must be altered in order to activate proto-oncogenes to oncogenes .
- **Tumor suppressor genes:** Genes that restrict cell cycle progression and promote programmed cell death. These genes suffer loss-of-function and recessive mutations, so both

alleles must be altered in order to inactivate a tumor suppressor gene.

Mutations in driver genes increase cell division, angiogenesis and telomere stabilization, whereas they allow cancer cells to ignore apoptotic signals. Moreover, some mutations affect DNA repairing enzymes, favoring the acquisition of new mutations [14].

Cancer is a complex disease that results from the interaction between the genotype and the environment. The major part of cancers are sporadic, due to the acquisition of somatic mutations in driver genes. These mutations are caused by mutagens such as tobacco, UV radiation or viruses. Nevertheless, there are also hereditary cancers caused by germline mutations that can be transferred to the progeny. In hereditary cancers, the genetic component is stronger than the environmental one, meaning that the carrier has an increased risk of developing cancer at an early age. Moreover, hereditary cancers are more aggressive and several tumors can arise simultaneously in the same or different tissues [15].

One of cancer's main characteristics is genetic heterogeneity, which can be classified as intratumoral and intertumoral. Intratumoral heterogeneity refers to the existence of distinct cell clones with different mutations within the same tumor. On the contrary, intertumoral heterogeneity occurs when two tumors, of the same (intrapatient) or different patients (interpatient), present distinct mutations [2]. Due to this mutational heterogeneity, precision medicine has become an interesting approach to treat cancer.

2.2 Personalized medicine and PanDrugs

Personalized or precision medicine is an emergent field whose aim is to make improved medical decisions for each patient according to thousands of people's epidemiological, clinical and genomics data [1]. The potential advantages of this approach include better disease prevention, more accurate diagnosis, more effective treatments and safer drug prescription, minimizing adverse drug reactions. Consequently, a reduction of healthcare expenditures is expected when personalized medicine becomes a routine practice. [16].

Personalized medicine takes into account genetic variability and tries to give the best treatment to each patient. Genetic variants are changes in the DNA sequence with respect to the reference genome and variants with a frequency $> 1\%$ in the population are known as polymorphisms. Single Nucleotide Polymorphisms (SNPs) are the most common type of genetic variants. It has been estimated that there are approximately 3.3 million SNPs in a person's genome [17]. Genetic variants can be classified according to the tissue they affect as [5]:

- **Germinal variants:** They occur in germinal cells. These variants are inherited from the progenitors and can be transmitted to the progeny if a mutant cell participates in the fertilization process. Consequently, all descendant's cells will carry this variation.
- **Somatic variants:** They occur in developing somatic cells and therefore, they can not be transmitted to progeny. The mutant cell will divide and generate a clone of cells with the same variant. This type of mutations are the ones that can cause sporadic cancers.

Variant calling is the process of identifying the specific genetic variants of an individual. This process consists of three steps [18]:

1) **Carry out a Whole Genome Sequencing (WGS) or a Whole Exome Sequencing (WES) and obtain a FASTQ file.**

   WGS and WES are two different Next-Generation Sequencing (NGS) techniques. WGS consists in sequencing all the genome of an individual, whereas WES only sequences the protein-coding regions. Consequently, WES is less costly than WGS. FASTQ files contain the short nucleotide sequences and qualities resultant from the sequencing process.

2) **Align the sequences to a reference genome and obtain SAM or BAM files.**

   Next, the short sequences in the FASTQ file are mapped against the reference sequence. The result of the alignment is saved in SAM (Sequence Alignment Map) or BAM (Binary Alignment Map) files. SAM format stores the plain-text version of the binaries contained in BAM files.

3) **Identify the differences between the sequenced genome and the reference and write a VCF file.**

   A VCF (Variant Calling File) stores information about the genetic variants of the individual. This kind of files can be given as input to PanDrugs.

PanDrugs is a bioinformatics platform for providing personalized drug prescription to cancer patients. This platform mines PanDrugsdb, a database with information about the implication of genes in different types of cancer and the effectiveness of drugs that target those mutated genes. The current version of PanDrugsdb stores 56,297 drug-target associations obtained from 4,804 genes and 9,092 compounds.

When the user inputs the type tumor and a list of mutated genes or a VCF with somatic variants, PanDrugs computes a GScore (Gene Score) and a DScore (Drug Score) using this information. GScores range from 0 to 1 and indicate the biological relevance and level of evidence of the implication of a particular gene or genetic variant in cancer. DScores range between -1 and 1 and estimate target's resistance (negative values) or sensitivity (positive values) to the drug. Then, PanDrugs ranks all the drugs in PanDrugsdb according to these two scores. The drugs with higher DScores that target genes with GScores closer to 1 are suggested as the best therapeutic candidates for that particular patient [4].

2.3 Pharmacogenomics and drug toxicity

Pharmacogenomics studies how the same drug administered to people with different germinal polymorphisms generates different responses. This variation in drug response may include distinct drug efficacy and toxicity. Polymorphisms can occur in intergenic regions but also in protein-coding sequences. Non-synonym polymorphisms cause changes in the amino acid sequence that may alter protein stability or function. When these polymorphisms affect drug metabolizing enzymes, they can increase the risk of side effects [6].

The metabolism of exogenous compounds (xenobiotics) takes place in liver cells and is divided in four phases. Phase 0 consists in the transport of xenobiotics into the cell. In Phase I, polar groups are added to xenobiotics in order to make them more hydrophilic, so they can be excreted through the urine. Compounds that are still not polar enough after Phase I undergo Phase II. In this second phase, the xenobiotics are conjugated to a larger polar group, increasing their

solubility. Finally, Phase III consists in the transport of the transformed xenobiotics outside the cell. Phases 0 and III are mediated by proteins that transport xenobiotics across cellular membranes.

Cytochrome P450 (CYP) is a superfamily of enzymes involved in the biosynthesis and degradation of endogenous compounds such as steroids, lipids, and vitamins. Moreover, they also metabolize exogenous compounds. These enzymes are subdivided into several families based on amino acid sequence similarities. CYP1, CYP2, CYP3 and CYP4 families play a major role in Phase I liver metabolism catalyzing the transformation of drugs. The genes that encode for these proteins have a high number of polymorphisms that affect drug toxicity. Moreover, several transmembrane proteins that transport xenobiotics in Phases 0 and III also have a great number of genetic variants that have been associated with differences in drug responses. There are two major transporter families: the Solute Carriers (SLC) and the ATP Binding Cassette (ABC) proteins. The former family is involved in Phase 0 metabolism whereas the later participates in Phase III [19].

In this work, we are particularly interested in the genetic variants that increase the risk of adverse drug reactions (ADRs) in response to drugs. Side effects are defined as phenotypical changes caused by the interaction of a drug with a molecule different from its target [7]. ADRs are important causes of morbidity and mortality, and occasion significant financial expenses to the healthcare system [8]. There are several databases that contain information about drug ADRs. The data contained in these resources has been obtained via Natural Language Processing (NLP) from two types of electronic documents with drug information: SPL (Structured Product Labeling) and SmPCs (Summary of Product Characteristics). SPLs are documents used by the FDA (Food and Drug Administration) for exchanging information about chemical products [20] whereas SmPCs contain information about drugs approved by EMA (European Medicines Agency) [21]. These two formats contain all the adverse reactions reported during the clinical trials of a particular drug. The ADRs can be found in different sections according to their level of severity. Moreover, the frequency of each ADR can also be extracted from these documents.

3 Databases

The aim of this chapter is to provide a brief description of the databases that were used in this project. Due to the great number of sources, we considered convenient to write a entire chapter aside from Materials and Methods (Chapter 4).

3.1 Terminologies for standardizing ADRs

3.1.1 MedDRA

MedDRA (Medical Dictionary for Drug Regulatory Activities) is a standardized hierarchical medical terminology developed in the late 1990s. MedDRA hierarchy is structured in five levels, ordered from more specific to more general as Low Level Terms (LLT), Preferred Terms (PT), High Level Terms (HLT), High Level Group Terms (HLGT) and System Organ Classes (SOC).

LLT level contains synonyms and lexical variants (such as words in different order, American or British forms, abbreviations or singular and plural forms) of a unique medical concept, the PT term. Moreover, a LLT can be a more precise term for a PT (i.e DRUG DEPENDENCE and HEROIN ADDICTION). Any PT contains at least itself as LLT and is grouped into one or more HLTs based upon anatomy, pathology, physiology, etiology or function. Related HLTs are contained in the same HLGT level. Finally, HLGTs are grouped by manifestation site, etiology or purpose into one SOC. Each PT is assigned to a primary SOC but can be grouped under other secondary SOCs. However, each LLT is linked to only one PT. In this project we worked with ADRs that belonged to LLT and PT levels.

MedDRA is broadly used through all the phases of development of a drug, from clinical trials to postmarketing surveillance. This allows to unify the ADRs of different drug products and therefore to create consistent databases of adverse events. Moreover, MedDRA has been translated and it is maintained in eleven languages, including English and Spanish, but also Portuguese, Italian, French, German, Czech, Dutch, Hungarian, Chinese and Japanese. Each MedDRA term is associated to a unique eight digit code that is conserved among the different translations, avoiding the loss of information if more than one language is used.

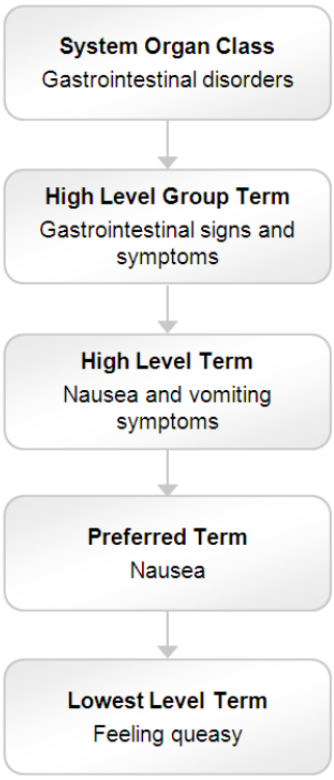


Figure 3.1: MedDRA hierarchy

Example retrieved from `www.meddra.org/how-to-use/basics/hierarchy`

MedDRA also accepts feedback from users, who can propose changes in the existing terminology, improvements in the structure of MedDRA, new terms to be included or even translation corrections. MedDRA terminology is updated twice a year, in March and September and can be freely downloaded for non-commercial purposes. For industry, a license must be purchased. MedDRA terminology is available at `www.meddra.org` [22].

3.1.2 UMLS

UMLS (Unified Medical Language System) is another biomedical terminology developed by the United States (US) National Library of Medicine. In this work, we did not use this database directly, but IntSide, SIDER, SPL-ADR-200db and PROTECT contained information derived from UMLS.

UMLS is formed by three knowledge sources that are updated quarterly: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus is the base of UMLS and integrates several biomedical vocabularies from different areas, including MedDRA terminology. UMLS groups synonym terms into the same concept and represent the relationships between them, as well as cross-references to external data. The Semantic Network is a catalogue of the different semantic types (such as drugs, diseases or adverse events) and relationships, including hierarchical (e.g. “isa“, “part of“), or associative (e.g. “location of“, “caused by“) relationships. The SPECIALIST Lexicon contains the lexical variation of UMLS terms [23].

3.2 Databases for standardizing drug names

3.2.1 STITCH

STITCH (Search Tool for Interacting Chemicals) is a database of known and predicted functional and physical interactions between proteins and chemicals. The current version of STITCH contains 430,000 chemicals and 9,643,763 proteins from 2,031 organisms [24]. Each chemical in the database is given a STITCH ID. This ID starts with CIDs or CID0 if the compound is a stereoisomer and CIDm or CID1 if it is a flat compound (i.e. merged stereoisomers). In this project we did not used STITCH directly, but SIDER database utilized the STITCH IDs to identify the different drugs.

STITCH database can be downloaded for free. Moreover, it can be accessed through its API or its web at `www.stitch.embl.de`.

3.2.2 PubChem

PubChem is a chemical compound database owned by the US National Institutes of Health (NIH) that contains information about drugs and other small chemicals, as well as proteins, nucleotides, lipids, carbohydrates and modified macromolecules. It is an open database, meaning that everybody can upload his or her data and that the rest of the people will have free access to it. PubChem data sources include government agencies, chemical vendors, journal publishers, researchers and curation efforts.

This resource organizes the data into three inter-linked databases: Substance, Compound and BioAssay. Substance database contains all the information uploaded by the different data sources. All the submitted Substances with the same chemical structure are grouped into the same Compound database record. A PubChem Compound record includes a brief description of the drug, as well as its synonyms, patent identifiers, other IDs and scientific articles that mention the compound. The BioAssay database contains information about biological assay experiments performed on substances.

PubChem is the largest database of public chemical information. Nowadays, it contains 249,392,519 Substances, 97,177,104 Compounds and 1,067,516 BioAssays. Most of the data in PubChem can be directly downloaded, accessed programmatically or through its web page at `www.pubchem.ncbi.nlm.nih.gov`. However, some datasets can not be downloaded in bulk due to license agreements [25].

In this project, we have accessed the PubChem Compound database through its webpage in order to standardize the drug terms that were not directly found in PanDrugsdb.

3.2.3 DrugBank

DrugBank database is a resource developed by the Wishart Research Group at the University of Alberta that contains information about drugs and drug targets. DrugBank is one of the most widely used reference drug resources in the world and its current version stores 7,800 different entries. Each drug entry contains more than 200 different fields, including a drug description, drug synonyms, commercial names in the US, Canada and the European Union (EU), chemical structure and formula, mechanism of action, metabolism, drug interactions and drug targets.

DrugBank can be accessed through its website at `www.drugbank.ca`. Moreover, bulk data is freely available for non-commercial use, though a permission is needed in order to download the database. If someone wishes to use DrugBank data for commercial profit, a license must be

purchased. Moreover, there are two datasets with unrestricted access that contain DrugBank IDs, names, synonyms and structures to allow the integration of DrugBank in other projects [26].

In this work, we have used DrugBank webpage to search drug terms that were not directly found in PanDrugsdb.

3.2.4 DrugCentral

DrugCentral is an online resource that contains information for active pharmaceutical ingredients approved by the FDA and other drug regulatory agencies. DrugBank contains more than 4,444 active pharmaceutical ingredients and 20,617 drug synonyms.

Each record in this database includes a brief description of the drug, as well as its structure, synonyms and brand names. Moreover, other information such as indications, adverse events, mechanism of action, bioactivity against targets, information about drug repurposing and links to external databases is provided. DrugCentral was created by Division of Translational Informatics at University of New Mexico in collaboration with the IDG (Illuminating the Druggable Genome). The regulatory agencies from US (FDA), EU (EMA) and Japan (PMDA, Pharmaceuticals and Medical Devices Agency) are periodically monitored in order to incorporate new approved drugs to DrugCentral. This database is a free resource that can be accessed online at `www.drugcentral.org` or downloaded in a relational database format [27].

In this project, we have used DrugCentral webpage to search drugs that could not be directly found in PanDrugsdb.

3.2.5 EMA Medicines database

EMA (European Medicines Agency) Medicine database is a catalogue of commercial drugs approved in the EU. Each entry in the database is formed by an EPAR (European Public Assessment Report) summary of a medicine. EPARs are reports of the drugs authorized in the EU that also contain questions and answers for the general public. Some of these questions address the therapeutic indications of a drug, the routes of administration, the mechanisms of action, the drug risks and benefits and why was that drug approved. Moreover, each entry in EMA Medicine Database includes the brand name, the active principle and the package insert of the drug. In addition, information about changes, safety reviews or withdrawal of the corresponding medicine is provided [28]. EMA Medicine database is freely available at `www.ema.europa.eu/en/medicines`.

In this project, we have accessed EMA Medicines database through its web page in order to standardize the drug terms that were found neither in PanDrugsdb, nor in PubChem, nor in DrugBank nor in DrugCentral.

3.3 Drug-ADR and pharmacogenomics data sources

We downloaded five source databases:

- **Databases that contained drug-ADR associations:** SIDER (Side Effect Resource), IntSide, SPL-ADR-200db and PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium).
- **PharmGKB (Pharmacogenomics Knowledgebase):** Database with information about genetic variants that contribute to a higher susceptibility to suffer ADRs in response to drug treatments.

The data contained in these resources was mainly retrieved via NLP from two types of electronic documents with drug information: SPLs (Structured Product Labelings) and SmPCs (Summaries of Product Characteristics).

3.3.1 SIDER

SIDER (Side Effect Resource) is a database of drugs and ADRs created by members of the European Molecular Biology Laboratory (EMBL). Since it was the first free resource of its kind, it has been used as data source for other databases such as IntSide. SIDER's data was obtained from SPLs using NLP. The current release, SIDER 4, contains data on 1,430 drugs, 5,880 ADRs and 140,064 drug-ADR pairs. This database is composed by several files. In this project we downloaded three of them: drug_names.tsv, meddra_all_se.tsv and meddra_freq.tsv. A fragment of the merged files is shown below.

Table 3.1: SIDER (subset)

STITCH ID	DRUG	MedDRA PT	MedDRA LLT	FREQUENCY
CID100000143	LEUCOVORIN	PRURITUS	PRURITUS	POSTMARKETING
CID100000085	CARNITINE	ABDOMINAL PAIN	ABDOMINAL PAIN	9%
CID100000143	LEUCOVORIN	RASH	RASH	POSTMARKETING
CID100000085	CARNITINE	ABDOMINAL PAIN	ABDOMINAL PAIN	6%
CID100000247	BETAINE	DECREASED APPETITE	ANOREXIA	UNCOMMON

1) STITCH flat compound ID; 2) Drug name; 3) MedDRA PT term; 4) MedDRA LLT term; 5) PT or LLT frequency.

SIDER can be browsed interactively at `www.sideeffects.embl.de`, were all data is also available for free download. In addition, there is a Github repository at `www.github.com/mkuhn/sider` where to report database errors, so the users can filter them out and the authors can improve SIDER in the upcoming versions [29].

3.3.2 IntSide

Intside is a web server created by the Structural Bioinformatics and Network Biology Group at the Institute for Research in Biomedicine of Barcelona. One of IntSide's main data sources is

SIDER. It provides a catalog of 1,175 side effects caused by 996 drugs. The drugs in IntSide are classified into eight biological or chemical groups:

- **Biological:** Biological processes, molecular functions, pathways, protein interactions and therapeutic targets.
- **Chemical:** Fragments, scaffolds and structural terms.

The web server displays the common biological and chemical fields for all drugs that cause the same ADR. These networks allow to infer molecular mechanisms that lead to a particular adverse reaction.

Intside is available at `www.intside.irbbarcelona.org` and its data can also be downloaded for free [30]. In this work we used the file intside_drug_effects.tsv. Its structure is shown in the table below.

Table 3.2: IntSide

EFFECT ID	ADR NAME	DRUG ID	DRUG NAME	FREQ
C0013922	EMBOLISM	CID100004205	MIRTAZAPINE	NULL
C0030193	PAIN	CID100444033	CICLESONIDE	NULL
C0013378	DYSGEUSIA	CID100054454	SIMVASTATIN	NULL
C1384353	INFESTATION	CID100151165	APREPITANT	NULL
C0042109	URTICARIA	CID100000727	LINDANE	NULL

1) UMLs concept ID; 2) MedDRA LLT term; 3) Flat compound STITCH ID; 4) Drug name, 5) Frequency (all fields are null).

3.3.3 SPL-ADR-200db

SPL-ADR-200db is a database created by Dina Demner-Fushmanthat *et al.* that contains ADR information for 200 FDA approved drugs. This dataset was obtained by manually annotating the adverse reactions mentioned in the SPLs of each drug. SPL-ADR-200db contains 5,098 distinct ADRs that have been mapped to UMLS and MedDRA terms. Moreover, information about the section of the SPL from where the ADR was extracted is available [31]. This information was used to estimate the severities of the ADRs in our dataset. SPL-ADR-200db was split into training and test sets and used in the Text Analysis Conference (TAC) 2017 Adverse Drug Reaction Extraction from Drug Labels task. The goal of this challenge was to create NLP applications for extracting ADR information from SPLs. SPL-ADR-200db is freely available at `www.osf.io/9hsxq/`, from where we downloaded the file FinalReferenceStandard200Labels.csv. A subset of this table is shown below.

Table 3.3: SPL-ADR-200db (subset)

Drug Name	PT ID	MedDRA PT	LLT ID	MedDRA LLT
NUCYNTA	10000125	ABNORMAL DREAMS	10000125	ABNORMAL DREAMS
NUCYNTA	10001497	AGITATION	10001497	AGITATION
NUCYNTA	10002198	ANAPHYLACTIC REACTION	10002218	ANAPHYLAXIS
NUCYNTA	10002424	ANGIOEDEMA	10002424	ANGIOEDEMA
NUCYNTA	10002855	ANXIETY	10002855	ANXIETY

1) Drug name; 2) MedDRA PT ID; 3) MedDRA PT term; 4) MedDRA LLT ID, 5) MedDRA LLT term.

3.3.4 PROTECT

The PROTECT ADR database (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium ADR Database) is a downloadable Excel file that contains all ADRs included in the SmPCs of medicines authorized in the EU. The current database has been updated up to 30 June 2016 and contains 917 drugs, 2,004 LLTs and 3,153 PTs.

This file can be downloaded for free from `www.imi-protect.eu/adverseDrugReactions.shtml` [32]. A subset of the table that was used in this project is shown below.

Table 3.4: PROTECT (subset)

PRODUCT	MedDRA LLT	MedDRA PT	PT ID	FREQUENCY
ZALVISO		CONSTIPATION	10010774	4
IMATINIB ACCORD		HAEMATOMA	10018852	3
SOMAVERT		HUNGER	10020466	3
EPISALVAN		INFLAMMATION OF WOUND	10054923	0
ZALASTA		LIBIDO DECREASED	10024419	4

1) Commercial name; 2) MedDRA LLT term; 3) MedDRA PT term; 4) MedDRA PT ID, 5) Frequency (0 = Unknown; 1 = Very rare; 2 = Rare; 3 = Uncommon; 4 = Common; 5 = Very common).

3.3.5 PharmGKB

PharmGKB (Pharmacogenomics Knowledgebase) is database that contains information about the influence of genetic variants in drug responses. PharmGKB's variants are classified as SNPs and haplotypes (group of SNPs that are inherited together). SNPs are named using their reference SNP (rs) ID number. These drug-variant-ADR relationships have been extracted from the literature using NLP techniques and manual curation. The current PharmGKB database contains a total of 21,332 annotations.

Each drug-variant-ADR annotation is assigned a level of evidence that measures the confidence in the association as determined by the PharmGKB curators. The levels of evidence range from 1 to 4, depending on how many articles report the association. Annotations of level 1 are the ones with more significant studies supporting them and annotations with level 4 are

based in individual case reports or studies or *in vitro* molecular or functional assays. Levels 1 and 2 are subdivided in groups A and B. Levels 1A and 2A include those associations that are used in clinics or are particularly well documented. 1B and 2B annotations do not meet these conditions but have the same level of evidence as sublevels 1A and 2A, respectively.

PharmGKB is a publicly available online resource that can be accessed at `www.pharmgkb.org` [33] . During this project, we downloaded clinical_ann_metadata.tsv, clinicalVariants.tsv and relationships.tsv files and merged them in order to obtain a unique table whose structure is shown below.

Table 3.5: PharmGKB (subset)

DRUG	GENE	VARIANT	MEDDRA LLT	LLT ID	EV
CALCIUM	VDR	rs731236	ASTHMA	10003553	
GRANISETRON	ABCB1	rs1045642	NEOPLASM	10028980	3
CODEINE	CYP2D6	CYP2D6*10	ARRHYTHMIA	10003119	
CAPTOPRIL	ACE	rs1799752	DISORDER KIDNEY	10013231	
FLUVOXAMINE	ABCB1	rs2032583	MAJOR DEPRESSION	10057840	3

1) Drug name; 2) Gene symbol; 3) Variant (SNP or Haplotype); 4) MedDRA LLT term; 5) MedDRA LLT ID; 6) Level of evidence of the association.

3.4 PanDrugsdb

The current version of PanDrugsdb stores 56,297 drug-target associations obtained from 4,804 genes and 9,092 compounds. We have used a subset of this database that includes information about drugs and their targets. A fragment of this table is shown below.

Table 3.6: PanDrugsdb (subset)

GENE	SOURCE	STANDARD	SHOWN	FAMILY
BRCA1	PX-12	141400-58-0	141400-58-0	OTHER
KIT	QUIZARTINIB	QUIZARTINIB	QUIZARTINIB	RECEPTOR TYROSINE KINASE_INHIBITOR (KEGG), FLT3 INHIBITOR(Cmap)
PIK3C2G	NUTLIN-3	NUTLIN-3	NUTLIN-3	MDM INHIBITOR(Cmap)
CDK11B	PHA-793887	PHA-793887	PHA-793887	CDK INHIBITOR(Cmap)
FOSL1	TPA	TPA	TPA	OTHER

1) Gene symbol; 2) Name of the drug in the original source; 3) Standard drug name; 4) Name shown by PanDrugs; 5) Drug family.

This table was utilized during drug standardization process. Initially, we searched for coincidences between source drug terms and fields 2, 3 or 4. If there was any match, the drug term

was standardized to the name shown by PanDrugs. Then, the terms that did not match any column underwent manual standardization. Again, if some of their synonyms matched fields 2, 3 or 4, they were standardized to the name shown in the 4th column of Table 3.6. Moreover, we used this table to retrieve all tyrosine kinase receptor inhibitors and the names of drugs directed against *EGFR* gene product. We did so in order to assess if ToxScores varied among members of the same drug family or with the same target.

4

Materials and Methods

4.1 Materials

In this work, we used Ubuntu 14.04.5 LTS, Python 2.7.6 and R 3.4.4. Moreover, we utilized several databases (See Databases, Chapter 3) for obtaining and standardizing the data:

1. **Data sources:** We extracted the data from SIDER, IntSide, SPL-ADR-200db, PROTECT and PharmGKB.
2. **Drug term standardization:** We accessed PubChem, DrugBank, DrugCentral and EMA databases in order to standardize the drug terms that were not found in PanDrugsdb.
3. **ADR term standardization:** We used MedDRA database to standardize the ADR terms at the LLT level and to group them in their corresponding PT.

In addition, we utilized PanDrugsdb to retrieve drugs that targeted *EGFR* gene or that were tyrosine kinase receptor inhibitors. Moreover, we input VCF files containing somatic data of the primary tumor and the metastasis of a paraganglioma patient to PanDrugs. Then, we extracted patient germinal data from another VCF file and re-ranked PanDrug's output according to the precomputed ToxScores.

Finally, the plots were drawn using the following R packages: ggplot2, ggrepel, gridExtra, RColorBrewer, scales, waffle, UpSetR and eulerr.

4.2 Methods

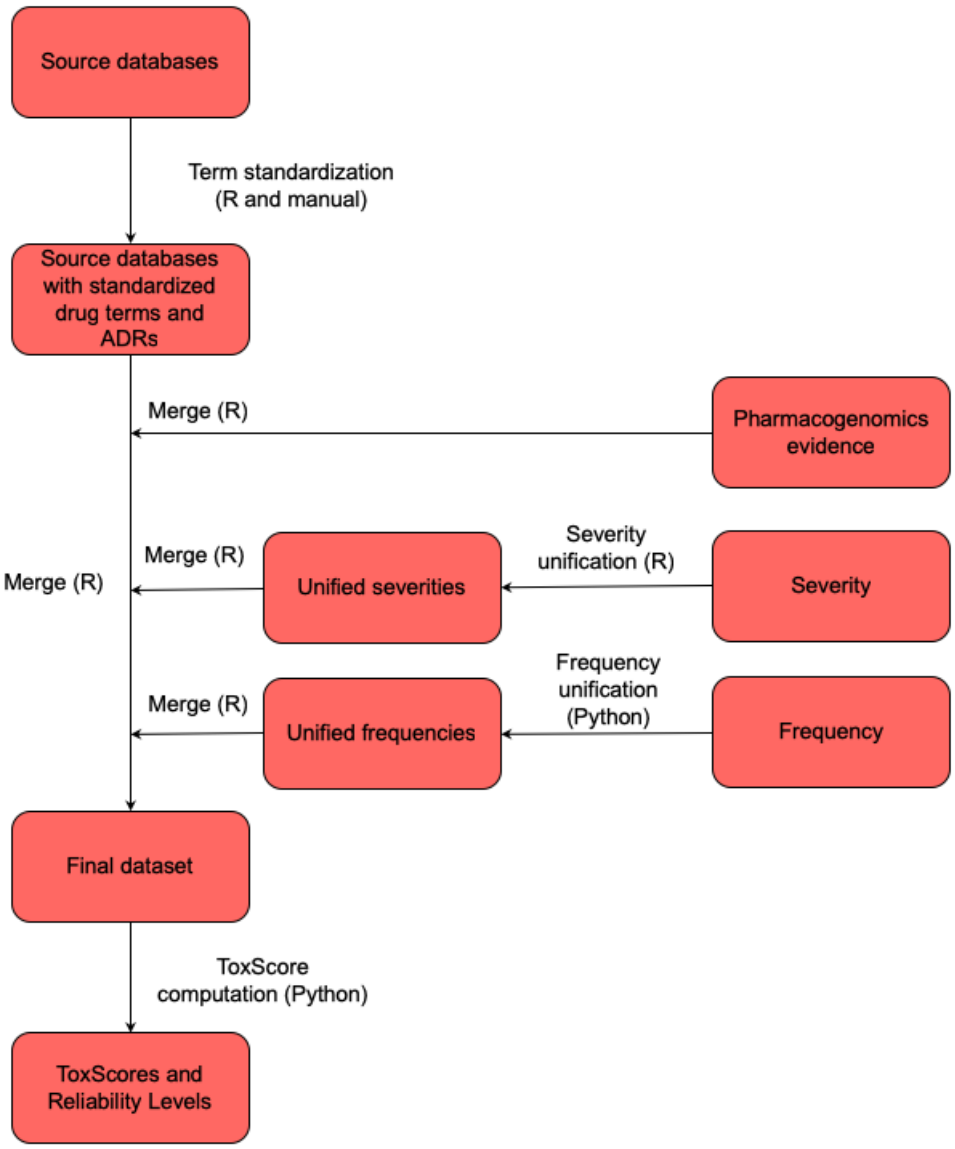


Figure 4.1: Workflow: from source databases to ToxScores

Main steps followed to obtain ToxScores from source databases. Programming languages used in each step are indicated in parentheses.

4.2.1 Drug term standardization

The same compound can be named using its active principle, chemical nomenclature or commercial name. Thus, in order to avoid redundant information in our final table, it was necessary to standardize the medicines retrieved from the five source databases.

The drug names that matched any PanDrugsdb compound were standardized to the name shown in PanDrugs ranking. Those terms that were not found in PanDrugsdb were manually standardized using PubChem, DrugBank, DrugCentral and EMA databases. For these terms we chose, if available, the synonym that matched the name used by PanDrugs.

4.2.2 ADR standardization

MedDRA is a hierarchical medical terminology structured in five levels, ordered from more specific to more general as Low Level Terms (LLT), Preferred Terms (PT), High Level Terms (HLT), High Level Group Terms (HLGT) and System Organ Class (SOC). In this project we worked with ADRs that belonged to LLT and PT levels. LLTs are synonyms and lexical variants of a unique medical concept, the PT term. Moreover, an LLT can be a more precise term for a PT. Each PT contains at least itself as LLT and each LLT is linked to only one PT. For more information, please refer to MedDRA section in Databases (Chapter 3, Section 3.1.1).

In the five source databases, ADRs were described using MedDRA terms. IntSide and PharmGKB databases only included LLT terms whereas SIDER, SPL-ADR-200db and PROTECT also contained the PT level. However, some PTs in the later databases were not associated to any LLT. As we wanted our final dataset to include both PT and LLT levels for all entries, we had to fill in empty PT cells in IntSide and PharmGKB data and blank LLT cells in SIDER, SPL-ADR-200db and PROTECT databases.

First, we asked MedDRA for permission to download their database. After we received the data, we created a table with all MedDRA PTs and their corresponding LLTs, as well as the ID codes for both PT and LLT terms (see Table 4.1). In order to fill in empty PT cells, we used this table to map all LTTs from the five source databases with their corresponding PT. Blank LLT cells were filled in with the same name shown in the PT level.

Table 4.1: MedDRA PTs and LLTs

PT ID	MedDRA PT	LLT ID	MedDRA LLT
10016747	FLAIL CHEST	10063696	PARADOXICAL CHEST MOVEMENT
10043871	TINEA NIGRA	10043871	TINEA NIGRA
10028034	MOUTH ULCERATION	10045352	ULCERATION OF MOUTH
10068322	ORAL PAPILLOMA	10062604	ORAL WART
10003402	ARTHROPOD STING	10003402	ARTHROPOD STING
10011469	CRYING	10069592	PERSISTENT CRYING

Still, there was redundancy among the LLT terms within the same PT. Two LLTs could be referring to the same ADR but the words were in different order (i.e PAIN KNEE and KNEE PAIN), written in British or American forms (i.e. HAEMATOMA and HEMATOMA), abbreviated (i.e. ALT INCREASED and ALANINE AMINOTRANSFERASE INCREASED), or written in singular or plural forms (i.e. MUSCLE CRAMP and MUSCLE CRAMPS). Moreover, two LLTs could be synonyms (i.e. MALIGNANT NEOPLASM and CANCER). Hence, we manually standardized all the LLTs within the same PT in our data. We preferably kept as standard terms those LLTs that coincided with the name of the PT, were written in singular and British forms and did not contained abbreviations. In order to asses whether two LLTs described the same medical condition, we used SIDER's web page, which includes a brief description of the ADRs contained in the database and a list of MedDRA synonyms.

4.2.3 ADR frequencies

Frequency data was extracted from PROTECT and SIDER databases. Frequencies were expressed in two ways: as a percentage or as a discrete value, following the terminology of CIOMS (Council for International Organizations of Medical Science) [34] (see Table 4.2). We used the later format to express the frequencies of ADRs in our final table.

Table 4.2: CIOMS terminology for ADR frequencies

Discrete Value	**Percentage**
Very Common	$freq \geq 1/10$
Common or Frequent	$1/10 > freq \geq 1/100$
Uncommon or Infrequent	$1/100 > freq \geq 1/1000$
Rare	$1/1000 > freq \geq 1/10000$
Very Rare	$freq < 1/10000$

In PROTECT, the frequencies of LLTs were encoded with a number from 0 to 5. Each number corresponded to a frequency measured in the discrete scale (0 = Unknown, 1 = Very rare, 2 = Rare, 3 = Uncommon, 4 = Common and 5 = Very Common). However, in SIDER, ADR frequencies were expressed in the two formats, for both LLT and PT levels. Moreover, after ADR standardization there were frequency duplicates for the same pair drug-ADR. As we did not have access to the original percentages of the categorical values, we decided to express all frequencies in a discrete scale. Also, we kept both LLT a PT frequencies.

First, we catalogued frequency data as PT frequencies or LLT frequencies. Note that some LLTs and PTs coincided. In those cases, the same frequency was included in both groups. Then, we transformed the PROTECT encoded data to their corresponding discrete values. Next, we converted the percentages. If the same drug-ADR pair had several frequency percentages, we transformed the mean of all values to discrete format.

This way we obtained six frequency levels, ordered from higher to lower as VERY COMMON, COMMON, UNCOMMON, RARE, VERY RARE and POSTMARKETING. POSTMARKETING frequencies were retrieved from SIDER database and corresponded to frequencies reported once the drug had been commercialized. As POSTMARKETING frequencies were obtained from individual reports and other factors could had influenced the manifestation of the ADR (such as patient's diet, life style, interaction with other drugs, etc.), we considered POSTMARKETING frequency the lowest level.

Finally, if there were still duplicates within the categorical values, we kept the most abundant term. If there were ties, we chose the level that expressed higher frequency.

4.2.4 ADR severities

Severity data was obtained from SPL-ADR-200db. Different from frequencies, ADR severities were considered independent of the drug. This way, each ADR was given a unique severity value (i.e SUDDEN DEATH as response to 5-Fluorouracil or Acamprosate was assigned to the maximum level of severity in both cases).

On our final dataset, ADR severities were grouped in three categories: ADVERSE REACTIONS, WARNINGS AND PRECAUTIONS AND BOXED WARNINGS. They were named after the sections in the SPL from where the ADRs were retrieved. ADVERSE REACTION section includes all the adverse events reported in clinical trials, from mild to severe. In WARNINGS AND PRECAUTIONS, the serious or clinically significant ADRs are named. Finally, the BOXED WARNINGS appear in a black box that highlights the most serious or even life-threatening adverse side effects. According to this, we ordered the different severity values from higher to lower as BOXED WARNINGS, WARNINGS AND PRECAUTIONS and ADVERSE REACTIONS. Note that the ADRs that appear in BOXED WARNINGS are also included in WARNINGS AND PRECAUTIONS. In the same way, ADRs named in these two sections can

also be found in ADVERSE REACTIONS [35]. When different categories were assigned to the same ADR, we kept the higher level as its severity value.

4.2.5 ToxScore computation

The ToxScore represents the noxiousness of a drug in response to a specific genetic variant. It was computed using four parameters:

- **Drug-ADR overlapping:** The proportion of sources that reported each drug-ADR association. The maximum value of overlapping was 1 and the minimum was 0.2. Note that this variable would never be equal to 0, as all drug-ADR associations were, at least, retrieved from PharmGKB.
- **Evidence of the pharmacogenomics association:** The level of reliability reported by PharmGKB for each annotation. It ranged from 0.25 to 1 (see Table 4.3). Not available (NA) data were forced to be 0.
- **Frequency:** The frequency of each ADR for a specific drug. We gave priority to LLT frequency. If this value was missing, then we used the PT frequency. The discrete values were assigned to a number from 0.1 to 1 (see Table 4.3). If an entry did have neither LLT nor PT frequency data, this parameter was forced to be 0.
- **Severity:** The severity of each ADR. The categorical values were assigned to a number from 0.33 to 1 (see Table 4.3). NA data were forced to be 0.

The ToxScore was computed using the following formula:

$$ToxScore\ = 0.3 \times PGx\ evidence\ + \sum_{i=1}^{i} \frac{0.1 \times \frac{ADR\ sources}{total\ sources} + 0.3 \times frequency + 0.3 \times severity}{N} \tag{4.1}$$

Where i is each drug's ADR and N is the total number of different ADRs of a particular drug. We weighted the four variables so the maximum value of the ToxScore was 1. Note that ADR frequency, severity and pharmacogenomics evidence were given the same weight, which was higher than the one assigned to database overlapping. This reflects the fact that we considered the evidence, frequency and severity more important than the overlapping between databases to assess the noxiousness of a drug in response to a genetic variant. The minimum value of the ToxScore was 0.02, which corresponded to drug-variant pairs that were only reported in PharmGKB and had neither pharmacogenomics evidence nor frequency nor severity data available.

Ideally, for the same patient, a drug with a ToxScore closer to 1 would be more noxious than another compound with a lower ToxScore. However, there were many missing data. As a result, lower ToxScores did not necessarily meant less noxiousness, but rather reflected absence of data. For this reason, we computed a Reliability Level for each ToxScore (Equation 4.2).

$$Reliability\ = \sum_{i=1}^{i} \frac{0.3 \times number\ of\ NA}{N} = \begin{cases} A & \text{if } 0 \leq Reliability\ < 0.3 \\ B & \text{if } 0.3 \leq Reliability\ < 0.6 \\ C & \text{if } 0.6 \leq Reliability\ \leq 0.9 \end{cases} \tag{4.2}$$

Table 4.3: Parameters for ToxScore computation

	Discrete Value	Number
Pharmacogenomics evidence	NA	0
	4	0.25
	3	0.5
	2B	0.75
	2A	0.75
	1B	1
	1A	1
Frequency	NA	0
	POSTMARKETING	0.1
	VERY RARE	0.28
	RARE	0.46
	UNCOMMON	0.64
	COMMON	0.82
	VERY COMMON	1
Severity	NA	0
	ADVERSE REACTIONS	0.33
	WARNINGS AND PRECAUTIONS	0.66
	BOXED WARNINGS	1

The result of this equation was ranged from 0 to 0.9, as the maximum number of NAs per entry was 3. This result was encoded into a categorical value in order to facilitate its comprehension. This way, each ToxScore was assigned a Reliability Level equal to A, B or C. ToxScores in A category were the most reliable, since the majority of entries for that particular drug-variant pair were complete. ToxScores in B level were moderately reliable, because the majority of entries lacked of pharmacogenomics evidence, frequency or severity data. Finally, ToxScores with a Reliability of C were computed with many missing data, thus they could not be trusted.

4.2.6 Plots

In Experiments and Results (Chapter 5) we have included the following types of plots:

- **Histograms:** Represent ranges of the data using a predefined number of bars (bins). The height of each bar is proportional to the number of data points that are found within that range. This plot is used to show the distribution of the data, although the number of selected bins can alter its shape. If data is divided into a very small (many bars) or a very large (few bars) number of bins, the distribution showed by a histogram can differ greatly from the true distribution.
- **Density plots:** A variation of the histogram that shows a smoothed distribution of the data. In these plots, the distribution's shape is not determined by the number of bins. The peaks indicate where the values are concentrated.

- **Boxplots:** Boxplots represent the dispersion and distribution of the data. The line inside the box indicates the median and, the box fragments above and below this line, represent the third and the first quartiles respectively. The total height of the box corresponds to the interquartile range. The whiskers represent the variability outside the 1st and 3rd quartiles and their extremes the maximum and minimum values of the distribution. Outliers are indicated as points outside the box and in line with the whiskers. In addition, we have represented the mean value of the distribution using a diamond.

- **UpSet plots and Venn Diagrams:** These plots are equivalent and represent the logical relations between different sets.

  - In Venn diagrams, each set appears as an ellipse of a different color. The area of each ellipse has been scaled according to the number of elements that the set contains. The overlapping between ellipses indicates the intersection between the sets.

  - UpSet plots are another form of representing the same concept. In these plots, the size of the sets is represented with blue horizontal lines. Moreover, the intersections of the sets are marked with black dots in a matrix. Above each matrix column, a vertical bar indicates the number of terms found in that set intersection. We believe that this plot is particularly useful with more than 4 different sets. In those cases, we find an UpSet plot clearer than a Venn Diagram.

- **Doughnut Plots and Waffle Plots:** They are equivalent to pie charts and are used to represent percentages.

  - Waffle plots consist in a grid of 10x10 where each square corresponds to a unit. In waffle plots, the percentages have been rounded to the closest integer in order to fill all the grid. The exact percentages are shown in the legend.

  - Doughnuts plots are equivalent to pie charts. The percentages are represented as arcs of the circumference of a pie plot.

- **Enrichr Bar Graph:** Enrichr is an open source enrichment analysis tool freely available at `www.amp.pharm.mssm.edu/Enrichr`. Enrichr accepts a list of genes as input and returns a rank of enriched terms that can be visualized as a bar graph. The plot reported in this dissertation represents the top 10 GO (Gene Ontology) biological processes that are overrepresented in the genes of our final dataset. These terms were sorted by decreasing p-value, which was obtained after applying a Fisher's exact test. The longer and lighter the bars the more significant the term [36].

All plots but Enrichr bar graph were done using R and ggplot2 itself or other dependent packages.

5

Experiments and Results

5.1 Drug term standardization

First, we downloaded SIDER, IntSide, SPL-ADR-200db, PROTECT and PharmGKB databases. Then, in order to merge all the data sources by common drug and ADRs, we performed a previous standardization step.

At the beginning, we had 2,868 different drug terms. We found out that 2,125 of these terms were already contained in PanDrugsdb. For the rest of drug terms, we manually searched for synonyms in PubChem, DrugBank, DrugCentral and EMA databases (See Chapter 4, Section 4.2.1). After the standardization process, the number of unique drugs dropped to 1,815.

As Figure 5.1 shows, 257 out of these 743 manually annotated drug terms were found in PubChem, DrugBank and DrugCentral, whereas 162 drugs had entries in EMA. Note that we only searched in EMA when we could not find an entry for the drug in the other three databases. This is the reason why EMA and PubChem, DrugBank and DrugCentral sets are mutually exclusive. The terms looked up in EMA were mainly mixture drugs and vaccines contained in PROTECT database. Moreover, we noted that PubChem and DrugBank had 72 common entries that were not found in DrugCentral. The size of this intersection was remarkable when we compared it with the number of common terms between PubChem and DrugCentral (15) or DrugBank and DrugCentral (8). In addition, only 4 terms were found exclusively in DrugCentral. From this data, we concluded that the majority of terms in DrugCentral were redundant and their synonyms could have been retrieved from PubChem or DrugBank. Among these three databases, PubChem seemed to be the more complete since it contained the highest number of non-overlapping entries.

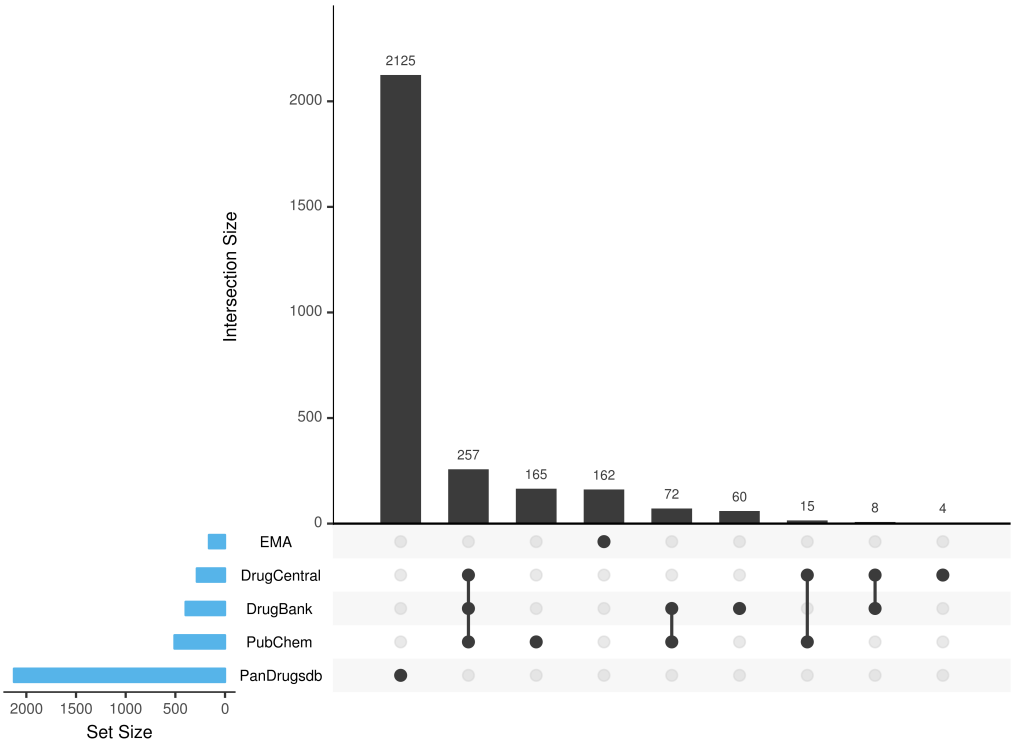


Figure 5.1: Sources from where drug synonyms where retrieved during drug standardization process

2,125 of the original drug names were already in PanDrugsdb. Only the terms that were not found in PanDrugsdb underwent manual standardization using the other databases. This is why PanDrugsdb set does not intersect with the other groups. Drug terms were first searched in PubChem, DrugBank and DrugCentral and, if no match was found in either database, the term was then queried in EMA. For this reason, terms retrieved from EMA do not overlap with the drug names found in the other databases. The majority of terms found in DrugCentral were also found in PubChem, DrugBank or both. Among PubChem, DrugBank and DrugCentral databases, the former was the one that contained most non-overlapping drug entries.

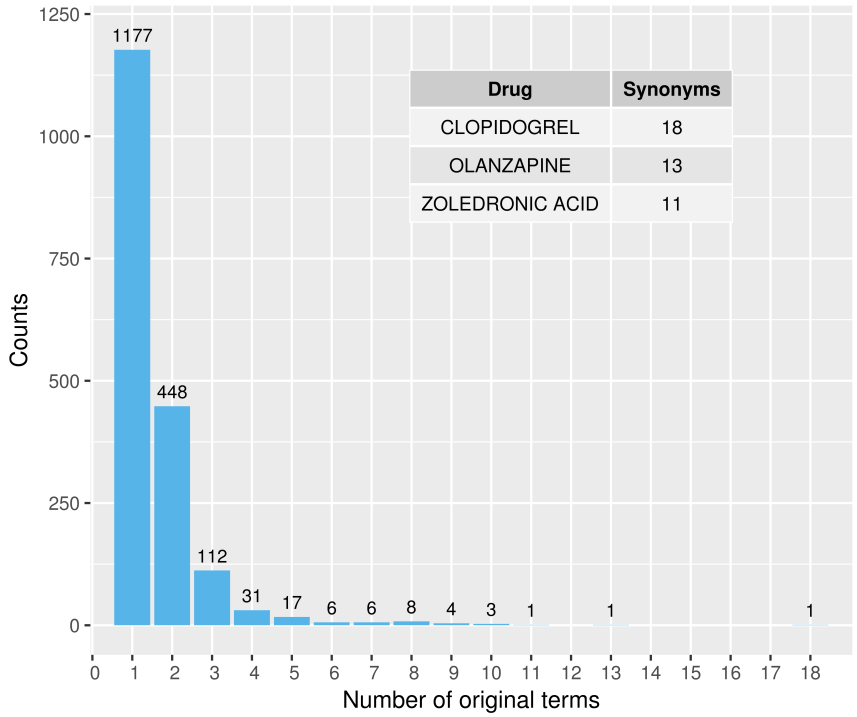


Figure 5.2: Number of original terms per standardized drug name

The distribution of the number of original drug terms per standard drug name was logarithmic. The majority of standardized drug terms included one or two original names retrieved from the five source databases. The standardized drug terms showed in the table were the ones that comprised the greater number of original terms. The majority of the terms standardized as Clopidogrel, Olanzapine and Zoledronic Acid were formed by the name of the drug and the pharmaceutical company that distributed it.

The majority of standardized drug names corresponded to one or two of the original terms retrieved from the five source databases. However, some of them such as Clopidogrel, Olanzapine and Zoledronic Acid encompassed many original terms (see Figure 5.2).

The major part of the terms standardized as Clopidogrel, Olanzapine and Zoledronic Acid were actually formed by the name of the compound followed by the name of the pharmaceutical company that produced it:

- **Clopidogrel:**

  Clopidogrel Acino, Clopidogrel Apotex, Clopidogrel BGR, Clopidogrel, Clopidogrel HCS, Clopidogrel Krka, Clopidogrel Krka d.d., Clopidogrel Mylan, Clopidogrel Ratiopharm, Clopidogrel Ratiopharm GmbH, Clopidogrel TAD, Clopidogrel Teva, Clopidogrel Teva Pharma, Clopidogrel Zentiva, Grepid, Iscover, Plavix, Zyllt.

- **Olanzapine:**

  Olanzapine Apotex, Olanzapine Cipla, Olanzapine Glenmark Europe, Olanzapine Glenmark, Olanzapine Mylan, Olanzapine, Olanzapine Teva, Olazax Disperzi, Olazax, Zalasta, Zypadhera, Zyprexa, Zyprexa Velotab.

- **Zoledronic Acid:**

  Aclasta, Zoledronate, Zoledronic Acid Accord, Zoledronic Acid Actavis, Zoledronic Acid Hospira, Zoledronic Acid Medac, Zoledronic Acid Mylan, Zoledronic Acid Teva Pharma, Zoledronic Acid Teva, Zoledronic Acid, Zometa.

Thus, if we had to standardize more drug terms in the future, it would be a good idea to create a dictionary of pharmaceutical companies. Using this dictionary, we would match the terms with such terminations and keep only the unmatched substrings, reducing the number of drug names that would have to be manually annotated.

In addition, we associated each LLT to a PT term according to MedDRA hierarchy and standardized the LLTs (See Chapter 4, Section 4.2.2). Initially, we had 11,319 different LLTs and, after the standardization process, we obtained 7,215 unique LLTs that were grouped into 5,147 PTs.

5.2 Statistics of the final table of drug-variant-ADR associations

After we had merged all five databases by their common drug names and ADRs, had subset the entries with information about PanDrugsdb medicines and had removed the duplicates, we obtained a final table of drug-variant-ADR associations. This table contained information about the genetic variants that conferred more susceptibility to suffer ADRs in response to the treatment with a specific drug. Moreover, for some entries, the table included the severities and/or frequencies of the ADRs and, in the majority of cases, a level of evidence of the pharmacogenomics association. Last, all entries had information about the sources from where the drug-ADR association was retrieved.

The final table contained 821 different genes and 2,294 unique variants associated with 2,017 distinct ADRs at the LLT level. These LLTs were included in 2,003 non-identical MedDRA PTs. As Figure 5.3 shows, the majority of variants in this final table were SNPs (82.82%) and the rest corresponded to haplotypes (groups of SNPs that are inherited together). Moreover, the majority of genes where those variants occurred encoded for enzymes that metabolized xenobiotics, drugs and steroids (Figure 5.4).

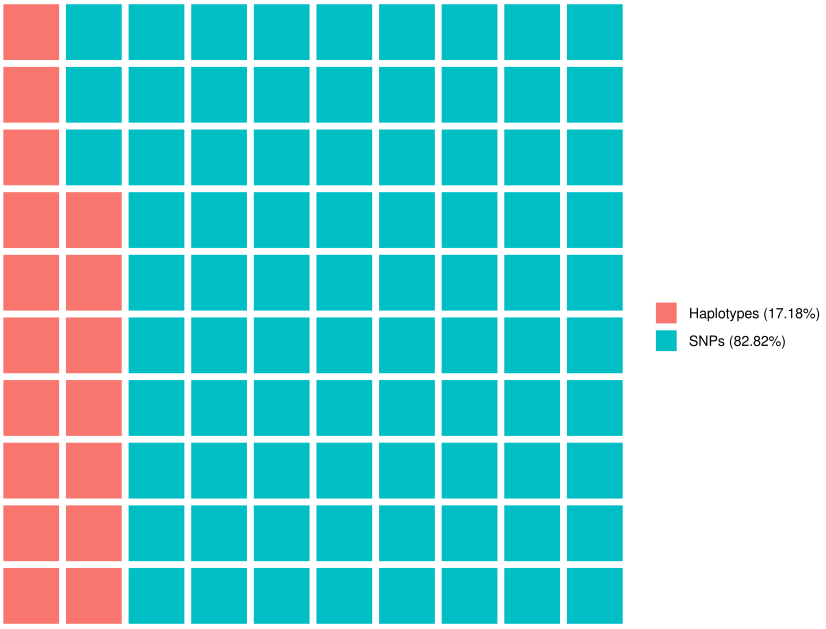


Figure 5.3: Types of variants in our data

82.82% of the variants were SNPs and the other 17.18% haplotypes. Haplotypes are defined as a group of SNPs that are inherited together.

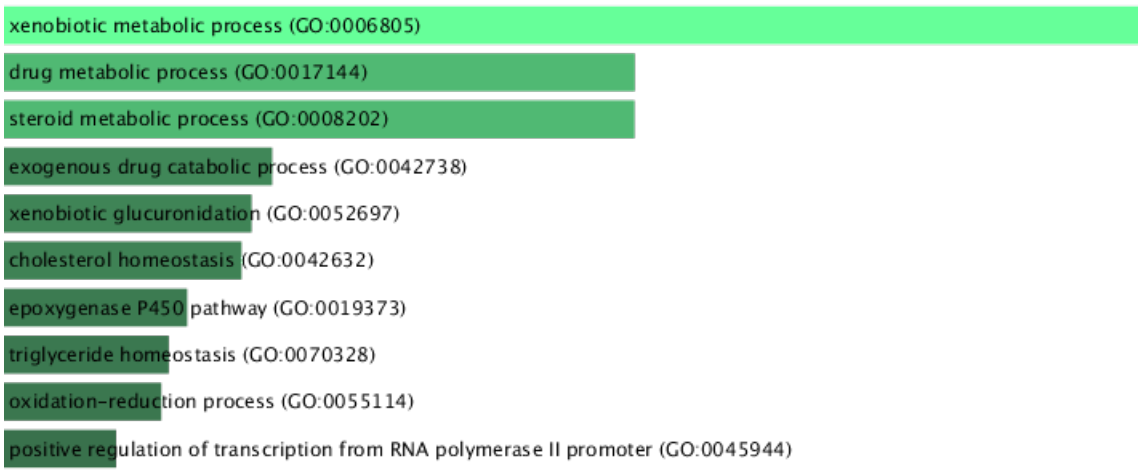


Figure 5.4: Biological processes overrepresented in our data

The majority of genes in our dataset participated in the metabolism of xenobiotics, drugs and steroids. This ranking of Gene Ontology (GO) biological processes was obtained using Enrichr and ordered by decreasing p-value after performing a Fisher's exact test.

In total, this table reported 64,027 drug-variant-ADR associations that comprehended 495 PanDrugs medicines, meaning that we had germline-associated pharmacogenomics data for 5.44% of the drugs included in PanDrugsdb. As Figure 5.5 shows, the majority of these drugs were rhodopsin family antagonists (6.56%) or agonists (3.79%) or inhibitors of oxidoreductases (5.45%). Rhodopsin family members are GPCRs (G Protein-Coupled Receptors) of class A. GPCRs are proteins with seven transmembrane domains that transduce extracellular signals into the cell and mediate many physiological processes. GPCR superfamily is divided into six classes: A, B, C, D, E and F, being class A the largest and best studied. Rhodopsin-like GPCRs play a major role in tumor biology, and that is the reason why many drugs target mutated forms

of these receptors [37].

Oxidoreductases are enzymes that catalyze the transference of electrons from one molecule to another. They play a crucial role in redox homeostasis, protecting cells from oxidative damage. Moreover, they are a key part of the electron transport chain and are involved in the energy metabolism. The enzyme NQO1 (NAD(P)H:Quinone Dehydrogenase 1) appears mutated in several types of cancer and inhibitors of oxidoreductase activity are used in order to counteract its upregulation [38].

Figure 5.5: Main drug families in our data

The majority of drugs in our final dataset were rhodopsin family antagonists (6.56%) or agonists (3.79%) or inhibitors of oxidoreductases (5.45%).

We checked the number of different LLTs within the same PT level (Figure 5.6). Most PTs only comprised one LLT, which corresponded to the PT level itself (See Chaper 3, Section 3.1.1 for more information). Only 61 PTs out of the 2,003 that formed this table included more than one LLT term. The PT with higher number of LLTs was DRUG DEPENDENCE. DRUG DEPENDENCE's LLTs were DRUG DEPENDENCE, OPIOID TYPE DEPENDENCE, COCAINE DEPENDENCE and HEROIN ADDICTION. These terms gave further and more precise information about the PT level they belonged to. Thus, we concluded that the standardization process was satisfactory.

Next, we assessed the number of ADRs reported for each drug. In Figure 5.7, the ADRs are expressed as LLTs (a) or PTs (b). As it was expected, the shape and mean values were virtually the same for the two logarithmic distributions. A drug caused, on average, around 35 different ADRs in presence of a specific variant. Nevertheless, many drugs had more than 100 reported and unique ADRs. The drugs with the highest number of associated adverse events were bortezomib, nilotinib and dasatinib. These three compounds are small molecule inhibitors used to treat white blood cancers. They enter cancer cells an provoke their death by interfering with cellular processes. All of them have been proved to be very effective for cancer treatment and, consequently, have DScores equal to 1 for most target genes. However, as other cancer drugs, they are very aggressive and can cause a large number of ADRs.

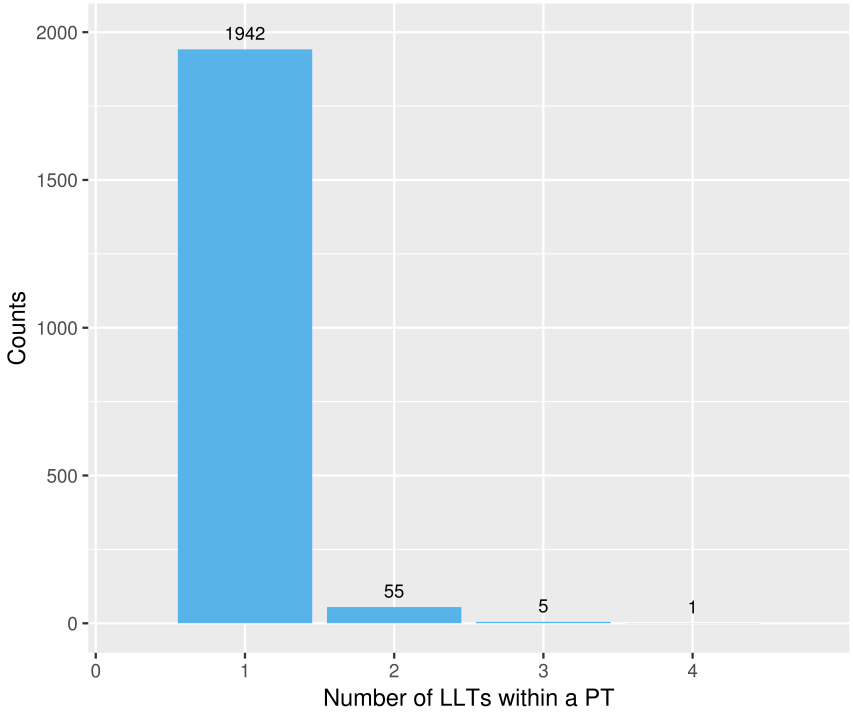
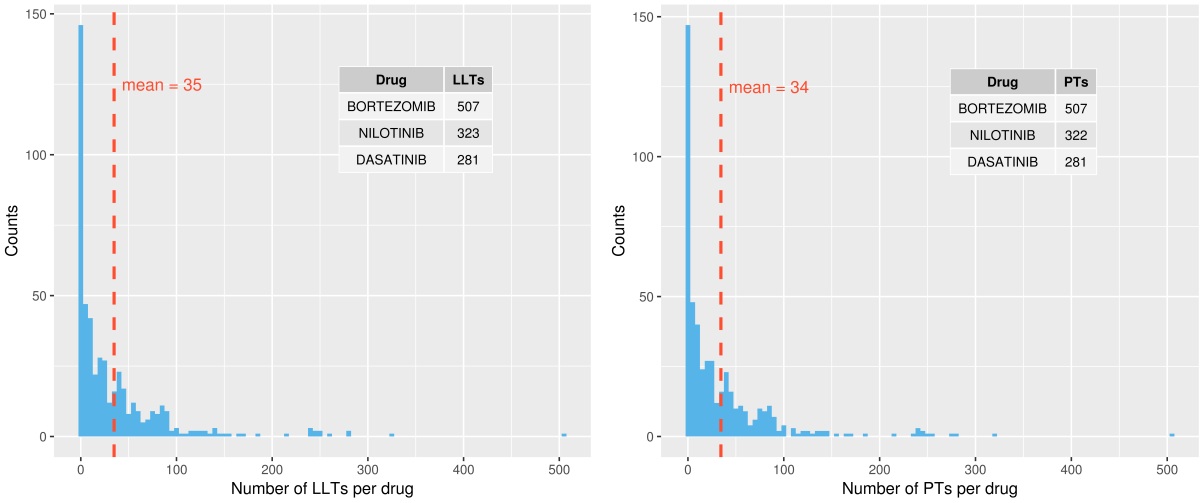


Figure 5.6: Number of LLTs per PT

Logarithmic distribution of the number of LLTs contained within the same PT. The majority of PTs comprised a single LLT. The PT with maximum number of LLTs was DRUG DEPENDENCE.



(a) Number of LLTs per drug

(b) Number of PTs per drug

Figure 5.7: Number of ADRs per drug

Logarithmic distribution of the number of ADRs per drug. A drug caused, on average, around 35 different ADRs in presence of a specific variant. Bortezomib, nilotinib and dasatinib were the medicines with the highest number of adverse drug reactions.

Regarding the pharmacogenomics variants and genes, the distribution was also logarithmic (see Figure 5.8). The mean number of pharmacogenomics variants found within a single gene was 3, although there were genes that clearly surpassed this value. The genes with more reported pharmacogenomics variants were *CYP2D6* and *CYP2C9*. These two genes codify for proteins

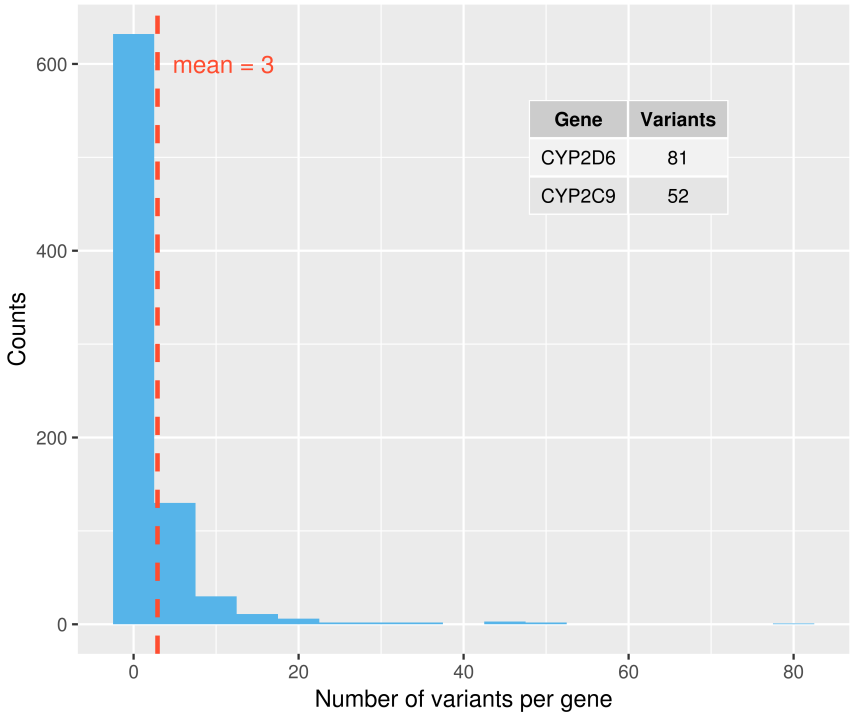


Figure 5.8: Number of pharmacogenomics variants per gene

Logarithmic distribution of the number of pharmacogenomics variants per gene. On average, a gene contained three variants inside its sequence. Nevertheless, cytochromes had an increased variability.

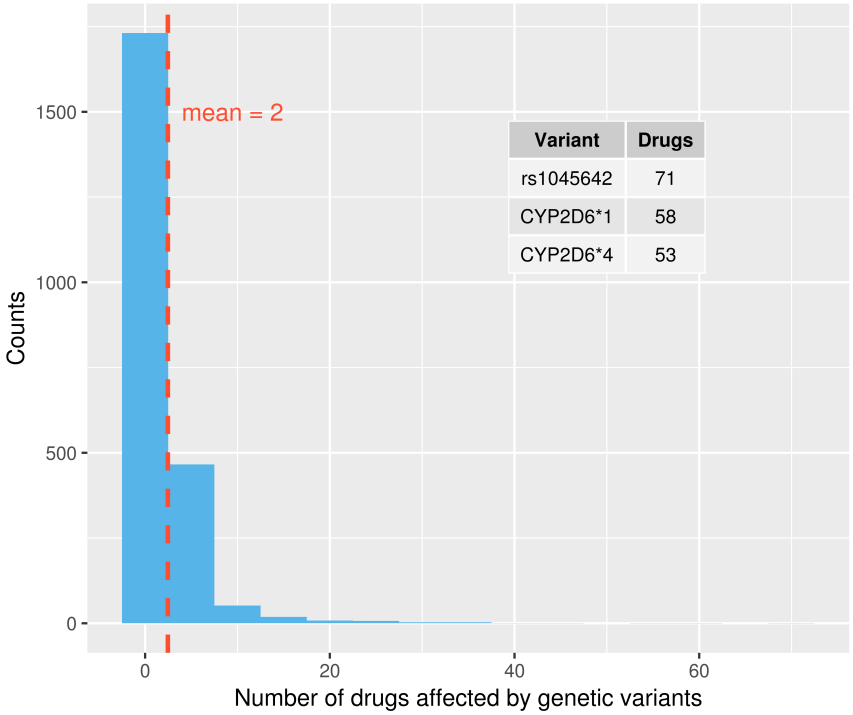


Figure 5.9: Number of drugs affected by a single genetic variant

The mean value of the logarithmic distribution of pharmacogenomics variants per drug was equal to 2. The variants that affected to most drug responses were found in *CYP2D6* and *ABCB1* genes.

that belong to the cytochrome 2 (CYP2) family. This protein family is crucial in the Phase I of drug metabolism and has many polymorphisms that affect cytochrome enzymatic activity.

Finally, we analyzed the number of drugs affected by the same pharmacogenomics variant (Figure 5.9). The majority of genetic variants caused ADRs for few drugs. On average, each variant affected two drugs. However, some variants of the *CYP2D6* gene affected a larger number of compounds. Moreover, we found that SNP rs1045642, which is located in the *ABCB1* gene, also influenced the responses of many drugs. *ABCB1* gene encodes for an ATP Binding Cassette (ABC) transporter that is involved in the Phase III of drug metabolism and pumps chemicals out of the cell. Its overexpression is a mechanism of drug resistance in cancer cells.

All the histograms shown above followed a logarithmic distribution, which reflected the biological nature of these data.

5.3 Parameters used in ToxScore computation

ToxScores were computed using four different parameters: the overlapping of drug-ADR information between the source databases, the evidence of each drug-variant-ADR association and the severity and frequency of the ADRs caused by a particular drug. Each ToxScore represented the noxiousness of a drug in presence of a specific genetic variant. For more information about ToxScore computation, please refer to Chapter 4, Section 4.2.5.

5.3.1 Overlapping between the source databases

Figure 5.10 shows the overlapping of drug-ADR associations found in the five different databases. As PharmGKB was our only source of pharmacogenomics data, all the entries in our dataset matched PharmGKB information. This was also the reason why PharmGKB was the larger set, followed by SIDER, IntSide, PROTECT and SPL-ADR-200db.

Although PharmGKB was our main source of data and many entries were only found in this database (Figure 5.10, bar 1), there was a partial overlapping of drug-ADR associations between PharmGKB and the other databases. SIDER shared many common entries with Intside (bar 2), PROTECT (bar 6) or both (bar 4). In fact, the majority of IntSide entries overlapped with SIDER's ones and only 101 drug-ADR associations were found in other databases (75 found exclusively in PharmGKB and 26 matching both PROTECT and PharmGKB). This was due to the fact that IntSide was created using SIDER as a main data source. Moreover, a large proportion of PharmGKB entries was found only in SIDER (bar 3) or in PROTECT (bar 5). Regarding SPL-ADR-200db, data was mostly shared with PROTECT and SIDER and only 3 entries were exclusively found in PharmGKB. Finally, 27 drug-ADR associations were reported in all databases.

The overlapping between the databases was used as a measure of the proof of drug-ADR associations. In order to compute the ToxScore, we gave the maximum overlapping value to those entries reported in all databases and the minimum value to those found only in PharmGKB. Thus, looking at this plot we knew that 5,219 entries would have the minimum value of overlapping and 27 would have an overlapping of 1.

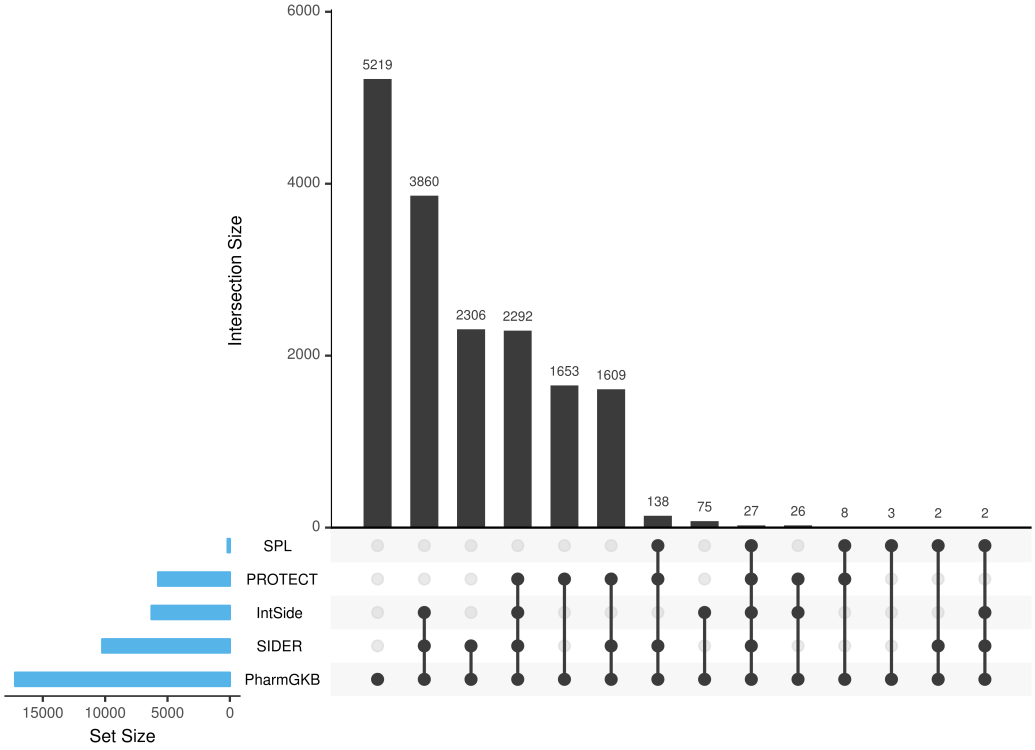


Figure 5.10: Overlapping of drug-ADR information between the source databases

PharmGKB was our only source of pharmacogenomics data, so all entries belonged to PharmGKB set. Associations found only in PharmGKB were given the minimum overlapping value. PharmGKB information partially overlapped with other databases. Aside from PharmGKB, the other four databases also shared information with each other. Since SIDER was a main source of data of IntSide database, the majority of IntSide entries were found in SIDER. SPL-ADR-200db data was also contained in SIDER and PROTECT. Only 3 SPL-ADR-200db entries were exclusively found in PharmGKB. 27 drug-ADR associations were reported in all databases and were assigned the highest overlapping value. Intersections with 0 size were dropped from the plot.

5.3.2 Pharmacogenomics evidence, frequency and severity

The other three parameters corresponded to pharmacological and pharmacogenomics data retrieved from the source databases. Some of this information was expressed as discrete values, so in order to compute the ToxScore we had to convert it to numeric format (see Chapter 4, Section 4.2.5).

The pharmacogenomics evidences of drug-variant-ADR associations were retrieved directly from PharmGKB. This variable took values between 1 and 4, being associations with an evidence of 1 the most reliable. We conserved these levels in order to compute the ToxScore. Thus, we considered that associations with a higher level of evidence contributed more to the noxiousness of a drug in presence of a specific variant. Levels 1 and 2 in PharmGKB were subdivided into A and B groups. These subgroups provided additional information about whether these associations were used in clinics or were particularly well documented. However, A and B subgroups did not indicate higher or lower evidence within the same level. Thus, we considered levels 1A and 1B and 2A and 2B equivalent. Further information about evidence levels can be found in PharmGKB database explanation (Chapter 3, Section 3.3.5).

As Figure 5.11 shows, an 87.88% of the entries in our final dataset had available pharmacogenomics evidence data. The majority of these entries belonged to levels 3 and 4 and only a 4.04% of the associations (3.95% in 1A and 0.09% in 1B) had the highest level of pharmacogenomics evidence.

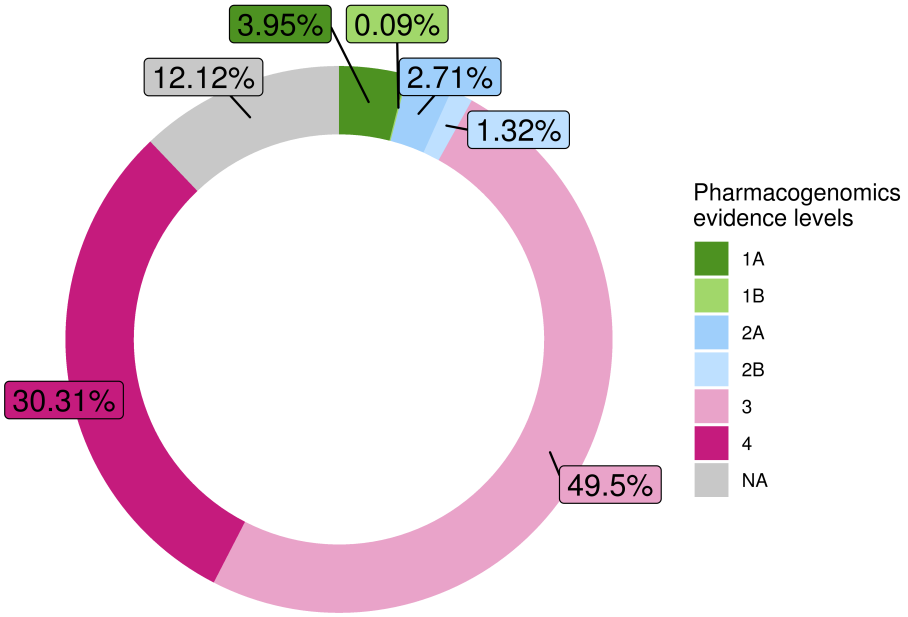


Figure 5.11: Proportion of pharmacogenomics evidence levels in our dataset

Pharmacogenomics evidence levels from PharmGKB were ordered from higher to lower as 1A/1B, 2A/2B, 3 and 4. Only a 12.12% of the entries did not have available data for this magnitude. The majority of drug-variant-ADR associations had a low level of evidence.

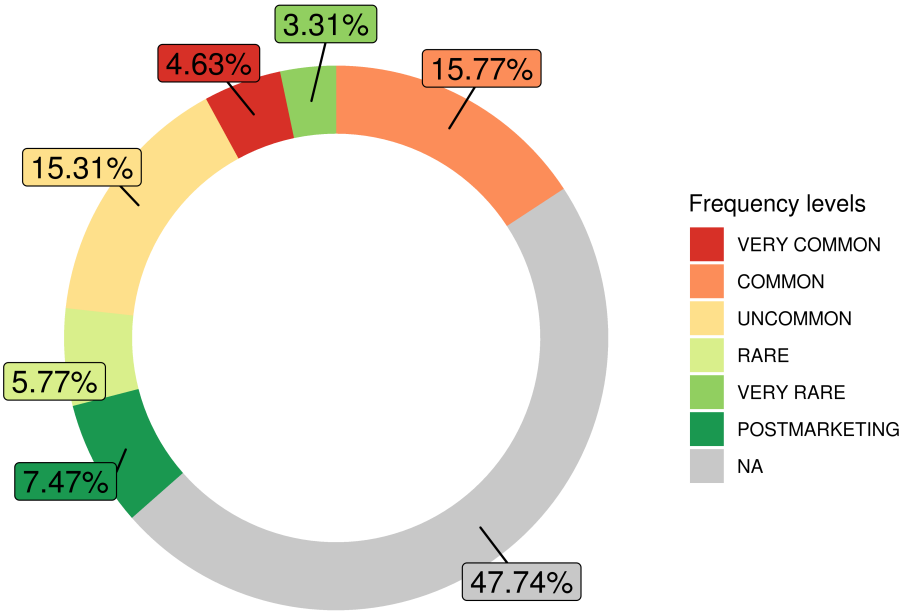


Figure 5.12: Proportion of frequency levels in our dataset

More than half of the data had available frequency data. The most common frequencies were COMMON and UNCOMMON. VERY RARE and VERY COMMON levels were the ones with fewer data.

Severity data was obtained from SPL-ADR-200db and frequencies were downloaded from SIDER and PROTECT. It is important to note that an ADR's severity is independent of the drug that causes it. However, the frequency of the same ADR varies between different chemicals.

Frequencies were expressed either as percentages (in SIDER) or discrete values (in SIDER and PROTECT). These discrete values corresponded to ranges defined by CIOMS. In order to being able to compare the data, we decided to express all frequencies as discrete values. For more information, please refer to Chapter 4, Section 4.2.5). We obtained six frequency levels, ordered from most frequent to less frequent as VERY COMMON, COMMON, UNCOMMON, RARE, VERY RARE and POSTMARKETING. As Figure 5.12 shows, more than half of the data had available frequency information. Among these entries, the major part of ADR frequencies were COMMON and UNCOMMON. Moreover 7.47% of the entries had POSTMARKETING frequencies reported after clinical trials. VERY RARE and VERY COMMON levels were the ones with fewer data.

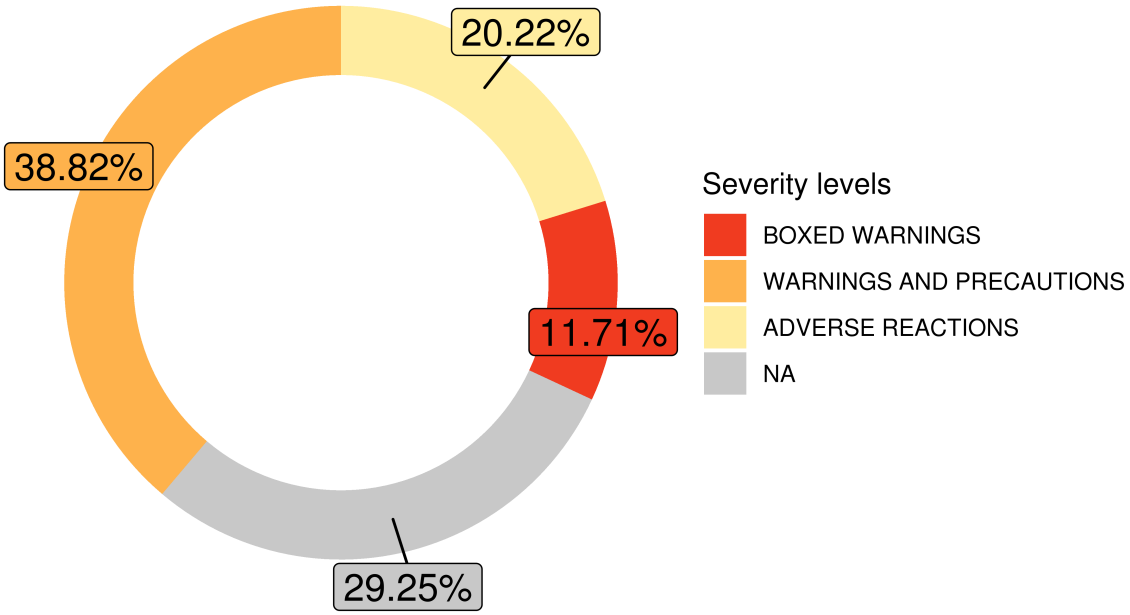


Figure 5.13: Proportion of severity levels in our dataset

Severity levels from SPL-ADR-200db were ordered from higher to lower as BOXED WARNINGS, WARNINGS AND PRECAUTIONS and ADVERSE REACTIONS. A 29.25% of the entries did not have available data for this magnitude. The majority of drug ADRs had an intermediate level of severity and 11.71% were very severe or even life-threatening.

The levels of severity were named after the section of the SPL from where the ADR was retrieved (see Chapter 4, Section 4.2.4). These levels were order from higher to lower as BOXED WARNINGS, WARNINGS AND PRECAUTIONS and ADVERSE REACTIONS. For ToxScore computation, we considered drugs with more severe ADRs more noxious. As Figure 5.13 shows, a 70.75 % of the entries had associated severity information, being the second level of severity the

most abundant. BOXED WARNINGS section includes the most serious or even life-threatening adverse effects, so it is remarkable that an 11.71% of the drug-variant-ADR associations had this value of severity.

Finally, we assessed the number of drug-variant-ADR associations that had available data for all the parameters, a part of them or only one. Note that all entries had at least a value for overlapping magnitude, so none of the drug-variant associations had a ToxScore of 0. In Figure 5.14 we represented a Venn Diagram of the number of entries that had available data for any of the parameters.

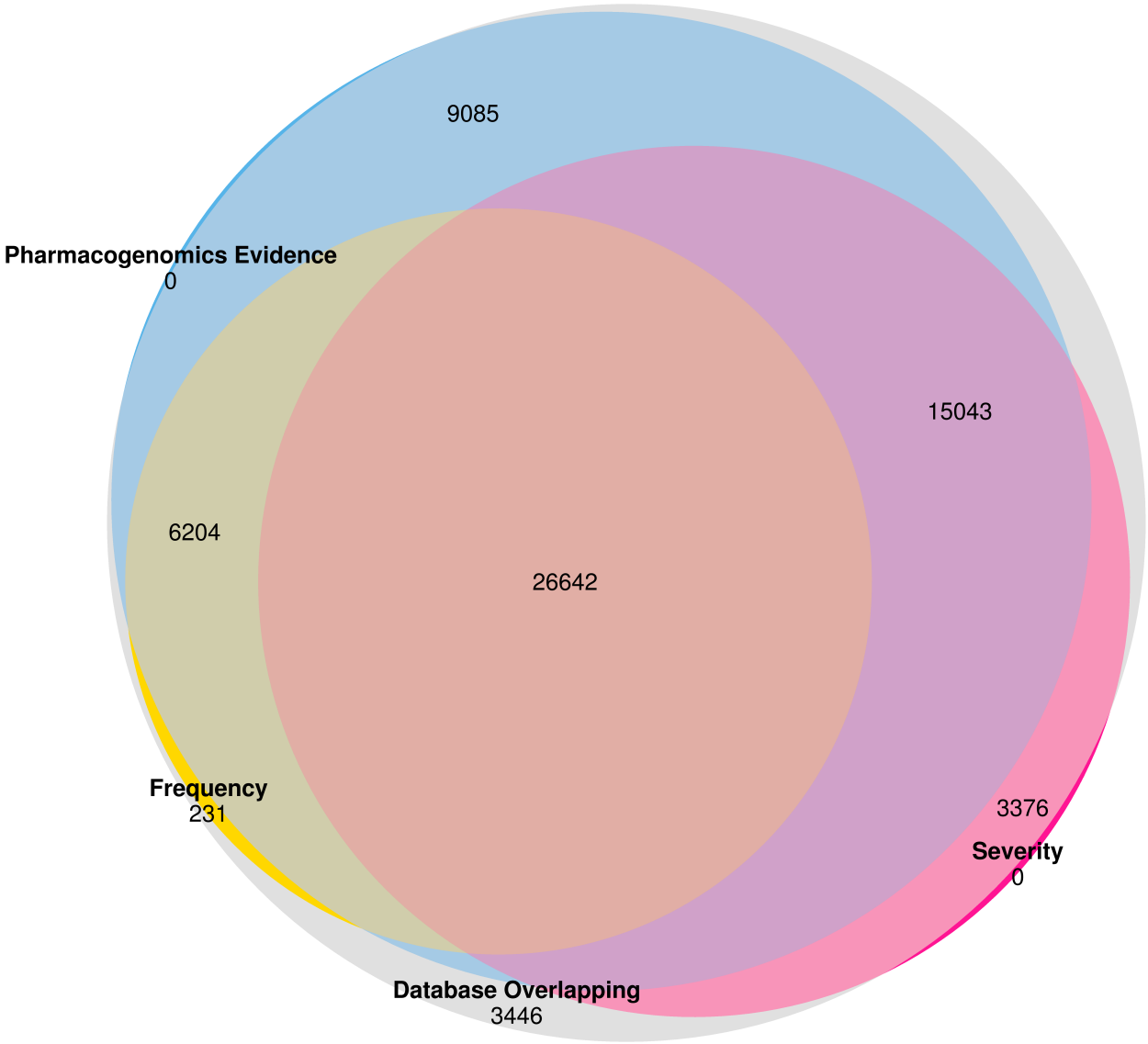


Figure 5.14: Available data for each ToxScore parameter

Overlapping set is colored in grey, pharmacogenomics evidence in blue, severity in pink and frequency in yellow. Database overlapping was the biggest set, followed by pharmacogenomics evidence, then severity and finally frequency. This result was consistent with the ones shown in the doughnut plots. 26,642 entries in our dataset had information available for all four parameters. 21,247 associations (15,043 + 6,204) had values for three parameters. 3,376, 9,085 and 231 entries had information for two parameters: overlapping and severity, pharmacogenomics evidence or frequency respectively. Finally, 3,446 associations had only overlapping data available. All entries in our dataset had values for overlapping parameter.

This Venn Diagram was consistent with the results reported in the previous doughnut plots: the lower the percentage of NAs the bigger the set. Aside from overlapping magnitude, the larger number of available data corresponded to pharmacogenomics evidence set, followed by ADR severity and frequency. We noted that the four sets had many common entries. In fact, 26,642 entries (41.61%) contained information for all four parameters and 21,247 (33.18%) drug-variant-ADR associations had information for three of them. 9,085 (14,19%), 3,376 (5,27%) and 231 (0.36%) entries contained information for overlapping and pharmacogenomics evidence, severity or frequency respectively. Finally, 3,446 (5.38%) associations had only overlapping data available.

5.4 ToxScore computation

Then, we computed 5,651 ToxScores for each drug-variant pair in our dataset (see Equation 4.1). This score ranged from 0.02 to 1 and indicated the noxiousness of a drug in presence of a specific genetic variant. ToxScores equal to 0.02 corresponded to entries found only in PharmGKB and with no pharmacogenomics evidence, severity or frequency data available. As there were many NA values for these three parameters, we computed a Reliability Level to accompany the ToxScores (see Equation 4.2). Reliability Levels could be equal to A, B or C and indicated the amount of missing data. ToxScores with a reliability level of A were calculated missing few data while ToxScores of type C were computed using only the overlapping parameter, thus they were considered unreliable. A portion of the table of drug-variant ToxScores is shown below.

Table 5.1: ToxScores

DRUG	VARIANT	ToxScore	LEVEL OF RELIABILITY
RAPAMYCIN	CYP3A4*22	0.45	A
PROPOFOL	rs58597806	0.26	B
SEVOFLURANE	rs193922876	0.32	C
DESFLURANE	rs193922772	0.32	C
METOPROLOL	CYP2D6*41	0.81	A
PLATINUM	rs11615	0.3	B

We drew a boxplot of the ToxScores grouped according to their Reliability Levels (Figure 5.15). 873 ToxScores (15.45%) belonged to group A, 2,406 (42.58%) had a Reliability Level of B and 2,372 (41.97%) were unreliable. Regarding the distributions of the three Reliability Levels, we noted several differences. First, the mean and median values were different. This was expected because, due to equations 4.1 and 4.2, the lower the number of available data the lower the ToxScore and Reliability values. Second, data dispersion in A level was greater than in B, being C the group with the lowest variability. Moreover, level C had the highest number of outliers. Finally, the distributions of the three Reliability Levels partially overlapped and their shapes were different. While A level's distribution was nearly normal, distributions of B and C groups were left and right skewed respectively. This skewness was very marked for the later group.

Based on these results we resolved to exclude ToxScores with a Reliability Level of C and only take in partial consideration those with a Reliability of B. The ToxScores in A category, in addition to be the most reliable ones, had the wider distribution. Their values ranged approximately between 0.25 and 0.75, with some ToxScores above 0.8.

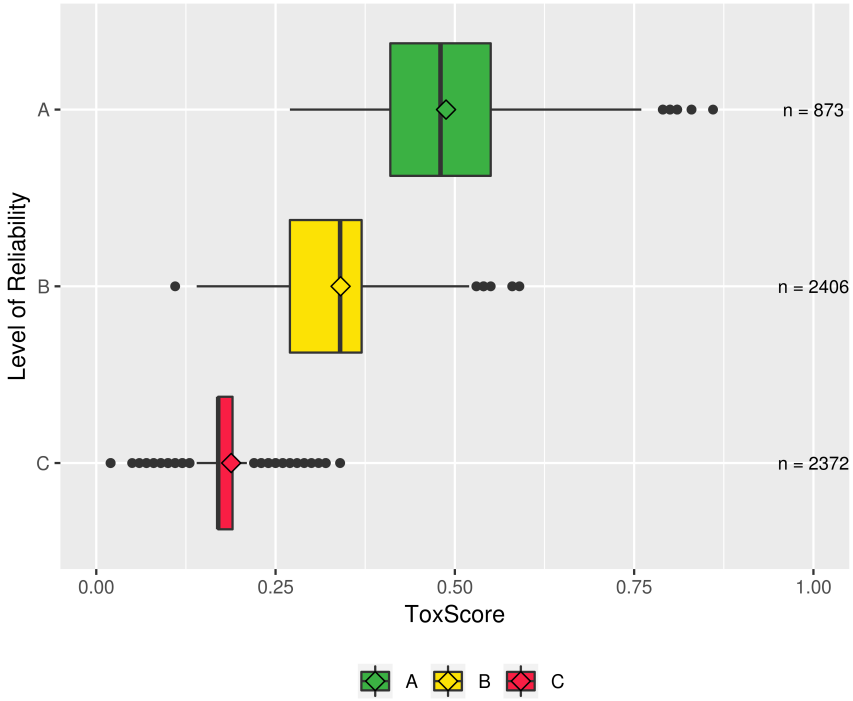


Figure 5.15: Boxplots of ToxScores grouped by Reliability Level

873, 2406 and 2372 ToxScores were assigned to Reliatibity Levels A, B and C respectively. Group A had the highest variability and was normally distributed. B an C distributions were left and right skewed, respectively. C level was the one with most outliers. The lower the ToxScore value, the lower its Reliatibity Level.

5.5 ToxScore comparison among drugs with the same target or within the same family of compounds

Next, we checked whether there were differences among the toxicity levels of the drugs with a common target. From PanDrugsdb (See Table 3.6) we subset all drugs directed against *EGFR* (Epidermal Growth Factor Receptor) gene that had a computed ToxScore. We chose this gene because it appears mutated in different types of cancer and is targeted by a large amount of compounds. Its product plays a key role in cancer cell proliferation. EGFR is a membrane protein with kinase activity that is constitutively active in cancer cells. Its mutated form does not need to bind the ligand in order to initiate a signaling cascade that enables cancer cells to continuously divide.

Moreover, we also analyzed tyrosine kinase receptor inhibitors, one major drug family among medicines that target *EGFR* (see Figure 5.16). Tyrosine kinase receptor inhibitors target the activated form of EGFR, counteracting the proliferation signal.

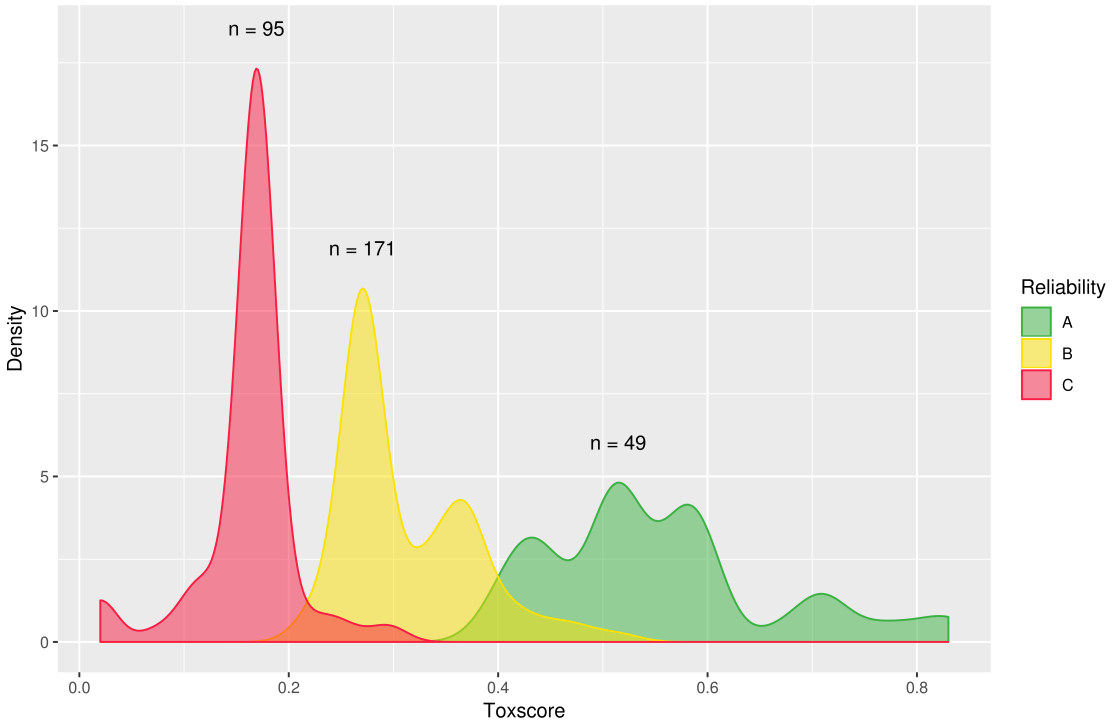
Figure 5.16: Drug families targeting *EGFR*

Tyrosine kinase receptor inhibitors (in red) were one of the most abundant drug families among medicines targeting *EGFR* (8.21%).

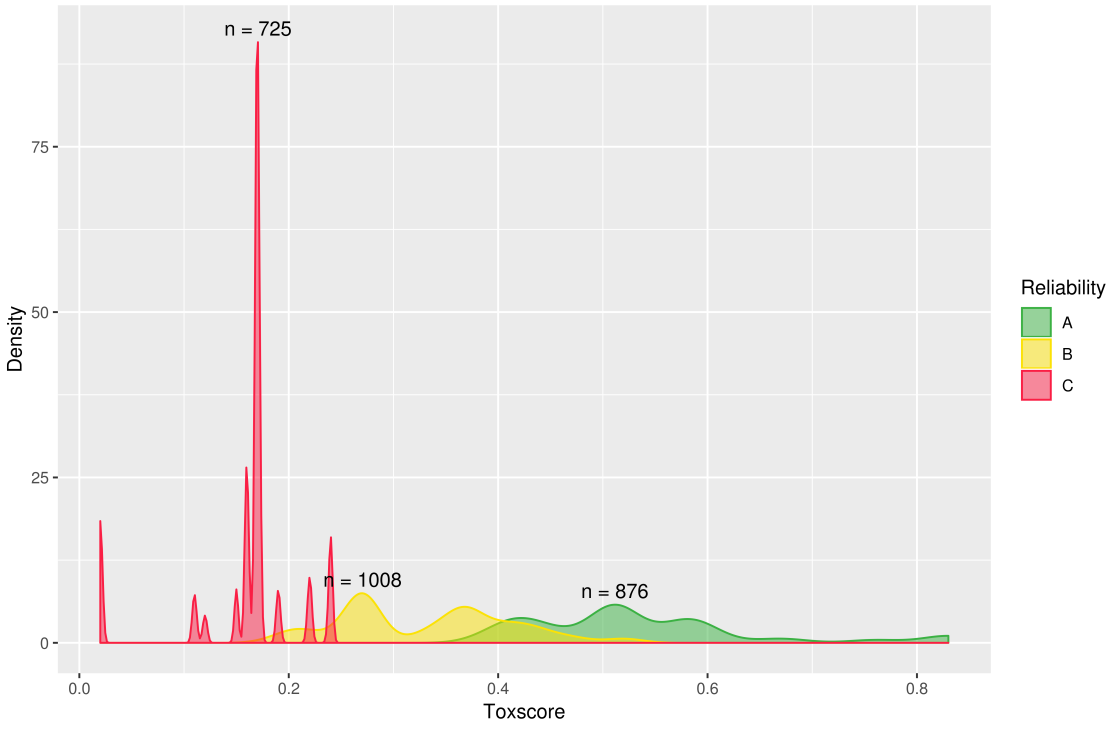
ToxScore distributions of the drugs targeting *EGFR* gene are shown in Figure 5.17 (a). From this distribution plot we concluded that the drugs with a common target had different ToxScores. Moreover, we could corroborate that ToxScore values also varied between the members of the same drug family (Figure 5.17 (b)).

The two distribution plots were consistent with the boxplot displayed in previous section. In both types of plots, the majority of ToxScores belonged to level B and the minority to level A. However, none of the distributions were normal. In fact, all of them had many nodes, specially the distributions of level C within the same family of compounds. Modes indicate values were a high number of points concentrate. These peaks were due to the fact that ToxScore parameters had few possible values. The lower the number of available data the lower the number of combinations and thus, the lower the number of possible ToxScore values. This explains that the distributions had more marked peaks as the Reliability Level decreased.

From these distribution plots, we concluded that the drugs with a common target or within the same family of compounds had different ToxScores. This implied that, for the same therapeutic target, there were drugs that were less harmful for a patient with a specific set of germinal variants. Thus, it would be preferable to prioritize these drugs in case that the efficacy against the target (represented by the DScore) was the same.



(a) *EGFR*



(b) Tyrosine kinase receptor inhibitors

Figure 5.17: ToxScore distributions of drugs directed against the same gene or belonging to the same compound family

The majority of ToxScores had a Reliability Level of C. ToxScores of class A had a wider range of possible values but did not follow a normal distribution. The lower the Reliability Level, the lower the number of possible ToxScore values and the more marked the peaks in the density distribution.

5.6 Reordering PanDrugs compound ranking according to germinal variants with known ToxScores

We determined whether the ranking of drugs returned by PanDrugs was modified when taking into account the ToxScores. In order to do so, we used Whole Exome Sequencing (WES) data from a paraganglioma patient. Paraganglioma is a rare type of neuroendocrine cancer in which 30% of the tumors are hereditary. The majority of these hereditary tumors have mutations in *REF* proto-oncogene, von Hippel-Lindau (*VHL*) tumor suppressor or genes that codify for Succinate Dehydrogenase (SDH) subunits. This patient data was lent by Paradifference Foundation, a non-profit organization whose aim is to finance paraganglioma research in order to find a cure for this illness. Unfortunatelly, paragangliomas are very rare and there is still little knowledge about the molecular mechanisms that influence this type of cancer. As a consequence, the rankings reported target genes that had low GScores. Still, we wanted to evaluate the effect of the germinal component in PanDrugs ranking.

First, variant calling was performed in order to detect somatic and germline mutations. Next, we input PanDrugs the somatic variations of the primary tumor and the metastasis, obtaining two rankings of therapeutic candidates. Then, we reordered PanDrugs output using the precomputed ToxScores for the drugs in the ranking and the germinal variants of the patient. These rankings are reported in Tables 5.2 and 5.3 (primary tumor) and Tables 5.4 and 5.5 (metastasis).

PanDrugs rankings were sorted by decreasing DScore and GScore. Reordered rankings were sorted first by decreasing DScore and Reliability Level, then by increasing ToxScore and finally by decreasing GScore. Drugs with the same DScore and without reported ToxScore were prioritized over drugs with evidence of toxicity for that patient. ToxScores with lower Reliability Levels had lower values, but these lower scores did not mean that the drugs were less toxic. That is why, in order to avoid confusion, we ranked the drugs with the same DScore but with ToxScores of type C last. Moreover, the patient had several variants that conferred susceptibility to ADRs in response to the same drug. In those cases, we kept the highest ToxScore with the highest Reliability Level for each drug.

Considering only somatic variants, the best available therapeutic candidates to treat the primary tumor were tamoxifen and everolimus. Nevertheless, when we reordered the ranking using the ToxScores, we detected that this patient had a germinal variant (rs1045642, AG genotype) that could cause ADRs in response to these drugs. This result is a good example of how the same variant can influence different drug responses. In fact, rs1045642 is a polymorphism of *ABCB1* gene and was reported to be the variant with most drug associations in our dataset (see Figure 5.9). rs1045642 variant has been associated with a higher risk of adverse events in breast cancer patients treated with tamoxifen and everolimus. Although this patient suffered from another type of tumor, he or she had the same germinal variant reported in breast cancer patients, so it was probable that this person would also experience drug toxicity in response to these two drugs.

Tamoxifen and everolimus had ToxScores with a Reliability of A, but tamoxifen's score was higher. This result indicated that everolimus would cause less toxic ADRs than tamoxifen. Nevertheless, tamoxifen's DScore and its targets' GScore were the highest. For this reason, tamoxifen did not lose its first position in the reordered ranking. Moreover, we noticed that when we took into account the ToxScores, some drugs in the ranking changed their positions. Dabrafenib and sorafenib were switched, as well as cobimetinib and temsirolimus; midostaurin and thalidomide and tretinoin, docetaxel and etoposide. Drugs were penalized in the ranking for having lower Reliability Level and higher ToxScore than other drugs with the same DScore. Although all drugs in the ranking had a DScore over the recommended threshold for therapeutic benefit (DScore > 0.6), none of their target genes had enough influence in paraganglioma, as

reported by the low GScores.

Given these results, we would recommend tamoxifen for compassionate treatment of the primary tumor of this patient. Although it was the drug with higher DScore, it targeted genes with a low GScore (0.4598), so this treatment won't be very effective.

Regarding the metastasis result considering only the somatic mutations, few drugs that targeted genes with very low GScores were reported. rs1045642 also appeared associated to etopoxide, but with an unreliable ToxScore. Consequently, etoposide was penalized in the ranking for having a lower Reliability Level than other drugs with the same DScore. Moreover, docletaxel was the first drug reported in PanDrugs ranking. Since DScores do not vary for the same drug and the rankings were ordered by this field, the fact that docletaxel had a higher position in Tables 5.4 and 5.5 implied that drugs reported for treating metastasis were less effective than drugs in the ranking of the primary tumor. Moreover, docletaxel's response was influenced by rs7311358 (AA genotype), a variant in *SLCO1B1* gene, a member of SLC (Solute Carrier) family. This gene encodes for OATP1B1 (Organic Anion Transporting Polypeptide 1B1). This protein transports toxins and drugs from blood into liver cells during the Phase 0 of drug metabolism.

In metastasis ranking, the first four drugs were reordered when using the ToxScore information. The patient did not had any variant that conferred susceptibility to ADRs in response to topotecan and thus, this drug was prioritized over the other three compounds with the same DScore. However, none of the drug in the ranking targeted any gene product with a high GScore. Thus, we would not recommend any of these medicines to treat the metastasis of this patient.

In conclusion, these results showed that our ToxScore was able to detect drugs that may have toxic effects due to patient's germinal variants. These variants were reported aside of the drug name, so the user could consult the literature in order to obtain more information. Moreover, in case that several drugs had the same DScores but different ToxScores, the ranking penalized the drugs with more toxic effect. This way, when we reordered the list of therapeutic candidates to treat metastasis, the order of the first drugs in the ranking was altered. The new ranking prioritized topotecan, a drug with the same efficacy and molecular target than other reported compounds but less toxic for the patient. Nevertheless, the first position in the primary tumor ranking still corresponded to tamoxifen, a drug that might cause toxic effects to the patient. This was due to the fact that tamoxifen was the most effective drug, as it was reflected by its DScore.

In view of these rankings, we would not recommend any drug to treat the metastasis whereas tamoxifen may be suitable for compassionate treatment of the primary tumor. However, as reported by the ToxScore, rs1045642 variant might increase tamoxifen's toxic effects, so the patient should be monitored in order to minimize the occurrence of ADRs.

Table 5.2: PanDrugs ranking results of somatic mutations obtained from primary tumor versus control samples

RANK	GENE(s)	DRUG	DScore	GScore
1	ALPK2|GGT1|MUC16|RICTOR	TAMOXIFEN	0.941	0.4598
2	GNAQ|PBRM1|RICTOR	EVEROLIMUS	0.94	0.1388
3	COL1A1|LILRB3|PIK3R2	COPANLISIB	0.934	0.356
4	ARID4B|NFATC1	LENALIDOMIDE	0.922	0.3329
5	GNAQ	TRAMETINIB	0.92	0.0625
6	GNA12|GNAQ	SORAFENIB	0.92	0.0625
7	GNA12|GNAQ	DABRAFENIB	0.92	0.0625
8	GNA12|GNAQ	REGORAFENIB	0.916	0.0625
9	COL1A1|GNAQ	BOSUTINIB	0.913	0.1145
10	GNA12|GNAQ	SORAFENIB TOSYLATE	0.913	0.0625
11	RICTOR	TEMSIROLIMUS	0.91	0.1388
12	GNAQ	COBIMETINIB	0.91	0.0625
13	CASP9	NIRAPARIB	0.907	0.2537
14	IFNAR2	RUXOLITINIB	0.907	0.2022
15	CASP9	OLAPARIB	0.906	0.2537
16	CASP9	RUCAPARIB	0.904	0.2537
17	COL1A1	DASATINIB	0.904	0.1145
18	NFATC1	THALIDOMIDE	0.903	0.3329
19	RICTOR	MIDOSTAURIN	0.903	0.1388
20	GPC3	ARSENIC TRIOXIDE	0.902	0.125
21	NFATC1	POMALIDOMIDE	0.901	0.3329
22	COL1A1	PONATINIB	0.901	0.1145
23	ARID4B|MUC16|PGP	DOCETAXEL	0.832	0.2854
24	EIF4B|MUC16|PGP	ETOPOSIDE	0.832	0.2854
25	ARID4B|CEACAM1|TMEM40	TRETINOIN	0.832	0.1616

Red and green rows indicate changes in the positions between the two rankings. Drugs in red descend and drugs in green ascend after ranking reordering using the ToxScores.

Table 5.3: Re-ranking of PanDrugs results (primary tumor) using ToxScores

RANK	GENE(s)	DRUG	DScore	GScore	VAR	ToxScore	RELIAB
1	ALPK2|GGT1|MUC16|RICTOR	TAMOXIFEN	0.941	0.4598	rs1045642(A;G)	0.65	A
2	GNAQ|PBRM1|RICTOR	EVEROLIMUS	0.94	0.1388	rs1045642(A;G)	0.52	A
3	COL1A1|LILRB3|PIK3R2	COPANLISIB	0.934	0.356			
4	ARID4B|NFATC1	LENALIDOMIDE	0.922	0.3329			
5	GNAQ	TRAMETINIB	0.92	0.0625			
6	GNA12|GNAQ	DABRAFENIB	0.92	0.0625			
7	GNA12|GNAQ	SORAFENIB	0.92	0.0625	rs2306283(G;G)	0.83	A
8	GNA12|GNAQ	REGORAFENIB	0.916	0.0625			
9	COL1A1|GNAQ	BOSUTINIB	0.913	0.1145			
10	GNA12|GNAQ	SORAFENIB TOSYLATE	0.913	0.0625			
11	GNAQ	COBIMETINIB	0.91	0.0625			
12	RICTOR	TEMSIROLIMUS	0.91	0.1388	rs2032582(A;C)	0.17	C
13	CASP9	NIRAPARIB	0.907	0.2537			
14	IFNAR2	RUXOLITINIB	0.907	0.2022			
15	CASP9	OLAPARIB	0.906	0.2537			
16	CASP9	RUCAPARIB	0.904	0.2537			
17	COL1A1	DASATINIB	0.904	0.1145			
18	RICTOR	MIDOSTAURIN	0.903	0.1388			
19	NFATC1	THALIDOMIDE	0.903	0.3329	rs735482(A;C)	0.27	B
20	GPC3	ARSENIC TRIOXIDE	0.902	0.125			
21	NFATC1	POMALIDOMIDE	0.901	0.3329			
22	COL1A1	PONATINIB	0.901	0.1145			
23	ARID4B|CEACAM1|TMEM40	TRETINOIN	0.832	0.1616			
24	ARID4B|MUC16|PGP	DOCETAXEL	0.832	0.2854	rs7311358(A;A)	0.58	A
25	EIF4B|MUC16|PGP	ETOPOSIDE	0.832	0.2854	rs2291075(T;T)	0.17	C

Red and green rows indicate changes in the positions between the two rankings. Drugs in red descend and drugs in green ascend after ranking reordering using the ToxScores. SNPs are referred with their rs code. Patient's genotype is written inside parenthesis. Semicolons separate the two alleles.

Table 5.4: PanDrugs ranking results of somatic mutations obtained from metastasis versus control samples

RANK	GENE(s)	DRUG	DScore	GScore
1	MUC16	DOCETAXEL	0.812	0.125
2	MUC16	TAMOXIFEN	0.812	0.125
3	MUC16	ETOPOSIDE	0.812	0.125
4	MUC16	TOPOTECAN	0.812	0.125
5	MUC16	OREGOVOMAB	0.513	0.125
6	HCAR3	NICOTINIC ACID	0.312	0.429
7	MUC16	CYCLOSPORIN A	0.212	0.125
8	MUC16	SOFITUZUMAB VEDOTIN	0.0008	0.125
9	MUC16	ABAGOVOMAB	0.0008	0.125
10	MUC16	B43.13	0.0004	0.125
11	MUC16	OREGOVAMAB	0.0004	0.125
12	MUC16	BUSERELIN ACETATE	0.0004	0.125
13	MUC16	SODIUM BUTYRATE	0.0004	0.125
15	PIEZO2	1080622-86-1	-0.0004	0.0625
15	PIEZO2	ENZASTAURIN	-0.411	0.0625

Red and green rows indicate changes in the positions between the two rankings. Drugs in red descend and drugs in green ascend after ranking reordering using the ToxScores.

Table 5.5: Re-ranking of PanDrugs results (metastasis) using ToxScores

RANK	GENE(s)	DRUG	DScore	GScore	VAR	ToxScore	RELIAB
1	MUC16	TOPOTECAN	0.812	0.125			
2	MUC16	DOCETAXEL	0.812	0.125	rs7311358(A;A)	0.58	A
3	MUC16	TAMOXIFEN	0.812	0.125	rs1045642(A;G)	0.65	A
4	MUC16	ETOPOSIDE	0.812	0.125	rs1045642(A;G)	0.17	C
5	MUC16	OREGOVOMAB	0.513	0.125			
6	HCAR3	NICOTINIC ACID	0.312	0.429			
7	MUC16	CYCLOSPORIN A	0.212	0.125	rs2032582(A;C)	0.42	A
8	MUC16	SOFITUZUMAB VEDOTIN	0.0008	0.125			
9	MUC16	ABAGOVOMAB	0.0008	0.125			
10	MUC16	B43.13	0.0004	0.125			
11	MUC16	OREGOVAMAB	0.0004	0.125			
12	MUC16	BUSERELIN ACETATE	0.0004	0.125			
13	MUC16	SODIUM BUTYRATE	0.0004	0.125			
14	PIEZO2	1080622-86-1	-0.0004	0.0625			
15	PIEZO2	ENZASTAURIN	-0.411	0.0625			

Red and green rows indicate changes in the positions between the two rankings. Drugs in red descend and drugs in green ascend after ranking reordering using the ToxScores. SNPs are referred with their rs code. Patient's genotype is written inside parenthesis. Semicolons separate the two alleles.

6

Discussion and Future Work

In this work, we have integrated PharmGKB pharmacogenomics data and the information from four ADR knowledge bases to compute a toxicity score for 495 different compounds in PanDrugsdb and 2,294 variants in PharmGKB.

This ToxScore is unique for each drug-variant pair an reflects the possible noxiousness of the drug if it is administered to a patient with that particular genetic variation. Toxscores were computed taking into account the pharmacogenomics evidence of the drug-variant-ADR, as well as the number of ADR databases where those adverse events were reported, their severity and frequency.

ToxScores range from 0.02 and 1. The minimum value of this score is not 0 because each ADR was at least reported in PharmGKB. Around 40% of drug-variant-ADR associations had available information for the four parameters used to compute the ToxScore. However, the rest of entries lacked of pharmacogenomics evidence, severity and/or frequency data. When these values were not available for a high proportion of ADRs of a specific drug-variant pair, the resultant ToxScore was inevitably lower. Nevertheless, in those cases, a lower score did not mean that the drug was less noxious in presence of that variant. For this reason, we computed a Reliability value that accompanies each ToxScore. This Reliability Level can be either A, B or C and reflects the amount of data that was missing when computing the corresponding ToxScore. Thus, drugs with well documented pharmacogenomics interactions that cause sever and frequent ADRs will have ToxScores closer to 1 with a Reliability Level of A. On the contrary, Toxscores grouped in level C are not reliable at all, because they were computed using many missing data.

Once we obtained these ToxScores, we demonstrated that its value varied among drugs with the same target or that belonged to the same family of compounds. This result justifies our aim to incorporate pharmacogenomics information about toxicity to PanDrugs. Since drugs with the same therapeutic use can have different toxicity depending on the germinal variants of an individual, it seems reasonable to choose the less harmful medicine for each cancer patient.

Finally, we validated this ToxScore with real data from a paraganglioma patient. Our ToxScore was able to detect drugs that might had toxic effects due to patient's germinal variants. These variants were reported aside of the drug name, so the user could consult the literature to obtain further information. Moreover, in case that several drugs had the same DScores but different ToxScores, the ranking penalized the drugs with more toxic effect.

In conclusion, in this work we computed a metric of drug noxiousness due to pharmacogenomics interactions. This metric, which we have called ToxScore, varied among drugs with the same therapeutic use. Our ToxScore was able to detect drugs with risk of being toxic for a particular patient. In consequence, PanDrugs ranking was reordered and therapeutic candidates with the same effectiveness but lower toxicity were suggested in order to minimize the risk of adverse drug reactions.

In the future, we shall extend the number of PanDrugsdb compounds with pharmacogenomics information. We are particularly interested in DrugBank, which has incorporated pharmacogenomics data in its last release. Moreover, we plan to reduce the number of not available data for severity and frequency parameters by downloading DrugCentral's SPL sections and ADR frequencies. In addition, we will explore other weights for the four ToxScore parameters and validate the different combinations with patient's data in order to refine ToxScore formula.

Data availability

Our final dataset and ToxScores can be retrieved from
`www.mega.nz/#F!uLYzDKgC!PXIWnaJmd0xydMslVzGvLw`.

Abbreviations

- **ABC**: ATP Binding Cassette
- **ADR**: Adverse Drug Reaction
- **BAM**: Binary Alignment Map
- **BU**: Bioinformatics Unit; Unidad de Bioinformática
- **CIOMS**: Council for International Organizations of Medical Science
- **CNIO**: Spanish National Research Center; Centro Nacional de Investigaciones Oncológicas
- **CYP**: Cytochrome P450
- **DNA**: Deoxyribonucleic Acid
- **DScore**: Drug Score
- **EGFR**: Epidermal Growth Factor Receptor
- **EMA**: European Medicines Agency
- **EMBL**: European Molecular Biology Laboratory
- **EPAR**: European Public Assessment Report
- **EPS**: Escuela Politécnica Superior
- **EU**: European Union
- **FDA**: Food and Drug Administration
- **GO**: Gene Ontology
- **GPCR**: G Protein-Coupled Receptor
- **GScore**: Gene Score
- **HLGT**: High Level Group Term
- **HLT**: High Level Term
- **IDG**: Illuminating the Druggable Genome
- **LLT**: Low Level Term
- **MedDRA**: Medical Dictionary for Drug Regulatory Activities
- **NA**: Not Available
- **NADPH**: Nicotinamide Adenine Dinucleotide Phosphate

- **NGS**: Next-Generation Sequencing
- **NIH**: National Institutes of Health
- **NLP**: Natural Language Processing
- **NQO1**: NAD(P)H:Quinone Dehydrogenase 1
- **OATP1B1**: Organic Anion Transporting Polypeptide 1B1
- **PanDrugsdb**: PanDrugs Database
- **PGx**: Pharmacogenomics
- **PharmGKB**: Pharmacogenomics Knowledgebase
- **PMDA**: Pharmaceuticals and Medical Devices Agency
- **PROTECT**: Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium
- **PT**: Preferred Term
- **rs**: reference SNP
- **SAM**: Sequence Alignment Map
- **SDH**: Succinate Dehydrogenase
- **SIDER**: Side Effect Resource
- **SING**: Next Generation Computer Systems Group; Sistemas Informáticos de Nueva Generación
- **SLC**: Solute Carrier
- **SmPC**: Summary of Product Characteristics
- **SNP**: Single Nucleotide Polymorphism
- **SOC**: System Organ Class
- **SPL**: Structured Product Labeling
- **STITCH**: Search Tool for Interacting Chemicals
- **TAC**: Text Analysis Conference
- **UAM**: Universidad Autónoma de Madrid
- **UMLS**: Unified Medical Language System
- **US**: United States
- **VCF**: Variant Calling File
- **VHL**: von Hippel-Lindau
- **WES**: Whole Exome Sequencing
- **WGS**: Whole Genome Sequencing

Bibliography

[1] Biankin AV. The Road to Precision Oncology. *Nat Genet*, 2017;49(3):320-321. doi: 10.1038/ng.3796.

[2] Grzywa TM, Paskal W, and WÅĆodarski PK. Intratumor and Intertumor Heterogeneity in Melanoma. *Transl Oncol*, 2017; 10(6):956-975. doi: 10.1016/j.tranon.2017.09.007.

[3] Cancer Research UK. Treatment for Cancer. 2017 (Last accessed: 11-18-2018). `www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment`.

[4] Piñeiro-Yáñez E, Reboiro-Jato M, Gómez-López G, Perales-Patón J, Troulé K, et al. Pandrugs: A Novel Method to Prioritize Anticancer Drug Treatments According to Individual Genomic Data. *Genome Med*, 2018; 10(1):41. doi: 10.1186/s13073-018-0546-1.

[5] Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, and Gelbart WM. Somatic Versus Germinal Mutation. *An Introduction to Genetic Analysis. 7th edition*, 2000. New York: W. H. Freeman. ISBN-10: 0-7167-3520-2. `www.ncbi.nlm.nih.gov/books/NBK21894/`.

[6] Apellániz-Ruiz M. Identification of Genetic Markers Predictive of Paclitaxel-Induced Neuropathy. *Doctoral Thesis*, pages 39–40, 2016. `www.repositorio.uam.es/handle/10486/675066`.

[7] Zhou M, Chen Y, and Xu R. A Drug-Side Effect Context-Sensitive Network Approach for Drug Target Prediction. *Bioinformatics*, 2018. doi: 10.1093/bioinformatics/bty906.

[8] Patton K and Borshoff DC. Adverse Drug Reactions. *Anaesthesia*, 2018; 73 Suppl 1:76-84. doi: 10.1111/ anae.14143.

[9] National Cancer Institute (NCI). What is Cancer? 2015 (Last accessed: 11-14-2018). `www.cancer.gov/about-cancer/understanding/what-is-cancer`.

[10] Galceran J, Ameijide A, Carulla M, Mateos A, Quirós JR, et al. Cancer Incidence in Spain, 2015. *Clin Transl Oncol*, 2017; 19(7):799-825. doi: 10.1007/s12094-016-1607-9.

[11] Guan X. Cancer Metastases: Challenges and Opportunities. *Acta Pharm Sin B*, 2015; 5(5):402-18. doi: 10.1016/j.apsb.2015.07.005.

[12] Martincorena I and Campbell PJ. Somatic Mutation in Cancer and Normal Cells. *Science*, 2015; 349(6255):1483-9. doi: 10.1126/science.aab4082.

[13] Chow AY. Cell Cycle Control by Oncogenes and Tumor Suppressors: Driving the Transformation of Normal Cells into Cancerous Cells. *Nature Education*, 2010; 3(9):7. `www.nature.com/scitable/topicpage/cell-cycle-control-by-oncogenes-and-tumor-14191459`.

[14] Fouad YA and Aanei C. Revisiting the Hallmarks of Cancer. *Am J Cancer Res*, 2017; 7(5):1016-1036. `www.ncbi.nlm.nih.gov/pmc/articles/PMC5446472/`.

[15] Rahner N and Steinke V. Hereditary Cancer Syndromes. *Dtsch Arztebl Int*, 2008; 105(41):706-14. doi: 10.3238/arztebl.2008.0706.

[16] Gómez-López G, Dopazo J, Cigudosa JC, Valencia A, and Al-Shahrour F. Precision Medicine Needs Pioneering Clinical Bioinformaticians. *Brief Bioinform*, 2017. doi: 10.1093/bib/bbx144.

[17] Karki R, Pandya D, Elston RC, and Ferlini C. Defining "Mutation" and "Polymorphism" in the Era of Personal Genomics. *BMC Med Genomics*, 2015;8:37. doi: 10.1186/s12920-015-0115-z.

[18] European Bioinformatics Institute (EBI). What is variant calling? Last accessed: 2-13-2019. `www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction/variant-identification-and-analysis/what-variant`.

[19] Stanley LA. Drug Metabolism. *Pharmacognosy*, pages 527–545, 2017. doi:10.1016/b978-0-12-802104-0.00027-5.

[20] Food and Drug Administration (FDA). Structured Product Labeling Resources. 2018 (Last accessed: 2-13-2019). `www.fda.gov/forindustry/datastandards/structuredproductlabeling/`.

[21] European Medicines Agency (EMA). Draft Presentation: Summary of Product Characteristics. Last accessed: 2-13-2019. `www.ema.europa.eu/documents/presentation/presentation-summary-product-characteristics_en.pdf`.

[22] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Meddra version 21.1. 2018. `www.meddra.org`.

[23] Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Res*, 2004;32(Database issue):D267-70. doi: 10.1093/nar/gkh061.

[24] Szklarczyk D D, Santos A, von Mering C, Jensen LJ, Bork P, et al. STITCH 5: Augmenting Protein-Chemical Interaction Networks with Tissue and Affinity Data. *Nucleic Acids Res*, 2016; 44(D1):D380-4. doi: 10.1093/nar/gkv1277.

[25] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, et al. Pubchem Substance and Compound Databases. *Nucleic Acids Res*, 2016;44(D1):D1202-13. doi: 10.1093/nar/gkv951.

[26] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res*, 2018;46(D1):D1074-D1082. doi: 10.1093/nar/ gkx1037.

[27] Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, et al. DrugCentral: Online Drug Compendium. *Nucleic Acids Res*, 2017;45(D1):D932-D939. doi: 10.1093/nar/gkw993.

[28] European Medicines Agency (EMA). Medicines. Last accessed: 1-29-2019. `www.ema.europa.eu/en/medicines`.

[29] Kuhn M, Letunic I, Jensen LJ, and Bork P. The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res*, 2016;44(D1):D1075-9. doi: 10.1093/nar/gkv1075.

[30] Juan-Blanco T, Duran-Frigola M, and Aloy P. IntSide: A Web Server for the Chemical and Biological Examination of Drug Side Effects. *Bioinformatics*, 2015;31(4):612-3. doi: 10.1093/bioinformatics/ btu688.

[31] Demner-Fushman D, Shooshan SE, Rodriguez L, Aronson AR, Lang F, et al. A Dataset of 200 Structured Product Labels Annotated for Adverse Drug Reactions. *Sci Data*, 2018;5:180001. doi: 10.1038/sdata.2018.1.

[32] European Medicines Agency (EMA) and Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium Work Package 3 (PROTECT WP3). PROTECT ADR database. 2017. `www.imi-protect.eu/adverseDrugReactions.shtml`.

[33] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, et al. Pharmacogenomics Knowledge for Personalized Medicine. *Clin Pharmacol Ther*, 2012;92(4):414-7. doi: 10.1038/clpt. 2012.96.

[34] World Health Organization (WHO). Definitions. Last accessed: 12-20-2018. `www.who.int/medicines/areas/quality_safety/safety_efficacy/trainingcourses/definitions.pdf`,.

[35] U.S. Department of Health, Human Services (HHS), Food and Drug Administration (FDA), et al. Guidance for Industry. Labeling for Human Prescription Drug and Biological Products – Implementing the PLR Content and Format Requirements. pages 7–13, 2013. `www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM075082.pdf`.

[36] Chen EY, Tan CM, Kou Y, Duan Q Q, Wang Z, et al. Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. *BMC Bioinformatics*, 2013;14:128. doi: 10.1186/1471-2105-14-128.

[37] Bar-Shavit R, Maoz M, Kancharla A, Nag JK, Agranovich D, et al. G Protein-Coupled Receptors in Cancer. *Int J Mol Sci*, 2016;17(8). pii: E1320. doi: 10.3390/ijms17081320.

[38] US National Library of Medicine. *NQO1* gene. 2019 (Last accessed: 2-11-2019). `www.ghr.nlm.nih.gov/gene/NQO1`.