UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXV

New approaches on statistical modeling for drug safety data

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof.ssa Giovanna Boccuzzo

Co-supervisore: Prof. Nicholas Tatonetti

Dottorando: Pietro Belloni

25 Febbraio 2023

Abstract

Adverse events associated with drugs are one of the leading causes of morbidity and mortality in the world, and statistics has always been an essential tool to contrast them. In this thesis, we focus on the set of statistical models and techniques used in pharmacovigilance, i.e. the detection of adverse effects of drugs after they have been approved and placed on the market. The first part of this thesis will describe the process that results in the approval of a drug by pharmacovigilance authorities. Next, the typical pharmacovigilance data collection system, based on the spontaneous report of adverse drug events, will be illustrated. The statistical models used mainly in the analysis of spontaneous data (known as disproportionality models) in the literature will then be reviewed and commented on. In the second part of the thesis, a new model for pharmacovigilance data will be proposed. This model, based on a lasso-penalized regression, is designed to analyze pharmacovigilance data and find new associations between drugs and adverse drug events, including drug-drug interactions, that may cause adverse events themselves. The model was tested on both simulated and real data. In the third part of the thesis, a new approach to statistics applied to pharmacovigilance is discussed. We show how the ability to find new associations between drugs and adverse events can be increased by including information from the biochemical structure of drugs. Specifically, techniques peculiar to natural language processing were used to map a drug into an embedding space of latent variables that describes its biochemical characteristics. The use of these latent variables, when properly combined with spontaneous data, can be a turning point in pharmacovigilance procedures.

Sommario

Gli eventi avversi associati ai farmaci sono una delle principali cause di malattia e decesso al mondo e la statistica è, da sempre, uno strumento essenziale per contrastarli. Questa tesi si concentra sull'insieme di modelli e tecniche statistiche usate in fase di farmacovigilanza, ovvero l'individuazione degli effetti avversi di farmaci dopo che questi sono stati approvati e messi sul mercato. Nella prima parte della tesi verrà descritto il processo che porta all'approvazione di un farmaco da parte delle autorità di farmacovigilanza. Successivamente, si illustra il sistema di raccolta dati tipico della farmacovigilanza, basato sulla raccolta spontanea delle segnalazioni degli effetti collaterali di farmaci. Verranno quindi passati in rassegna e commentati i modelli statistici principalmente usati nell'analisi dei dati spontanei (noti come modelli di disproporzionalità) presenti in letteratura. Nella seconda parte della tesi verrà proposto un nuovo modello per l'analisi dei dati di farmacovigilanza. Questo modello, basato su una regressione con penalizzazione lasso, è stato pensato per analizzare i dati di farmacovigilanza e trovare nuove associazione fra farmaci ed effetti avversi, includendo anche le interazioni fra i farmaci, che possono a loro volta provocare degli effetti avversi. Il modello è stato testato sia su dati simulati che su dati reali. Nella terza parte della tesi si discute di un nuovo approccio alla statistica applicata alla farmacovigilanza. Si mostra come la capacità di trovare nuove associazione fra farmaci ed eventi avversi può essere incrementata includendo le informazioni provenienti dalla struttura biochimica dei farmaci. In particolare, sono state usate tecniche proprie del *natural langue processing* per proiettare un farmaco in uno spazio di variabili latenti che ne descrive le caratteristiche biochimiche. L'utilizzo di queste variabili latenti, se debitamente affiancato ai dati spontanei, può essere un punto di svolta nelle procedure di farmacovigilanza.

How many young people have become scientists because of you?
As of today there is one more.
A Piero Angela, un maestro per tutti noi.

Acknowledgements

This work was made possible by the great support received from my supervisor, Prof. Giovanna Boccuzzo (University of Padua), and my co-supervisor, Prof. Nicholas Tatonetti (Columbia University). I extend my gratitude to the coordinator of the PhD Course Prof. Nicola Sartori and the entire administrative office. During my time as a PhD Student, many colleagues, both from the University of Padua and Columbia University, contributed directly and indirectly to the success of my work. It is not possible to mention them all, but thanks in particular to Margherita (my older sister), Matteo, Corrado, Emanuele, Laura, Mattia, Dafne, Harry, C.J., Davide, Giovanna and all my friends from Scienze Statistiche and DBMI. Special thanks to Beatrice, my friends and my family – especially Antonio, who contributed like no one else to my scientific education.

The greatest thanks to my colleagues from cycle XXXV. I have always admired your cleverness, perseverance and empathy, and I am very proud to have gone through this journey with you. Before starting, I knew that I would have some colleagues; now that we have finished, I know that I had a second family. Touqeer, Beppe, Erika, Nicolas, Cristian, Marco, Caizhu, Fariborz, I wish you all the best. This work is because of you.

Contents

List of Figures

List of Tables

Introduction

Overview

For their very nature, drugs not only provide benefits, but also cause harm. An adverse drug event (ADE) is defined as (Edwards and Aronson, 2000):

> "An appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product."

Most of these events are discovered during the long phases of drug development and are therefore already known by the time a drug is launched on the market. Other events, on the other hand, are only noticed in the post-marketing phase, making them a major source of morbidity and even mortality worldwide, sometimes even causing a drug to be withdrawn from the market – some notable cases are described in McBride (1961); Charatan (2001); Qureshi *et al.* (2011). This claim is widely established in the literature and has been the subject of many studies for several decades (Bates *et al.*, 1995; Ross, 2001; Shojania *et al.*, 2002; Nebeker *et al.*, 2004).

Statistical analysis plays a crucial role in helping to discover new associations between drugs and ADEs by analyzing spontaneous data collected during the post-marketing phase. Spontaneous data have the peculiarity of having only cases and not controls, so unique statistical models that simulate the presence of a denominator have to be used. For more than two decades, pharmacovigilance authorities have used a variety of statistical techniques, known as *disproportionality models*, to systematically scan drug safety data.

For this purpose, basic statistical methods such as *reporting odds ratio* and *proportional reporting ratio* were deployed (Evans *et al.*, 2001; Rothman *et al.*, 2004).

These methods are extremely simple and fast to evaluate the association between a drug and an ADE as they require only a few calculations. Furthermore, their result is always interpretable from an epidemiological point of view. On the other hand, basic methods such as the reporting odds ratio are inconvenient if they have to be used to scan a very large database, such as spontaneous pharmacovigilance databases. Indeed, this would require the construction of a number of frequency tables equal to the number of drug-ADE pairs, which is often extremely high.

Hence, some other models (mostly Bayesian) have been introduced with the aim of analyzing spontaneous large-scale pharmacovigilance data. The best known model is *gamma-Poisson shrinkage model*, introduced by DuMouchel in the late 1990s (DuMouchel, 1999) and then later developed with several extensions (DuMouchel and Pregibon, 2001; Fram *et al.*, 2003). Another important model, with a shrinkage mechanism similar to the previous model, is the *Bayesian confidence propagation neural network* (Bate *et al.*, 1998; Norén *et al.*, 2006).

The advantages of Bayesian disproportionality models over classical disproportionality models are many. For example, they use a prior distribution to introduce a useful shrinkage to correct the estimated association between a drug and an ADE; they can be modified to account for drug-drug interaction (Szarfman *et al.*, 2002); they can adjust the estimates to control for demographic variables (DuMouchel and Harpaz, 2012) and they are faster at scanning an entire database. However, they are computationally more intensive: the implementation of these methods requires many more calculations than the reporting odds ratio or the proportional reporting ratio. In addition, their outcome is not easily interpreted due to shrinkage. The measures they produce are not exactly interpretable as known epidemiological measures, as is the case with classical methods (e.g. the reporting odds ratio can be interpreted as a classical odds ratio).

Finally, with the predominance of machine learning and deep learning in every area of data analysis, many uses of this class of models have been observed in recent years. For example, tree models, such as random forest and boosting, are discussed in Pham *et al.* (2019) and lasso and class-imbalanced subsampling lasso are successfully deployed by Ahmed *et al.* (2016). Some authors have recently presented tools that combine disproportionality models (especially the gamma-Poisson shrinkage model) with statistical techniques from other fields, such as the Synthetic Minority Oversampling Technique (SMOTE) correction or the E-M algorithm (Xiao *et al.*, 2018; Wei *et al.*, 2020).

The study of statistical models for disproportionality analysis aims at identifying new associations between drugs and ADEs more accurately. Therefore, contributions on this topic are valuable because they can help pharmacovigilance authorities, research institutes, pharmaceutical companies, or other stakeholders better identify new associations between drugs and adverse reactions. In this thesis, we present some contributions in the area of statistics applied to pharmacovigilance.

Main contributions of the thesis

A drug safety signal is defined as an association between a drug and an adverse drug event found using disproportionality analysis. The task of disproportionality models is the reporting (signaling) of new drug-ADE associations to pharmacovigilance authorities, which might proceed to verify the association using other data. The signals coming from drug-drug interactions are also important, but at the same time are more difficult to detect using the models mentioned above, for this reason there are fewer models dedicated to interactions detection in the literature. The original contribution of this thesis is an attempt to improve the way signals are generated, without neglecting the presence of interactions and through the use of alternative data sources.

First, we introduce a new model for the identification of drug safety signals. Specifically, we identify drug-ADE pairs and drug-drug interaction pairs using a *lasso* regression, which uses the Bayesian Information Criterion (BIC) for variable selection. Next, we extend the model for the selection of variable interactions to capture ADEs associated with a drug-drug interaction.

To assess the performance of the model, we use a simulation study. Then, the method is tested on real data from the FDA Adverse Event Reporting System database, the database of spontaneous reports managed by the U.S. pharmacovigilance authority Food and Drug Administration (FDA). We compare the results with a state-of-the-art Bayesian disproportionality model as a benchmark, and we find that our results are competitive, although we noticed several difficulties in establishing the accuracy of disproportionality models because of the lack of gold standards in this field.

We notice that the proposed model, like many other disproportionality models, does not show outstanding performance. Indeed, as recently shown in some

comparative studies on one of the very few gold standards available, very complex disproportionality models do not exceed simpler ones in accuracy and precision (Pham *et al.*, 2019). Similarly, more sophisticated machine learning models often have lower performance than simple reporting odds ratios. The reason for these results is due to many biases induced by the poor quality of the spontaneous data on drug safety.

To overcome the difficulties imposed by spontaneous data as the sole data source, we introduce a new technique to predict the presence of adverse drug events. Typically, surveillance is based on the disproportionality analysis of spontaneous reporting system databases, but their voluntary nature causes multiple biases that induce a limited predictive performance of statistical models. Alternative data sources can help overcome this limitation.

We used data on the biochemical structure of the drugs alongside spontaneous pharmacovigilance data to obtain a better overall performance. To represent the chemical structure of the drug's active ingredients, we used MACCS vectors and SMILES strings. The Molecular Access System (MACCS) is a 166-bit mathematical representation of a chemical compound, obtained with only binary features. The Simplified Molecular Input Line Entry System (SMILES) is a textual representation of a chemical compound, obtained by an algorithm that translates a two-dimensional graph (the structure of the molecule) into plain text.

We used MACCS vectors as a set of latent binary features to predict the presence of a latent adverse event. We used SMILES strings to derive an embedding space using a model similar to the Bidirectional Encoder Representations from Transformers (BERT), known as ChemBERTa. We compared the predictive power of these two sets of latent features and found that the ChemBERTa embedding space provides higher performance. Then we combined the features obtained from the embedding space with data from the FAERS spontaneous database to predict the presence of an adverse event with a performance equal to or better than the usual disproportionality models. Since statistical models used in disproportionality analysis are limited by the spontaneous nature of the data, we can conclude that the use of an endogenous data source reduces the bias and leads to better results.

The thesis is concluded with a discussion of the results obtained. As innovative as our results are, much remains to be done to improve the performance of statistical models used in pharmacovigilance, both from the perspective of data

sources and modeling. Therefore, many future developments may bring interesting contributions to this field, which remains of paramount importance to public health.

Chapter 1

Disproportionality analysis: the role of statistics in drug safety

1.1 The drug development process

From the conception of a drug, every stage of its development is focused on safety. In this chapter, we provide a general introduction to the drug safety procedure through each of its phases, with special attention to the post-marketing phase.

As stated by the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA), the classical drug development process consists of several steps:

1. **Drug discovery**. The potential number of possible biochemical compounds that could be used as a drug is extremely large. Today, very sophisticated techniques are used to filter out possible compounds and focus on those that are really promising, including models of machine learning, deep learning, and natural language processing (NLP) models (Chen *et al.*, 2018).

2. **Preclinical research step**. Once a promising compound is found, several experiments are conducted in a laboratory environment to assess its mechanisms of action, benefits, toxicity, and interaction with other compounds. During the preclinical phase, many studies are conducted both in vitro (controlled environment) and in vivo (animal testing).

3. **Clinical research step**. The crucial step in the development of any drug is clinical research and consists of a series of trials carried out by administering

the potential drug to people. Typically, clinical research is divided into three phases; these differ in sample size and study duration.

(a) Phase I. The potential drug is tested on healthy volunteers[1], with a sample size typically smaller than 80 and a study duration of some months. The main goal of phase I studies is to collect data on the interactions of the potential drug with the human body. Preclinical in vivo data is used to assess the best dosage and the trade-off between beneficial effects and side effects. The first adverse drug events, especially the more serious ones, are usually noticed in this phase. Special attention is paid to how the potential drug is absorbed, metabolized, and disposed of by the body. The FDA estimates that $\sim 70\%$ potential drugs can move on to the next stage (FDA, 2018a).

(b) Phase II. The potential drug is tested on patients with the target disease, with a sample size typically on the order of hundreds and a study duration that can range from a few months up to two years. The dosage is the one recommended during therapy and is based on dara from phase I and preclinical stage. In some cases, a control group is involved, with placebo or standard treatment administration. Similarly to phase I, phase II is also aimed at drug safety; the sample size is not large enough to demonstrate whether the drug is effective, but many adverse events are noticed compared to the previous phase. The FDA estimates that around $\sim 33\%$ potential drugs can move on to the next stage.

(c) Phase III. The last phase is the core of the clinical research step. The clinical dosage of the potential drug is tested in several thousands of patients with the target disease in the presence of a control group. The study duration is usually one to four years, but it can be even longer. A larger sample size allows for different subpopulations; therefore, statistical analyses can be more complex and allow for the evaluation of unexplored aspects of both the risks and benefits of the potential drug. Most of the drug safety data are provided by phase III studies. The

[1]With the exception of cancer drugs, which are always tested on patients, given their serious side effects.

FDA estimates that approximately 25% - 30% potential drugs will successfully complete phase III.

The phases of clinical research described above are considered standard procedures by pharmacovigilance authorities. However, nowadays, with the advancement of precision medicine and the use of increasingly developed biomarkers, many clinical trials can follow alternative procedures – some well-known examples being basket and umbrella trials, typically used in oncology (Park *et al.*, 2020).

4. **Agency approval and post-marketing surveillance**. Once enough data have been collected from the clinical research phases and the safety and efficacy of the drug have been proven, the company that developed the potential drug can submit an approval to market it from the competent pharmacovigilance authority.

   (a) Review. The pharmacovigilance authority starts a detailed review of the whole process. A panel of experts is assembled to evaluate both the efficacy and safety of the potential drug using all available data. Since most drugs have known adverse reactions from the clinical phases, the panel must consider the trade-off between positive and negative effects. If this trade-off is considered beneficial, the drug can be commercialized. The review process can be expedited in different ways. For example, the FDA provides a fast track for breakthrough therapies or drugs that fulfill an unmet medical need. A comprehensive description of FDA accelerated approvals is available on the agency's website (FDA, 2018b).

   (b) Post-marketing surveillance. Although clinical research provides many safety data, once a drug is on the market, more adverse drug reactions previously ignored can be observed. This phenomenon is due to the fact that the drug is administered in quantities that far exceed the sample sizes of the clinical stages, so the probability of observing rare cases increases considerably. In addition, the drug now acts in an uncontrolled environment. Even though phase III studies collect data from a heterogeneous population, it is impossible to control the enormous number of variables in the world outside the controlled context of a trial. For that reason, pharmacovigilance authorities, such as the FDA and the EMA,

constantly monitor drugs already available on the market so that they are ready to recall them (temporarily or permanently) if a new adverse reaction makes the trade-off between positive and negative effects no longer beneficial. The post-marketing surveillance process is also known as *phase IV* to emphasize its continuity with the clinical research phase.

The method used for this surveillance practice consists of two steps: the collection of spontaneous data and their analysis with statistical disproportionality models. Both of these steps will be explored in more detail later in this thesis.

Phase IV continuously monitors a drug after its approval; therefore, the availability of up-to-date safety data is of paramount importance.

1.2 The key role of spontaneous databases

The major pharmacovigilance authorities collect data needed for drug safety analyses in large databases known as *spontaneous databases*. Each authority maintains a different database, but this database does not necessarily collect only reports from its territory, as it can also collect reports from different territories. For example, the database maintained by the FDA, known as FAERS (FDA Adverse Event Reporting System), contains data not only from the United States but also from other states around the world.

Moreover, FAERS is the only large spontaneous database whose download is totally public. Therefore, it is often used to develop and test new statistical and data mining techniques on pharmacovigilance data. Data access of other spontaneous databases (like the EMA database EudraVigilance) is granted only under special conditions. For this reason, whenever reference is made to drug safety data in this thesis, it means data from the FAERS database, unless otherwise specified.

1.2.1 Relevance and issues of spontaneous data

Since their creation, spontaneous databases have been shown to be of great use in pharmacovigilance (Lu, 2009). However, they have very special characteristics that make their analysis particularly complex and unusual (Zorych *et al.*, 2013).

As the name suggests, spontaneous databases have a spontaneous data collection mechanism: anyone can submit an adverse drug reaction report. Most records are submitted by physicians or patients, but no one is required to report an adverse drug reaction[2]; spontaneity itself is the main cause of the problems of drug safety data collection. Some problems will be explored in more detail in the following sections.

1.2.1.1 Absence of control data

The most immediate consequence of spontaneous data collection is the absence of control data; only data on the drug taken and the symptoms that have occurred since its administration are collected. For example, there are no data on patients who did not experience a certain symptom after receiving a drug or data on patients who did not receive the drug and experience the symptom. This implies that it is impossible to use the many statistical methods for case-control studies or to calculate the classical epidemiological measures of occurrence (e.g., prevalence, incidence) and association (e.g., risk ratio, odds ratio). Ad hoc models, called disproportionality models, must be used as an alternative.

1.2.1.2 Underrepresentation, Weber effect and notoriety bias

Only a small fraction of ADEs is collected in spontaneous databases, which suffer from a constant underrepresentation of the cases count. In addition, after a drug is released on the market, there is increasing concern about possible ADEs, which are reported more frequently. However, after the drug has been on the market for a certain period of time, concern about side effects wanes, leading to even greater underrepresentation.

This phenomenon is also known as the Weber effect, first highlighted by Weber (1984) and then further explored on spontaneous data by other authors. The Weber effect is a cycle that affects almost all drugs, which see a rapid increase in spontaneous reports for about the first year and a half to two years after marketing. Subsequently, a decrease is observed, despite the fact that prescriptions for the drug will continue to increase for a long time (Wallenstein and Fife, 2001; Hartnell and Wilson, 2004).

[2]An exception is made for some pharmaceutical industries, which are required by the FDA to report the ADEs of their products already on the market (FDA, 2022).

In some cases, the number of spontaneous reports may undergo another sharp increase induced by the sudden notoriety of a drug and/or its side effect. This notoriety can result from a new scientific discovery related to the drug or even from a piece of news that raises awareness among patients and healthcare staff. This second phenomenon is known as notoriety bias or publicity bias (Neha *et al.*, 2019).

Therefore, as some authors argue, the reliability of spontaneous data is high only in the first four years after the drug is introduced to the market (Stephenson and Hauben, 2007).

1.2.1.3 Other issues of spontaneous data

Spontaneity induces incompleteness of data: an inexperienced reporter (someone who does not work in health care or someone who has never reported before) can easily miss some details of their report, resulting in the presence of missing values. For the same reason, some sources of records are considered more reliable than others. Another problem already debated in the literature is the presence of duplicate records, and drug-naming issues can also be observed, as manufacturers may use different names for the same drug.

If missing data and duplicate records can be addressed with statistical techniques (Banda *et al.*, 2016), the drug-naming issue could be solved by using biochemical data from the active ingredient of the drug.

1.2.1.4 Is it still worth using spontaneous databases?

Despite the multiple biases that affect spontaneous databases, there is no alternative data source to completely replace them. However, these biases must be taken into account, which is why increasingly sophisticated disproportionality models have been developed over time; caution is needed throughout the process, always remembering that any results obtained from these models will be affected by the problems of data spontaneity. So, as some authors propose, it is still worthwhile to use spontaneous databases, but those who use them for research must always keep in mind their limitations (Stephenson and Hauben, 2007). Although abandoning spontaneous data is currently unthinkable, integrating them with data from alternative sources could lead to a more precise identification of ADEs.

1.2.2 The FDA Adverse Event Reporting System (FAERS)

In order to support the FDA post-marketing safety surveillance for pharmaceuticals, a spontaneous database has been created. As stated by the FDA: "The FDA Adverse Event Reporting System (FAERS) is a database that contains adverse event reports, medication error reports and product quality complaints resulting in adverse events that were submitted to FDA" (FDA, 2018c).

FAERS has collected FDA spontaneous reports since the fourth quarter of 2012, replacing the previous AERS database. The database is updated quarterly and can be freely consulted and downloaded; to date (December 2022), the total number of available records slightly exceeds 18 million.

FAERS database consists of six different tables, each containing different information.

1. **Demographic** Table with the data of the patient to whom the drug was administered, such as age, sex, or country of residence.

2. **Drug** Table with drug data, such as its commercial name, the name of the active principle, the dosage, the route of administration, or the batch identification number.

3. **Reaction** Table with description of adverse drug reaction events. Adverse events are encoded using MedDRA terminology, which details the type of reaction (Brown, 2007).

4. **Outcome** Table with data on the outcome of the patient after ADE (e.g. hospitalization, death...).

5. **Report sources** Table with details on the ADE reporter.

6. **Therapy** Table with details on the therapy, if any, to which the drug administered pertained.

These tables are connected by primary keys, as shown in Figure 1.1.

So far, the number of FAERS records has grown over time. In 2013 (the first full year of data gathering), there were about 1.07 million registered cases. In 2022, the registered cases were about 2.33 million. Typically, the vast majority of reporters are equally divided between drug consumers and healthcare professionals (Figure 1.2a). About two-thirds of the cases are permanently from the United States, while the remaining third are from other countries (Figure 1.2b).

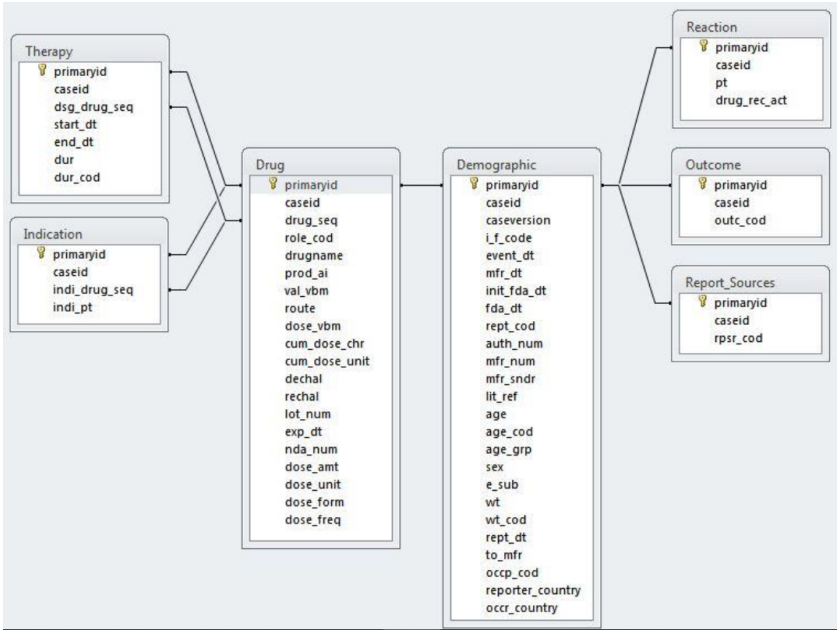

FIGURE 1.1: Relational structure of FAERS database.

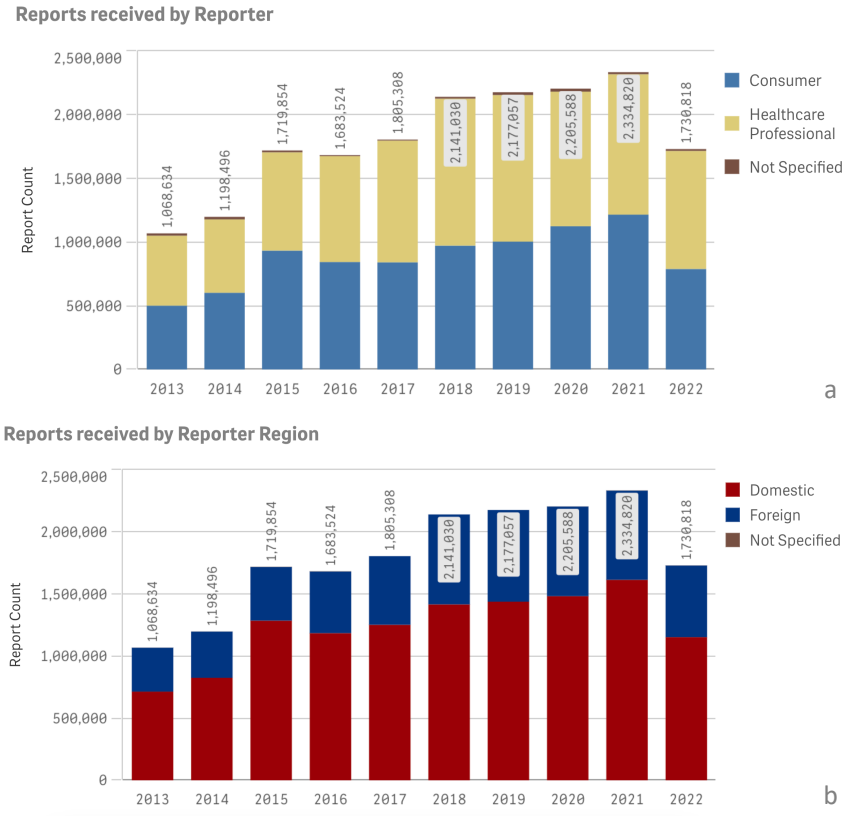

FIGURE 1.2: a) Number of FAERS reports by type of reporter from 2013 to 2022, third quarter. b) Number of FAERS reports by reporter region (Domestic: U.S. only, Foreign: rest of the world) from 2013 to 2022, third quarter.

Since the amount of spontaneous data is growing, we need increasingly accurate statistical models for its analysis. The next chapter is devoted to a review of the most prominent disproportionality models in the field literature.

1.3 Review of statistical models for disproportionality analysis

1.3.1 Background and notation

The underlying idea of any statistical application in pharmacovigilance is that if a drug and an ADE are observed together a disproportionate number of times, then there could be an association between the two (which is why it is called *disproportionality analysis*). Once an association is identified in a spontaneous database, it is the responsibility of pharmacovigilance authorities to investigate the association using other data such as, for example, data from previous clinical trials.

Suppose that a spontaneous database has n rows corresponding to n ADE reports. We can denote three sets of binary variables (or features): $(X_{i,1}, X_{i,2}, \ldots, X_{i,p})$ indicates whether the administration of a certain drug appears in the i-th report, $(Y_{i,1}, Y_{i,2}, \ldots, Y_{i,q})$ indicates the presence in the i-th report of a certain ADE, and $(Z_{i,1}, Z_{i,2}, \ldots, Z_{i,r})$ denotes a set of demographic information concerning the patient to whom the i-th report is referred (such as age or gender). Consequently, the total data size will be $(p + q + r)n$.

Given a drug X_k and an adverse drug event Y_l, a naive approach to the problem is to define a disproportionality when the probability of observing the adverse event given the drug is greater than a fixed baseline value B. Therefore, when

$$\Pr\left(Y_l = 1 | X_k = 1\right) = \frac{N_{k,l}}{N_k} \geq B \tag{1.1}$$

with $N_k = \sum_{i=1}^{n} X_{i,k}$ and $N_{k,l} = \sum_{l=1}^{n} X_{i,k} Y_{j,l}$. Since it is difficult to consider all drug-ADE pairs, an approach to speed up association mining would be to use an association rule algorithm. The literature on association rules is highly developed, especially in other fields, such as basket analysis or text mining (Agrawal

and Srikant, 1994). For example, the most famous algorithm for finding association rules (Apriori) can be used successfully in spontaneous pharmacosurveillance databases. However, these methods are heavily influenced by the choice of threshold B, which would be an arbitrary threshold that varies from time to time depending on the drug being considered. Furthermore, some authors have made criticisms regarding the difficulties of association rule algorithms in adjusting for events with different frequencies (see, for example, Silverstein *et al.* (1998)).

As seen, it is difficult to approach the problem as a simple search for association rules. Thus, since the late 1990s, the literature has mainly developed around two classes of models: basic disproportionality methods (which we also refer to as *frequentist models*) and *Bayesian models*.

1.3.2 Basic methods

In a classic case-control study, the simplest way to assess an association between dichotomous variables is to use a contingency table. From the contingency table, many association measures can be calculated, such as the odds ratio or the relative risk. Statistical tests can also be used, both with an approximate null distribution, such as the χ^2 test, or with an exact distribution, such as Fisher's exact test (Ahmed *et al.*, 2010). Since the main limitation of spontaneous drug safety data is the absence of control data (patients who took drugs but had no adverse events, or patients who had adverse drug events without taking drugs), the idea underlying the basic methods is indeed to build some sort of surrogate to the controls.

A surrogate contingency table leverages the fact that spontaneous databases are very large, and so a lot of different drugs and adverse drug events are recorded. Given a drug X_k and an adverse drug event Y_l, a table can be constructed using reports from patients who did not receive the drug X_k and did not experience the side effect Y_l (Table 1.1).

TABLE 1.1: Surrogate contingency table for drug X_k and adverse drug event Y_l.

	Drug X_k	Other drugs	Total
ADE Y_l	n_{11}	n_{10}	$n_{1.}$
Other ADEs	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

From this table, several measures of association can be calculated. The relative risk and the odds ratio, commonly used in epidemiology, when calculated from spontaneous data are referred to as *proportional reporting ratio* (PRR) and the *reporting odds ratio* (ROR)

$$\mathrm{PRR} = \frac{n_{11}/(n_{11}+n_{01})}{n_{10}/(n_{10}+n_{00})} \tag{1.2}$$

$$\mathrm{ROR} = \frac{n_{11}n_{00}}{n_{10}n_{01}} \tag{1.3}$$

described, for example, in Evans *et al.* (2001) and Rothman *et al.* (2004).

The PRR is the ratio between the occurrence of Y_l among all ADEs recorded after the administration of X_k over recorting the event of another ADE after the administration of any other drug. The ROR is the ratio between the ratio of the count of X_k reported with Y_l over the count of X_k reported with other ADEs and the ratio of the count of other drugs reported with Y_l over the count of other drugs reported with other ADEs. Both indicators are greater than 1 if the drug is associated with the adverse drug event. Although the two measures are very similar (the odds ratio is asymptotically equivalent to the relative risk for small probabilities), ROR is generally preferred in pharmacovigilance (Waller *et al.*, 2004). The interpretation of ROR and PRR is equivalent to their counterparts in case-control studies. For example, if ROR = 2.53, then reports with Y_l observed after the administration of drug X_k occur 2.53 times more frequently than reports with Y_l observed after the administration of other drugs.

For both statistics, the confidence interval (CI) can be computed, useful to identify which signals are significant; a signal is considered significant when the lower bound of the interval is greater than 1. Since both PRR and ROR have zero as the lower bound and no upper bound, their distribution is asymmetric and the confidence interval is constructed using a Gaussian approximation of their logarithm. For example, a confidence interval with level α for the PRR statistics is given by

$$\mathrm{CI}(\alpha) = \exp\left\{\log(\mathrm{PRR}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{01}}-\frac{1}{n_{11}+n_{10}}-\frac{1}{n_{01}+n_{00}}}\right\} \tag{1.4}$$

and, similarly, with level α for the ROR statistics is given by

$$\mathrm{CI}(\alpha) = \exp\left\{\log(\mathrm{ROR}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}\right\} \tag{1.5}$$

with $z_{\alpha/2}$ quantile of the standard normal distribution.

PRR and ROR statistics can take into account demographic variables using a Mantel-Haenszel approach. If we treat the demographic variables as confounders variables and suppose that they generate S strata, we denote by n_s the number of observations in stratum s and by $a_s = n_{11_s}$, $b_s = n_{10_s}$, $c_s = n_{01_s}$, $d_s = n_{00_s}$ the components of the surrogate contingency table restricted to the single stratum. The adjusted ROR (aROR) is defined as

$$\mathrm{aROR} = \frac{\sum_{s=1}^{S} \frac{a_s d_s}{n_s}}{\sum_{s=1}^{S} \frac{b_s c_s}{n_s}} \tag{1.6}$$

and its confidence interval with level α is defined as

$$\mathrm{CI}(\alpha) = \exp\left\{\log(\mathrm{aROR}) \pm z_{\alpha/2}\sqrt{\frac{\sum_{s=1}^{S} \frac{(a_s+b_s)(c_s+d_s)(a_s+c_s)(b_s+d_s)}{n_s^2(n_s-1)}}{\sum_{s=1}^{S} \frac{a_s b_s}{n_s} \quad \sum_{s=1}^{S} \frac{c_s d_s}{n_s}}}\right\}. \tag{1.7}$$

The derivation of the adjusted PRR and its confidence interval is similar.

The main advantages of PRR and ROR are their ease of implementation. To measure the association between a drug and an adverse effect, it is sufficient to construct the surrogate contingency table and perform some simple calculations. Moreover, their interpretation is straightforward. However, there are also several drawbacks, mostly originating from the extreme unbalance of the surrogate contingency table. It is easy to notice how their confidence intervals are subject to the assumption of Gaussianity on the logarithm and the consequent need to have the counts in the surrogate contingency matrix sufficiently large. Although it is unlikely since the data size is usually large, if the frequencies are particularly low, n_{10} or n_{01} can be zero and it would not be possible to calculate the statistics. Finally, these methods can consider only a drug and an ADE at a time and are difficult to implement if we want to detect the association between the interaction between two different drugs and an ADE. Nevertheless, PRR and ROR are currently used for the analysis of spontaneous drug safety data (Ang *et al.*, 2016; Shan *et al.*, 2020; Diaby *et al.*, 2021).

Once a surrogate contingency table for a drug and ADE is obtained, association tests can also be performed. For example, χ^2 and Yule's Q tests of independence can be used for association mining purposes (Silverstein *et al.*, 1998; van Puijenbroek *et al.*, 2002). In addition, a probabilistic approach can be used. The probability of the number of records with a specific drug-ADE pair can be calculated as a mean of the Poisson probability - see, for example, Tubert *et al.* (1992), but this approach requires the strong assumption that there is no relationship between ADE and the drug.

Association tests still suffer from the same disadvantages as association measures, first and foremost the challenge of including drug-drug interaction. Their use in the literature is moderate compared to PRR and ROR, which are by far the most widely used non-Bayesian methods in data analysis (Montastruc *et al.*, 2011).

1.3.3 Bayesian models

1.3.3.1 Gamma-Poisson Shrinkage model

One way to mitigate the extreme unbalance of the surrogate contingency table is to place a prior distribution on the counts of the cells of the tables. For this reason, the most widely used models for spontaneous drug safety data (besides PRR and ROR) are Bayesian. A popular and intuitive Bayesian model is known as the *Gamma-Poisson Shrinkage* (GPS) (DuMouchel, 1999).

Let $N_i = \sum_{l=1}^{n} X_{i,l}$ be the random variable counting the number of total reports with X_i, and let $N_{i,j} = \sum_{l=1}^{n} X_{i,l}Y_{j,l}$ be the random variable counting the number of times drug X_i has been recorded with ADE Y_j. A basic association measure between X_i and Y_j is the spontaneous data equivalent to the SMR, or the indirect standardized mortality rate, known in drug safety under the name *relative reporting rate* (RR),

$$\mathrm{RR}_{i,j} = N_{i,j}/\mathrm{E}\left(N_{i,j}\right) = N_{i,j}/E_{i,j} \tag{1.8}$$

where $E_{i,j}$ is the expected counts of reports with X_i and Y_j observed together and serves as baseline. If $\mathrm{RR}_{i,j} \gg 1$ an association between drug and ADE is supported by the data. Under the null hypothesis of independence between X_i

and Y_j

$$E_{i,j} = \Pr(X_i = 1) \Pr(Y_j = 1)\, n = \frac{N_i N_j}{n}. \tag{1.9}$$

It is reasonable to assume the count of reports as Poisson distributed $N_{i,j} \sim \text{Pois}(\mu_{i,j})$, with $\mu_{i,j} = \lambda_{i,j} E_{i,j}$. The parameter $\lambda_{i,j}$ acts as a multiplicative factor on $E_{i,j}$: if $\lambda_{i,j} > 1$ then $RR_{i,j} \gg 1$.

On $\lambda_{i,j}$, a mixture of gamma distributions can be assumed as a prior distribution. The density function of the mixture is defined as

$$\pi(\lambda_{i,j}; \alpha_1, \beta_1, \alpha_2, \beta_2, P) = g(\lambda_{i,j}; \alpha_1, \beta_1)\, P + g(\lambda_{i,j}; \alpha_2, \beta_2)\,(1 - P) \tag{1.10}$$

with $\alpha_1, \beta_1, \alpha_2, \beta_2 > 0$, $P \in (0, 1)$ and g density of the gamma distribution. Since the family of gamma distributions is conjugated with the Poisson distribution, the posterior distribution is a mixture of gamma itself and the marginal distribution of each $N_{i,j}$ is a mixture of two negative binomial distributions with parameters α and $E_{i,j}/(\beta + E_{i,j})$. The posterior distribution for $\lambda_{i,j}$ can be derived as

$$\begin{aligned}\pi(\lambda_{i,j} | N_{i,j} = n_{i,j}) = {} & g(\lambda_{i,j}; \alpha_1 + n_{i,j}, \beta_1 + E_{i,j})\, Q_n \\ & + g(\lambda_{i,j}; \alpha_2 + n_{i,j}, \beta_2 + E_{i,j})\,(1 - Q_n)\end{aligned} \tag{1.11}$$

with Q_n posterior probability of $\lambda_{i,j}$ coming from the first component of the mixture given $N_{i,j} = n_{i,j}$. Thanks to Bayes' theorem

$$Q_n = \frac{f(n; \alpha_1, \beta_1, E_{i,j})\, P}{f(n; \alpha_1, \beta_1, E_{i,j})\, P + f(n; \alpha_2, \beta_2, E_{i,j})\,(1 - P)} \tag{1.12}$$

where $f(\cdot)$ is the density of a negative binomial distribution with parameters α and $E_{i,j}/(\beta + E_{i,j})$. The final number of hyperparameters is five $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, P)$ which are usually estimated via maximum likelihood

$$\text{L}(\theta) = \prod_{ij} \left[f(N_{i,j}; \alpha_1, \beta_1, E_{i,j})\, P + f(N_{i,j}; \alpha_2, \beta_2, E_{i,j})\,(1 - P) \right]. \tag{1.13}$$

The posterior means of $\lambda_{i,j}$ and $\log \lambda_{i,j}$ are

$$\text{E}(\lambda_{i,j} | N_{i,j} = n_{i,j}) = \frac{\alpha_1 + n}{\beta_1 + E} Q_n + \frac{\alpha_2 + n}{\beta_2 + E}(1 - Q_n) \tag{1.14}$$

and

$$\begin{aligned}\mathrm{E}\left(\log \lambda_{i,j} | N_{i,j} = n_{i,j}\right) = &\left[\Psi(\alpha_1 + n) - \log(\beta_1 + E)\right] Q_n \\ &+ \left[\Psi(\alpha_2 + n) - \log(\beta_2 + E)\right](1 - Q_n) \quad (1.15)\end{aligned}$$

where Ψ is the digamma function (derivate of the logarithm of $\Gamma(x)$). The quantity

$$\begin{aligned}\mathrm{EB}\log 2_{i,j} &= \mathrm{E}\left(\log_2(\lambda_{i,j} | N_{i,j} = n_{i,j})\right) \\ &= \mathrm{E}\left(\log(\lambda_{i,j} | N_{i,j} = n_{i,j})\right) / \log 2\end{aligned} \quad (1.16)$$

is the empirical Bayesian (EB) counterpart of $\log_2(RR_{i,j})$. To obtain a quantity on the same scale as the RR, the exponential transformation of the EB can be computed

$$\mathrm{EBGM}_{i,j} = 2^{\mathrm{EB}\log 2_{i,j}} \quad (1.17)$$

where EBGM stands for *empirical Bayes geometric mean*.

Since the posterior distribution of $\lambda_{i,j}$ is known, its percentile can be calculated to obtain a credibility interval for the EBGM measure. For example, if we are interested in a 95% interval, we can obtain the EBGM lower and upper bound using the 5^{th} and 95^{th} percentile of $\lambda_{i,j}$. Then, if $\lambda_{i,j}^{0.05} > 1$ the data support the hypothesis of an association between the drug and the ADE involved.

$\mathrm{EBGM}_{i,j}$ and $\mathrm{RR}_{i,j}$ have the same behavior and interpretation, the difference being that if $N_{i,j}$ is small, $\mathrm{EBGM}_{i,j}$ decreases regardless of the baseline causing a shrinkage (Church and Hanks, 1990). The introduction of this shrinkage component is crucial and brings several positive features in parameter estimation, especially when the target drug or ADE is not reported frequently (DuMouchel, 1999).

Due to this shrinkage, GPS overcomes a drawback of PRR and ROR. Moreover, it provides an interpretable measure, although the interpretation is not as immediate as its frequentist counterpart. Like PRR and ROR, the main downside of the EBGM measure is that it can be applied to only one drug-ADE pair at a time and is difficult to deploy if the goal is to test drug-drug interactions. The measure can be adjusted for demographic covariates with stratification, but this kind of adjustment is computationally intensive since its implementation requires many calculations (5-dimensional likelihood maximization). To overcome

the problem of taking into account drug-drug interaction, the *Multi-Item Gamma-Poisson Shrinkage* (MGPS) model was later proposed (Szarfman *et al.*, 2002).

1.3.3.2 Bayesian Confidence Propagation Neural Network

A *Bayesian Confidence Propagation Neural Network* (BCPNN) is a two-layer feed-forward neural network that was originally proposed to associate drugs and ADEs (Lansner and Holst, 1996; Bate *et al.*, 1998). In this network, the units in the first stratum correspond to the drug variables and those in the second stratum correspond to the ADE variables. It turns out that the network is transparent, the value of the weights corresponds to a quantity known as *information component* (IC). Given a drug X_i and an adverse drug event Y_j, we can show that

$$\mathrm{IC}_{i,j} = \log_2 \frac{\Pr(X_i, Y_j)}{\Pr(X_i)\Pr(Y_j)} \tag{1.18}$$

and if $\mathrm{IC}_{i,j} > 0$ an association between the drug and the ADE is plausible. For the purpose of computing the information component, the estimation of the neural network can be ignored.

It is reasonable to assume that the four cells in the surrogate contingency table are generated by a multinomial distribution $\mathrm{Mn}(n, p_{11}, p_{10}, p_{01}, p_{00})$ with probability mass function

$$p(n_{11}, n_{10}, n_{01}, n_{00}; n, p_{11}, p_{10}, p_{01}, p_{00}) = \frac{p_{11}^{n_{11}} p_{10}^{n_{10}} p_{01}^{n_{01}} p_{00}^{n_{00}}}{n_{11}! n_{10}! n_{01}! n_{00}!} n!. \tag{1.19}$$

A Dirichlet distribution

$$f(p_{11}, p_{10}, p_{01}, p_{00}; \alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00}) = \frac{p_{11}^{\alpha_{11}-1} p_{10}^{\alpha_{10}-1} p_{01}^{\alpha_{01}-1} p_{00}^{\alpha_{00}-1}}{B(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00})} \tag{1.20}$$

is assumed as a prior distribution on the parameters $p_{11}, p_{10}, p_{01}, p_{00}$, with

$$B(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00}) = \prod_{i=1}^{4} \Gamma(\alpha_i) / \Gamma(\sum_{i=}^{4} \alpha_i). \tag{1.21}$$

Since the Dirichlet distribution is conjugate to the multinomial distribution, the posterior distribution will be $\mathrm{Dir}(\gamma_{11}, \gamma_{10}, \gamma_{01}, \gamma_{00})$ with $\gamma_{i,j} = \alpha_{i,j} + n_{i,j}$. Due to the properties of the conjugate model, the marginal distributions for p_{11}, $p_{1.}$ and

$p_{\cdot 1}$ are beta-distributed and can be easily calculated:

$$p_{11} \sim \text{Beta}(\gamma_{11}, \gamma_{10} + \gamma_{01} + \gamma_{00}), \tag{1.22}$$

$$p_{1\cdot} \sim \text{Beta}(\gamma_{11} + \gamma_{10}, \gamma_{01} + \gamma_{00}), \tag{1.23}$$

$$p_{\cdot 1} \sim \text{Beta}(\gamma_{11} + \gamma_{01}, \gamma_{10} + \gamma_{00}). \tag{1.24}$$

Therefore, under the assumption of independence between $p_{1\cdot}$ and $p_{\cdot 1}$, an approximation of the expected value of the information component can be obtained via a Taylor expansion

$$\text{E}\left(\text{IC}_{i,j}\right) \approx \log_2 \frac{\text{E}\left(p_{11}\right)}{\text{E}\left(p_{1\cdot}\right)\text{E}\left(p_{\cdot 1}\right)} = \log_2 \frac{\gamma_{11}(\gamma_{11} + \gamma_{10} + \gamma_{01} + \gamma_{00})}{(\gamma_{11} + \gamma_{10})(\gamma_{11} + \gamma_{01})}. \tag{1.25}$$

Since the closed form of the distribution of the information component is unknown, to estimate the bounds of its credibility intervals Monte Carlo simulations are usually used.

After the introduction of the information component as a measure of disproportionality, the BCPNN model was improved using the *moderating prior distributions*

$$\alpha_{11} = \frac{1}{2} \quad \alpha_{10} = \frac{1}{2}\frac{n_{\cdot 0} + 0.5}{n_{\cdot 1} + 0.5} \quad \alpha_{01} = \frac{1}{2}\frac{n_{0\cdot} + 0.5}{n_{1\cdot} + 0.5} \quad \alpha_{00} = \frac{1}{2}\frac{n_{\cdot 0} + 0.5}{n_{\cdot 1} + 0.5}\frac{n_{0\cdot} + 0.5}{n_{1\cdot} + 0.5}. \tag{1.26}$$

Norén *et al.* (2006) showed that the use of the moderating prior distribution is equivalent to adding an additional data set where the drug and ADE co-occur half the time and where the marginal frequencies are the same as in the original data. The use of these prior distributions generates an effect similar to GPS shrinkage leading to better estimates.

The information component has the same advantages and drawbacks as the GPS. Both models are more computationally demanding than frequentist methods, but the increasing power of modern computers has induced most pharmacovigilance authorities to integrate them in their phase IV pipeline since they provide better estimation than PRR and ROR (DuMouchel and Harpaz, 2012; Pham *et al.*, 2019).

1.3.4 Regression models and data mining methods

Basic methods (such as PRR and ROR) and Bayesian methods (such as GPS and BCPNN and their extensions) are currently used by pharmacovigilance authorities, such as FDA or EMA. Nevertheless, today modern research is looking for new approaches that, although not developed for disproportionality analysis, fit the data well.

An easy way to tackle spontaneous data analysis is through logistic regression (DuMouchel *et al.*, 2008). Binary variables that indicate the presence of drugs in the reports act as covariates and those that indicate ADE are used as dependent variables in the regression form

$$\log \frac{\Pr(Y_j = 1)}{\Pr(Y_j = 0)} = \alpha + \sum_{i=1}^{p} \beta_i X_i \tag{1.27}$$

with Y_j ADE and $X_1, X_2, \ldots, X_p$ drugs. The associations between drugs and ADE are given by $\beta_1, \beta_1, \ldots, \beta_p$, which can be tested with the well-known regression coefficient tests.

Such an approach has two main advantages. First, it allows us to analyze the associations between all drugs and a given side effect at once; this is of great help to take into account co-prescribed drugs. For example, if drug X_i is associated with ADE Y and drugs X_i and X_j are often administered together, it is likely that the methods based on the surrogate contingency table incorrectly associate X_j with Y. In this case, drug X_j is defined as *innocent bystander* (Dijkstra *et al.*, 2020). Because a regression associates an independent variable with a covariate net of the effect of the other covariates, the risk of generating signals from innocent bystanders is reduced. Second, demographic variables can also be included in the regression without using stratification. Considering drug interactions can be computationally onerous, since including interactions among all (or only some) covariates greatly increases the number of parameters in the regression.

Logistic regression can be combined with GPS to obtain the *regression-adjusted GPS* (RGPS) (DuMouchel and Harpaz, 2012). In RGPS, the usual t-test used to test the significance of the regression parameters is replaced by GPS, making the regression better for low frequencies and unbalanced surrogate contingency tables. RGPS has been shown to perform better than both logistic regression and GPS (Harpaz *et al.*, 2013). However, the computational cost of the model is very high

and the difficulty of analyzing drug-drug interactions remains.

An alternative to logistic regression and RGPS may be the use of the logistic lasso and adaptive lasso (Ahmed *et al.*, 2016). The lasso and adaptive lasso allow for a better selection of drugs significantly associated with ADEs, introduce a shrinkage similar to that introduced by GPS and BCPNN, and can be easily modified to account for interactions. For this reason, these models will be discussed in detail in the next chapter.

In addition to regression models, other methods have been used, such as algorithms developed to perform data mining and machine learning. Tree models and combinations of trees were used: Random Forest, Gradient Boosting, and AdaBoost have been successfully used on spontaneous data, sometimes combined with SMOTE correction (Wei *et al.*, 2020; Chandak and Tatonetti, 2020). However, as much as using tree models has proven useful, the results are not competitive with the models listed above.

1.4 Discussion

The collection and statistical analysis of pharmacovigilance data are somewhat atypical. Even after a drug has received the necessary approvals to be put on the market, it still needs to be monitored to notice any adverse event that went unnoticed during the drug development stages. The way drug safety data are collected during phase IV is spontaneous, which is why ad hoc statistical models have been developed for their analysis. These models, known as disproportionality models, aim to generate a signal when a drug and an adverse event have been reported together a suspected number of times in spontaneous databases.

Disproportionality models are typically classified into basic models and Bayesian models. Models developed in the late 1990s, despite undergoing some modifications over time, are still used by major pharmacovigilance authorities[3]. In particular, the problem of drug-drug interaction analysis remains open since not all models can effectively address it.

[3]Currently, FDA performs data mining using PRR and MGPS and EMA using ROR (FDA, 2018d; EMA, 2016).

Chapter 2

A hierarchical lasso-BIC model for drug-drug interaction detection

2.1 Introduction

As seen in Chapter 1, the statistical methods currently used in the analysis of drug safety data are basic (frequentist) or Bayesian disproportionality models. However, as technology progresses, new methods are being introduced to attempt to make the process of identification of ADEs more efficient. As mentioned above, the alternatives that have been most explored in recent years are machine learning algorithms and data mining methods. Among these methods, one that has a suitable structure for drug safety data analysis is the logistic regression with *lasso* penalization presented in Tibshirani (1996) and developed later with contributions from different authors.

Both stepwise variable selection and lasso regression are commonly used methods for variable selection in logistic regression. However, the best approach for variable selection depends on the specific data and research question at hand. Stepwise selection can help to identify a smaller set of variables that are most predictive of the outcome, which can improve the model's interpretability and reduce overfitting. Moreover, stepwise variable selection may be more appropriate when there are relatively few predictors and a clear hypothesis about which predictors are important. Lasso regression, on the other hand, can help to select a smaller

set of variables that are most predictive of the outcome and can handle collinear predictors better than stepwise selection. In general, lasso regression tends to be more robust and effective for variable selection when there is a large number of predictors or when predictors are highly correlated (Harrell, 2017; Steyerberg *et al.*, 2010).

The main reason why a lasso-penalized regression is here preferred to a stepwise logistic regression is that the lasso introduces a penalty into the parameter estimation that allows, at the same time, variable selection and shrinkage of the coefficients. This procedure results in a better variable selection than the classic logistic regression with stepwise selection of parameters in this specific field. Lasso also turns out to be a computationally efficient solution to fit a logistic regression with high-dimensional data.

First, we will present the lasso method as a tool for disproportionality analysis. Then, we will discuss a new adaptive lasso model proposed by Courtois *et al.* (2021) based on the Bayesian information criterion (BIC) for variable selection (lasso-BIC).

Drug-drug interaction is currently estimated to be responsible for approximately 30% of adverse drug events (Noguchi *et al.*, 2019); therefore, we will extend the model to address the problem of drug-drug interaction using a method specifically designed to identify interactions between variables proposed by Yuan and Lin (2006) and Lim and Hastie (2015).

The performance of the model will be compared with simulation studies. An initial simulation study is presented to evaluate the ability of the lasso method to identify pairs of drug-ADE. Subsequently, the ability of the model to identify not only individual pairs but also possible drug-drug interactions will be investigated in a second simulation study. Finally, an application to spontaneous data from the FAERS database will be presented.

2.2 Methods

2.2.1 Lasso and logistic lasso for spontaneous data

In a general context, let (x_i, y_i) be N pairs where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\intercal$ are covariates and y_i is the response variable for the ith element. Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^\intercal$,

the lasso estimates are defined as

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \qquad \text{s.t.} \qquad \sum_{j=1}^{p} |\beta_j| \leq s \tag{2.1}$$

where $s \geq 0$ is referred to as *hyperparameter* or *regularization parameter*. Notice that we can rewrite the minimization problem in the Lagrangian form

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{2.2}$$

where λ is a hyperparameter with a one-to-one correspondence with s (Hastie *et al.*, 2013).

Since, for each value of s or λ, we can reparametrize the intercept α by standardizing the covariates to obtain $\hat{\alpha} = \bar{y}$, we can then ignore the intercept without loss of generality. The hyperparameter controls the amount of shrinkage applied to the estimates. Let $\hat{\beta}_j^0$ be the ordinary least-squares estimates and let $s_0 = \sum_{j=1}^{p} |\hat{\beta}_j^0|$, some value of $s < s_0$ will cause shrinkage of the solutions toward zero, while some other coefficients will be exactly zero.

The hyperparameter in the lasso regression can be selected in different ways, depending on the data with which the model is used. Some examples of methods for the choice of s are cross-validation, generalized cross-validation, and the use of estimators of prediction error (such as the *Akaike information criterion* or the *Bayesian information criterion*).

In the context of spontaneous pharmacovigilance data, N denotes the number of spontaneous reports and p denotes the number of drugs (used as covariates). Let X be the matrix $N \times p$ that has, as columns, the binary variables that indicate the presence of the drugs in each report and y the binary vector of length N indicating if the adverse drug event of interest is present in the reports. We can fit a logistic lasso regression such as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[-\ell(\beta, y, X) + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{2.3}$$

with $\ell(\beta, y, X)$ log-likelihood of the logistic model

$$\log \frac{\Pr(y_i = 1)}{\Pr(y_i = 0)} = \sum_{j=1}^{p} \beta_j x_{ij}. \tag{2.4}$$

The nonzero estimated coefficients are coupled with the drugs associated with the ADE. Furthermore, since we are not interested in detecting the protective effect of drugs on ADE, we can discard strictly negative coefficients. The value of the remaining strictly positive coefficients indicates the level of association between the drug with which they are associated and the ADE of interest.

There are several strategies for selecting hyperparameters using spontaneous data. The first is classic cross-validation. Each cycle of cross-validation involves partitioning the data into n_f subsets, known as folds. Specifically, the $n_f - 1$ subsets are used to fit the model (the model is estimated based only on these data), and the remaining subset is used to validate the model obtained through a metric chosen to evaluate its predictive performance, such as a measure calculated from a confusion matrix. The procedure is repeated f times, so that all subsets of the data serve both as train sets and as test sets. In a lasso regression, cross-validation is performed for each potential λ value. The best λ will be the one associated with the best performance according to the chosen metric.

Another approach, based on permutations, was proposed by Sabourin *et al.* (2015). Let π_l be a permutation of the set $\{1, 2, \dots N\}$, and let y_{π_l} be the permuted version of y according to π_l. If a lasso regression is fitted using the original covariates as regressors and each permutation of y as a response variable, we can obtain $\lambda_{\mathsf{min}}(y_{\pi_l})$, the smallest value of the hyperparameter such that zero covariates are selected in the regression on the permuted response variable y_{π_l}. Then, we can use the median value of the vector $(\lambda_{\mathsf{min}}(y_{\pi_1}), \lambda_{\mathsf{min}}(y_{\pi_1}), \dots, \lambda_{\mathsf{min}}(y_{\pi_K}))$ as selected λ. K is the maximum number of permutations, the authors suggest setting $K = 20$.

An alternative approach, proposed by Courtois *et al.* (2021), relies on the Bayesian information criterion (BIC). For a candidate λ_0 in the set of all potential λ values, we can compute

$$\mathrm{BIC}_{\lambda_0} = -2\ell_{\lambda_0} + |\hat{\beta}_{\lambda_0} \neq 0| \log(N) \tag{2.5}$$

with ℓ_{λ_0} log-likelihood of a logistic model, whose covariates are those associated

with the strictly positive coefficients of the lasso logistic regression fitted with λ_0. The λ associated with a lower BIC is preferred. The authors showed that this method of selecting the hyperparameter, compared to the other two methods, is particularly suitable for spontaneous pharmacovigilance data, especially in the presence of many drugs (high dimensionality).

2.2.2 Adaptive lasso extension

An optimal variable selection procedure should have the following properties (known as oracle properties): identify the correct subset of true predictors and produce unbiased estimates. Fan and Li (2001) showed that the regular lasso regression does not have these properties, and in some situations, the selection of variables done by the model may be inconsistent. Indeed, it is observed that with the same penalty value for all covariates, the lasso regression tends to over-penalize the most important ones and may produce estimates that are biased.

Adaptive lasso is an alternative implemented to improve variable selection, consisting of the use of *adaptive weights* (AW) to penalize covariates differently from the usual lasso penalization (Zou, 2006). The minimization process is then defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[-\ell(\beta, y, X) + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right] \tag{2.6}$$

where $\ell(\beta, y, X)$ is, once again, the log-likelihood of the model 2.4 and the penalty applied to the jth covariate is defined by $\omega_j \lambda = \lambda_j$. The variable x_j will be penalized according to the weight ω_j: the higher the value, the lower the chances that the variable will be selected.

Courtois *et al.* (2021) proposed an AW suitable for the lasso-BIC procedure. Weights are defined as

$$\omega_j^{lb} = \begin{cases} |\hat{\beta}_j^{BIC}|^{-1} & \text{if} \quad \hat{\beta}_j^{BIC} \neq 0 \\ \infty & \text{if} \quad \hat{\beta}_j^{BIC} = 0 \end{cases} \tag{2.7}$$

where $\hat{\beta}_j^{BIC}$ is the jth coefficient estimated with a lasso-BIC regression (without AW). The use of these weights is further justified by the fact that they lead to an estimator similar to the one recently proposed by Li *et al.* (2021). The authors demonstrate several favorable characteristics of this estimator. Specifically, it

is reliable for variable selection tasks, exhibits an oracle property for parameter estimation, and has a grouping property for highly correlated covariates.

2.2.3 Hierarchical lasso for drug-drug interaction detection

The models mentioned above use lasso regression to mine spontaneous pharmacovigilance data to find associations between drugs and ADEs. However, none of them are extended to include associations between drug-drug interactions and ADE. It is possible to use a lasso regression to select drug-drug interactions associated with the ADE of interest by including the statistical interactions between the covariates in the model and selecting them using the lasso penalty. However, this approach may lead to a great computational cost since the dimensionality would increase disproportionately; by including interactions between variables, the regression matrix becomes

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} & x_{11} : x_{12} & x_{11} : x_{13} & \dots \\ x_{21} & x_{22} & \dots & x_{2p} & x_{21} : x_{12} & x_{21} : x_{13} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Np} & x_{N1} : x_{12} & x_{N1} : x_{13} & \dots \end{bmatrix}. \tag{2.8}$$

A more efficient way is to adopt an interaction selection approach similar to the hierarchical lasso, which can be combined with the adaptive lasso and hyperparameter selection based on the BIC. If we want to determine whether there is an association between drugs x_1, x_2, and an ADE y considering also the interaction $x_{1:2}$, a minimization problem using a constrained grouped lasso logistic loss has to be solved (Yuan and Lin, 2006):

$$\begin{aligned}(\hat{\alpha}_1, \hat{\tilde{\alpha}}_1, \hat{\alpha}_2, \hat{\tilde{\alpha}}_2, \hat{\alpha}_{1:2}) = \underset{\alpha_1, \tilde{\alpha}_1, \alpha_2, \tilde{\alpha}_2, \alpha_{1:2}}{\operatorname{argmin}} & -\Big[y^{\mathsf{T}}\Big(x_1\alpha_1 + x_2\alpha_2 + \Big[x_1 \; x_2 \; x_{1:2}\Big] \Big[\tilde{\alpha}_1 \; \tilde{\alpha}_2 \; \alpha_{1:2}\Big]^{\mathsf{T}}\Big) \\ & - \log\Big(\exp\Big\{x_1\alpha_1 + x_2\alpha_2 + \Big[x_1 \; x_2 \; x_{1:2}\Big] \Big[\tilde{\alpha}_1 \; \tilde{\alpha}_2 \; \alpha_{1:2}\Big]^{\mathsf{T}}\Big\}\Big)\Big] \\ & + \lambda\left(\|\alpha_1\|_2 + \|\alpha_2\|_2 + \sqrt{2\left(\|\tilde{\alpha}_1\|_2^2 + \|\tilde{\alpha}_2\|_2^2\right) + \|\alpha_{1:2}\|_2^2}\right)\end{aligned} \tag{2.9}$$

under the sets of constraints

$$\sum_{i=1}^{2} \alpha_1^i = 0 \qquad \sum_{i=1}^{2} \alpha_2^i = 0 \qquad \sum_{i=1}^{2} \tilde{\alpha}_1^i = 0 \qquad \sum_{i=1}^{2} \tilde{\alpha}_2^i = 0 \tag{2.10}$$

and

$$\sum_{i=1}^{2} \alpha_{1:2}^{ij} = 0 \text{ for a fixed } j \qquad \sum_{j=1}^{2} \alpha_{1:2}^{ij} = 0 \text{ for a fixed } i. \tag{2.11}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. The final estimates are given by $\hat{\beta}_i = \hat{\alpha}_i + \hat{\tilde{\alpha}}_i$ (for $i = 1, 2$) and $\hat{\beta}_{1:2} = \hat{\alpha}_{1:2}$.

Furthermore, we can use *Theorem 1* from Lim and Hastie (2015), which proves that the constrained hierarchical lasso loss (2.9) is equivalent to the simpler unconstrained loss

$$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_{1:2}) = \underset{\beta_1, \beta_2, \beta_{1:2}}{\operatorname{argmin}} - \Big[y^{\mathsf{T}} \left(x_1\beta_1 + x_2\beta_2 + x_{1:2}\beta_{1:2} \right) + \log \left(\exp \left\{ x_1\beta_1 + x_2\beta_2 + x_{1:2}\beta_{1:2} \right\} \right) \Big] + \lambda \left(\|\beta_1\|_2 + \|\beta_2\|_2 + \|\beta_{1:2}\|_2 \right) \tag{2.12}$$

therefore, the optimization problem is much easier to solve. This loss can be generalized to include the desired number of drugs and the interactions between them. To select the best λ, the lasso-BIC procedure can be used.

2.3 Simulations

2.3.1 Simulations without interactions

To test the ability of the hierarchical lasso-BIC to detect associations between both drugs, drug-drug interactions, and ADE, we used some simulated data to mimic the structure of spontaneous drug safety data. Data were simulated via *directed acyclic graph* (DAG), using the algorithm proposed by Dijkstra *et al.* (2020) modified to also generate drug-drug interactions.

First, we investigate the ability of the model to detect drug-ADE pairs (excluding drug-drug interactions) using a plain lasso regression trained with a 10-fold cross-validation as a benchmark comparison. The simulated data set consists of 10000 reports, with a number of drugs equal to 10 and a number of ADEs equal to 10, and the number of pairs of drug-ADE that are associated is 10 out of 100.

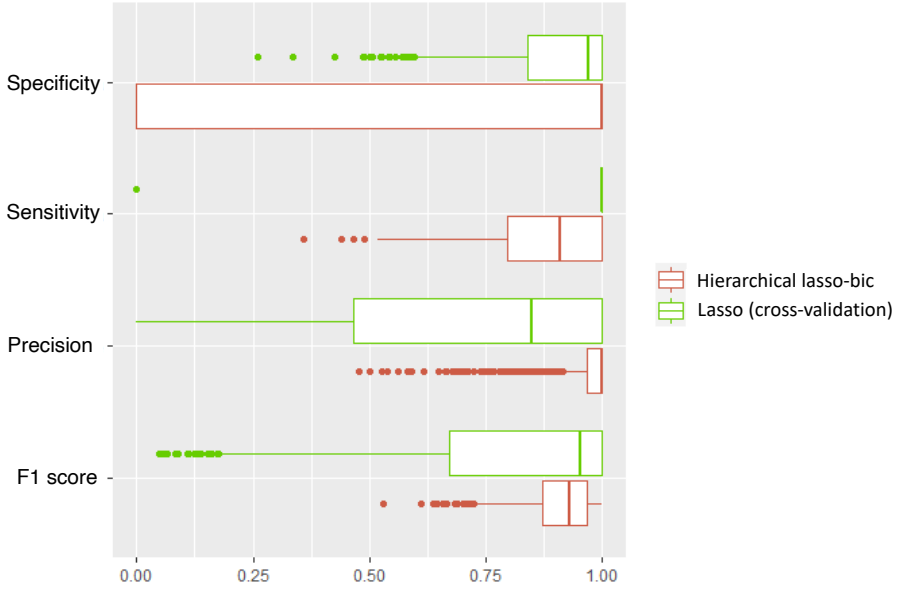

FIGURE 2.1: Boxplots of 100 replications of a 10000 reports simulation with 10 associated drug-ADE couples.

We repeat the entire simulation 100 times. For each simulated ADE, the confusion matrix was calculated to determine the performance of the models. From the confusion matrix, we calculated four statistical measures: specificity, sensitivity, precision, and the F1 score (harmonic mean between precision and sensitivity) defined as follows

$$\begin{aligned}
\text{specificity} &= \frac{\text{true negative}}{\text{true negative} + \text{false positive}} \\
\text{sensitivity} &= \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \\
\text{precision} &= \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \\
\text{F1 score} &= 2\,\frac{\text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}}
\end{aligned}$$

where *false positive* is the number of drug-ADE couples identified by the model but not truly associated, while *false negative* is the number of drug-ADE couples not identified by the model but truly associated.

The results are shown in Figure 2.1. Both models appear to perform well, particularly in terms of precision and the F1 score. The hierarchical lasso-BIC, however, has a wide range of specificity since it has a specificity equal to one in the majority of cases and a specificity equal to zero in fewer cases (31 out of 100).

On the other hand, the lasso with the hyperparameter selected via cross-validation has a sensitivity equal to one in almost every simulation (97 out of 100). Although high sensitivity or specificity may seem like a good indicator of model performance, the fact that it is always zero or one is not always a good sign. This is presumably because the number of negatively associated drug-ADE pairs is difficult to detect, leading to zero false negatives or false positives. However, we can notice that, concerning the precision and F1 score values, the hierarchical lasso-BIC turns out to be the one that gives the best performance and the least variability in the 100 simulations.

2.3.2 Simulations with interactions

To further investigate the ability of hierarchical lasso-BIC regression to detect associations between drug-drug interactions and ADEs, we simulated an additional data set with 10000 reports, a number of drugs equal to 10, and the number of adverse drug events equal to 4. To obtain various scenarios, drugs are associated with ADE in different ways:

- the first adverse drug event (ADE1) is associated with five different drugs but no drug-drug interactions,
- the second and third adverse drug events (ADE2, ADE3) are associated with five different drugs and one interaction between two of them,
- the fourth adverse drug event (ADE4) is associated with five different drugs and all interactions between all the drugs.

As the number of possible drug-ADE pairs has increased significantly compared to the previous simulation and, with it, also the size of the design matrix needed to estimate the models, we fitted only the hierarchical lasso-BIC.

As we can notice from the model performance (Table 2.1) the results trace those obtained in the case without interactions. The specificity once again takes on problematic values, this time focusing only on values close to zero; the measure reaches zero in the scenario in which all interactions between drugs are associated with the ADE. This confirms the difficulty of the model in identifying true negatives, i.e., pairs that are not truly associated. Fortunately, the importance of detecting not associated pairs is not a priority over detecting associate pairs; in

TABLE 2.1: Performance of the hierarchical lasso-BIC model on 10000 simulated reports and four different scenarios of the association of drug and drug-drug interactions.

Adverse drug event	Specificity	Sensitivity	Precision	F1 score
ADE1	0.00	1.00	0.90	0.95
ADE2	0.18	0.92	0.91	0.91
ADE3	0.19	0.91	0.93	0.92
ADE4	0.13	0.94	0.89	0.91

other words, sensitivity is to be considered more important than specificity in this context. The other performance measures are better and indicate a good ability of the model to identify truly associated couples.

These simulations are useful to define whether the hierarchical model is suitable for pharmacovigilance data. However, unfortunately, simulating the complexity of pharmacovigilance data is very difficult, as there are many drugs in spontaneous databases but, at the same time, their frequency is small compared to the total number of records. For example, in the 2019 FAERS data, there are about 58 thousand drugs against more than two million reports (even fewer are the adverse effects, approximately 12 thousand). Simulating such a large number of reports, drugs, and adverse drug events is computationally difficult and, at the same time, would not guarantee the same amount of complexity reached in non-simulated data, especially if simulations need to be repeated several times to control their variability (Dijkstra *et al.*, 2020). Therefore, in addition to knowing the performance of a model on simulated data, it is necessary to test it on real data.

2.4 Spontaneous data application

After verifying through simulation studies that a hierarchical lasso-BIC regression can be used for pharmacovigilance data, we will use it to analyze real spontaneous data. Specifically, the model will be applied to the 2019 FAERS data (1727296 unique records) to test its potential to identify associated pairs of drug and adverse drug events and drug-drug interactions.

To define the model performance, it is necessary to use a gold standard. In the case of pharmacovigilance, this should be a set of drug-ADE pairs whose association is known a priori. Having this kind of gold standard is difficult since

the only solution is to resort to manually curated data sets of established drug-ADE pairs (Ryan *et al.*, 2013). Unfortunately, to date, there are no manually curated data sets that include drug-drug interactions.

For this reason, the *TwoSIDES* data set is used. TwoSIDES is a data collection of drug-drug-ADE relationships, extracted from publicly available health records (Tatonetti *et al.*, 2012b). Although TwoSIDES does not represent a proper gold standard, it still allows us to dispose of associations retrieved from sources external to the spontaneous data. TwoSIDES only contains information on drug-drug interactions, a similar data set for associations between single drugs and ADE is OffSIDES, created by the same authors with the same criteria.

From the 2019 FAERS data, we selected six different adverse drug events: *bronchitis*, *anemia*, *myocardial infarction* and three types of liver disease (*chronic kidney failure*, *kidney injury* and *acute kidney injury*). We search for possible associations between these ADEs and all drugs and drug interactions in the FAERS data. The OffSIDES database is used as the gold standard for the associations between drugs and adverse drug events, while TwoSIDES is used for interactions. The results of the classification are contained in Table 2.2.

TABLE 2.2: Performance of the hierarchical lasso-BIC model in the 2019 FAERS data. OffSIDES and TwoSIDES are used as gold standard for adverse drug event detection (OffSIDES) and drug-drug interaction detection (TwoSIDES).

Adverse drug event	Specificity	Sensitivity	Precision	F1 score
Bronchitis	0.09	0.58	0.49	0.53
Anemia	0.08	0.67	0.53	0.59
Myocardial infarction	0.14	0.48	0.43	0.45
Chronic kindey failure	0.09	0.57	0.48	0.52
Acute kidney injury	0.12	0.46	0.42	0.44
Kidney injury	0.03	0.30	0.40	0.34

The values of all indicators are lower than those obtained with the simulated data. As expected, specificity takes low values, while the other metrics (sensitivity, precision and F1 score) are at higher levels, signaling that the model is less likely to detect a false negatively associated pair than a false positively associated pair.

2.5 Discussion

Analysis of spontaneous data is crucial to discover new associations between drugs and ADEs and becomes particularly important if includes also the effect of drug-drug interactions. Methods currently used by pharmacovigilance authorities (such as the Proportional Reporting Ratio, the Reporting Odds Ratio, the Gamma-Poisson Shrinkage, and the Bayesian Confidence Propagation Neural Network) struggle to account for interactions, which complicate the model fitting, limit their efficiency, and increase the computational cost.

For this reason, we chose to use a penalized regression approach. The lasso penalty, which has already been proven to be useful in the pharmacovigilance context, was used in combination with the hyperparameter selection method based on the BIC index. To account for interactions more efficiently, the model was modified using the hierarchical group-lasso approach.

We tested the model in several simulated scenarios to evaluate its performance; first, data were simulated with associations between single drugs and ADEs (no interaction). In this scenario, the lasso-BIC hierarchical model performs well in terms of sensitivity (which in most cases falls in the interval [0.50, 1]), and precision (which mostly takes values close to one). Specificity takes discordant values (zero or one); this suggests that the model has a reduced ability to classify negatively associated pairs. However, a high F1 score leads to an overall positive evaluation of the model, which has better performance than a lasso model with the hyperparameter selected through cross-validation.

Next, data were simulated including drug-drug interactions. Again, the specificity is at much lower values than the sensitivity, indicating that the model has the same weaknesses regarding false negative detection that were found in the case without interactions. However, the precision and the F1 score are still at high levels, indicating a good fit of the model to the simulated data.

Since the simulated data are not enough to evaluate the performance of the model, data from the FAERS spontaneous database (year 2019) were also used to try to predict associations between drugs, drug-drug interactions and six ADEs (bronchitis, anemia, myocardial infarction, chronic kidney failure, acute kidney injury and kidney injury). In the absence of a gold standard, we used OffSIDES and TwoSIDES databases, which collect data from adverse drug events and drug-drug interactions from publicly available health records.

In this case, the observed model performance is poorer than the one observed in the simulated contexts. Specificity has, again, low values, while sensitivity has moderate values indicating that the model has more difficulty in predicting truly associated pairs than the simulated data context; similarly, the accuracy and the F1 score have smaller values than simulations.

Other attempts to identify drug-drug interactions have been developed over time, using approaches other than the use of spontaneous data, such as the biochemical similarity between drugs (Vilar *et al.*, 2014; Kim and Tatonetti, 2021). However, the approach presented here has the advantage of being regression-based; it is therefore very practical for quickly associating drugs and their interactions with an ADE through the use of spontaneous data. Furthermore, it could be easily adapted to include external variables, such as demographic characteristics of the patient or information on who filed the report, without resorting to stratification procedures.

However, it is worth mentioning that the performance of the model on real data is not high enough to justify its use by pharmacovigilance authorities, which require high-precision and already well-established instruments. So, the hierarchical lasso-BIC model could be used as a tool for an initial filter of pharmacovigilance signals to then implement methods such as BCPNN (or other Bayesian disproportionality methods) for a second identification of significant pairs.

Chapter 3

Improving adverse drug event prediction using biochemical features extracted with ChemBERTa

3.1 Introduction

As mentioned in previous chapters, careful monitoring of drug safety is essential to detect ADEs that may follow drug administration. Many drugs' ADEs are discovered during clinical trial phases, particularly during phases II and III, but the relatively low sample size used in those stages causes a variety of infrequent effects to go unnoticed. Therefore, identifying associations between drugs and adverse events during the post-marketing phase (phase IV) is of paramount importance. Spontaneous databases such as FAERS are imperfect tools, but, despite the limitations (first among all the lack of control data), their analysis remains the main instrument for detecting adverse effects in a post-marketing setting.

Part of the literature is devoted to comparing the performance of the aforementioned models (Bate and Evans, 2009; Harpaz *et al.*, 2013; Ding *et al.*, 2020). In particular, Pham *et al.* (2019) performed a multimodel comparison on the manually curated reference standard set provided by the Observational Medical Outcomes Partnership (Ryan *et al.*, 2013); their findings prove that no model is capable of high performance, reaching a maximum AUROC < 0.70. The reason why many

models do not perform well is due to the fact that the data are heavily biased, as argued extensively in Chapter 1. The spontaneous nature of the data causes a number of biases including lack of controls (and, consequently, scarcity of gold standard databases), under-representation, publicity bias, Weber effect, presence of many confounding variables, and infeasibility of proving causality. The use of alternative and unbiased data sources to supplement spontaneous databases can increase the performance of ADE prediction.

In this chapter, we propose to process data from the biochemical structure of drugs using a deep learning model to support classic spontaneous pharmacovigilance data. Machine learning and deep learning have recently been used to analyze the structure of drugs to predict chemical properties or perform molecule generation and drug discovery, but its use to discover new ADEs is still unexplored (Hirohara *et al.*, 2018; Arús-Pous *et al.*, 2019; Jo *et al.*, 2020; Manne, 2021). Therefore, we extracted valuable features from the drugs of the OMOP reference standard set using two different mapping systems, one based on MACCS fingerprint vectors and one based on SMILES strings. Then, we used these features to predict whether there is an association between drugs and adverse effects contained in the OMOP database.

3.2 Data

3.2.1 OMOP reference set

A sensitive issue related to pharmacovigilance data is related to the lack of gold standard databases. To overcome this problem, manually curated data are used, such as the OMOP reference standard set introduced by Ryan *et al.* (2013). Its latest version consists of 183 drugs and 4 ADEs (Acute Kidney Injury, Acute Liver Injury, Acute Myocardial Infarction, and Gastrointestinal Bleed) extracted from the drug's product labels to collect 165 positively associated and 134 negatively associated drug-ADE pairs. The OMOP reference set has been used to test the new approach presented later.

3.2.2 Alternative use of FAERS data

The FAERS (FDA Adverse Event Reporting System), already fully described in Chapter 1 is the main spontaneous reporting database completely available to the

public and therefore is the most widely used to obtain information on drugs ADE. We downloaded the four quarters of the 2019 FAERS raw data (1727296 unique records) using the `faers.db` R library (Lanera *et al.*, 2022), we chose not to use more recent data to avoid bias related to the COVID-19 pandemic. From the raw data, we selected the drug-ADE pairs found in the OMOP reference set. Then we selected only the most frequent ($> 0.1\%$) adverse events. We treated those events as variables to represent each selected drug in a manner similar to Tatonetti *et al.* (2012a) and Lorberbaum *et al.* (2015). So, we obtained a matrix having as rows 398 drug-ADE pairs, as columns 198 frequent adverse drug events and, in each cell, the value of the relative frequency of the adverse event with respect to each drug.

3.2.3 MACCS vectors

There is more than one way to encode endogenous information from a chemical compound, one of the most widely used is the representation of molecules as binary vectors, also known as chemical fingerprints.

The most popular fingerprint encoding is the Molecular Access System (MACCS), a 166-bit representation of a molecule introduced in the early 2000s (Durant *et al.*, 2002). MACCS fingerprint forms a mathematical representation of a chemical compound allowing, for example, the calculation of dissimilarity measures between molecules. Furthermore, since each bit of the binary vector represents a chemical feature of the compound, MACCS fingerprints can be interpreted as a feature space that describes the drug endogenously (Kuwahara and Gao, 2021). This chemical representation has been used to estimate the absorption, distribution, metabolism, and excretion (ADME) properties of a molecule, but has never been used to predict ADEs (Shen *et al.*, 2010).

We derived fingerprint vectors using `rcdk` 3.6.0 (Guha, 2007), an R interface to the CDK Java framework for chemoinformatics (Steinbeck *et al.*, 2003). This framework allows us to load molecular information from the public ChEMBL database (Gaulton *et al.*, 2017) and obtain a MACCS fingerprint for each drug in the OMOP reference set.

Although a MACCS vector can be interpreted, its understanding is quite complex, because each bit represents a substructure of a molecule, an atom property, or an atomic bond property. For example, bit # 49 denotes the presence of an

electric charge and # 60 denotes the presence of sulfur monoxide. A complete list of the correspondence between bits and molecular features can be found in Table A.1.

3.2.4 SMILES strings

Another way to retrieve endogenous drug data is through the Simplified Molecular Input Line Entry System (SMILES) and consists of a text string that describes the compound (Weininger, 1988). Since SMILES are derived from the 2D graphical representation of a molecule, the representation is not unique and can be done using a number of equivalent SMILES strings.

Therefore, some algorithms were developed to generate unique (also known as *canonical*) SMILES strings. Different algorithms have different canonical SMILES, but they ensure a unique representation of a chemical compound as long as there is consistency in the use of the algorithm (O'Boyle, 2012). The canonicalization algorithm we used is the one developed by the CDK framework; it was chosen because of its popularity and because we also used it for MACCS vector extraction (Steinbeck *et al.*, 2003). For example (Figure 3.1), the ethanol molecule (CH_3CH_2OH) has three possible SMILES (CCO, OCC and C(O)C) but its canonical SMILES obtained via the CDK algorithm is unique (CCO). From now on, any reference

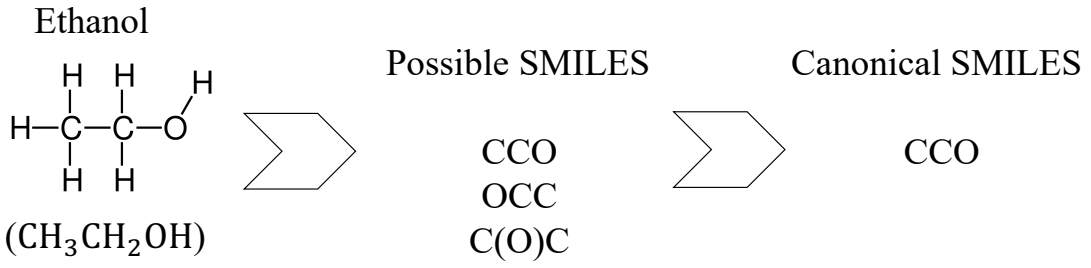

FIGURE 3.1: The ethanol molecules represented using SMILES string.

to SMILES will be intended as canonical SMILEs obtained using the CDK algorithm. Similarly to what we did for the MACCS fingerprints, we obtained the drugs' SMILES from the ChEMBL database using the CDK framework.

3.3 Embedding SMILES strings with ChemBERTa

3.3.1 Transformer models for embedding space representation

SMILES strings can be viewed as a chemical language capable of defining the composition of a drug, its molecules, atoms, and bonds, with a simple dictionary (made up of ASCII symbols) and syntactic rules. Similarly to what is done with language texts, a set of SMILES strings can be processed with a Natural Language Processing (NLP) algorithm to form an embedding space where drugs can be mapped using a set of latent features.

We choose a Bidirectional Encoder Representations from Transformers (BERT) class model for the embedding space representation. Since its introduction in 2018, BERT has established itself as the best model for self-supervised representations of text, outperforming pre-existing NLP models in several tasks (Devlin *et al.*, 2018; Raffel *et al.*, 2020). BERT takes full advantage of a transformer architecture with L number of layers and A number of attention heads (Vaswani *et al.*, 2017). Firstly, a vast unlabelled text corpus is employed to perform a pretrain, then the model is finetuned throughout task-specific labelled data.

During the self-supervised pretraining phase, the model aims to achieve both the masked language and the next sentence prediction tasks. To perform the masked language prediction task, a random sample of the input text is masked and a cross-entropy loss is minimised to predict the masked tokens; similarly, to perform the next sentence prediction task, some negative and positive examples are generated from the text corpora to try to predict, using a binary loss, whether two segments follow each other. The whole model is optimised using the well-known `Adam` optimisation algorithm, details of the parameters chosen for optimisation are contained in the original article (Kingma and Ba, 2014).

3.3.2 ChemBERTa usage to predict adverse drug events

Since its emergence, BERT has been adapted to different text corpora, including clinical text corpora for use in pharmacoepidemiology; a notable case is MTTLADE, a transformer for the extraction of adverse events from clinical text (drissiya El-allaly *et al.*, 2021). The pretraining phase has been modified as well. For example, the recent model RoBERTa does not employ next sentence prediction and focusses only on the masking pattern to obtain better performance (Liu *et al.*, 2019).

We adapted RoBERTa to create an embedding space for the SMILES chemical language. In 2020, RoBERTa has been pre-trained on 100k SMILES strings from the `zinc15` database (Irwin and Shoichet, 2005; Sterling and Irwin, 2015), with $L = 6$ and $A = 12$, resulting in ChemBERTa (v. 1), a transformer model for chemical compounds available on HuggingFace (Chithrananda, 2020). The generated latent space has been used in other biochemical tasks (prediction of brain barrier permeability, clinical toxicity, ability to inhibit HIV replication, stress-response pathway activation), where its performance approaches the baseline results (Chithrananda *et al.*, 2020). Therefore, we assume that this space can also be used in ADE prediction.

After mapping the drugs of the OMOP reference set on a 768-dimensional latent space generated by ChemBERTa, we fine-tuned the model to predict whether or not a drug is associated with an ADE.

3.3.3 Parsing algorithms, software and libraries

Before feeding the SMILES string to ChemBERTa, we performed a text parsing. Parsing is a common pre-processing phase in NLP, useful to separate the raw string of text into smaller components based on some syntactic rules. Classical parsing algorithms cannot be used in this setting because they are based on grammatical linguistic rules. Therefore, we made use of a specific parsing algorithm described in Appendix C of Sidorova and Garcia (2015).

To generate the latent space and map the drugs in it, we use Python 3.8.2 with the `transformers` 4.18.0 library (Wolf *et al.*, 2020). To generate the MACCS fingerprints, analyze the latent features, perform the classification task, and plot the results, we used R 4.1.0 "Camp Pontanezen" (R Core Team, 2022).

3.4 Results

3.4.1 Comparison between MACCS and ChemBERTa features

We used the pre-trained ChemBERTa model available on HuggingFace, fed with the SMILES strings derived from the drugs found in the OMOP reference set. The result is a matrix with, as rows, 398 drug-ADE pairs and, as columns, 768 latent features describing the related chemical compounds. We also extracted MACCS fingerprint vectors from the same chemical compounds, which encode information about drugs in a set of 166 binary features.

To compare the predictive power of these two sets of features, we used them to predict the presence of the individual ADEs listed in the previous section. We trained a support vector machine classifier and evaluated using a 70/30 train/test split, repeated 5 times. To evaluate the model performance, we calculated both the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). We also obtained a 95% confidence interval for the evaluation metrics using 1000 bootstrap replications.

We saw a substantial equivalence between the results obtained from the two sets of features (Figure 3.2, Table A.2). Nevertheless, we chose to proceed considering only the ChemBERTa set of features because they provide greater flexibility because of their continuous nature. The computational times of the classification task are only slightly slower when ChemBERTa features are used, probably because of the higher dimensionality.

3.4.2 ADE prediction with ChemBERTa features and FAERS data

We joined the ChemBERTa feature matrix with the one obtained from the 2019 FAERS data to further increase the classification accuracy. We trained the same model with only the ChemBERTa features and then only the 2019 FAERS data features and compared it with the entire set of features.

Regarding acute myocardial infarction (AMI) ADE, the classification performance obtained with just ChemBERTa features (AUROC: 0.61 [0.38 – 0.80]; AUPRC: 0.66 [0.35 – 0.89]) is lower than the one obtained with FAERS features

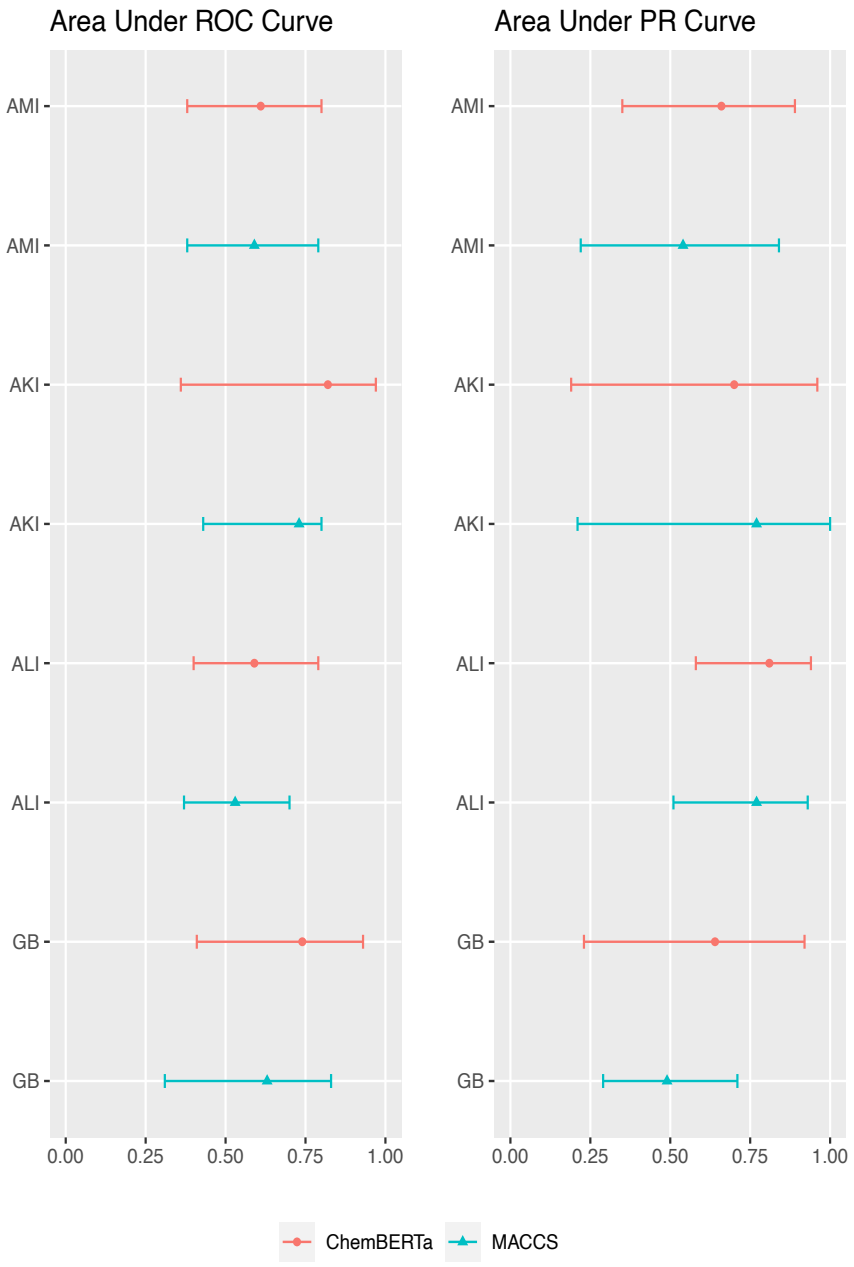

FIGURE 3.2: Result of the support vector machine classifier on the OMOP Gold Standard Database. Comparison between the set of features originated from the ChemBERTa and MACCS fingerprint vectors. AMI: Acute Myocardial Infarction, AKI: Acute Kidney Injury, ALI: Acute Liver Injury, GB: Gastrointestinal Bleed.

(AUROC: 0.81 [0.58 – 1]; AUPRC: 0.83 [0.52 – 0.95]) or the entire set of FAERS + ChemBERTa features (AUROC: 0.78 [0.62 – 0.95]; AUPRC: 0.85 [0.68 – 0.96]). About acute kidney injury (AKI), acute liver injury (ALI), and gastrointestinal bleed (GB) ADEs, the classification performance using the FAERS + ChemBERTa set of features is generally higher and more precise - with smaller confidence intervals - than the one obtained with a single set of features (Figure 3.3, Table A.3). This result is observed in both the AUROC and the AUPRC.

Regarding the AKI ADE, we observed a 0.11 increase in the AUROC and a 0.13 increase in the AUPRC with respect to the use of the ChemBERTa features alone, and a 0.08 increase in the AUROC and a 0.05 increase in the AUPRC with respect to the use of the FAERS data alone. Regarding the ALI ADE, we observed a 0.59 increase in the AUROC and a 0.10 increase in the AUPRC with respect to the use of the ChemBERTa features alone, and a 0.14 increase in the AUROC and a 0.08 increase in the AUPRC with respect to the use of the FAERS data alone. Finally, regarding the GB ADE, we observed a 0.16 increase in the AUROC and a 0.05 increase in the AUPRC with respect to the use of the ChemBERTa features alone, and a 0.09 increase in the AUROC and a 0.02 increase in the AUPRC with respect to the use of the FAERS data alone.

3.5 Discussion

The sole statistical analysis of the spontaneous pharmacovigilance data can be used to predict ADE, but there is room for improvement. Therefore, we made use of endogenous data to incorporate and improve the usual disproportionality analysis models. The data found in the chemical structure of drugs convey implicit information about the compounds of their active ingredients and can be used in ADE prediction task.

Our results suggest two novel conclusions. The first is that data from the biochemical structure of a drug constitute a data source that can be used to predict ADE. Specifically, the use of features obtained from SMILES strings has predictive power similar to that of pharmacovigilance data alone. To extract those features, we mapped the chemical compound of a drug active ingredient to a latent feature space generated by ChemBERTa, a BERT-like transformer model. The latter is that these features, combined with spontaneous data, lead to better performance in ADE forecasting.

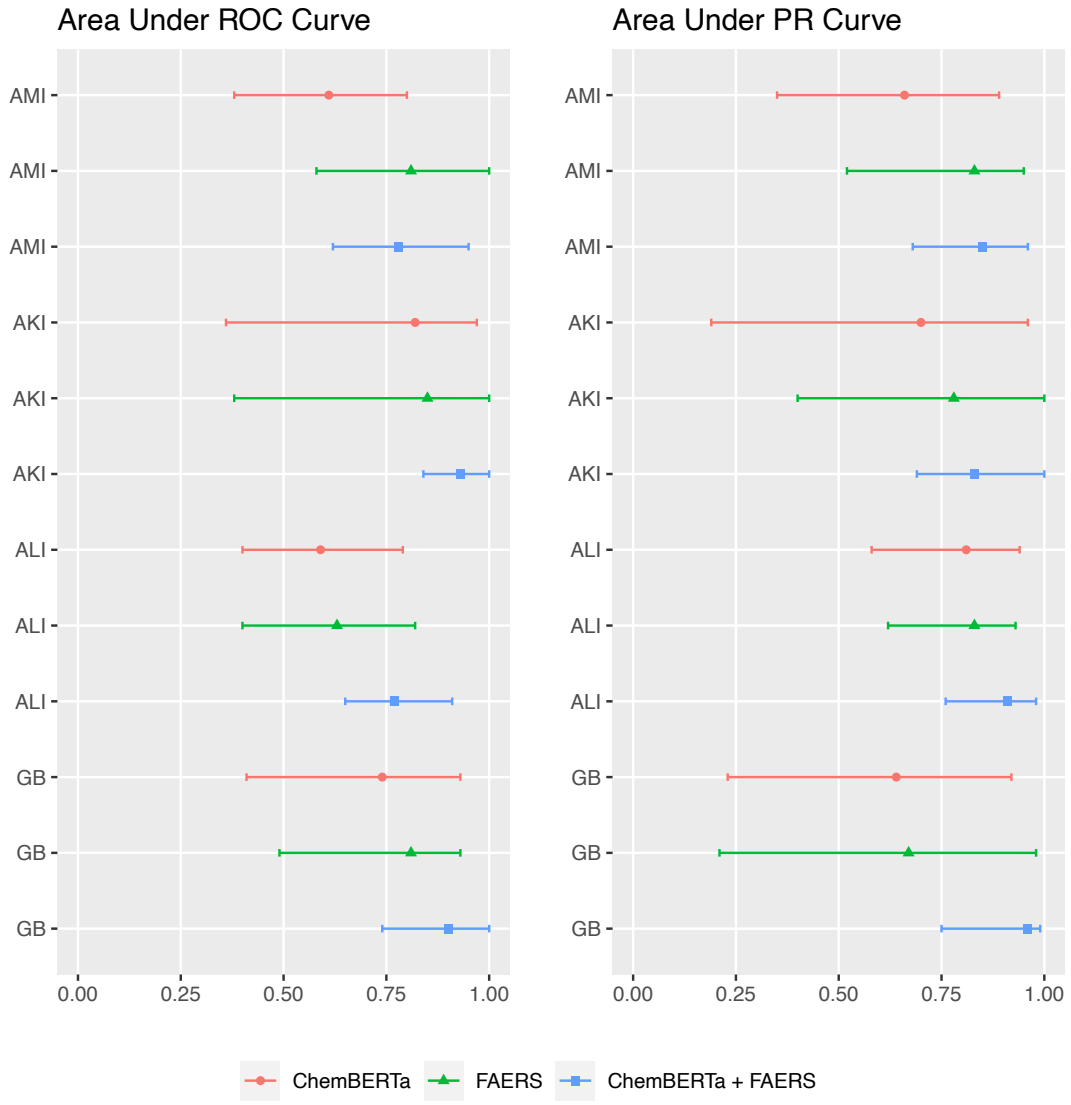

FIGURE 3.3: Result of the support vector machine classifier on the OMOP Gold Standard Database. Comparison between the set of features originated from ChemBERTa, 2019 FAERS data, and both above. AMI: Acute Myocardial Infarction, AKI: Acute Kidney Injury, ALI: Acute Liver Injury, GB: Gastrointestinal Bleed.

Taking these encouraging results into account, many unanswered questions remain. We know that spontaneous data are extremely biased and that data from the biochemical structure of drugs improve the overall data quality, leading to greater accuracy. But with this work it was not possible to determine why this happens, that is, which biases (if any) are mitigated by the presence of endogenous data. Moreover, because of the relatively small size of the OMOP reference set, it was not possible to study in detail the impact of biochemical data on different drug classes to investigate if there are common structures that contribute to the detection of ADE. Finally, we could not create any visualisation of the embedding space obtained, as common techniques (such as PCA or t-SNE) failed to produce any meaningful drug clusters.

This study suggests several possible further steps. A structured pipeline can be designed for pharmacovigilance agencies to systematically include biochemical data in the routine already implemented in ADE discovery. Furthermore, existing Bayesian disproportionality models (such as BCPNN and GPS) could be modified to include these data in their prior assumptions. Moreover, with more labelled data available, it might be possible to investigate which bias is resolved using endogenous data.

Our results show that the use of alternative data sources in pharmacoepidemiology can improve the outcome of the classical disproportionality models based only on the analysis of spontaneous data. In addition, it is particularly useful in this framework, where data can often be underrepresented, to have a way to leverage a constant, always available source of data such as the one obtained from the structure of the chemical compound. We suggest a combined use of the two data sources to achieve better performance in an ADEs detection procedure.

Conclusions

Discussion

This manuscript investigated several aspects of statistical applications in the field of pharmacovigilance. The drug safety monitoring process does not end once the drug has been released on the market, but goes on through continuous monitoring. These surveillance procedures are essential to guarantee the safety of patients and consumers, especially in a context, which is the case of many countries, of population aging and consequent reliance on increased medical and drug treatment.

In Chapter 1, we described the two pillars of post-marketing drug safety: spontaneous data and disproportionality analysis. Spontaneous databases are of paramount importance, as they are the main tools currently used by pharmacovigilance authorities to collect data on drugs during phase IV. However, their availability depends on the spontaneous reporting of adverse drug reactions, and this spontaneity leads to the presence of numerous biases. For this reason, spontaneous data consist of only cases and no controls, they suffer from constant underrepresentation and are subject to publicity bias and several confounding variables. To analyze this particular type of data, specially designed models (disproportionality models) are used. In this thesis, we reviewed the main statistical models currently used for disproportionality analysis, discussed their strengths and weaknesses, and pointed out their limitations.

In Chapter 2, we developed an innovative model for the analysis of pharmacovigilance data with the goal of not only finding new associations between drugs and adverse drug events, but also including drug-drug interactions. The model is a lasso-penalized regression, with interaction selection based on the hierarchical-grouped lasso and hyperparameter selection obtained via BIC index optimization. We tested the performance of the model on both simulated and real data coming from the publicly available spontaneous pharmacovigilance database maintained

by the FDA (FAERS). The results obtained show that the model is capable of identifying adverse events and drug interactions, but its performance is not good enough to be used alone in the analysis of spontaneous data.

The aim of Chapter 3 was to go beyond the paradigms of disproportionality analysis by proposing a new approach for the identification of new associations between drugs and adverse drug events. Since most of the limitations of the disproportionality models arise from the many biases inherent in spontaneous data (Stephenson and Hauben, 2007), we propose the use of endogenous data to improve the prediction of adverse drug events. The data used are extracted from the biochemical structure of the drugs; we used both MACCS vectors and SMILES strings to represent the drug structure. SMILES strings have been innovatively treated as a language and mapped into an embedding space of latent variables using ChemBERTa, a deep learning transformer model. In our analysis, we showed that the use of data extracted with ChemBERTa from the chemical structures of drugs combined with spontaneous data leads to a better performance in the detection of adverse drug events. This procedure, if properly standardized, could be of great help to the pharmacovigilance authorities during phase IV of drug surveillance.

Future directions of research

Some future developments of the research work presented in this thesis ought to be mentioned. Regarding the approach presented in Chapter 2, several improvements could be made to the model. Being based on regression, it could be modified to take into account the demographic variables of the subjects who experienced the adverse drug event, including the proper covariates. Furthermore, to obtain better predictive performance, the BIC-based hyperparameter selection could be replaced with the permutation-based selection mentioned in Section 2.2.1, since it was considered a valid alternative to the BIC selection by some authors in a non-interaction scenario (Sabourin *et al.*, 2015; Courtois *et al.*, 2021).

The novel approach proposed in Chapter 3 paves the way for many more developments in pharmacovigilance. First, it can be extended to also address drug-drug interactions, incorporating, for example, the large-scale prediction procedures discussed by Vilar *et al.* (2014). In addition, another method can be developed to include patient demographic characteristics in predictive features. Similarly to the

hierarchical lasso-bic model, the procedure that includes the biochemical data of the drugs can also be validated using the OffSIDES and TwoSIDES data frames, which contain more data than the reference set already used. The inclusion of biochemical features of drugs in the pharmacovigilance process to take place must be included in an automated workflow. An interesting future development could be the study of a standardized algorithm to be implemented in phase IV that merges spontaneous data with biochemical data when a new data batch arrives in a pharmacovigilance data set. Finally, new Bayesian disproportionality models that include the drug's biochemical structures could be investigated Our proposed procedure ultimately combines the two types of data; a Bayesian model, on the other hand, would have the advantage of incorporating the endogenous information as a priori component, which is very consistent with factual reality.

Appendix

Appendix A

A.1 List of MACCS 116-bit features

TABLE A.1: List of MACCS features with description and corresponding SMARTS (SMILES arbitrary target specification). SMARTS is a language for substructural patterns of chemical compounds, closely related to SMILES.

No.	Description	SMARTS of pattern
1	ISOTOPE	('?',0)
2	ISOTOPE Not complete	('[#103,#104]',0)
3	Group IVa,Va,VIa Periods 4-6	('[Ge,As,Se,Sn,Sb,Te,Tl,Pb,Bi]',0)
4	actinide	('[Ac,Th,Pa,U,Np,Pu,Am,Cm,Bk,Cf,Es,Fm,Md,No,Lr]',0)
5	Group IIIB,IVB	('[Sc,Ti,Y,Zr,Hf]',0)
6	Lanthanide	('[La,Ce,Pr,Nd,Pm,Sm,Eu,Gd,Tb,Dy,Ho,Er,Tm,Yb,Lu]',0)
7	Group VB,VIB,VIIB	('[V,Cr,Mn,Nb,Mo,Tc,Ta,W,Re]',0)
8	QAAA@1	('[!C;!c;!#1]1~*~*~*~*1',0)
9	Group VIII	('[Fe,Co,Ni,Ru,Rh,Pd,Os,Ir,Pt]',0)
10	Group IIa	('[Be,Mg,Ca,Sr,Ba,Ra]',0)
11	4M Ring	('*1~*~*~*~*1',0)
12	Group IB,IIB	('[Cu,Zn,Ag,Cd,Au,Hg]',0)
13	ON(C)C	('[O,o]~[N,n](~[C,c])~[C,c]',0)
14	S-S	('[S,s]-[S,s]',0)
15	OC(O)O	('[O,o]~[C,c](~[O,o])~[O,o]',0)
16	QAA@1	('[!C;!c;!#1]1~*~*~*1',0)
17	CTC	('[C,c]#[C,c]',0)
18	Group IIIA	('[B,Al,Ga,In,Tl]',0)
19	7M Ring	('*1~*~*~*~*~*~*1',0)
20	Si	('[Si]',0)

21	C=C(Q)Q	('[C,c]=[C,c](~[!C;!c;!#1])~[!C;!c;!#1]',0)
22	3M Ring	('*1~*~*~*1',0)
23	NC(O)O	('[N,n]~[C,c](~[O,o])~[O,o]',0)
24	N-O	('[N,n]-[O,o]',0)
25	NC(N)N	('[N,n]~[C,c](~[N,n])~[N,n]',0)
26	C$=C($A)$A	('[C,c]=;@[C,c](@*)@*',0)
27	I	('[I]',0)
28	QCH2Q	('[!C;!c;!#1]~[CH2]~[!C;!c;!#1]',0)
29	P	('P',0)
30	CQ(C)(C)A	('[C,c]~[!C;!c;!#1](~[C,c])(~[C,c])~*',0)
31	QX	('[!C;!c;!#1]~[F,Cl,Br,I]',0)
32	CSN	('[C,c]~[S,s]~[N,n]',0)
33	NS	('[N,n]~[S,s]',0)
34	CH2=A	('[CH2]=*',0)
35	Group IA	('[Li,Na,K,Rb,Cs,Fr]',0)
36	S Heterocycle	('[$(S@*),$(s@*)]',0)
37	NC(O)N	('[N,n]~[C,c](~[O,o])~[N,n]',0)
38	NC(C)N	('[N,n]~[C,c](~[C,c])~[N,n]',0)
39	OS(O)O	('[O,o]~[S,s](~[O,o])~[O,o]',0)
40	S-O	('[S,s]-[O,o]',0)
41	CTN	('[C,c]#[N,n]',0)
42	F	('F',0)
43	QHAQH	('[!C;!c;!#1;H,H2,H3,H4]~*~[!C;!c;!#1;H,H2,H3,H4]',0)
44	OTHER	('?',0)
45	C=CN	('[C,c]=[C,c]~[N,n]',0)
46	BR	('Br',0)
47	SAN	('[S,s]~*~[N,n]',0)
48	OQ(O)O	('[O,o]~[!C;!c;!#1](~[O,o])(~[O,o])~*',0)
49	CHARGE	('[-,-2,-3,-4,+,+2,+3,+4]',0)
50	C=C(C)C	('[C,c]=[C,c](~[C,c])~[C,c]',0)
51	CSO	('[C,c]~[S,s]~[O,o]',0)
52	NN	('[N,n]~[N,n]',0)
53	QHAAAQH	('[!#6;!#1;!H0]~*~*~*~[!#6;!#1;!H0]',0)
54	QHAAQH	('[!#6;!#1;!H0]~*~*~[!#6;!#1;!H0]',0)
55	OSO	('[O,o]~[S,s]~[O,o]',0)
56	ON(O)C	('[O,o]~[N,n](~[O,o])~[C,c]',0)
57	O Heterocycle	('[$(O@*),$(o@*)]',0)

58	QSQ	(’[!C;!c;!#1]~[S,s]~[!C;!c;!#1]’,0)
59	Snot%A%A	(’[S,s]!:*:*’,0)
60	S=O	(’[S,s]=[O,o]’,0)
61	AS(A)A	(’*~[S,s](~*)~*’,0)
62	A$!A$A	(’*@*!@*@*’,0)
63	N=O	(’[N,n]=[O,o]’,0)
64	A$A!S	(’*@*!@[S,s]’,0)
65	C%N	(’[C,c]:[N,n]’,0)
66	CC(C)(C)A	(’[C,c]~[C,c](~[C,c])(~[C,c])~*’,0)
67	QS	(’[!C;!c;!#1]~[S,s]’,0)
68	QHQH	(’[!#6;!#1;!H0]~[!#6;!#1;!H0]’,0)
69	QQH	(’[!C;!c;!#1]~[!#6;!#1;!H0]’,0)
70	QNQ	(’[!C;!c;!#1]~[N,n]~[!C;!c;!#1]’,0)
71	NO	(’[N,n]~[O,o]’,0)
72	OAAO	(’[O,o]~*~*~[O,o]’,0)
73	S=A	(’[S,s]=*’,0)
74	CH3ACH3	(’[CH3]~*~[CH3]’,0)
75	A!N$A	(’*!@[N,n]@*’,0)
76	C=C(A)A	(’[C,c]=[C,c](~*)~*’,0)
77	NAN	(’[N,n]~*~[N,n]’,0)
78	C=N	(’[C,c]=[N,n]’,0)
79	NAAN	(’[N,n]~*~*~[N,n]’,0)
80	NAAAN	(’[N,n]~*~*~*~[N,n]’,0)
81	SA(A)A	(’[S,s]~*(~*)~*’,0)
82	ACH2QH	(’*~[CH2]~[!#6;!#1;!H0]’,0)
83	QAAAA@1	(’[!C;!c;!#1]1~*~*~*~*~*1’,0)
84	NH2	(’[NH2]’,0)
85	CN(C)C	(’[C,c]~[N,n](~[C,c])~[C,c]’,0)
86	CH2QCH2	(’[CH2][!C;!c;!#1][CH2]’,0)
87	X!A$A	(’[F,Cl,Br,I]!@*@*’,0)
88	S	(’[S,s]’,0)
89	OAAAO	(’[O,o]~*~*~*~[O,o]’,0)
90	QHAACH2A	(’[!#6;!#1;!H0]~*~*~[CH2]~*’,0)
91	QHAAACH2A	(’[!#6;!#1;!H0]~*~*~*~[CH2]~*’,0)
92	OC(N)C	(’[O,o]~[C,c](~[N,n])~[C,c]’,0)
93	QCH3	(’[!C;!c;!#1]~[CH3]’,0)
94	QN	(’[!C;!c;!#1]~[N,n]’,0)

95	NAAO	(’[N,n]∼*∼*∼[O,o]’,0)
96	5 M ring	(’*1∼*∼*∼*∼*∼*1’,0)
97	NAAAO	(’[N,n]∼*∼*∼*∼[O,o]’,0)
98	QAAAAA@1	(’[!C;!c;!#1]1∼*∼*∼*∼*∼*∼*1’,0)
99	C=C	(’[C,c]=[C,c]’,0)
100	ACH2N	(’*∼[CH2]∼[N,n]’,0)
101	8M Ring or larger	(’[r8,r9,r10,r11,r12]’,0)
102	QO	(’[!C;!c;!#1]∼[O,o]’,0)
103	CL	(’Cl’,0)
104	QHACH2A	(’[!#6;!#1;!H0]∼*∼[CH2]∼*’,0)
105	A$A($A)$A	(’[!C;!c;!#1]@*(@*)@*’,0)
106	QA(Q)Q	(’[!C;!c;!#1]∼*(∼[!C;!c;!#1])∼[!C;!c;!#1]’,0)
107	XA(A)A	(’[F,Cl,Br,I]∼*(∼*)∼*’,0)
108	CH3AAACH2A	(’[CH3]∼*∼*∼*∼[CH2]∼*’,0)
109	ACH2O	(’*∼[CH2]∼[O,o]’,0)
110	NCO	(’[N,n]∼[C,c]∼[O,o]’,0)
111	NACH2A	(’[N,n]∼*∼[CH2]∼*’,0)
112	AA(A)(A)A	(’*∼*(∼*)(∼*)∼*’,0)
113	Onot%A%A	(’[O,o]!:*:*’,0)
114	CH3CH2A	(’[CH3]∼[CH2]∼*’,0)
115	CH3ACH2A	(’[CH3]∼*∼[CH2]∼*’,0)
116	CH3AACH2A	(’[CH3]∼*∼*∼[CH2]∼*’,0)
117	NAO	(’[N,n]∼*∼[O,o]’,0)
118	ACH2CH2A >1	(’*∼[CH2]∼[CH2]∼*’,1)
119	N=A	(’[N,n]=*’,0)
120	Heterocyclic atom >1	(’[!C;!c;R]’,1)
121	N Heterocycle	(’[(*N*@*),(n@*)]’,0)
122	AN(A)A	(’*∼[N,n](∼*)∼*’,0)
123	OCO	(’[O,o]∼[C,c]∼[O,o]’,0)
124	QQ	(’[!C;!c;!#1]∼[!C;!c;!#1]’,0)
125	Aromatic Ring >1	(’?’,0)
126	A!O!A	(’*!@[O,o]!@*’,0)
127	A$A!O >1	(’*@*!@[O,o]’,1)
128	ACH2AAACH2A	(’*∼[CH2]∼*∼*∼*∼[CH2]∼*’,0)
129	ACH2AACH2A	(’*∼[CH2]∼*∼*∼[CH2]∼*’,0)
130	QQ >1	(’[!C;!c;!#1]∼[!C;!c;!#1]’,1)
131	QH >1	(’[!#6;!#1;!H0]’,1)

132	OACH2A	('[O,o]~*~[CH2]~*',0)
133	A$A!N	('*@*!:[N,n]',0)
134	X (HALOGEN)	('[F,Cl,Br,I]',0)
135	Nnot%A%A	('[N,n]!:*:*',0)
136	O=A>1	('[O,o]=*',1)
137	Heterocycle	('[!C;!c;R]',0)
138	QCH2A>1	('[!C;!c;!#1]~[CH2]~*',1)
139	OH	('[OH,OH2,OH3]',0)
140	O >3	('[O,o]',3)
141	CH3 >2	('[CH3]',2)
142	N >1	('[N,n]',1)
143	A$A!O	('*@*!@[O,o]',0)
144	Anot%A%Anot%A	('*!:*:*!:*',0)
145	6M ring >1	('*1~*~*~*~*~*~*1',1)
146	O >2	('[O,o]',2)
147	ACH2CH2A	('*~[CH2]~[CH2]~*',0)
148	AQ(A)A	('*~[!C;!c;!#1](~*)~*',0)
149	CH3 >1	('[CH3]',1)
150	A!A$A!A	('*!@*@*!@*',0)
151	NH	('[N!H0]',0)
152	OC(C)C	('[O,o]~[C,c](~[C,c])~[C,c]',0)
153	QCH2A	('[!C;!c;!#1]~[CH2]~*',0)
154	C=O	('[C,c]=[O,o]',0)
155	A!CH2!A	('*!@[CH2]!@*',0)
156	NA(A)A	('[N,n]~*(~*)~*',0)
157	C-O	('[C,c]-[O,o]',0)
158	C-N	('[C,c]-[N,n]',0)
159	O>1	('[O,o]',1)
160	CH3	('[CH3]',0)
161	N	('[N,n]',0)
162	Aromatic	('a',0)
163	6M Ring	('*1~*~*~*~*~*~*1',0)
164	O	('[O,o]',0)
165	Ring	('[R]',0)
166	Fragments	('?',0)

A.2 Result of the support vector machine classifier on the OMOP Gold Standard Database.

TABLE A.2: Result of the support vector machine classifier on the OMOP Gold Standard Database. Comparison between the set of features originated from ChemBERTa and MACCS fingerprint vectors.

ADE	Feature set	AUROC	lowAUROC	uppAUROC
AMI	ChemBERTa	0.61	0.38	0.80
AMI	MACCS	0.59	0.38	0.79
AKI	ChemBERTa	0.82	0.36	0.97
AKI	MACCS	0.73	0.43	0.80
ALI	ChemBERTa	0.59	0.40	0.79
ALI	MACCS	0.53	0.37	0.70
GB	ChemBERTa	0.74	0.41	0.93
GB	MACCS	0.63	0.31	0.83
		AUPRC	lowAUPRC	uppAUPRC
AMI	ChemBERTa	0.66	0.35	0.89
AMI	MACCS	0.54	0.22	0.84
AKI	ChemBERTa	0.70	0.19	0.96
AKI	MACCS	0.77	0.21	1
ALI	ChemBERTa	0.81	0.58	0.94
ALI	MACCS	0.77	0.51	0.93
GB	ChemBERTa	0.64	0.23	0.92
GB	MACCS	0.49	0.29	0.71

Table A.3: Result of the support vector machine classifier on the OMOP Gold Standard Database. Comparison between the set of features originated from ChemBERTa, 2019 FAERS data and both above.

ADE	Feature set	AUROC	lowAUROC	uppAUROC
AMI	ChemBERTa	0.61	0.38	0.80
AMI	FAERS	0.81	0.58	1
AMI	ChemBERTa + FAERS	0.78	0.62	0.95
AKI	ChemBERTa	0.82	0.36	0.97
AKI	FAERS	0.85	0.38	1
AKI	ChemBERTa + FAERS	0.93	0.84	1
ALI	ChemBERTa	0.59	0.40	0.79
ALI	FAERS	0.63	0.40	0.82
ALI	ChemBERTa + FAERS	0.77	0.65	0.91
GB	ChemBERTa	0.74	0.41	0.93
GB	FAERS	0.81	0.49	0.93
GB	ChemBERTa + FAERS	0.90	0.74	1
		AUPRC	lowAUPRC	uppAUPRC
AMI	ChemBERTa	0.66	0.35	0.89
AMI	FAERS	0.83	0.52	0.95
AMI	ChemBERTa + FAERS	0.85	0.68	0.96
AKI	ChemBERTa	0.70	0.19	0.96
AKI	FAERS	0.78	0.40	1
AKI	ChemBERTa + FAERS	0.83	0.69	1
ALI	ChemBERTa	0.81	0.58	0.94
ALI	FAERS	0.83	0.62	0.93
ALI	ChemBERTa + FAERS	0.91	0.76	0.98
GB	ChemBERTa	0.64	0.23	0.92
GB	FAERS	0.67	0.21	0.98
GB	ChemBERTa + FAERS	0.69	0.75	0.99

Bibliography

Agrawal, R. and Srikant, R. (1994) Fast algorithms for mining association rules. Volume 1215, pp. 487–499. Citeseer.

Ahmed, I., Dalmasso, C., Haramburu, F., Thiessard, F., Broët, P. and Tubert-Bitter, P. (2010) False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics* **66**, 301–309.

Ahmed, I., Pariente, A. and Tubert-Bitter, P. (2016) Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research* **27**, 785–797.

Ang, P. S., Chen, Z., Chan, C. L. and Tai, B. C. (2016) Data mining spontaneous adverse drug event reports for safety signals in singapore–a comparison of three different disproportionality measures. *Expert opinion on drug safety* **15**, 583–590.

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H. and Engkvist, O. (2019) Randomized smiles strings improve the quality of molecular generative models. *Journal of Cheminformatics* **11**, 71.

Banda, J. M., Evans, L., Vanguri, R. S., Tatonetti, N. P., Ryan, P. B. and Shah, N. H. (2016) Data descriptor: A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific Data* **3**, 1–11.

Bate, A. and Evans, S. J. W. (2009) Quantitative signal detection using spontaneous adr reporting. *Pharmacoepidemiology and Drug Safety* **18**, 427–436.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A. and Freitas, R. M. D. (1998) A bayesian neural network method for adverse drug

reaction signal generation. *European journal of clinical pharmacology* **54**, 315–321.

Bates, D. W., Cullen, D. J., Laird, N., Petersen, L. A., Small, S. D., Servi, D., Laffel, G., Sweitzer, B. J., Shea, B. F. and Hallisey, R. (1995) Incidence of adverse drug events and potential adverse drug events: implications for prevention. *Jama* **274**, 29–34.

Brown, E. (2007) Medical dictionary for regulatory activities (meddra). *Pharmacovigilance: Second Edition* **20**, 168–183.

Chandak, P. and Tatonetti, N. P. (2020) Using machine learning to identify adverse drug effects posing increased risk to women. *Patterns* **1**, 100108.

Charatan, F. (2001) Bayer decides to withdraw cholesterol lowering drug. *BMJ: British Medical Journal* **323**, 359.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. and Blaschke, T. (2018) The rise of deep learning in drug discovery. *Drug Discovery Today* **23**, 1241–1250.

Chithrananda, S. (2020) Chemberta: Training a bert-like transformer model for masked language modelling of chemical smiles strings. date accessed: 19/08/2022. `https://huggingface.co/seyonec/ChemBERTa-zinc-base-v1`.

Chithrananda, S., Grand, G. and Ramsundar, B. (2020) Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885* .

Church, K. and Hanks, P. (1990) Word association norms, mutual information, and lexicography. *Computational linguistics* **16**, 22–29.

Courtois, E., Tubert-Bitter, P. and Ahmed, I. (2021) New adaptive lasso approaches for variable selection in automated pharmacovigilance signal detection. *BMC medical research methodology* **21**, 1–17.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* .

Diaby, V., Almutairi, R. D., Chen, Z., Moussa, R. K. and Berthe, A. (2021) A pharmacovigilance study to quantify the strength of association between the

combination of antimalarial drugs and azithromycin and cardiac arrhythmias: implications for the treatment of covid-19. *Expert review of pharmacoeconomics outcomes research* **21**, 159–168.

Dijkstra, L., Garling, M., Foraita, R. and Pigeot, I. (2020) Adverse drug reaction or innocent bystander? a systematic comparison of statistical discovery methods for spontaneous reporting systems. *Pharmacoepidemiology and Drug Safety* **29**, 396–403.

Ding, Y., Markatou, M. and Ball, R. (2020) An evaluation of statistical approaches to postmarketing surveillance. *Statistics in Medicine* **39**, 845–874.

DuMouchel, W. (1999) Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. *The American Statistician* **53**, 177–190.

DuMouchel, W., Fram, D., Yang, X., Mahmoud, R. A., Grogg, A. L., Engelhart, L. and Ramaswamy, K. (2008) Antipsychotics, glycemic disorders, and life-threatening diabetic events: a bayesian data-mining analysis of the fda adverse event reporting system (1968–2004). *Annals of Clinical Psychiatry* **20**, 21–31.

DuMouchel, W. and Harpaz, R. (2012) Regression-adjusted gps algorithm (rgps). *Oracle Health Sci* .

DuMouchel, W. and Pregibon, D. (2001) Empirical bayes screening for multi-item associations. pp. 67–76.

Durant, J. L., Leland, B. A., Henry, D. R. and Nourse, J. G. (2002) Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* **42**, 1273–1280.

Edwards, I. R. and Aronson, J. K. (2000) Adverse drug reactions: definitions, diagnosis, and management. *The Lancet* **356**, 1255–1259.

drissiya El-allaly, E., Sarrouti, M., En-Nahnahi, N. and Alaoui, S. O. E. (2021) Mttlade: A multi-task transfer learning-based method for adverse drug events extraction. *Information Processing Management* **58**, 102473.

EMA (2016) Eudravigilance system overview. date accessed: 15/11/2022. https://www.ema.europa.eu/en/human-regulatory/research-development/pharmacovigilance/eudravigilance/eudravigilance-system-overview.

Evans, S. J. W., Waller, P. C. and Davis, S. (2001) Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety* **10**, 483–486.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**, 1348–1360.

FDA (2018a) Clinical research phase studies. accessed: 15/11/2022. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research.

FDA (2018b) Fast track, breakthrough therapy, accelerated approval, priority review. date accessed: 15/11/2022. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/fast-track-breakthrough-therapy-accelerated-approval-priority-review.

FDA (2018c) Questions and answers on fda's adverse event reporting system (faers). date accessed: 15/11/2022. https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers.

FDA (2018d) Data mining at fda. date accessed: 15/11/2022. https://www.fda.gov/science-research/data-mining/data-mining-fda-white-paper#disproportionality.

FDA (2022) Providing submissions in electronic format postmarketing safety reports - guidance for industry.

Fram, D. M., Almenoff, J. S. and DuMouchel, W. (2003) Empirical bayesian data mining for discovering patterns in post-marketing drug safety. pp. 359–368.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E. *et al.* (2017) The chembl database in 2017. *Nucleic acids research* **45**(D1), D945–D954.

Guha, R. (2007) Chemical informatics functionality in r. *Journal of Statistical Software* **18**, 1–16.

Harpaz, R., DuMouchel, W., LePendu, P., Bauer-Mehren, A., Ryan, P. and Shah, N. H. (2013) Performance of pharmacovigilance signal-detection algorithms for the fda adverse event reporting system. *Clinical Pharmacology and Therapeutics* **93**, 539–546.

Harrell, F. E. (2017) Regression modeling strategies. *Bios* **330**(2018), 14.

Hartnell, N. R. and Wilson, J. P. (2004) Replication of the weber effect using postmarketing adverse event reports voluntarily submitted to the united states food and drug administration.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2013) *The Elements of Statistical Learning*. Springer. ISBN 978-0-387-84857-0.

Hirohara, M., Saito, Y., Koda, Y., Sato, K. and Sakakibara, Y. (2018) Convolutional neural network based on smiles representation of compounds for detecting chemical motif. *BMC Bioinformatics* **19**, 526.

Irwin, J. J. and Shoichet, B. K. (2005) Zinc a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **45**, 177–182.

Jo, J., Kwak, B., Choi, H.-S. and Yoon, S. (2020) The message passing neural networks for chemical property prediction on smiles. *Methods* **179**, 65–72.

Kim, C. and Tatonetti, N. (2021) Prediction of adverse drug reactions associated with drug-drug interactions using hierarchical classification. *bioRxiv* p. 2021.02.10.430512.

Kingma, D. P. and Ba, J. (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kuwahara, H. and Gao, X. (2021) Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach. *Journal of Cheminformatics* **13**, 27.

Lanera, C., Belloni, P. and Guidone, N. (2022) faers.db: Faers database in r. R package version 0.0.0.9001.

Lansner, A. and Holst, A. (1996) A higher order bayesian neural network with spiking units. *International Journal of Neural Systems* **7**, 115–128.

Li, N., Peng, X., Kawaguchi, E., Suchard, M. A. and Li, G. (2021) A scalable surrogate l0 sparse regression method for generalized linear models with applications to large scale data. *Journal of Statistical Planning and Inference* **213**, 262–281.

Lim, M. and Hastie, T. (2015) Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics* **24**, 627–654.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Lorberbaum, T., Nasir, M., Keiser, M. J., Vilar, S., Hripcsak, G. and Tatonetti, N. P. (2015) Systems pharmacology augments drug safety surveillance. *Clinical Pharmacology Therapeutics* **97**, 151–158.

Lu, Z. (2009) Information technology in pharmacovigilance: Benefits, challenges, and future directions from industry perspectives. *Drug, healthcare and patient safety* **1**, 35.

Manne, R. (2021) Machine learning techniques in drug discovery and development. *International Journal of Applied Research* **7**, 21–28.

McBride, W. G. (1961) Thalidomide and congenital abnormalities. *Lancet* **2**, 90927–90928.

Montastruc, J. L., Sommet, A., Bagheri, H. and Lapeyre-Mestre, M. (2011) Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database. *British Journal of Clinical Pharmacology* **72**, 905–908.

Nebeker, J. R., Barach, P. and Samore, M. H. (2004) Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. *Annals of internal medicine* **140**, 795–801.

Neha, R., Subeesh, V., Beulah, E., Gouri, N. and Maheswari, E. (2019) Existence of notoriety bias in fda adverse event reporting system database and its impact on signal strength. *Hospital Pharmacy* **56**, 152–158.

Noguchi, Y., Nagasawa, H., Tachi, T., Tsuchiya, T. and Teramachi, H. (2019) Signal detection of oral drug-induced dementia in chronic kidney disease patients using association rule mining and bayesian confidence propagation neural network. *Die Pharmazie-An International Journal of Pharmaceutical Sciences* **74**, 570–574.

Norén, G. N., Bate, A., Orre, R. and Edwards, I. R. (2006) Extending the methods used to screen the who drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in medicine* **25**, 3740–3757.

O'Boyle, N. M. (2012) Towards a universal smiles representation - a standard method to generate canonical smiles based on the inchi. *Journal of Cheminformatics* **4**, 22.

Park, J. J. H., Hsu, G., Siden, E. G., Thorlund, K. and Mills, E. J. (2020) An overview of precision oncology basket and umbrella trials for clinicians. *CA: A Cancer Journal for Clinicians* **70**, 125–137.

Pham, M., Cheng, F. and Ramachandran, K. (2019) A comparison study of algorithms to detect drug–adverse event associations: Frequentist, bayesian, and machine-learning approaches. *Drug Safety* **42**, 743–750.

van Puijenbroek, E. P., Bate, A., Leufkens, H. G. M., Lindquist, M., Orre, R. and Egberts, A. C. G. (2002) A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety* **11**, 3–10.

Qureshi, Z. P., Seoane-Vazquez, E., Rodriguez-Monguio, R., Stevenson, K. B. and Szeinbach, S. L. (2011) Market withdrawal of new molecular entities approved in the united states from 1980 to 2009. *Pharmacoepidemiology and drug safety* **20**, 772–777.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67.

Ross, S. D. (2001) Drug-related adverse events: a readers' guide to assessing literature reviews and meta-analyses. *Archives of internal medicine* **161**, 1041–1046.

Rothman, K. J., Lanes, S. and Sacks, S. T. (2004) The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and drug safety* **13**, 519–523.

Ryan, P. B., Schuemie, M. J., Welebob, E., Duke, J., Valentine, S. and Hartzema, A. G. (2013) Defining a reference set to support methodological research in drug safety. *Drug safety* **36**, 33–47.

Sabourin, J. A., Valdar, W. and Nobel, A. B. (2015) A permutation approach for selecting the penalty parameter in penalized model selection. *Biometrics* **71**, 1185–1194.

Shan, W., Hong, D., Zhu, J. and Zhao, Q. (2020) Assessment of the potential adverse events related to ribavirin-interferon combination for novel coronavirus therapy. *Computational and Mathematical Methods in Medicine* **2020**, 1391583.

Shen, J., Cheng, F., Xu, Y., Li, W. and Tang, Y. (2010) Estimation of adme properties with substructure pattern recognition. *Journal of Chemical Information and Modeling* **50**, 1034–1041.

Shojania, K. G., Duncan, B. W., McDonald, K. M. and Wachter, R. M. (2002) Safe but sound: patient safety meets evidence-based medicine. *Jama* **288**, 508–513.

Sidorova, J. and Garcia, J. (2015) Bridging from syntactic to statistical methods: Classification with automatically segmented features from sequences. *Pattern Recognition* **48**, 3749–3756.

Silverstein, C., Brin, S. and Motwani, R. (1998) Beyond market baskets: Generalizing association rules to dependence rules. *Data mining and knowledge discovery* **2**, 39–68.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. and Willighagen, E. (2003) The chemistry development kit (cdk): an open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences* **43**, 493–500.

Stephenson, W. P. and Hauben, M. (2007) Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiology and drug safety* **16**, 359–365.

Sterling, T. and Irwin, J. J. (2015) Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling* **55**, 2324–2337.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. and Kattan, M. W. (2010) Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* **21**(1), 128.

Szarfman, A., Machado, S. G. and O'neill, R. T. (2002) Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fda's spontaneous reports database. *Drug safety* **25**, 381–392.

Tatonetti, N. P., Fernald, G. H. and Altman, R. B. (2012a) A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *Journal of the American Medical Informatics Association* **19**, 79–85.

Tatonetti, N. P., Ye, P. P., Daneshjou, R. and Altman, R. B. (2012b) Data-driven prediction of drug effects and interactions. *Science Translational Medicine* **4**, 125ra31 LP – 125ra31.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

Tubert, P., Bégaud, B., Péré, J.-C., Haramburu, F. and Lellouch, J. (1992) Power and weakness of spontaneous reporting: a probabilistic approach. *Journal of clinical epidemiology* **45**, 283–286.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser and Polosukhin, I. (2017) Attention is all you need. *Advances in neural information processing systems* **30**.

Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C. and Tatonetti, N. P. (2014) Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature protocols* **9**, 2147–2163.

Wallenstein, E. J. and Fife, D. (2001) Temporal patterns of nsaid spontaneous adverse event reports. *Drug safety* **24**, 233–237.

Waller, P., Puijenbroek, E. P. V., Egberts, A. C. G. and Evans, S. (2004) The reporting odds ratio versus the proportional reporting ratio:'deuce'. *Pharmacoepidemiology and drug safety* **13**, 525–526.

Weber, J. C. P. (1984) Epidemiology of adverse reactions to nonsteroidal antiinflammatory drugs. *Advances in Inflammation Research. 1984.* .

Wei, J., Lu, Z., Qiu, K., Li, P. and Sun, H. (2020) Predicting drug risk level from adverse drug reactions using smote and machine learning approaches. *IEEE Access* **8**, 185761–185775.

Weininger, D. (1988) Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28**, 31–36.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R. and Funtowicz, M. (2020) Transformers: State-of-the-art natural language processing. pp. 38–45.

Xiao, C., Li, Y., Baytas, I. M., Zhou, J. and Wang, F. (2018) An mcem framework for drug safety signal detection and combination from heterogeneous real world evidence. *Scientific Reports* **8**, 1806.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67.

Zorych, I., Madigan, D., Ryan, P. and Bate, A. (2013) Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Statistical Methods in Medical Research* **22**, 39–56.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 1418–1429.

Pietro Belloni

CURRICULUM VITAE

Personal Details

Date of Birth: May 1, 1994
Place of Birth: Padova, Italy
Nationality: Italian

Contact Information

University of Padova
Department of Statistical Sciences
via Cesare Battisti, 241
35121 Padova, Italy
phone: +39 049 827 4174
e-mail: belloni@stat.unipd.it
website: www.pietrobelloni.com

Current Position

Since October 2019; (expected completion: January 2023)
PhD Student in Statistical Sciences, University of Padova
Thesis title: New approaches on statistical modeling for drug safety data
Supervisor: Prof. Giovanna Boccuzzo (University of Padova)
Co-supervisor: Prof. Nicholas Tatonetti (Columbia University)

Research interests

- Epidemiology
- Pharmacoepidemiology
- Medical Statistics
- Social Statistics

Education

October 2016 – March 2019
Master degree (*laurea magistrale*) in Statistical Sciences
University of Padova, Department of Statistical Sciences
Title of dissertation: "Staging Cancer Through Text Mining of Pathology Records"
Supervisor: Prof. Giovanna Boccuzzo
Final mark: 107/110

October 2013 – September 2016
Bachelor degree (*laurea triennale*) in Statistics, Economics and Finance
University of Padova, Department of Statistical Sciences
Title of dissertation: "An Approach to Combine Biomarkers"
Supervisor: Prof. Gianfranco Adimari
Final mark: 102/110

Visiting periods

February 2022 – August 2022
Columbia University
New York City, NY, USA
Supervisor: Prof. Nicholas Tatonetti

August 2017 – January 2018
Aarhus University
Aarhus, Denmark
Supervisor: -

Further education

September 2022
Summer School in Social Statistics
University of Naples Federico II

October 2021
Statistical aspects of Deep Neural Networks
University of Milano - Bicocca

September 2021
Summer School in Social Statistics
University of Padova

Work experience

April 2019 – September 2019
European Institute of Oncology
Research fellow

Computer skills

- **R**: data wrangling, statistical modeling, data mining & machine learning, package development, data visualization (*advanced skills*)
- **Python**, **SAS**: statistical modeling (*basic skills*)
- **MS Office Suite** (*advanced user*)
- **Latex**, **Markdown**: document creation and editing (*proficient skills*)
- Machine learning libraries: **Keras**, **TensorFlow**, **Thorch** (*basic skills*)
- Version control: **GitHub** (*usual user*)

Language skills

- **Italian**: mother tongue
- **English**: fluent (Listening: C1; Reading: C1; Speaking: C1; Writing: B2)

Publications

P. Belloni, M. Silan, G. Cuman. "Fake news spreading and sentiment of Italians during the first COVID-19 lockdown." In: M. Misuraca, G. Scepi, M. Spano (eds). *Proceedings of the 16th conference on statistical analysis of textual data.* Vadistat Press (2022) https://doi.org/10.13140/RG.2.2.27575.39846

M. Tagliabue, R. De Berardinis, **P. Belloni**, S. Gandini, D. Scaglione, F. Maffini, R. A. Mirabella, S. Riccio, G. Giugliano, R. Bruschini, F. Chu, M. Ansarin. "Oral tongue carcinoma: prognostic changes according to the updated 2020 version of the AJCC/UICC TNM staging system." *Acta Otorhinolaryngologica Italica* (2022) https://doi.org/10.14639/0392-100X-N2055

S. Burlina, M. G. Dalfrà, **P. Belloni**, S. Ottanelli, F. Mecacci, G. Mello, A. Lapolla. "Can the First Fasting Plasma Glucose Test in Pregnancy Predict Subsequent Gestational Complications?" *International Journal of Endocrinology* (2022) https://doi.org/10.1155/2022/9633664

R. De Berardinis, M. Tagliabue, **P. Belloni**, S. Gandini, D. Scaglione, F. Maffini, S. Margherini, S. Riccio, G. Giugliano, R. Bruschini, F. Chu, M. Ansarin. "Tongue cancer treatment and oncological outcomes: The role of glossectomy classification." *Surgical Oncology* (2022) https://doi.org/10.1016/j.suronc.2022.101751

D. Alterio, M. Augugliaro, M. Tagliabue, R. Bruschini, S. Gandini, L. Calabrese, **P. Belloni**, L. Preda, F. A. Maffini, G. Marvaso, A. Ferrari, S. Volpe, M. A. Zerella, O. Oneta, I. Turturici, A. Ombretta, F. Ruju, M. Ansarin, R. Orecchia, B. A. Jereczek-Fossa. "The T-N tract involvement as a new prognostic factor for PORT in locally advanced oral cavity tumors." *Oral Diseases* (2021) https://doi.org/10.1111/odi.13885

M. Tagliabue, **P. Belloni**, R. De Berardinis, F. Chu, S. Zorzi, C. Fumagalli, L. Santoro, S. Chiocca, M. Ansarin. "A systematic review and meta-analysis of the prognostic role of age in oral tongue cancer." *Cancer Medicine* (2021) https://doi.org/10.1002/cam4.3795

P. Gnagnarella, S. Raimondi, V. Aristarco, H. Johansson, F. Bellerba, F. Corso, S. P. De Angelis, **P. Belloni**, S. Caini, S. Gandini. "Ethnicity as modifier of risk for Vitamin D receptors polymorphisms: Comprehensive meta-analysis of all cancer sites". *Critical reviews in oncology/hematology* (2020) https://doi.org/10.1016/j.critrevonc.2020.103202

P. Belloni, G. Boccuzzo, S. Guzzinati, I. Italiano, C. R. Rossi, M. Rugge, M. Zorzi. "Staging Cancer Through Text Mining of Pathology Records." In: Mariani P., Zenga M. (eds). *Studies in Classification, Data Analysis, and Knowledge Organization.* Springer, Cham (2019) https://doi.org/10.1007/978-3-030-51222-4_4

Conference presentations

Identification of neighborhood clusters on data balanced by a poset-based approach
StatisticAll, Italian festival of statistics and demography. Treviso, 2022 (talk)

Not only pharmacovigilance data: the use of biochemical features in drug adverse events prediction
Statistical methods and models for complex data. Padova, 2022 (poster)

Fake news spreading and sentiment of Italians during the first COVID-19 lockdown
CNDSS 2021. Naples, 2021 (talk)

Impact of lockdown on the everyday life of Italians: some evidence from a sentiment analysis
XLIV AIE Congress – Epidemiology for Epidemic. Online, 2020 (poster)

Fecal microbiota, Vitamin D, serum biomarkers of inflammation and diet: a complex interactive network influencing colorectal cancer risk and prognosis
Royal Statistical Society International Conference. Belfast, 2019 (poster)

Extraction of cancer information from pathology clinical records using text mining
International Conference on Data Science and Social Research. Milan, 2019 (talk)

Teaching experience

February 2023 – May 2023 (programmed)
Issues and methods for population and society
Master degree (*laurea magistrale*) in Statistical Sciences
Lecturer, 21 hours
University of Padova
Instructor: Prof. Margherita Silan

December 2022 – January 2023
Laboratory of data analysis for social research
Master degree (*laurea magistrale*) in Cultural Pluralism, Social Change and Migrations
Teacher in charge, 21 hours
University of Padova
Instructor: Prof. Pietro Belloni

Thesis co-supervision

Enrico Bovo, 2023 (expected)
Mortality profile in the ULSS 6 Euganea: the role of social and environmental factors in a spatial clustering
Master degree (*laurea magistrale*) in Statistical Sciences, University of Padova
Supervisor: Prof. Giovanna Boccuzzo

Matteo Cortivo, 2022
Data mining approaches for pharmacovigilance applied to adverse event reporting systems
Master degree (*laurea magistrale*) in Statistical Sciences, University of Padova
Supervisor: Prof. Giovanna Boccuzzo

References

Giovanna Boccuzzo
Full Professor
University of Padova
Department of Statistical Sciences
Via Cesare Battisti 241, Padova, Italy
boccuzzo@stat.unipd.it

Nicholas Tatonetti
Associate Professor
Columbia University
Department of Biomedical Informatics
622 W 168th St., New York City, NY, USA
nick.tatonetti@columbia.edu