Università degli Studi di Milano - Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di laurea in Informatica

A Data Analytics Framework for Medical Prescription Pattern Dynamics

Relatore: Prof. Francesco Archetti
Correlatore: Prof. Paolo Mariani
Tutor aziendale: Dott. Gaia Arosio

Relazione della prova finale di:
Ilaria Battiston
Matricola 816339

Anno Accademico 2018-2019

To Ilaria Giordani, Paolo Mariani, Gaia Arosio, Francesco Archetti, Antonio Candelieri and the whole CMR team, for accompanying and helping me through this journey of personal and professional growth.

To Marina Avitabile, Fabio Stella, Alberto Dennunzio and Alberto Leporati, for using their passion to inspire and encourage me to pursue my favourite subjects.

To unixMiB, for everything we accomplished during these years, the powerful connections we built and the skills we acquired, being a solid and tight group.

To Martina Restelli, Marco D'Antino, Luca Toma and Laura Nesossi, for being great friends and giving me emotional support with endless patience.

Abstract

Research on prescription pattern dynamics exploits medical information to highlight the issues concerning the growing antibiotic resistance among the Italian country, using a subset of data regarding the Campania region.

After a clear understanding of the problem and the available information, along with its quality and feasible outcomes, a first analytical approach allows reconstruction of diagnosis and prescriptions global trending, from which is possible to extract patient journeys and underline changes.

A deeper insight on antibiotics verifies national studies on overprescription, giving additional information on general practitioners' behaviour, brands popularity and seasonality of prescriptions, obtained with time series analysis and clustering.

Contents

1 Summary **4**

2 Application domain **5**

3 Goals definition **16**

4 Data description **25**

5 Information loss **31**

6 Global analytics **36**

7 Patient journey **47**

8 Antibiotic trends analytics **55**

9 k-means approach **68**

10 Graph analysis **78**

11 Assessments of results and future directions **93**

Chapter 1

Summary

This research work aims to support existing national studies on the medicine consumption in Italy and the growing problem of antibiotic resistance, through exploratory analysis and clustering on healthcare data.

After explaining the Italian sanitary system along with functioning of the pharmaceutical products' market, an overview of healthcare data is given, underlying its collection, potentialities and issues in the first chapters.

The available dataset comprehends medical data on patients with related prescriptions and diagnoses in the past 18 years in the Italian region Campania.

Since health data consists in large quantities of information, an in-depth understanding of its structure and value is required, assessing the risk of progressive information loss and constructing subsets of clean values to perform statistics in chapters 4-5.

Global analysis in chapter 6 allows to identify anomalies to focus on, focussing on antibiotic consumption in chapter 8: prescription trends are highlighted in relation with national research, explaining potential causes of unusual patterns.

The patient-doctor relationship is reconstructed in chapter 7, resulting in a consistent dataset of medical history to apply clustering algorithms and analyse co-prescriptions.

Clustering is performed according to prescribing habits based on doctors and diagnoses, detecting communities of prescribers with similar conduct, medicines preference and over-prescribing.

Chapters 9-10 focus on various approaches of clustering based on antibiotic prescriptions and comparing results, using the subset of information and the obtained trends.

The scope of the research is limited to prescription patterns, yet healthcare data has many other applications: future additional work is required to give a complete insight, considering a wider spectrum of medicines and relating them to diagnoses.

Chapter 2

Application domain

This chapter gives a general idea of the antibiotic resistance problem, describing it on a national level in the Italian Sanitary system context.

It then provides an exhaustive description of the current practices of healthcare data collection, classification and standardisation, introducing its analytical purposes.

2.1 Antibiotic resistance and misuse

Antimicrobial resistance is a rising global problem which threatens the effective prevention and treatment of an ever-increasing range of infections caused by bacteria, parasites, viruses and fungi [1].

Microorganisms exposed to antimicrobial drugs develop the ability to *defeat substances designed to kill them*, making infections persist in the body due to the unsuccessful action of agents.

This issue threatens public health, causing higher healthcare costs to treat patients and potentially compromising surgeries and chemotherapy results due to the ineffectiveness of antibiotics. No one can completely avoid the risk of resistant infections, but some people are at greater risk than others (for example, people with chronic illnesses) [2].

Antimicrobial resistance occurs naturally over time, usually through genetic changes. However, **the misuse and overuse of antimicrobials is accelerating this process**. In many places, antibiotics are overused and misused in people and animals, and often given without professional oversight [1].

Infections such as the common cold or sore throats are often countered with antibiotics, which have no effect against viruses and could as well put patients at the risk of suffering adverse reaction [3].

Those critical issues are worsened by the fact that at the moment there are no antibiotic drugs in development, and no trials in the past 30 years led to commercialisation of new antimicrobial medicines [4].

The success rate for clinical drug development is, in fact, low; historical data show that, generally, only 1 in 5 infectious disease products that enter human testing (phase 1 clinical

trials) will be approved for patients [5].

Therefore, to minimise the development of resistance, **contributing factors must be reduced**, optimising the use of drugs. This work requires effective surveillance and follow-up of consumption, at a local and national level.

To be effective in the long run, work to optimise use of antibiotics must influence the prescribing practices of individual physicians. The goal is "rational use", i.e. the correct patient receives the correct antibiotic at the correct dose and for the correct duration of treatment, in accordance with evidence-based guidelines. Over-prescribing should be avoided, preferring a conservative approach without resulting in under-prescribing [6].

Such work should be carried out close to the prescriber, something which also requires high resolution prescription data, down at the level of **individual prescribers**.

2.2 Italian National Sanitary Service

The **Italian National Sanitary Service** (SSN) consists the complex of *functions*, *activities* and *healthcare services* offered by the State. It is based on **subsidiarity**, a general principle of the European Union law, stating that a central authority should have a subsidiary function, performing only those tasks which cannot be performed at a more local level[1].

SSN is articulated in different responsibility levels divided among the State, Regions, institutions and organisations, along with private structures and the **Health Ministry**, which coordinates the national sanitary plan.

Citizens benefit of healthcare services paying a related **ticket** [7], that represents the established way to contribute to expenses. It is used for:

- Specialist examinations;
- First-aid help in non-emergency situations;
- Thermal care.

Sanitary assistance on the territory is free, in fact general practitioners' visits are exempted from payment of tickets (while additional services such as certificates may require a fee). A **general practitioner** (GP) is a way for the citizen to access SSN in terms of global care and health education.

GPs manage types of illness that present in an *undifferentiated way* at an *early stage of development*, which may then require urgent intervention. Their duties are not confined to specific organs of the body, and they have particular skills in treating people with multiple health issues [8].

According to WONCA (World Organization of Family Doctors), they are responsible of supplying *integrated and continuative care*. Some fundamental skills and activities [8] to pursue this goal are:

- Communication with patients;

[1] Oxford Dictionary

- Management of the practice;
- Clinical tasks;
- Problem solving;
- Holistic modelling.

In Italy, GPs have a crucial role in preventing diseases, understanding the symptoms, introducing patients to therapeutical approaches and monitoring the development or regression of illnesses [9]. Primary aid is guaranteed through diagnoses, prescription, therapy and basic levels of assistance.

Visits take place at the medical office, according to methodologies established by the doctor, generically on appointment or at the patient's domicile. After a visit, a common task of the general practitioner in agreement with SSN is prescribing drugs or further medical checks through a prescription.

Prescriptions are healthcare documents that govern the plan of care for an individual patient [10], consisting in written authorization to purchase a specific medicine from a pharmacy.

Drugs are dispensed according to the national guidelines, with the following regimes:

- **OTC** (Over The Counter), not subject to medical prescription;
- **RR** (Repeatable Prescription), to be sold after presenting a prescription;
- **RNR** (Non-Repeatable Prescription), to be sold after presenting a prescription which has to be renewed each time;
- RL (Limitative Prescription), used in hospitals and clinics;
- RMR (Ministerial Tracing Prescription), for narcotics and psychotropic substances.

2.2.1 Pharmaceutical products and prescriptions

The **Italian Pharmacy Agency** (AIFA) is the public institution for regulatory activity of medicines in Italy. Its main duty consists in all the activities related to the regulatory process of drugs, registering and authorising them to commercialization with a negotiated price.

Before a drug can be sold in pharmacies among the Italian territory, it must have received the **authorisation** by AIFA: each medicine is subject to checks regarding chemical, pharmaceutical, biological, toxicological, and clinical aspects and researches, to see if it satisfies security and efficacy standards [11].

After passing all the quality controls, a product is assigned an unique **AIC code** for it to be identified with its specific details.

AIFA guarantees uniformity and equity of the pharmaceutical system, coordinating national and local authorities such as Regions. Procedures are based on *safeness*, *innovation* and *accessible healthcare*: pharmaceutical costs are regulated in a context of financial compatibility with industry competitiveness, while pursuing goals of economic balance and population' safeguard.

From the year 2000 every form of participation to sanitary expenses from citizens has been abolished [7], yet most Regions have introduced special **drugs classes** with a fixed quota for each medical prescription or package, to remedy the profit deficit.

A medicine only sold after presenting a medical prescription is defined *ethical*. Currently existing categories of ethical drugs according to reimbursement are [12]:

- A, entirely at the expense of SSN, comprehending essential medicines and the ones for chronic diseases;
- H, at the expense of SSN only in a hospital environment;
- C, fully paid by citizens according to the brand price.

Complete lists for classes A and H are publicly available, and each single drug or active principle can be individually looked up on the official AIFA database. Prescriptions can fall into any category, while OTCs follow a C regime.

Generic drugs have reduced costs, and to have a brand product the citizen must explicitly ask for it and pay the additional price [7].

2.2.2 Antibiotic consumption in Italy

Italy is the European nation with the *highest antibiotic resistance mortality* ($\sim$10.000 deaths every year), in terms of infection caused by resistant bacteria [13]. The public health relevance of this issue is high, because of the considerable epidemiologic on the population (increment of morbidity and mortality) and the heavy social burden (workdays loss, usage of diagnostic procedures) [4].

AIFA published the national report "Antibiotic use in Italy 2017", providing consumption and expenses data at national and regional level. The report allows to identify areas of potential inappropriateness and promote a comparison between Regions with the aim of improving prescriptions and antibiotic use.

Resistance to methicillin goes up to 38%, which is a non-negligible emergency; furthermore, people affected by MRSA (methicillin-resistant Staphylococcus aureus) have 64% more probability of death compared to individuals who didn't develop an antibiotic resistant infection.

Antibiotic resistant infections are diffused in all age ranges, but particularly affect the extremes (individuals aged 75 and more years).

The problem is aggravated by the fact **antibiotics are drugs sold over the counter** in Italy, therefore individuals are able to get medicines according to their own judgement without a prescription by an expert — even if their diagnosis doesn't involve an antibiotic cure.

AIFA recommends actions to raise awareness among the population [15]:

- Using antibiotics only if prescribed by a doctor;
- Completing the therapy without interrupting it;
- Avoiding taking several antibiotics in short spans of time.

Patients are not the only ones contributing to antibiotic resistance: another aspect of it is the **lack of ethical prescriptions** from general practitioners, with the poor control of sanitary system workers by the authorities.

For instance, 75% of cases is due to infections related to sanitary assistance, and the need to contrast those actions, especially in patient care structures, is rising.

More than 85% of doses have been provided by SSN (National Sanitary Service), both from private or public pharmacies and public sanitary structures. 90% of those doses are due to prescriptions from general practitioners or paediatricians.

Geographical analysis confirms a **greater consumption in the southern and centre regions** compared to the north. **Campania** is the region with most antibiotic usage and the highest average costs, although values have suffered a slight decline since 2016.

There is a marked increase in consumes between seasons, in particular between the summer months and the winter ones, related to the flu symptoms peaks observed in winter during the years. A relevant part of seasonal prescriptions could be avoided, since a considerable amount of infections spreading in the cold months is viral.

OMS, in the 20th WHO List of Essential Medicines (2017), groups antibiotics in three categories, with the aim of guiding their prescribing:

1. *Access*, which should always be used as a first-choice treatment (penicillins with broad spectrum);
2. *Watch*, antibiotics with higher risk to induce resistance (cephalosporins, macrolides);
3. *Reserve*, antibiotics to be given in serious diseases after other unsuccessful alternatives, with exclusive hospital usage.

The antibiotic classes with the most use prevalence in Italy are penicillins, comprehending **amoxicillin and beta-lactose inhibitors** (clavulanic acid), macrolides and cephalosporins, having a major detach of usage from all the other antibiotics.

The association between amoxicillin and clavulanic acid suggests a probable **over-use** in cases where only prescribing amoxicillin could have a minor impact on resistance with a selective spectrum of actions.

Furthermore, 40% of prescriptions in 2017 did not involve a first-choice (*access*) antibiotic, with a growing tread from North to South.

Equivalent drugs usage is still a minor percentage: 70,1% of consumption is composed by brand drugs with expired license, and Campania is again one of the regions with the lowest incidence of generic medicines.

Inappropriate consuming and antibiotic abuse can be countered only with a global *one health* approach, promoting interventions for responsible use of medicines in all fields [4].

The focal point of monitoring and implementing initiatives to improve prescriptive appropriateness is represented by **general practitioners**, due to those being the main source of antibiotic prescriptions.

2.3 Healthcare data

Healthcare data is defined as t*hat information used to provide, manage and/or report the services used across the entire healthcare system.* Its origin is the encounter between a patient and a provider, who will record the service rendered, the conditions of the service, patient information and clinical information [16].

To trim costs and maximise productivity and value, healthcare entities are turning to their data and decision support tools to validate quality initiatives: data may not be captured completely or accurately, with incorrect information or keys and missing referrals.

Recent advances in health information technology have expanded the scope of health data. Advances in health information technology have fostered the *eHealth* paradigm, which has expanded the collection, use, and philosophy of health data.

eHealth is a recent healthcare practice, defined as *a set of technological themes in health* today, more specifically based on commerce, activities, stakeholders, outcomes, locations, or perspectives [17].

Health information is understood and appraised among Electronic Health Records, prescription methodologies, health knowledge managements and information systems. The main concern is the confidentiality of the data, standardised through coding techniques.

Electronic Health Records is a widespread application of big data in medicine. Patients have their own digital record which includes demographics, medical history, allergies, laboratory test results etc. Records are shared via secure information systems and are available for providers from both public and private sector [18].

2.3.1 In Italy

Italy has different typologies of data collection, distinguished by both the source and the practical applications. Production is gathered in three different channels (public, pharmaceutical and private), with eventual derivations.

Public data

Public data is produced by *public institutions*, using the national registries forms submitted by taxpayers and personal data of services users.

Ministero della Salute (Health Ministry) is the national entity to promote health as a fundamental right, and owns an open data system containing information about chronic and rare diseases, public structures and medical devices classification.

The purpose of public datasets is mainly *informative*, since data is already presented in an aggregated form.

Personal data of patients is collected by **Agenzia delle Entrate**, having the duty of collecting taxes and revenues. Sanitary expenses are detracted from taxes, including partial costs of pharmaceutical products.

Agenzia delle Entrate produces a preventive and consumptive balance sheet every year, elaborating results of various duties and concessions.

Another national source of healthcare information is **ISTAT** (National Institute of Statistics), collecting data, microdata and metadata to make analysis and classifications. Its main areas concern suicides, road accidents and mental health.

Public data consists in a small part compared to all the existing datasets concerning Italy, which are currently private.

Pharmaceutical data

Pharmaceutical data has as its source the activity of pharmacies, interacting with the market while buying and selling both OTC products and prescribed ones.

Promofarma is the main commercial society dedicated to prescriptions data collecting, offering:

- Studies on pharmaceutical expense and profitability;
- Outsourcing services for local network handling.

It organises information, to then aggregate and send it to public services, publishing private reports on consumption trends. Analysis is made according to various granularity levels.

Data is electronically transmitted to Ministero dell'Economia e delle Finanze (Economy and Finance Ministry) and Ministero della Salute (Health Ministry), and all pharmacies must monthly submit records[2]. About 400 millions of prescriptions are sent every year.

Private service providers

Private healthcare data is produced by individuals (i.e. general practitioners) or structures agreeing to use a proprietary storage system, submitting information to companies.

Instances of private healthcare industry services providers are:

- IMS Health[3], one of the largest vendors of data, collecting electronic health records and prescription data:
  - IQVIA[4] integrates data science with human science, automatizing health processes with information services;
- Complion[5], a document management and workflow platform for clinical research sites;
- Millenium[6], offering a suite of proprietary systems of family medicine applications on a national level.

[2]Promofarma official website
[3]IMS Health Italy official website
[4]IQVIA official website
[5]Complion official website
[6]Millewin official website

Those channels make a preprocessing action, outputting derived data to be linked with evidence and analytics for concrete decisions making.

Results are sold to research centres and pharmaceutical companies to make custom analytics based on specific needs and desired outcomes.

2.4 Data classification

Analytical processes require precise **standards** concerning data integrity and availability, since decision-making processes must be reliable and transparent.

Healthcare data, during the collection and parsing processes, needs to be classified according to national or international **standards**, in a way that information cannot be misinterpreted and fields can be mapped to wider categories. Some standardization exists in the way data is captured.

2.4.1 ICD

ICD (International Classification of Diseases) is the foundation for global identification of health trends and statistics, and the international standard for reporting *diseases* and *health conditions*. It is the diagnostic classification standard for all clinical and research purposes [1], maintained by the **World Health Organization** (WHO).

ICD defines the universe of diseases, disorders, injuries and other related health conditions, listed in a comprehensive, hierarchical fashion that allows for:

- Easy storage, retrieval and analysis of health information for evidenced-based decision-making;
- Sharing and comparing health information between hospitals, regions, settings and countries;
- Data comparisons in the same location across different time periods [29].

The ICD is revised periodically and is currently in its 10th version, while in Italy the latter has only been adopted to classify death causes [30]. Until a complete upgrade, the diagnostic system is standardised using **ICD-9**.

ICD-9

ICD-9, officialized in 1978, is the 9th version of the International Classification of Diseases, ordering diseases and traumas in groups according to defined criteria, allowing a common language to code information related to morbidity and mortality for comparisons and statistics [30].

ICD-9-CM is an adaption to ICD-9 used in Italy to assign diagnostic and procedure identifiers, providing additional morbidity detail. Diagnoses are extended with codes for surgical, diagnostic and therapeutical procedures.

It is composed by a group of three digits followed by up to two optional ones adding further details, separated with a dot:

1. The first number (001-999) represents the **macro-category** based on the type of the disease or the injury they describe;
2. The second group provides more specific information about the *type*, *location*, and *severity* of the disease or injury.

There are also two sets of alphanumeric codes in ICD-9-CM. E-codes describe external causes of injury, while V-codes describe factors that influence health status and/or describe interactions with health services [31].

Example: 414.12, falling in diseases of the circulatory system (390-459), specifically into ischaemic heart disease (410-414).

- 414: other forms of chronic ischaemic heart disease;
- 1: aneurysm and dissection of heart;
- 2: dissection of coronary artery.

2.4.2 ATC

The **Anatomical Therapeutic Chemical Classification System** (ATC) is a pharmaceutical products coding system adopted worldwide, controlled by World Health Organisation.

Medicines are divided into different groups according to to the **organ** or **system** on which they act, their **therapeutic intent** or nature, and the drug's **chemical characteristics**. Different brands share the same code if they have the same active principle and indications.

One single drug can have multiple codes, since ATC also comprehends instructions regarding administration or use, and a code can represent more than one active ingredient.

The ATC classification is composed by seven alphanumeric symbols split into five adjacent hierarchical sets, defined **levels** and having the following structure [32]:

1. One letter indicating the anatomical/pharmacological main group among 14;
2. Two digits indicating the therapeutic subgroup;
3. One letter indicating the pharmacological subgroup;
4. One letter indicating the chemical subgroup;
5. Two digits indicating the chemical substance.

The ATC system also includes many *defined daily doses* (DDDs). This is a measurement of the assumed average maintenance dose per day for a drug used for its main indication in adults.

Alterations in ATC classification can be made when the main use of a product has clearly changed, and when new groups are required to accommodate new substances or to achieve

better specificity in the groupings. Changes are twice annually submitted to the WHO official database.

Example: C03BA12.

- C: cardiovascular system;
- 03: diuretics;
- B: low-ceiling diuretics, excluding thiazides;
- A: sulfonamides, plain;
- 12: Clorexolone.

2.4.3 AIC

AIC represents **authorisation to admission to commerce** of a medicine, and is a 9-digit code conceded by AIFA after a careful check of safeness and efficacy. It is a sort of "identity card" of the drug, since it contains the *essential characteristics* defining it[7]. Different brands are identified by different AIC.

AIC establishes:

- The drug name;
- Its composition (active principle);
- Description of the fabrication method;
- Therapeutical instructions, contraindications and adverse reactions;
- Dosage and way of administration;
- Conservation measures;
- Characteristics of the product and its packaging;
- Brochure;
- Risks evaluation for the environment.

Every possible modification of those characteristics involves a further request for authorization to AIFA. Official databases such as Fedefarma contain information related to every product, along with its eventual expire of authorisation.

2.4.4 Privacy concerns

Healthcare is a highly regulated industry where the ability to achieve, maintain and efficiently demonstrate **regulatory compliance** improves the organizations' overall security posture, allowing them to focus on patient care and improved outcomes.

When implemented with complimentary solutions, data classification can play a pivotal role in managing regulated data with precision, effectiveness and a level of efficiency that

[7] AIFA definition

allows healthcare organizations the opportunity to properly focus on their core mission [22].

The **General Data Protection Regulation** (GDPR) recognises data concerning health as a special category of data, and provides a definition for health data for data protection purposes. Specific safeguards for personal health data and for a definitive interpretation of the rules that allows an effective and comprehensive protection of such data have been addressed by the GDPR.

Processes that foster innovation and better quality healthcare need robust data protection safeguards in order to maintain the trust and confidence of individuals in the rules designed to protect their data[8].

2.5 Healthcare analytics

Healthcare analytics is a field of growing importance which helps understanding the statistical perspective of results collected in the healthcare area.

Extracting insights can be a complex challenge: health big data gives a huge *volume* and *variety* of information, therefore accessing the resources in a quick way is necessary.

Other issues to deal with are *veracity*, *validity* and *viability*, fundamental characteristics to ensure reliable and relevant analytics. Checking for integrity and quality can be difficult to verify without domain knowledge [21].

One of the possible applications, given the concerning issue of antibiotic resistance, is explaining the situation through **statistics** and **trends**, obtaining practical results alongside theoretical scientific research.

Big data can assess the appropriateness of prescribing through the existing classification systems, comparing their patterns within time ranges. A general view is essential to identify specific unusual changes, extending national studies with a perspective centred on products and reduced geographical areas.

There are five main necessary fields for analysis:

1. **Spatial data**, in different granularity levels;
2. **Personal data** of patients;
3. **Temporal data**, for time-series analysis;
4. **Pharmacotherapeutic data**, classifiable according to different identification codes;
5. **Diagnostic data**, for cross-validation of diagnoses.

The two main risks encountered while doing analysis are **information loss** and **inappropriate prescribing**, that compromise the quality of statistics. Data may be incomplete, biased or filled with noise: another goal of analytics is to contrast incompleteness and incorrectness, obtaining coherent and clear results.

[8] General Data Protection Regulation

Chapter 3

Goals definition

This chapter implements theoretical knowledge into a practical context, applying healthcare analytics to the Italian sanitary system through specific methodologies and illustrating results to expect.

It gives an overview of the potential problems and concerns, such as information loss and privacy, and ways to address them.

3.1 Methodologies

Reported work is performed within the project "Territorial Analysis on Local Antibiotic Resistance through Data Analytics Techniques", by Consorzio Milano Ricerche, concerning medical data in the region Campania.

This part of the research is specifically focussed on the **doctor-patient relationship**: person-centred care concerning *diagnoses* and *prescriptions*, analysing their changes according to habits, physiological aspects, time and geographical area of both interested parts.

There are several external factors influencing market trends of medicines, for instance the advent of **generic drugs**, having the same active principle and bioequivalence with brand medicines, but lower prices thanks to their dissociation from pharmaceutical companies.

The concept of generic drug has been introduced in Italy in 1995, to be legally formalised in 1996 [14], yet it has initially been perceived by general practitioners and pharmacists as a mere instrument to save money at the expense of quality [27].

Only in the past 10 years, advertisement efforts have been made to make the population aware of the strict quality checks and the reliability of generic drugs, starting a slow switching process. If sales of a brand product decrease, then, it might have been replaced with its equivalent.

This is only an instance of event which cannot be predicted using the raw data: information about *market withdrawals*, *advertisement* or *economic availability* need to be **cross-checked** with analytical results.

Considering a **wide time span** is essential to have an overall idea of data quality, information loss and potentiality of available resources. All those factors contribute to progressive rescaling of the dataset:

1. Detailed analytics must be done according to restricted areas (pathologies, products);
2. Different time spans can be compared, going deeper into the general view;
3. Incorrect or unclear information has to be removed, causing retention of total records whose amount gets narrower while cleaning is in progress (*funnel* effect).

The available dataset is modelled in a **relational schema**, which allows organization of records in structures (tables) to maintain integrity and compatibility between different kind of fields. This allows flexibility using dynamic views and query optimisation based on set theory.

For a better fruition of the workload, table names have been changed to standard English keywords.

3.2 Considerations

Since health big data can provide a high variety of information, it is required to have some clear *objectives* in mind to keep on track and avoid losing focus.

The research is centred on the **changes of prescription patterns of antibiotics**: this is a wide goal, and more information is needed to achieve results. It is necessary to narrow the field down, only concentrating on some **classes** of medicines, and define subgroups of GPs and patients in a restricted amount of time.

To obtain constraints the more objective as possible, some additional analysis can be useful. For instance, focussing on chronic patients is a relatively fast way to reduce the huge amount of rows to elaborate, but causes the loss of most information.

The first step to take, having a deep understanding of the data, is recognising the extent and impact of the **progressive information loss**, to define the final amount of clear records. Trying to fix mistakes is a risk, since the outcome could be incorrect, therefore deleting is the most practical option.

Removing unclear and futile data may not be enough to have consistent results: 18 years of data is a wide range, and *splitting the dataset* or deciding to only consider a smaller scope can be beneficial for the analysis quality.

Some constraints can be imposed on general practitioners as well: since the results have to be coherent and accurate, it's best practice to only consider **active GPs** with a **constant number of patients** (to be defined).
This would partially remedy the fact that doctors may have different approaches to the same disease.

The creation of cohorts (chronic patients) among the statistical population is an example of **cluster sampling**.

After the initial parsing, there will be a rough draft of the final result which then will be subject of the following steps:

1. Further analysis on data correctness (*record linkage*);
2. Elaboration of the statistics and *time series clustering*.

Another relevant instance for analysis is the **subset** of diseases to consider: choices have to be made according to *external studies*, *marketing researches* and further *discoveries on the provided data*. Focussing on the **most common ones** is a guideline to start.

Having an idea of which illnesses and prescription have unstable patterns might give a better vision, and can be done through statistics on the whole database.

Some examples of analytics are:

- Most common diseases through the years;
- Most common *chronic* diseases through the years;
- Changes of the number of prescriptions for diseases in the same area;
- Changes of prescriptions based on the patient phenotype or market trends.

An obstacle to perceiving the meaning of results is the **restricted domain knowledge**: to compensate, confronting some experts in the field is required. The team comprehends computer scientists, statisticians, biologists and healthcare workers.

3.3 Practical methods

Having a better vision of the medical domain and the growing issues in the Italian system still requires the underline of **practical objectives** and **expected outcomes**, before approaching analytical procedures.

The first essential statistic to extract is about **progressive information loss**: this makes possible to estimate how much data can give reliable and complete results, compared to the total. It is defined *progressive* since the more iterations of cleaning are being made, the more data is going to be lost.

Parsing remaining data is essential to have a stable functioning of database interrogations. Fields need to be checked for correctness using lookup tables, record linkage or regular expression matching; text has to be cast to numbers to apply mathematical operations, and dates can be divided into months and years.

Performances require a high level of **optimisation** due to the huge workload of information, obtained through targeted expressions, strict constraints and better memory management of records.

A global overview of the data allows to detect anomalies or trends of specific areas to focus on: potential analysis pertains dividing patients according to distinctive characteristics (gender, age group) to observe variations of diagnoses and prescriptions.

After collecting the first batch of results and checking them using external knowledge, information with *unusual patterns* is outlined, and further examination is made on a subset of features.

Final outcomes are then subject of more advanced techniques, which include:

- Time series analysis;
- Clustering of trajectories;
- Graph algorithms.

3.4 Approaches description

While systematically applying methodologies, it is essential to make a difference based on **use case**s and types of algorithms, to have an overview of different techniques and results to expect.

3.4.1 Descriptive statistics

Descriptive statistics are used to summarize and describe basic features of information, to give a full picture of a sample without generalizing beyond the data in hand.

They are used to offer an insight without getting into conclusions, presenting values in a manageable form and simplifying large amounts of records. Such summaries can be quantitative (*statistics*) or visual (*graphs*), giving basis to extend the research with more specific approaches.

Some instances of commonly used measurements are **central tendency**, **variability** and **dispersion**, which allow to broke down datasets and re-purpose each attribute into a smaller description. Indices give information on how a value is distributed, how spread-out it is and what shape it assumes.

Descriptive statistics allows to have a general overview of the available health data to then restrict the domain and identify areas of interest.

3.4.2 Time series analysis

There are two main goals of **time series analysis**: identifying the *nature* of a phenomenon represented by the sequence of observations, and *forecasting* (predicting future values of the time series variable).

Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, it can interpreted and integrated with other data.

Regardless of the depth of understanding and the validity of interpretation (theory) of the phenomenon, the identified pattern can be used to predict future events [24].

Antibiotic resistance is the most immediate example of use case: observing **trends** during years and months helps understanding the growth (or recess) of the issue and its development.

3.4.3 Exploratory analysis

Exploratory data analysis (EDA) is a strategy of data analysis that emphasizes maintaining an open mind to alternative possibilities. EDA is a philosophy or an attitude about how data analysis should be carried out, rather than being a fixed set of techniques [28].

The approach, similarly to descriptive statistics, consists in *summarizing the main characteristics* of a dataset to eventually apply statistical models to search for mathematical relationships between variables.

It is difficult to obtain a clear cut answer from "messy" human phenomena, and thus the exploratory character of EDA is very suitable to medical research.

This is a systematic way to investigate relevant information from multiple perspectives: in many stages of inquiry, the working questions are non-probabilistic and the focal point should be the data at hand rather than the probabilistic inference in the long run. Hence, prematurely adopting a specific statistical model would hinder from considering different possible solutions.

The key point of EDA is the emphasis placed on using data to suggest hypotheses to test, rather than confirming existing hypotheses. Causes of observed phenomenon can be pinpointed, assessing assumptions for statistical inference through the appropriate statistical tools.

Techniques consist in **plotting features** to visualize their behaviour, to then extract the most relevant ones through dimensionality reduction and projecting trends.

Exploratory analysis gives a better insight on the information obtainable from the healthcare data, understanding unusual trends and defining detailed objectives.

3.4.4 PCA

Principal Component Analysis is mostly applied as a tool in exploratory data analysis for making *predictive models*. It is one of the most widely used methods to *reduce dimensionality* of large datasets while still preserving most information and variability.

PCA allows to find **meaningful projections of features**, using an unsupervised approach, finding the subspace of largest variance and calculating the eigenvectors to project data into a subspace.

This translates into finding new variables that are linear functions of those in the original dataset, converting data into a linear mapping with less dimensions, that successively maximize variance and with features that are uncorrelated with each other [19].

Reduced variables can be used for an easier interpretation, and successive methodologies such as clustering are applied to subsets to improve computational time. The obtained

attributes are subject of further research in the healthcare field and multi-dimensional data visualisation.

3.4.5 Clustering

Data clustering techniques are descriptive data analysis techniques that can be applied to multivariate data sets to uncover the structure present in the data.

They are particularly useful when classical second order statistics (the sample mean and covariance) cannot be used. Namely, in exploratory data analysis, one of the assumptions that is made is that no prior knowledge about the dataset, and therefore the dataset's distribution, is available. In such a situation, data clustering can be a valuable tool.

Data clustering is a form of unsupervised classification, as the clusters are formed by *evaluating similarities and dissimilarities* of intrinsic characteristics between different cases, and the grouping of cases is based on those emergent similarities and not on an external criterion.

Also, these techniques can be useful for datasets of any dimensionality over three, as it is very difficult for humans to compare items of such complexity reliably without a support to aid the comparison [20].

Clustering healthcare data offers the opportunity to **differentiate elements** such as patients and doctors according to their characteristics and behavioural patterns, highlighting trends to possibly make predictions for the future.

k-means

***k*-means** clustering belongs to partitioning-based techniques grouping, which are based on the iterative relocation of data points between clusters. It is used to divide either the cases or the variables of a dataset into non-overlapping groups, or clusters, based on the characteristics uncovered [20].

The goal is to produce groups of cases/variables with a **high** degree of similarity within **each group** and a **low** degree of similarity between **groups**.

The objective is therefore to *minimise* the following equation:

$$C(z, \mu) = \sum_i ||x_i - \mu_{z_i}||^2$$

z_i are the assignment variables, which can take values $z_i = 1, \ldots, K$ where K is an arbitrary number of clusters.

k-means clustering is very useful in exploratory data analysis and data mining in any field of research, and as the growth in computer power has been followed by a growth in the occurrence of large data sets.

A good cluster analysis is both efficient and effective, in that it uses as few clusters as possible while still capturing all statistically important clusters.

The practical approach uses the algorithm of *Hartigan and Wong*, the most popular implementation, which searches for the partition of data space with locally optimal within-cluster *sum of squares of errors* (SSE).

The Hartigan method examines every item in each cluster at random, calculates the distance to centroids and assigns it to the optimal partition, taking into account the motion during re-assignment.

k-means can be used to classify prescription patterns, having the chance to construct a feature matrix with time series data.

3.4.6 Graph algorithms

While graphs originated in mathematics, they are also a pragmatic and high fidelity way of modelling and analyzing data. The objects that make up a graph are called **nodes** or vertices and the links between them are known as **relationships**, links, or edges.

Graph algorithms are a subset of tools for graph analytics. Graph pattern–based querying is often used for *local data analysis*, whereas graph computational algorithms usually refer to more global and iterative analysis.

Graph algorithms provide one of the most potent approaches to analyzing **connected data** because their mathematical calculations are specifically built to operate on relationships. They describe steps to be taken to process a graph to discover its general qualities or specific quantities.

Based on the mathematics of graph theory, graph algorithms use the relationships between nodes to infer the organization and dynamics of complex systems [23].

Healthcare data can be represented using a graph, identifying main attributes along with their behaviour and interactions.

Betweenness centrality

Centrality algorithms are used to understand the roles of particular nodes in a graph and their impact on that network. They're useful because they identify the *most important nodes* and help understanding group dynamics such as credibility, accessibility, the speed at which things spread, and bridges between groups.

Betweenness Centrality is a way of detecting the amount of influence a node has over the flow of information or resources in a graph. It is typically used to find nodes that serve as a bridge from one part of a graph to another.

The Betweenness Centrality algorithm first calculates the *shortest (weighted) path* between every pair of nodes in a connected graph. Each node receives a score, based on the number of these shortest paths that pass through the node. The more shortest paths that a node lies on, the higher its score.

The betweenness centrality of a node is calculated by adding the results of the following

formula for all shortest paths:

$$B(u) = \sum_{s \neq u \neq t} \frac{p(u)}{p}$$

u is the selected node, p is the total number of shortest paths between nodes s and t, and $p(u)$ is the number of shortest paths passing through u.

Betweenness can be used to measure antibiotics' influence among co-prescriptions.

Degree centrality

Degree Centrality is the simplest of the algorithms, and counts the *number of incoming and outcoming relationships* from a node. It is used to find **popular instances** in a graph, concerning immediate connectedness and near-term probabilities.

The **degree** of a node is in fact the number of direct relationships it has, calculated for in-degree (incoming edges) and out-degree (out-coming edges).

Degree is another indicator to analyse the influence of single products to highlight most common ones.

Community detection

A **community** is a group of individuals having the same particular characteristic, sharing or having certain attitudes and interests in common[1].

Community formation is common in all types of networks, and identifying them is essential for evaluating group behaviour and emergent phenomena. Most real-world networks exhibit substructures (often quasi-fractal) of more or less independent subgraphs.

Connectivity is used to *find communities* and *quantify the quality* of groupings, uncovering structures and tendencies of groups to attract or repel others.

The general principle in finding communities is that *its members will have more relationships within the group than with nodes outside their group*. Identifying these related sets reveals **clusters** of nodes, isolated groups, and network structure. This information helps to infer **similar behaviour** or preferences of peer groups, estimate resiliency, find nested relationships, and prepare data for other analyses.

Modularity algorithms optimize communities locally and then globally, using multiple iterations to test different groupings and increasing coarseness. **Louvain Modularity**, in particular, is used for looking at grouping quality and hierarchies.

It maximizes the presumed accuracy of groupings by comparing relationship weights and densities to a defined estimate or average, revealing hierarchies at different scales with distinct levels of granularity and aggregating into super-communities.

Community detection in a healthcare graph schema can help dividing instances according to their habits, giving a structure in groups with different patterns.

[1] Lexico dictionary definition

Similarity detection

Similarity algorithms work out which nodes most resemble each other by using various methods to compare items like node attributes. It is useful for dense graphs, giving additional insights to clustering.

Jaccard Similarity, a term coined by Paul Jaccard, measures similarities and differences between sample sets with discrete attributes, assigning a **coefficient** to each pair of nodes. It is defined as the size of the intersection divided by the size of the union of two sets.

Similarity can be applied to recognise prescription patterns, finding doctors with the same characteristics.

3.5 Tools

Modern technologies make possible to process big data with reduced costs: there are plenty of data stores, development and integration tools for each research purpose.

Since the project requires a relational structure along with statistical computing and machine learning, analysing and elaborating the health data is made through:

- **PostgreSQL**[2], an open source object-relational database management system known for its robustness and reliability:
  - Indexes and Common Table Expressions are useful to avoid huge computational times;
  - Database interrogations and functions give a schema overview to then apply further grouping and filtering.
- **Neo4j**[3], a native graph database which gives data a different representation, processing entities as nodes while highlighting their connections:
  - Queries are expressed using **Cypher**;
  - The additional plugin Graph Algorithms returns implemented, parallel version of common network problems.
- **R**[4], a free software environment for statistics and graphics computing, offering a wide range of techniques and formulae:
  - *ggplot2* is a package to create and visualise plots;
  - Clustering is performed using embedded functions.

Due to the amount of sensitive data, detailed results are going to be omitted: the final conclusions will be a product of aggregation and schematisation.

[2]PostgreSQL official website
[3]Neo4j official website
[4]R project official website

Chapter 4

Data description

This chapter aims to give a full understanding of the database on which analytics is performed, getting into a first level of technical detail with the description of fields, ER model, primary keys, number of records and a basic example of analysis.

From now on, all images and tables are obtained using the dataset illustrated below.

4.1 Dataset description

The available database is collected, managed and complied with current privacy regulations, by *Consorzio Nazionale delle Cooperative Mediche* (CNCM), using software provided by Dedalus[1], market leader of the clinical software area, supporting doctors and their processes through its society Millennium.

Research is conducted using data collected by an interface used by general practitioners to track interactions between them and the population benefiting of the national healthcare system.

The database contains recorded **medical history of patients** using healthcare services in the southern-Italy region **Campania**, focussing on the doctor-patient relationship.

It is composed of pseudonymised data by encryption before acquisition and is related to eighteen years data about outpatient visits at GPs' ambulatories located in several Health Districts.

Since the database is not regulated by local laws, it encounters a greater risk of inaccuracy, unlike pharmacies or tax registers.

General practitioners are the only responsible of filling values, therefore there is *no assurance of completeness and correctness of data*: mistakes are common, as well as missing information. A part of patients journey happens in hospitals or specialised medical offices, and those records aren't present since belonging to external sources.

Only a part of the whole amount of drugs (and examinations) require a written prescription: most medicines are given over the counter, and there is no certainty that a patient is

[1] Dedalus official website

going to buy that specific drug or the generic equivalent. Linkage between prescriptions and actual purchase is missing.

Furthermore, **not all prescriptions are ethical**: antibiotic resistance is an ascertained issue, and general practitioners can be influenced by pharmaceutical companies, resulting in *lack of objectiveness*.

Data is **highly sensitive**: despite encryption of all names, a considerable amount of geographical information is available. To avoid cross-checking using location, for results to be published patients or doctors should be aggregated in groups whose numerousness exceeds a fixed value (rule of thumb states 3).

Other sensitive information such as email addresses and passwords to log in the system is irrelevant for analytics, therefore it is safe to remove anything not strictly related to the research purposes.

4.2 Database overview

The database used for analytics contains data on medical histories of patients between **January 2000** and **October 2018**. This leads to some observations:

1. The year 2018 is present only up to June, so it cannot be used while making time series within years (there is going to be a drop of values due to incompleteness, which may lead to wrong conclusions);
2. A timespan of 20 years is too wide to make consistent analytics;
3. Early dated records might contain outdated or incomplete information.

Global inferences have been made with the entire dataset, while the need of detailed recent reports leads to the decision of using a limited range of years for prescription pattern changes and patient journey.

The research work has been done on only a part of the original Millewin database, consisting in **4 tables**. There is information available on **general practitioners**, **patients**, **diagnoses** and **prescriptions**: each macro-category is included in a separated table, therefore identifying the relationship between fields is necessary to make interrogations with co-joined data.

The 4 tables with their sizes are:

- *patients*, 1 015 618 tuples;
  Basic information about patients, identified by an encrypted UID;
- *patients_ doctors*, 1 015 618 tuples;
  Extension of *patients* with the same key, containing more detailed information about patient-doctor relationships and linkage with GPs identifiers:
  - There are 1 356 general practitioners;
- *diagnoses*, 15 460 199 tuples;
  Information about diagnoses and relative description;

- *prescriptions*, 118 716 403 tuples;
  Information about therapies and prescribed medicines.

It is noticeable that the number of rows is varying: there are more prescriptions than diagnoses, since the first tend to happen more often.

Each diagnosis and prescription is uniquely distinguished by the triplet {*patient*, *doctor*, *date*}. Dates are at level of timestamp, making each one different from the others (it is improbable to have a diagnosis or prescription for the same patient, by the same doctor and at the same exact moment).

Analysis is performed using dates in the *YYYY-MM-DD* format, since non-unique data still allows to aggregate results and identify patterns. There are several different prescriptions for the same patient made on the same day.

4.3 Description of the tables

Before being able to work with the data, it is essential to understand its structure, functioning and behaviour within time.

Below is reported a brief description of the 4 tables, along with the main fields used for analytics, statistics and machine learning.

4.3.1 *patients*

The table *patients* includes information about patients. To ensure privacy dealing with sensitive data, there are no full names: fields are **encrypted** as a 22-character string containing letters, numbers and special symbols.

Other relevant fields are:

- *birthdate*, date of birth;
- *death*, eventual date of death;
- *birth_municipality*, name (and code) of the birth municipality;
- *gender*, birth sex;
- *convention*, type of convention with the Italian insurance system.

4.3.2 *patients_doctors*

The table *patients_doctors* contains information similar to *patients*, with additional fields focussing on their relationship with the general practitioners, which are essential to link and analyse data:

- *userid*, **encrypted** UID of the general practitioner of the patient;
- *date*, date of beginning of the doctor-patient relationship;

- *postcode*, zip-code of the patient (for geographical analysis);
- *province*, province of the patient;
- *revocation*, eventual date of termination of the doctor-patient relationship (a patient changing GP).

All the IDs of the GPs, along with all other data on GPs, are stored in an external table *users* (the research has been made considering a subset of the original DB). The latter does not contain any other information relevant for analysis, since active doctors can be extracted from other tables.

4.3.3 *diagnoses*

The table *diagnoses* comprehends the diagnoses associated to patients and relative GPs. Each diagnosis is defined by its **ICD-9** code, an international identifier for diseases maintained by the World Health Organization.

Summary of most important features:

- *id*, patient ID from *patients*;
- *userid*, corresponding to *doctor* in *patient_doctor* (general practitioner ID);
- *date*, date of insertion of the diagnosis in the database;
- *last_update*, timestamp of last edit of the tuple;
- *description*, a textual description of the diagnosis;
- *IDC9*, code of the diagnosis according to the ICD-9 standards.

4.3.4 *prescriptions*

The table *prescriptions* contains the prescribed medicines for each patient. There is no linkage between diagnoses and prescriptions in the database, therefore additional work is required to detect correlation.

Each prescription is defined by an **ATC code**, from the Anatomical Therapeutical Chemical classification system maintained by the World Health Organization. Furthermore, there are **active principle code** and **authorisation for commerce code**(AIC).

Fields summary:

- *id*, patient ID;
- *userid*, corresponding to *pa_medi* in *nos_002* (general practitioner ID);
- *date*, date of insertion of the prescription in the database;
- *last_update*, timestamp of last edit of the tuple;
- *AIC*, AIC code (authorisation for commerce);
- *description*, textual description of the medicine with dosage;

- *pieces*, number of boxes;
- *active_principle*, active principle code;
- *ATC*, ATC code;
- *euro*, price.

4.4 ER model

After identifying the main fields, it is possible to build and include an **ER model**[2] (figure 4.1), to provide immediate comprehension of how entities are related, visualising information to identify keys and unique fields, essential for joining data.

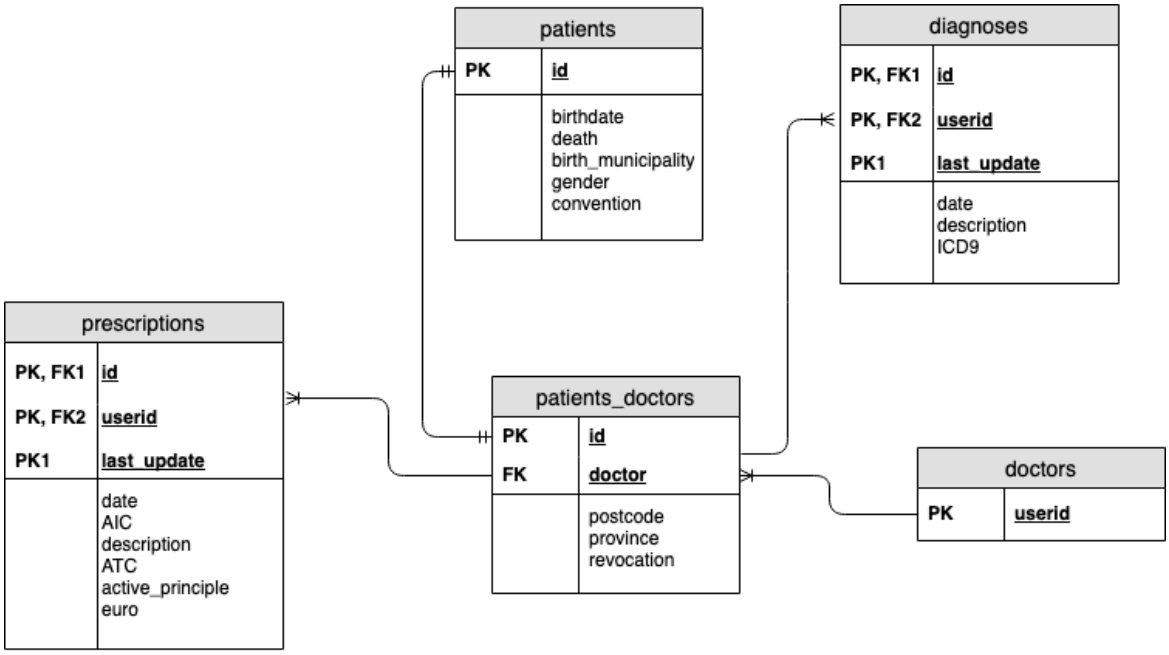

Figure 4.1: ER model

4.5 Example of preliminary analysis

Analysis on patients' gender, on 1 015 618 distinct patients, with approx. percentages:

- Males: 479 556, 47%;
- Females: 533 256, 52%;
- Other or null, 2 806, $< 1\%$.

Percentages of males and females are coherent with ISTAT data of 2018, showing 48% males and 52% females in the whole region Campania[3], confirming the validity of data.

[2]Image made with `draw.io`
[3]ISTAT 2018

The 1% of missing data introduces the **information loss** issue, which will be examined in detail in the next part of the research.

4.5.1 Variations through time

To give a general idea on variations of available information and magnitude orders, a **snapshot** of 2010-2017 is used to show patterns and differences.

It is expected for data to progressively increase, since more and more general practitioners have familiarised with the electronic recording systems within years.

A first analytical approach consists in counting the number of patients getting at least a diagnosis for each year:

	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Patients	329 715	320 253	316 431	320 948	324 920	323 441	330 641	326 987
Age mean	47,43	47,89	48,44	48,68	48,98	49,78	50,12	50,54
Age s. d.	20,7	20,81	20,84	20,86	20,87	20,84	20,85	20,79
Women	185 035	179 700	178 248	180 304	182 286	181 313	185 729	183 419
Men	144 016	140 330	137 514	139 780	141 727	141 158	143 931	142 363

Table 4.1: Number of patients with diagnoses

Another example can be obtained counting the number of prescriptions each year:

	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
P.	7 326 923	7 108 288	7 269 855	7 541 539	7 777 753	7 847 313	7 856 246	7 785 325

Table 4.2: Number of patients with prescriptions

Patients getting diagnoses tend to remain stable, yet prescriptions have a difference of more than 450 000 records, confirming the hypothesis of information registering increase while raising the concern of an effective general overprescribing.

Chapter 5

Information loss

This chapter aims to give a qualitative assessment of the whole database through an initial analysis, giving an overall idea on how impactful is the progressive information loss.

After assessing the correctness and completeness of records, some fields will be eventually excluded from the analysis, while others that cannot be removed will have a major impact on the results.

Having missing fields, particularly in the process of joining tables, can lead to a **cumulative augmentation** of the information loss: empty data such as the patient date of birth or gender will cause the deletion of the entire patient, in case analytics is centred on pathologies by age or gender.

Joining is in fact an operation which requires *all fields of reference to be present*, combining entries of the selected tables.

5.1 General overview

Before starting running queries, there are some aspects to consider involving issues which sometimes cannot be addressed just with database interrogations:

1. The geographic information is sometimes imprecise and hard to comprehend, since it consists in text fields;
2. Some diagnoses descriptions don't match with the corresponding ICD-9 code;
3. A consistent amount of missing data is originated in case a general practitioner doesn't prescribe anything but makes other operations (medical certificates, examinations and such);
4. Hospital prescriptions are missing;
5. Some medicines are given over the counter without requiring a prescription, therefore there is no entry in the DB;
6. General practitioners might prescribe a medicine to a different patient than the one who has the pathology (relatives, friends, ...);

7. There is inappropriate prescribing, antibiotic resistance, misuse or over-use of medicines;
8. A patient can change doctor, so the approach to the same disease may vary.

Furthermore, there have been noticed some common instances of incorrect data:

- Some dates don't fall in the acceptable range (e.g. 1999, 2034, ...):
  - A date is considered wrong if it falls before 01-01-2000 or after 10-01-2018;
- A consistent amount of fields are empty or null.

It is important to remember that some records might be affected by more than one incorrect field, therefore when calculating total information loss it is essential to intersect different subsets, to see which records they have in common, instead of just summing the numbers of corrupted rows.

Overall, the loss on single records is not a relevant issue since the total amount of rows allows safe removal, yet the cumulative augmentation (funnel analysis) gives a progressive deletion of information.

5.2 Information loss on records

5.2.1 *patients* and *patient_ doctors* (1 million of tuples)

Summary

Both tables contain similar data related to patients, with a $1:1$ correspondence between primary keys, therefore joining is an elementary operation and there is no information loss.

- Patients with null or empty gender: 2 752 $\rightarrow$ 0.27%;
- Patients with gender different from M and F: 54 $\rightarrow$ 0.005%;
- Patients with beginning of the patient-doctor relationship outside the accepted range: 226 858 $\rightarrow$ 22.33%;
- Patients with null province: 99 981 $\rightarrow$ 9.84%.

A subset of patients is creating through an auxiliary view to highlight the final progressive data loss compared to the original tables, joining *patients* and *patients_ doctors* according to those constraints:

1. Join on equal patient code (*id*);
2. Not null date of birth;
3. Dates between 2000 and 2018;
4. Existing and not null gender;
5. Existing and not null province.

The total rows respecting all those constraints are 713 352, so approximatively 300 000 tuples (patients) have been deleted.

This result implies that during analysis there will be at least 1/3 of the data which is going to be removed due to incompleteness and inaccuracy: not considering patients will imply deleting their diagnoses and prescriptions as well.

A waffle chart is shown on figure 5.1, highlighting the total tuples and the impact of every restriction, standardised on a scale from 1 to 100:

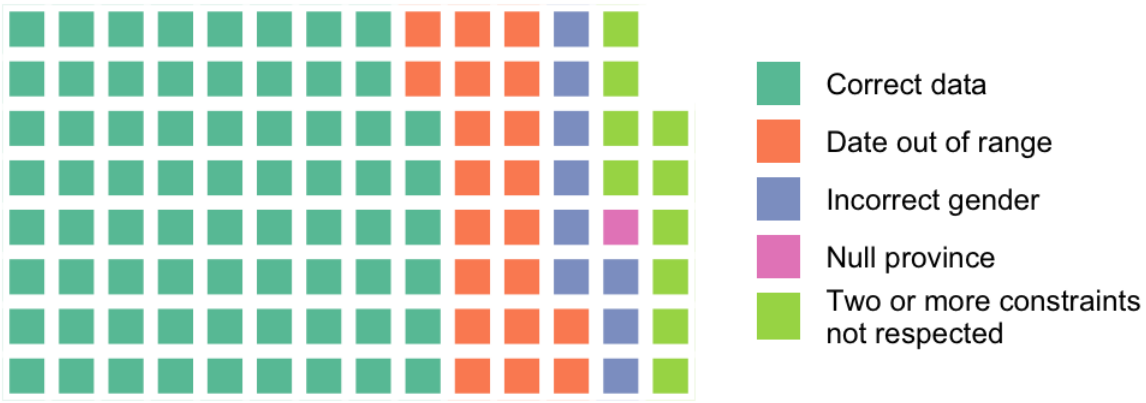

Figure 5.1: Information loss on patients, waffle plot

There are no null provinces and birthdates. The bigger impact is caused by dates falling outside the acceptable range, meaning patients started treatment with their general practitioner earlier than 2000.

5.2.2 *diagnoses* (15 millions of tuples)

ICD-9

An ICD-9 code is correct if it is present in the official ICD-9 database [25]. General practitioners may use different formats and notations, so before removing codes each one has been subject to preprocessing and parsing.

An ICD-9 code is considered wrong if there is no match in the official DB after any of those transformations:

1. Removal of the dot;
2. Addition of 0 at the beginning;
3. Addition of 0 at the end;
4. Removal of the last 0.

There are 76 distinct incorrect ICD-9 codes, a negligible amount considering the total records consisting in 9 895 distinct codes.

Summary

- Incorrect ICD-9 codes: 76 → 0.0005%;
- Null or empty ICD-9 codes: 2 103 169 → 13.02%;
- Null or empty descriptions: 656 067 → 4.24%;
- Dates out of range: 807 985 → 5.23%.

Empty descriptions almost always correspond to empty ICD-9 codes, therefore the total information loss on descriptions is absorbed by null codes, resulting in the same 13.02% united percentage.

5.2.3 *prescriptions* (118 millions of tuples)

ATC code

The ATC code, unlike ICD-9, has a univocal format (a numeric string of 7 digits), and no parsing is needed. All the codes have been checked with the ontologies in a well-known biology portal [26], and an ATC is considered incorrect if there is no match.

There are 1 750 292 non-existent codes, which are caused by:

1. Alteration of the codes within the years without updating the database;
2. Single prescriptions indexed using the superclass code;
3. Codes only recognised by local pharmacies.

The latter is the most common reason: there are up to 90 000 occurrences of a single unofficial code. Those numbers, despite being high, are not majorly impacting results, considering the total amount of 118 millions of records.

Summary

- Incorrect ATC codes: 1 750 292 → 1.47%;
- Null or empty ATC codes: 1 577 749 → 1.33%;
- Null or empty AIC codes: 1 310 719 → 1.1%;
- Null or empty descriptions: 101 248 410 → 85.28%;
- Dates out of range: 3 472 119 → 2.92%.

Clearly, the prescription description is not a field which can be used for reliable analytics since empty values prevail.

The field *pieces* (number of boxes) might be useful to check for appropriate prescribing, but since the field is a string it requires casting and parsing to integer, and it could be more relevant to focus on prescription patterns.

5.3 Total information loss

Overall, total information loss is shown according by its belonging table.

Table	Records	Percentage		Usable records
patients	302 266	29%	→	713 352
diagnoses	2 403 202	15.5%		13 056 997
prescriptions	5 560 191	4.7&		113 156 212

Table 5.1: Information loss summary

The aggregated results are not final: each single output from interrogations will be influenced by all the different invalid fields (i. e. a patient could have all correct data yet a missing prescription AIC), creating a progressive deletion effect.

Overall, percentages do not compose the majority of tuples, since tables have magnitude order of millions.

Chapter 6

Global analytics

This chapter concerns global analytics, used to have an initial view on the real value of the data: goals include understanding what kind of information can be extracted and setting practical fields of interest to make deeper research.

The methodology consists in the application of **exploratory analysis** and **descriptive statistics** on the existing dataset.

To have a general idea, the considered approach involves the *whole dataset*: all the available information is included, consisting in the entire time range of 18 years and the totality of patients with related records. Missing or incorrect data is then removed, without imposing additional constraints.

Analysis is performed using:

- Diagnoses;
- Prescriptions;
- Patients' phenotype.

Performed operations are simple, such as aggregating, ranking and counting, to avoid high computational costs while obtaining clear and comparable results.

Prescriptions are additionally grouped and measured along with diagnoses, to identify eventual linkage between most popular ones, and focus on discrepancies.

The main purpose of global analytics is highlighting unusual patterns and cross-checking information with external studies, confirming anomalies or restricting the field to existing areas of concern.

6.1 Most frequent diseases

Selection of a subset of diseases is possible after preliminary analysis, of which the most immediate regards frequent appearances in the database.

There are 9 895 distinct ICD-9 in the *diagnoses* table: of those, the 10 most popular ones in terms of count get selected and shown in table 6.1.

#	ICD-9	Count	Description
1	799.9	441 131	Other unspecified cause of morbidity*
2	401.9	305 476	Unspecified essential hypertension
3	462	301 575	Acute pharyngitis
4	521.0	274 624	Dental caries
5	595.9	226 428	Cystitis, unspecified
6	464.1	221 803	Acute tracheitis
7	466.0	200 408	Acute bronchitis
8	780.7	183 721	Malaise and fatigue
9	530.81	182 897	Esophageal reflux
10	724.2	176 921	Lumbago

Table 6.1: Most frequent diseases in 18 years

The most common ICD-9 (*full name: other unknown or unspecified cause of morbidity and mortality) corresponds to a generic unspecified disease, therefore cannot be subject of detailed analysis.

Other diseases vary through chronic and not, different interested age ranges and different organs of the body: further information can be obtained imposing additional constraints on classification.

6.1.1 Most frequent diseases based on age

Since some diseases are likely to appear within specific ages, patients are grouped into **ranges** and the most common ICD-9 are counted based on the difference between birthdate and prescription date.

Age ranges

The classification method follows *international standards* by United Nations [33], developed on the basis of existing national practices and recommendations concerning age.

Health classification, in particular, identifies the following ranges to the third level of detail:

1. Less than 1 year of age;
2. 1-14 years of age;
3. 15-24 years of age;
4. 25-44 years of age;
5. 45-64 years of age;
6. 65+ years of age.

Since the available data concerns general practitioners and not paediatricians, although patients in the younger ages are present there is no warranty of completeness of the doctor-patient relationship, therefore the first two ranges have to be removed.

Results

The most common diagnosis for each age range is:

Range	ICD-9	**Patients**	**Description**
3	462	49 096	Acute pharyngitis
4	799.9	119 265	Other unspecified cause of morbidity
5	401.9	140 510	Unspecified essential hypertension
6	401.9	123 031	Unspecified essential hypertension

Table 6.2: Most frequent diagnosis for age range

Since range 4 (25-44) has an unspecified disease as the most common, the 2nd and 3rd results are listed:

1. 462, 97 040 patients, acute pharyngitis;
2. 521.0, 83 948 patients, dental caries.

Hypertension is more common among older patients, while young adults are mostly affected by pharyngitis and caries.

Overall, the most popular diagnoses according to age reflect the general rankings.

6.1.2 Most frequent diseases based on gender

Since the database contains more female patients than males, it is expected to have a gap between amounts of diagnoses.

The most frequent diagnoses are however **unrelated** to gender: there are no relevant differences implying a kind of patients is most likely subject to a disease in the considered subset.

Analytics regarding gender have to be performed choosing a different batch of illnesses, taking the ones which are proven to affect one gender rather than the other.

6.1.3 Most common diagnoses based on ICD-9 class

Making ranks from data on a general level is useful to identify most common diagnoses overall, yet only a limited part of all diseases is taken into account.

Information is sliced only according to aggregated total value, without additional considerations and therefore excluding a wide part of the totality.

Since ICD-9 classification divides diagnoses in **classes** according to the interested part of the body, extracting the most common illness according to its class allows a better understanding of how important is each group and its **influence** or relation to the prescriptions. Data can be therefore used to make comparisons.

The whole span of 18 years is used to obtain results, having removed records subject to progressive information loss and incompleteness.

ICD-9 analysis show:

Description	Diagnoses	Count
Infectious and parasitic diseases	Herpes zoster	28 202
Neoplasms	Benign neoplasm of skin	41 725
Endocrine and metabolic diseases	Pure hypercholesterolemia	54 935
Blood and blood-forming organs	Anaemia, unspecified	71 689
Mental disorders	Dysthymic disturb	43 303
Nervous system	Acute otitis media	65 776
Circulatory system	Unspecified essential hypertension	204 995
Respiratory system	Acute pharyngitis	212 704
Digestive system	Dental caries	186 540
Genitourinary system	Cystitis, unspecified	157 684
Pregnancy and childbirth	Caesarean delivery	3 144
Diseases of the skin	Dermatitis, unspecified	93 178
Musculoskeletal system	Lumbago	116 289
Congenital anomalies	Spondylolisthesis	883
Conditions in the perinatal period	Subdural and cerebral hemorrhage	587
Symptoms, signs	Other unspecified cause of morbidity	262 867
Injury and poisoning	Allergy, unspecified	74 773
External causes of injury	Pregnant state, incidental	56 169

Table 6.3: Most popular diagnoses for ICD-9 class

Number of prescriptions varies from less than 1 000 to more than 250 000, showing the discrepancy of amount of diagnoses among different classes.

Results reflect general analytics, giving additional insight on other categories having consistent values yet not so high to enter the top rankings, such as dermatitis, anaemia and otitis.

6.2 Chronic illnesses

A chronic illness is defined as *a disease or condition that usually lasts for 3 months or longer and may get worse over time*[1]. They can usually be controlled but not cured.

The major limitation of the database is the lacking of fields representing chronic patients, since illnesses can last any amount of time yet be diagnosed only once, therefore counting and ranking does not give relevant information.

Even identifying a subset of chronic patients based on amount of prescriptions or phenotypes would not result in consistent outcomes, because popular diseases such as hypertension or pharyngitis are spread among all patients and a common reason for periodic visits of the general practitioner.

[1]Cancer.gov definition

A patient living with a chronic disease can, in fact, have episodes of sporadic disturbs, therefore there is not a clear distinction between diagnoses and illnesses.

According to researches on chronic diseases in Italy, they tend to occur in older adults, aged 45 or more: the total patients of the dataset falling in this category in 2018 is more than half, 625 309.

Some diffused chronic illnesses are:

- Hypertension, with 305 476 diagnoses (almost half of the older adult patients);
- Arthritis, with 14 110 diagnoses;
- Allergy, with 100 744 diagnoses;
- Osteoporosis, with 141 671 diagnoses;
- Chronic bronchitis, with 70 537 diagnoses (unrelated to bronchitis);
- Asthma, with 121 252 diagnoses;
- Crohn's disease, with 859 diagnoses;
- Fibromyalgia, with 5 818 diagnoses;
- Arrhythmia, with 21 394 diagnoses;
- Tumour (all, ICD-9 codes 140-239), 288 659 cases.

Although there is a chance that the same disease gets diagnosed more than once to the same patient, values are still high compared to 1 million of patients.

6.3 Gender-influenced diseases

Analytics on most frequent diagnoses according to patient's gender do not show additional relevant information, hence outcomes cannot be obtained using popularity as measure. Prescriptions do not vary either, since their nature is preventive and not curative.

Gender-related diseases might not require a considerable amount of visits of the general practitioner, because either of their chronic aspect or their brief duration.

However, there are important *biological and behavioural differences* between categories. They affect manifestation, epidemiology and pathophysiology of many widespread diseases and the approach to health care.

Since gender affects a wide range of physiological functions, it has an impact on a wide range of diseases including those of the **cardiovascular**, pulmonary and autoimmune systems, as well as diseases involving gastroenterology, hepatology, nephrology, endocrinology, haematology and neurology [34].

Those differences should reflect on the existing data, showing discrepancies between the number of diagnoses for males and females considering the same disease. To verify the hypothesis, a subset of ICD-9 is selected based on external research, and controls are performed.

It is important to consider that female patients are about 4% more than males.

Description	Diagnoses to males	Diagnoses to females
Cystitis (all)	59 276	173 556
Major depression	1 141	1 483
Anxiety states	13 926	23 455
Substance abuse	1 054	142
Erectile dysfunction	3 944	75
Osteoporosis	10 129	82 655
Anaemia, unspecified	20 426	51 318
Myocardial infarction	3 672	1 399

Table 6.4: Diseases counts based on gender

Some categories such as drug abuse include all ICD-9 subsets because of their sparsity, while most common illnesses are counted according to specific typologies to avoid dispersion of information.

It is clear that the selected diseases affect a specific gender than the other: females are likely to get cystitis, anxiety, osteoporosis and anaemia, while males are most subject to substance abuse and erectile dysfunction.

The 75 diagnoses of erectile dysfunction to women are a bright example of wrong diagnosing, and are described as "decrease of sexual desire" or even "frigidity". Those records have to be excluded from analysis of results, showing information loss cannot always be predicted.

Heart attack is an interesting field to look deeper into: researches show difference of signs between women and men, therefore the bigger number of male diagnoses might be caused by females misinterpreting symptoms.

Overall, studies related to Campania comply with available information on how diseases affect distinct genders.

6.4 Most common prescriptions based on ATC class

The purpose of analysing most common prescriptions is to highlight similarities between diseases, and verify whether popular products changed within the past year.

The considered time range is 16 years, divided in spans of 8: 2002-2009 and 2010-2017, to have two comparable pictures of data.

Analytics is performed on the first level of each ATC class, and after extracting each pharmaceutical product, its usage is linked with the most common diagnoses, to understand eventual mutual relationships and connections.

The most common ATC codes for category are counted according to their category (table 6.5).

Class	Description	ATC '02-'09	N. '02-'09	ATC '10-'17	N. '10-'17
A	Metabolism	Omeprazole	717 559	Pantoprazole	2 350 368
B	Blood	Acetylsalicylic acid	1 661 035	Acetylsalicylic acid	1 939 484
C	Cardiovascular	Nitroglycerine	891 586	Atorvastatine	1 379 261
D	Epidermis	Calcipotriol	45 545	Calcipotriol	69 505
G	Urinary system	Tamsulosin	288 840	Tamsulosin	494 191
H	Hormones	Levothyroxine	706 394	Levothyroxine	838 499
J	Infections	Amoxicillin	870 151	Amoxicillin	1 522 031
L	Tumours	Tamoxifen	36 208	Methotrexate	62 324
M	Muscular system	Nimesulide	1 074 632	Ketoprofen	740 165
N	Nervous system	Paroxetine	177 918	Escitalopram	348 080
P	Antiparasitic	Hydroxychloroquine	27 977	Hydroxychloroquine	52 150
R	Respiratory	Beclometasone	412 688	Beclometasone	531 337
S	Sensory organs	Timolol	162 444	Timolol	237 196
V	Various	Oxygen	103 379	Oxygen	246 050

Table 6.5: Most frequent prescription per AIC class

6.4.1 Comparison of results

Alimentary tract and metabolism

The A class shows some potentially concerning results: values of prescriptions of those drugs have tripled from 2002-2009 to 2010-2017. Omeprazole has been substituted by Pantoprazole as top prescribed, yet its prescriptions grew considerably. Both products are for eophageal reflux, which is one of the most diagnoses diseases.

A brief time series analysis on the original *prescriptions* table helps to visualise the growth of the two medicines:

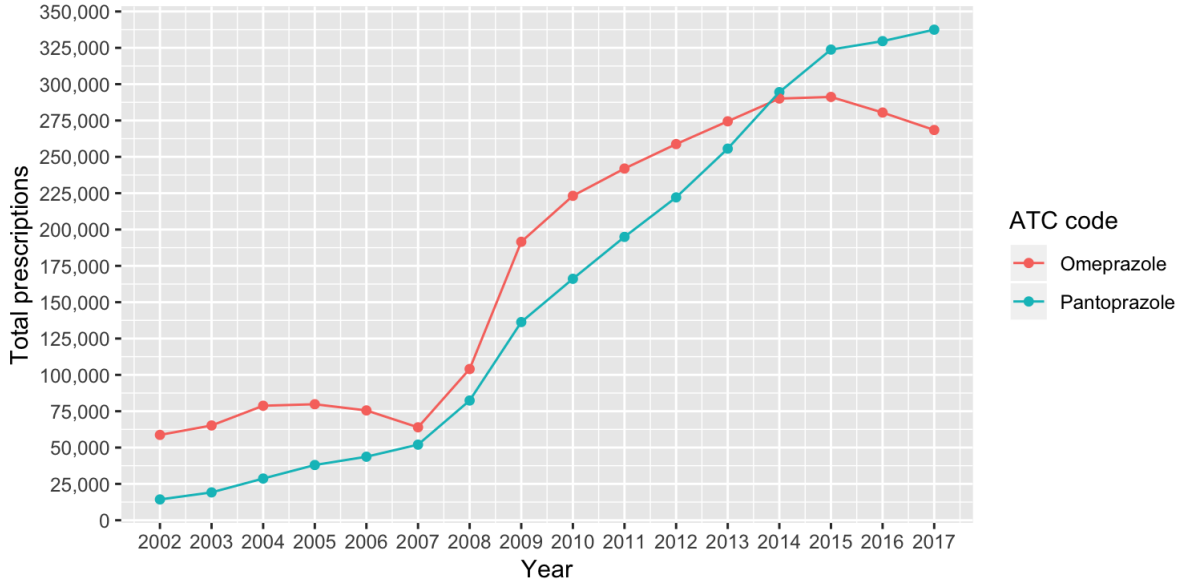

Figure 6.1: Omeprazole and Pantoprazole trends

The trend is subject to a drastic change starting from 2007, with a switch of most popular in 2014.

Further examinations confirm that, despite other alimentary tract medicines have lower values, the number of prescriptions of a consistent amount of those is higher than 1 million, yet the raising trend only concerns omeprazole and pantoprazole.

The hypotheses for changes in patterns are:

1. A bigger amount of information falling in the 2010-2017 range;
2. An effective increase in esophageal reflux and therefore prescriptions related to it.

There are approximatively 18 more millions of tuples overall in 2010-2017, but considering a total amount of 118 millions, the unusual growth cannot be justified by information loss.

Diagnoses of esophageal reflux within years are then counted:

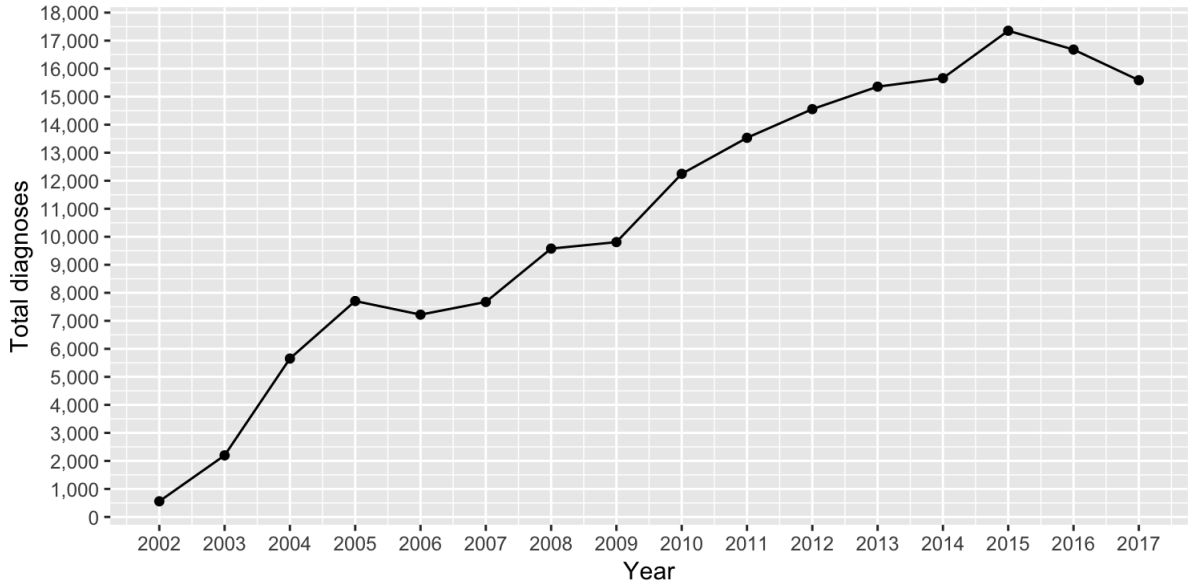

Figure 6.2: Esophageal reflux trends

Values started from less than 1 000 and got to almost 18 000, yet numbers are still too small to explain 300 000 yearly prescriptions.

Other approaches to have a deeper understanding can involve reconstructing the patients history, or extracting the common diagnoses of patients with a prescription of omeprazole or pantoprazole.

Blood and blood forming organs

Blood class has as major product acetylsalicylic acid, which is the active principle of aspirin. This case most likely regards cardio-aspirin, used to prevent and treat heart diseases and as anti-thrombotic.

Cardiovascular system

Atorvastatin is the most common medicine in 2010-2017, already confirmed by gender-related analysis. Before 2010, however, the highest-ranked prescriptions was nitroglycerin, followed by amlodipine and simvastatin.

Numbers suggest a consistent switch of products while dealing with cardiovascular system disease. Values increased, confirming the national issue of heart diseases as principal death cause [35].

Those classes of active principle are used for cholesterol control as well, therefore this can have an influence on the prescription trends.

Nitroglycerin lost popularity, probably because of its heavy side effects, unclearness of mechanism or potential danger while interfering with other medicines or defibrillation.

Dermatologicals

Skin medicines aren't subject of relevant mutations: the most popular is still calcipotriol (calcipotriene), a form of vitamin D.

The usage is mainly for psoriasis, which is discordant with the top ranked unspecified dermatitis in the ICD-9 analysis.

There are 30 768 entries in the whole database for ICD-9 codes concerning psoriasis (696.*), approximately $\frac{1}{3}$ respecting to dermatitis and $\frac{1}{4}$ of prescriptions, yet psoriasis is a chronic disease and only gets diagnosed once.

Genito-urinary system and sex hormones

Urinary system diseases have as most prescribed drug tamsulosin, for benign prostatic hyperplasia. The most diagnosed issue is cystitis, yet it doesn't usually get treated with a medicine cure.

There are 63 368 cases of prostatic hyperplasia in the whole database, which compared to the $\sim$700 000 total prescriptions doesn't explain such high value. It is not a chronic disease, nor prescribed to women and kids, therefore this phenomenon should be looked further into.

Systemic hormonal preparations, excluding sex hormones and insulins

Levothyroxine is one of the most prescribed products in the United States, for hypothyroidism: more than 12% of the US population is affected by this disease, while in Italy the percentage is around 10%.

The discrepancy may be due to different lifestyle or lack of diagnoses: in the Campania database there are 29 184 cases in 18 years, which is a small amount compared to the top diagnoses.

Antiinfectives for systemic use

Amoxicillin, belonging to the penicillins group, is the most used medicine for infections: its scope covers ear, pharynx, airways, skin, teeth and urinary system infections.

Generally penicillins cover almost the entire amount of prescriptions for infections: this can be related to the high quantity of diagnoses concerning otitis, pharyngitis, laryngitis, bronchitis and tracheitis.

Number of prescriptions between the two time ranges has doubled, which consists in a starting point to affirm the spread of antibiotic resistance and lack of ethical prescriptions.

Antineoplastic and immunomodulating agents

Prescriptions for the antineoplastic category tend to vary: from 2010 there is a prevalence of methotrexate, used to treat tumours, especially leukaemia but also psoriasis, regional enteritis and rheumatoid arthritis.

Previously, the most used product is tamoxifen, for breast cancer. All the other common prescriptions consist in inhibitors, which could mean cases of breast cancer is decreasing.

The database shows 3 805 cases in 2002-2009 and 4 086 in 2010-2017, yet the values are biased because of the bigger amount of information related to the most recent years.

Musculo-skeletal system

The popular prescriptions for the muscular system are the same within years, yet their amount drastically changes.

Nimesulide is a nonsteroidal anti-inflammatory drug used for acute pain treatment in orthodontic, rheumatological and gynaecological fields. The medicine has encountered various collateral effects on an European level.

Toxicity and adverse reactions explain the decrease in the number of prescriptions and the switch to Ketoprofene, another nonsteroidal anti-inflammatory drug to treat rheumatological arthritis and osteoarthritis.

Nervous system

Powerful antidepressants lead the area of nervous system drugs: although the two products are different in the considered time spans, the class is the same (inhibitors of serotonin receptors).

Prescriptions values are increasing, despite the number of dysthymia diagnoses going from 22 656 to 20 647, therefore other illnesses treated with antidepressant could be spreading (depression, anxiety).

Antiparasitic products, insecticides and repellents

Hydroxychloroquine is the only product in this category which is consistently prescribed and passes the 10 000 entries: this is because although it being an antimalarial, it gets used for rheumatological arthritis and lupus.

Respiratory system

Different time spans show no relevant differences in the amount of prescriptions of drugs for the respiratory system, yet beclometasone has a detach from all the other medicines (almost double the amount).

Sensory organs

Timolol is prescribed for hypertension, one of the most diagnosed diseases overall. There is coherence between diagnoses and prescriptions, yet a difference of approximately 100 000: cases must be increasing.

Various

Most common various product is oxygen, having a wide spectrum of uses including cardiac insufficiency, chronic bronchitis and bronchopneumonia: crossing diagnoses happening the same day of this prescription would give an overall idea of the leading diagnoses.

6.5 Conclusions

Overall, data show a general trend of increase of both diagnoses and prescriptions. In most cases, this can be attributed to the presence of more information in the recent years, yet some specific instances raise concerns.

Research could get detailed results through a time series analysis on:

- Esophageal reflux and metabolism prescriptions;
- Antibiotic resistance;
- Tumours and benign hyperplasia.

Chapter 7

Patient journey

This chapter defines the patient journey, collection of historical data of patients from the beginning of treatment. After performing information loss analytics to validate the approach, couples of diagnoses-prescriptions are identified to observe their relationship.

After a preliminary analysis to underline the main focus areas, there is enough information to begin reconstructing the **patient journey**, a complete set of data including a patient's records and relationships with primary healthcare, to then identify patterns and changes.

An objective definition of patient journey can be created using the following guidelines:

1. Patients with **complete medical history** for a fixed amount of years;
2. Records with patient, prescribing GP, diagnosis and prescription on the **same date**;
3. Only **first-time** diagnoses and prescriptions considered.

The imposed criteria is strict: taking diagnoses and prescriptions on the same day means removing *all prescriptions* following the first diagnosis. In other words, all the instances of patients coming back to their GP to renew a prescription (e. g. for a chronic disease) have been deleted.

This approach can be useful to extract a cohort of patients beginning their treatment, and analyse the variations of first-time prescriptions, especially for chronic illnesses. It's important to notice that **not all doctors** may have patients getting new diagnoses.

7.1 Imposed criteria

Aside from completeness and correctness of the data, there are more restrictions to maintain consistency:

- The prescribing general practitioner should not change in the time range;
- The patient should not be deceased;
- There should be a sanitary convention;
- The general practitioner should be active.

All those constraints can be checked using the related fields in the database: *revocation* for interruption of the relationship, *death* for death and *convention* for the sanitary convention.

The table *users* contains all the IDs of active general practitioners in 2018, therefore joining it with other tables is the best method to remove all the rows with an inactive GP. Data is up to date, and since the patient journey includes 2018 and requires consistency of history (the focus is on the most recent information) there is no need to check for active GPs in the previous years.

The biggest risk is again the **loss of information**: the impact of data cleansing is heavy, and the obtained results might not give an insightful perspective.

7.2 Examples of data cleaning

An initial data cleaning has been made on the whole database to have a first understanding of the potential information loss.

In this case, having such a big amount of tuples is useful: it is possible to remove a considerable percentage of them without losing generality and still having numerous samples.

Information on the general loss is already available thanks to the specific analysis on each single field, therefore the shown data cleaning will only consider patients and GPs.

The active general practitioners are **432**: this result has been retrieved counting the different IDs in *users* (438) and removing the ones not present in *patients_ doctors* (6).

About half of patients is going to be lost, due to not respecting the consistency criteria. Starting from a million of records, concrete results are still obtainable.

7.2.1 ICD-9

ICD-9 codes have been removed after a comparison with the official WHO database.

The 2 669 312 removed tuples have the following issues:

- Wrong code: 0.005%;
- Empty code: 75.4%;
- Empty description: 24.6 %.

The total number of rows is 15 460 199, of which 17% must be deleted due to falling among one of the above categories.

7.2.2 ATC

ATC have been removed according to the same procedure as ICD-9.

The 3 328 041 removed tuples have the following issues:

- Wrong code: 52.6%;
- Empty code: 47.4%.

Those consist in 21,5% of total.

Further statistics related to data loss are extracted after having a first set of results, comparing individual values to the original after each step of restriction imposition.

7.3 Patient-based approach

The first approach consists in testing with an **arbitrary range constraint**: all dates must fall in the span between 2010 and 2018.

The first analysis has pure research and testing purposes, to understand the impact of cutting the dataset in terms of information loss. All the previously introduced criteria must be considered as well: there must be a continuous doctor-patient relationship between active GPs and non-deceased patients with sanitary conventions.

To summarise, the obtained slice of data comprehends only patients starting their journey from 2010, having diagnosis and prescription on the same date by an active GP.

The outcome is a patient journey table containing data from 2000 to 2018, with a total amount of **144 618** tuples: this means that there are roughly 150k first-time diagnoses and prescriptions to patients.

7.3.1 Results breakdown

Seeing that the starting tables had number of rows in the order of millions, some deeper analysis is necessary to figure out the uses of this significant loss.

The 144 618 complete tuples are composed by:

- 27 733 patients;
- 422 general practitioners;
- 1 381 unique diagnoses;
- 904 unique prescriptions.

Further uses for those low values can be found counting how many dates would not fall in the considered range. The percentage of records with date earlier than 2010 in each table is:

- Patients: 75%;
- Diagnoses: 54.6%;
- Prescriptions: 44.7%.

It can be seen that most data loss is used by patients, most likely starting their treatment prior to 2010.

Patients

The following *pie chart* illustrates the data loss on patients according to the criteria defined in the previous section.

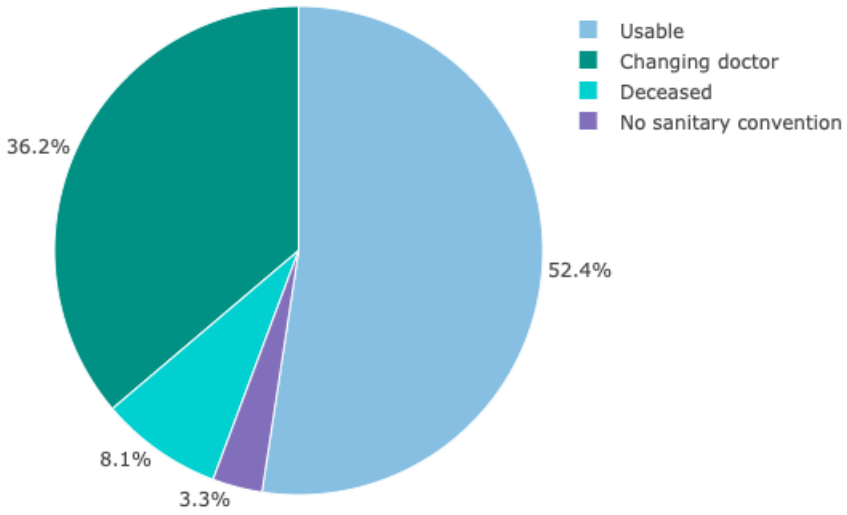

Figure 7.1: Information loss on patients

The total amount of unusable patients is 58 415, not enough to use a loss of 75%: the reason is therefore cutting off all data before 2010.

8 years is a too wide range to obtain a consistent patient journey, and such information loss isn't negligible: the conclusion of the first approach is that introducing time boundaries is something which needs an accurate control, to avoid missing out most of the data.

7.4 Prescription-based approach

Since extracting a slice of records according to their date span is an abrupt approach, careful parameter tuning and detailed constraints introduction is necessary to have a solid and consistent amount of information.

The focus is on the **number of prescriptions**: given a small range of years, only patients with *at least one new prescription* are going to be considered. All the previous criteria must be respected (continuous doctor-patient relationship between active GPs and non-deceased patients with sanitary conventions).

This methodology allows cluster sampling without having to remove half of the dates: criteria based on the number of prescriptions creates another patient cohort which can be used to accurately select rows from the other tables.

The proposed time range is **2016-2018**: pharmaceutical companies generally use the last two years of sales, so picking the last three years gives additional information without

compromising the consistency of data.

The outcome is a patient journey consisting of 1 465 005 tuples: almost 10 times the previous result. This leads to two important statements:

1. The time range is appropriate, since the number is large enough to make analysis without loss of generality;
2. The new imposed criterion gives more consistent data and the possibility to build time series.

Further cleaning is required to link diagnoses and prescriptions, since there is not an univocal correspondence: multiple diagnosis and prescriptions may be associated to the same date.

7.4.1 Results breakdown

The 1 465 005 complete tuples are composed by:

- 230 381 patients;
- 412 general practitioners;
- 4 324 unique diagnoses;
- 1 280 unique prescriptions.

Only 7% of the total prescriptions have been taken into account, yet a million and half is still a consistent amount.

Funnel graph of patients

A funnel chart is used to visualize the progressive reduction of data as it passes from one phase to another. Data in each of these phases is represented as different portions of 100% (the whole amount of patients)[1].

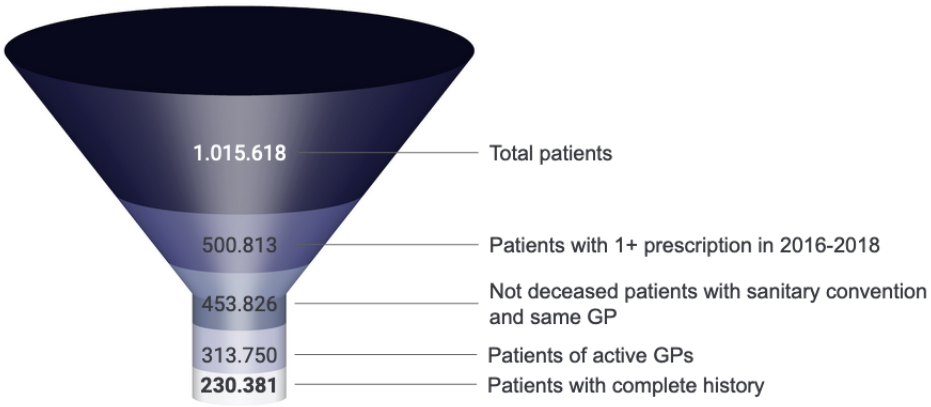

Figure 7.2: Information loss on patients, funnel graph

[1]Funnel chart, Fusioncharts

Imposing every restriction made the patients number decrease: starting from a million, the remaining ones are 1/4 of it.

The constraint removing the biggest part consists in having at least one prescription, meaning that half of patients is generally healthy and most prescriptions are related to a smaller subset of individuals.

7.4.2 Examples of analysis

Resulting information, despite needing further checking and lookup, is a starting point for time series analysis.

Prescription indicators

There are 1 465 005 prescriptions, 1 280 of which are unique.

Average new prescriptions per patient:

Year	Average
2016	3.26
2017	3.37
2018 (incomplete)	3.06
All years	5.16

Table 7.1: Average new prescriptions

The range goes from 1 to 129 new prescriptions associated with diagnosis in three years.

Most common couples

The 5 most common diagnoses and prescriptions happening on the same day are displayed in the table below.

Diagnosis	Prescription	Count
Vitamin D deficiency, unspecified	Colecalciferol	13 122
Periapical abscess without sinus	Amoxicillin and beta-lactamase inhibitor	7 761
Esophageal reflux	Pantoprazole	6 093
Diarrhea, unspecified	Rifaximin	6 000
Cystitis, unspecified	Fosfomycin	5 885

Table 7.2: Most common couples of diagnosis-prescription

The first one is a type of vitamin D, and Pantopranzole is a medicine for digestive tract diseases: all the others are antibiotics.

This prior analysis already shows a concerning amount of antibiotic prescriptions.

Most common antibiotics

Antibiotic prescriptions compose 20% of total, which consists in 300 503 instances. The most popular ones are:

1. Amoxicillin and beta-lactose inhibitors, 62 560 prescriptions;
2. Ciprofloxacin, 23 762 prescriptions;
3. Levofloxacin, 22 595 prescriptions;
4. Rifaximin, 21 680 prescriptions;
5. Ceftriaxone, 19 523 prescriptions.

7.5 Results comparison

Comparing the two patient journey outcomes through graphs is a good way to visualize changes and improvements.

7.5.1 Changes in data composition

Figure 7.3 shows barplots highlighting the changes between data categories in the two approaches.

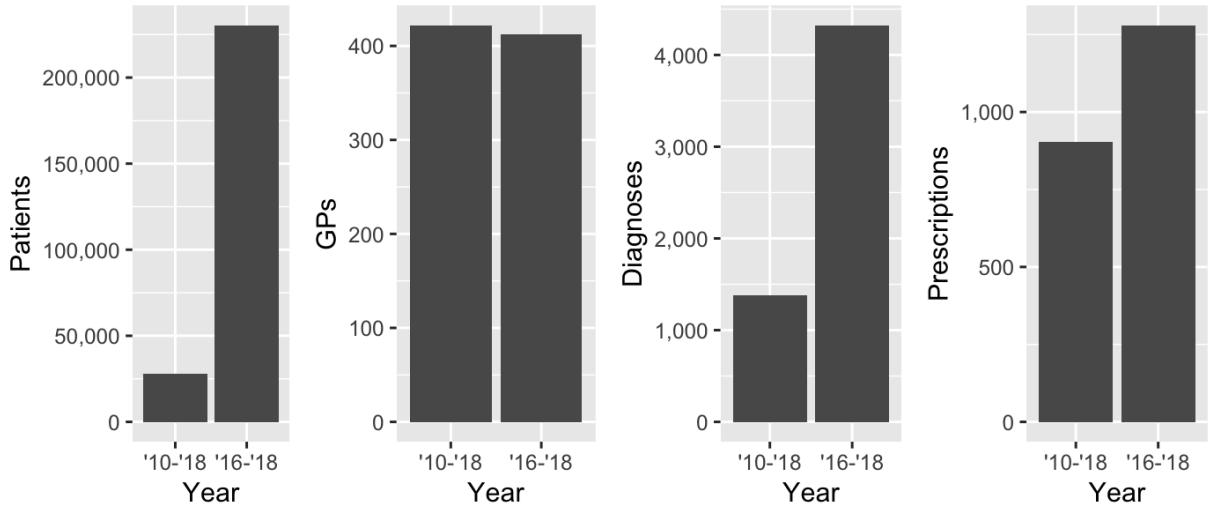

Figure 7.3: Information loss barplots

General practitioners' number is stable, yet all the other categories are subject to a substantial improvement in the prescription-based approach.

Patients, in particular, change from 27 733 to 230 301.

7.5.2 Improvement on patients information loss

A pie chart for data loss on patients in the range 2016-2018 has been made to compare with the previous one.

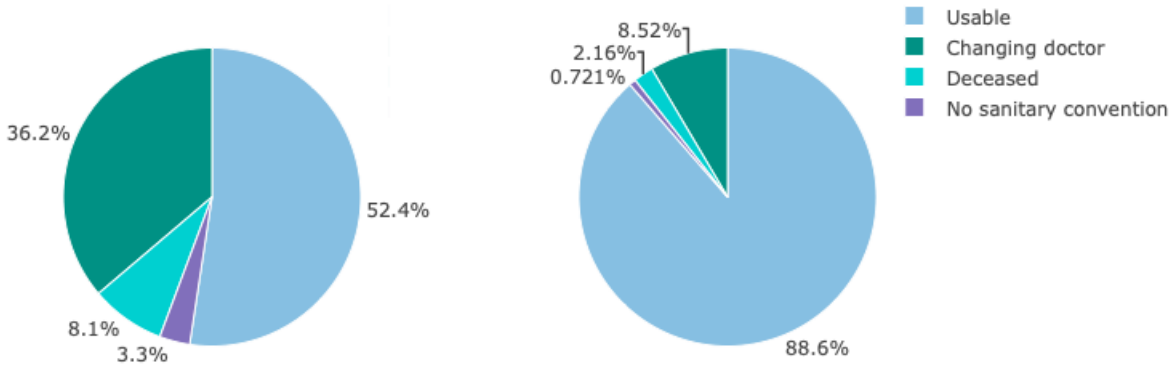

Figure 7.4: Information loss comparison pie charts, 2010-2018 on left, 2016-2018 on right

It is confirmed that the number of usable patients has noticeably increased: using a smaller time span reduces the chances of death and change of GP.

7.6 Future work

Further analysis is needed to assess an insightful perspective focussed on specific diseases, including *prescriptions following the first related diagnoses* and removing unrelated ones having the same date.

Examples of future work are:

- Definition of a criterion so that a patient is considered chronic (minimum amount of prescriptions);
- Selection of GPs and patients having a number of diagnoses and prescription within a range (for data consistency);
- Deeper analysis of prescription changes during the years;
- Analysis on specific diseases (digestive system and tumours);
- Time series analysis and clustering approaches.

Since research shows a considerable amount of antibiotic instances, the data is linking to the issue of antibiotic resistance. More detailed analysis using time series can give a wider perspective.

Chapter 8

Antibiotic trends analytics

This chapter deals with antibiotic resistance, highlighting values through aggregation and visualisation of time spans to verify trends and variations.

Having obtained a general analytical view on couples of first-time diagnoses and prescriptions, and cross-checking those results with global trends, the research can focus on **antibiotic prescriptions** to identify **changes of patterns.**

Defined time frames in limited temporal windows are selected to make a detailed trajectory analysis related to prescriptive appropriateness compared to antibiotic resistance, without going into details of pathologies.

Such analytics are useful for research, highlighting rising health issues alongside the AIFA reports, and to support pharmaceutical companies using the potentiality of healthcare data. Having information about AICs (different for each product) allows to give a new insight not only on ethical matters, but also on the *global market trends.*

8.1 Identifying antibiotics

Antibiotics, also known as antibacterials, are *medications produced by microorganisms that destroy or slow down the growth of bacteria* [36]. They include a range of powerful drugs and are used to treat diseases caused by bacteria.

A doctor can prescribe a broad-spectrum antibiotic to treat a wide range of infections. A narrow-spectrum antibiotic is only effective against a few types of bacteria. In some cases, a healthcare professional may prescribe to prevent rather than treat an infection, as might be the case before surgery.

ATC codes related to antibiotics divided by class are:

- A01AB, A02BD, A07A;
- D01, D06, D07C, D09AA, D10AF;
- G01;
- J, the whole class;

- R02AB;
- S01A, S02A, S03A (removing hormones).

8.2 Subset extraction

The first criteria to impose, considering analytics is going to be made on antibiotic prescriptions, is the actual presence of them. This can be reworded by removing all data related to patients who never had a prescription of an ATC code among the ones previously listed.

There are 794 267 patient with at least one antibiotic prescription in the whole time range (2000-2018), whose total prescriptions consist in 114 044 470 tuples. This implies the remaining ∼200 000 patients only have ∼4m of prescriptions, which is probably justified by a healthier physical state.

Since antibiotics patterns are subject to major changes during years, a big amount of data can be dispersive: pharmaceutical companies make analysis only considering 2-3 years, yet for research purposes an *intermediate range* is the most informative.

A good time span consists in 10 years, considering most recent ones: since 2018 has to be excluded due to incompleteness, 2008-2017 is the final choice.

Progressive data loss has to be considered, imposing additional constraints of completeness and accuracy. The latest version of the used dataset is a subset of the table prescriptions, with the following restrictions:

- AIC corresponding to an antibiotic;
- Prescription date between 2008-01-01 and 2017-12-31;
- Active general practitioners;
- Patients with usable information about gender, date of birth and location.

The obtained record set is composed by 8 386 057 tuples, each representing a single prescription.

8.3 Global statistics

Global statistics help to have a general overview of the data, aggregating values to identify most impactful changes and the broad behaviour.

Exploration analysis is performed again, extracting central tendency and dispersion to then perform time series analytics.

8.3.1 Average prescriptions

The average number of prescriptions has been calculated only considering the number of patients with at least one antibiotic prescription from 2008 onwards (670 634).

Year	Mean	Standard deviation
2008	2.72	2.73
2009	2.75	2.78
2010	2.72	2.76
2011	2.69	2.73
2012	2.68	2.76
2013	2.74	2.84
2014	2.79	2.88
2015	2.79	2.79
2016	2.77	2.91
2017	2,72	2,84

Table 8.1: Statistics on patients with antibiotic prescriptions

Values are calculated only considering the number of patients who received at least one prescription in the determined year. Results are shown in table 8.1.

Values show the mean tends to slowly increase, and although there are patients getting up to 30 antibiotic prescriptions each month, the standard deviation is low: the mode both for year and month is 1, hence lower values weigh more.

Statistics related to the number of patients with **only one** antibiotic prescription, making statistics comparing time spans between 2008-2017:

Interval	Mean for year	Standard deviation
Year	118 301.4	2 831.65
Month	40 130.83	6 004.2

Table 8.2: Statistics on patients with one prescription

8.3.2 Yearly prescriptions

The number of yearly antibiotic prescriptions is:

Year	Antibiotics
2008	813 337
2009	846 067
2010	797 472
2011	777 582
2012	760 280
2013	798 320
2014	812 594
2015	802 323
2016	776 756
2017	727 723

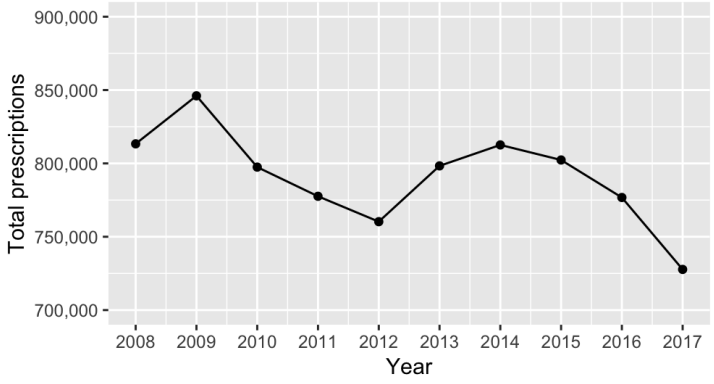

Table 8.3: Yearly antibiotic prescriptions

The 2012 fall might be caused by a substantial decrease of the approved antibiotics by AIFA [37], but is noticeable that the number rises starting from 2013 due to antibiotic resistance.

8.3.3 Seasonal prescriptions trends

A barplot is the most informative way to visualise trending of number of prescriptions and highlight seasonality.

For winter and summer months, the amount of patients is crossed with the total prescriptions. The x-axis shows the number of prescription, while y shows how many patients got that number. A logarithmic scale us used to standardise the quantity of patients having a smaller number of prescriptions (1-2), contrasting skewness towards large values.

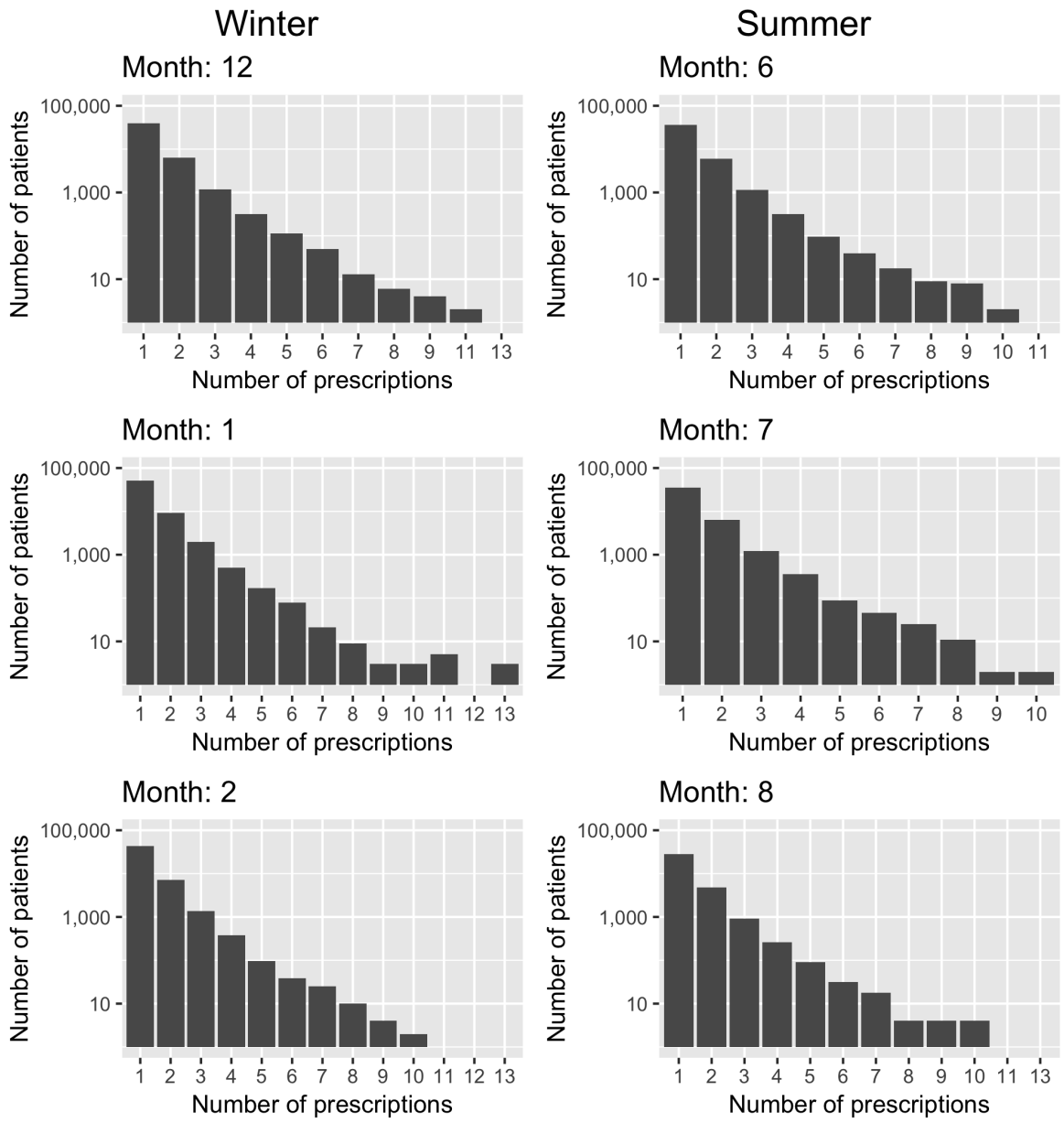

Figure 8.1: Number of prescriptions based on patients (2017)

The considered year is 2017. The majority of patients gets only one prescription in all months, yet values change for instance from 28 000 to 40 000 from August to January.

Furthermore, the number of patients having more than 2 prescriptions increases in winter: they are around 3 000 in January and less than 1 500 in August.

8.4 ATC rankings

For each year and each month, the 3 most common antibiotics are extracted, according to their ATC code. Rankings show different popular antibiotics for each year, therefore there are more than 3 labels in the plots.

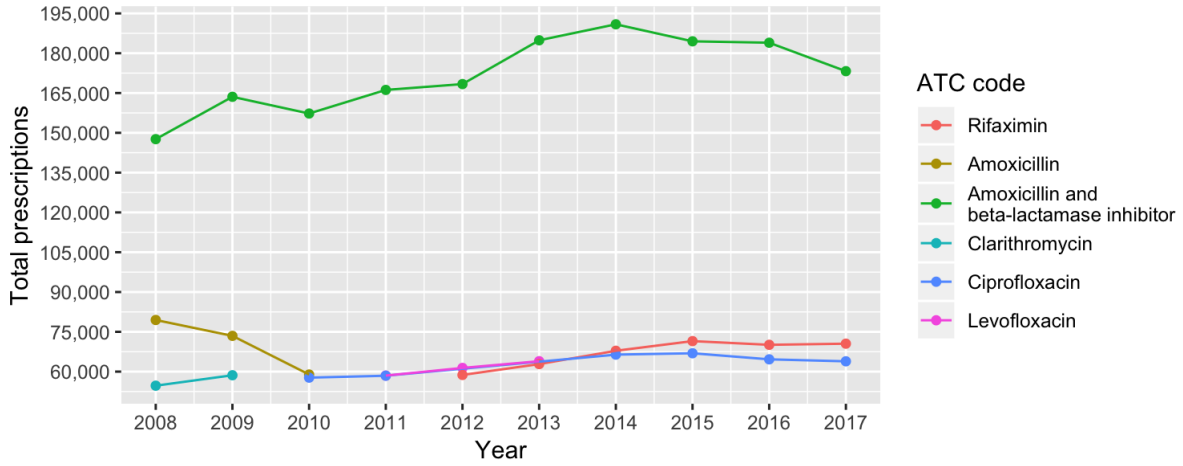

Figure 8.2: Top ATC for years

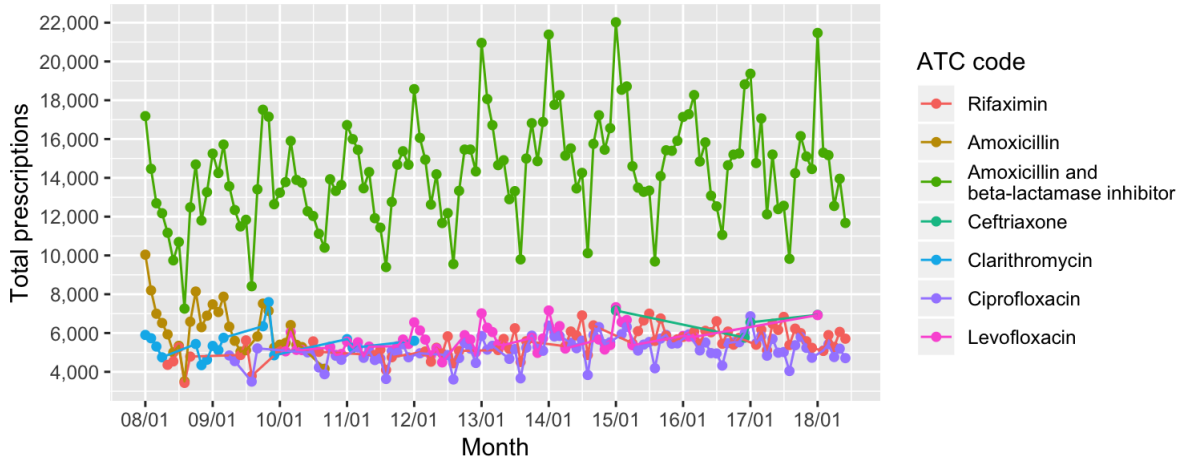

Figure 8.3: Top ATC for month

Most ATCs follow a similar pattern: the amount decreases in 2010-2012 (years of economic crisis and revoked authorisations), to then have a small peak in 2016 and get lower again in 2017, similarly to the global amount of prescriptions.

The most noticeable unusual trend corresponds to **amoxicillin and beta-lactose inhibitors** (ATC J01CR02), which considerably detaches from the others. Amoxicillin is one of the most popular group of drugs, confirmed by global analytics on the database, AIFA reports and USA trends [38].

It is evident that the monthly prescriptions follow a seasonal trend: the number of prescription rises during the winter and falls during the summer.

8.4.1 Digestive tract antibiotics

Since ICD-9 codes have some unusual trends, comparing them with antibiotics might provide additional information. The interested area comprehends ICD-9 class A, diseases of the digestive tract, whose diagnoses have doubled in the span of 10 years.

Antibiotics to treat this subset of diseases include:

- Nyastatin;
- Rifaximin;
- Paromomycin.

The purpose of analysing digestive tract antibiotics is the highlight of peaks of prescriptions, to understand whether they follow the same trends as Omeprazole and Pantoprazole (chapter 5.4). Those values are subject of an increase of more than 100 000 prescriptions from 2010 to 2017.

Cross-checking related codes with the antibiotics dataset, rifaximin is indeed one of the most prescribed drugs, while its numbers are still considerably inferior to amoxicillin (the most popular antibiotic). Seasonality is still present, yet less pronounced.

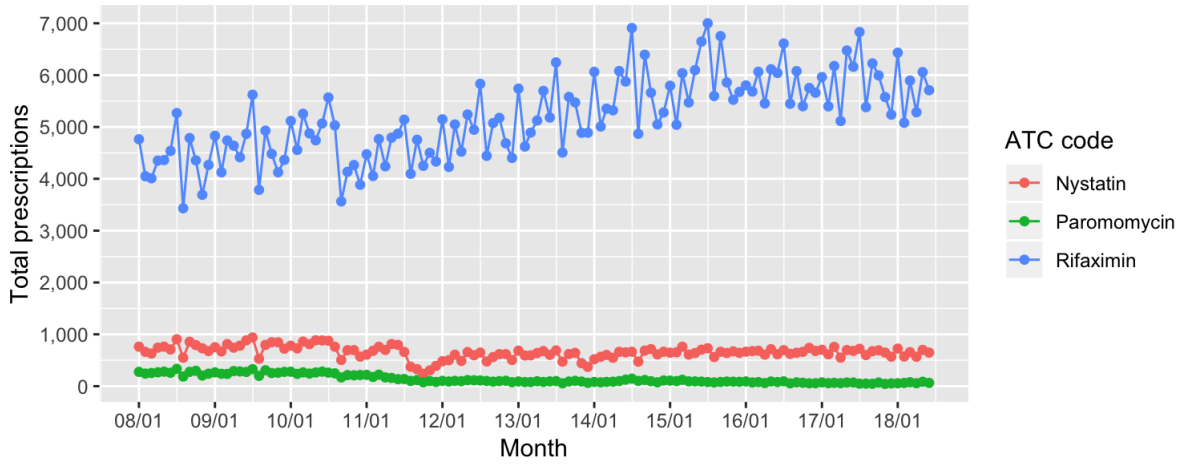

Figure 8.4: Seasonality of digestive tract antibiotics

In 2017, all prescriptions of those antibiotics are 79 138, although there is a growth of about 10 000 from 2010. This value is still not comparable to 350 000 prescriptions of Pantoprazole, therefore it is fair to assume trends of antibiotics for digestive tract are not related to the considerable increase of other medicines.

8.5 AIC rankings

AIC rankings are useful to have an insight not only on active principle classes and treatment of diseases, but also on pharmaceutical products, companies and popularity within general practitioners.

AIC is, in fact, different for each medicine based on brand and dosage, despite having the same AIC and belonging to the same class.

This allows to identify external causes for unusual trends, such as marketing campaigns, aggressive advertisement and pressure to use a specific drug rather than an equivalent one.

Prescriptions with AICs do not have a public lookup table, yet drugs can be singularly searched on a public database by AIFA. Analytics have been made on the antibiotics subset of data, using the 5 most prescribed medicines.

Trends are shown for both months and years, focussing on the gradual growth.

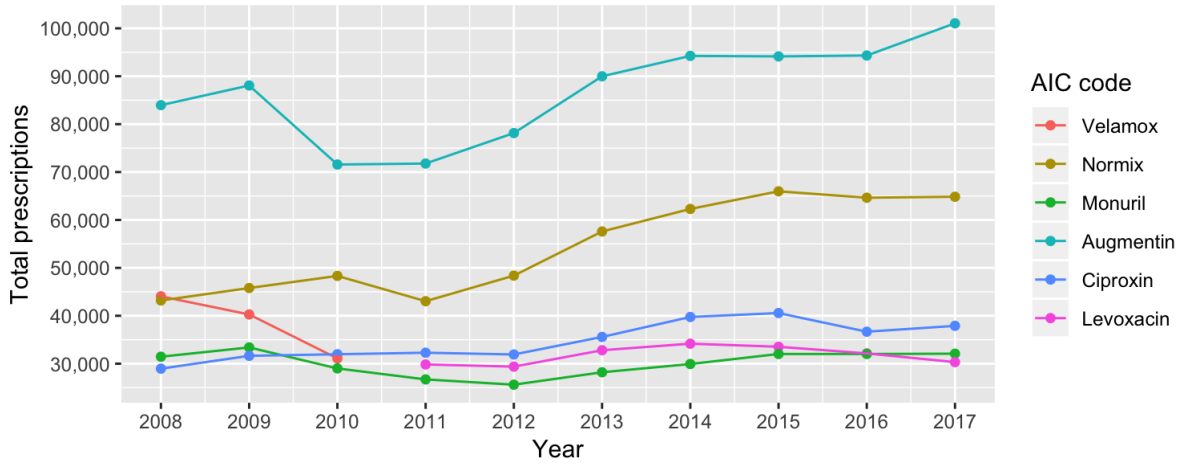

Figure 8.5: Top AICs for years

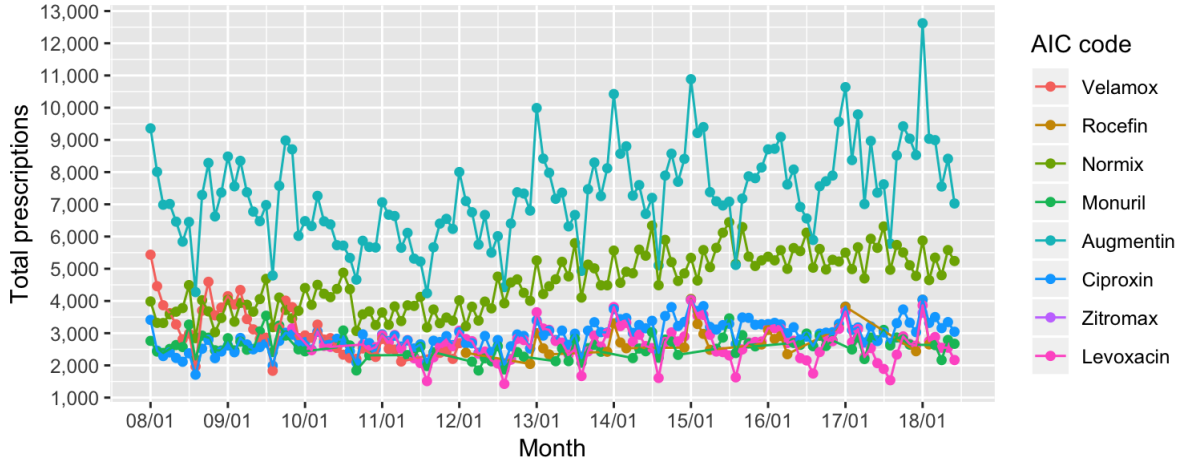

Figure 8.6: Top AICs for months

Augmentin is the top ranked prescription, detaching from the others and with a progressive increase starting from 2010: values go up to 100 000 prescriptions in 2017, and January 2017 is the month with the maximum number.

Having a difference of 30 000 between 2010 and 2017 without a correspondent increase of diagnoses shows frequent episodes of antibiotic resistance and overprescribing.

A better understanding of magnitude orders can be obtained counting the total prescriptions in 10 years of the most common AICs:

Code	Prescriptions	Antibiotic
026089019	934 942	Augmentin 875 mg + 125 mg 12 coated tablets
025300029	586 782	Normix 200 mg 12 coated tablets
026664021	373 720	Ciproxin 500 mg 6 coated tablets
033940038	264 604	Levoxacin 500 mg 5 coated tablets
025680024	229 816	Monuril 3 g 2 sachets
023097102	132 787	Velamox 1 g 12 dispersible tablets

Table 8.4: Total prescriptions of top antibiotics

The drug Velamox has an opposite behaviour compared to the others: while most medicines are increasing their number of prescriptions, Velamox is subject to **progressive decrease** until disappearance from the most prescribed drugs.

8.5.1 An insight on Velamox

Velamox, according to AIFA, is an amoxicillin-based drug used for respiratory trait, ear and genital infections. It is sold in different packages, whose the most popular is the 1 g dispersible tablets with 12 tablets.

Since the general trending of most prescribed antibiotics only gives a partial vision of the comparisons between products, a complete chart is extracted using Velamox and three other popular drugs: Augmentin, Normix and Levoxacin. The difference is evident.

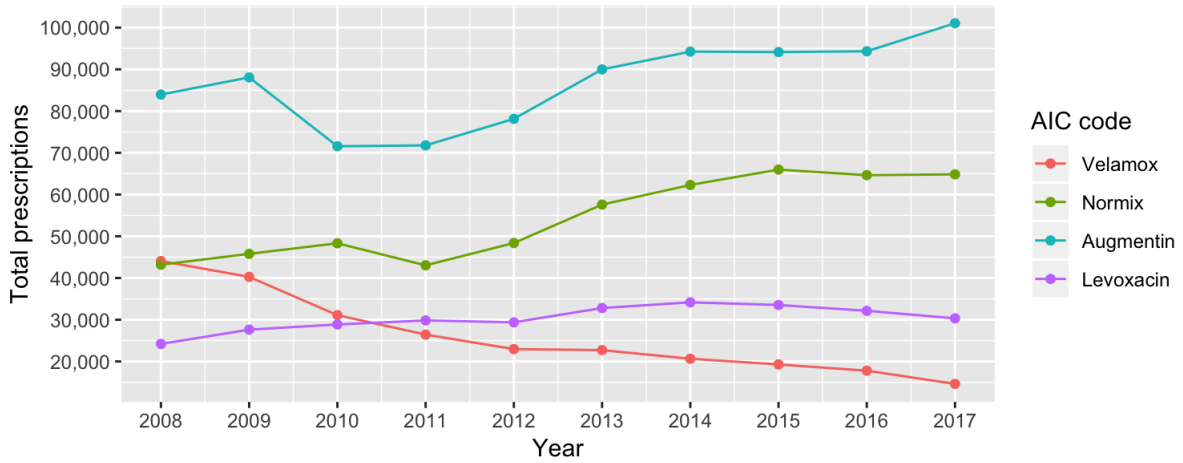

Figure 8.7: Velamox trends for years

The first hypothesis of such a drastic drop of a product is its switch with another one, changing prescriptive habits of doctors. If that was the case, graphs related to the specific subset of patients would have a correspondent rise of values belonging to another antibiotic.

Velamox in 2008, the year of maximum amount of prescriptions, has been given to approximately 32 000 patients. Extracting only the top 3 antibiotic prescriptions of those, in the last 10 years, along with Velamox, shows no substitution by another drug.

The fall starts in 2009, having a drop of 30 000 units, and gradually continues until 2017 with only 3 037 prescriptions.

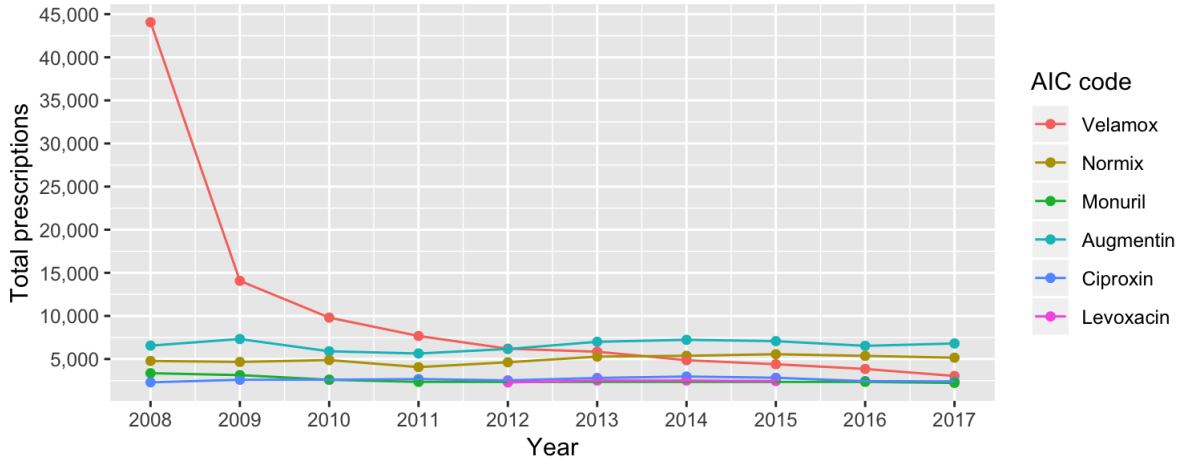

Figure 8.8: Prescriptions of Velamox subset of patients

Additional analysis is made on the subset of 32 000 patients:

- 1 500 are dead;
- 8 000 have changed GP;
- 18 000 are females;
- 14 000 are males;
- 16 000 are born before 1960.

General practitioners assigned to the subset of patients have generally decreased their prescriptions quota, roughly halving it comparing 2008 to 2017.

Another hypothesis is the general decrease of the active principle, of which Velamox is the main representative medicine.

ATC trends related to Velamox's AIC, however, show usual patterns: since the same ATC is linked with more products, the curve is still leaning upwards, yet there is an offset equal to the decrease in Velamox.

Since analysis on a subset of patients gives little insight, a wider time range can give additional information on a complete picture of global trends.

Further knowledge is obtained counting the total number of Velamox prescriptions among all patients from 2000 to 2017 (figure 8.9).

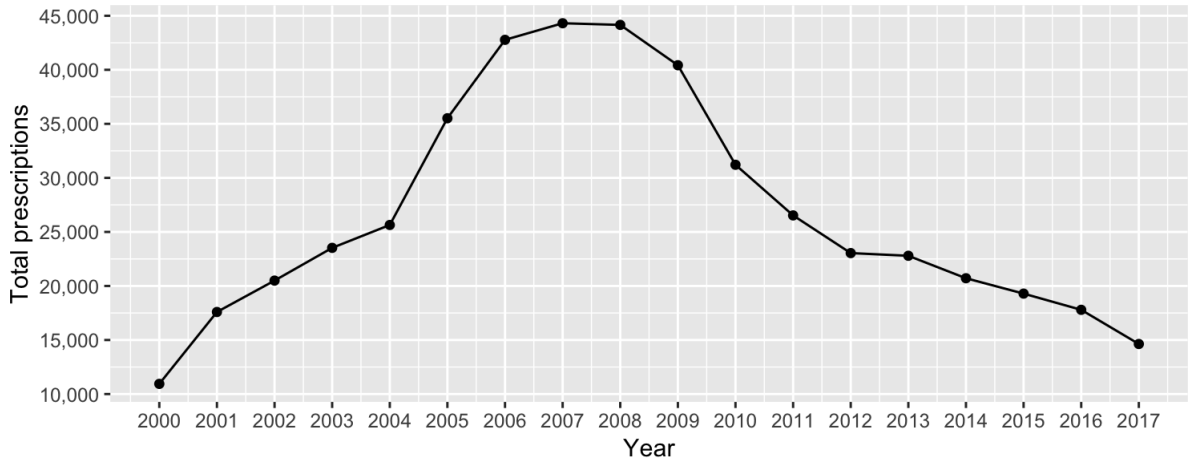

Figure 8.9: Velamox trends in 2000-2017

Velamox has acquired popularity from 2005 to 2010, when antibiotic resistance still hadn't been recognised as a problem and before the advent of generic drugs, yet this does not provide an explanation for the decrease.

AIFA has also revoked authorisation for most instances of Velamox packages, only leaving three out of ten. The company which produces the drug, Mediolanum Farmacy, has been acquired by Neopharmed Gentili in 2018.

Summarising, changes in Velamox prescriptions are partially caused by:

- Doctors gradually abandoning it after authorisation revoking;
- Patients dying or switching doctor;
- Substitutions with other drugs, none of whose is prescribed enough to have a significant trend, such as generics;
- Different advertisement campaigns after the company acquisition.

8.5.2 An insight on Augmentin

After analysing a drug with an unusual decrease, the focus can shift to the opposite side, considering the one with a noticeable increase.

Augmentin is the most prescribed antibiotic, with an almost constant positive trend.

2010 and 2011 have lower values, possibly explained by the Italian economic crisis or the diffusion of generic drugs. Starting from 2013 numbers continue to rise, reaching their highest peak in 2017.

Most Augmentin consumers are adults in the range of 45-64 years of age, following global trends previously obtained. The range of 15-24 years of age does not contribute to the progressive increase.

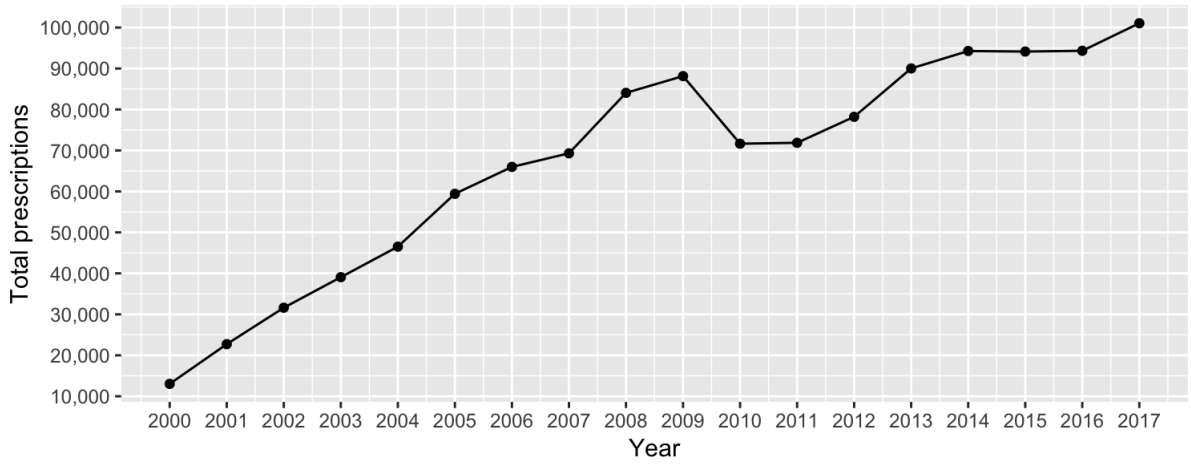

Figure 8.10: Augmentin trends in 2000-2017

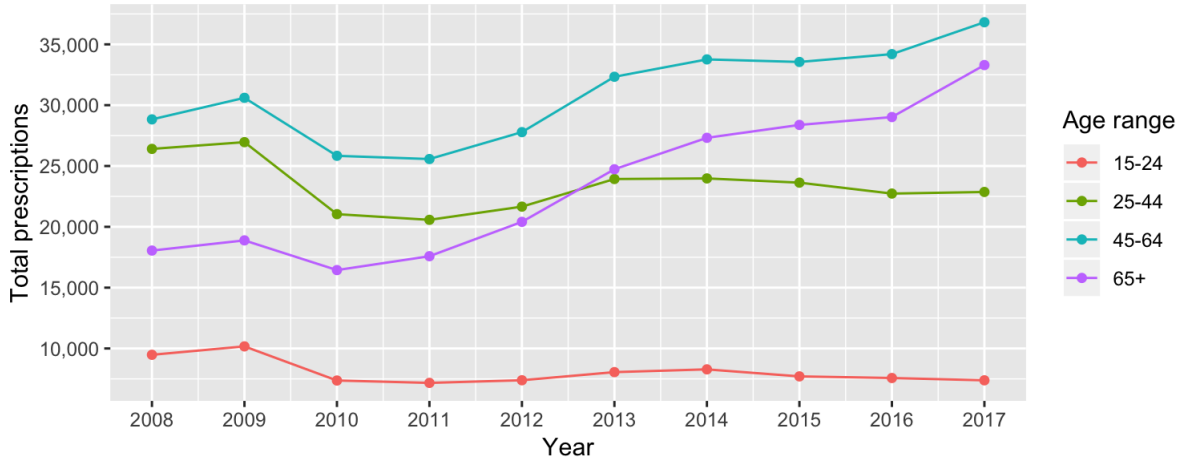

Figure 8.11: Augmentin trends in for age and year

8.5.3 Analysis on geographical area

There is information loss on geographical area of patients, therefore it is necessary to extract a further subset of antibiotic prescriptions, applying an additional level of cleansing.

Patients with complete and correct province of residence fields are selected, obtaining:

- 590 472 patients, 74,34%;
- 7 458 312 prescriptions, 88,94%.

Campania has five provinces: Naples, Avellino, Caserta, Benevento and Salerno. Analytics are expected to show an uniform distribution of prescriptions among those, yet outcomes are skewed towards Naples (significantly bigger) and other smaller cities.

Prescriptions come from 932 different cities, while there are 550 distinct cities in Campania. This implies a slice of patients having residence in a different place outside the region.

Results might be altered by the non-uniform provenance of doctors and patients (mostly

from the Naples area), and other big municipalities have large population although not qualifying as provinces.

Another factor of influence consists in doctors inserting the general CAP of the region (80100), which corresponds to Naples, in the database, instead of looking for the specific one of each city.

Results can be summarized listing the 5 municipalities for total prescriptions in 2008-2017:

Postcode	City	Prescriptions
063049	Naples	2 854 056
063002	Afragola	618 519
063024	Castellammare di Stabia	497 309
063005	Arzano	314 640
063035	Gragnano	210.960

Table 8.5: Prescriptions per province

It can be seen that none of those cities aside from Naples belongs to the provinces. The population of each smaller place is around 50 000 inhabitants, making the average prescriptions per individual high.

Discrepancies among provinces are also shown by the following plot, identifying trends:

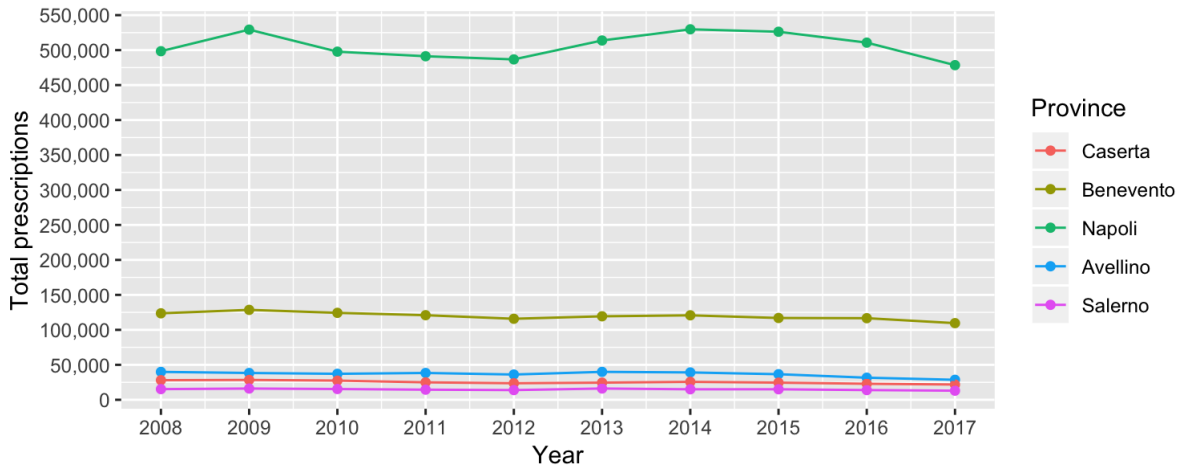

Figure 8.12: Prescriptions trends per province

Pharmacies

The number of pharmacies for each province with related population (ISTAT 2018) confirms the gap between Naples and the others:

1. 857 pharmacies in the Naples area, 3 101 002 inhabitants;
2. 375 pharmacies in the Salerno area, 1 101 763 inhabitants;

3. 254 pharmacies in the Caserta area, 923 445 inhabitants;
4. 161 pharmacies in the Avellino area, 421 523 inhabitants;
5. 107 pharmacies in the Benevento area, 279 127 inhabitants.

8.6 ATC and AIC correlation

Since both ATC and AIC trends show seasonality and unusual increases of amoxicillin-based products, both are compared to underline potential correlation.

Statistics related to amoxicillin and beta-lactose inhibitors and Augmentin are extracted and plotted together.

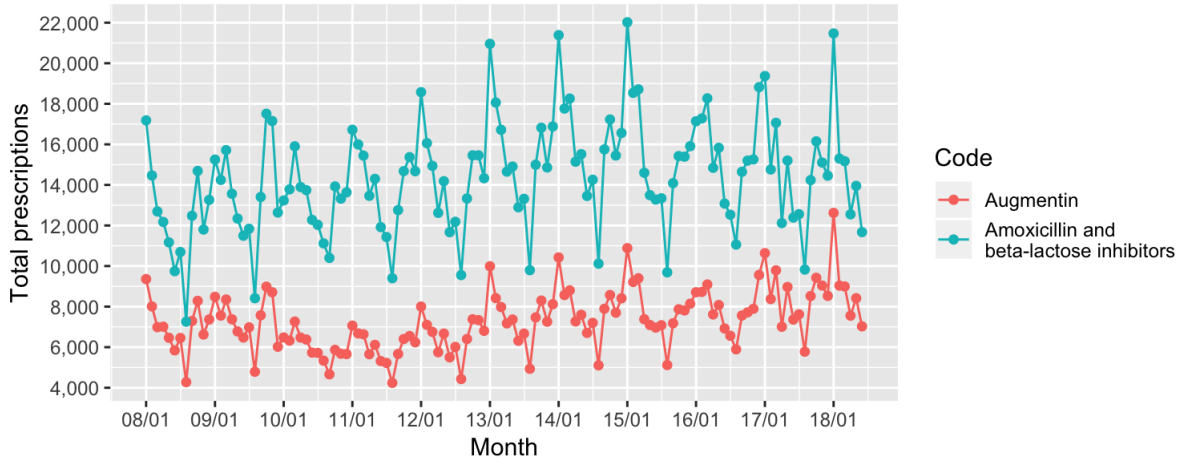

Figure 8.13: AIC and ATC trends

The two sets of values have a **Pearson correlation** of 0.88107, enough to affirm there is correlation between them:

- Amoxicillin and beta-lactose inhibitors mean: 14 369.53;
- Augmentin mean: 7 308.57;
- Mean of differences between couples of points: 7 011.5.

About 50% of amoxicillin and beta-lactose inhibitors prescriptions is composed by Augmentin, a relevant amount confirming its popularity.

Chapter 9

k-means approach for cluster analysis

This chapter illustrates the implementations of the k-mean algorithm on a matrix of general practitioners and prescriptions for antibiotic, to classify them according to their behaviour.

***k*-means clustering** is the most commonly used unsupervised machine learning algorithm for dividing a given dataset into k clusters.

This method related to antibiotic resistance research consists in applying a time series clustering approach to *group general practitioners according to their prescriptive habits of antibiotics*, using medicines as profiling instrument to obtain models of usage for each product.

Additional layers (factors) to consider while profiling are space, time, structure and personal information of patients, to highlight potential behaviours to classify doctors and co-prescription activities.

Data has therefore to be subject of parsing, grouping values b to chosen features and selecting subsets based on consistency, maximising potential given information.

Records will be mapped into a n-dimensional matrix which will be fed to the algorithm, obtaining a cluster of belonging for each element. Dimensions will represent **individuals**, **relevant features** and **time**, yet time slices will be taken separately to be able to make comparisons and predictions.

Results aim to identify whether general practitioners are "loyal" prescribers: having most prescribed antibiotics, checking whether there exist typologies of patterns according to active and illustrative variables.

k-means is the chosen clustering algorithm, since the approach is non-hierarchical and simple: its minimum computation complexity and ease of use make it the most suitable algorithm to perform an initial clustering on raw data, reducing the space into disjoint smaller sub-spaces to then perform additional analysis.

After identifying the clusters and linking each one with a specific prescription patten, it is possible to discriminate between GPs with *changing* and *constant* behaviours among time.

9.1 Range of time

Since the aim of clustering is identifying prescription patterns and their changes, extracting **different time slices** allows better insights on evolutions of trends.

To avoid the impact of seasonality, having values increasing in the winter months, data is collected and aggregated in a range of **one year**: the final table contains cumulative results and features, without distinction between different times of the year.

Comparisons are made selecting two snapshots of data according to different years, and visualising outcomes to understand whether values have shifted cluster (general practitioners have changed habits).

Data in two years needs to be distant enough to ensure the presence of potential changes, yet consistent enough to prevent information loss.

Selected years are **2010** and **2017**, 2017 being the latest complete time range and 2010 being the first year considered for antibiotics analysis.

9.2 Dataset construction

Before being able to apply a clustering algorithm, data has to be cleaned, arranged and aggregated: having a format of single prescription records would not give the desired outcome, since grouping has to be made according to *counts in years*.

Transposing information from a row to a column visualisation allows to emphasise the values of each feature.

The most important constraint while constructing the matrix is consistency of data: all general practitioners must be active in both years, to make comparison possible.

The number of constantly active doctors is 372, having respectively 228 022 and 214 316 patients in 2010 and 2017. This implies a data loss of about 75% in both sets.

Features to consider are:

- Antibiotic prescriptions;
- Patients data;
- Other prescriptions' habits data.

To limit horizontal expansion of the matrix, a limited number of attributes are taken into account, chosen by their relevancy:

- Antibiotic prescriptions are broken and counting according to the **8 most popular products** (Augmentin, Normix, Ciproxin, Levoxacin, Monuril, Velamox, Rocefin, Zitromax);
- Patients data only includes **gender** and most common **age range** in $\{1, 2, 3, 4\}$.

Other prescriptions' data has been limited to the *total number of prescriptions* (not strictly related to antibiotics) for each year. Information about the total amount of

prescriptions might be helpful to understand the percentage of antibiotics respecting to the whole, and therefore their influence.

According to the goal of the analysis, every sample is represented as a **13-dimensional vector**, with each row containing information on a single general practitioner.

Displayed below is an extract of the final table for 2017:

#	**Doctor ID**	**Aug.**	**Normix**	...	**M. patients**	**F. patients**	**Age**	**Prescriptions**
1	DRTQRGJ	5	123	...	303	397	3	17 862
2	DSJRY7B	242	131	...	322	399	4	15 845

Table 9.1: k-means matrix extract

Input matrix has dimensions of $372 \times 13 \times 2$, where the latter is an additional time dimension which distinguishes between 2000 and 2017, removed splitting the dataset to run the algorithm separately in the two years.

9.3 Features selection

Having too many features, despite the initial grouping, may lead to overfitting and having general practitioners clustered according to irrelevant information.

Among all features, some are relevant for the outcome of the algorithm, while others are purely descriptive to be checked after clustering results.

All the amounts of *prescriptions* for each antibiotic (columns 1-9) have to be considered to extract similarity, and scaled to normalise numbers. Those count as **active variables**, and clustering will be performed considering the subset of belonging.

Doctors' IDs are removed during computation, to then be added back to map each row with its belonging individual.

The other attributes are **illustrative**, and will be attached to clustering results, to offer further information to be compared after having a general idea of each doctor's group.

9.4 Optimal number of clusters

Determining the most suitable number of clusters in a data set is a fundamental issue in k-means clustering, which requires the user to specify the number of clusters k to be generated. There are various approaches to pick the best value, yet the method is ultimately subjective [40].

9.4.1 `NbClust`

`NbClust` is a R package which provides 30 indices for determining the number of clusters and proposes the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods.

Both datasets have been tested with `NbClust` using number of clusters in $\{2, 20\}$.

Results for 2010:

```
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 7 proposed 8 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 2 proposed 15 as the best number of clusters
* 1 proposed 20 as the best number of clusters

***** Conclusion *****

* According to the majority rule, the best number of clusters is 2
```

Results for 2017:

```
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 1 proposed 3 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 3 proposed 9 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 1 proposed 15 as the best number of clusters
* 1 proposed 18 as the best number of clusters
* 2 proposed 19 as the best number of clusters
* 5 proposed 20 as the best number of clusters

***** Conclusion *****

* According to the majority rule, the best number of clusters is 2
```

It can be seen that indices give 2 as optimal number of clusters for both matrices.

9.4.2 Elbow and silhouette

Direct methods consist in optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named elbow and silhouette methods, respectively.

A detailed trend on the optimal number of clusters can be obtained applying respectively the WSS (Within Sum of Squares) and silhouette indexes in R.

The basic idea behind partitioning methods is to define clusters such that the total intra-cluster variation (or total within-cluster sum of square, WSS) is minimized. The total WSS measures the compactness of the clustering, which should be as small as possible.

The Elbow method looks at the total WSS as a function of the number of clusters: one should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

Average silhouette method computes the average silhouette of observations: it measures how similar a sample is to the others belonging to the same cluster, compared to how similar it is to the others in different clusters, estimating the distance between clusters for different values of k.

The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values [40].

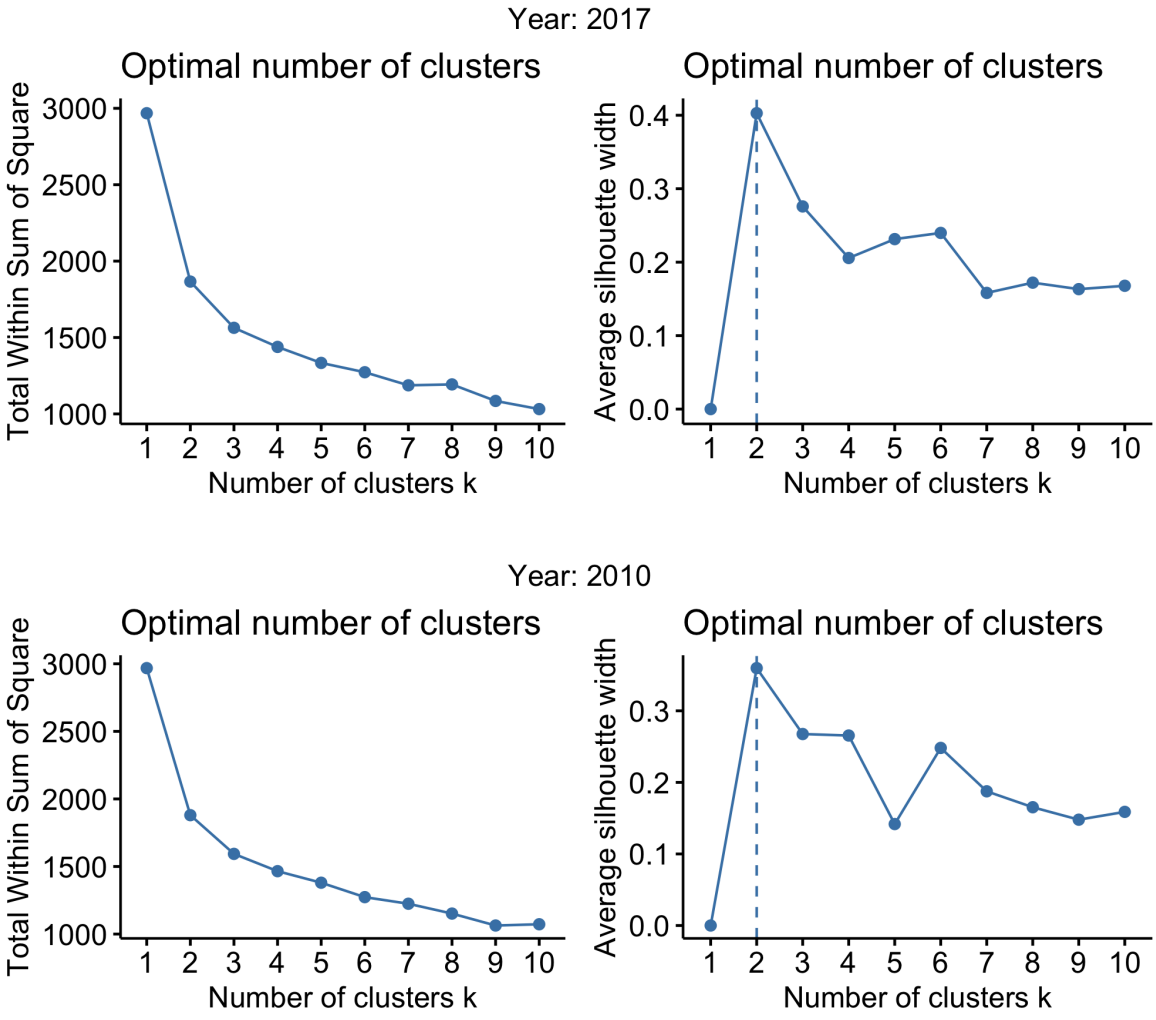

Figure 9.1: WSS and silhouette indexes for k-means

It can be seen in the figure that both indexes show 2 as the best number, coherently with `NbClust`, yet 5 or 6 clusters give an acceptable score as well.

Having 2 clusters increases the risk of classification according to **irrelevant features**, such as large and small prescribers, giving no other useful information.

Therefore, the algorithm is subject of an **additional run** with an arbitrary amount of 4 clusters, observing how many individuals compose each cluster and eventually adjusting the value of k.

9.5 Results

9.5.1 2 clusters

Year 2010

#	Augmentin	Normix	Ciproxin	Levox.	Mon.	Vel.	Roc.	Zitr.	Size
1	64,44	52,95	36,96	37,39	32,23	34,45	25,88	25,15	232
2	224,59	131,29	86,29	80,10	80,25	99,74	62,08	54,15	140

Table 9.2: k-means with 2 clusters, 2010

Year 2017

#	Augmentin	Normix	Ciproxin	Levox.	Mon.	Vel.	Roc.	Zitr.	Size
1	95,69	70,46	49,27	42,37	37,51	21,74	36,17	32,14	242
2	316,50	185,61	104,90	86,72	89,07	42,25	77,11	79,42	130

Table 9.3: k-means with 2 clusters, 2017

Considerations

The two clusters have a consistent different between the **means** of antibiotic prescriptions' amounts, especially Augmentin and Velamox, which are the most popular.

This leads to considering two clusters as *heavy* prescribers and *light* ones, as expected.

Comparison between total prescriptions of members of each cluster:

Year	Cluster	Mean	SD
2010	1	9 043,41	5 120,81
2010	2	18 228,35	5 314,28
2017	1	10 126,52	5 392,88
2017	2	19 755,14	5 360,95

Table 9.4: Comparison of total prescriptions within clusters

Standard deviation of total amount of prescription is overall the same, which implies clustering has consistent and related results. Mean between first and second cluster tends to increase during the years, yet is still close.

This confirms the hypothesis of grouping according to the number of prescriptions.

The amount of general practitioners in the clusters is almost the same from 2010 to 2017, with 78 doctors swapping clusters in total:

- 44 switched from large prescribers to small ones;
- 34 switched from small prescribers to large ones.

The difference of 10 switching doctors is the same as the difference between cluster 1 in 2010 and cluster 1 in 2017.

General practitioners overall remained mostly stable with their amount of prescriptions, yet some switched habits while the mean of total prescriptions increased. This most likely implies that doctors classified as large prescribers still consistently increased their numbers, so that the mean grew by roughly 1 000 in 7 years, although some doctors started to prescribe less.

Clusters for each year are shown performing a PCA among the considered features (number of prescriptions for each antibiotic) with the `fviz_cluster` function from package `factoextra`.

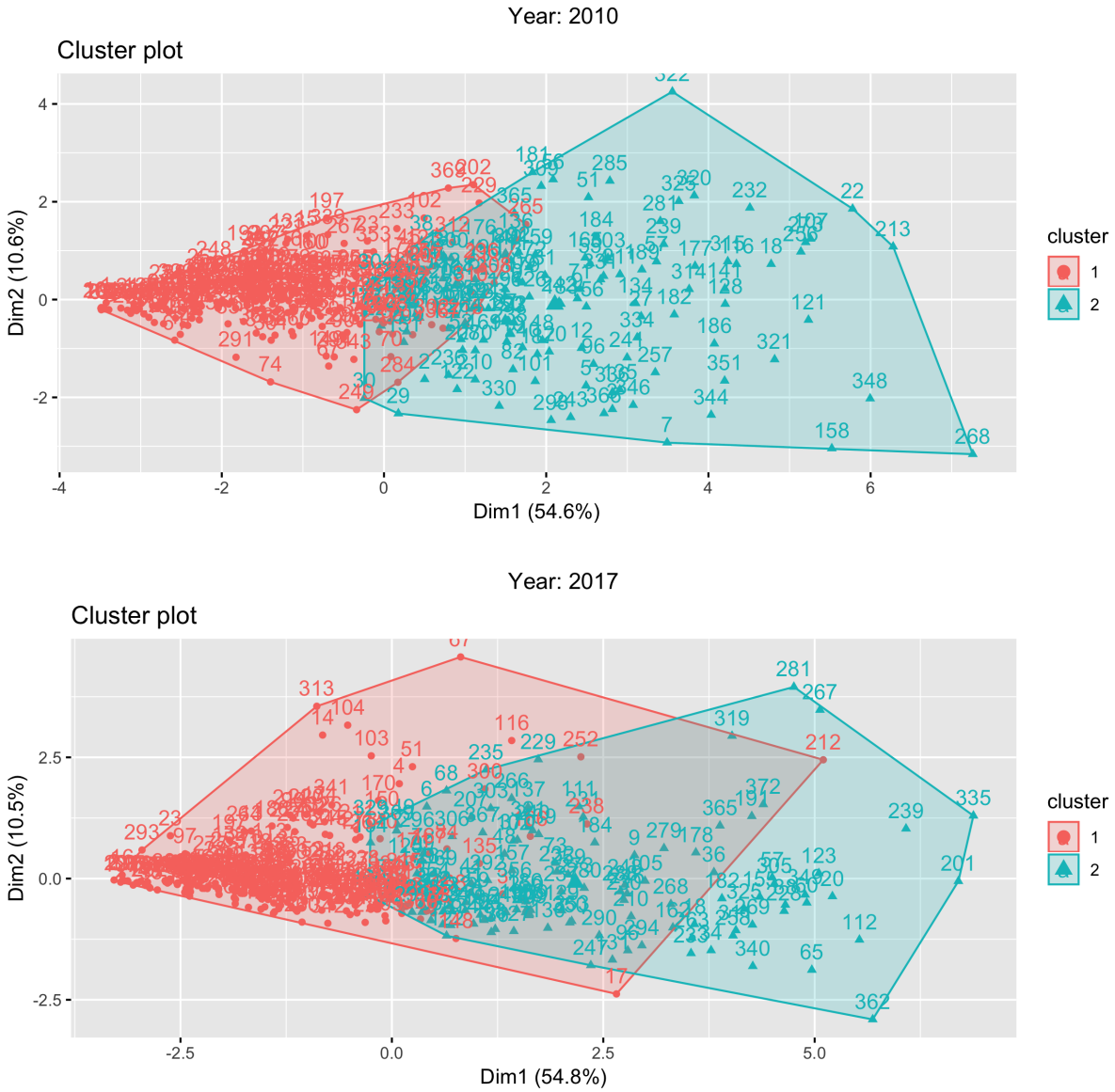

Figure 9.2: PCA with 2 clusters

Each number represents the row corresponding to a general practitioner, while the axes are principal components. Clusters tend to overlap because of the dimensionality reduction.

9.5.2 4 clusters

Year 2010

#	Augmentin	Normix	Ciproxin	Levox.	Mon.	Vel.	Roc.	Zitr.	Size
1	291,35	187,93	107,84	94,42	102,02	159,28	70,68	65,37	45
2	86,85	120,96	68,14	70,03	65,63	76,75	48,13	47,63	61
3	195,29	85,38	69,41	61,93	61,60	57,77	50,10	43,20	103
4	48,27	37,02	27,59	30,61	23,15	25,51	20,97	19,14	163

Table 9.5: k-means with 4 clusters, 2010

Year 2017

#	Augmentin	Normix	Ciproxin	Levox.	Mon.	Vel.	Roc.	Zitr.	Size
1	397,54	236,54	124,08	101,05	107,93	55,01	87,51	112,71	59
2	32,57	165,30	89,50	90,31	76,88	47,65	72,69	69,80	26
3	228,18	119,93	78,85	63,74	63,15	28,99	59,01	46,01	117
4	78,24	52,33	39,33	33,88	28,84	16,92	28,36	25,03	170

Table 9.6: k-means with 4 clusters, 2017

Considerations

According to the cluster means for each antibiotic, general practitioners within clusters can be classified as:

1. Large prescribers;
2. Strong preference for Normix;
3. Strong preference for Augmentin;
4. Small prescribers.

It can be seen that Augmentin means for cluster 3 is higher than the Normix one for cluster 2: Augmentin is most likely the over-prescribed drug.

Similarities in clusters composition:

- 28 doctors being large prescribers in 2010 stayed constant in 2017;
- 12 doctors being Normix prescribers in 2010 stayed constant in 2017;
- 55 doctors being Augmentin prescribers in 2010 stayed constant in 2017;
- 122 doctors being small prescribers in 2010 stayed constant in 2017.

Similarities show 20-40 doctors for each cluster switching prescription habits, with a total of 155 elements belonging to a different groups. Considering 372 original individuals, this consists in 41,6%.

Cluster size varies, with cluster 2 being subject of a consistent decrease in the number of elements from 2010 to 2017, while all the others increase size. Normix prescribers, then, switched to another category.

Comparison between total prescriptions of members of each cluster:

Year	Cluster	Mean	SD
2010	1	22 244,29	4 540,88
2010	2	15 587,75	4 328,30
2010	3	14 910,14	4 574,42
2010	4	7 131,60	4 326,86
2017	1	23 154,69	4 523,31
2017	2	17 481,92	5 370,08
2017	3	15 042,37	4 699,17
2017	4	8 459,83	4 601,97

Table 9.7: Comparison of total prescriptions within clusters

Standard deviation tends to stay constant, and mean for large and small prescribers is widely different, therefore the predictions seem to be correct. Mean between Normix and Augmentin prescribers is similar, meaning that both general practitioners having this behaviour do not tend to overprescribe other medicines.

To have a detailed view of changes between Normix and Augmentin preferences, those groups are compared to each other and the major prescribers cluster:

- 26 Normix prescribers in 2010 switched to Augmentin in 2017;
- 4 Augmentin prescribers in 2010 switched to Normix in 2017;
- 5 Normix prescribers in 2010 switched to large prescribers in 2017;
- 22 Augmentin prescribers in 2010 switched to large prescribers in 2017.

Normix prescribers progressively switched to large prescribers (cluster 1) or having Augmentin preference (cluster 3), and all Augmentin averages increased: in particular in cluster 1, those have an additional 100 prescriptions in 2017.

It is safe to assume that although small prescribers represent a considerable percentage of the total, Augmentin prescriptions are being pushed upwards, most likely because of antibiotic resistance and general practitioners' tendency to overprescribe.

Performing a run of the algorithm only using Augmentin and Normix values shows that, despite 2010 being the same, Normix prescriptions in 2017 aren't relevant enough to be the main characteristic of a cluster.

Normix has therefore lost popularity within years: this is confirmed by antibiotics' trends graphs, showing a constant value for Normix prescriptions in the last years while Augmentin has a steady increase.

All the other antibiotics have constant trends as well, aside from the general raise from 2012 to 2015, yet values are consistently smaller, hence why the algorithm does not assess particular importance to them.

Clusters for each year are again shown performing a PCA with `fviz_cluster`.

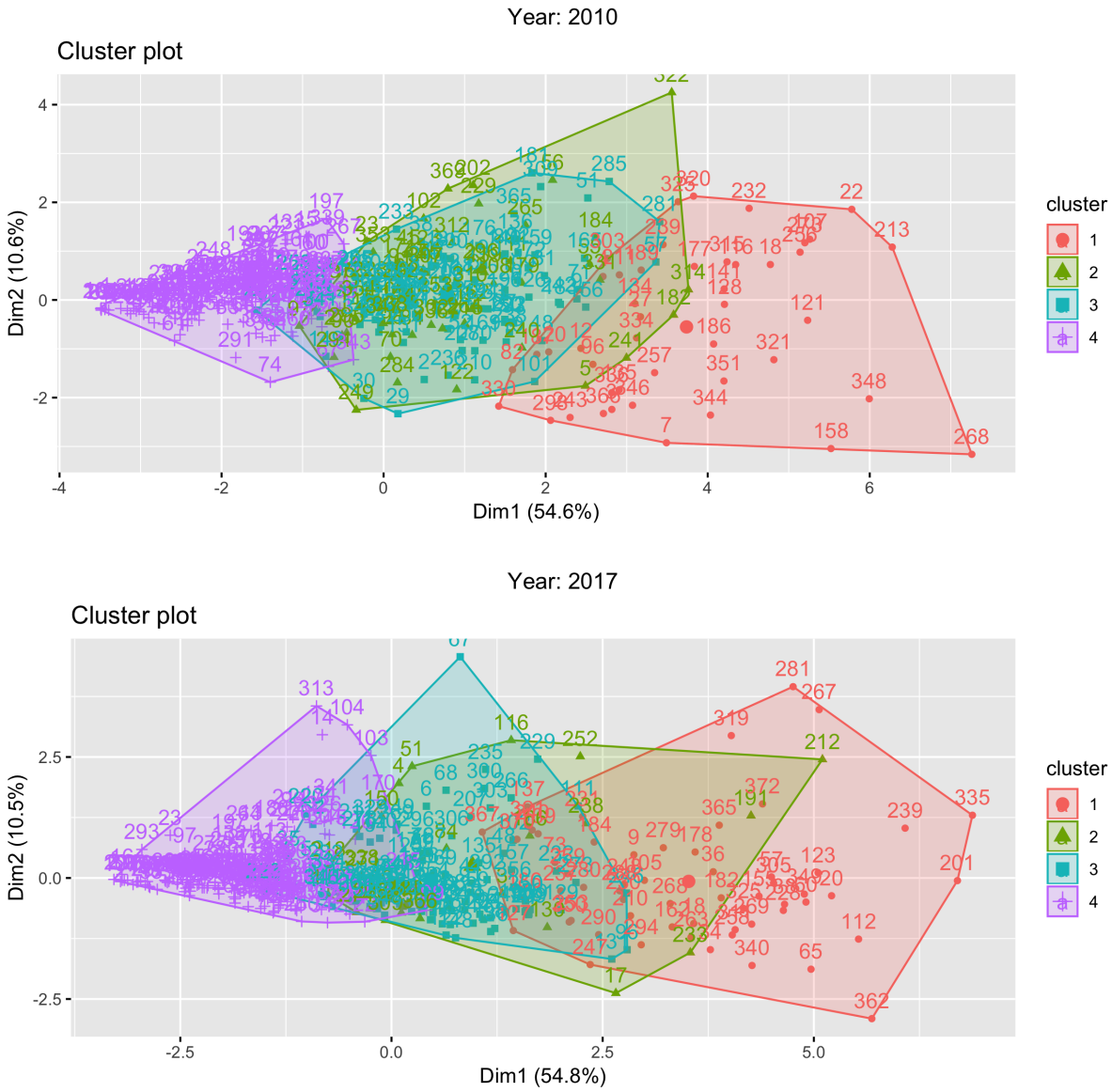

Figure 9.3: PCA with 4 clusters

Chapter 10

Graph analysis

This chapter performs a final analysis on antibiotic patterns through a graph database, applying graph algorithms to identify similarities between prescriptions and popularity among prescriptions. Results are then compared with k-means clustering.

10.1 Graph databases

Graph databases are data management systems allowing persistent representation of entity and relationship in a graph structure, implementing the *Property Graph Model* efficiently down to the storage level.

A **graph** $G =< E, V >$ is an abstract data type showing connections (edges E) between pairs of vertices (V). Nodes identify entities and their properties, while relationships are joining attributes between tables with eventual additional characteristics.

Unlike other databases, *relationships take first priority.* A graph database is purpose-built to handle highly connected data, providing great performance, flexibility and frictionless development.

Queries allow to match pattern of nodes and relationships in a graph, providing ACID transaction compliance without specifying details on how to implement operations. Graph-crossing and related algorithms are highly efficient.

10.1.1 An overview on Neo4j

Neo4j is an online graph database management system with Create, Read, Update and Delete (CRUD) operations working on a graph data model. The data model for a graph database is also significantly simpler and more expressive than those of relational or other NoSQL databases.

In Neo4j, everything is stored in the form of an **edge**, **node**, or **attribute**. Each node and edge can have any number of attributes, and both nodes and edges can be labelled. Labels can be used to narrow searches, improving speed.

Queries are written using **Cypher**, a declarative graph query language that allows for expressive and efficient querying and updating of the graph.

Cypher is inspired by a number of different approaches and builds on established practices for expressive querying, with SQL-inspired keywords and high-level semantics [23].

10.2 Prescription coupling

An excessive usage of antibiotics causes death of microorganisms in the human body which provide to maintaining immune cells and killing certain oral infections [39].

To equilibrate the intestinal flora, lactic ferments are often taken together with antibiotics, so that new "good" bacteria can restore the probiotic action.

If this hypothesis is correct, the dataset will show antibiotic prescriptions **paired with other drugs**, on the same date — or it will highlight potential linkings between infections and other pathologies receiving a specific prescription.

10.3 Goals

Goals of analytics through graphs is completion of antibiotic patterns changes and patient journey, providing a different point of view on those two important aspects altogether.

This part of the research aims to focus on:

- Co-prescriptions, understanding whether specified couples of drugs are often prescribed together;
- Clustering in communities, to identify similar kinds of doctors according to their prescription history;
- Centrality measures of nodes, to highlight particularly important entities in the graph.

10.4 Practical approach

10.4.1 Relational database structure

The available data comprehends patients, general practitioners, and their prescriptions in the time span from 2000 to 2018, located in Campania. Summarising the amount of records for each entity:

- 888 219 patients;
- 2 486 doctors;
- 118 716 403 prescriptions;
- 33 523 drugs.

Due to the amount and veracity of data, identifying a subset of records is useful to have detailed and targeted results, removing dispersive information and leaving a restricted pool of prescriptions, setting acceptability conditions.

Since analytics are aimed to identify **antibiotic prescription patterns**, similarly to past approaches, a new dataset has been extracted, imposing the following constraints:

1. AIC corresponding to an antibiotic;
2. Prescription date between 2008-01-01 and 2017-12-31;
3. Active general practitioners;
4. Patients with usable information about gender, date of birth and location.

This leads to obtaining a new model, composed by:

- 670 634 patients;
- 1 377 doctors;
- 8 386 057 prescriptions;
- 2 802 antibiotics.

To allow analytics on patient journey and co-prescriptions, it is necessary to access all the prescriptions assigned to all patients belonging in the subset. A major extraction is performed from the main table, comprehending:

1. Identifier of patients who received at least one antibiotic prescription **on the same date** as a generic prescription;
2. Prescription date between 2008-01-01 and 2017-12-31.

This reduces the number of other prescriptions, adding drugs not belonging to the antibiotic class. Duplicates, mistakes and empty fields are removed.

Information loss is displayed in the figure below:

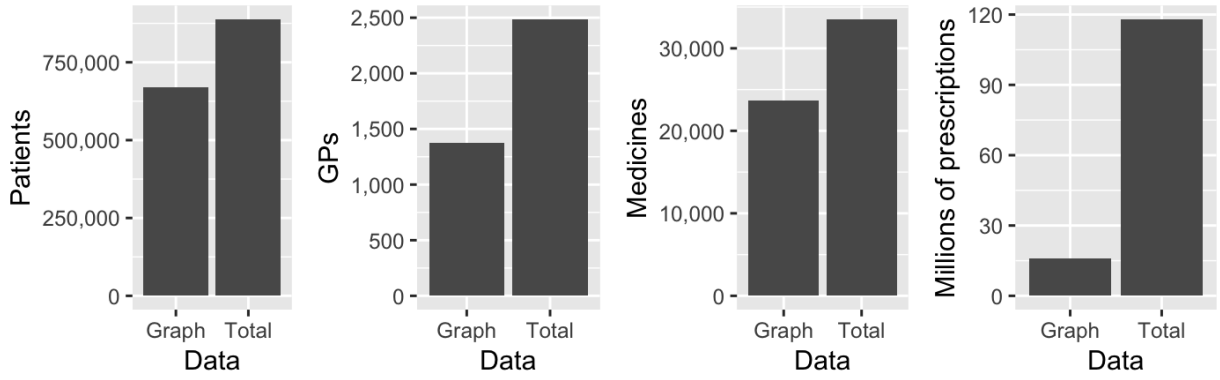

Figure 10.1: Information loss barplot

Values of prescriptions and medicines are aggregated in the plot, including both antibiotics and other drugs. The amount of removed prescriptions is most likely caused by the strict constraint of couples having the same date.

The resulting structure is an unweighted directed graph, with a data composition of:

- 670 634 patients;
- 1 377 doctors;
- 8 328 272 prescriptions of antibiotics;
- 2 465 antibiotics;
- 7 587 009 other prescriptions;
- 21 248 other medicines.

10.4.2 Migration of the database and graph modelling

The database has to be structured following the SQL to Cypher practices and guidelines, assigning nodes and relationships in an appropriate way considering the existing dataset and the related goals. The entity-relationship model translates with the following nodes and attributes:

- Patient;
  - ID, birthdate, gender;
- Doctor;
  - ID;
- Antibiotic;
  - AIC code, ATC code;
- Medicine (anything not Antibiotic);
  - AIC code, ATC code;
- Prescription;
  - patient, doctor, date, drug;
- OtherPrescription (not Antibiotic Prescription);
  - patient, doctor, date, drug.

All nodes are imported, and main indexes are created for optimisation of queries speed. Relationships are then created according to IDs and AIC codes:

- Prescription − TO → Patient;
- Prescription − FROM → Doctor;
- Prescription − OF → Antibiotic;
- OtherPrescription − TO → Patient;
- OtherPrescription − FROM → Doctor;
- OtherPrescription − OF → Medicine.

10.5 Visualisation and analytics

10.5.1 Sample graph

A sample graph is obtained using the `apoc` functions.

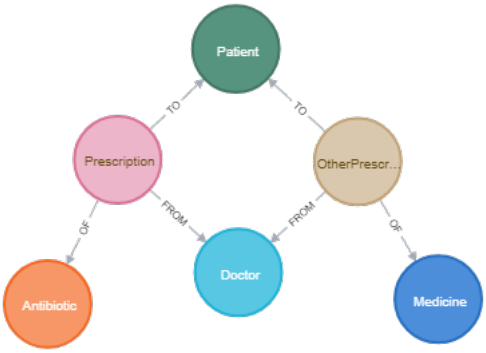

Figure 10.2: Sample graph

10.5.2 Examples

An example of graph subset can be obtained extracting one of the individuals with the most antibiotic prescriptions (the 10th in descending order), a male patient born in 1943, and his associated drugs and doctors. The prescriptions history in 10 years is displayed below.

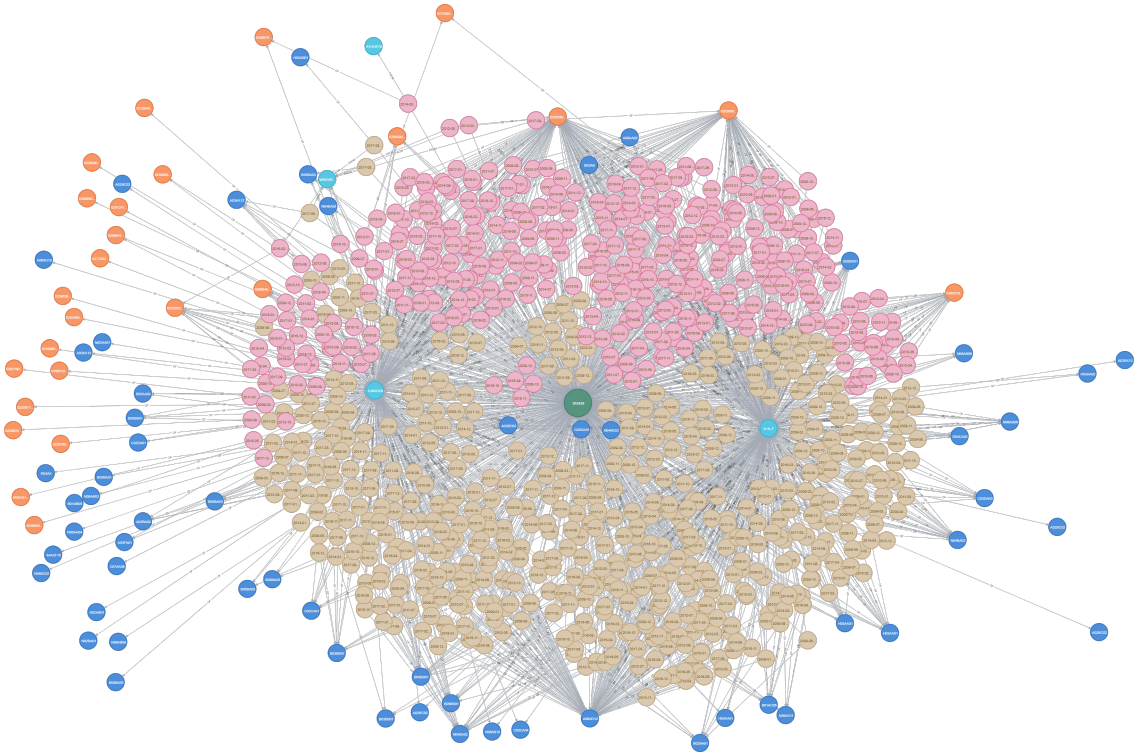

Figure 10.3: View focussed on patient

The graph displays:

1. 1 patient, the green node in the centre;
2. 360 antibiotic prescriptions, the pink nodes;
3. 524 other prescriptions, the brown nodes;
4. 4 doctors, the light blue nodes;
5. 25 antibiotics, the orange nodes;
6. 59 other medicines, the blue nodes.

From this first *patient-centred* visualisation of the graph is possible to identify different behaviours of general practitioners: each one of them is linked to specific antibiotics and medicines.

To have a view focussed on *prescriptions*, the same procedure is applied extracting the 500th antibiotic in descending order according to number of prescriptions, corresponding to Locabiotal spray bottle 15 ml (50 mg / 5 ml).

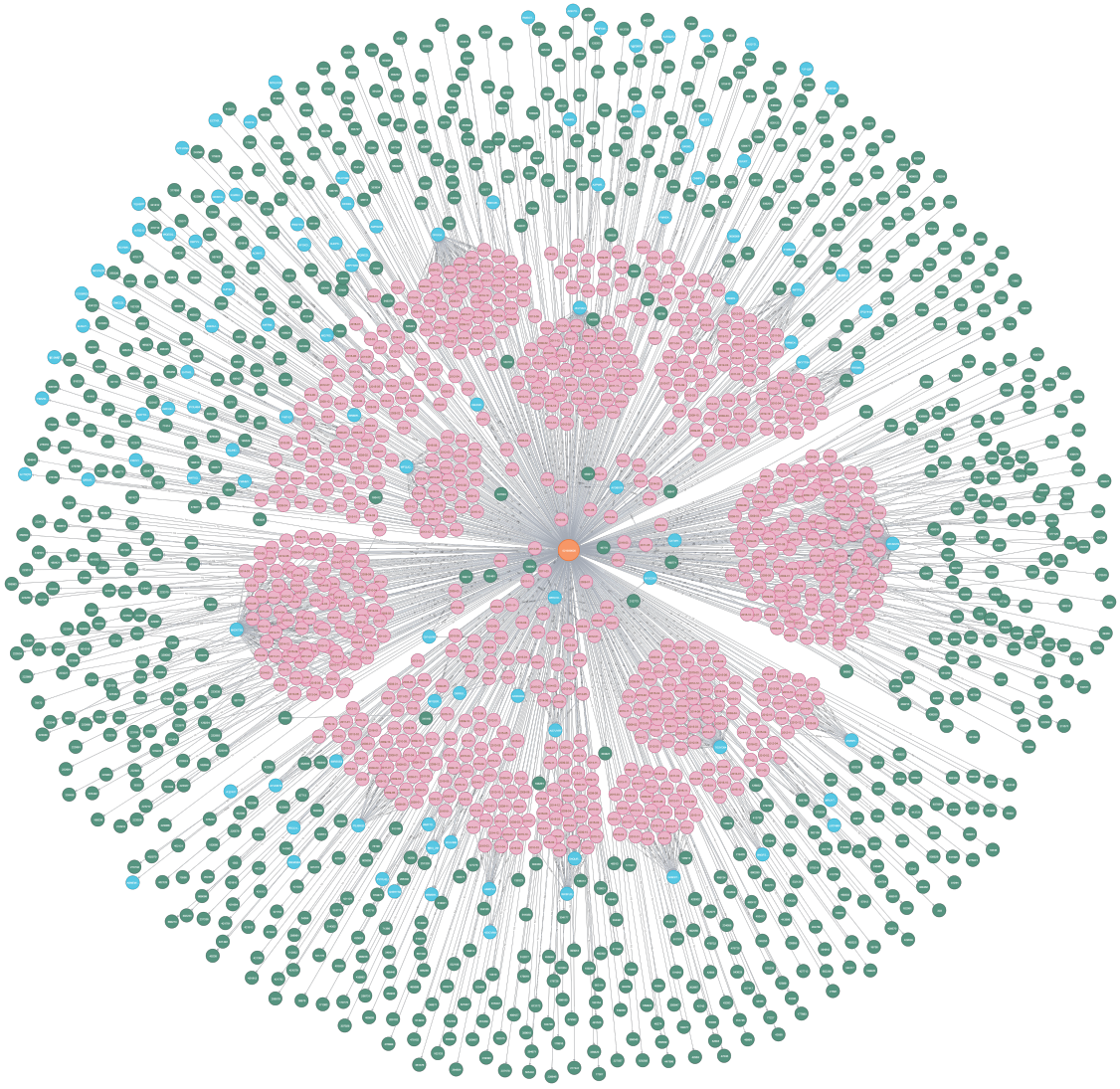

Figure 10.4: View focussed on antibiotic

The previous graph (10.4) displays:

1. 1 antibiotic, the orange node in the centre;
2. 896 antibiotic prescriptions, the pink nodes;
3. 107 doctors, the light blue nodes;
4. 817 patients, the green nodes.

The same antibiotic is prescribed by several doctors, although only 7% of total. There are nodes single-handedly taking a relevant slice of prescriptions (left center and right center), yet most of them is not a habitual prescriber.

Those two examples allow to have a general idea of how nodes interact with each other, grouping in clusters.

10.5.3 Graph statistics

A first set of global and local statistics are used to get the first insight on the graph and its components, using 10 years of data.

Number of nodes	16 611 005
Number of relationships	47 745 841
Average prescriptions per doctor	11 557,93
Standard deviation	15 457,6
Maximum per doctor	96 841
Minimum per doctor	1
Average prescriptions per patient	23,73
Standard deviation	40,46
Maximum per patient	1 567
Minimum per patient	1

Table 10.1: Global statistics

10.5.4 Projecting a co-prescription graph

Since the restriction on generic prescriptions involves having the same date and patient of another antibiotic prescription, couples are analysed adding a relationship between Antibiotic and Medicines in the main graph.

After counting the number of repetitions for each couple Antibiotic-Medicine, the first 100 most popular ones are used to couple nodes, with the amount as property of the relationship PRESCRIBED_WITH.

Visualising the newly created relationships, two connected components are highlighted in figure 7.5 (orange antibiotics, blue other medicines).

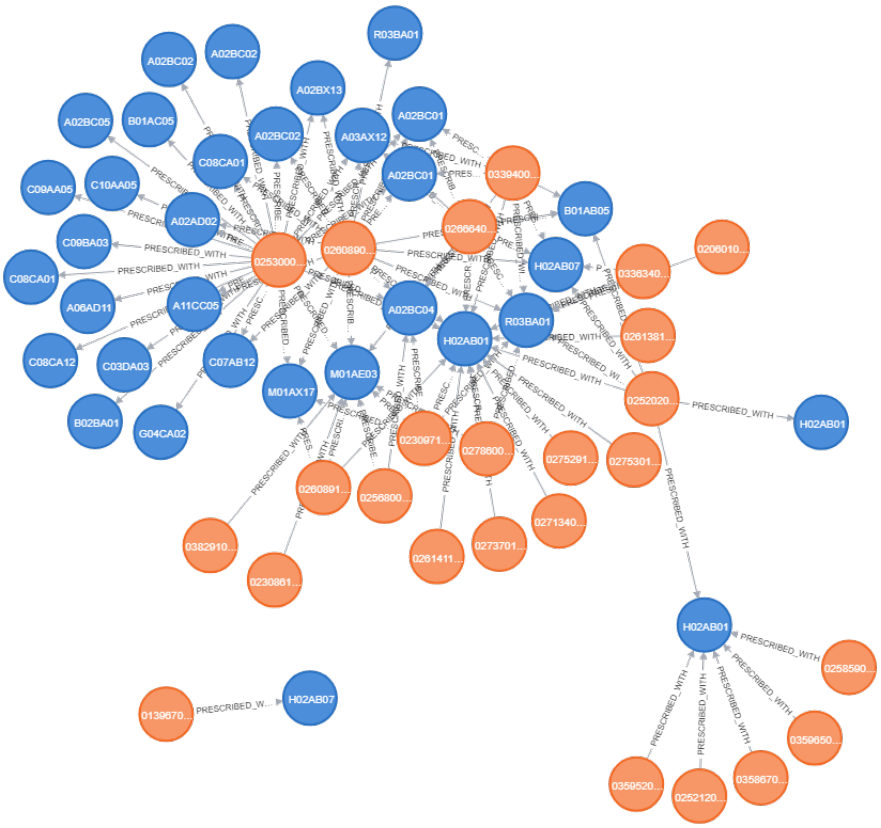

Figure 10.5: Co-prescriptions graph

The component with only one couple corresponds to Plaquenil - Deltacortene, prescribed together approximately 5 000 times. Plaquenil can be used as antibiotic (antimalarial), but it is mostly given to treat arthritis, while Deltacortene is a corticosteroid for rheumatisms.

The range of values varies from 4 145 to 38 153 in the timespan of 10 years. The 5 most popular co-prescriptions are:

1. Augmentin - Oki;
2. Rocefin - Bentelan;
3. Augmentin - Bentelan;
4. Normix - Cardioaspirin;
5. Augmentin - Aulin.

Bentelan is a corticosteroid which cannot be used without an antibiotic in presence of systemic (concerning the whole organism) infections, since it is an immunosuppressive drug, and this would explain the frequent co-prescriptions.

All the **antibiotics** are among the most prescribed ones, which justifies their presence in the co-prescriptions as well.

10.5.5 Projecting prescriptive habits

Having a detailed view of doctors' most common prescriptions gives another insight on how antibiotics are related.

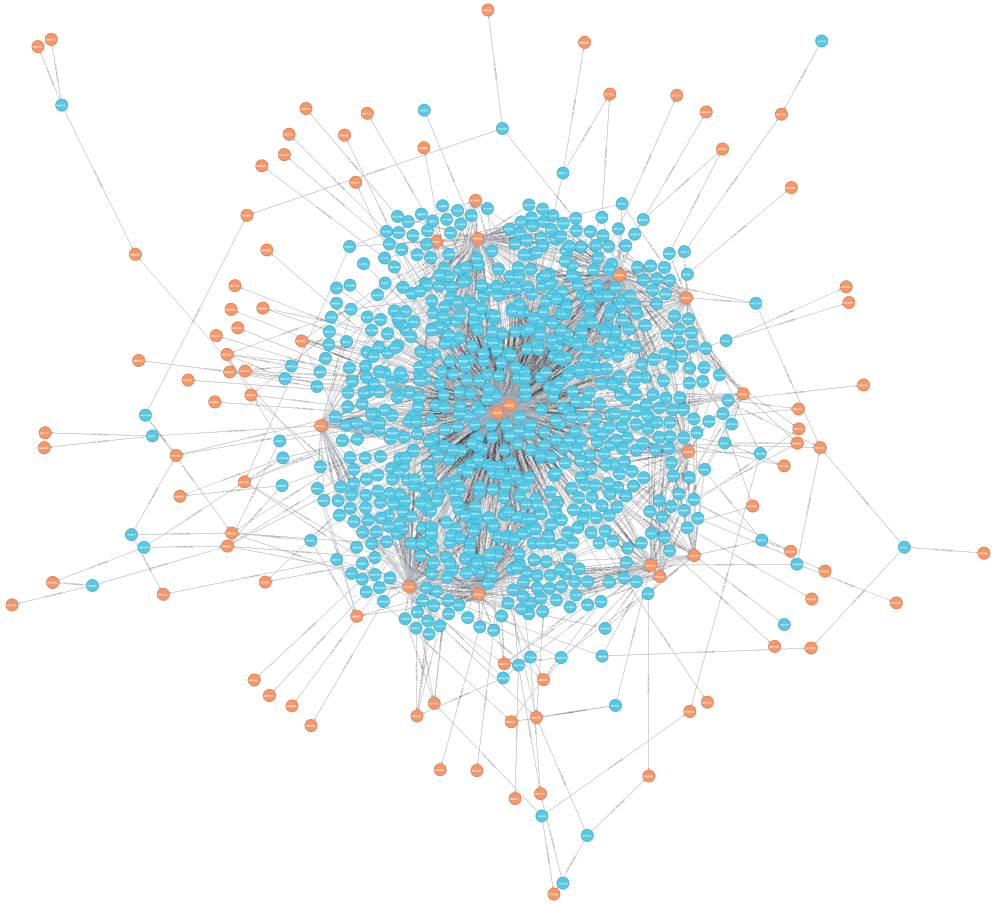

Figure 10.6: Most prescribed antibiotics graph

The three most prescribed antibiotics for each doctor are taken, along with their count, and a new relationship OFTEN_PRESCRIBED between Doctor and Antibiotic is created. Only amounts of prescriptions greater or equal to 50 are taken into account, for consistency of results, resulting in 2 353 links.

The obtained graph displays:

1. 809 doctors, the light blue nodes;
2. 95 antibiotics, the orange nodes.

There are a few antibiotics in the centre (Augmentin and Normix), linked to a large amount of doctors, and external antibiotic nodes having rare prevalence.

Comparing the quantity of antibiotics and doctors and their relationships, it can be seen that most doctors prescribe a very restricted set of antibiotics.

To have a better understanding of the popularity of specific drugs, an auxiliary relationship PRESCRIBED_BY_SAME_DOCTOR is created between Antibiotics.

Antibiotics are interlinked with new relationships:

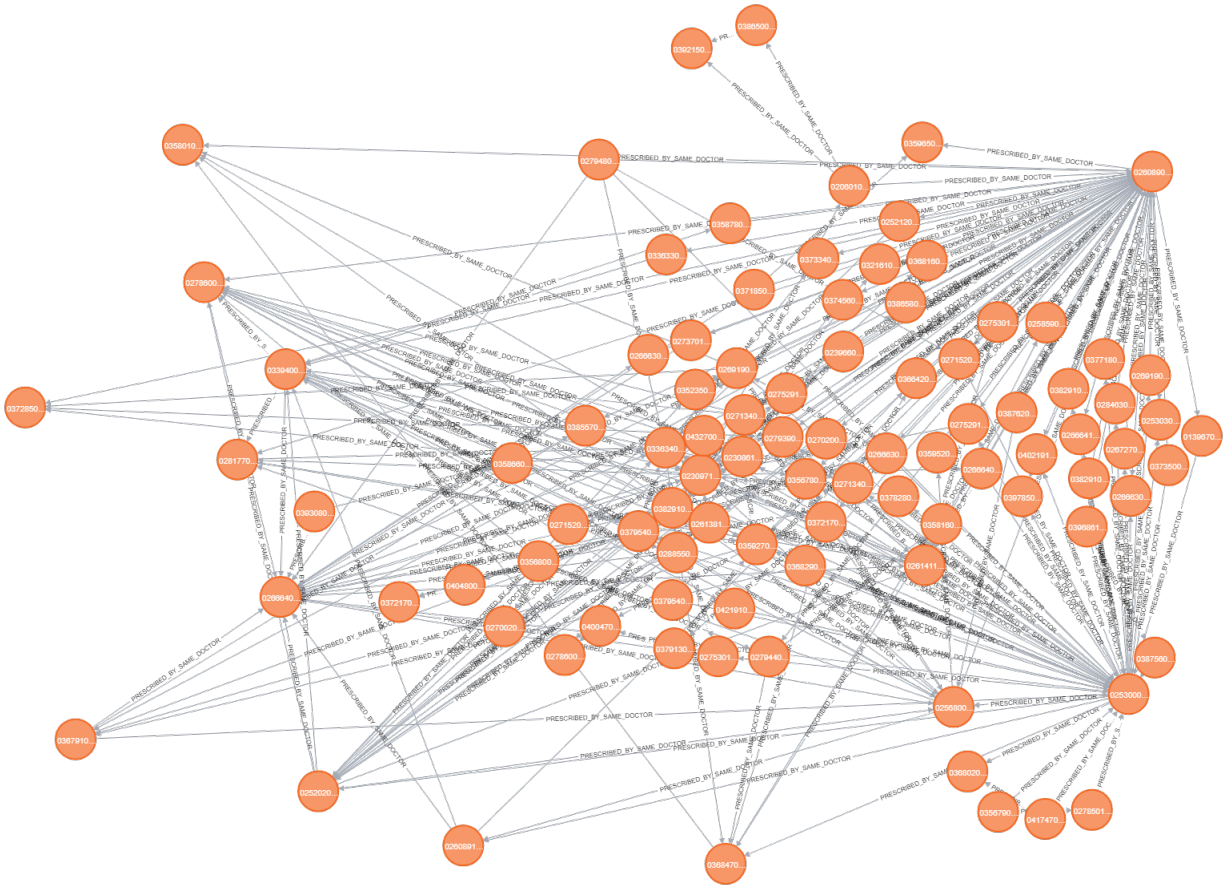

Figure 10.7: Antibiotics prescribed by same doctors

95 nodes are connected by 294 relationships.

The previous graph confirms the discrepancy between popularity of antibiotics. Nodes on the top left have a higher number of connections, while ones on the bottom right tend to have fewer prescriptions.

One hypothesis consists in having a subset of common antibiotics shared between most general practitioners, and another one of individual preferences.

10.6 Graph algorithms

The final part of this research consists in wrapping up existing results and compare them to the ones obtained with a different methodology, applying graph algorithms.

Graph algorithms are the powerhouse behind analytics for connected systems. These algorithms use the connections between data to evaluate and infer the organization and dynamics of real-world systems[1].

[1]Neo4j official documentation

10.6.1 Centrality algorithms

Betweenness

The Betweenness Centrality algorithm calculates the shortest (weighted) path between every pair of nodes in a connected graph, using the breadth-first search algorithm.

Each node receives a score, based on the number of these shortest paths that pass through the node. Nodes that most frequently lie on these shortest paths will have a higher betweenness centrality score.

Betweenness among the top 5 **antibiotics**:

Antibiotic	**Betweenness**
Augmentin (tablets)	2 562
Normix	1 852,74
Velamox	683,78
Ciproxin	562,71
Augmentin (bottles)	489,93

Table 10.2: Betweenness statistics for antibiotics

As expected, the *most prescribed* antibiotics have the highest betweenness scores: since they often get prescribed, the number of relationships involving the nodes is high, increasing the probability of a shortest path crossing them.

Degree

Degree Centrality is the simplest of all the centrality algorithms. It measures the number of incoming and outgoing relationships from a node, analysing its influence.

Degree among the top 5 *antibiotics*, according to co-prescriptions:

Antibiotic	**Degree**
Augmentin (tablets)	32
Normix	22
Ciproxin	14
Augmentin (bottles)	12
Velamox	10

Table 10.3: Degree statistics for antibiotics

Results are similar to Betweenness Centrality, as expected: most prescribed products have the largest number of paths.

Another approach of Degree Centrality analysis involves general practitioners, to understand their influence as prescribers.

Calculating degree among **doctors** is useful to determine whether the top antibiotics prescribers are also the top prescribers, using the top 5 doctors.

Doctor	Degree (antibiotics)	Degree (medicine)	Total degree	Patients
1	45 625	51 221	96 846	4 466
2	42 554	41 594	84 148	2 022
3	35 645	22 383	58 028	1 816
4	34 587	40 315	74 902	2 028
5	34 380	28 763	63 143	1 874

Table 10.4: Degree statistics for doctors

Seeing the obtained results, the overprescribing of some general practitioners is clear: most of them makes a number of antibiotic prescriptions equal to others having double the amount of patients.

Doctors prescribing too much and not strictly when needed is one of the main causes of antibiotic resistance in Italy, therefore further analysis is required to assess changes in patients numbers and break down results in shorter time spans.

10.6.2 Graph sampling

Machine learning algorithms need a heavy amount of resources and computational time, therefore running clustering on a graph with nodes in the magnitude order of millions is not the best practice to extract insightful information in limited time.

Imposing small time ranges and only relevant features (lower bound on prescriptions) improves not only machine usage, but also accuracy of data.

A new subset is extracted, selecting all data from the year 2017 of patients having arbitrarily at least **10 antibiotic prescriptions** during that year.

The new database is composed by:

- 8 228 patients;
- 115 443 antibiotic prescriptions;
- 775 doctors;
- 946 antibiotics;
- 181 016 other prescriptions;
- 4 654 other medicines.

Having a time range of one year and only considering patients getting antibiotic prescriptions rather often does not offer additional information on antibiotic resistance, yet it allows to obtain relevant patterns of prescribing habits.

The output of graph processing will be used to apply computationally expensive graph algorithms.

10.6.3 Similarity detection

Jaccard similarity is computed among **doctors**, considering the Cartesian product of nodes (775×775) with a threshold of 0.4 as coefficient. The result consists in about 500 values, which show the presence of clusters. The considered parameter to determine similarity is antibiotic prescribing (relationship between doctors and antibiotics), therefore linked nodes have the same habits.

The algorithm computes for each node, the most similar other node, according to the maximum Jaccard coefficient.

Doctors are then assigned a link, connecting similar pairs. Since a single doctor can be the most similar to several ones, nodes tend to group in clusters: connected components have a high probability of having common prescription patterns.

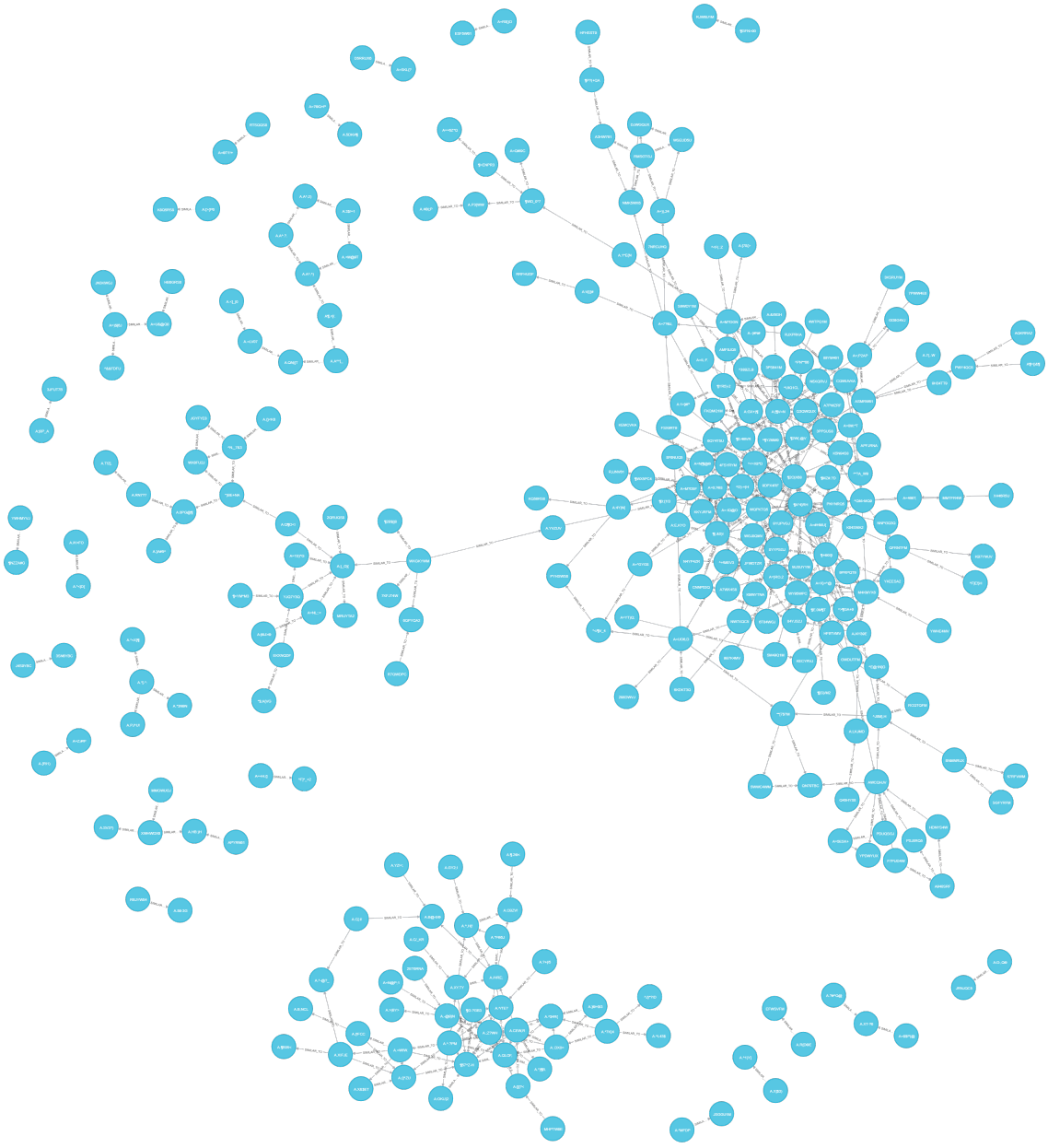

Figure 10.8: Similarity on doctors

A large cluster on the right (figure 7.8) groups most of general practitioners, confirming the hypothesis of popular antibiotics prescribed by the majority of individuals.

Going into a higher level of detail, Augmentin and Normix are the most common prescriptions, yet a single doctor *often leans towards one* of the two instead of giving them both to patients.

There also is a pair of doctors connecting different components of the graph, acting as a "bridge" between prescribers of different antibiotics.

10.6.4 Community detection

The concept of community refers to the structure (topological and relational aspects) of dense sub-components of a graph. Node attributes are considered as well, to obtain additional inferences on their membership. Data is linked with discrete attributes, and applications are based on graphs properties.

Identifying communities allows to understand the global functioning and relationships between individuals with the same features. Different groups interact between one another, and can be related, giving information on the global schema.

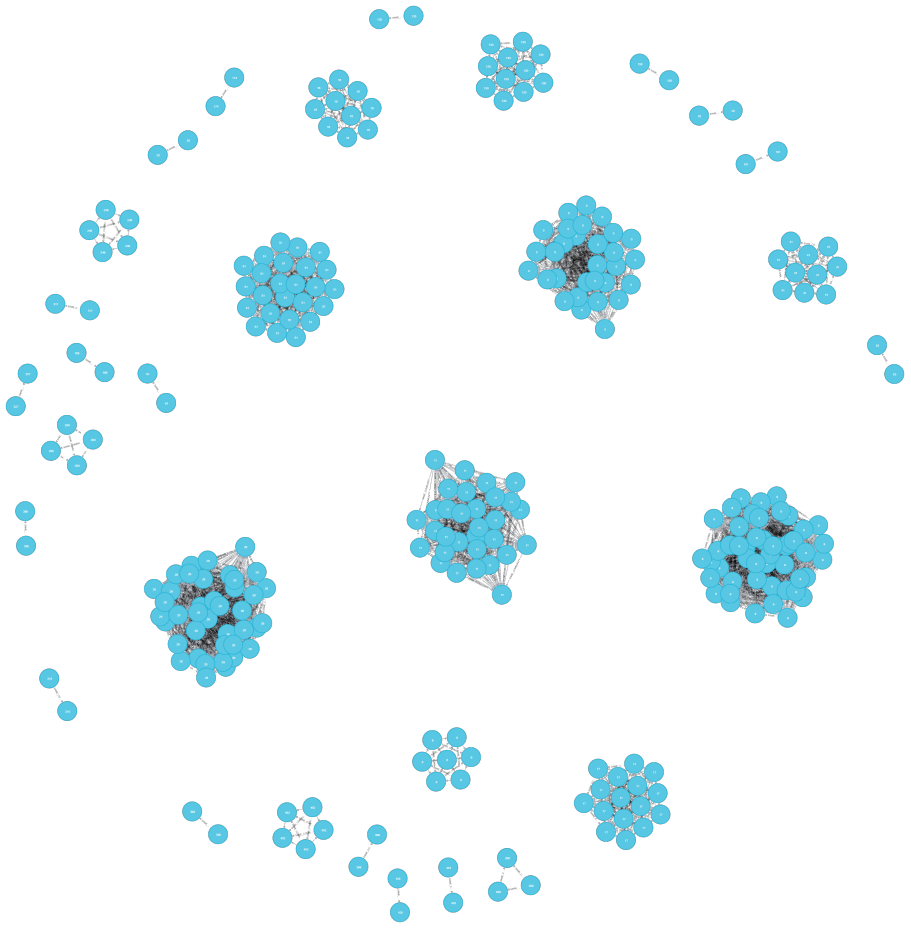

Figure 10.9: Communities of doctors

Communities are identified between **doctors** according to their similarity, based on antibiotic prescriptions and resulting in 526 communities, 495 of which are composed by a

singlet. The presence of clusters implies that groups of doctors have the same prescriptive habits, while others do not fall in any specific category.

Results are similar to the k-means approach, although the considered dataset is significantly smaller: each community is distinguished by relevant characteristics, taking into account additional factors such as the amount of different products.

The first 5 communities in terms of numerosity are:

Community	**Doctors**	**Characteristic**
9	48	Large Normix prescriptions
26	41	Rare prescribers
11	31	All antibiotics prescribers, large amount
$0 \cup 51$	$30 + 25$	All antibiotics prescribers, medium amount
17	16	Large Augmentin prescribers

Table 10.5: Communities statistics

As expected, Normix and Augmentin prescribers are different classes of doctors: since both antibiotics are wide spectrum, the preference is most likely influenced by promotional factors or personal bias.

10.7 Considerations

Graph databases offer a completely different perspective compared to relational ones, allowing to understand behaviour of nodes and linkage between them.

Despite not having a relationship between each pair of node, the whole graph can be efficiently crossed through **paths**, displaying different views focussed on nodes or their type while linking them to the whole structure.

Building targeted datasets and extracting features is immediate, adding relationships with eventual weight and filtering according to properties.

Those datasets or query results are fed to graph algorithms, which automatically stream or add resulting properties to then make further considerations without having to export data in a format readable by programming languages.

The main issue is encountered while *handling large amount of data*: performance with 200k nodes is optimal, yet scaling in the order of millions makes computationally expensive algorithms impossible to run in an acceptable amount of time.

Results are consistent with other machine learning approaches and exploration analytics, providing different methodologies to improve speed and comprehension.

Concrete outcomes after an approach with graphs consist in a further confirmation of the issue of antibiotic resistance, comparing doctors' degree to their number of patients and analyzing their prescriptive habits which, although clear, are concerning.

Co-prescriptions also highlight a major prevalence of a smaller set of products, influenced by the pressure of pharmaceutical companies and majorly prescribed for common diseases.

Chapter 11

Assessments of results and future directions

The research ultimately accomplishes its key objective, confirming AIFA reports on antibiotic resistance in Italy: *the amount of yearly prescriptions grew of almost 50 000 in the last 10 years*, despite the number of patients remaining stable and the lack of general spread of illness.

The same outcome is also shown by the reconstruction of patient journeys, analysing first-time prescriptions and diagnoses among the complete history of patients.

Graph analytics gives an additional insight on co-prescriptions and popularity of antibiotics, identifying ones prescribed together and comparing values to the amount of patients of each doctor.

The most popular antibiotic corresponds to **amoxicillin and beta-lactamase inhibitors**, active principle on which is based **Augmentin** — both trends are main subjects of the significant increase.

Prescriptive patterns among general practitioners allow to distinguish between classes of habits, characterised by preference for Augmentin or Normix, and the amount of antibiotics doctors tend to give.

Additional analysis should be performed on variations through time, understanding the relationship between number of patients and number of prescriptions to check whether general practitioners are effective over-prescribers. It is necessary to consider the composition effect, since values can be affected by patients switching doctors.

Those results were obtained only after an extensive preprocessing of data, optimisation of queries and analysis on progressive information loss: healthcare data is in fact difficult to handle without considering its quantity and variety, both on relational and graph models.

This work is only the beginning of a more detailed research: restricting the domain to antibiotics implies the need of future analytics expanding the scope, for instance to other concerns raised while making global statistics such as the abnormal prescription of digestive trait medicines.

Bibliography

[1] WHO, Antimicrobial resistance fact sheet

[2] CDC, Antibiotic / Antimicrobial Resistance

[3] P. Little, C. Gould, I. Williamson, G. Warner, M. Gantley, A. L. Kinmonth, Reattendance and complications in a randomised trial of prescribing strategies for sore throat: the medicalising effect of prescribing antibiotics, *BMJ*

[4] AIFA, antibiotic usage in Italy report

[5] PEW, Antibiotics Currently in Global Clinical Development

[6] Public Health Agency of Sweden, 2014, Swedish work on containment of antibiotic resistance

[7] Portale della Salute, Esenzioni dal ticket

[8] WONCA, 2011, Italy - La Definizione Europea della Medicina Generale / Medicina di Famiglia

[9] Ministero della Salute, Medici di medicina generale

[10] S. M. Belknap, H. Moore, S. A Lanzotti, P. R. Yarnold, M. Gets et al., 2008, Application of Software Design Principles and Debugging Methods to an Analgesia Prescription Reduces Risk of Severe Injury From Medical Use of Opioids, *ASCPT*

[11] AIFA, Autorizzazione all'immissione al commercio

[12] Legislazione italiana, Provvedimento CUF 30.12.1993

[13] Repubblica, Antibioticoresistenza: in Italia il primato europeo di decessi

[14] Legge 28 dicembre 1995, n. 549, Misure di razionalizzazione della finanza pubblica

[15] AIFA, La resistenza agli antibiotici, emergenza mondiale: il primo rapporto globale del WHO

[16] M. C. Tzourakis, 1996, The Healthcare Industry and Data Quality, *MIT Documents*

[17] H. Oh, C. Rizo, M. Enkin, A. Jadad, 2005, What Is eHealth: A Systematic Review of Published Definitions, *Journal of Medical Internet Research*

[18] M. Lebied, 12 Examples of Big Data Analytics In Healthcare That Can Save People, *Datapine*

[19] I. T. Joliffe, J. Cadima, 2016, Principal component analysis: a review and recent developments, *Philosophical Transactions of the Royal Society*

[20] L. Morrissette, S. Charter, 2013, The k-means clustering technique: General considerations and implementation in Mathematica, *University of Ottawa*

[21] Understanding the Many V's of Healthcare Big Data Analytics, J. Bresnick, *Health IT Analytics*

[22] G. Stockley, Data Classification in Healthcare, *Boldon James*

[23] M. Needham, A. E. Hodler, Graph Algorithms: Practical Examples in Apache Spark and Neo4j, 2019, *O' Reilly*

[24] StatSoft Inc, 2013, Electronic Statistics Textbook, *Tulsa*

[25] Ministero della Salute, Manual ICD-9-CM, Italian version, 2007

[26] Bioportal, ATC ontologies

[27] AIFA, Medicinali equivalenti: qualità, sicurezza ed efficacia

[28] C. H. Yu, 2017, Exploratory Data Analysis, *Oxford Bibliographies*

[29] WHO, ICD classification

[30] Ministero della Salute, Classificazioni ICD

[31] Medical Billing and Coding Online, ICD-9-CM codes database

[32] WHOCC, ATC levels

[33] United Nations, Provisional Guidelines on Standard International Age Classifications

[34] V. Regitz-Zagrosek, 2015, Sex and gender differences in health, *EMBO reports*

[35] ANSA, Malattie cardiovascolari causano il 44% dei decessi, numeri in aumento

[36] C. Nordqvist, What to know about antibiotics, *Medical News Today*

[37] Aboutpharma, Antibiotici, continua il calo della ricerca e sviluppo secondo l'Ocse

[38] ClinCalc, The top 300 of 2019

[39] Infection Control Today, Study Shows Antibiotics Destroy Immune Cells and Worsen Oral Infection

[40] Datanovia, Determining The Optimal Number Of Clusters: 3 Must Know Methods