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

grazie CNR <3

ricordiamoci di mettere laura nei ringraziamenti

Abstract

Research on prescription pattern dynamics aims to highlight the issues concerning the raising 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 possible uses, 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

Chapter 1

Introduction

1.1 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.[26]

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).[27]

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.[26]

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.[29]

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 discoveries of new antimicrobial medicines.[51]

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 without resulting in under-prescribing.[28]

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**.

1.2 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.[30]

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,[31] which 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 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.[22]

According to WONCA (World Organization of Family Doctors), they are responsible of supplying integrated and continuative care. Some fundamental skills and activities[22] to pursue this goal are:

- Communication with patients;
- 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.[35] 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,[36] consisting in written authorization to purchase a specific medicine from a pharmacy.

Drugs are dispensed according to the national guidelines,[33] 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.

1.2.1 Pharmaceutical products and prescriptions

The Italian Pharmacy Agency (AIFA) is the public institution for regulatory activity of drugs 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.[41]

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,[31] 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:[32]

- 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, while 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.[31]

1.2.2 Antibiotic consumption in Italy

Italy is the European nation with the highest antibiotic resistance mortality (10.000 deaths every year), in terms of infection caused by resistant bacteria.[43] 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).[19]

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.[51]

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:[4]

- 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 and 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 20$^{\text{word or phrase}th}$ 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 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.[19]

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.

1.3 Healthcare data

Healthcare data is defined as that 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.[20]

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.[21]

Health information is understood and appraised among Electronic Health Records, prescribing, health knowledge managements and information systems. The main concern is the condifentiality 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.[49]

1.3.1 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, 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.[10] 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, one of the largest vendors of data, collecting electronic health records and prescription data:
  - IQVIA integrates data science with human science, automatizing health processes with information services;
- Complion, a document management and workflow platform for clinical research sites;
- Millenium, offering a suite of proprietary systems of family medicine applications on a national level.

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.

1.4 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.

1.4.1 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,[26] 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.[37]

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.[38] 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.[38]

ICD9-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.[39]

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.

1.4.2 ATC

The Anatomical Therapeutic Chemical Classification System (ATC) is a drug 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:[42]

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.

1.4.3 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.[40] 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.

1.4.4 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 allows healthcare organizations the opportunity to properly focus on their core mission.[11]

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.[12]

1.5 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.[1]

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:[5]

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.

Chapter 2

Goals definition

2.1 Methodologies

This 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,[3] yet it has initially been perceived by general practitioners and pharmacists as a mere instrument to save money at the expense of quality.[18]

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).

The 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.

2.2 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's 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 further 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, so 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.

2.3 Practical methods

Having a better vision of the medical domain and the rising issues in the Italian system, before approaching analytical procedures it is necessary to underline practical objectives and expected outcomes.

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 correct 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 (sex, 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.

2.4 Approaches description

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

2.4.1 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.

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

2.4.2 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.[14]

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

2.4.3 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.[7]

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.

2.4.4 PCA

PCA is mostly used as a tool in exploratory data analysis and 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, translating data into a linear mapping with less dimensions, that successively maximize variance and with features that are uncorrelated with each other.[8]

Reduced variables can be used for an easier interpretation and visualisation, 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.

2.4.5 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.[9]

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.[9]

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.

2.4.6 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 and

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.[13]

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 us understand 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 and out-degree.

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

Community detection

Community formation is common in all types of networks, and identifying them is essential for evaluating group behaviour and emergent phenomena.

Connectedness is a core concept of graph theory that enables a sophisticated network analysis such as finding communities. 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. Evaluating different types of communities within a graph can uncover structures, like hubs and hierarchies, 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 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. This strategy identifies community hierarchies and provides a broad understanding of the overall structure.

Louvain Modularity is used for looking at grouping quality and hierarchies. It finds clusters by moving nodes into higher relationship density groups and aggregating into super-communities.

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.

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

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.

2.5 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**,[23] an open source object-relational database management system known for its robustness and reliability:
  - The development platform to interface with the web server containing the dataset is **PgAdmin 4**;
  - 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**,[24] 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**,[25] 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.

Reports and slide-shares have been created and accessed using the **Google Suite**.

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.

Chapter 3

Data description

3.1 Dataset description

The available dataset is provided by Dedalus, market leader of the clinical software area, supporting doctors and their processes through its society Millennium.

Dedalus offers a wide range of solutions, such as surgical journey management, drug management systems, tracking for healthcare and enterprise resource planning.[48]

Millenium is a specific national infrastructure focussing on primary healthcare, using Dedalus products to develop national and regional level projects.

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 region Campania, focussing on the doctor-patient relationship.

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 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 must 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.

3.2 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, so it's necessary to identify the relationship between fields.

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;
- *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's 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.

3.3 Description of the tables

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

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

3.3.1 *patients*

The table *patients* includes information about patients. To ensure privacy dealing with sensitive data, there are no full names: everything is **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;
- *sex*, birth sex;
- *convention*, type of convention with the Italian insurance system.

3.3.2 *patients_doctors*

The table *patients_doctors* contains information similar to *pazienti*, 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.

3.3.3 *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.

3.3.4 *prescriptions*

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

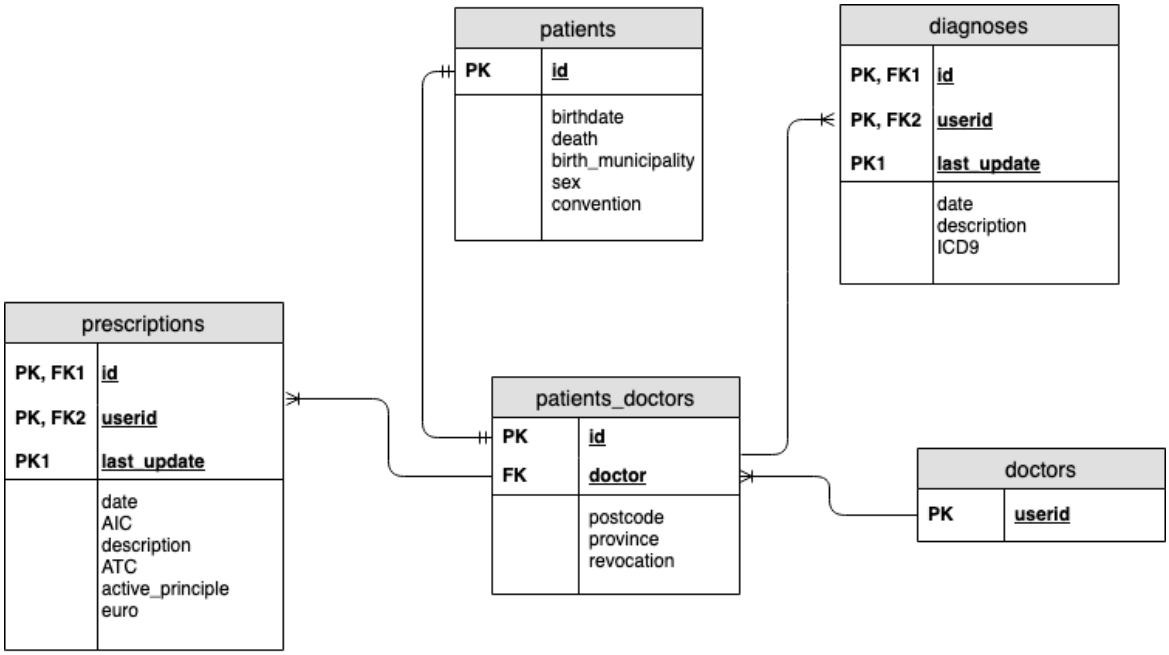
Each prescription is defined by an **ATC code**, from the Anatomical Therapeutical Chemical classification system maintained by the World Health Organization. Furthermore, there are **active principle code** and **authorisation for commerce code**(AIC).

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.

3.4 ER model

After identifying the main fields, it is possible to build and include an **ER model**,[6] to provide immediate comprehension of how entities are related, visualising information to identify keys and unique fields, essential for joining data.



3.5 Variations through time

To give a general idea of variations of patients and magnitude orders, a snapshot of 2010-2017 is used to show patterns and differences, counting the number of patients getting at least a diagnosis for each year:

	2010	2011	2012	2013	2014	2015	2016	2017
Number	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 st. dev.	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

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

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

Chapter 4

Information loss

The first analysis aims to give a **qualitative assessment** of the whole database, giving an overall idea on how impactful is the progressive information loss.

After understanding 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.[16]

4.1 General overview

Before starting running queries, there are some factors to consider and 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 while a general practitioner doesn't prescribe anything but makes other operations (medical certificates, examinations and such);
4. Prescriptions in hospitals 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.

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

4.2 Information loss on records

4.2.1 *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, so joining is an elementary operation and there is no information loss.

- Patients with null or empty sex: $2\ 752 \rightarrow 0,27\%$;
- Patients with sex 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 sex;
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 below, highlighting the total tuples and the impact of every restriction, standardised on a scale from 1 to 100:

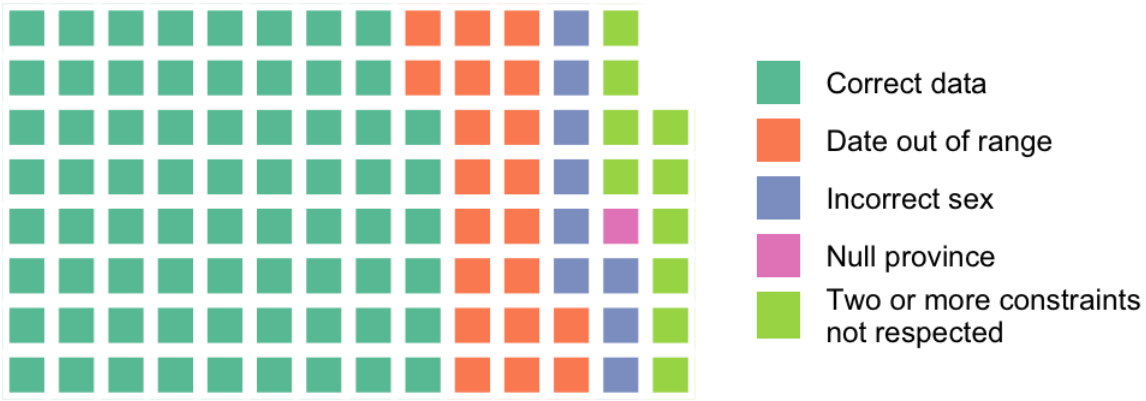


Figure 4.1: Information loss on patients, waffle plot, Campania Millennium database

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.

4.2.2 *diagnoses* (15 millions of tuples)

ICD-9

An ICD-9 code is correct if it is present in the official ICD-9 database.[15] General practitioners may use different formats and notations, so before removing codes each one has been subject to some 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 incorrect ICD-9 codes, a very small amount considering the total records.

Summary

- Incorrect ICD-9 codes: $76 \rightarrow 0,0005\%$;
- Null or empty ICD-9 codes: $2.103.169 \rightarrow 13.02\%$;
- Null or empty descriptions: $656.067 \rightarrow 4,24\%$;

- Dates out of range: $807.985 \rightarrow 5,23\%$.

4.2.3 *prescriptions* (118 millions of tuples)

ATC code

The ATC code, unlike ICD-9, has an 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,[17] 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, aren't really impacting results considering the total amount of 118 millions of records.

Summary

- Incorrect ATC codes: 1 750 292 $\rightarrow 0,1,47\%$;
- Null or empty ATC codes: 1 577 749 $\rightarrow 1,33\%$;
- Null or empty AIC codes: 1 310 719 $\rightarrow 1,1\%$;
- Null or empty descriptions: 101 248 410 $\rightarrow 85,28\%$;
- Dates out of range: 3 472 119 $\rightarrow 2,92\%$.

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's more relevant to focus on prescription patterns analysis.

Chapter 5

Global analytics

Chapter 6

Patient journey

After a preliminary analysis knowing the main focus area, there is enough information to begin reconstructing the **patient journey**, a complete set of data representing a patient history 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.

6.1 Imposed criteria

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

- The prescribing general practitioner mustn't change in the time range;
- The patient mustn't be deceased;
- There must be a sanitary convention;
- The general practitioner must 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, so joining it with other tables is the best method to remove all the rows with an inactive GP. Data is up to date, yet 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 enough prospective.

6.2 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's 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, so 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.

6.2.1 ICD-9

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

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

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.

6.2.2 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.

6.3 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 a cutting the the dataset in terms of information loss. All the previously introduced criteria must be considered as well, so 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 must comprehend 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.

6.3.1 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 causes of this significant loss.

The 144 618 complete tuples are composed by:

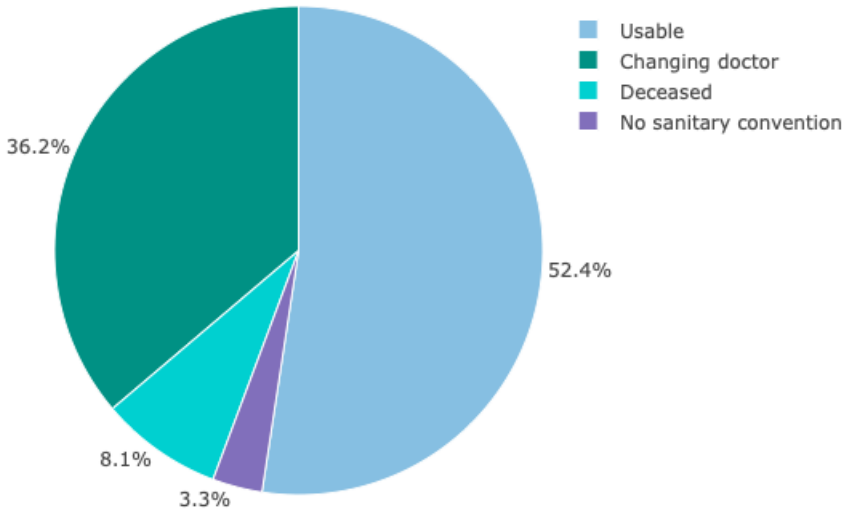
- 27 733 patients;
- 422 general practitioners;
- 1 381 unique diagnoses;
- 904 unique prescriptions.

Further causes 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%.

Patients

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



There is an enormous loss on patients: the possible reason might be that most patients started treatment earlier than 2010, plus diagnosis and relative prescription are in different dates.

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.

6.4 Prescription-based approach

Since extracting a slice of records according to their date span is an abrupt approach, careful parameters 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, so there must be a 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.

More cleaning is required to link diagnoses and prescriptions, since there is not a $1:1$ correspondence: multiple diagnosis and prescriptions may be associated to the same date. This can be done using a lookup table.

6.4.1 Results breakdown

The 1 465 005 complete tuples are composed by:

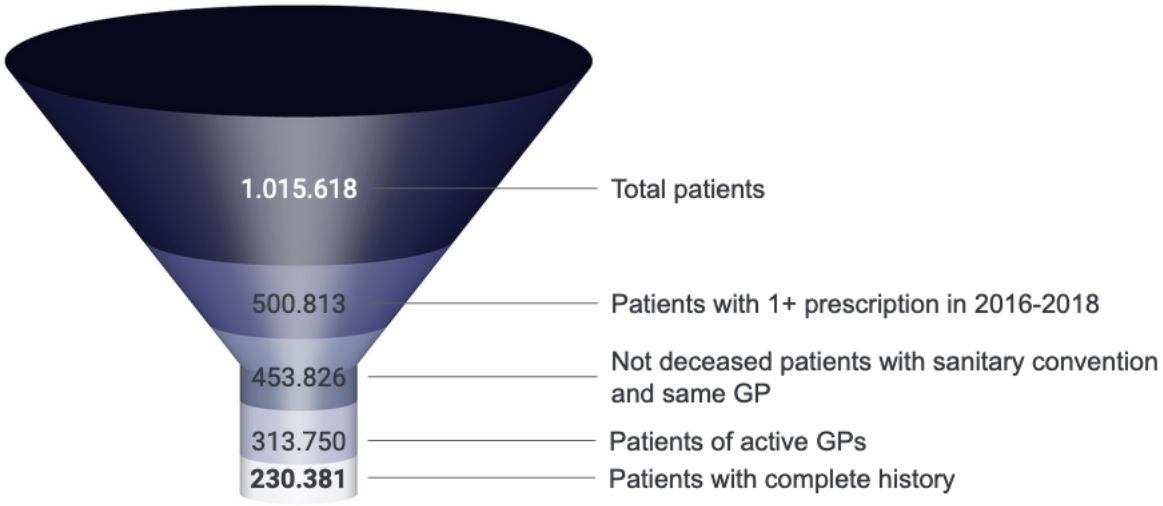
- 230 381 patients;
- 412 general practitioners;
- 4 324 unique diagnoses;
- 1 280 unique prescriptions.

Only 7% of the total prescription has been taken in consideration, yet a million and half is still a satisfying amount.

Funnel graph of patients

The 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).[52]

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



6.4.2 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:

2016	3,26
2017	3,37
2018 (incomplete)	3,06
All years	5,16

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 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

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:

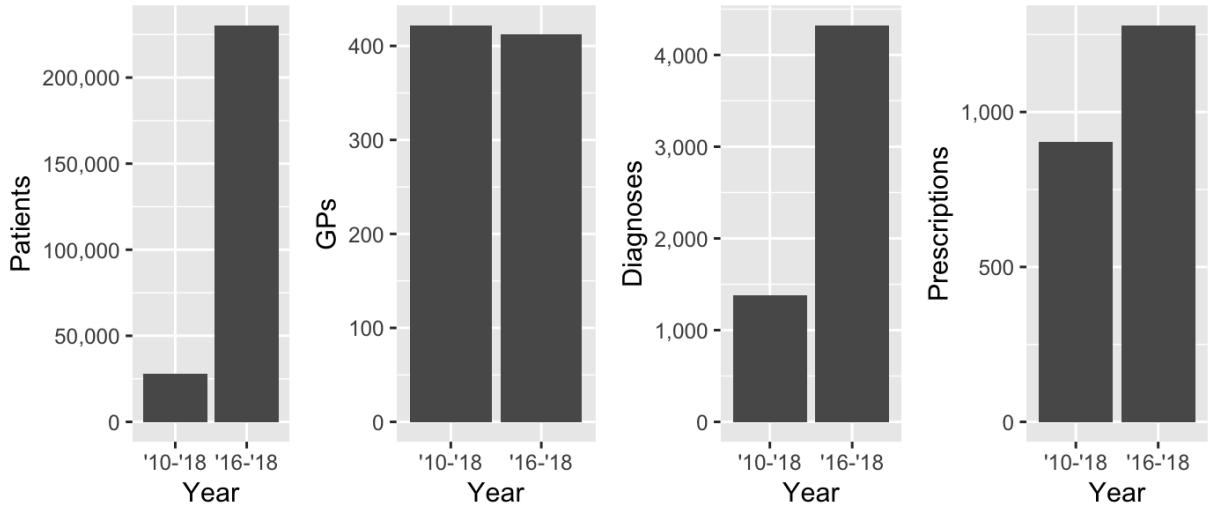
1. Amoxicillin and beta-lactose inhibitors, 62 560 prescriptions;
2. Ciprofloxacin, 23 762 prescriptions;
3. Levofloxacin, 22 595 prescriptions;
4. Rifaximin, 21 680 prescriptions;
5. Ceftriaxone, 19 523 prescriptions.

6.5 Results comparison

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

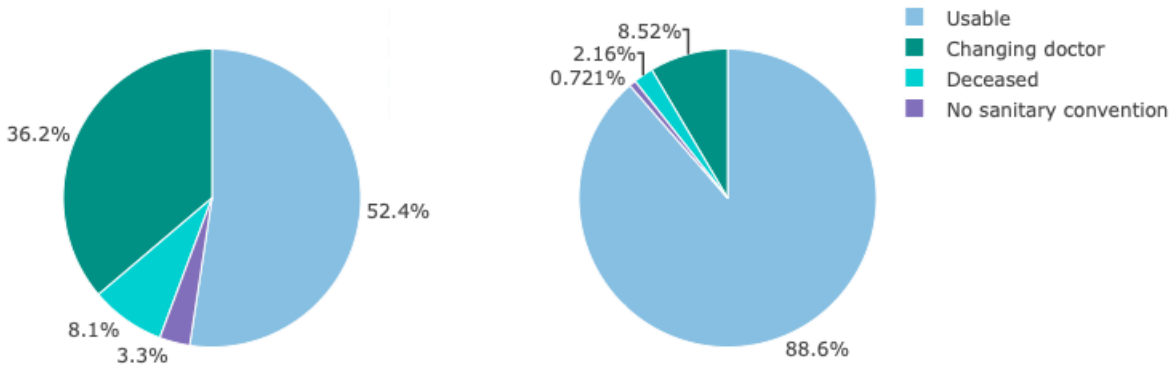
6.5.1 Changes in data composition

Below are shown barplots highlighting the changes between data categories in the two approaches.



6.5.2 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.



It's easy to see that the number of usable patients has noticeably increased: using a smaller time span reduces the chances of death and change of GP.

6.6 Future work

Further analysis is needed to assess an insightful perspective focussed on specific diseases, including prescriptions following the first related diagnoses and removing irrelevant 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 with the issue of antibiotic resistance. More detailed analysis with time series can give a wider perspective.

Chapter 7

k-means approach for cluster analysis

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 according to chosen features and selecting a subset according to consistency and potential given information.

Records will be mapped into an 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, check 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.

7.1 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 photographies 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.

7.2 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 represent a data loss of about 75% in both values.

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 sex and most common age range in $\{1, 2, 3, 4\}$.

Other prescriptions' data has been limited to the total number of prescriptions 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	Augmentin	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 7.1: k-means matrix extract, Campania Millennium database

Input matrix has dimensions of 372×13, and an additional time dimension which distinguishes between 2000 and 2017, which is going to be removed splitting the dataset to run the algorithm separately in the two years.

7.3 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.

7.4 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.[53]

7.4.1 `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
```

7.4.2 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.[53]

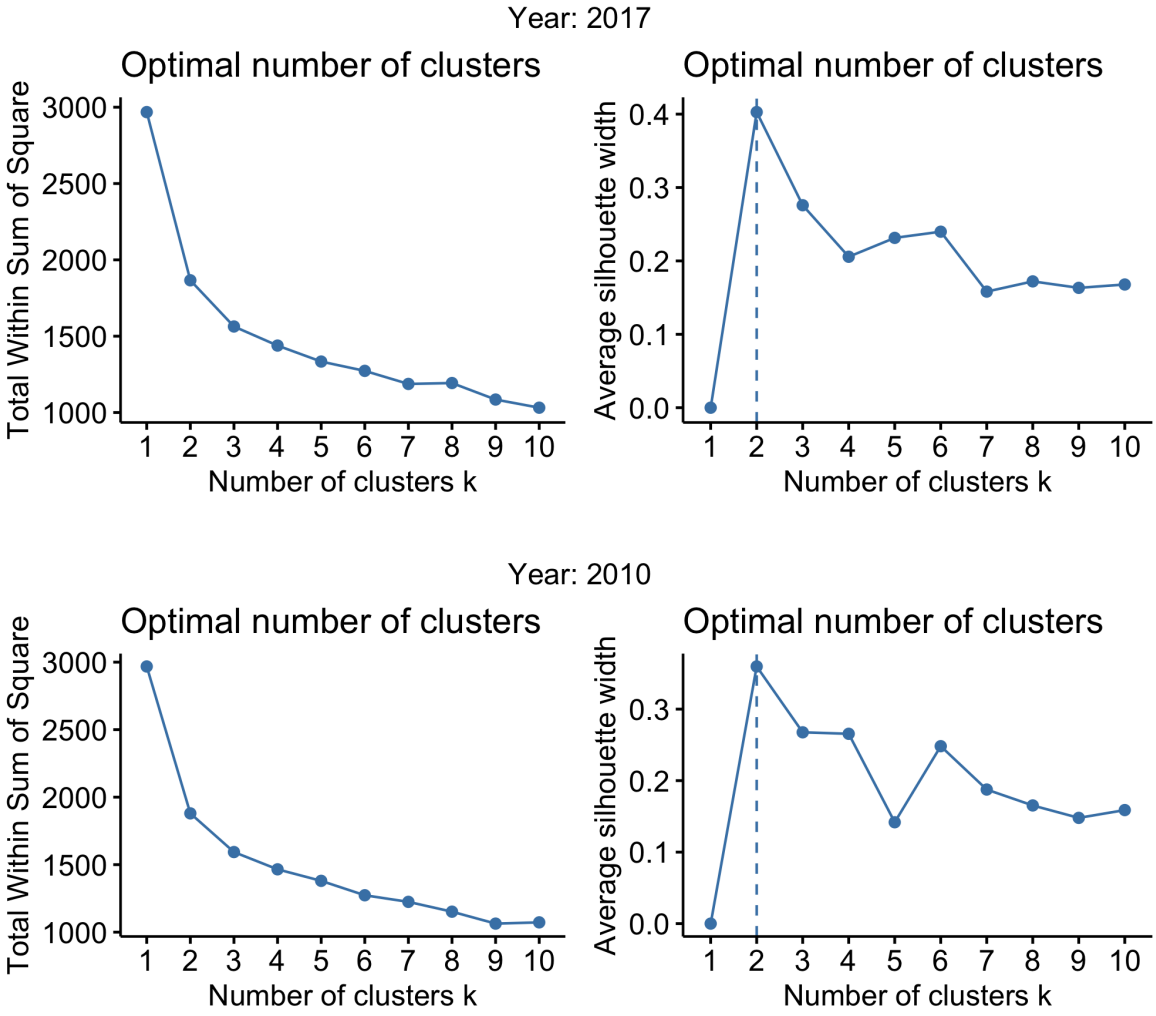


Figure 7.1: WSS and silhouette indexes for k-means, Campania Millennium database

It can be seen in the figure that both indexes show 2 as the best number, 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. Although distinctions are evident, there is no additional useful information obtained after grouping.

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.

7.5 Results

7.5.1 2 clusters

Year 2010

#	Augmentin	Normix	Ciproxin	Levoxacin	Monuril	Velamox	Rocefin	Zitromax	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 7.2: k-means with 2 clusters, 2010, Campania Millennium database

Year 2017

#	Augmentin	Normix	Ciproxin	Levoxacin	Monuril	Velamox	Rocefin	Zitromax	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 7.3: k-means with 2 clusters, 2017, Campania Millennium database

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 7.4: Comparison of total prescriptions within clusters, Campania Millennium database

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 means 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`.

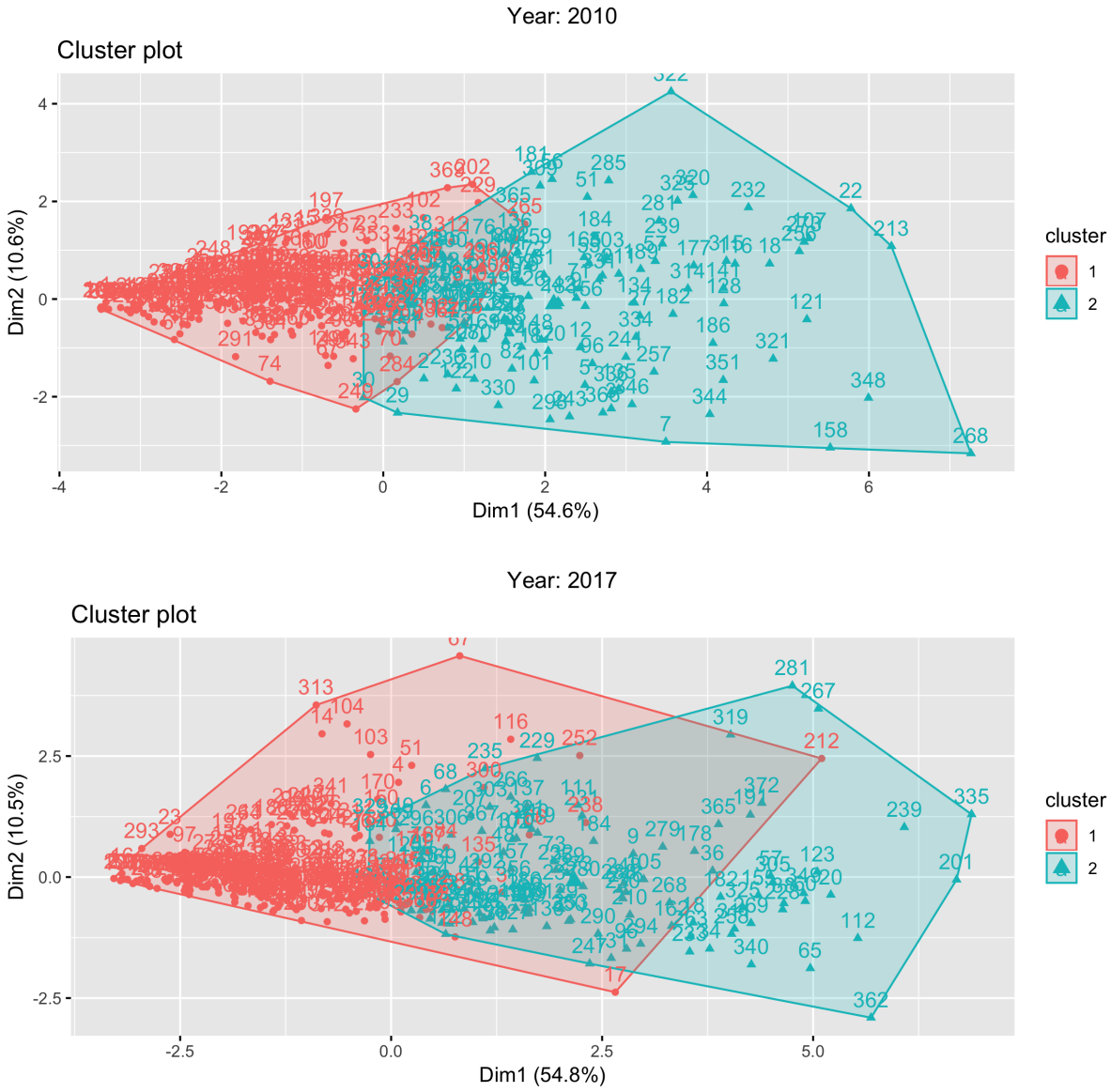


Figure 7.2: PCA with 2 clusters, Campania Millennium database

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.

7.5.2 4 clusters

Year 2010

#	Augmentin	Normix	Ciproxin	Levoxacin	Monuril	Velamox	Rocefin	Zitromax	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 7.5: k-means with 4 clusters, 2010, Campania Millennium database

Year 2017

#	Augmentin	Normix	Ciproxin	Levoxacin	Monuril	Velamox	Rocefin	Zitromax	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 7.6: k-means with 4 clusters, 2017, Campania Millennium database

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 7.7: Comparison of total prescriptions within clusters, Campania Millennium database

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`.

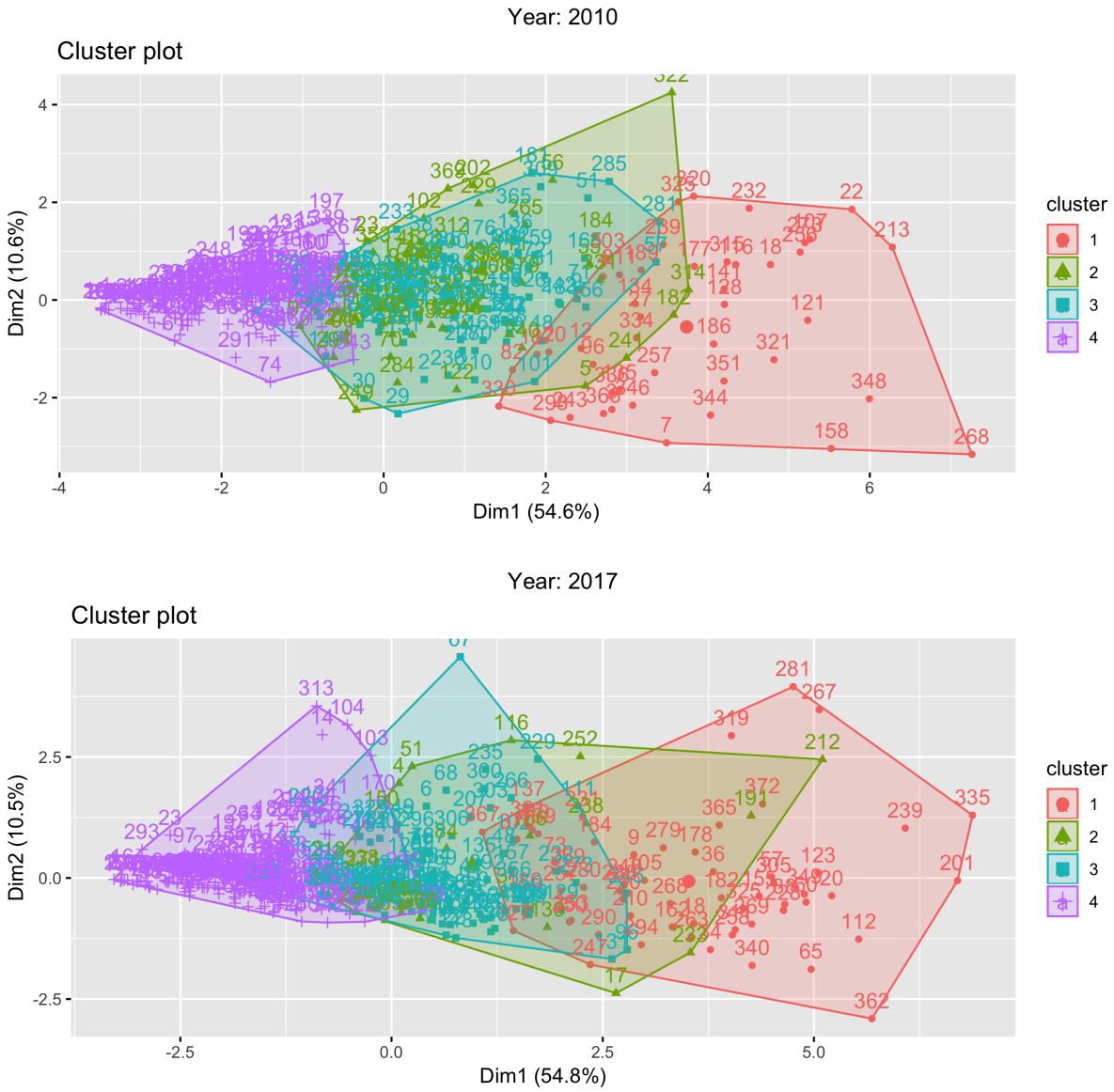


Figure 7.3: PCA with 4 clusters, Campania Millennium database

Chapter 8

Assessments of results and future directions

Bibliography

[1] https://healthitanalytics.com/news/understanding-the-many-vs-of-healthcare-big-data-analytics

[2] https://www.millewin.it/

[3] https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1995-12-29;549 art3!vig=

[4] http://www.agenziafarmaco.gov.it/content/la-resistenza-agli-antibiotici-emergenza-mondiale-il-pri

[5] Davide Castaldi, *Richiesta Dati CNCM per Analisi Appropriatezza Prescrittiva*, Consorzio Milano Ricerche, 2018.

[6] Made with draw.io.

[7] http://www.creative-wisdom.com/teaching/551/Reading_materials/Yu_EDA_Oxford.pdf

[8] https://royalsocietypublishing.org/doi/full/10.1098/rsta.2015.0202

[9] https://www.researchgate.net/publication/308020680_The_k-means_clustering_-technique_General_considerations_and_implementation_in_Mathematica

[10] http://www.promofarma.it/Lista-pagine/Documentazione/Art-50.aspx

[11] https://www.boldonjames.com/blog/data-classification-in-healthcare/

[12] https://edps.europa.eu/data-protection/our-work/subjects/health_en

[13] Graph Algorithms Neo4j

[14] http://www.statsoft.com/Textbook/Time-Series-Analysis

[15] Manuale ICD-9-CM versione italiana 2007. http://www.salute.gov.it/portale/documentazione/p6_2_2_-1.jsp?lingua=italiano&id=2251

[16] Davide Castaldi, *Allegato Tech DB Campania*, Consorzio Milano Ricerche, 2018.

[17] https://bioportal.bioontology.org/ontologies/ATC

[18] http://www.agenziafarmaco.gov.it/sites/default/files/medicinali_-equivalenti-qualita_sicurezza_efficacia.pdf

[19] http://www.aifa.gov.it/sites/default/files/Rapporto-L'uso_degli_antibiotici_in_-Italia_2017_0.pdf

[20] http://mitiq.mit.edu/iciq/Documents/IQ%20Conference%201996/Papers/TheHealthCareIndustryandDataQu

[21] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1550636/

[22] https://www.woncaeurope.org/sites/default/files/documents/Definizione%20WONCA%202011%20ita_-A4.pdf

[23] https://www.postgresql.org/

[24] https://neo4j.com/

[25] https://www.r-project.org/

[26] https://www.who.int/en/news-room/fact-sheets/detail/antimicrobial-resistance

[27] https://www.cdc.gov/drugresistance/about.html

[28] https://www.folkhalsomyndigheten.se/contentassets/dae82c7afd424a57b57ec81818793346/swedish-work-

[29] bmj.com/cgi/pmidlookup?view=long&pmid=9270458

[30] https://en.oxforddictionaries.com/definition/subsidiarity

[31] http://www.salute.gov.it/portale/esenzioni/dettaglioContenutiEsenzioni.jsp?lingua=italiano&id=46

[32] http://www.fcr.re.it/classificazione-dei-farmaci-ai-fini-della-rimborsabilita

[33] https://web.archive.org/web/20111129162006/http://www.farmaciadicello.it/ricetta-01.htm

[34] https://web.archive.org/web/20140611065109/http://www.woncaeurope.org/sites/default/files/docume

[35] http://www.salute.gov.it/portale/temi/p2_6.jsp?lingua=italiano&id=1698&area=tumori&menu=percorso

[36] https://ascpt.onlinelibrary.wiley.com/doi/full/10.1038/clpt.2008.24

[37] https://www.who.int/classifications/icd/en/

[38] http://www.salute.gov.it/portale/temi/p2_6.jsp?lingua=italiano&id=1982&area=statisticheSSN&menu=

[39] https://www.medicalbillingandcodingonline.com/icd-cm-codes/

[40] http://www.agenziafarmaco.gov.it/glossary/term/1432

[41] http://www.agenziafarmaco.gov.it/content/l%E2%80%99autorizzazione-all%E2%80%99immissione-commerc

[42] https://www.whocc.no/filearchive/publications/2019_guidelines_web.pdf

[43] https://www.repubblica.it/salute/medicina-e-ricerca/2019/03/13/news/antibioticoresistenza_-in_italia_il_primato_europeo_di_decessi-221467306/

[44] https://www.medicalnewstoday.com/articles/10278.php

[45] https://www.aboutpharma.com/blog/2019/01/10/antibiotici-continua-il-calo-della-ricerca-e-svilupp

[46] https://clincalc.com/DrugStats/Top300Drugs.aspx

[47] https://www.infectioncontroltoday.com/antibiotics-antimicrobials/study-shows-antibiotics-destroy

[48] https://www.dedalus.eu/

[49] https://www.datapine.com/blog/big-data-examples-in-healthcare/

[50] https://www.my-personaltrainer.it/Foglietti-illustrativi/Bentelan.html

[51] http://www.agenziafarmaco.gov.it/content/la-resistenza-agli-antibiotici-emergenza-mondiale-il-pr

[52] https://www.fusioncharts.com/resources/chart-primers/funnel-chart

[53] https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-meth

[54] https://neo4j.com/docs/pdf/neo4j-graph-algorithms-3.5.pdf