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

Titolo della tesi su più righe

Relatore: Prof. Francesco Archetti

Correlatore: Prof. Antonio Candelieri

Tutor aziendale: Dott. Gaia Arosio

Relazione della prova finale di:
Ilaria Battiston
Matricola 816339

Anno Accademico 2018-2019

Abstract

Descriptive abstract

Contents

Chapter 1

Introduction

Healthcare analytics is a field of growing importance which allows a deep understanding 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 various applications of this field is the change of patterns in the historical data: this research specifically focuses on **prescription pattern changes** on chronic patients, highlighting the development of some common medicines through time.

The considered healthcare data is a dump of a database created and handled by Millennium Srl,[2] a leader company in IT services for medicine. This has been obtained through an extraction procedure on the original DB, assigning unique names to each table and column.

There are five main necessary fields for analysis:[3]

1. **Spatial data**, in different granularity levels;
2. **Personal data** of patients;
3. **Temporal data**, in a range from 2000 to 2018, 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 the 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

Data description

2.1 Overview of the database

The database used for analytics contains data on medical histories of patients between **January 2000** and **October 2018.**

It's important to note that the research work has been done on only a part of the whole database, consisting in **5 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 recognise the relationship between fields.

The 5 tables and their relative sizes are:

- *pazienti*, 1.015.618 tuples;
  Basic information about patients, identified by an encrypted UID;
- *nos_002*, 1.015.618 tuples;
  Extension of *pazienti* with the same key, containing more detailed information and linkage with GPs;
- *cart_pazpbl*, 15.460.199 tuples;
  Information about diagnoses and relative description;
- *cart_terap*, 118.716.403 tuples;
  Information about therapies and prescribed medicines;
- *cart_accert*, 151.478.456 tuples;
  Information about medical checks and examinations.

It's easy to see that the number of rows is varying a lot: there are a lot more prescriptions than diagnosis, and 150 millions of medical examinations.

Each diagnosis, prescription and examination is uniquely distinguished by the triplet patient-GP-date, which maps to *codice–userid–time_last* in the database.

There are many other date fields, but *time_last* (last update of the record) is the only one of type *timestamp*, making each one different from the others (it's unlikely 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 there's no need of unique data: there are lots of different prescription for the same patient made on the same day.

2.2 Description of the tables

Below is reported a brief description of the 5 tables, along with the main fields used for analytics and their meaning.

2.2.1 *pazienti*

The table *pazienti* includes information about patients. To maintain privacy, there are no real names: everything is **encrypted** as a 22-character string containing letters, numbers and special symbols.

Other relevant fields are:

- *data_ open*, date of beginning of the doctor-patient relationship;
- *nascita*, date of birth;
- *decesso*, eventual date of death;
- *comune_ di_ nascita*, name (and code) of the birth municipality;
- *sesso*, birth sex;
- *pa_ convenzione*, type of convention with the Italian insurance system.

2.2.2 *nos_ 002*

The table *nos_ 002* contains the same information as *pazienti*, with additional fields which are essential to analyse data:

- *pa_ medi*, **encrypted** UID of the general practitioner of the patient;
- *pa_ cap*, zip-code of the patient (for geographical analysis);
- *pa_ pro*, province of the patient;
- *pa_ drevoca*, 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 will be used to identify active doctors, but doesn't contain any other useful information.

2.2.3 *cart_pazpbl*

The table *cart_pazpbl* 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.

Fields summary:

- *codice*, patient ID;
- *userid*, corresponding to *pa_medi* in *nos_002* (general practitioner ID);
- *data_open*, date of insertion of the diagnosis in the database;
- *time_last*, timestamp of last edit of the tuple;
- *nome_pbl*, a textual description of the diagnosis;
- *cp_code*, code of the diagnosis according to the ICD-9 standards.

2.2.4 *cart_terap*

The table *cart_terap* contains the prescribed medicines for each patient. There is no linkage between diagnosis 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**.

Fields summary:

- *codice*, patient ID;
- *userid*, corresponding to *pa_medi* in *nos_002* (general practitioner ID);
- *data_open*, date of insertion of the prescription in the database;
- *time_last*, timestamp of last edit of the tuple;
- *co_codifa*, AIC code (authorisation for commerce);
- *co_des*, textual description of the medicine;
- *te_npezzi*, number of boxes;
- *te_attivo*, active principle code;
- *co_atc*, ATC code;
- *euro*, price.

2.2.5 *cart_accert*

The table *cart_accert* includes all the medical checks and examination for each patient. The first goals of the research only included diagnoses and prescriptions, hence why the

comprehension of this table isn't as deep as the others.

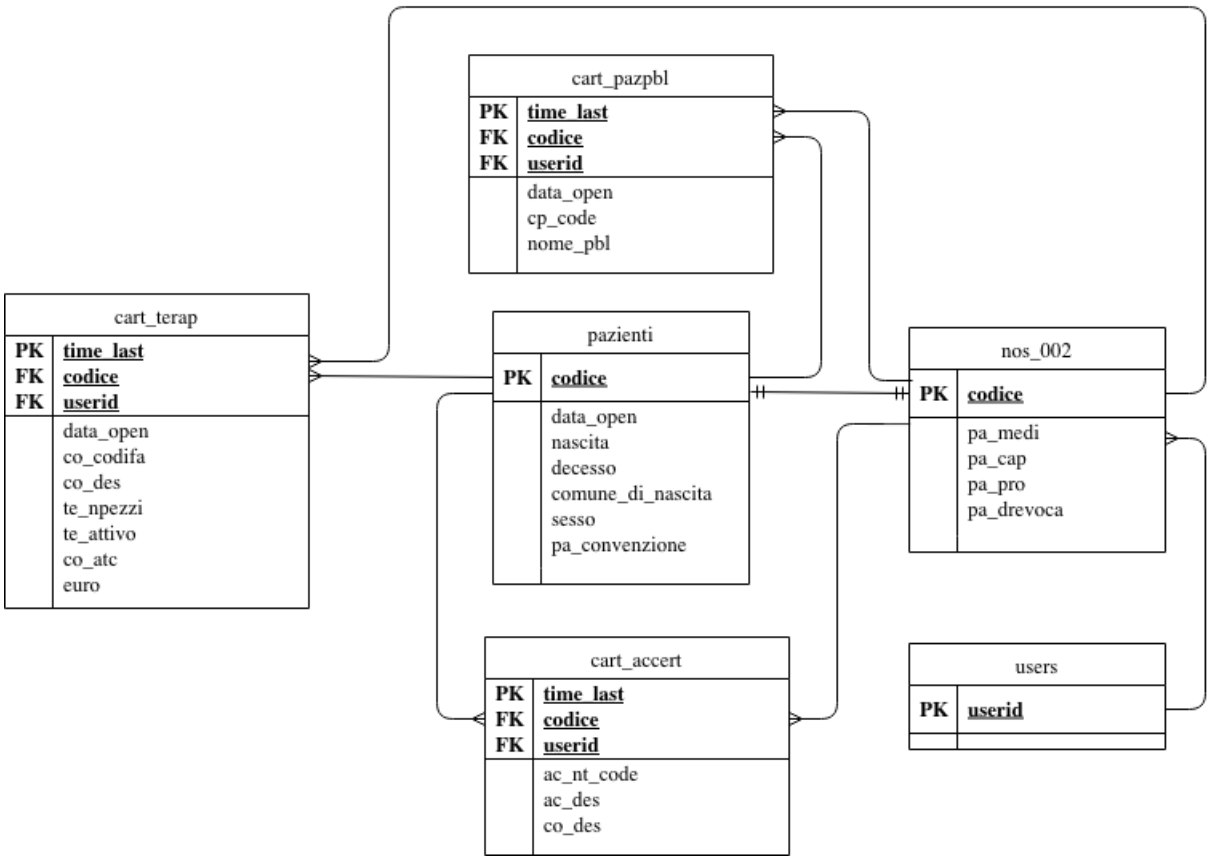
Each examination is defined by an **ICD-9-CM** code, an extension of the ICD-9 database with standard procedures.

Fields summary:

- *codice*, patient ID;
- *userid*, corresponding to *pa_ medi* in *nos_ 002* (general practitioner ID);
- *data_ open*, date of insertion of the examination in the database;
- *time_ last*, timestamp of last edit of the tuple;
- *ac_ nt_ code*, ICD-9-CM code;
- *ac_ des*, textual description;
- *cod_ ese*, exemption code for the examination.

2.3 ER model

After identifying the main fields, it's useful to build and include an **ER model**[4] to have a better understanding of the tables and their relationship.



Chapter 3

Goals definition

Since health big data can be so vary and informative, it's important 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 on chronic patients**: 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 diseases, and define how a patient can be considered chronic.

To obtain constraints the more objective as possible, some further analysis can be useful. 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 best 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 consistent, it's good 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 with record linkage;
2. Elaboration of the statistics and time series clustering.

Another relevant instance for analysis is the **subset** of chronic 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 diagnoses based on the patient gender and age.

An obstacle to perceiving the meaning of results is the restricted domain knowledge: to compensate that, it's useful to confront some experts in the field. The team comprehends computer scientists, statisticians, biologists and healthcare workers.

The tools to analyse and elaborate the health data are:

- **PostgreSQL**, for data management and querying;
  - The software project to interface with the web server is **PgAdmin 4**;
  - All queries need to be optimized to avoid huge computational times, using indexes and Common Table Expressions;
- **R**, for machine learning and graphics computing;
  - There are many ways to create plots;
  - Possible uses are time series analysis and trajectory clustering.

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

Due to the amount of sensitive data, detailed results won't be shared: the final conclusions will be a product of aggregation and schematisation.

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 take 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.[6]

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;
2. Some diagnoses descriptions don't match with the corresponding ICD-9 code;
3. A lot of missing data is originated when 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 don't need prescriptions, so there is no entry in the DB;
6. General practitioners might prescribe a medicine to a different patient than the one who actually needs it (relatives, friends, ...);
7. There is inappropriate prescribing, 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 patterns of incorrect data:

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

Overall, the loss on single data 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 *pazienti* and *nos_ 002* (1 million of tuples)

Summary

Both tables contain similar data related to patients, with a $1:1$ correspondence between primary keys, so joining them 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: $99.981 \rightarrow 9,84\%$.

pazienti_ ok

An auxiliary view *pazienti_ ok* has been created to highlight the information loss augmentation on patients and be compared to the original tables.

The view is derived by the union of *pazienti* and *nos_ 002*, with the following conditions:

1. Join on equal patient code (*codice*);
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 lost.

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

4.2.2 *cart_pazpbl* (15 millions of tuples)

ICD-9

An ICD-9 code is correct if it's in the official ICD-9 database.[5] 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 none 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 *cart_terap* (118 millions of tuples)

ATC code

The ATC code, unlike ICD-9, has only one possible format (a numeric string of 6 digits), and no parsing is needed. All the codes have been checked with the ontologies in a well-known biology portal,[7] 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 relevant 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 AC 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 *te_ npezzi* (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 important to focus on chronic patients analysis.

4.2.4 *cart_ accert* (151 millions of tuples)

Summary

- Null or empty examination codes: $208.939 \rightarrow 0,14\%$;
- Null or empty descriptions: $9.377 \rightarrow 0,006\%$;
- Dates out of range: $4.451.626 \rightarrow 2,94\%$.

The official ICD-9-CM database is available for future checks on correctness of the codes.

Chapter 5

Patient journey based on first-time prescriptions

After the preliminary analysis and deciding the main focus area, there are enough information to begin reconstructing the **patient journey**. The goal of this process is to highlight changes in *prescription patterns for chronic diseases*, starting from the medical history of patients.

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 very 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 for a chronic disease have been deleted.

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

5.1 Imposed criteria

Aside from the completeness and correctness of the data, there are more restrictions to maintain the 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: *pa_ drevoca* for interruption of the relationship, *decesso* for death and *pa_ convenzione* for the sanitary convention.

The table *users* contains all the IDs of active general practitioners, so joining it with other tables is enough to remove all the rows with an inactive GP. Data is up to date, but since the patient journey will include 2018 (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 prospective.

5.2 Initial data cleaning

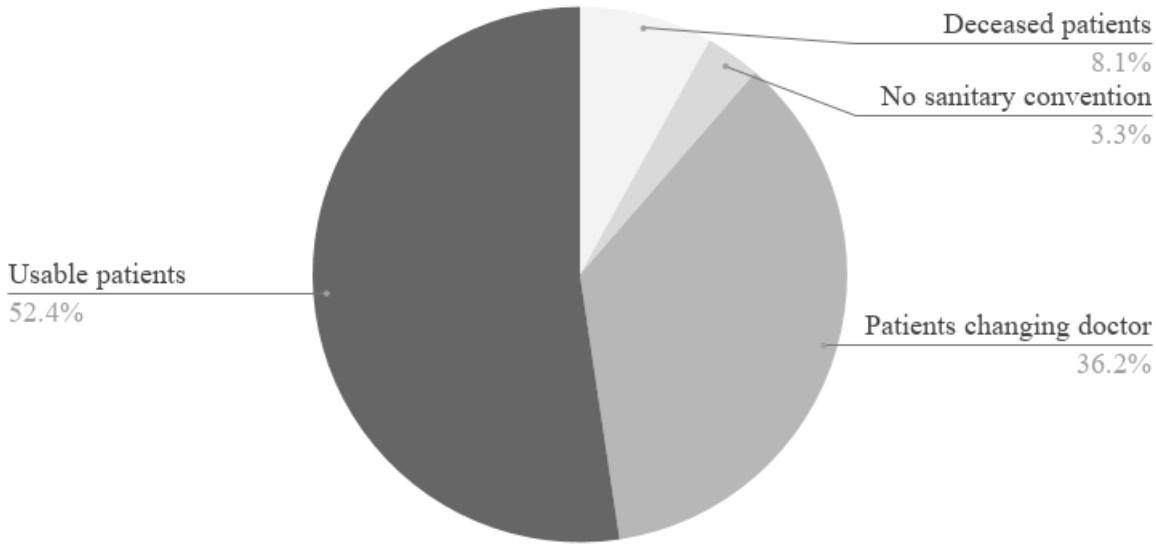
An initial data cleaning has been made on the whole database to have a first understanding of the potential information loss.

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 that weren't present in *nos_ 002* (6).

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



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.

5.3 First approach

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

The analysis has pure research purposes, to understand the impact of a cut of 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.

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 of first-time diagnosis and prescriptions to patients.

5.3.1 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 huge 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%;
- Diagnosis: 54,6%;
- Prescriptions: 44,7%.

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.

5.4 Second approach

Since just removing according to the date is an abrupt approach, it's necessary to tune parameters and introduce more detailed constraint, to have a bigger 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.

5.4.1 Results breakdown

The 1.665.005 complete tuples are composed by:

- 230.381 patients;
- 422 general practitioners;
- 1.381 unique diagnoses;
- 904 unique prescriptions.

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

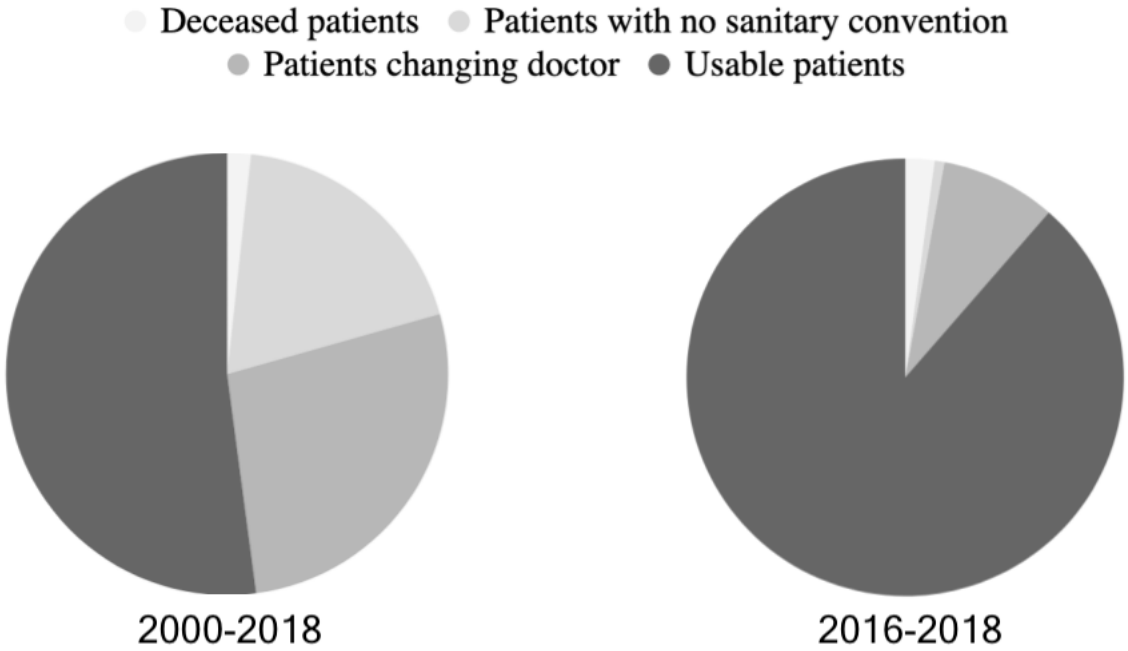
5.5 Results comparison

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

5.5.1 Changes in data composition

5.5.2 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.

5.5.3 Funnel graph of patients

A funnel graph is useful to see how imposing every restriction made the patients number decrease: starting from a million, in the end there only is about 1/4 of it.

Bibliography

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

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

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

[4] Made with draw.io.

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

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

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