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.