Integrating the Italian National Healthcare Registries

Alexandre Crivellari 902064

Muluken B. Megersa

Alessandro Rota

# Summary

Lorem Ipsum

# Index

Summary 1

Index 1

Introduction 1

Datasets and Acquisition Methods 1

Process Overview 3

Software Architecture 3

Conclusions and Issues Faced 3

Future Developments 3

Operational Guide 3

References 3

# Introduction

The topic of scarcity of medical personnel is greatly discussed both on a national [1] and European level [2] and is one of the main political topics being discussed in recent years in mostly all developed countries.

There are many reasons behind why such shortage has happened in the first place: the **increasing ageing of the population** (especially in the European Union, where fertility and natality rates have dropped), along with an **inadequate planning on medical education** across all countries, has led to a situation where **there are more doctors and medical practitioners going into retirement** (or early retirement, due to stress and unsatisfactory working conditions, that will not be addressed in this introduction) **than there are newly educated professionals** entering the healthcare market.

While this stands true also in the Italian environment, as confirmed both from field experts and public authorities, a **clear and publicly accessible collection of data about the medical population is still missing.** While all medical practitioners (doctors and other healthcare professionals) are obliged to be inscribed in publicly available records, these records are not meant to provide information on an aggregated level, as their purpose is to just provide verifiability for any person to possess all the requirements to act as a medical practitioner.

In addition, there is not only “*one”* public registry of medical practitioners *per se*: in reality, the National Registry of Medical Doctors (which will be discussed further in the next section) acts as a collector of all the different Local Registries from all the Italian provinces, thus **bringing differences on what information is collected and in what format is stored**.

The main objective of this project, therefore, has been to find a way to **understand how the Italian medical population is structured** and to **make the data available for exploration, aggregated analysis and search for insights**. By doing so, we believe it is possible to improve how planning on medical education is performed for the near future, as well as gaining a deeper knowledge on how threatening the situation really is.

# Datasets and Acquisition Methods

First of all, the group has analyzed how the information on medical professionals is presented across the different public registries.

Since 2023, both the National Federation of Medical Practitioners and Dentists (*“Federazione Nazionale Ordine dei Medici Chirurghi e Odontoiatri”, FNOMCEO* [3]) and the National Federation of Medical Professionals (which includes *“Federazione Nazionale Ordine Fisioterapisti Italiani” - FNOFI* [4]for physiotherapists and *“Tecnici Sanitari di Radiologia Medica e Professionisti Sanitari delle Terapie, Riabilitazione e Prevenzione” – TSRM PSTRP* [5]for all the other medical professions, aside from nurses) are available online for public access and for checking whether an individual is actually licensed to perform medical care as a certified healthcare professional.

This has been a welcomed change. Prior to this it was already possible to perform said check, but the information about an individual’s belonging to a specific registry was seldom fragmented across all the registries across the different provinces (especially with regards to the FNOFI and TSRM-PSTRP registries). By merging all provincial registries into one, unique public registry, checking for a medical practitioner’s credentials has become much easier for the population (although with some issues on the quality of data, which will be discussed later on).

A particular characteristic about how a single search is performed in both FNOMCEO and FNOFI-TSRM-PSTRP makes this project viable, since **it is possible to obtain information about any particular professional by just providing his or her surname**. Any search made by providing a surname returns, in fact, a tabular list of all medical professionals with such surname, along with some of his or her personal data, in detail:

For **FNOMCEO**:

* *Name*
* *Surname*
* *Date of Birth*
* *Place of Birth*
* *Date of Graduation*
* *Name of Degree*
* *Date of Qualification*
* *Session No. of Qualification*
* *Province of Registration*
* *Number of Registration*
* *Date of Specialization 1 (if any)*
* *Name of Specialization 1 (if any)*
* *(any other additional Specialization)*
* *(any other additional Special Qualification)*

For **FNOFI-TSRM-PTSRP**:

* *Name*
* *Surname*
* *Date of Birth*
* *Place of Birth*
* *Date of Graduation*
* *Name of Medical Profession*
* *Date of Qualification*
* *Province of Registration*
* *Number of Registration*
* *(any other additional Specialization)*
* *(any other additional Special Qualification)*

Since there is no available API for accessing this data, and since the goal is to reverse-engineer the underlying complete datasets for aggregate operations on all the medical population, the idea is therefore to **obtain a comprehensive list of all the surnames in Italy** and use it to **scrape the complete datasets via Python script,** collecting all the information about the medical professionals one surname at a time, for both registries.

One dataset, then, is still needed: **the complete list of all surnames in Italy** to use as keys for our search on FNOMCEO and FNOFI-TSRM-PSTRP. Luckily an Italian website, Cognomix [6], seems to already have the information needed: a comprehensive list of more than 42.000 surnames, that would hypothetically seem to be enough to scrape both registries (the expected amount on Medical Doctors and Dentists in Italy is around 400.000 individuals, while the expected amount of other medical professionals is around 69.000 for FNOFI and 173.000 for TSRM-PSTRP).

# Data Acquisition and Storage

***Data Acquisition and Storage -*** The logical process and the datasets acquisition is described as follows:

1. First, the list of all Italian surnames is obtained via scraping with a Python script using Selenium on Cognomix.com. The complete list of 42.666 records is stored locally as a light CSV file for later.
2. Via another Python script and using Selenium and BeautifulSoup, the csv list of all Italian surnames is used on <https://portale.fnomceo.it/cerca-prof/index.php> (FNOMCEO Registrry search engine) to obtain a tabular HTML list of all doctors with any given surname. Each record of the HTML page is then stored structurally onto a JSON file as a Document object.
3. The same is being done for <https://albo.alboweb.net/registry/search> (TSRM-PSTRP) and <https://albo.alboweb-fnofi.net/registry/search> (FNOFI). Both registries are absolutely identical in the form structure and HTML logic and will parallelly be stored onto different JSONs.

The reason for why **the group has adopted a NoSQL approach** is quite straightforward: given that any doctor or medical professional may have zero, one or multiple specializations in any medical field, and given that the national registries are the result of a (not completely well-performed) integration of several different local registries with some differences between them, a too structured SQL approach would have probably been both potentially **inefficient** (for all the individuals without any specialization) and **dangerous** (we do not know what is the maximum number of specializations obtained within all Italian doctors). With a Document-based approach, on the other hand, it is possible to maintain the data structure needed for accommodating any information regarding each record. Additionally, a Document-based approach will make the future queries on the integrated databases much more performative: for example, when searching for only a specific medical specialization (e.g.: dermatology) we won’t have to look for all medical professionals from FNOFI or TSRM-PSTRP (that could never be also dermatologists). Finally, still via Python script the different JSONs obtained during the different scraping sessions are appended together and prepared for integration.

# Data Quality Assessment and Cleaning

The nature of the work has a few characteristics that must be considered for evaluating data quality both on a subjective basis and an objective one.

The level of original fragmentation of the dataset, caused by both a fragmentation of the different medical professions along with the different record keeping structure in place in the different provinces, has made impossible to have *one, certain, consistent, non-variable and officially accepted* number of medical professionals at any given time. There is, since 2023, a “certain enough” number of individuals registered at FNOFI (69.848), at FNOMCEO (*around* 357.000) and at TSRM-PSTRP (*around* 173.000), as declared by the respective organizations. For this reason, in terms of evaluating the **completeness** of the dataset, it has been decided to find out if the data scraped by the three registries was far from these expected results or not. Naturally, assuming an *open world* scenario, we are defining exclusively an **Object Completeness** (as it is the case with NoSQL exercises such as this one).

With this in mind:

* FNOMCEO dataset resulted in 357.084 records, which is *exactly* the expected amount of data (100%);
* FNOFI dataset resulted in 48.313 records, which accounts for 69,13% of the expected records;
* TSRM-PSTRP dataset resulted in 91.559 records (52,92% of the expected results).

The reason for this incomplete collection of records for FNOFI and TSRM-PSTRP does not come surprising: while FNOMCEO has centralized their National Registry since 2006, this has not been the case for the other two registries, who have begun their process of data centralization only in the end of 2023. For this reason, it may still take a while for all the local registries (provinces) to integrate their data into the new national ones.

In terms of **Temporal** quality of dataset, it is important to guarantee that the data is up to date both in terms of new professionals inscribed and in terms of other professionals retiring or withdrawing from the medical profession (for incompatibility with other activities or for any other change in their status). While it would be extremely difficult to have any information on the latter change of status (it is, in fact, possible to maintain the medical practice even after retirement), it is important to **keep the dataset Current** if it is going to be used for updated insights. The last update of records for both FNOMCEO and FNOFI-TSRM-PSTRP falls between 19th of January and 15th of February and, given that no additional inscription sessions have been done in the meantime, we can consider the database up to date. We will not consider any Timeliness quality of the database, since the goal of the project is to receive insights on a “enough-updated” dataset of medical professionals and we do not need a strict more recent update of that data (which is not bound to frequent change).

In terms of **Consistency**, especially in terms of **integrity constraints**, the group has made sure that all information was correctly defined at the source level. Specifically, we have made sure that **all records presented the correct format for all main attributes** (as in: all birth dates had to be in the same date format, all registration numbers had to be numbers[[1]](#footnote-1), all the other attributes had to be strings).

There has been, however, an issue in terms of **Consistency with different representations of the same data**: for example, Radiologic Technologists (*Tecnico Sanitario di Radiologia Medica*, or TSRM for short) had been classified as “TSRM” or “T.S.R.M.”, depending on the registry of initial provenance. For this reason, a Python dictionary has been manually applied to all records with the purposes of **cleaning the data**, thus eliminating the redundancies and allowing for the correct query execution.

# Exploratory Analysis

Min-Max for Last\_Update

Distribution for age: FNOMCEO, FNOFI, ALBOWEB, with comments

Table count for all different medical specializations and each of the 18 medical professions (so, divided between FNOMCEO and FNOFI-TSRM-PSTRP)

# Database Integration

How?

# Software Architecture

Languages, Frameworks and tools used.

# Conclusions and Issues Faced

Before

# Future Developments

Aside from the issue on Completeness of the dataset (which should actually solve itself when all the local province registries will correctly input their data into the correct national database), an interesting development would be to deduce a point of contact **with every single professional**, thus **enriching** the dataset with additional, valuable information, such as an **e-mail address**. In particular, given that:

1. All medical professionals are obliged to have a PEC address (*Posta Elettronica Certificata*);
2. Such PEC address must be publicly available on *Indice Nazionale Indirizzi PEC* (INIPEC, <https://www.inipec.gov.it/cerca-pec>);
3. It is possible to make a search for a PEC address by inputting the Registry name (FNOMCEO, FNOFI or TSRM-PSTRTP) and the inscription number for each individual of the dataset;

It would be theoretically possible to scrape the PEC addresses for all practitioners (considering, of course, the presence of a reCAPTCHA in Inipec search). Naturally this would have a whole lot of implications, especially on a GDPR perspective, that will not be discussed in this report but that must be considered in evaluating this idea.

Another possible development for this project is to make the dataset available for public consultation with an **Open Data access**, in order to make this aggregated overview of the state of the healthcare market available for public consultation and further analysis. It would of some use, for example, for the public (and very much current) discussion on the **seat availability in medical universities**, a very controversial topic at the present moment.

# Operational Guide

The public repository and history for the whole project can be found on Github, at the link provided in the Reference paragraph of this report [7].

The repository contains different Jupyter Notebooks, along with some files used for indexing (which will be further explained later on this chapter) and the different folders for the different registries.

For replicating the project, it is required to **first delete the following files**:

* last\_index.txt
* fnofi\_last\_index.txt
* alboweb\_last\_index.txt

The reason for this is that, depending on the machine used and on the online availability of the different websites (or, rephrased: the probability of receiving an error during the scraping process from the websites, due to *memory errors* or *request denials*), **the Python script saves a useful index** for each iteration during the process, so that if anything happens during the scraping activity (i.e., loss of connectivity) all data acquired until that moment is still saved, and the scraping can begin again from that precise point. This has proven extremely useful, as **each registry requires around 18-20 hours for being completely scraped** (this is also the result of limiting the amount of requests sent to the website at the same time).

At this point there are **four main Python scripts** available for use:

1. The first one is *scraping\_surename.ipynb*, which will scrape all surnames from cognomix.com and provide the most complete list of surnames in Italy for the subsequent scraping phase;
2. Once the scraping from Cognomix is done and we have obtained and saved the surname.csv file (the keys for the main scraping phase), it’s possible to run the scraping for the three registries by running the other Python scripts:
   1. *scraping\_fnomceo\_data* for FNOMCEO;
   2. *scraping\_fnofi\_data* for FNOFI;
   3. *scraping\_alboweb\_data* for TSRM-PSTRP.
3. If any scraping is interrupted on any registry, the user has to run again the final cell (containing the *main()* function) on the stopped Jupyter notebook: the scraping will resume from where it was interrupted.
4. After around 19 hours for each notebook, several jsons will be created for each registry.

# References

[1] Pennisi F, Minerva M, Dalla Valle Z, Odone A, Signorelli C. *Genesis and prospects of the shortage of specialist physicians in Italy and indicators of the 39 schools of hygiene and preventive medicine*, 2023. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/37695180/>

[2] Henley J, Connolly K, Jones S, Giuffrida A. *‘A ticking time bomb’: healthcare under threat across western Europe*, 2022. Retrieved online from <https://www.theguardian.com/society/2022/dec/14/a-ticking-time-bomb-healthcare-under-threat-across-western-europe>

[3] <https://portale.fnomceo.it/cerca-prof/index.php>

[4] <https://albo.alboweb-fnofi.net/registry/search>

[5] <https://www.tsrm-pstrp.org/>

[6] <https://www.cognomix.it/origine-cognomi-italiani>

[7] <https://github.com/MulukenMegersa/data_management_project_2024.git>

1. Identification numbers are, of course, strings and do not have numerical properties. For the purpose of control on integrity constraints, however, this has been the fastest way to make sure that these strings did not have any alphabetical characters or whitespaces in them. [↑](#footnote-ref-1)