1. Introduction

The aim of this project is to create a structured dataset containing regulatory and clinical information on licensed drugs in Italy. As part of the project, publicly available official data provided by the Italian Medicines Agency (AIFA) were systematically collected, including both lists of prescription drugs and their corresponding clinical descriptions.

The data collection process was largely carried out manually. For each drug, relevant sections were extracted from the “Riassunto delle Caratteristiche del Prodotto (RCP)” documents through individual searches. The project specifically focused on "Class A" and "Class H" drugs, which were merged, cleaned, and standardized into a unified data structure.

The resulting dataset is intended to enable meaningful insights in both healthcare analytics and pharmaceutical policy development. It aims to serve as a robust resource for the development of clinical decision support systems, drug safety analyses, and therapeutic evaluations.

1. Sources Used
   1. Official Data Sources

AIFA Drug Lists Portal (<https://www.aifa.gov.it/liste-dei-farmaci>)

The primary data source for the project was the Class A and Class H drug lists provided by the Italian Medicines Agency (AIFA). These lists present officially authorized drugs in Italy in a structured manner, including key regulatory information such as active ingredient, drug name, marketing authorization holder, and AIC code. This information served as the initial dataset and was later enriched with clinical data.

AIFA Drug Search and RCP Portal (<https://medicinali.aifa.gov.it/it/#/it/>)

To access clinical information on drugs, AIFA’s online search platform was used. Each drug was queried in the system using its AIC code to retrieve the "Riassunto delle Caratteristiche del Prodotto (RCP)" documents. These documents are official clinical resources prepared in accordance with European and Italian pharmaceutical regulations, containing sections such as dosage, indications, contraindications, side effects, and interactions. Only the mandatory sections were systematically extracted for the project.

* 1. Software and Technological Tools Used

**Python & Selenium:**

To improve time efficiency during the clinical data extraction phase, the Python programming language was used in combination with the Selenium library. Selenium was employed to automate web browser interactions—such as inputting AIC codes into the portal, navigating to specific pages, and retrieving targeted content. This allowed for an automated data retrieval process instead of manual browsing.

**Microsoft Excel & Power :**

Query Excel served as the foundation for data manipulation in the project. Drug lists were merged, unnecessary columns were removed, and a clean data structure was prepared in this environment. With the help of the Power Query add-in, large datasets were filtered, empty cells removed, and transformation operations carried out quickly and accurately.

**Power BI :**

The constructed dataset was imported into Power BI for visualization. Key outputs—such as the distribution of active substances, frequency of therapeutic indications, and patterns of side effects—were presented in a user-friendly manner via graphical dashboards. Power BI’s filtering and interactive analysis capabilities contributed to making the data more meaningful.

**PostgreSQL :**

The dataset was also restructured in a relational database environment. Using PostgreSQL, data normalization was performed and logical relationships were established between tables. This setup enabled the grouping of drugs based on specific attributes and allowed for the extraction of analytical insights through SQL queries.

1. Methodology
   1. Collection and Preparation of Regulatory Data

The first phase of the study began with obtaining the drug lists published by the Italian Medicines Agency (AIFA). Two separate Excel files, containing drugs categorized under Class A (prescription and reimbursable drugs) and Class H (for hospital use), were downloaded from AIFA’s official website (https://www.aifa.gov.it/liste-dei-farmaci). These files contained approximately 12,600 rows of data in total.

However, considering the time and resource constraints of the project, a representative sample of drugs was selected from each class instead of using the entire dataset. Only a total of 300 drugs were used in the study, chosen through random sampling to preserve representativeness. This approach aimed to make the data collection process more manageable and to accelerate subsequent analyses.

The selected data were simplified and structured to include the following fields:

* Active Ingredient (Principio Attivo)
* Group Description (Descrizione Gruppo)
* Drug Name and Packaging (Denominazione e Confezione)
* Marketing Authorization Holder (Titolare AIC)
* AIC Code (AIC)
* Equivalent Group Code (Codice Gruppo Equivalenza)
* Class (Class A or H)

These merged datasets in Excel were then cleaned using Power Query by removing empty rows and duplicate entries. The resulting simplified list served as the foundation for the main table, which would later be enriched with clinical information.

* 1. Collection of Clinical Information (RCP Data)

In the second phase of the study, the goal was to obtain clinical information for each drug. To achieve this, the Italian Medicines Agency's drug search portal (https://medicinali.aifa.gov.it/it/#/it/) was used to search for each drug by its AIC code. This process was automated using the Selenium library instead of being carried out manually.

* The developed automation process consisted of the following steps:
* AIC codes for the 300 selected drugs were sequentially entered into the portal using Selenium.
* Each drug's detail page was accessed.
* The link to the “Riassunto delle Caratteristiche del Prodotto (RCP)” PDF on the page was automatically detected.
* The ATC code displayed on the same page was also automatically extracted.
* The obtained PDF link and ATC code were saved into a new CSV file, matched with the corresponding drug AIC.

At this stage, the RCP documents were retrieved as PDFs, but full-text extraction of the content was not performed yet. Text processing of the clinical content was carried out in the next step.

This newly created CSV file was then merged with the previously prepared base drug list (see Section 3.1) using the AIC code. As a result, for each drug, both the regulatory data (e.g., active ingredient, group, marketing authorization holder) and the related RCP PDF link and ATC code were integrated into a single row, forming a unified structure.

Following this merge, the resulting table included:

* Identification and regulatory information
* Link to the clinical document (PDF)
* ATC classification

This structure served as the foundational framework for the next phase, which involved processing each PDF's content and segmenting it into clinical sections(4.1 – 6.2).

* 1. Structuring and Finalizing the Dataset

The third and final phase of the project involved transforming all collected regulatory and clinical information into a structured dataset. This process combined automated steps with manual interventions.

The RCP PDF links obtained via Selenium in the previous phase were systematically downloaded and saved locally. Although the initial goal was to automatically extract the textual content from these PDFs, the presence of watermarks and embedded subtext layers in the documents prevented reliable automation.

As a result, each PDF document was manually reviewed, and the following mandatory clinical sections were individually read and manually entered into the dataset:

4.1 Therapeutic Indications

4.2 Dosage and Method of Administration

4.3 Contraindications

4.4 Special Warnings and Precautions

4.5 Interactions with Other Medicinal Products

4.6 Fertility, Pregnancy, and Lactation

4.7 Effects on Ability to Drive and Use Machines

4.8 Undesirable Effects

4.9 Overdose

6.2 Incompatibilities

Each section was defined as a separate column in the dataset, and the contents were entered in full text. However, in cases where any of these sections were missing or inaccessible in a drug’s RCP document, that drug was excluded from the dataset. This ensured that only drugs with complete clinical documentation were included.

As a result, a structured dataset was created in which each row represents a single drug and contains both regulatory and clinical information. This dataset was then imported into Power BI and integrated into a PostgreSQL relational database system, making it ready for both visual analysis and SQL-based querying.

* 1. Data Visualization and Transformation

Once the dataset was finalized, it was imported into the Power BI environment for analysis and presentation. Since most of the clinical texts consisted of long paragraphs, a transformation process was applied through Power Query to make this content analyzable. Using a Python script integrated within Power Query, the clinical sections were segmented into smaller parts. Key concepts—such as keywords, mentioned drug names, frequencies of side effects, and similar elements—were extracted from the texts. As a result, the textual data was converted into numerical indicators and structured in a way suitable for visualization.

Finally, an interactive dashboard was created in Power BI, enriched with visual elements such as charts, distribution views, and filters to facilitate user-friendly data analysis. In particular, metrics such as:

* Distribution by active ingredient
* Frequency of therapeutic indications
* Overall frequencies of side effects

were made easily analyzable.

* 1. Database Integration

The finalized dataset was transferred to a PostgreSQL environment to achieve a more organized and manageable structure. The data was formatted to comply with a relational database model, with key fields preserved and appropriate tables created. This transfer enabled more systematic access to the data and established a suitable infrastructure for future analyses.

1. Choices Made

During the project, several decisions were made based on time management, technical constraints, and data processability.

* Instead of using the full dataset of approximately 12,600 rows provided by AIFA, a total of 300 drugs were selected to represent both Class A and Class H. This ensured a data-focused yet manageable sample size.
* To obtain clinical information, RCP documents in PDF format were downloaded in bulk. However, due to watermarks and layered structures within the files, automated text extraction was not feasible. As a result, all clinical content was entered manually.
* Drugs with missing or inaccessible clinical information were excluded from the dataset. Only records containing complete information were included in the project.
* To facilitate analysis of the clinical texts, a Python integration within Power Query was used to process and simplify the content by extracting keywords.
* The dataset was also prepared for use in Power BI and PostgreSQL environments, thereby creating the necessary infrastructure for advanced analysis and visualization processes.

1. Difficulties Encountered

Throughout the project, various technical and operational challenges were encountered. In particular, the data collection and text extraction phases proved to be more labor-intensive than initially expected.

* The fact that the RCP documents were in PDF format significantly complicated the text extraction process. Watermarks and hidden content layers within the PDFs made automated text reading unsuccessful.
* As a result, all clinical information had to be entered manually, which made the process extremely time-consuming.
* In some cases, drug pages on the AIFA portal failed to load or were missing ATC/RCP information, which led to the exclusion of certain drugs from the dataset.
* The length and density of the clinical content caused performance issues and slowdowns when working in environments such as Excel.
* Finally, preparing the dataset for integration into Power BI and PostgreSQL took considerable time to ensure compatibility, but the process was ultimately completed successfully.

1. Conclusion

In this study, a structured dataset containing regulatory and clinical information on licensed drugs in Italy was successfully created. Based on open-source data provided by AIFA, information on 300 drugs was carefully selected and compiled through a manual processing approach. Due to the PDF format of the clinical content and its unsuitability for automated text extraction, the content was collected manually; only drugs with complete information were included in the dataset. The collected data was then structured to be compatible with both visualization tools (Power BI) and database integration (PostgreSQL). This project not only provided hands-on experience with real-world data but also involved a multifaceted data processing workflow requiring the integrated use of various tools. The resulting dataset holds significant potential for use in health policy development, clinical decision support systems, and academic research.