

Construction of a Dataset on Italian Medicines

Objective

The aim of this assignment is to **construct a structured dataset** based on official data provided by the Italian Medicines Agency (AIFA), including information about **regulatory and clinical aspects of** medicines that are currently authorized in Italy.

The dataset is expected to be built **primarily through manual data extraction (so the students will have to extract the data as is, i.e., in Italian)**, simulating the work of a data extractor. However, students with the appropriate skills may also use **automated tools** (e.g., web scraping or text parsing), **provided they clearly document the process** in their final report.

Overview of the Workflow

Step 1 – Downloading and Merging Medicine Lists

1. Visit the AIFA website: <https://www.aifa.gov.it/liste-dei-farmaci>

Lista farmaci di classe A e H per consentire a tutti gli Operatori la prescrizione per principio attivo

Informazioni sul dato: Liste di Trasparenza - Lista farmaci di classe A e H per consentire a tutti gli Operatori la prescrizione per principio attivo

Descrizione:	Tabelle contenenti l'elenco dei farmaci di fascia A, dispensati dal Servizio sanitario Nazionale, ordinati per principio attivo e per nome commerciale, al fine di consentire, per tutti gli Operatori sanitari, la prescrizione per principio attivo disposta dall'articolo 15, comma 11-bis, del decreto legge 6 luglio 2012 n.95, convertito con modificazioni dalla Legge 7 agosto 2012 n. 135. Disponibile anche la tabella di farmaci di classe H.
Data ultimo aggiornamento:	14/02/2025
Tipo file: CSV	Download (Classe A - per principio attivo) al 01-08-2024 Download (Classe A - per nome commerciale) al 01-08-2024 Download (Classe H - per principio attivo) al 01-08-2024 Download (Classe H - per nome commerciale) al 01-08-2024

2. Download the Excel files corresponding to:
 - **Class A medicines** (about 10400+ rows);
 - **Class H medicines** (about 2100+ rows).
3. **Open both files and merge** them into a **single spreadsheet** (CSV format). You should **retain only the following columns**:
 - **Active Ingredient** (Principio Attivo)
 - **Group Description** (Descrizione Gruppo)
 - **Medicine Name and Packaging** (Denominazione e Confezione)
 - **Marketing Authorization Holder** (Titolare AIC)
 - **AIC Code** (AIC) – this is a unique identifier for each medicine

- **Equivalence Group Code** (Codice Gruppo Equivalenza)
- **Class** (A or B)


An extract from the Table obtained:

1	Principio Attivo	Descrizione Gruppo	Denominazione e Confezione	Prezzo al pubblico €	Titolare AIC	AIC
2	Acamprosato	ACAMPROSATO 333MG 84 CAMPRAL*84 cpr riv gastrores 333 mg		33,77	BRUNO FARMACEUTICI SpA	034208013
3	Acarbosio	ACARBOSIO 100MG 40 UN ACARBOSIO*40 cpr 100 mg		5,63	TECNIMEDE SOC.TECNICO-MED.S.A.	039716182
4	Acarbosio	ACARBOSIO 100MG 40 UN ACARBOSIO*40 cpr 100 mg		5,63	DOC GENERICI Srl	044155024
5	Acarbosio	ACARBOSIO 100MG 40 UN ACARPHAGE*40 cpr 100 mg		5,63	BRUNO FARMACEUTICI SpA	038835144
6	Acarbosio	ACARBOSIO 100MG 40 UN GLUCOBAY*40 cpr 100 mg		8,04	BAYER SpA	026851016
7	Acarbosio	ACARBOSIO 50MG 40 UNI ACARBOSIO*40 cpr 50 mg		5,63	TECNIMEDE SOC.TECNICO-MED.S.A.	039716170
8	Acarbosio	ACARBOSIO 50MG 40 UNI ACARBOSIO*40 cpr 50 mg		5,63	DOC GENERICI Srl	044155012
9	Acarbosio	ACARBOSIO 50MG 40 UNI ACARBOSIO*40 cpr 50 mg		5,63	AUROBINO PHARMA ITALIA Srl	047612027
10	Acarbosio	ACARBOSIO 50MG 40 UNI ACARPHAGE*40 cpr 50 mg		5,63	BRUNO FARMACEUTICI SpA	038835043
11	Acarbosio	ACARBOSIO 50MG 40 UNI GLUCOBAY*40 cpr 50 mg		8,04	BAYER SpA	026851028
12	Acarbosio	ACARBOSIO 50MG 40 UNI GLUCOBAY*40 cpr 50 mg		7,49	GMM FARMA Srl	045430016
13	Acarbosio	ACARBOSIO 50MG 40 UNI GLUCOBAY*40 cpr 50 mg		7,49	FARMA 1000 Srl	045461011
14	Acarbosio	ACARBOSIO 50MG 40 UNI GLUCOBAY*40 cpr 50 mg		7,49	NEW PHARMASHOP Srl	047923014
15	Acetolololo	ACEBUTOLOLO 400MG 30 SECTRAL*30 cpr 400 mg		10,71	CHEPLAPHARM ARZNEIMITTEL GMBH	022155057
16	Acetolofenac	ACECLOFENAC 100MG 30 AIRTAL*orale sosp polv 30 bust 100 mg		7,73	ALMIRALL S.A.	032773032
17	Acetolofenac	ACECLOFENAC 100MG 30 GLADIO*orale polv 30 bust 100 mg		7,91	ABIOGEN PHARMA SpA	031220027
18	Acetolofenac	ACECLOFENAC 100MG 30 KAFENAC*orale sosp polv 30 bust 100 mg		7,69	ALMIRALL S.A.	031842026
19	Acetolofenac	ACECLOFENAC 100MG 40 ACECLOFENAC*40 cpr riv 100 mg		5,64	ACCORD HEALTHCARE SLU	042403042
20	Acetolofenac	ACECLOFENAC 100MG 40 ACECLOFENAC*40 cpr riv 100 mg		5,64	EG SpA	043259035
21	Acetolofenac	ACECLOFENAC 100MG 40 AIRTAL*40 cpr riv 100 mg		10,26	ALMIRALL S.A.	032773020
22	Acetolofenac	ACECLOFENAC 100MG 40 GLADIO*40 cpr riv 100 mg		10,57	ABIOGEN PHARMA SpA	031220015
23	Acetolofenac	ACECLOFENAC 100MG 40 KAFENAC*40 cpr riv 100 mg		10,26	ALMIRALL S.A.	031842014
24	Acenocumarolo	ACENOCUMAROLO 1MG 2 SINTROM*20 cpr 1 mg		1,91	MERUS LABS LUXCO II SARL	011782024
25	Acenocumarolo	ACENOCUMAROLO 4MG 2 SINTROM*20 cpr 4 mg		2,03	MERUS LABS LUXCO II SARL	011782012

Example of table rows
and their respective
columns

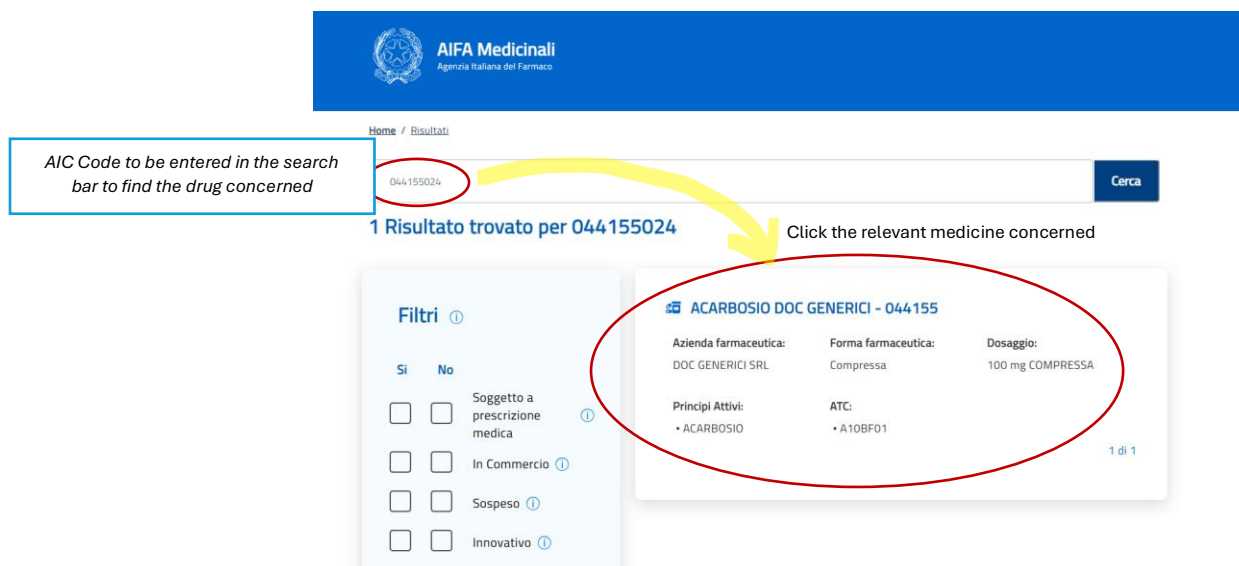
4. Clean the dataset by:

- Removing any **empty rows**;
- Eliminating **duplicate entries**.

 **Note:** Each group will be required to extract and collect a representative sample of approximately 1,300 drugs. Assignment to each group of the drugs to be extracted will be based on the completion of a spreadsheet that will be uploaded to the course's Teams channel.

Step 2 – Retrieving Clinical Information from the Summary of Package Leaflet

1. Go to the following AIFA portal: <https://medicinali.aifa.gov.it/it/#/it/>
2. Use the **AIC code** from the merged file to **search for each medicine**, one at a time.



The screenshot shows the AIFA Medicinali portal interface. At the top, the header reads "AIFA Medicinali Agenzia Italiana del Farmaco". Below the header, the search bar contains the AIC code "044155024", which is circled in red. A yellow arrow points from this code to the search results. The search results show "1 Risultato trovato per 044155024". The result is for "ACARBOSIO DOC GENERICI - 044155". A red circle highlights the details of this result, including the manufacturer "DOC GENERICI SRL", the form "Compresa", the dosage "100 mg COMPRESSA", and the active ingredient "ACARBOSIO". A filter sidebar on the left shows various filters like "Soggetto a prescrizione medica", "In Commercio", "Sospeso", and "Innovativo".

3. Open the **product page** and locate the document known as:

- **“Summary of Product Characteristics”** (Riassunto Caratteristiche Prodotto), also referred to as the **package leaflet** or **“bugiardino”**.

The screenshot shows the product page for ACARBOSIO DOC GENERICI 100 mg COMPRESSA. The page is divided into several sections. On the left, there is a link to the 'Foglio Illustrativo' (package leaflet). In the center, the 'Principi Attivi' (Active Ingredients) section lists 'ACARBOSIO'. To the right of this, the 'ATC' (Anatomical Therapeutic Chemical) code is listed as 'A10BF01 - ACARBOSIO'. On the far right, there is a link to the 'Riassunto Caratteristiche Prodotto' (Summary of Product Characteristics). Below the main product information, there is a section for 'Confezioni' (Packaging) which includes details about the '100 MG COMPRESSE' (100 mg tablets) and the '40 COMPRESSE IN BLISTER PVC/PE/PVDC-AL' (40 tablets in blister pack).

ACARBOSIO DOC GENERICI 100 mg COMPRESSA

Azienda titolare: DOC GENERICI SRL

[Foglio Illustrativo](#)

[Riassunto Caratteristiche Prodotto](#)

Principi Attivi

- ACARBOSIO

ATC

- A10BF01 - ACARBOSIO

Vie di somministrazione

- ORALE

Forma farmaceutica

Compresse

Confezioni

"100 MG COMPRESSE" 40 COMPRESSE IN BLISTER PVC/PE/PVDC-AL

AIC

AIC 044155024

Data autorizzazione

11/02/2017

Classe di rimborsabilità

A Totale Carico Del Ssn.

Regime di Fornitura

RR Soggetto a Prescrizione Medica

IN COMMERCIO

4. Extract the **ATC** (as shown in the figure) by entering it as a new column in the spreadsheet and **manually open** and **extract** the contents of the following **mandatory sections**:

Required Sections to Include:

- **4.1** Therapeutic indications
- **4.2** Posology and method of administration
- **4.3** Contraindications
- **4.4** Special warnings and precautions for use
- **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 (side effects)
- **4.9** Overdose
- **6.2** Incompatibilities

5. Also extract the **leaflet URL** for each drug, e.g. in the previous case, the URL would be: <https://medicinali.aifa.gov.it/it/#/it/organizzazione/898/farmaci/44155/stampati/Fl> and save it in a column called **URL**.

Ultimately, the schema is:

Column Header	Description
AIC	Unique identifier code for the medicine (Autorizzazione Immissione in Commercio)
Principio Attivo	Active ingredient of the medicine
Descrizione Gruppo	Therapeutic group or category
Denominazione e Confezione	Commercial name and packaging description
Titolare AIC	Marketing authorization holder (usually a pharmaceutical company)
Codice Gruppo Equivalenza	Code indicating the equivalence group, if applicable
Classe	Reimbursement class (e.g., A or H)
ATC	Anatomical Therapeutic Chemical code (from the Summary of Product Characteristics)
Indicazioni terapeutiche	Therapeutic indications (section 4.1 of the SPC)
Posologia e modalità di somministrazione	Dosage and method of administration (section 4.2)
Controindicazioni	Contraindications (section 4.3)
Avvertenze e precauzioni	Warnings and special precautions (section 4.4)
Interazioni	Interactions with other medicines (section 4.5)
Fertilità, gravidanza e allattamento	Fertility, pregnancy, and lactation (section 4.6)
Effetti su guida e macchinari	Effects on ability to drive and use machines (section 4.7)
Effetti indesiderati	Undesirable effects or side effects (section 4.8)
Sovradosaggio	Overdose information (section 4.9)
Incompatibilità	Incompatibilities (section 6.2)
URL	Direct link to the official product leaflet (SPC) on the AIFA website

Step 3 – Final Dataset Construction

2. Create a CSV file. Each **row should correspond to a single medicine identified with AIC**.
3. For very long text content, you may:
 - Paste the full text directly into the corresponding cell of the spreadsheet
4. If any required section is **not available** for a given medicine, write: "Not available".

Deliverables (All Mandatory)

Each student or group must submit the following 3 items:

1. **Final dataset** in CSV format.
2. A **written report** (maximum 4–5 pages) including:
 - Sources used;
 - Detailed explanation of the methodology adopted;
 - Choices made (e.g., number of medicines, additional sections, how long texts were handled);
 - Any difficulties encountered during the process;
 - *(Optional)* Code or scripts used, only if automated tools were applied (e.g., web scraping).
3. In addition to the above deliverables, **students must:**
 - **Import the final dataset into Power BI** and create a dashboard that visualizes and summarizes key aspects of the medicines (e.g., distribution by active ingredient, therapeutic indication frequency, side effects prevalence);
 - **Build a relational database** (e.g., using **ORACLE**, MySQL, PostgreSQL, or SQLite) by appropriately normalizing the data;
 - **Create and execute at least +5 SQL** queries aimed at producing analytical outputs (e.g., medicines grouped by equivalence group, most frequent contraindications, medicines by ATC classification).