PROJET UE-3

GUNDUZ Maxime

KAJOKA Felicite

Master 2 Informatique Biomédicale

2024-2025

Conception d'un entrepôt de données standardisé pour l'étude d'un cas clinique

Introduction :

L'hypertension artérielle (HTA) est une maladie chronique qui affecte selon l'OMS 1,28 milliard de personnes dans le monde en 2023, dont 46% ignorent leur condition (référence 1). Cette maladie se caractérise par une pression anormalement élevée exercée sur les parois des vaisseaux sanguins, ce qui les endommage avec le temps et entraîne des complications graves telles que des accidents vasculaires cérébraux (AVC), l'infarctus du myocarde ou l'insuffisance cardiaque.

Bien que l'HTA puisse être un diagnostic primaire dans certains contextes, elle est souvent sous-représentée dans les bases de données cliniques centrées sur des patients gravement malades, comme MIMIC-III, où les diagnostics prioritaires sont ceux des états aigus. Cette situation complique l'identification directe des stratégies thérapeutiques utilisées pour l'HTA et impose une approche indirecte.

Dans ce contexte en prenant en compte les données provenant de "MIMIC-III Clinical Database Demo" sur la dernière version 1.4, notre objectif est de concevoir un entrepôt de données standardisé qui puisse permettre d'identifier les médicaments les plus prescrits pour les patients souffrant d'HTA.

Matériels et Méthodes :

Selection des fichier CSV :

Pour commencer, il a été indispensable d'analyser les données sous forme de tableaux CSV proposées par la "MIMIC-III Clinical Database Demo" afin d'identifier les fichiers les plus pertinents pour notre étude. Ensuite, nous avons procédé à une sélection des colonnes les plus importantes dans chacun des fichiers retenus. Dans le cadre de notre étude, les tableaux CSV seront désignés par le terme "Table"

Standardisation des colonnes au format OMOP :

Après notre sélection des fichiers CSV avec les colonnes d'intérêt, nous avons utilisé les logiciels "WhiteRabbit" et "Rabbit in a Hat" pour standardiser les noms des colonnes, en passant du format initial "MIMIC-III" au format standardisé OMOP CDM version 5.4.

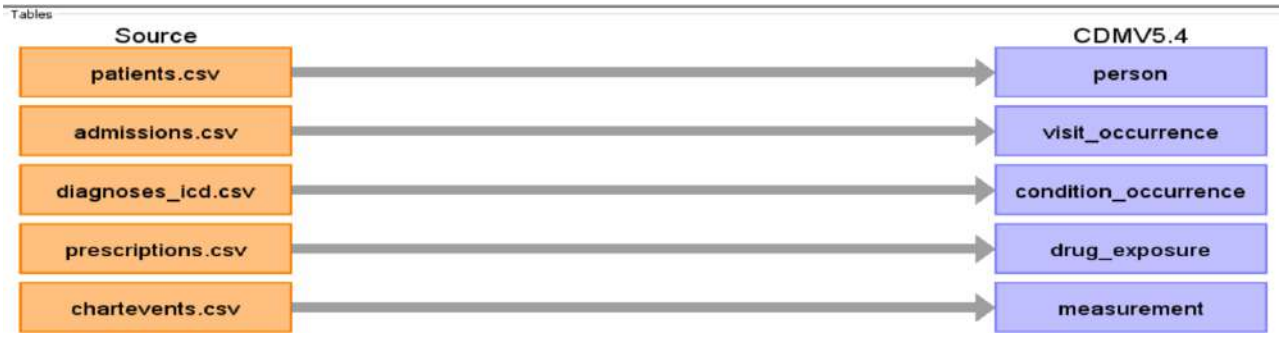


Figure illustrant la correspondance OMOP des noms des Table

Cette représentation illustre nos différents mappages des noms des tables "MIMIC-III" vers leur équivalent au format OMOP.

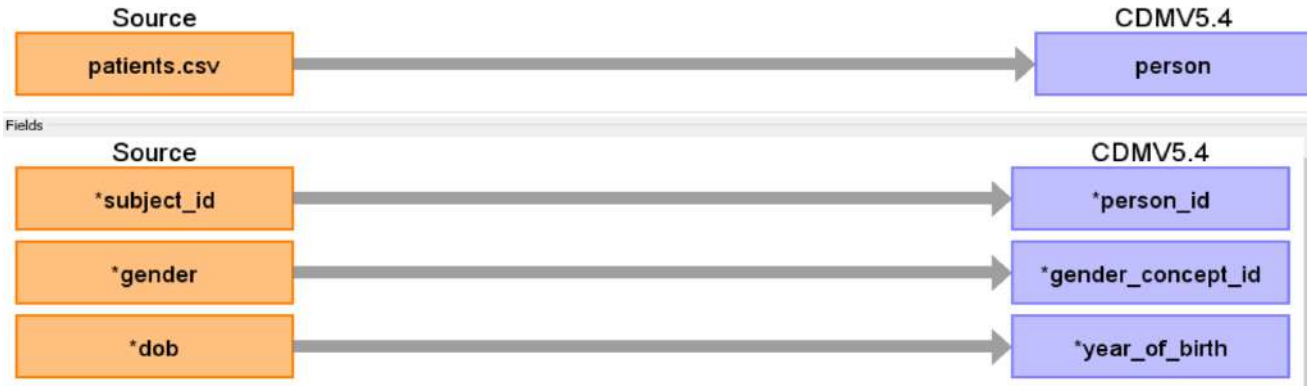


Figure illustrant la correspondance OMOP des noms des colonnes de Patient

Notre entrepôt nécessite une table dédiée à la représentation des patients. Nous avons choisi de définir chaque patient de manière unique grâce à un identifiant, **"person_id"**, qui garantit son unicité au sein de l'entrepôt. Chaque patient est également caractérisé par des attributs tels que son sexe et sa date de naissance.

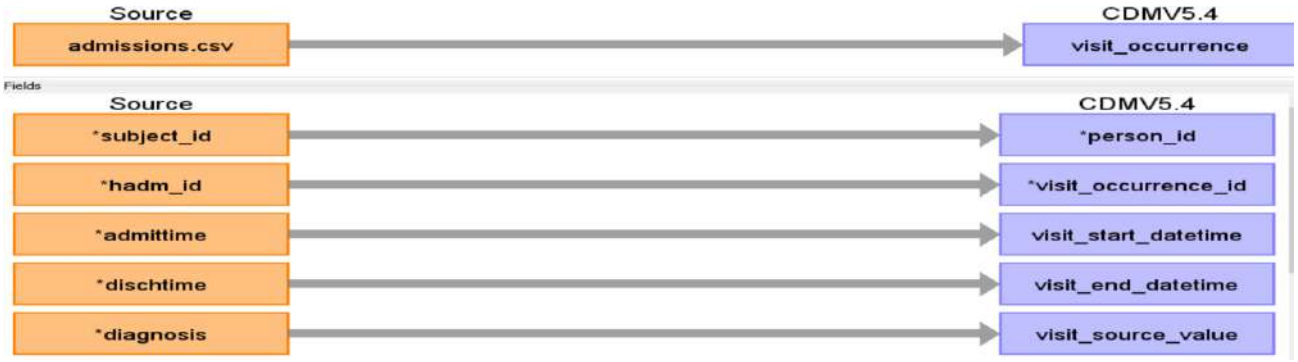


Figure illustrant la correspondance OMOP des noms des colonnes des admissions

Nous avons également besoin d'une table qui représente les différentes admissions des patients. Cette table devra avoir un identifiant comme "hadm_id" qui sera "visit_occurrence_id" afin d'identifier une hospitalisation unique d'un patient.

Cette table va nous permettre d'identifier les patients qui sont hospitalisés à travers "person_id" et à identifier le motif d'hospitalisation à travers les diagnostics "visit_source_value".

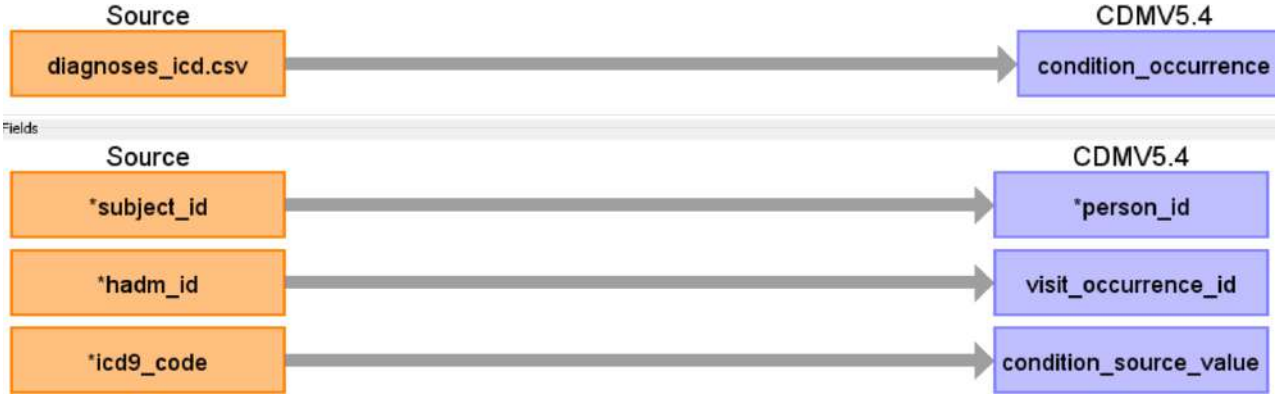


Figure illustrant la correspondance OMOP des noms des colonnes des diagnostic

Cette table permet d'identifier les diagnostics (codés selon la classification ICD) associés aux hospitalisations des patients.

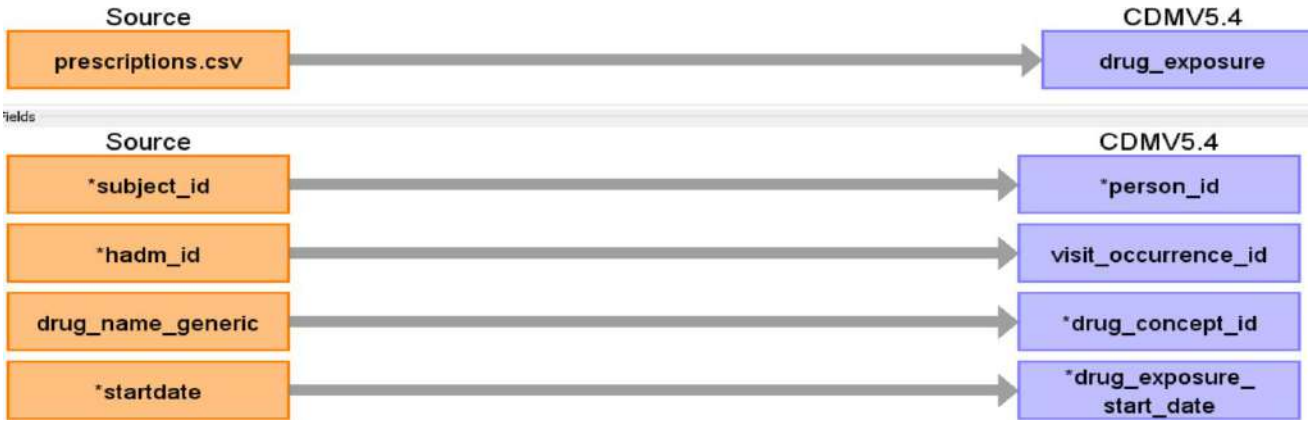


Figure illustrant la correspondance OMOP des noms des colonnes des prescriptions

Nous avons besoin d'une table représentant les différentes prescriptions de médicaments sur différentes hospitalisations. Pour cela, nous avons décidé de garder les identifiants des patients ainsi que des hospitalisations. Nous avons également besoin de savoir le nom du médicament.

Nous avons choisi d'utiliser **"drug_name_generic"** à la place de **"drug"**, car les noms génériques permettent une meilleure interopérabilité et sont couramment utilisés dans les bases de données cliniques. Cette colonne a été associée à **"drug_concept_id"**, qui représente le concept standardisé d'un médicament dans le modèle OMOP. Par ailleurs, nous avons décidé de conserver la colonne **"startdate"**, mappée à **"drug_exposure_start_date"** dans OMOP, pour représenter la date de début de l'exposition ou de la prescription du médicament.

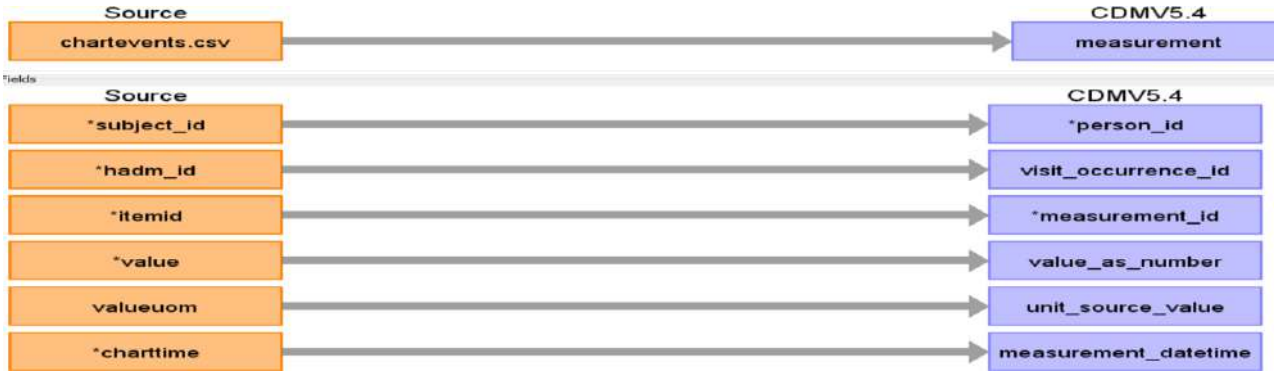


Figure illustrant la correspondance OMOP des noms des colonnes des mesures

Nous avons une table représentant les différentes mesures réalisées de l'hypertension artérielle sur différents patients par leur hospitalisation. Nous avons les identifiants des personnes, et des hospitalisations, ainsi que l'identifiant unique des mesures à travers "itemid" ("measurement_id" dans OMOP). Nous avons également des colonnes sur la représentation des mesures, comme la valeur de la mesure qui est une valeur (value_as_number) et son unité (unit_source_value), ainsi que la date prescrite de la mesure (measurement_datetime).

Standardisation des concepts au format OMOP :

Après la standardisation OMOP des colonnes, il a fallu standardiser les concepts au format OMOP. Pour cela, il a fallu identifier les différents concepts et faire un choix.

Nous avons identifié, à partir d'une analyse de fichiers CSV, des patients admis souffrant de maladies dont l'HTA est un facteur de risque, et nous avons sélectionné les patients souffrant de ces maladies et qui souffrent probablement d'HTA :

- STROKE/TIA : Accident vasculaire cérébral (AVC)
- INFERIOR MYOCARDIAL INFARTUS\CATH : Infarctus inférieur du myocarde
- CONGESTIVE HEART FAILURE : Insuffisance cardiaque congestive

Nous n'avons pas de patients qui sont hospitalisés pour l'HTA du fait qu'aucun patient ne soit réellement hospitalisé pour cette raison.

On a pu déduire les patients souffrant de ces maladies et obtenir la liste de leurs médicaments face à l'HTA, qui sont : 'Metoprolol', 'Lisinopril', 'Hydralazine', 'Hydralazine HCl', 'Captopril', 'Metoprolol XL', 'Atenolol', 'Losartan Potassium', 'Diltiazem', 'Spironolactone', 'Labetalol'.

Nous avons les différents concepts des maladies et des médicaments qu'on va devoir standardiser.

De là, nous pouvons avec "Athena" obtenir tous les "CONCEPT ID" sur la ligne "Non-standard to Standard map (OMOP)" pour toutes les colonnes représentant un concept en recherchant les différents concepts. Nous avons eu les concepts suivants (si le champ est vide, cela signifie que la colonne n'est pas un concept) :

Colonnes OMOP CDM version 5.4	***CONCEPT ID ("Non-standard to Standard map (OMOP)"***
person_id	
gender_concept_id	- FEMALE => Concept id= 8532 - MALE => Concept id = 8507
year_of_birth	

Table PERSON Standardisée au format OMOP

Colonnes OMOP CDM version 5.4	***CONCEPT ID ("Non-standard to Standard map (OMOP)"***
person_id	
visit_occurrence_id	
visit_start_datetime	
visit_end_datetime	
visit_source_value	- Congestive heart failure => CONCEPT ID = 319835 - Old inferior myocardial infarction => CONCEPT ID = 4121467 - Stroke/TIA => CONCEPT ID = 4053371

Table VISIT_OCCURRENCE Standardisée au format OMOP

Colonnes OMOP CDM version 5.4	***CONCEPT ID ("Non-standard to Standard map (OMOP)"***
person_id	
visit_occurrence_id	
condition_source_value	

Table CONDITION_OCCURRENCE Standardisée au format OMOP

Colonnes OMOP CDM version 5.4	***CONCEPT ID ("Non-standard to Standard map (OMOP)"***
person_id	
visit_occurrence_id	
drug_concept_id	- Metoprolol : ndc = 51079080120 => CONCEPT ID = 40167218 - Lisinopril : ndc = 310013039 => CONCEPT ID = 19003830 - Losartan potassium : ndc = 71610037045 => CONCEPT ID = 40185280 - Spironolactone : ndc = 51079010320 => CONCEPT ID = 19079658 - Hydralazine : ndc = 63323061401 => CONCEPT ID = 40174776 - Labetalol : ndc = 182820289 => CONCEPT ID = 40169683 - Captopril : ndc = 904504561 => CONCEPT ID = 19074672 - Atenolol : ndc = 51079075920 => CONCEPT ID = 19018811 - Diltiazem : ndc = 51079074520 => CONCEPT ID = 1328689 - Hydralazine HCl : ndc = 517090125 => CONCEPT ID = 40174776 - Metoprolol XL : ndc = 186109039 => CONCEPT ID = 40166831
drug_exposure_start_date	

Table DRUG_EXPOSURE Standardisée au format OMOP

Pour la table "DRUG_EXPOSURE", il a fallu identifier à partir de "PRESCRIPTIONS.csv" les différents codes "ndc" afin de rechercher les codes "ndc" sur "Athena" pour retrouver les CONCEPT ID.

Colonnes OMOP CDM version 5.4	***CONCEPT ID ("Non-standard to Standard map (OMOP)"***
person_id	
visit_occurrence_id	
measurement_id	
value_as_number	
unit_concept_id	- mm [Hg] : CONCEPT ID = 8876
measurement_datetime	

Table MEASUREMENT Standardisée au format OMOP

Une fois que nous avons obtenu le plan pour standardiser nos tables au format OMOP à partir de nos tables d'origine, nous avons réalisé des traitements sur les tables des fichiers CSV que nous avons pris comme base.

Transformation des fichiers CSV avec R :

Nous avons dans un premier temps importé nos fichiers CSV d'origine dans Rstudio avec R sous la forme de dataframes, puis nous avons, pour tous les dataframes, supprimé les colonnes non pertinentes. Après cela, nous avons commencé à traiter table par table tous nos dataframes.

Nous ferons le renommage des colonnes à la fin pour la standardisation des colonnes.

- Table Visit_Occurence :

Nous avons extrait tous les patients dont la colonne "diagnosis" correspond à l'une de nos 3 maladies, et avons modifié les résultats de la colonne "diagnosis" par leur "CONCEPT ID" qui leur correspond. Ceci nous permet d'avoir les patients ("subject_id") souffrant d'au moins une de ces maladies.

- Table Person :

À partir de notre travail fait préalablement, on a pu filtrer tous les patients ("subject_id") souffrant d'au moins une de ces maladies et les extraire pour ne garder que ceux-ci.

On a mis à jour les valeurs de la colonne "gender_concept_id" en fonction de leur "CONCEPT ID" et on a extrait uniquement l'année pour la colonne "dob".

- Table Condition_Occurence :

Nous avons uniquement extrait les patients qui nous intéressent.

- Table Drug_Expore :

On a extrait uniquement les patients qui nous intéressent ainsi que tous les médicaments qui nous intéressent, et nous avons, pour la colonne "drug_name_generic", remplacé les noms des médicaments par la valeur de "Concept ID" appropriée. Nous avons, pour la colonne "startdate", éliminé l'heure.

- Table Measurement :

Nous avons filtré sur la colonne "valueuom" pour ne garder que les lignes avec la valeur "mmHg", puis nous avons filtré nos patients pour ne garder que ceux qui nous intéressent. Nous avons par la suite remplacé la valeur "mmHg" par son CONCEPT ID.

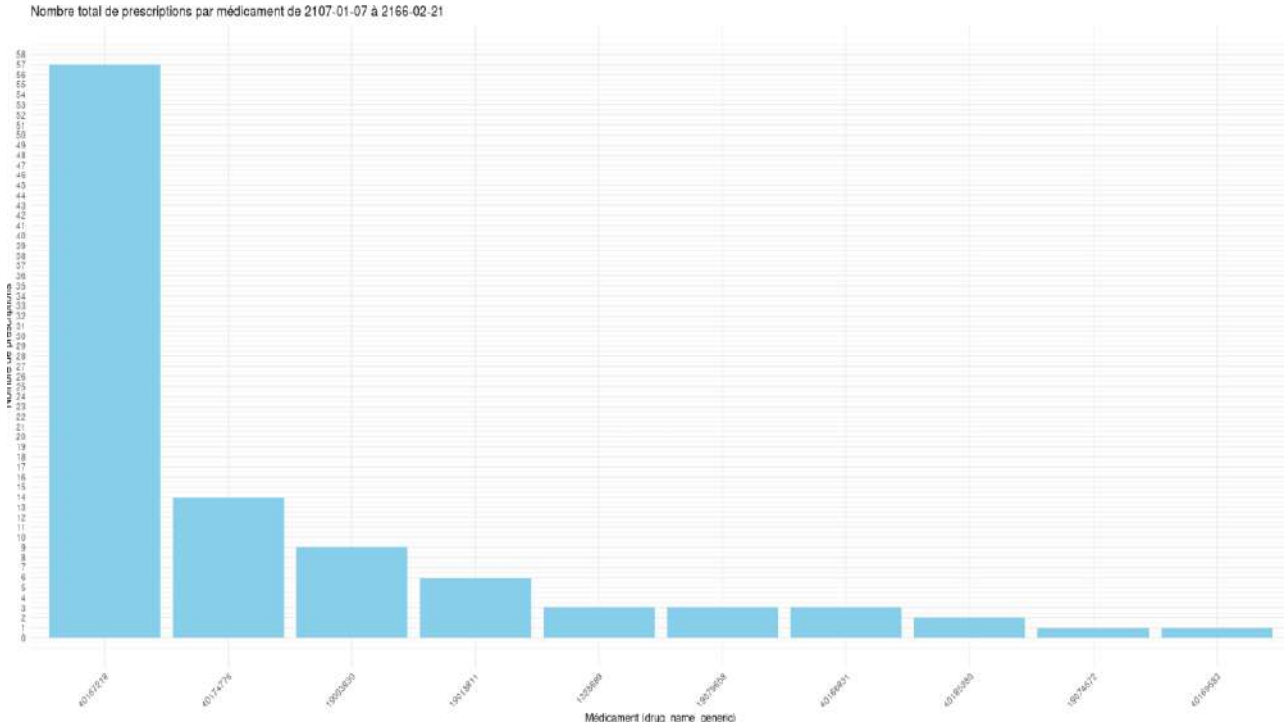
Nous avons ensuite fait le renommage des colonnes sur tous les dataframes pour correspondre au format OMOP, puis nous avons exporté nos dataframes sous la forme de fichiers CSV afin de les utiliser pour faire des requêtes logiques.

Nous avons également réalisé un histogramme représentant le nombre de prescriptions de nos médicaments sur l'ensemble de nos données.

Enfin, nous avons importé nos nouveaux CSV sur le site "csvfiddle.io", qui nous a permis de faire des requêtes SQL directement sur les CSV.

Résultat :

Nous avons pu exploiter les données de notre entrepôt pour répondre à notre contexte. Nous avons représenté sous la forme d'un histogramme tous nos médicaments pour visualiser le nombre de prescriptions par médicament, et nous avons eu les résultats suivants :



Nous pouvons voir que le médicament "Metoprolol" (40167218) est un médicament extrêmement prescrit (57 fois), tandis que le "Hydralazine HCl" (40174776) a été prescrit 14 fois.

Pour approfondir notre étude, on peut s'intéresser au nombre de médicaments prescrits par patient, et obtenir avec la requête SQL suivante :

SELECT person_id AS patient, drug_concept_id AS médicament, COUNT(*) AS prescription FROM DRUG_EXPOSURE GROUP BY person_id, drug_concept_id ORDER BY person_id, drug_concept_id;

Le resultat suivant :

	patient	médicament	prescription
1	10026	19003830	2
2	10026	40167218	3
3	10026	40174776	2
4	10088	19003830	4
5	10088	19018811	1
6	10088	19074672	1
7	10088	40166831	3
8	10088	40167218	23
9	10088	40174776	4
10	10111	1328689	1
11	10111	40167218	11
12	10111	40174776	7
13	10111	40185280	2
14	10124	1328689	2
15	10124	19003830	3
16	10124	19018811	5
17	10124	19079658	3
18	10124	40167218	20
19	42075	40174776	1
20	43870	40169683	1

De là, nous pouvons constater que le patient “10088” s’est fait prescrire 23 fois la “Metoprolol" (40167218) sur l’ensemble de ses consultations.

Conclusion et Discussion :

Nous avons analysé les médicaments les plus prescrits pour les patients souffrant d'hypertension artérielle (HTA). Pour cela, nous avons étudié les tables issues de la "MIMIC-III Clinical Database Demo", afin de sélectionner celles les plus pertinentes pour notre contexte.

Dans un premier temps, nous avons filtré uniquement les colonnes d'intérêt et renommé ces dernières au format OMOP. Ensuite, nous avons identifié les différents concepts pour trouver les Concept ID adéquats.

Nous avons ainsi transformé nos tables CSV avec R pour obtenir un mini entrepôt de données interopérable, exploitable dans notre contexte et par d'autres instituts.

Cependant, nous avons remarqué que les données issues de la "MIMIC-III Clinical Database Demo" sont probablement fictives. En effet, nous avons constaté des dates de naissance provenant de la fin du 21ème siècle (par exemple, le patient 43870 est né en 2097). Pour l'avenir de ce projet, il serait plus pertinent de prendre en compte un plus grand nombre de patients et de médicaments afin d'obtenir des résultats plus fidèles aux données de la "MIMIC-III Clinical Database Demo".

Référence :

1 : https://www.who.int/fr/news-room/fact-sheets/detail/hypertension#:~:text=Principaux%20faits,d'hypertension%20l'ignorent.