# Intégration et utilisation secondaire des données de santé hospitalières hétérogènes : des usages locaux à l'analyse fédérée

Romain Griffier

université de BORDEAUX

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

# DOCTEUR
# DE L'UNIVERSITÉ DE BORDEAUX

ECOLE DOCTORALE EDSP2

SOCIÉTÉS, POLITIQUE, SANTÉ PUBLIQUE

Santé publique Option Informatique et Santé

Par **Romain GRIFFIER**

## Intégration et utilisation secondaire des données de santé hospitalières hétérogènes : des usages locaux à l'analyse fédérée

Sous la direction de :
Co-directeur : **Vianney JOUHET**
Co-directrice : **Fleur MOUGIN**

Soutenue le 18 décembre 2024

Membres du jury :

| | | | |
|---|---|---|---|
| Pr Leslie GRAMMATICO-GUILLON | PU-PH | CHRU de Tours | Rapportrice |
| Dr Bastien RANCE | MCU-PH | Hôpital Européen Georges Pompidou AP-HP | Rapporteur |
| Pr Rodolphe THIEBAUT | PU-PH | Université de Bordeaux | Examinateur |
| Pr Pascal STACCINI | PU-PH | Université Côte d'Azur | Examinateur |
| Dr Vianney JOUHET | PH | CHU de Bordeaux | Co-directeur |
| Pr Fleur MOUGIN | PU | Université de Bordeaux | Co-directrice |

# Table des matières

# Table des figures

# Liste des tableaux

# Glossaire

**AAP** : Appel À Projets
**API** : *Application Programming Interface*
**ARS-NA** : Agence Régionale de Santé Nouvelle-Aquitaine
**BPI** : Banque Publique d'Investissement
**CDM** : *Common Data Model*
**CDW** : *Clinical Data Warehouse*
**CépiDC** : Centre d'épidémiologie sur les causes médicales de DéCès
**CERS** : Centre Éthique et Recherche en Santé
**CESREES** : Comité Éthique et Scientifique pour les Recherches, les Études et les Évaluations dans le domaine de la Santé
**CHU** : Centre Hospitalier Universitaire
**CNAM** : Caisse Nationale d'Assurance Maladie
**CNIL** : Commission Nationale de l'Informatique et des Libertés
**CNSA** : Caisse Nationale de Solidarité pour l'Autonomie
**CRC** : *Clinical Research Chart* (*Data Repository Cell*)
**CSE** : Comité Scientifique et Éthique
**DARWIN** : *Data Analysis and Real World Interrogation Network*
**DARWIN-CC** : *DARWIN Coordination Center*
**DDL** : *Data Definition Language*
**DGOS** : Direction Générale de l'Offre de Soins
**DICOM** : *Digital Imaging and COmmunications in Medicine*
**DNIM** : Direction du Numérique
**DMP** : Dossier Médical Partagé
**DP** : *Data Partner*
**DPI** : Dossier Patient Informatisé
**DPO** : *Data Protection Officer* (Délégué à la protection des données personnelles)
**DRCI** : Direction de la Recherche Clinique et de l'Innovation
**eCRF** : *electronical Case Report Form*
**EDS** : Entrepôt de Données de Santé
**EEDS** : Espace Européen des Données de Santé
**EHDEN** : *European Health Data & Evidence Network*
**EHRs** : *Electronic Health Records*
**EMA** : *European Medicines Agency*
**ESEA** : e-Santé En Action
**ETL** : *Extract, Transform and Load*
**FDA** : *Food and Drug Administration*
**FHIR** : *Fast Healthcare Interoperability Resources*
**GCS** : Groupement de Coopération Sanitaire
**GHT** : Groupement Hospitalier de Territoire
**GIP** : Groupement d'Intérêt Public

**GT** : Groupe de Travail
**GUI** : *Graphic User Interface*
**HDH** : *Health Data Hub*
**HMA** : *Heads of Medicines Agencies*
**i2b2** : *Informatics for Integrating Biology & the Bedside*
**IA** : Intelligence Artificielle
**IAM** : Informatique et Archivistique Médicales (Unité)
**IEP** : Identifiant d'Épisode du Patient
**IHU** : Institut Hospitalo-Universitaire
**IMI 2** : *Innovative Medicines Initiative 2*
**Insee** : Institut National de la Statistique et des Études Économiques
**IoT** : *Internet of Things* (internet des objets)
**IPP** : Identifiant Permanent du Patient
**IRIS** : Îlots Regroupés pour l'Information Statistique
**LDAP** : *Lightweight Directory Access Protocol*
**LIL** : Loi Informatique et Libertés
**LIMS** : *Laboratory Information Management System*
**LOINC** : *Logical Observation Identifiers Names & Codes*
**MDPH** : Maison Départementale pour les Personnes Handicapées
**ML** : *Machine Learning*
**NABM** : Nomenclature des Actes de Biologie Médicale
**NEL** : *Named-Entity-Linking*
**NER** : *Named-Entity Recognition*
**NDA** : Numéro de Dossier Administratif
**NIP** : Numéro Identifiant Patient
**NLP** : *Natural Language Processing*
**OHDSI** : *Observational Health Data Sciences and Informatics*
**OIDC** : *OpenID Connect*
**OLAP** : *Online Analytical Processing*
**OLTP** : *Online Transaction Processing*
**OMOP-CDM** : *Observational Medical Outcomes Partnership - Common Data Model*
**ONT** : *Ontology Management Cell*
**P4DP** : *Platform for Data in Primary care*
**PACS** : *Picture Archiving and Communication System*
**PMSI** : Programme de Médicalisation des Systèmes d'Information
**PROs** : *Patient-Reported Outcomes*
**RDBMS** : *Relational DataBase Management System*
**RGPD** : Règlement Général de Protection des Données
**RNIPH** : Recherche N'Impliquant pas la Personne Humaine
**RSSI** : Responsable de la Sécurité des Systèmes d'Information
**RWD** : *Real World Data*
**RWE** : *Real World Evidence*

**SHRINE** : *Shared Health Research Informatics NEtwork*
**SI-DEP** : Système d'Information de DEpistage Populationnel
**SIG** : Système d'Information Géographique
**SIH** : Système d'Information Hospitalier
**SNDS** : Système National des Données de Santé
**SNIIRAM** : Système National d'Information Inter-Régimes de l'Assurance Maladie
**SQL** : *Structured Query Language*
**TAL** : Traitement Automatique des Langues
**UE** : Union Européenne

# Contributions scientifiques

Les travaux présentés dans ce manuscrit ont fait l'objet de différentes contributions scientifiques.

Le travail décrit dans le chapitre 2 a fait l'objet d'une **communication orale dans une conférence internationale avec proceedings** :

> **Griffier, R**, Cossin, S, Konschelle, F, Mougin, F, Jouhet, V, Data Element Mapping in the Data Privacy Era. Studies in Health Technology and Informatics. 2022 May 25 ; 294:332-6. DOI : 10.3233/SHTI220469

Le travail présenté dans le chapitre 3 est en cours de reviewing dans une **revue internationale avec comité de lecture**. Le preprint de l'article est disponible :

> **Griffier, R**, Mougin, F, Jouhet, V, Integrating healthcare data in an i2b2 model persisted through Elasticsearch. 2024 Aug 25. DOI : 10.2196/preprints.65753

De manière plus large, la contribution de l'environnement EDS du CHU de Bordeaux dans des réseaux fédérés décrit dans le chapitre 4 a fait l'objet de plusieurs publications scientifiques dans des **revues internationales avec comité de lecture** :

— Études réalisées au sein du réseau 4CE :

> Brat, GA, Weber, GM, Gehlenborg, N, Avillach, P, Palmer, NP, Chiovato, L, International electronic health record-derived COVID-19 clinical course profiles : the 4CE consortium. NPJ digital medicine. 2020 ; 3:109. DOI : 10.1038/s41746-020-00308-0

> Weber, GM, Hong, C, Palmer, NP, Avillach, P, Murphy, SN, Gutiérrez-Sacristán, A, International Comparisons of Harmonized Laboratory Value Trajectories to Predict Severe COVID-19 : Leveraging the 4CE Collaborative Across 342 Hospitals and 6 Countries : A Retrospective Cohort Study. medRxiv : The Preprint Server for Health Sciences. 2021 Feb 5 :2020.12.16.20247684. DOI : 10.1101/2020.12.16.20247684

> Moal, B, Orieux, A, Ferté, T, Neuraz, A, Brat, GA, Avillach, P, Acute respiratory distress syndrome after SARS-CoV-2 infection on young adult population : International observational federated study based on electronic health records through the 4CE consortium. PloS One. 2023 ; 18:e0266985. DOI : 10.1371/journal.pone.0266985

— Étude réalisée au sein du réseau DARWIN-EU :

Du, M, Dernie, F, Català, M, Delmestri, A, Man, WY, Brash, JT, Treatment of systemic lupus erythematosus : Analysis of treatment patterns in adult and paediatric patients across four European countries. European Journal of Internal Medicine. 2024 Aug 11 :S0953–6205(24)00344-3. DOI : 10.1016/j.ejim.2024.08.008

Dernie, F, Corby, G, Robinson, A, Bezer, J, Mercade-Besora, N, **Griffier, R,** Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU). Pharmacoepidemiology and Drug Safety. 2024 Nov 12 ; 33:e70042. DOI : 10.1002/pds.70042

# CHAPITRE 1

# Introduction et contexte

# Sommaire

Ces dernières années ont été marquées par une augmentation sans précédent de la disponibilité des données numériques. Dans le domaine de la santé, la prolifération des technologies de l'information, dont l'un des objectifs est de faciliter la communication entre les professionnels de santé, génère de très grandes quantités de données. Les données de santé représentent environ 30% des données disponibles au format numérique dans le monde [8]. Ces données massives, qui sont généralement produites pour des objectifs précis, peuvent également être utilisées pour d'autres usages dans le cadre de l'utilisation secondaire des données de santé.

## 1.1 Utilisation secondaire des données de santé

La définition d'utilisation secondaire des données de santé n'est pas consensuelle [9]. De manière simple, elle désigne l'utilisation des données de santé à des fins autres que celles pour lesquelles les données ont été initialement collectées. Cette utilisation secondaire couvre un large éventail de domaines [10], notamment le suivi de l'activité des établissements, le phénotypage [11, 12], la recherche [5], l'enrichissement des données des registres épidémiologiques [13], la qualité et la sécurité des soins [14] ou la surveillance épidémiologique [15, 16].

Dans cette section, nous introduisons la notion de « données de vie réelle » et nous présentons le panorama des données de vie réelle disponible en France, formant le « patrimoine des données de santé ».

### 1.1.1 Données de vie réelle

Dans le cadre de l'utilisation secondaire des données de santé, les données utilisées sont qualifiées de « données de vie réelle » (*Real World Data* ou RWD en anglais), par

opposition aux données collectées dans un cadre contrôlé au cours de projets de recherche (essais cliniques). La *Food and Drug Administration* (FDA) a défini les RWD comme des « *données relatives à l'état de santé du patient et/ou à la prestation de soins de santé collectées en routine à partir de diverses sources. Les données issues des établissements de santé, les données relatives aux remboursements des frais médicaux, les données provenant de registres de pratiques ou de maladies et les données recueillies à partir d'autres sources (telles que les technologies numériques de santé) qui peuvent fournir des informations sur l'état de santé sont des exemples de RWD* » [17].

Dans [18], les auteurs listent les différentes sources de RWD (Figure 1.1). En plus des données issues des établissements de santé et des données de remboursement des prestations de soins, sont considérées comme RWD les données qui, par nature ou par destination, peuvent être relatives à l'état de santé d'un individu. On peut citer par exemple les données issues des applications de e-santé, les données déclarées par les patients (*Patient-Reported Outcomes* ou PROs), les données environnementales ou encore les données issues des réseaux sociaux.

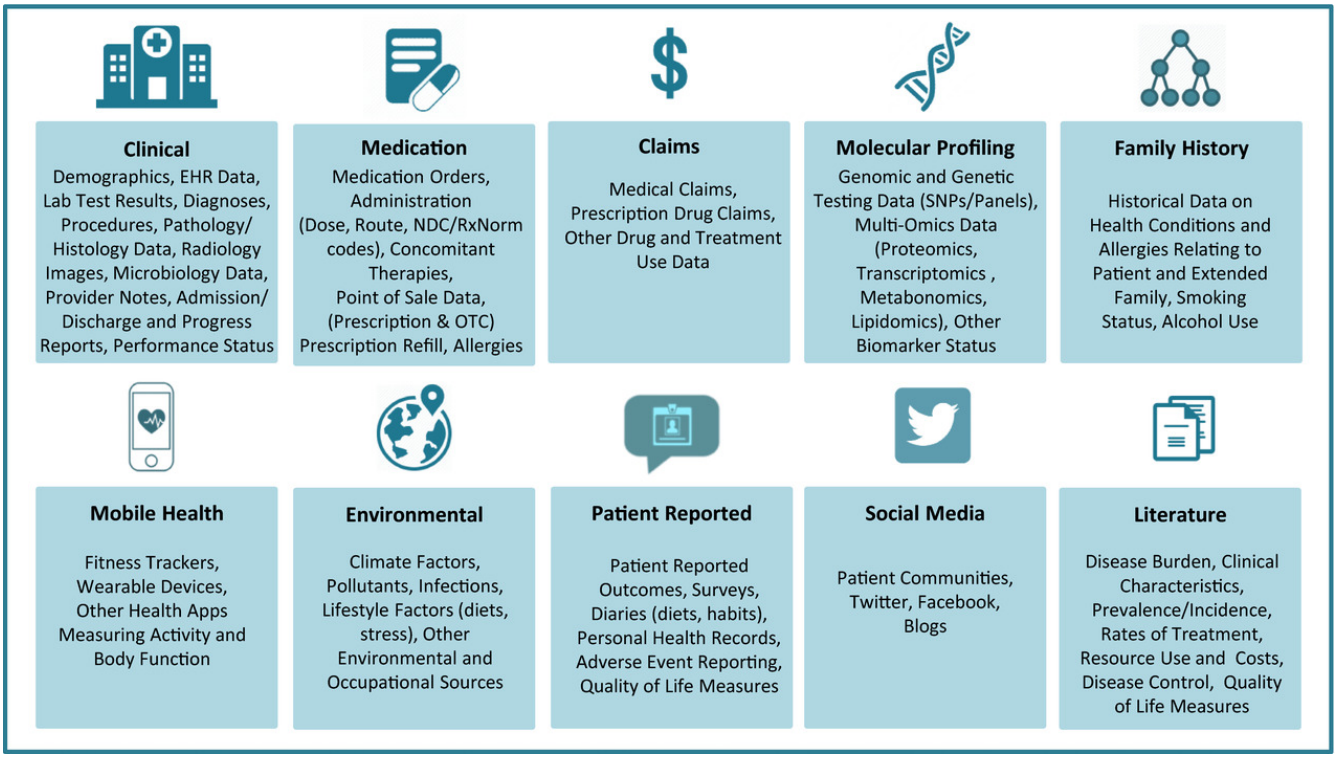


FIGURE 1.1 – Différents types de données de vie réelle

Extrait de : [18]

Les sources de données qui sont présentées dans [18] sont similaires à celles constituant le « patrimoine des données de santé en France », détaillé dans le rapport de préfiguration de la Plateforme des Données de Santé [19].

### 1.1.2 « Patrimoine des données de santé » en France

En 2018, la mission de préfiguration de la Plateforme des Données de Santé [19], également appelée *Health Data Hub* (HDH), identifie les sources de données de santé qui constituent le « patrimoine des données de santé » en France. On peut regrouper ces sources en cinq grandes catégories :

1. Les données collectées à un niveau régional ou national ;
2. Les données de recherche et de surveillance ;
3. Les données produites par les patients ;
4. Les données produites dans un contexte de médecine de ville ;
5. Les données produites dans un établissement de santé ;

#### *1.1.2.1 Données collectées au niveau national / régional*

**Système National des Données de Santé**

Le Système National des Données de Santé (SNDS) a été créé par la loi de modernisation du système de santé du 26 janvier 2016 [20], comme un enrichissement de la base de données du Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM, [21]).

Ainsi, le SNDS contient :

— Les données du SNIIRAM (données de remboursement de l'assurance maladie, tous régimes confondus) ;
— Les données issues des établissements de santé, via le Programme de Médicalisation des Systèmes d'Information (PMSI) ;
— Les causes médicales de décès, recensées par le Centre d'épidémiologie sur les causes médicales de DéCès (CépiDC) ;
— Les données médico-sociales des Maisons Départementales pour les Personnes Handicapées (MDPH) et de la Caisse Nationale de Solidarité pour l'Autonomie (CNSA) ;
— Un échantillon des données d'organismes d'assurance maladie complémentaire ;
— Les données relatives à la COVID-19, avec les bases de données « Vaccin Covid » et « SI-DEP » (Système d'Information de DEpistage Populationnel).

**Dossier Pharmaceutique**

Le dossier pharmaceutique a été créé par la loi du 30 janvier 2007 relative à l'organisation de certaines professions de santé du code de la santé publique [22]. Il recense les médicaments délivrés en ville pour les bénéficiaires de l'assurance maladie ayant consenti à son ouverture. De manière plus précise, il contient :

— L'antériorité sur 3 ans des médicaments courants ;
— L'antériorité sur 5 ans des médicaments biologiques [1] ;
— L'antériorité sur 23 ans des vaccins.

#### Dossier Médical Partagé

Le Dossier Médical Partagé (DMP) est prévu par la loi n°2004-810 du 13 août 2004 relative à l'assurance maladie [23]. La Caisse Nationale d'Assurance Maladie (CNAM) est en charge de la gouvernance et du déploiement du DMP. Le DMP correspond à un dossier médical disponible au niveau national. Son objectif est de regrouper l'ensemble des données médicales concernant un patient, notamment afin de favoriser la continuité des soins entre la ville et l'hôpital. Les différents producteurs de données de santé (médecin de ville, pharmacien, établissement de santé, etc.) sont en charge de l'alimentation du DMP.

#### Bouquet de services numériques régionaux

Au niveau régional, de nombreux « bouquets de services numériques » sont en cours de déploiement pour permettre aux professionnels d'accéder, depuis un point d'entrée unique, à des outils numériques régionaux partagés.

Par exemple en Nouvelle-Aquitaine, le projet « Krypton » [2], géré par le Groupement d'Intérêt Public (GIP) e-Santé en Action (ESEA), en lien avec l'Agence Régionale de Santé Nouvelle-Aquitaine (ARS-NA), permet l'échange et le partage sécurisé d'examens d'imagerie réalisés entre les structures de santé publiques et privées adhérentes du territoire.

### *1.1.2.2 Données de recherche et de surveillance*

#### Systèmes-fils du SNDS

Les systèmes-fils correspondent à des bases de données appariées à un échantillon du SNDS. C'est le HDH, via le Comité Éthique et Scientifique pour les Recherches, les Études et les Évaluations dans le domaine de la Santé (CESREES) et la Commission Nationale de l'Informatique et des Libertés (CNIL), qui instruit les demandes de création des systèmes-fils. En 2018, il existait plus de 200 systèmes-fils [19].

#### Données issues des grandes cohortes nationales

Les études de cohorte sont des études épidémiologiques observationnelles incluant un suivi longitudinal d'un groupe homogène de patients. L'objectif est de mesurer la survenue d'un évènement de santé dans le temps (mesure des cas incidents) et

1. Un médicament biologique correspond à un médicament synthétisé ou extrait d'une source biologique (par opposition aux médicaments issus de la chimie de synthèse).
2. `https://www.esea-na.fr/gip/krypton`

d'identifier les caractéristiques qui diffèrent entre le groupe d'individus ayant présenté l'évènement et le groupe d'individus n'ayant pas présenté l'évènement. En 2019 en France, on comptait plus de 250 cohortes [24].

Au niveau national, on peut par exemple citer la cohorte CONSTANCES. Il s'agit d'une « cohorte épidémiologique généraliste constituée d'un échantillon représentatif de 200 000 adultes âgés de 18 à 69 ans à l'inclusion, consultant des centres d'examens de santé de la sécurité sociale » [25]. L'objectif principal de la cohorte CONSTANCES est d'étudier les déterminants professionnels et sociaux de santé.

**Données des registres épidémiologiques et des registres de pratique**

Un registre correspond à « un recueil continu et exhaustif de données nominatives intéressant un ou plusieurs événements de santé dans une population géographiquement définie, à des fins de recherche et de santé publique, par une équipe ayant les compétences appropriées » [26].

Les données des registres sont des données d'une grande qualité, consolidées, multisources, qui se veulent exhaustives à l'échelle d'un territoire donné [27]. Parmi les registres épidémiologiques, on peut citer en particulier les registres des cancers [28], qui existent depuis les années 1970, ou encore les registres d'anomalies congénitales [29].

#### *1.1.2.3 Données produites par les patients*

Les patients sont également des producteurs de données. On peut en particulier citer les données issues des dispositifs médicaux, des objets connectés (internet des objets - *Internet of Things* ou IoT [30] en anglais) et les PROs [31]. La FDA définit les PROs comme « toute mesure de l'état de santé du patient qui est rapportée directement par le patient, sans interprétation du médecin ou d'une tierce personne » [32]. Ces données sont particulièrement intéressantes puisqu'elles capturent des évènements de santé en dehors du système de soins.

#### *1.1.2.4 Données produites dans un contexte de médecine de ville*

Différents types de données sont couramment produits dans le contexte de la médecine de ville, notamment :

— Les données collectées par les médecins généralistes et spécialistes en ville ;
— Les données des laboratoires de biologie de ville ;
— Les données des laboratoires de cytologie et d'anatomie pathologiques ;
— Les données d'imagerie de ville, incluant les images et les comptes rendus d'imagerie.

#### *1.1.2.5 Données produites dans un contexte hospitalier et de groupements hospitaliers de territoires*

Dans les établissements de santé, les données numériques relatives aux patients sont organisées au sein des Systèmes d'Information Hospitaliers (SIH). On appelle Dossier Patient Informatisé (DPI) le(s) logiciel(s) utilisé(s) pour tracer les éléments relatifs à la santé et à la prise en charge des patients. Les données des DPI peuvent être produites à l'échelle d'un seul établissement de santé (ex : un centre hospitalier), mais également à l'échelle de plusieurs établissements de santé qui partagent un SIH commun, par exemple à l'échelle d'un Groupement Hospitalier de Territoire (GHT).

Au sein des DPI, les données de santé numériques sont appelées EHRs (*Electronic Health Records* en anglais). Les EHRs sont définis comme « *une collection longitudinale de données numériques de santé concernant un patient [...]* » [33] générés au cours d'une ou plusieurs venues dans n'importe quel contexte de prestation de soins (hospitalisation, consultation, délivrance, etc.).

Les EHRs peuvent être utilisés à des fins dites « secondaires » [34, 35]. Un focus spécifique sur l'utilisation secondaire des données hospitalières est présenté en section 1.2.

L'ensemble de ces données constitue le « patrimoine des données de santé » en France. La dichotomie entre les données issues de la recherche interventionnelle et les RWD correspond notamment à une différence de qualité des données.

### 1.1.3 Qualité des données de vie réelle

Là où les données issues de la recherche sont collectées d'une manière contrôlée dans le cadre d'études interventionnelles, les RWD souffrent d'un manque de qualité très bien décrit dans la littérature [36, 37, 38]. En particulier, la qualité des données au sein des DPI est connue pour ne pas être au niveau des standards habituels de la recherche [39, 40, 41].

De nombreux travaux concernant l'évaluation et l'amélioration de la qualité des RWD existent [42, 43]. Dans [44], l'auteur propose les sept critères suivants pour évaluer la qualité des RWD :

- L'**exactitude des données** (*accuracy*) : la donnée disponible est-elle une mesure valide du phénomène qu'elle cherche à capturer ?
- L'**exhaustivité des données** (*completeness*) : les données disponibles sont-elles suffisamment complètes pour répondre à la question posée ?
- La **cohérence / uniformité des données** (*consistency*) : dans le cas d'études multicentriques, les données collectées sont-elles uniformes / standardisées ?

— La **provenance des données** (*provenance*) : est-on capable d'identifier la provenance de chaque donnée (notamment à des fins d'audit) ?
— La **traçabilité** (*traceability*) : les changements de localisation, de responsable ou de valeur dans les données sont-ils tracés ?
— La **validation des données** (*validation*) : l'outil à l'origine de la mesure est-il validé ? Cela concerne en particulier les données issues de l'IoT ou des PROs ;
— L'**évaluation globale de l'adéquation des données à l'objectif visé** (*fit-for-purpose data*) : les données disponibles permettront-elles de répondre à l'objectif ? Il s'agit ici d'une évaluation globale de la qualité des données, en prenant en compte les aspects de qualité, de pertinence et de disponibilité de ces dernières.

### 1.1.4 Preuves en vie réelle

L'utilisation secondaire des données a pour objectif de produire des **connaissances** à partir des **informations** disponibles au sein de **données** de vie réelle. Il est essentiel de définir ces trois concepts bien qu'il n'en existe pas de définition consensuelle [45]. Dans la suite du document, nous utiliserons les définitions proposées par Blumentritt et Johnston [46] :

— **Données** : une donnée est un élément brut livré en dehors de tout contexte (un nombre, une image, un texte, un son, etc.).

  Par exemple, « 39°C » est une donnée. L'absence de contexte fait que cette donnée seule n'a pas de sens (il pourrait en effet s'agir aussi bien d'une température atmosphérique, d'une température corporelle ou de la température de cuisson d'un aliment) ;

— **Information** : une information est une donnée contextualisée.

  Dans l'exemple, rajouter du contexte à la donnée « 39°C » permet d'obtenir une information : « il fera 39°C demain à Bordeaux » et « le patient a une température de 39°C » sont deux informations différentes que l'on peut obtenir à partir de la même donnée. La contextualisation de la donnée en une information, fiable et vérifiable, est donc un élément crucial vers la connaissance ;

— **Connaissance** : la connaissance correspond à l'appropriation de l'information. C'est de l'accumulation de l'information, fiable et vérifiable, que naît la connaissance.

  Dans l'exemple, en l'absence de connaissance, l'information n'entraînera pas de modification de notre comportement vis à vis de cette dernière. Au cours de notre expérience de vie, nous avons développé des connaissances qui, si nous nous trouvons à Bordeaux demain, nous feront porter des vêtements adaptés à une forte température. Dans un contexte de prise en charge d'un patient, notre expérience médicale nous fera prescrire à un patient un antipyrétique afin de faire baisser sa

température.

Bien que les RWD ne soient pas exemptes de limites, elle revêtent cependant un intérêt important pour la production de connaissances complémentaires à celles produites dans un contexte de recherche interventionnelle [47, 48].

Dans le contexte des RWD, les connaissances générées sont qualifiées de « preuves en vie réelle » (*Real World Evidence* ou RWE en anglais). La FDA définit les RWE comme : « *preuve clinique concernant l'utilisation, les avantages ou les risques potentiels d'un produit médical dérivés de l'analyse des RWD* » [17]. Ainsi, la production de RWE à partir des RWD nécessite de traiter des **données** pour en extraire des **informations** qui seront susceptibles de générer de nouvelles **connaissances**.

## 1.2 Utilisation secondaire des données de santé hospitalières

Les DPI sont aujourd'hui largement déployés dans les établissements de santé [49, 50]. L'objectif principal des DPI est de permettre le partage d'informations entre les professionnels de santé à des fins cliniques et de facturation.

Dans les DPI, le recueil des données est longitudinal, en particulier pour les patients présentant une pathologie chronique et pour lesquels les contacts avec les établissements de santé sont nombreux. Ces données de suivi sont particulièrement intéressantes pour évaluer l'évolution d'une pathologie au cours du temps.

Au sein des DPI, les EHRs correspondent à des données très riches, détaillées, couvrant un large domaine (médecine, chirurgie, obstétrique, pédiatrie, urgences, etc.). Les DPI contiennent des données issues de différentes sources : les différents dossiers de spécialité, l'imagerie (au sein des PACS), la biologie, les données concernant les diagnostics et actes médicaux issues du PMSI, etc. De manière indirecte, les DPI peuvent également contenir des données produites en dehors de l'hôpital (ex : résultats d'examens de biologie réalisés en ville, courrier d'adressage du médecin traitant, etc.) mais cela reste marginal.

### 1.2.1 Données hospitalières et cas d'usage

L'utilisation secondaire des données de santé peut concerner des usages très différents [51, 52, 53, 54], incluant :

— Le **pilotage des établissements**, avec des objectifs tels que l'évaluation de l'activité, l'optimisation de l'utilisation des ressources hospitalières [55] ou encore l'**évaluation des parcours de soins** [56] ;

— Le **support aux activités de vigilance** [57, 58] ou d'**évaluation de la qualité des soins** [59] ;
— Le **support à l'activité de soins** [60], en particulier avec le développement de l'intelligence artificielle [61, 62] et de la médecine de précision [49, 62] ;
— La **recherche**, incluant l'activité de recherche elle-même, au travers de cohortes rétrospectives sur la base des EHRs [63, 64], la **surveillance épidémiologique** (cela a été particulièrement illustré pendant l'épidémie de COVID-19 ; [65, 66]), mais également les activités en support de la recherche, comme par exemple l'aide à l'**identification des patients pour l'inclusion dans des études** [67, 68, 69].

Bien que les données soient collectées en routine dans les DPI, l'utilisation secondaire des données de santé hospitalières présente un certain nombre de limitations et d'obstacles [70, 71, 72, 73]. On distingue en particulier :

1. Des **verrous liés aux données** ;
2. Des **verrous liés aux organisations**.

### 1.2.2 Verrous liés aux données

Les données de santé sont complexes par nature et de par leur volumétrie [74] ; on parle parfois de « *Big Data* », caractérisé par les « 4V » : volume, variété, vélocité et valeur. Outre cette complexité intrinsèque, on relève d'autres verrous liés aux données qui sont présentés dans la suite de cette section.

#### *1.2.2.1 Organisation interne des Dossiers Patients Informatisés*

L'objectif des DPI est de permettre l'échange d'informations entre professionnels de santé autour de la prise en charge d'un patient au cours d'une venue [3] dans un établissement de santé. Ainsi, les DPI sont optimisés pour afficher et saisir de l'information centrée sur le patient et/ou la venue.

Les données des DPI sont stockées dans des bases de données dédiées, la plupart du temps des bases de données relationnelles [75]. Les modèles qui sous-tendent le stockage des données dans les DPI sont conçus pour gérer un maximum de transactions [4] en parallèle : ils ont une organisation dite OLTP (*Online Transaction Processing*) [76].

Dans un objectif d'utilisation secondaire des données de santé, l'interrogation initiale des données de santé est plutôt centrée sur des critères clinico-biologiques. On cherche en effet à identifier des patients qui respectent des critères d'inclusion et de non

3. Une venue correspond à une interaction entre un patient et le système de santé. Dans un contexte hospitaliser, les venues correspondent classiquement à des hospitalisations ou des consultations.
4. Une transaction est un ensemble d'opérations de lecture ou de mise à jour sur une base de données.

inclusion. L'organisation des données qui permet d'interroger efficacement les données dans ce contexte est différente et dite OLAP (*Online Analytical Processing*) [76]. Dans les systèmes OLAP, les données sont organisées selon des schémas en étoile, en flocon ou selon d'autres modèles analytiques [77].

#### *1.2.2.2 Organisation en « silos » des données de santé au sein des Systèmes d'Information Hospitaliers*

Une autre des problématiques connues et documentées dans le contexte de l'utilisation secondaire des données de santé est celle des « silos » de données [78, 79]. Dans les SIH, cela fait référence au fait que les données sont stockées dans des bases de données différentes du fait d'une collecte réalisée avec de multiples outils dédiés à différents métiers.

En effet, les données issues de différents services tels que la radiologie, l'anatomie et cytologie pathologiques, la pharmacie, la biologie et les données cliniques sont susceptibles d'être collectées par différents logiciels spécialisés pour répondre au besoin de la prise en charge des patients. Ainsi le stockage de ces données est réalisé dans des bases de données différentes, spécifiques à chaque application et, par conséquent, hétérogènes.

Ce manque d'intégration rend l'utilisation secondaire des données complexe, en séparant les données pertinentes des patients dans différents « silos » de stockage. En particulier, cette séparation des données complique l'interrogation transversale des données disponibles, en limitant les données relatives à un seul individu à des sous-ensembles incomplets au sein de chaque application du SIH.

#### *1.2.2.3 Hétérogénéité des données de santé hospitalières*

Les données de santé sont des données au sein des DPI collectées par de nombreux professionnels de santé qui ont une vision différente des individus qu'ils prennent en charge. Ces données sont ainsi recueillies en routine au fil de l'eau et de façon longitudinale. Elles reflètent le point de vue à un moment donné des professionnels de santé concernant les patients pris en charge. Par ailleurs les données produites peuvent être multimodales (texte, image, vidéo, son, etc.) et multi-niveaux (données décrivant un patient, une cellule, des bactéries, une mutation, etc.). Ces caractéristiques liées à l'utilisation primaire (c'est-à-dire dans le cadre du soin) en font des données complexes et hétérogènes par nature. On définit classiquement deux types d'hétérogénéité :

1. L'**hétérogénéité syntaxique** (*syntaxic heterogeneity*, plus rarement appelée *syntactic heterogeneity*, en anglais) ;
2. L'**hétérogénéité sémantique** (*semantic heterogeneity* en anglais) ;

### Hétérogénéité syntaxique

L'hétérogénéité syntaxique correspond à la co-existence de formats de données et de modèles multiples pour l'organisation des données de santé [80]. Les données de santé sont en effet de différents types :

- **Données structurées codées à l'aide de terminologies standards**, telles que les diagnostics (codés à l'aide de la Classification Internationale des Maladies, 10$^{\text{ème}}$ révision ou CIM-10 ; [81]) ou les données relatives à la prescription ou à l'administration de médicaments (codées à l'aide de la Classification Anatomique, Thérapeutique et chimique ou ATC ; [82]) ;
- Les **données structurées codées à l'aide de terminologies locales** (ou d'interface) [83], telles que les résultats biologiques comme illustré au chapitre 2 ;
- Les **données semi-structurées**, correspondant aux données enregistrées dans les formulaires de soins. Ces informations semi-structurées correspondent à des données structurées ou en texte libre contextualisées par une forte organisation au sein des formulaires : les questions sont organisées en sections, elles-mêmes organisées en pages spécifiques au sein des formulaires. En plus du sens propre porté par les questions et les réponses, beaucoup de sens est porté par le contexte où se situe la question ;
- Les **données non structurées / en texte libre**, telles que les comptes rendus d'hospitalisation, les comptes rendus d'imagerie ou les ordonnances de sortie.
- Les **données plus complexes**, comme par exemple les données d'imagerie ou les données de génétique.

Cette hétérogénéité de format entraîne une hétérogénéité des modes de stockage et d'interrogation. Par exemple, les données d'imagerie médicale sont organisées selon la norme DICOM (*Digital imaging and communications in medicine* [84]) au sein des PACS (*Picture Archiving and Communication System*). Le standard d'interrogation de ces données est également défini dans le DICOM, qui est spécifique aux données d'imagerie. D'un autre côté, les données cliniques saisies en texte libre peuvent être stockées dans des entrées de formulaires au sein du DPI (le stockage est souvent réalisé dans une base de données relationnelle) ou dans des fichier binaires (.docx, .pdf, etc.). Ces deux modes de stockage sont associés à des modes d'interrogation différents.

### Hétérogénéité sémantique

Au sein des hôpitaux, afin de faciliter l'utilisation des outils par les professionnels de santé, les DPI déployés sont adaptés spécifiquement aux usages locaux. Cette adaptabilité permet notamment d'organiser la collecte des données en s'alignant sur les besoins des professionnels de santé dans leur contexte spécifique. Bien qu'essentielle pour l'optimisation de la prise en charge des patients, cette adaptabilité génère des dictionnaires de données (dans la suite du document, nous utiliserons le terme de

**métadonnées** pour désigner ces dictionnaires) de grande dimension et par conséquent introduit une hétérogénéité sémantique au sein des DPI.

L'hétérogénéité sémantique correspond au fait que des éléments de même nature peuvent être exprimés de manière différente (données codées selon des terminologies différentes, description d'un phénomène par des formulations différentes au sein de textes, etc.). Par exemple, un résultat de biologie pourra être codé selon une terminologie locale, spécifiquement conçue au sein d'un établissement, selon une terminologie nationale comme la Nomenclature des Actes de Biologie Médicale (NABM ; [85]), selon un standard international comme la LOINC® (Logical Observation Identifiers Names & Codes ; [86, 87]) ou pourra être reporté dans un document sous forme de texte libre.

De la même manière, lorsque des concepts sont extraits du texte libre par des méthodes de Traitement Automatique des Langues (TAL - *Natural Language Processing* ou NLP en anglais), par exemple avec des méthodes de reconnaissance d'entités nommées (*Named-Entity Recognition* ou NER en anglais) ou des méthodes d'annotation sémantique (*Named-Entity-Linking* ou NEL en anglais), il est possible de voir apparaître de l'hétérogénéité sémantique. Si la ressource terminologique utilisée est une ressource spécifique de la tâche, ou une ressource standard mais différente des ressources utilisées pour coder les données structurées, l'hétérogénéité sémantique augmente.

Le manque de standards de métadonnées, la non-utilisation de ces standards ou l'utilisation de standards différents couvrant un même domaine sont des freins à l'utilisation secondaire des données [88].

### 1.2.3 Verrous liés aux organisations

#### *1.2.3.1 Réglementation sur les données de santé à caractère personnel*

D'un point de vue réglementaire, les données de santé sont régies par la Loi Informatique et Libertés (LIL ; loi n°78-17 du 6 janvier 1978 [89]) et le Règlement Général de Protection des Données (RGPD) européen (règlement UE 2016/679 du Parlement européen et du Conseil du 27 avril 2016 [90]). Les données de santé sont définies dans l'article 4 du RGPD comme des « *données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne* » [91].

Le considérant n°35 du RGPD apporte les précisions suivantes : « *Les données à caractère personnel concernant la santé devraient comprendre l'ensemble des données se rapportant à l'état de santé d'une personne concernée qui révèlent des informations sur l'état de santé physique ou mentale passé, présent ou futur de la personne concernée.*

*Cela comprend des informations sur la personne physique collectées lors de l'inscription de cette personne physique en vue de bénéficier de services de soins de santé ou lors de la prestation de ces services au sens de la directive 2011/24/UE du Parlement européen et du Conseil au bénéfice de cette personne physique ; un numéro, un symbole ou un élément spécifique attribué à une personne physique pour l'identifier de manière unique à des fins de santé ; des informations obtenues lors du test ou de l'examen d'une partie du corps ou d'une substance corporelle, y compris à partir de données génétiques et d'échantillons biologiques ; et toute information concernant, par exemple, une maladie, un handicap, un risque de maladie, les antécédents médicaux, un traitement clinique ou l'état physiologique ou biomédical de la personne concernée, indépendamment de sa source, qu'elle provienne par exemple d'un médecin ou d'un autre professionnel de la santé, d'un hôpital, d'un dispositif médical ou d'un test de diagnostic in vitro* » [92].

Dans le cas d'une utilisation secondaire des données de santé, le traitement des données de santé à caractère personnel doit respecter des conditions prévues par la loi et être en accord avec l'intérêt des patients. Cela implique notamment :

— Une **transparence** concernant les usages des données, avec une information claire et accessible, à la fois collective et individuelle, avant la collecte de la **non-opposition** du patient pour le traitement des données personnelles le concernant ;

— La collecte et le traitement des données doivent être strictement limités à ce qui est nécessaire pour la réalisation du projet (**minimisation**), et leur conservation doit être temporaire, dans un environnement sécurisé ;

— Des mesures techniques et organisationnelles appropriées doivent être mises en place pour garantir la sécurité et la confidentialité des données. En particulier, la **pseudonymisation** consiste à remplacer ou supprimer les informations directement identifiantes (nom, prénom, adresse, date de naissance, etc.), de manière à ce que les données ne puissent pas être reliées à une personne sans informations supplémentaires. De nombreux travaux décrivent des méthodes de pseudonymisation des données dans le contexte de l'utilisation secondaire des données de santé [93, 94, 95] ;

— L'**exercice des droits** des usagers, tels que le droit à l'information, à l'opposition, à la limitation du traitement, à la portabilité, ainsi qu'à la rectification ou au retrait de leurs données.

Outre ce socle général indispensable pour chaque étude impliquant une utilisation secondaire des données de santé, la réglementation différencie :

— Les « **études internes** », réalisées au sein d'une équipe de soins. Ces études peuvent être réalisées sous le couvert de l'article 65.2 de la LIL [89]. Dans ce cadre, une simple déclaration dans le registre des traitements du délégué à la protection des données personnelles (*Data Protection Officer* ou DPO en anglais) de l'établissement promoteur est nécessaire ;

— Les « **études multicentriques** » [5], réalisées en dehors du périmètre d'une seule équipe de soins. Dans ce contexte, les démarches réglementaires correspondent à un **régime d'autorisation** avec une déclaration de l'étude au niveau du portail du HDH, qui transmet la demande d'autorisation après de la CNIL et du CESREES.

Si l'établissement responsable de traitement s'est engagé auprès de la CNIL à respecter les obligations de conformité à la **méthodologie de référence MR-004** [96], le régime d'autorisation n'est plus nécessaire [6]. Dans ce cas, les formalités déclaratives des projets en utilisation secondaire des données de santé sont simplifiées avec :

1. Une déclaration de l'étude dans le registre des traitements locaux via le DPO de l'établissement ;
2. Un avis du Comité Scientifique et Éthique (CSE) de l'établissement, qui valide la pertinence éthique et scientifique de l'étude ;
3. Un enregistrement du projet dans le registre des projets sur le site du HDH [7].

Au total, les démarches réglementaires encadrant l'utilisation secondaire des données de santé restent complexes, comme soulevé dans le rapport Marchand-Arvier en 2023 [97].

#### *1.2.3.2 Différents acteurs, différents besoins*

Un autre verrou organisationnel concerne la multiplicité des acteurs qu'il est nécessaire de fédérer autour des projets impliquant une utilisation secondaire des données de santé [73]. Ces différents acteurs ont des objectifs différents vis à vis de cette utilisation secondaire. En particulier, on peut citer :

— Les **établissements de santé**, au travers notamment des directions de ces structures, qui sont responsables des traitements réalisés sur les données issues du soin. Ils sont aussi responsables de la reconnaissance de la propriété intellectuelle qui peut être générée dans les projets en utilisation secondaire des données de santé ;

— Les **professionnels de santé**, qui assurent la collecte des données de santé. Même si les données à caractère personnel sont régies par le principe d'indisponibilité [8] [98], des « réflexes propriétaires », tels qu'évoqués dans [97], peuvent entraver l'utilisation des données par des tiers non impliqués dans la collecte initiale des données ;

5. « Multicentrique » au sens de la CNIL, c'est-à-dire les études multi-services mono-établissement ou les études multi-établissements.

6. Les méthodologies de référence sont des procédures simplifiées pour l'accès et le traitement de données de santé dans un contexte de recherche, pour les établissements s'étant engagés à respecter le cadre sécurisé défini dans ces méthodologies (il existe différentes méthodologies en fonction du type d'étude). Ces méthodologies de référence permettent d'alléger le circuit technico-réglementaire pour le traitement de ces données.

7. Le répertoire public des projets du HDH est disponible via : `https://www.health-data-hub.fr/enregistrer-projet`.

8. Ni le patient concerné, ni le professionnel de santé qui en a assuré la collecte ne sont « propriétaires » des données (ici de santé) à caractère personnel générées.

— Les **patients**, qui ont des droits sur les données les concernant bien qu'ils n'en soient pas propriétaires (droits d'opposition, droit à la rectification, droit à la suppression, etc.) et vis-à-vis desquels les chercheurs ont des devoirs, en particulier de transparence.

Au total, de nombreux verrous liés aux données ou organisationnels peuvent entraver l'utilisation secondaire des données de santé, en particulier en ce qui concerne les données issues du soin. Dans le contexte des établissements de santé hospitaliers, les Entrepôts de Données de Santé (EDS ; *Clinical Data Warehouses* ou CDWs en anglais) sont une solution permettant de faciliter l'utilisation secondaire des données de santé » issues du soin. Plus que de simples outils techniques, les EDS doivent être de véritables plateformes techniques et organisationnelles en support de l'utilisation secondaire des données de santé.

## 1.3 Entrepôt de Données de Santé hospitalier, un outil pour « libérer l'utilisation secondaire des données de santé » issues du soin

Dans le contexte biomédical, et plus particulièrement au sein des hôpitaux, la solution mise en œuvre pour résoudre les différents verrous liés aux données et organisationnels évoqués précédemment est celle des EDS [99, 100, 101, 102, 103, 104].

Selon [105], les EDS sont des « *plateformes utilisées pour l'intégration de plusieurs sources de données par le biais d'outils analytiques spécialisés qui facilitent le traitement et l'analyse des données* ». Il s'agit ainsi de plateformes permettant l'intégration de données de santé hétérogènes et la mise à disposition, pour les utilisateurs finaux, d'une vue intégrée de ces données [35] pour l'utilisation secondaire des données de santé.

En premier lieu, les EDS sont des plateformes d'intégration des données hétérogènes de santé. Dans la section suivante, deux modèles d'intégration *open source* sont présentés et comparés.

### 1.3.1 EDS, un objet technique : modèles d'intégration des données

De nombreux modèles d'intégration des données dans les EDS ont été proposés [50, 105]. En particulier, deux modèles de données *open source* sont largement utilisés dans le monde :

— *Informatics for Integrating Biology & the Bedside* (**i2b2**) [106] proposé par le département d'informatique biomédicale de l'université d'Harvard en 2007.

— *Observational Medical Outcomes Partnership - Common Data Model* (**OMOP-CDM**) [107], résultant d'un partenariat public-privé mis en place en 2008 et dirigé par la FDA.

#### *1.3.1.1 i2b2 - Informatics for Integrating Biology and the Bedside*

i2b2 est une plateforme open-source proposée par le département d'informatique biomédicale de l'Université d'Harvard, aux États-Unis, depuis 2007 [106]. Cette plateforme est composée d'un modèle de données, d'une couche d'applications et d'une couche d'API (interface de programmation d'application ou *Application Programming Interface* en anglais) pour assurer la communication entre les différents modules qui constituent la plateforme. L'objectif d'i2b2 est double :

— Proposer un modèle d'intégration de données hétérogènes de santé dans une base de données dédiée à l'utilisation secondaire des données de santé. Ainsi, i2b2 correspond à un modèle de persistance (stockage) des données pour les EDS [9] ;

— Proposer un ensemble de services au-dessus de cette base de données intégrée pour l'utilisation secondaire des données de santé.

L'application i2b2 est développée selon une architecture micro-services (*service-based architecture* en anglais) comme illustrée sur la Figure 1.2. Chaque web-service est appelé « cellule » (*i2b2 cell* en anglais). L'ensemble des services déployés lors d'une installation d'i2b2, couplé avec le modèle de persistance des données, est appelé « ruche » (*i2b2 hive* en anglais). Un requêteur est proposé au-dessus de l'application i2b2 [10], permettant notamment l'identification de patients / séjours sur la base de critères clinico-biologiques (Figure 1.3).

9. Le modèle de données utilisé par i2b2 est proposé au travers de la persistance des données dans une base de données relationnelle. Les fichiers DDL du modèle de données d'i2b2 sont disponibles via : `https://github.com/i2b2/i2b2-data`

10. Un requêteur de démonstration est disponible à l'adresse suivante : `https://www.i2b2.org/webclient/`

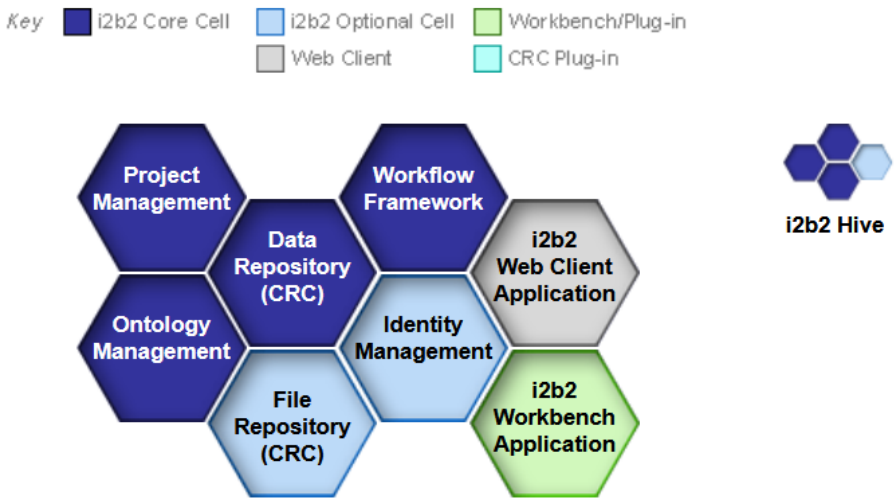


FIGURE 1.2 – Cellules et ruche d'i2b2

Extrait de : `https://www.i2b2.org/resrcs/hive.html`. Les hexagones correspondent à différentes « cellules » (*i2b2 cell*) qu'il est possible de déployer lors d'une installation d'i2b2. L'ensemble des cellules déployées lors d'une installation d'i2b2 correspond à la « ruche » (*i2b2 hive*). Les hexagones en bleu foncé correspondent aux cellules indispensables pour le bon fonctionnement de l'application i2b2 ; les hexagones en bleu ciel aux cellules optionnelles ; l'hexagone gris au front-end du requêteur proposé au-dessus d'i2b2 ; les hexagones verts aux plug-ins possibles dans i2b2. Chaque cellule expose un service, chargé d'un ensemble de fonctionnalités cohérentes pour un domaine particulier. Par exemple, la cellule « *Project Management* » correspond au service en charge des utilisateurs, des groupes et des projets ; la cellule « *Ontology Management* » correspond au service en charge des terminologies intégrées dans i2b2 ; la cellule « *Data Repository* » correspond au service de requêtage des données ainsi qu'au modèle utilisé pour la persistance des données dans i2b2 (données intégrées, données de requêtage, etc.).

Les données cliniques des patients sont intégrées dans le *Data Repository*. Le modèle d'intégration des données dé-normalisées cliniques dans i2b2 est basé sur un modèle en étoile [77], comme présenté dans la Figure 1.4). Ce modèle contient :

— Une table centrale de faits, `OBSERVATION_FACT`, au niveau de laquelle les données cliniques des patients sont intégrées sous forme d'observations. Une observation correspond à un point d'information concernant la santé d'un patient, recueillie lors d'une interaction entre le patient et le système de santé (au cours d'une hospitalisation ou d'une consultation, par exemple). Une observation peut correspondre à un élément structuré (diagnostic codé d'après une terminologie standard, résultat de biologie, etc.) ou à un élément non structuré (compte rendu d'imagerie, compte rendu d'hospitalisation, commentaire d'un résultat de biologie, etc.).
— Des tables périphériques de dimensions :

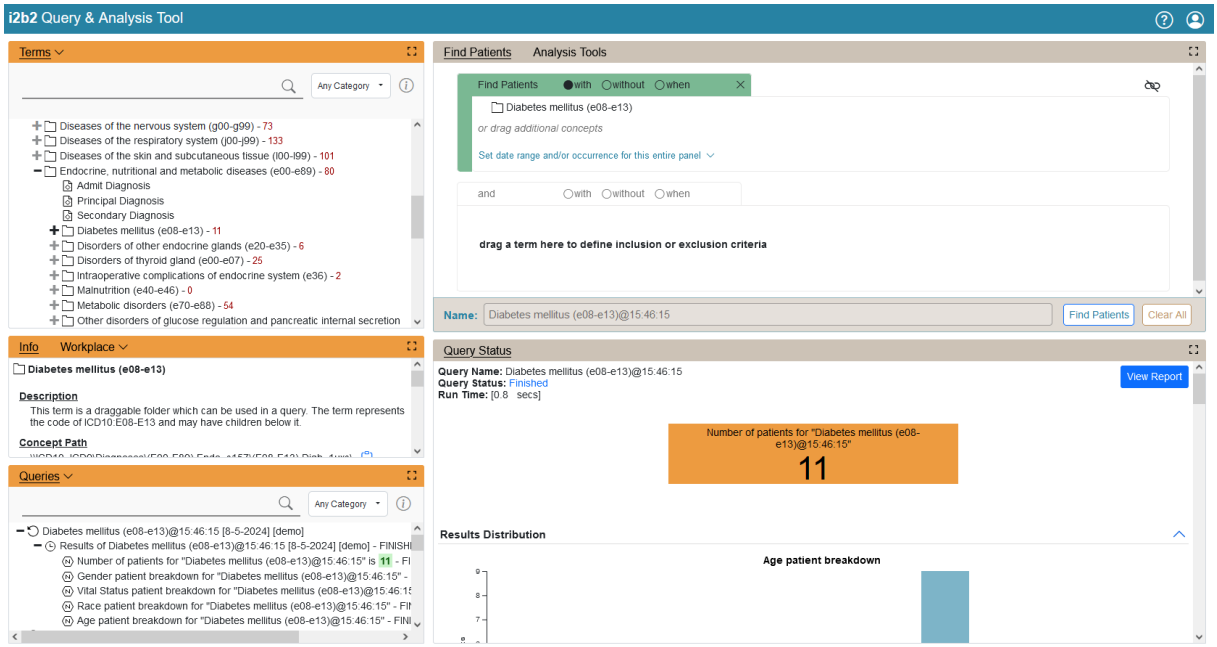


FIGURE 1.3 – Client Web d'i2b2

Extrait de : `https://www.i2b2.org/webclient/` Requêteur fourni au-dessus d'i2b2. Le requêteur permet 1) d'explorer les terminologies intégrées dans i2b2 au travers de la section « *Terms* » ; 2) de construire des requêtes de dénombrement de patients / séjours sur la base de critères clinico-biologiques dans la section « *Find Patients* » ; et 3) de visualiser les résultats de la requête dans la section « *Query Status* ». Ici, on dénombre l'ensemble des patients pour lequel un diagnostic de diabète (« *Diabetes mellitus* »), codé E08-E13 en CIM-10, est disponible.

- La table `PATIENT_DIMENSION`, qui contient les informations relatives aux patients (sexe, date de naissance, etc.) ;
- La table `VISIT_DIMENSION` qui contient les informations relatives aux venues (date d'entrée, date de sortie, etc.) ;
- Les tables contenant la description des valeurs possibles des colonnes de recherche (*lookup column*) de la table `OBSERVATION_FACT`. On distingue ici trois tables différentes :
  - La table `CONCEPT_DIMENSION`, qui décrit les concepts utilisés pour caractériser le sens principal d'une observation ;
  - La table `MODIFIER_DIMENSION`, qui décrit les valeurs possibles d'un niveau conceptuel secondaire, modifiant le niveau conceptuel principal ;
  - La table `PROVIDER_DIMENSION`, qui décrit les valeurs possibles des entités (soignants, unité de soins, etc.) à l'origine de l'observation.

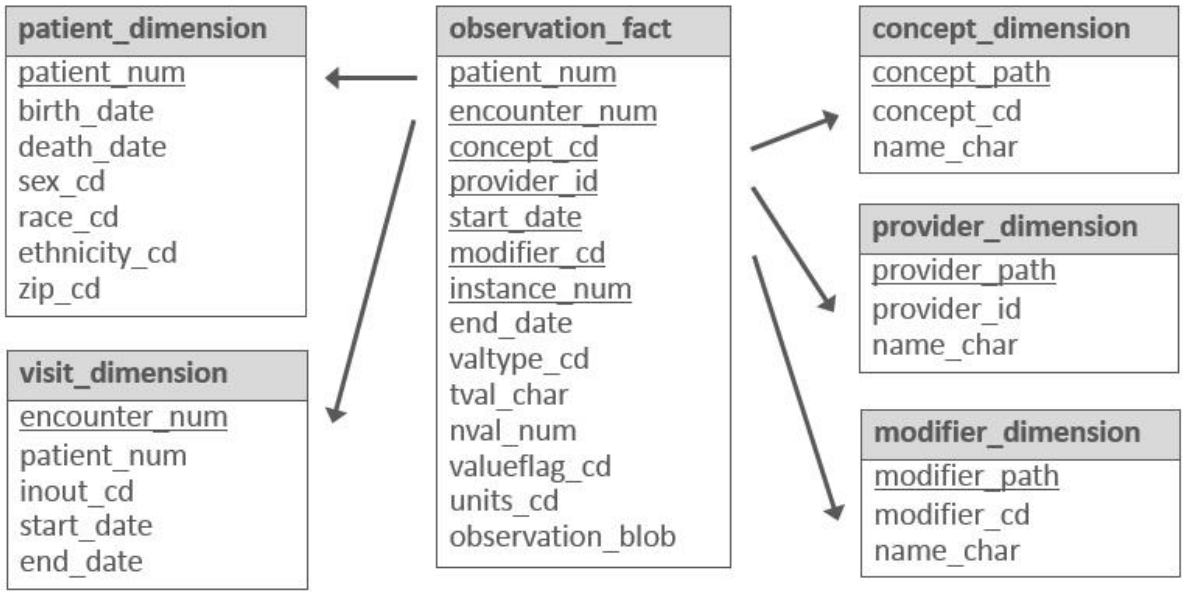


FIGURE 1.4 – i2b2 - Schéma en étoile

Extrait de : `https://community.i2b2.org/wiki/display/BUN/2.+Quick+Start+Guide`

#### *1.3.1.2 OMOP-CDM - Observational Medical Outcomes Partnership - Common Data Model*

OMOP-CDM est un modèle d'intégration des RWD en santé (données cliniques ou données médico-administratives) centré sur le patient [107]. Ce modèle a été proposé par OMOP, un partenariat public-privé entre la FDA, des universités, des responsables de données et l'industrie pharmaceutique fondé en 2008.

Depuis 2014, le modèle OMOP-CDM est maintenu et développé au sein de la communauté open-source *Observational Health Data Sciences and Informatics*[11] (OHDSI). La version actuelle d'OMOP-CDM est la 5.4[12].

En Europe, le modèle de données OMOP-CDM est promu par le réseau *European Health Data & Evidence Network* (EHDEN ; [108]). EHDEN a pour objectif de mettre en œuvre une fédération d'EDS standardisés suivant le modèle de données OMOP-CDM. Ce projet a initialement été financé par un projet *Innovative Medicines Initiative 2* (IMI 2) et est à présent maintenu par la *EHDEN foundation*.

Le modèle d'intégration de données d'OMOP-CDM est constitué de deux parties principales (Figure 1.5) :

11. Le site web d'OHDSI est disponible via : `https://ohdsi.org/`
12. La documentation du modèle est disponible à l'adresse suivante : `https://ohdsi.github.io/CommonDataModel/index.html`

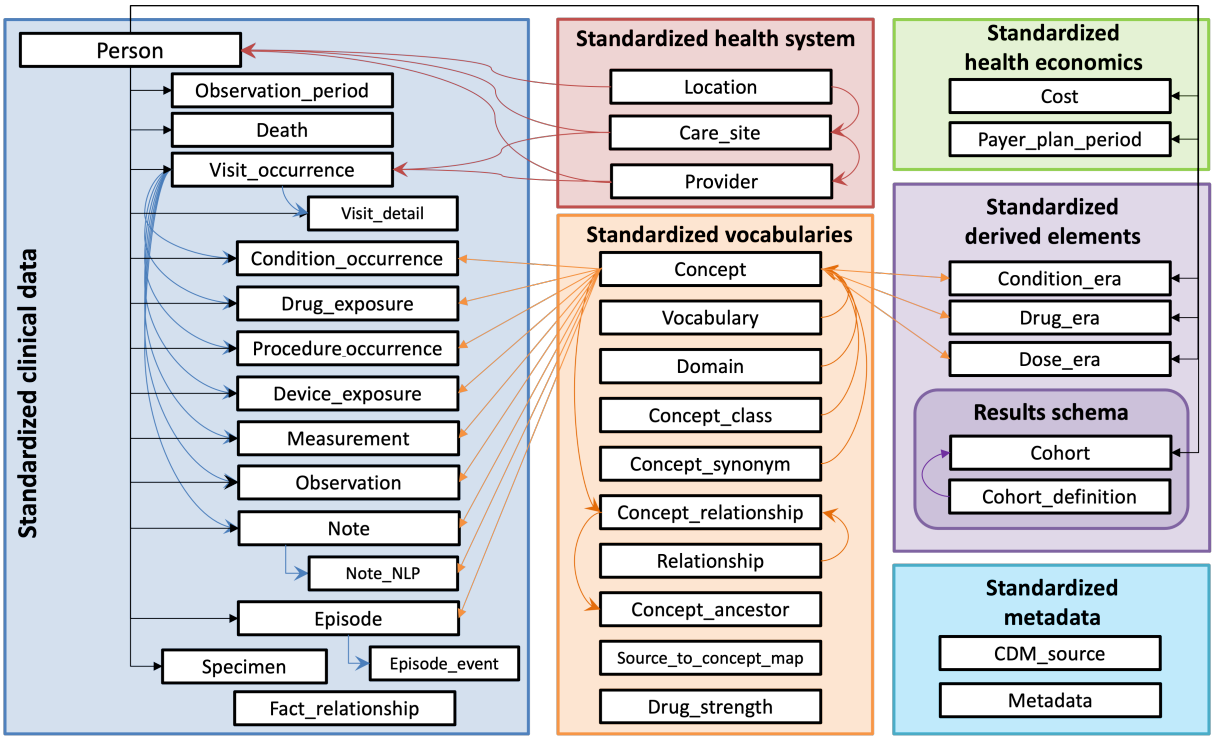


FIGURE 1.5 – Organisation du modèle de données d'OMOP-CDM

Extrait de : `https://ohdsi.github.io/CommonDataModel/`

— Les **données cliniques standardisées** (*Standardized Clinical Data* en anglais ; en bleu sur la Figure 1.5). Il s'agit de la partie du modèle de données d'OMOP-CDM qui permet l'intégration des données individuelles de santé.

L'ensemble des données intégrées dans OMOP-CDM est rattaché au patient (table `PERSON`) et à la venue (table `VISIT_OCCURRENCE`). Une table dénormalisée, spécialisée par domaine, permet l'intégration des différents types de données. Par exemple, la table `DRUG_EXPOSURE` permet l'intégration des données de prescription, d'administration ou de délivrance médicamenteuse ; la table `MEASUREMENT` permet l'intégration des données de mesures standardisées (résultat de biologie, mesures réalisées lors de l'examen physique d'un patient, etc.).

— Les **vocabulaires standardisés** (*Standardized Vocabularies* en anglais ; en orange sur la Figure 1.5). Il s'agit de la partie du modèle de données d'OMOP-CDM qui permet l'intégration des différents concepts utilisés dans le modèle. Dans cette partie du CDM, on retrouve notamment :

  — La table `CONCEPT`, qui contient l'ensemble des concepts disponibles dans la base de données OMOP. Cette table `CONCEPT` contient ainsi :

    — Les métadonnées locales issues de la (des) source(s) de données intégrée(s) dans OMOP-CDM. Ces métadonnées locales sont intégrées dans la table `CONCEPT` avec des identifiants supérieurs ou égaux à 2 000 000 000.

— Le vocabulaire proposé et maintenu par OHDSI [109] est téléchargeable via Athena[13]. Il s'agit d'un méta-modèle intégrant des terminologies nationales et internationales. Les différentes terminologies intégrées se chevauchant sur les domaines représentés, un « concept standard » par notion à représenter (*meaning* en anglais) est sélectionné. Quand cela est possible, les concepts non standards disponibles sont alignés vers ces concepts standards.

— La table `CONCEPT_RELATIONSHIP`, qui contient l'ensemble des alignements entre les concepts non standards (concepts locaux ou issus de terminologies intégrées mais non standards) et les concepts standards.

#### *1.3.1.3 Comparaison des modèles d'intégration d'i2b2 et d'OMOP*

i2b2 et OMOP sont des modèles d'intégration très proches. En particulier, i2b2 et OMOP sont basés sur une intégration des données observationnelles dans des tables de faits. Là où i2b2 ne contient qu'une seule table de faits (`OBSERVATION_FACT`), OMOP contient autant de tables de faits que de domaines médicaux couverts par le CDM (`MEASUREMENT`, `DRUG_EXPOSURE`, etc.) :

— Dans OMOP, cette spécialisation des tables de faits permet d'avoir des attributs spécifiques par domaine couvert (ex : pour le médicament, on va retrouver des attributs spécifiques autour du dosage du médicament, de la voie d'administration, etc.). Cependant, ces attributs spécifiques, différents d'un domaine à l'autre, aboutissent souvent à des colonnes non remplies [110].

— Dans i2b2, le mécanisme permettant la gestion de ces attributs complémentaires est géré par les *modifiers*. Il s'agit d'un attribut unique de la table `OBSERVATION_FACT`, permettant d'apporter des informations complémentaires à une observation principale, en créant des lignes de faits secondaires. Un exemple de cette gestion des *modifiers* est présenté dans le Tableau 1.1.

Tableau 1.1 – Gestion des attributs complémentaires (*modifiers*) dans i2b2

| ENCOUNTER_NUM | CONCEPT_CD | MODIFIER_CD | INSTANCE_NUM | NVAL_NUM |
|---|---|---|---|---|
| 10000 | PRESC:PARACETAMOL | @ | 20000 | |
| 10000 | PRESC:PARACETAMOL | PRESC:DOSE | 20000 | 1000 |
| 10000 | PRESC:PARACETAMOL | PRESC:ROUTE\|PO | 20000 | |
| 10000 | PRESC:PARACETAMOL | PRESC:ALD | 20000 | |

L'observation principale (celle pour laquelle le `MODIFIER_CD = '@'`) correspond à une administration de paracétamol. Trois observations secondaires spécifient la dose, la voie d'administration et le fait que la prescription est à la demande. C'est l'`INSTANCE_NUM` identique entre l'observation principale et les observations secondaires qui permet de faire le lien.

13. Un explorateur du vocabulaire proposé par OHDSI est disponible via : `https://athena.ohdsi.org/`

Une plus grande différence entre ces deux modèles d'intégration réside dans l'importance et la gestion des vocabulaires entre OMOP et i2b2 :

— OMOP, au travers d'OHDSI, met à disposition un vocabulaire contrôlé sur lequel il est nécessaire de s'aligner pour s'intégrer totalement dans le modèle OMOP. Le modèle OMOP-CDM est très fortement lié à ce vocabulaire. Intégrer des données dans OMOP sans les standardiser avec le vocabulaire standard d'OHDSI correspond à une intégration partielle des données dans OMOP. Ainsi, OMOP-CDM peut à la fois être vu comme un modèle d'intégration syntaxique et sémantique.

— i2b2, quant à lui, propose une organisation souple permettant une intégration des vocabulaires, en particulier des métadonnées locales, sous forme de hiérarchie. Des vocabulaires standards sont également proposés par i2b2, notamment par intégration de vocabulaires standards externes [111]. Dans le cadre de SHRINE [14] (*Shared Health Research Informatics Network*) [112], un vocabulaire contrôlé est également mis à disposition pour la standardisation des données nécessaire à l'exécution de requêtes de dénombrement distribuées.

Bien que l'utilisation d'un vocabulaire standardisé dans i2b2 soit possible, elle est beaucoup moins au centre du modèle qu'elle ne l'est dans OMOP-CDM. Ainsi, i2b2 peut être vu comme un modèle permettant principalement une intégration syntaxique des données.

### 1.3.2 EDS, un objet réglementaire en France : le référentiel EDS de la CNIL

En novembre 2021, la CNIL a mis en place un référentiel spécifique sur les EDS, appelé « Référentiel relatif aux traitements de données à caractère personnel mis en œuvre à des fins de création d'entrepôts de données dans le domaine de la santé » [113]. Ce référentiel définit les exigences en matière :

— De **gouvernance**. Le référentiel défini plus précisément deux organes de gouvernance différents :

1. Un comité de pilotage, en charge de déterminer les orientations stratégiques et scientifiques de l'entrepôt ;
2. Un comité scientifique et éthique (CSE), en charge de rendre un avis préalable à chaque utilisation secondaire des données de santé.

— De **catégories de données intégrées dans l'entrepôt** :

1. Les données directement identifiantes relatives aux patients ;

14. SHRINE permet de diffuser et de distribuer des requêtes de manière standardisée entre différents EDS au format i2b2, sans déplacement de données à caractère personnel. Seuls les résultats agrégés des requêtes (c'est-à-dire de dénombrement) sont partagés.

2. Les données pseudonymisées de santé relatives aux patients. Sont distinguées dans cette catégorie les données génétiques, les données de suivi de localisation et les autres données de santé (données cliniques, données biologiques, etc) ;
3. Les données relatives aux professionnels de santé ;

Sont associées à ces différents types de données des contraintes techniques et réglementaires d'accès différentes. Par exemple, les données pseudonymisées de santé et les données directement identifiantes relatives aux patients doivent être stockées dans des environnements isolés (cloisonnement logique et cryptographique) avec des habilitations d'accès différentes en fonction des utilisateurs.

— D'**information** auprès des patients et des professionnels de santé, d'**exercice des droits** et de **transparence**. En particulier, le référentiel rappelle la nécessité d'une information individuelle et préalable à chaque finalité, et le droit d'opposition pour les patients quant à l'utilisation secondaire des données de santé les concernant ;

— De **sécurité** des systèmes d'information relatifs aux entrepôts et aux espaces de travail spécifiques pour les projets de recherche. Ces éléments de sécurité recouvrent des éléments techniques (cryptographie, pare-feu, pseudonymisation, etc.), mais aussi des éléments de sensibilisation et de formation des utilisateurs.

Les données de santé sont soumises à une réglementation forte, de par leur sensibilité intrinsèque et du fait qu'il s'agisse de données à caractère personnel. Une solution intéressante pour favoriser l'utilisation secondaire des données de santé, en simplifiant les démanches technico-réglementaires, sont les réseaux fédérés d'EDS.

### 1.3.3 Réseaux fédérés d'Entrepôts de Données de Santé

Un réseau fédéré est une architecture de systèmes distribuée dans laquelle plusieurs nœuds coopèrent, tout en conservant leur autonomie. Chaque nœud dans le réseau conserve ses propres données et son propre contrôle opérationnel, mais collabore avec les autres nœuds pour partager certaines données ou exécuter des tâches communes [114]. Ainsi, dans un réseau fédéré, il n'y a pas de base de données centralisée unique où toutes les données sont stockées ou gérées. Ce sont les résultats des analyses produits localement qui sont partagés dans le réseau.

Les données individuelles pseudonymisées restant stockées localement sous la responsabilité des producteurs de données, les contraintes réglementaires liées au partage de ces données (RGPD) sont largement réduites, permettant une meilleure réactivité dans la production de résultats au sein des réseaux fédérés.

Dans le domaine de la santé, de nombreux réseaux fédérés ont été mis en place [112, 115, 3, 116, 117]. Dans la suite, nous décrivons les réseaux 4CE et DARWIN-EU.

#### *1.3.3.1 Réseau 4CE*

Le réseau 4CE (*Consortium for Clinical Characterization of COVID-19 by EHR*) est un réseau mis en place en 2020 par Isaac S. Kohane dans le contexte de la pandémie de COVID-19 [3]. L'objectif du réseau 4CE était la mise à disposition rapide de données agrégées ouvertes en lien avec la COVID-19, à partir des données hospitalières issues de multiples EDS hospitaliers dans le monde. Un total de 96 hôpitaux ont contribué à la première étude de 4CE, avec des premiers résultats mis à disposition dès avril 2020. Les données partagées consistaient en des données démographiques, des données épidémiologiques, des données relatives aux diagnostics et des données relatives à la biologie.

La Figure 1.6 décrit les flux de données au sein du réseau 4CE. Les établissements de santé qui ont contribué à ce réseau disposaient des données intégrées dans des EDS aux formats i2b2 ou OMOP. Au niveau de chaque site participant, les données disponibles en lien avec la COVID-19 étaient requêtées et agrégées dans un modèle de données commun [15]. Ces données agrégées étaient partagées au niveau du consortium, qui étaient en charge de les regrouper et de mettre à disposition des visualisations communes.

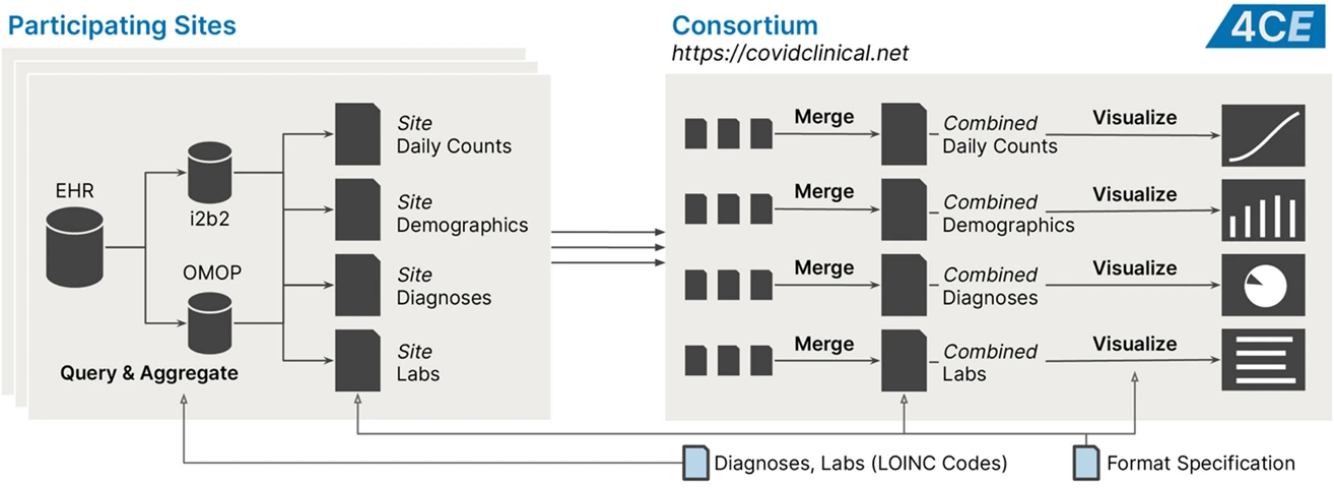


FIGURE 1.6 – Flux de données au sein du réseau fédéré 4CE

Extrait de : [3]

Au total, le réseau 4CE a contribué à la réalisation de 15 études, principalement autour de la caractérisation des populations atteintes de la COVID-19 [5, 3, 118, 119, 4, 120, 121, 122]. Certaines études avaient un focus spécifique dans la population pédiatrique

15. Les données étant agrégées au niveau du modèle de données commun, celui-ci ne contenait pas de données à caractère personnel.

[123, 124, 125] et d'autres sur le COVID long [126, 127, 128]. Une étude méthodologique sur l'estimation des modèles de Cox dans le cadre des réseaux fédérés a également été réalisée [129].

#### *1.3.3.2 Réseau DARWIN-EU*

Le réseau DARWIN-EU est un réseau fédéré européen dont l'objectif est l'évaluation de l'utilisation, de la sécurité et de l'efficacité des médicaments au sein de l'Union Européenne (UE), à partir de données observationnelles issues de bases de RWD disponibles sur le territoire [130, 131].

Le réseau DARWIN-EU a été mis en place suite aux recommandations de l'**HMA/EMA Big Data Task Force** (HMA : *Heads of Medicines Agencies*, réseau des agences nationales de régulation des médicaments en santé humaine et vétérinaire de l'UE ; EMA : *European Medicines Agency*, agence européenne du médicament) initiées en 2019. L'objectif de ce groupe de travail était de cartographier les « big data » disponibles / intéressants pour le régulateur et d'établir des recommandations sur les étapes à mettre en place pour améliorer l'usage de ces données. Parmi les 10 recommandations prioritaires[16] qui ont été émises, la recommandation n°1 était de « *mettre en place une plateforme durable permettant d'accéder aux données sur les soins de santé provenant de l'ensemble de l'UE et de les analyser (Data Analysis and Real World Interrogation Network - DARWIN)* ».

En Juin 2021, un appel d'offres pour la mise en place du *DARWIN Coordination Center* (DARWIN-CC) a eu lieu. DARWIN-CC a pour objectif la mise en place opérationnelle d'un réseau fédéré, l'identification et le recrutement de bases de RWD (appelées *Data Partners* ou DP) et la conduite des études pour soutenir la prise de décision réglementaire. C'est le centre médical *Erasmus University Medical Center* de Rotterdam, aux Pays-Bas, qui a remporté cet appel d'offres.

Concernant les DP, deux vagues de recrutement ont eu lieu en novembre 2022 et en mars 2024[17]. Au total, 20 DP représentant 13 pays différents de l'UE ont été recrutés dans le réseau DARWIN-EU, incluant des acteurs publics et privés (Figure 1.7). Au total, près de 150 millions de patients sont inclus dans ce réseau. La liste des DP intégrés dans le réseau DARWIN-EU est disponible en annexe (Tableau A.1).

En France, deux bases de RWD ont été intégrées dans DARWIN-EU : l'EDS OMOP du Centre Hospitalier Universitaire (CHU) de Bordeaux et une partie du Système

16. Les 10 recommandations prioritaires de l'HMA-EMA joint Big Data Task Force sont disponibles via : `https://www.ema.europa.eu/en/documents/other/priority-recommendations-hma-ema-joint-big-data-task-force_en.pdf`

17. La liste des DP est disponible via : `https://www.ema.europa.eu/en/documents/other/darwin-eu-data-partners-onboarded-phases-i-ii_en.pdf`

National des Données de Santé (SNDS).

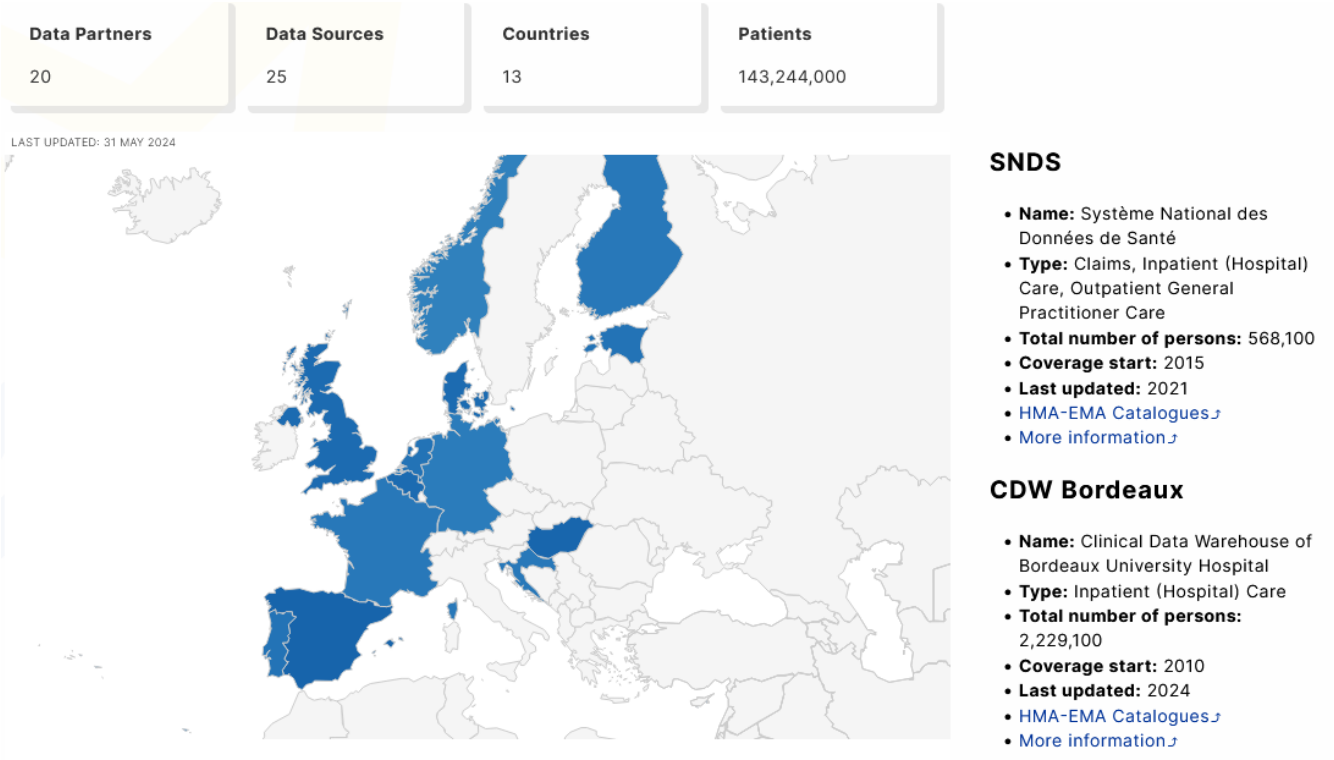


FIGURE 1.7 – Couverture du réseau DARWIN-EU

Extrait de : `https://www.darwin-eu.org/index.php/data/data-network`

Comme DARWIN-EU est un réseau fédéré (Figure 1.8), les données observationnelles restent au sein des différentes bases de RWD qui contribuent au réseau tandis que les scripts d'analyses, qui sont créés de manière centralisée au niveau du DARWIN-CC, sont partagés et exécutés localement au niveau des différents DP. Seules des données anonymes (données agrégées et offusquées [18]), sont renvoyées au centre de coordination qui agrège les résultats des différents sites participant à une étude.

Au total, 30 études ont été conduites au sein du réseau fédéré DARWIN-EU, dont 15 ont été achevées (en septembre 2024). Trois autres études sont d'ores et déjà prévues.

18. L'offuscation correspond à la non-présentation de dénombrements précis lorsque la taille du groupe dénombré est faible, souvent inférieure à 5 ou 10.

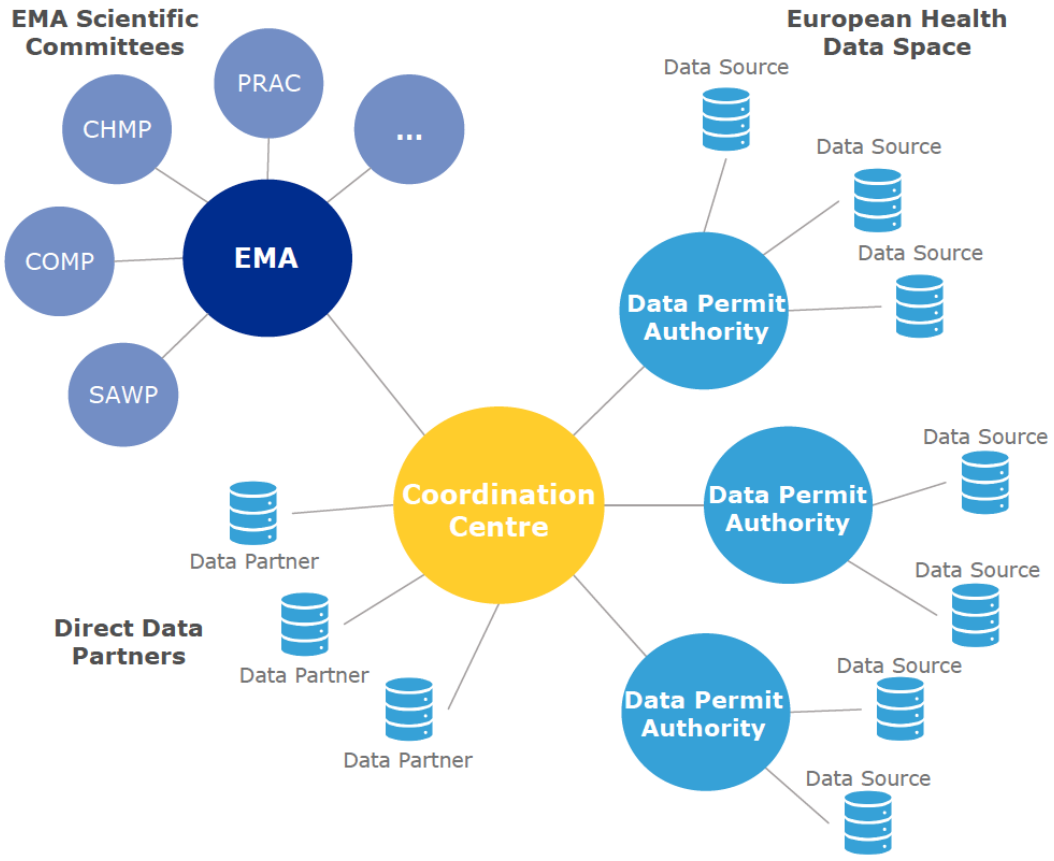


FIGURE 1.8 – Réseau fédéré DARWIN-EU

Extrait de : `https://www.darwin-eu.org/index.php/data/network-requirements`

## 1.4 L'Entrepôt de Données de Santé hospitalier du CHU de Bordeaux

Depuis 2018, le CHU de Bordeaux travaille à la construction de son EDS. Depuis le début, les grands objectifs de cet EDS ont été définis afin de construire une feuille de route à long terme :

1. **Éviter les « *boîtes noires* » dans le système**. Comme ces RWD sont utilisées pour produire des connaissances, il est essentiel d'avoir un contrôle complet de la chaîne de traitement des données, depuis l'intégration des données jusqu'à la production des résultats ;

2. **Assurer la transparence et renforcer la confiance entre les acteurs**. En raison du caractère sensibible des données de santé sous-jacentes [132], il est très important de créer des outils transparents afin de garantir la confiance entre les patients, les médecins et les chercheurs ;

3. **Accéder facilement aux données pour les utilisateurs finaux**, en clarifiant les circuits de demande. Les EDS sont conçus pour fournir des réponses en quelques secondes ; il serait contradictoire que les utilisateurs doivent attendre des mois pour mettre en œuvre le cas d'utilisation qu'ils envisagent ;

4. **Autonomiser les utilisateurs finaux**. Afin de mettre en place un système efficace, il est essentiel de donner aux utilisateurs finaux l'accès aux données et de leur permettre de mener leurs recherches de la manière la plus autonome possible ;

5. **Partager les données au-delà de l'institution**. Afin de tirer le meilleur parti des données disponibles, il a été identifié dès le départ que le partage des données en dehors de l'établissement était un objectif important. Cet objectif vise à permettre aux acteurs extérieurs à l'institution de produire des connaissances à partir des données disponibles au sein de l'EDS.

La description détaillée de l'environnement EDS du CHU de Bordeaux est proposée dans le chapitre 4 de ce manuscrit.

## 1.5 Contributions de ce travail de thèse

Ce travail de thèse s'inscrit dans l'environnement de l'EDS du CHU de Bordeaux, avec pour objectif d'aborder un certain nombre de problématiques liées à l'utilisation secondaire des RWD. Concrètement, il contribue à proposer des solutions et des méthodes pour résoudre des problèmes qui limitent l'utilisation secondaire des données de santé.

Dans ce travail, nous montrons comment un EDS, au-delà d'être un outil à destination d'utilisateurs finaux souhaitant produire des connaissances à partir des RWD, est encore aujourd'hui un objet de recherche en soi, nécessitant encore de développer des méthodes et des organisations nouvelles pour permettre une utilisation optimale des données.

Dans le but de réduire l'hétérogénéité sémantique des données intégrées dans l'EDS du CHU de Bordeaux, le **chapitre 2** présente un travail d'alignement de la biologie numérique basée sur les instances. Adresser la problématique de l'hétérogénéité sémantique vise à répondre aux enjeux de standardisation indispensables à l'usage en autonomie par les utilisateurs finaux d'une part, et à la mise en œuvre des réseaux fédérés d'autre part, répondant ainsi aux objectifs 4 et 5 de la feuille de route.

Dans le **chapitre 3**, un travail d'adaptation du modèle de données d'i2b2 pour permettre la persistance des données intégrées dans l'EDS dans une base de données NoSQL est présenté. Une évaluation de l'efficience de cette persistance et du requêtage est également proposée. Ces travaux sont essentiels pour répondre aux objectifs 3 et 4 de la feuille de route, notamment en favorisant l'interaction efficiente avec les données en autonomie pour les utilisateurs finaux. Il constitue la clef de voûte de l'usage local.

Enfin, le **chapitre 4** détaille les implémentations des deux EDS du CHU de Bordeaux, en support des usages locaux (objectifs 3 et 4) et fédérés (objectif 5). Cette partie montre comment les choix et les méthodes développées sur les aspects techniques, méthodologiques et organisationnels répondent aux objectifs énoncés ci-dessus.

# CHAPITRE 2

## Alignements des data elements numériques de biologie pour une standardisation de la biologie en LOINC

## Sommaire



# 2.1 Introduction

Comme dit précédemment, les données de santé produites dans le cadre des soins peuvent être réutilisées à de multiples fins (phénotypage [12, 133], recherche [11, 16], etc.). Bien que l'utilisation secondaire des données de santé revêt un intérêt majeur [34, 53, 134], elle est cependant complexe [135, 136, 137] (grands volumes de données, données cloisonnées, données non disponibles, etc.). Une des difficultés majeures réside dans l'hétérogénéité de la représentation des concepts médicaux, c'est-à-dire l'**hétérogénéité sémantique** [41, 138]. Différentes approches peuvent être utilisées pour réduire cette hétérogénéité. En particulier, de nombreux efforts se concentrent sur l'alignement des terminologies locales (ou terminologie d'interface [139]) à des standards

internationaux [140, 141, 142].

Dans la suite du document, les codes utilisés pour représenter les concepts de biologie dans les terminologies locales et les standards seront appelés **data elements**. Au sens de la norme ISO-11179-3 [143], un data element est « une unité de donnée considérée comme indivisible dans un contexte particulier ». D'une manière simple, un data element peut être vu comme l'une des représentations possibles d'une question conceptuelle (appelée **data element concept**). Par exemple, la question conceptuelle « Quelle est la valeur de la glycémie sanguine à jeun d'un patient ? » peut être représentée dans un formulaire du DPI par la question « Glycémie à jeun : » ou être représentée par le code « BIO:GLYCEMIE_A_JEUN » dans la terminologie d'interface d'un logiciel de biologie. Ainsi, aligner des terminologies de biologie peut être vu comme la réalisation d'un alignement entre data element.

En biologie, l'un des standards les plus utilisés dans le monde pour le codage des données biologiques est la *Logical Observation Identifiers Names and Codes* (LOINC®) [86, 87]. La LOINC® est une terminologie internationale, créée en 1994, pour le codage des observations de biologie médicale. En 1999, HL7 [1] (*Health Level Seven*) a désigné la LOINC comme le dictionnaire à privilégier pour nommer les tests de biologie dans les normes d'interopérabilité. Depuis son développement initial, la LOINC a vu son domaine s'étendre, lui permettant de couvrir, en plus de la biologie médicale, les observations cliniques (température, poids, etc.) et les documents.

Les différentes entrées de la LOINC sont appelées « concepts LOINC » (*LOINC term* en anglais). Ces derniers décrivent les éléments de biologie médicale selon les six attributs majeurs suivants (Figure 2.1) :

- Le « composant » (*component* en anglais) qui correspond à l'analyte mesuré (ex : glucose, créatinine, etc.) ;
- Le « milieu » (*system* en anglais) indiquant la nature de l'échantillon biologique dans lequel a été mesuré l'analyte (ex : serum ou plasma, sang artériel, urine, etc.) ;
- La « grandeur » (*property* en anglais) qui indique la nature de ce qui est mesuré (ex : concentration massique, concentration molaire, nombre, volume, etc.) ;
- L' « échelle » (*scale* en anglais) qui précise la nature de l'échelle de mesure (ex : quantitative, qualitative, ordinale, etc.). Il convient de noter que l'unité de mesure n'est pas spécifiée dans l'échelle : l'unité de mesure n'est pas un élément constitutif du code LOINC ;
- Le « temps » (*time aspect* en anglais) correspondant à la temporalité de la mesure (ex : instantanée, 1h après la prise d'un traitement, sur 24h, etc.) ;
- La « méthode » (*method* en anglais) qui décrit le type de technique utilisé pour effectuer l'analyse (ex : comptage manuel, dosage radio-immunologique, etc.). Cet

1. HL7 est une organisation en charge de la définition de spécifications techniques pour l'échange de données au sein des systèmes d'information hospitaliers.

attribut est optionnel et est particulièrement important quand la méthode affecte l'interprétation clinique du résultat.

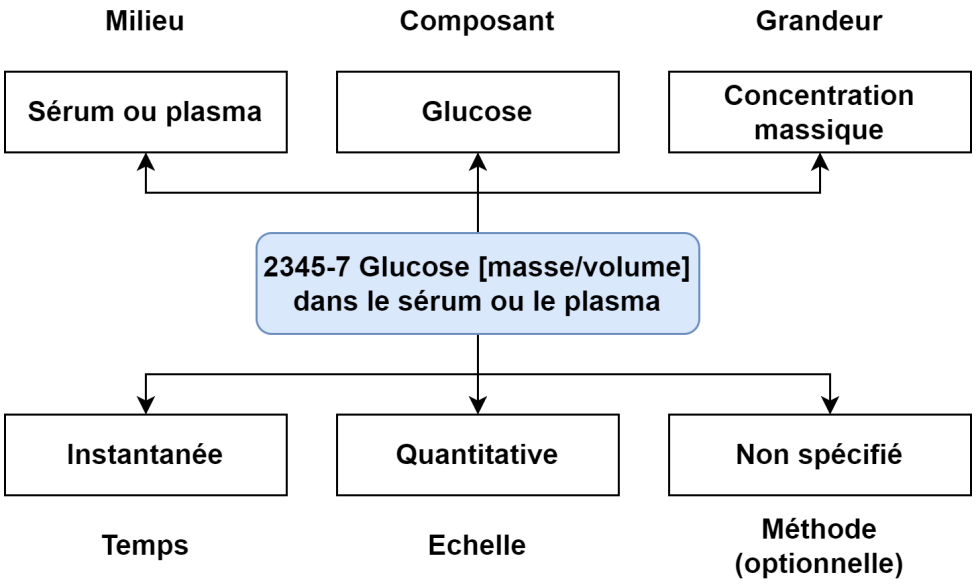


FIGURE 2.1 – Exemple d'un concept LOINC - Glucose sanguin (2345-7)

En plus de ses différents attributs, chaque concept LOINC est associé à un code, le « code LOINC », et trois libellés différents :

— Le*Fully-Specified Name*, qui correspond à la concaténation des six attributs majeurs du code LOINC (ex : *Glucose:MCnc:Pt:Ser/Plas:Qn*);

— Le *Long Common Name*, qui correspond à un libellé long, facilement lisible par un humain (ex : *Glucose [Mass/volume] in Serum or Plasma*) ;

— Le *Short Name*, qui correspond à un libellé court, en particulier utile pour les systèmes qui n'accepte pas des libellés trop long (ex: *Glucose SerPl-mCnc*).

Dans le SIH du CHU de Bordeaux, trois sources de biologie différentes ont été disponibles au cours du temps :

— TD-Synergy® : il s'agit du système de gestion des informations de laboratoire (*Laboratory Information Management System* ou LIMS en anglais) historique du CHU de Bordeaux, utilisé jusqu'en 2018 ;

— Glims® : il s'agit du LIMS qui a progressivement pris la suite de TD-Synergy® depuis 2018 ;

— DxCare® : il s'agit du DPI du CHU de Bordeaux. Les résultats produits par les deux LIMS ci-dessus sont envoyés à DxCare® dans le contexte du soin (c'est dans DxCare® que les soignants consultent les résultats de biologie). Pour chaque data element de biologie disponible dans chacun des deux LIMS, un data element de biologie est créé dans DxCare®. La correspondance entre les data element des

LIMS et ceux de DxCare® n'est pas disponible.

Les données issues de ces trois logiciels sont intégrées dans l'EDS du CHU de Bordeaux. Chacun des trois logiciels possède sa propre terminologie locale. De plus, au sein d'une même source de biologie, plusieurs data element différents peuvent être utilisés pour représenter les mêmes concepts de biologie. Cette multiplicité de data element pour représenter les mêmes concepts entraîne un degré élevé d'hétérogénéité sémantique pour la biologie au niveau de l'EDS du CHU de Bordeaux.

Une des sources de biologie est partiellement alignée avec la LOINC. Ainsi, aligner les data elements des différents logiciels de biologie pourrait permettre d'étendre le mapping avec la LOINC aux deux autres sources de données pour lequel les alignements ne sont pas encore disponibles. En outre, l'alignement des data elements de biologie des différentes sources - indépendamment d'un mapping vers la LOINC - pourrait impliquer que, dès qu'un alignement est disponible pour l'une des sources, il pourrait être propagé à tous les data elements alignés entre eux, limitant ainsi la charge d'un alignement avec la LOINC, qui est coûteux en termes de ressources et de temps [144].

L'alignement entre terminologie a été largement étudié dans la littérature ces dernières décennies [145, 146]. Différents cadres ont été proposés pour caractériser les différentes méthodes d'alignement, dont le plus commun a été proposé par Euzenat *et al.* [147], qui distingue les méthodes terminologiques (ou lexicales), les méthodes structurelles, les méthodes sémantiques et les méthodes basées sur des instances (ou extensionnelles).

Dans le contexte des EDS, qui agrègent de larges volumes de données de santé, les méthodes d'alignements basées sur les instances semblent revêtir un intérêt particulier. Cependant, avec le développement des méthodes d'apprentissage automatique (*Machine Learning* ou ML en anglais) et d'Intelligence Artificielle (IA), de nombreuses études ont soulevé des questions de confidentialité et de sécurité dans les méthodes d'IA [148, 149, 150].

Dans ce contexte, nous proposons ici une méthode d'alignement basée sur les instances [151] des data elements de biologie numérique sur la base de features agrégées totalement anonymes, permettant de contrôler les risques de compromission de la confidentialité des données intégrées dans l'EDS du CHU de Bordeaux.

## 2.2 Méthodes

L'objectif de ce travail était de développer une méthode d'alignement entre différents data elements de biologie numérique représentant le même concept LOINC. Cette tâche d'alignement entre deux data elements peut être considérée comme une tâche de **entity**

**linkage** [152], c'est-à-dire une tâche de classification de paires de data elements. On peut ainsi définir :

— Les « **paires positives** », c'est-à-dire les paires pour lesquelles les deux data elements constituant la paire représentent le même concept de biologie ;

— Les « **paires négatives** », c'est-à-dire les paires pour lesquelles les deux data elements constituant la paire représentent deux concepts de biologie différents.

Afin de réaliser ces alignements, une méthode en trois étapes successives a été proposée (Figure 2.2) :

1. Sélection des data elements au sein de l'EDS du CHU de Bordeaux et extraction des features agrégées (section 2.2.1) ;
2. *Blocking strategy* pour réduire la combinatoire des paires de data elements possibles et extraction des mesures de similarité entre les paires (section 2.2.2) ;
3. Parmi les paires sélectionnées, calcul des similarités entre ces paires et identification des mappings par apprentissage machine supervisé (section 2.2.3).

Une évaluation de la *blocking strategy* et des résultats obtenus avec les différents modèles d'apprentissage supervisé a ensuite été réalisée (section 2.2.4).

## 2.2.1 Sélection des data elements et extraction des features

Cette partie est décrite au niveau de la Figure 2.2A.

#### *2.2.1.1 Sélection des data elements de biologie numérique*

La première étape de la méthode correspond à une étape de sélection des data elements :

— La méthode d'alignement proposée ne concerne que les data elements rattachés à des données numériques. Ainsi, les data elements rattachés à des résultats structurés (ex : résultat d'une hémoculture bactérienne) ou en texte libre (ex : commentaire associé à un résultat de biologie) ont été exclus ;

— De plus, s'agissant de produire des features sur la base de données anonymes, les data elements rattachés à un faible nombre d'observations ont également été exclus. Un seuil de 10 000 observations disponibles a été utilisé pour filtrer les data elements avec un nombre d'observations jugé trop faible.

#### *2.2.1.2 Extraction des features de biologie numérique*

Sur la base de ce sous-ensemble de data elements correspondant à des résultats numériques de biologie pour lesquels au moins 10 000 observations étaient disponibles, des features agrégées ont été calculées. Afin d'éliminer les outliers, les features ont été

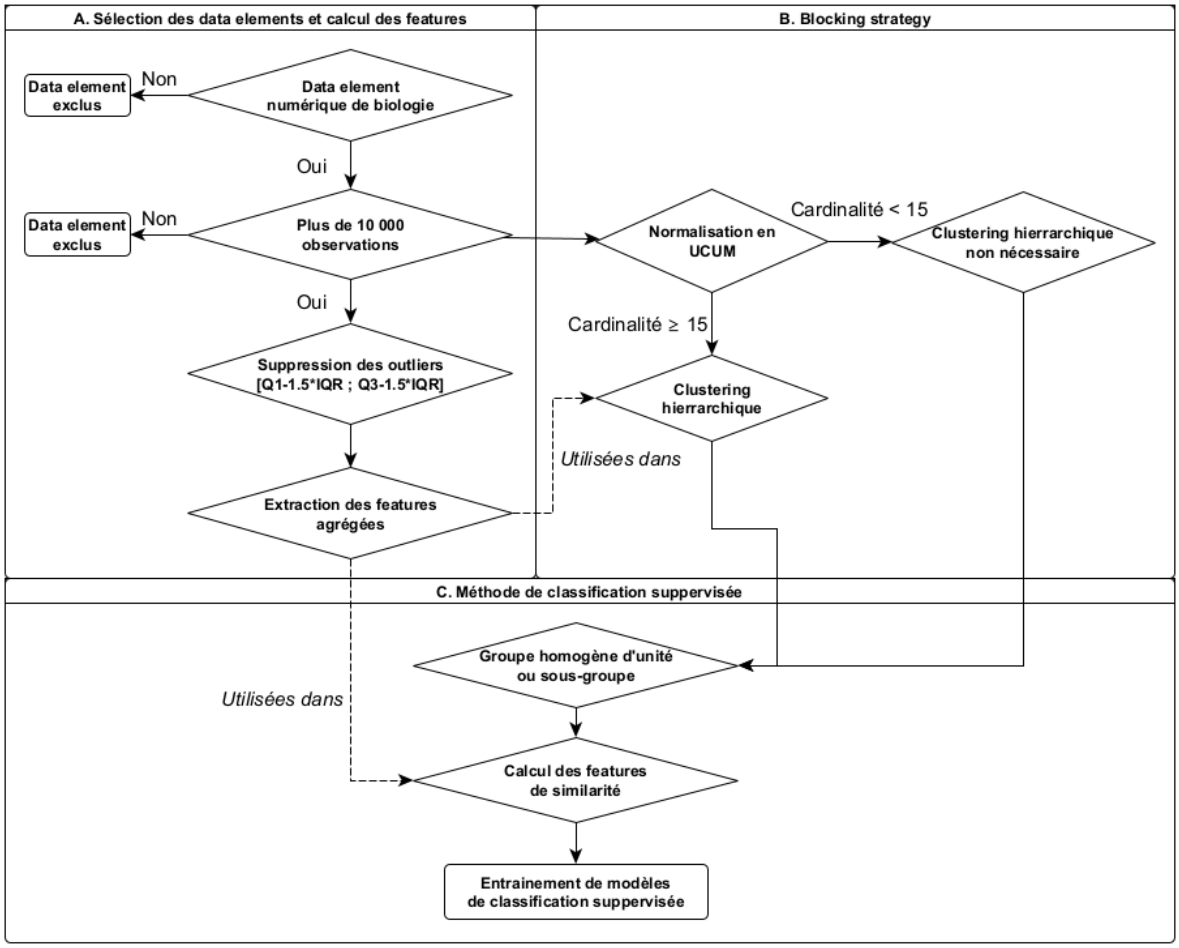


FIGURE 2.2 – Méthode d'alignement des data elements de biologie numérique en trois étapes

Au niveau de l'étape A., les data element avec un nombre d'observations disponibles important ($> 10\,000$) sont sélectionnés et des features agrégées sont extraites. L'étape B. correspond à une étape de blocking strategy permettant de regrouper les data element en cluster sur la base de leur unité et / ou d'une étape de clustering hiérarchique. L'étape C. correspond à l'entraînement de modèles de classifications supervisés sur la base des paires de data element possibles au niveau de chacun des clusters obtenus à l'étape B.

calculées uniquement pour le sous-ensemble des observations qui étaient incluses dans l'intervalle suivant :

$$Observations \in [Q_1 - 1,5 * IQR; Q_3 + 1,5 * IQR]$$

avec :

— $Q_1$ correspondant au premier quartile de distribution des données numériques ;
— $Q_3$ correspondant au troisième quartile de distribution des données numériques ;
— $IQR$ correspondant à l'intervalle inter-quartile, c'est-à-dire à $Q_3 - Q_1$ ;

Cette étape de filtrage des observations a été réalisée pour chacun des data elements (l'intervalle définissant quelles observations devaient être conservées était différent d'un data element à l'autre).

Sur la base de ces observations filtrées, les features agrégées suivantes ont été calculées :

— **Des indicateurs de position** : la moyenne, la médiane, le mode, le minimum, le maximum, les 1$^{er}$ et 3$^{e}$ quartiles, les 5$^{e}$ et 95$^{e}$ percentiles, les déciles ;
— **Des indicateurs de dispersion** : écart-type, variance, étendue, intervalle inter-quartile, coefficient de variation ;
— **Des indicateurs d'utilisation** : pourcentage de résultats anormaux, pourcentage de résultats au-dessus de la borne supérieure de normalité, pourcentage de résultats en-dessous de la borne inférieure de normalité, pourcentage de résultats produits en journée (entre 8h et 20h), pourcentage de résultats produits dans la nuit (entre 20h et 8h).

Par ailleurs, la **distribution échantillonnée** des données rattachée à chacun des data elements a été calculée (distribution calculée avec un échantillonnage de 100 intervalles).

Dans le contexte de cette tâche de entity linkage, il est nécessaire de calculer des métriques de similarité entre différentes paires de data elements. Afin de réduire la dimension des paires pour lesquelles ces métriques ont été calculées, la deuxième étape de la méthode d'alignement proposée correspond à une **blocking strategy**.

### 2.2.2 *Blocking strategy*

Cette partie est décrite au niveau de la Figure 2.2B.

Dans l'objectif de réaliser des alignements entre les data elements de chaque source et au sein de chaque source (un même concept de biologie pouvant être représenté par différents data elements au sein d'une même source), le nombre de paires de data elements

maximum $N_{paires}$ vaut :

$$N_{paires} = \frac{n\ (n-1)}{2}$$

où $n$ correspond à la somme du nombre de data elements de chaque source.

Au vu de la cardinalité (nombre de data elements) de chaque source, il est indispensable de réduire la dimension des comparaisons réalisées via une *blocking strategy*. En traitement automatique des langues, et plus particulièrement dans la tâche de entity linkage, la *blocking strategy* correspond à une étape de découpage de l'espace de recherche en « blocs » d'entités similaires. Seules les entités appartenant à un même bloc sont comparées entre elles, réduisant grandement la complexité en terme de calcul [153].

La *blocking strategy* que nous avons mise en place est basée sur deux étapes successives :

— Regroupement des data elements rattachés à une même unité de mesure. Afin de prendre en compte les variations possibles dans la façon dont sont disponibles les unités de mesure, une étape préalable de normalisation des unités selon la terminologie UCUM [2] [154] a été réalisée. A l'issue de cette première étape, nous obtenons des groupes homogènes de data elements en ce qui concerne l'unité de mesure.

— Découpage des groupes homogènes de data elements lorsque la cardinalité du groupe était supérieure à 15 data elements. Pour cela, une méthode d'apprentissage non supervisé par clustering hiérarchique (*hierarchical clustering* en anglais) a été utilisée. A l'issue de cette seconde étape, nous obtenons des groupes homogènes du point de vue de l'unité de mesure et avec une cardinalité plus faible.

C'est au sein de ces groupes de data elements que sont ensuite réalisés les alignements entre paires de data elements. Différents modèles d'apprentissage machine supervisé ont été entraînés à partir des features de similarité calculées entre les différentes paires d'un même groupe.

### 2.2.3 Identification des mappings par apprentissage machine supervisé

Cette partie est décrite au niveau de la Figure 2.2C.

2. UCUM est un système proposant une façon non ambiguë de représenter les unités de mesure.

#### *2.2.3.1 Calcul des métriques de similarité*

Au sein de chacun des sous-groupes obtenus suite à l'étape de *blocking strategy*, des métriques de similarité ont été calculées entre chacune des paires possibles. Les métriques de similarité calculées sont les suivantes :

— La valeur absolue de la différence $D$ entre la valeur de la *feature* $A$ disponible pour le data element 1 ($F_{A_1}$) et la valeur de la *feature* $A$ disponible pour le data element 2 ($F_{A_2}$) :

$$D_{F_{A_1},F_{A_2}} = |F_{A_1} - F_{A_2}|$$

— La différence relative $DR$ entre la valeur de la *feature* $A$ disponible pour le data element 1 ($F_{A_1}$) et la valeur de la *feature* $A$ disponible pour le data element 2 ($F_{A_2}$) :

$$DR_{F_{A_1},F_{A_2}} = \frac{|F_{A_1} - F_{A_2}|}{max(|F_{A_1}|, |F_{A_2}|)}$$

— Le pourcentage de recouvrement de l'aire sous la courbe de deux distributions échantillonnées $D_1$ et $D_2$.

Au total, 67 mesures de similarité différentes ont été calculées.

#### *2.2.3.2 Entraînement des modèles de classification des paires de data elements (apprentissage supervisé)*

Un *gold standard* contenant des alignements entre des data elements de biologie et des concepts LOINC a été constitué sur la base des alignements déjà disponibles au sein de l'EDS du CHU de Bordeaux et enrichi manuellement. Au total, 3 669 alignements étaient disponibles. Ces alignements étaient disponibles vers des codes LOINC décrivant :

— Le « composant » (ex : 'Sodium') ;
— Le « système » (ex : 'Ser/Plas' pour sérum ou plasma) ;
— Le « type de propriété » (ex : 'SCnc' pour concentration de substance) ;
— Le « type d'échelle » (ex : 'Qn' pour quantitatif) ;

Quand les informations d'« aspect temporel » ou de « type de méthode » étaient disponibles, le code était simplifié à la combinaison des quatre composants principaux ci-dessus.

Ces alignements ont été utilisés pour construire un jeu de données pour l'apprentissage supervisé (Figure 2.3) :

— Lorsqu'un code LOINC était aligné vers plusieurs data elements, l'ensemble des data elements impliqués ont été utilisés pour créer des paires de data elements. Seule une paire par couple de data elements était calculée (c'est-à-dire que si la paire $D_A \leftrightarrow D_B$ était calculée, la paire $D_B \leftrightarrow D_A$ ne l'était pas). Ces paires correspondent aux « **paires positives** » ;

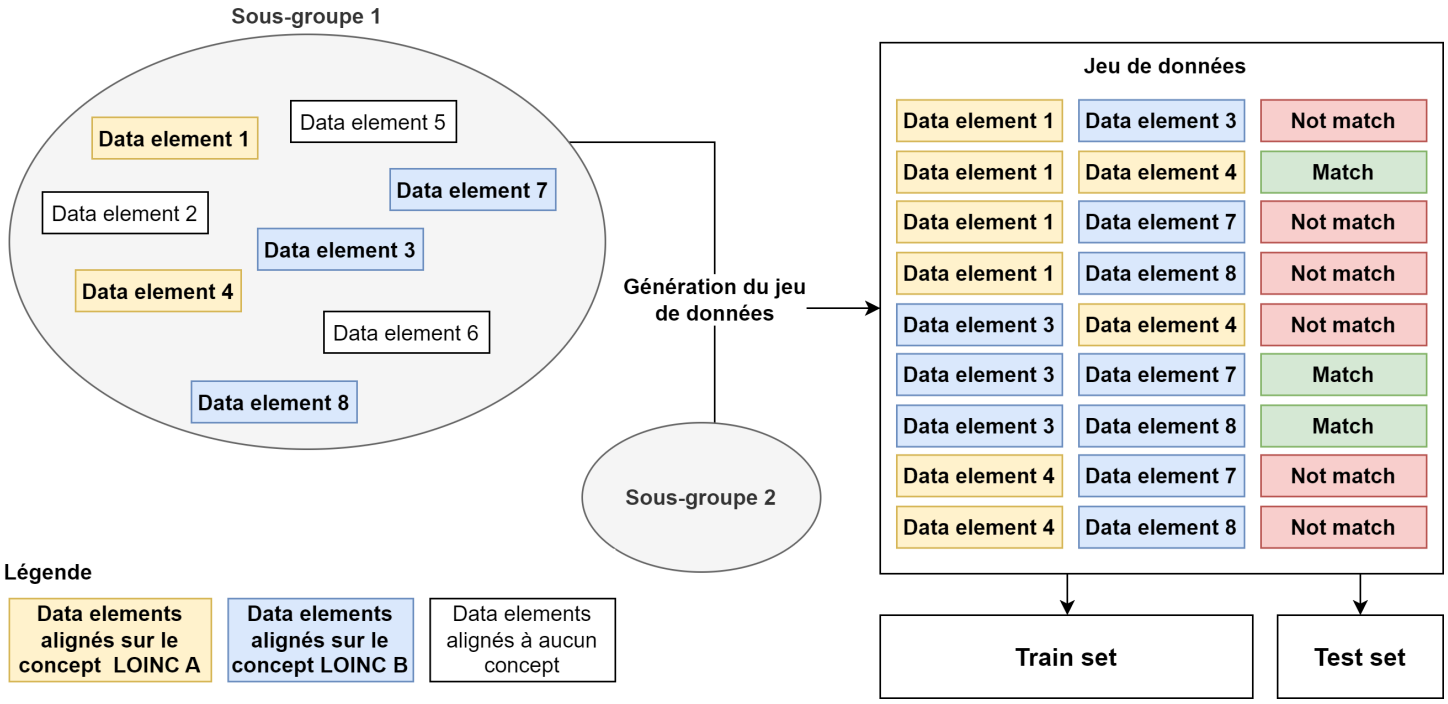


FIGURE 2.3 – Constitution du jeu de données pour la méthode d'alignement des data elements de biologie

— Au sein d'un même groupe (obtenu après l'étape de *blocking strategy*), lorsqu'un data element était aligné vers un code LOINC, l'ensemble des paires possibles entre ce data element et les autres data elements du même groupe mais rattachés à un autre code LOINC ont été calculées. Ces paires correspondent aux « **paires négatives** ».

Ce jeu de données a été découpé en un jeu d'entraînement (75%) et un jeu de test (25%). Le jeu d'entraînement a été utilisé pour entraîner différents modèles de classification supervisé, parmi lesquels une régression logistique, un réseau de neurones, une machine à vecteurs de support (*Support Vector Machine* ou SVM en anglais) et des forêts aléatoires (*Random Forest* ou RF en anglais). La liste des différents modèles entraînés est présentée au niveau des résultats (Tableau 2.3).

Les hyperparamètres des modèles ont été optimisés à partir du jeu d'entraînement. Afin d'éviter un sur-apprentissage lors de l'optimisation des hyperparamètres, une méthode de validation croisée (*cross-validation* en anglais) K-fold avec N=5 a été utilisée. Après optimisation des hyperparamètres, les différents modèles optimisés pour leur hyperparamètres ont été entraînés sur l'ensemble du jeu de données d'entraînement et évalués sur le jeu de données de test.

### 2.2.4 Évaluation

#### *2.2.4.1 Évaluation de l'étape de blocking strategy*

L'étape de *blocking strategy* a pour objectif de réduire le nombre de paires de data elements pour lesquelles il sera nécessaire de calculer des mesures de similarité. Cette réduction ne doit cependant pas séparer des data elements qui correspondent en réalité au même concept. Ainsi, la *blocking strategy* a été évaluée sur :

— Sa capacité à réduire le nombre de paires de data elements calculées en déterminant le nombre de paires calculées pour l'étape de classification et le nombre de data elements moyen par sous-groupe.

— Sa capacité à ne pas séparer des data elements qui correspondent en réalité au même concept LOINC. Cette évaluation a été réalisée sur la base du gold standard construit manuellement. Au sein du gold standard, le pourcentage de concepts LOINC mappés à au moins deux data elements et présents dans un seul cluster après l'étape de *blocking strategy* (c'est-à-dire que les data elements ont été correctement regroupés) a été calculé.

#### *2.2.4.2 Évaluation des différents modèles d'apprentissage supervisé*

L'évaluation des modèles de classification a été réalisée sur la base :

— Du **rappel** (*recall* en anglais) : il correspond à la capacité de l'algorithme à identifier l'ensemble des paires de data elements positives. Le rappel correspond à la sensibilité ;

— De la **précision** (*precision* en anglais) : il correspond à la proportion de vraies paires de data elements parmi les paires classées comme positives par l'algorithme. La précision correspond à la valeur prédictive positive ;

— De la **F-mesure** (également appelée F1) : elle correspond à la moyenne harmonique de la précision et du rappel :

$$F_1 = 2 * \frac{(Précision * Rappel)}{(Précision + Rappel)}$$

En plus de ces métriques, la courbe Rappel-Précision ainsi que son aire sous la courbe (*Area Under the Curve* ou AUC en anglais) ont été calculés.

## 2.3 Résultats

### 2.3.1 Description des données de biologie disponibles dans l'EDS du CHU de Bordeaux

Les données de biologie intégrées au sein de l'EDS du CHU de Bordeaux correspondaient à 793 442 947 observations représentées par 162 612 data elements (Figure 2.4). Les résultats de biologie numériques correspondaient à 582 808 292 observations (73,45%), représentées par 71 474 data elements (43,95%). Après filtrage des data elements avec au moins 10 000 observations, il restait 2 202 data elements (3,08%) représentant 569 150 002 observations numériques (97,66%). Parmi ces data elements, 338 (15,35%) n'étaient pas associés à une unité (représentant 26 850 086 observations). Les autres data elements étaient associés à 238 unités différentes.

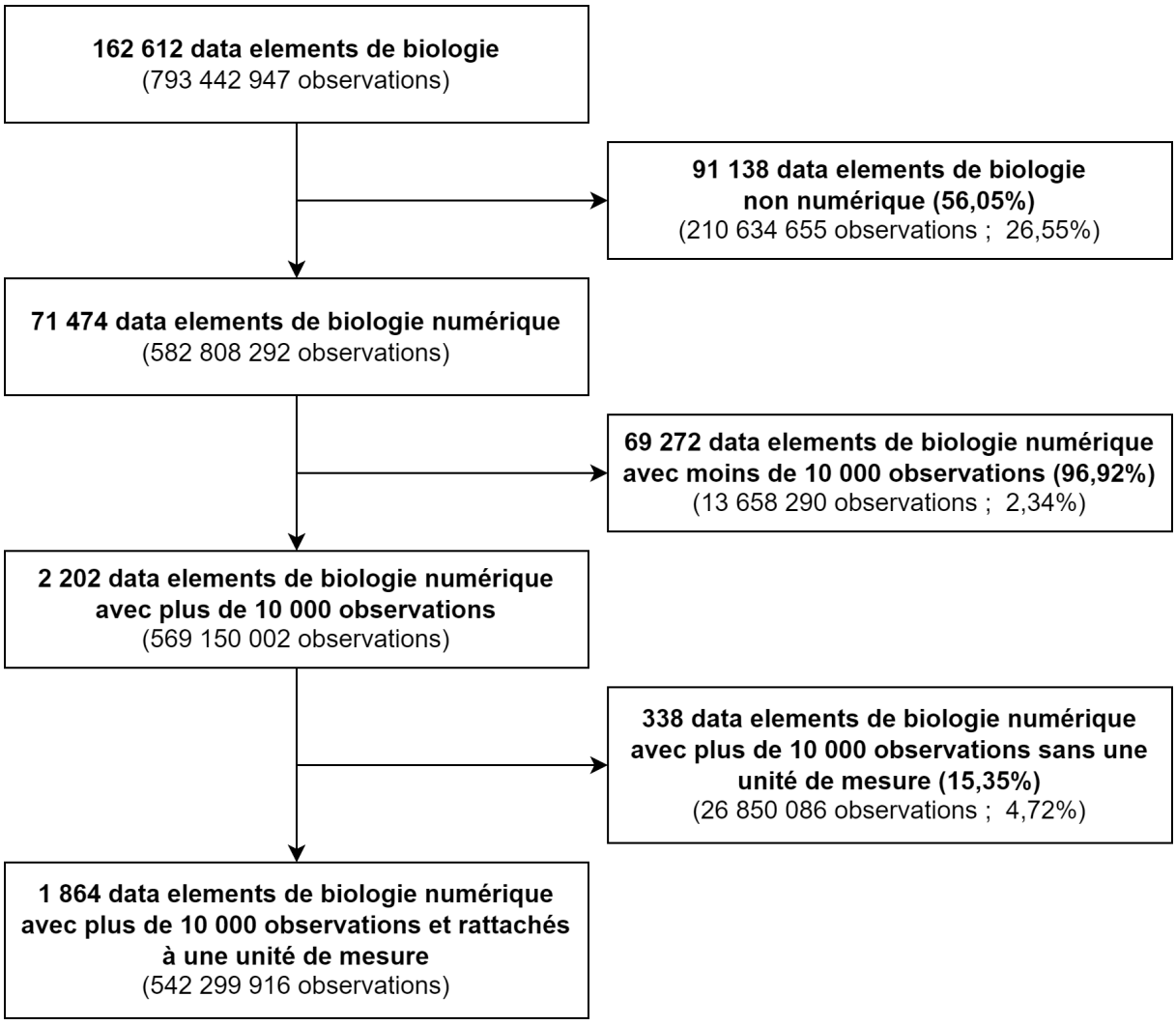


FIGURE 2.4 – Diagramme de flux des différents data elements de biologie

Le Tableau 2.1 présente la volumétrie des observations et des data elements pour chacune de ces trois sources. DxCare® est la source de données avec le plus grand nombre d'observations (on rappelle que DxCare® intègre les deux LIMS), avec un nombre d'observations correspondant à peu près à la somme des observations disponibles dans les deux autres sources (51,74% des observations). DxCare® est aussi la source présentant le plus d'hétérogénéité en terme de data elements (66,21% des data element).

Tableau 2.1 – Métriques décrivant les données de biologie disponibles dans l'Entrepôt de Données de Santé du CHU de Bordeaux

| **Source** | **Nombre d'observations** | **Nombre de data elements** |
|---|---|---|
| TD-Synergy® | 178 064 102 | 27 350 |
| Glims® | 204 856 427 | 42 127 |
| DxCare® | 410 522 418 | 136 184 |
| **Total** | **793 442 947** | **205 661** |

### 2.3.2 Impact de la *blocking strategy*

Après normalisation des unités en UCUM (Tableau 2.2), les data elements du jeu de données ont été regroupés en 59 groupes d'unités homogènes (six data elements n'ont pas pu être normalisés suivant UCUM, représentant 84 593 observations). La cardinalité moyenne au sein de ces groupes d'unités homogènes était de 37,34 (écart-type=86,00). 39 groupes d'unités homogènes avaient une cardinalité inférieure ou égale à 15 data elements et 20 groupes d'unités homogènes avaient une cardinalité supérieure à 15 data elements (moyenne=99,60 ; écart-type=127,95 ; maximum=413 pour l'unité de mesure « % »).

Le clustering hiérarchique a été réalisé sur les 20 groupes d'unités homogènes avec une cardinalité supérieure à 15 (Tableau 2.2). Il a permis de générer 126 clusters avec une cardinalité moyenne de 15,81 data elements (écart-type = 46,94 ; maximum=410).

Tableau 2.2 – Cardinalités (nombre de data elements) avant et après le clustering hiérarchique

| | Cardinalité après normalisation en UCUM[a]> 15 | | | |
|---|---|---|---|---|
| | Non | Oui, nécessité d'un HC[b] | | |
| | | Avant le HC[b] | Après le HC[b] | Total après la BS[c] |
| **N** | 39 | 20 | 126 | 165 |
| **Moyenne (e.t.)** | 5,41 (4,11) | 99,60 (127,95) | 15,81 (46,94) | 13,35 (41,27) |
| **Médiane [IQR]** | 5 [2 ; 9] | 40 [32 ; 92] | 3 [1 ; 9] | 3 [2 ; 9] |
| **Min ; Max** | 1 ; 15 | 18 ; 413 | 1 ; 410 | 1 ; 410 |

[a] *Unified Code for Units of Measure.*
[b] *Hierarchical clustering.*
[c] *Blocking strategy.*

Sur la base des 2 202 data elements de biologie numérique associés à plus de 10 000 observations, le nombre maximal de paires calculables était de 2 423 301. Suite à la *blocking strategy*, le nombre de paires de data elements à calculer était de 153 178 paires, soit une réduction de la complexité de 93,67%.

Parmi les 2 202 data elements de biologie numérique associés à plus de 10 000 observations, 1 166 (52,95%) étaient rattachés à un code LOINC dans le gold standard. À partir du gold standard, on retrouve que 94,36% des concepts associés à plusieurs data elements étaient regroupés dans un seul cluster à l'issue de l'étape de *blocking strategy*.

### 2.3.3 Identification des mappings par apprentissage machine supervisé

Les mesures de similarité entre chacune des paires de data elements ont été calculées au sein des 126 clusters issus du clustering hiérarchique et les 39 groupes d'unités homogènes avec une cardinalité inférieure à 15 data elements et pour lesquels les data elements étaient liés à un code LOINC dans le gold standard. Ainsi, les mesures de similarité ont été calculées pour 40 984 paires de data elements constituant le jeu de données pour l'apprentissage supervisé.

Le jeu de données a ensuite été découpé en :

— Un jeu de données d'entraînement, comprenant 29 015 paires de data elements, dont 2 542 paires positives (8,76%).

— Un jeu de données de validation, comprenant 9 672 paires de data elements, dont 872 paires positives (9,0%).

Sur la base du jeu de données d'entraînement, différents modèles de classification ont été entraînés. Une première étape d'optimisation des hyperparamètres a été effectuée sur la base de l'échantillon d'apprentissage par une méthode de cross-validation (K-fold, N=5).

Chacun des modèles optimisés pour leurs hyperparamètres a ensuite été entraîné sur la base de l'échantillon d'apprentissage avant d'être évalué sur l'échantillon de validation. Les résultats des modèles de classification sont présentés dans le Tableau 2.3 et les courbes de Rappel-Précision ($AUC_{PR}$) sont présentées dans la Figure 2.5.

TABLEAU 2.3 – Résultats des modèles de classification pour la tâche de entity linkage entre deux data elements

| | **Rappel** | **Précision** | **F-mesure** | $AUC_{PR}$ |
|---|---|---|---|---|
| **Régression logistique** | 0,510 | 0,673 | 0,581 | 0,679 |
| **k-Nearest Neighbors** | 0,820 | 0,824 | 0,822 | 0,830 |
| **Support-Vector Machine** | 0,692 | 0,802 | 0,743 | 0,819 |
| **Réseau de neurones** | 0,812 | 0,791 | 0,801 | 0,872 |
| **Arbre de décision** | 0,725 | 0,780 | 0,751 | 0,772 |
| **Extra-Trees** | 0,817 | 0,870 | 0,843 | 0,934 |
| **Forêt aléatoire** | **0,827** | **0,875** | **0,850** | **0,942** |
| **Gradient boosting** | 0,813 | 0,892 | 0,850 | 0,942 |

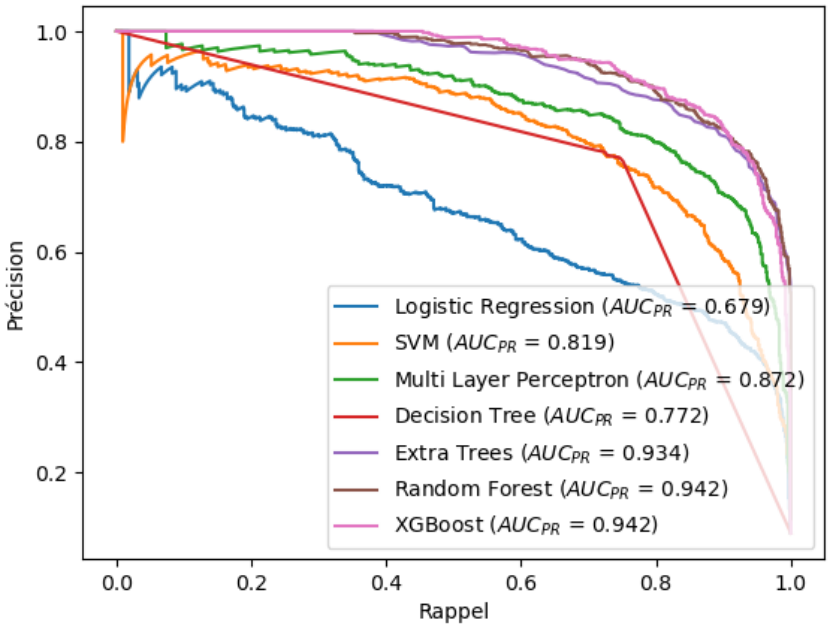


FIGURE 2.5 – Courbe Rappel-Précision ($AUC_{PR}$) des modèles de classification entraînés

Le modèle avec forêts aléatoires et celui avec Gradient Boosting ont donné des résultats similaires, avec une F-mesure à 0,850. S'agissant de diminuer les faux négatifs, nous avons choisi de privilégier le modèle avec le rappel le plus haut. Ainsi, les performances étaient les meilleures avec forêt aléatoire, avec un rappel à 0,827 et une précision à 0,875. La matrice de confusion de la forêt aléatoire est présentée dans la Figure 2.6. La courbe Rappel-Précision de la forêt aléatoire est présentée Figure 2.7, où on retrouve un $AUC_{PR} = 0,942$.

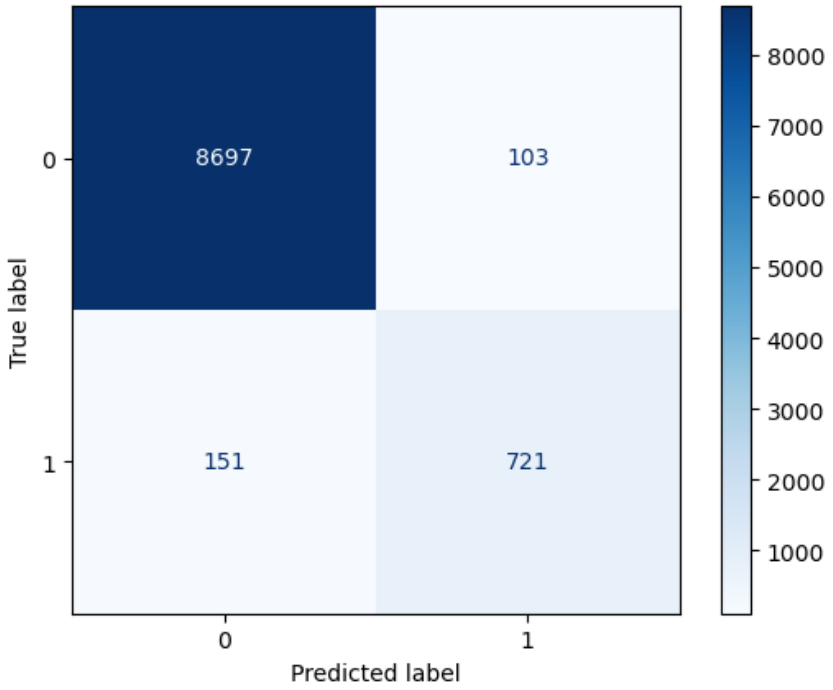


FIGURE 2.6 – Matrice de confusion du modèle avec forêt aléatoire

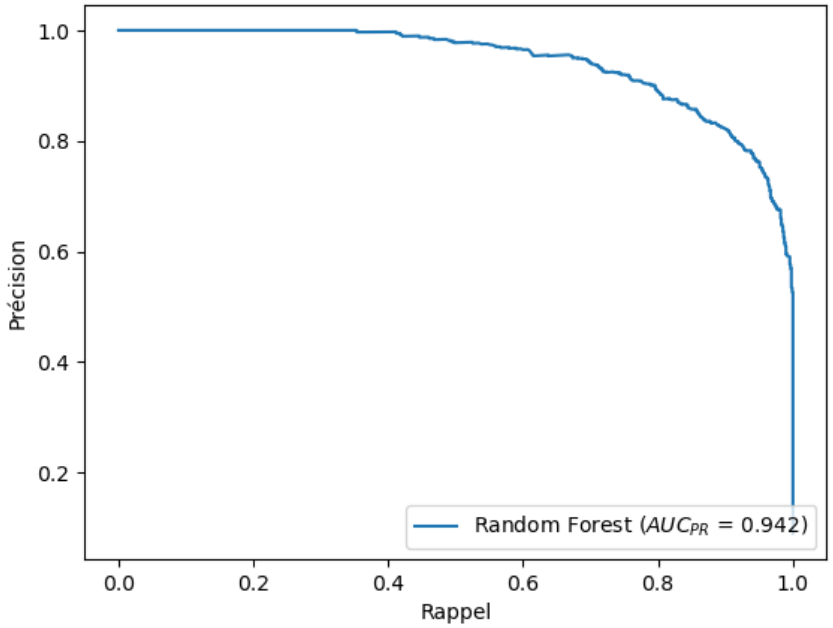


FIGURE 2.7 – Courbe Rappel-Précision ($AUC_{PR}$) de la forêt aléatoire

La Figure 2.8 présente l'importance de chacune des features dans le résultat du modèle avec forêt aléatoire. On retrouve parmi les features les plus informatives : la différence relative du 3e quartile (diffq3relative), le pourcentage de recouvrement calculé sur la base des deux distributions échantillonnées (densitymatch) et la différence relative de la moyenne (diffmoyrelative).

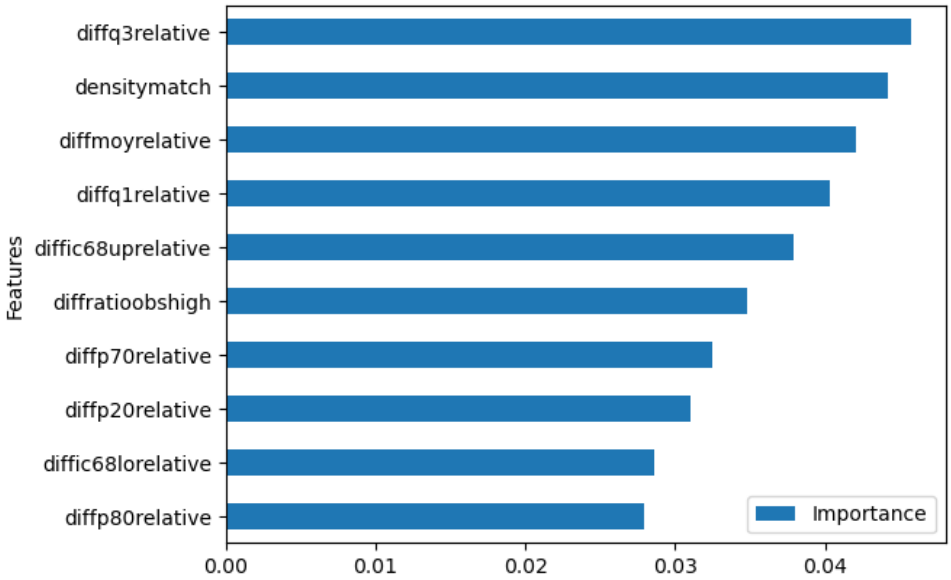


FIGURE 2.8 – Importance des features dans le modèle avec Forêt aléatoire

Dans le cas où le modèle de classification serait utilisé comme outil de triage, il serait acceptable que 50% des paires évaluées soient correctes (et donc que 50% des paires évaluées soient fausses). En fixant une précision minimum à 50%, le rappel du modèle avec forêt aléatoire est maximal (c'est-à-dire 1). La matrice de confusion correspondant à ce seuil est présentée en Figure 2.9.

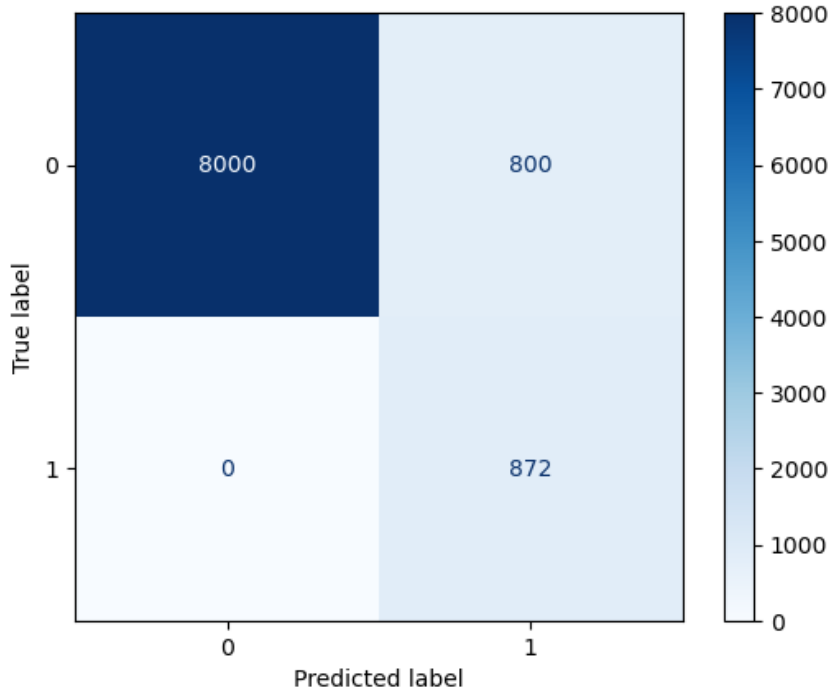


FIGURE 2.9 – Matrice de confusion du modèle avec forêt aléatoire - Précision fixée à 0,5 (seuil du modèle à 0,03)

## 2.4 Discussion

La LOINC est la terminologie qui a été choisie comme standard pour la biologie médicale dans le cadre du Ségur du numérique en santé [155]. L'alignement entre les terminologies locales de biologie et la LOINC représente une lourde charge [144], justifiant l'intérêt de développer des processus automatiques pour aider aux alignements. De plus, malgré la démocratisation de l'utilisation de la LOINC en routine en France, ces méthodes d'alignement automatique restent intéressantes pour aligner la biologie historique.

### 2.4.1 Apports

Dans le cadre de ce travail, nous proposons une méthode d'alignement des data elements de biologie numérique sur la base de features agrégées anonymes (méthode basée sur les instances). La méthode mise en œuvre comporte une étape de *blocking strategy* non supervisée en amont d'une tâche de classification supervisée.

#### *2.4.1.1 Blocking strategy*

L'étape de *blocking strategy* permet de réduire la dimension des comparaisons à réaliser lors de l'étape de mapping (réduction de 97,3% du nombre de comparaisons à réaliser) avec une perte raisonnable de paires candidates positives du gold standard (6,3%).

En modifiant la hauteur d'élagage de l'arbre généré par le clustering hiérarchique (c'est-à-dire, la *height*), il est possible de diminuer la proportion des mappings candidats positifs séparés dans différents clusters. Si l'on choisit une hauteur d'élagage qui n'entraîne pas de séparation de mappings candidats positifs du gold standard dans différents clusters, le nombre de paires à calculer est alors de 251 837 (soit 64,40% paires en plus par rapport à la valeur utilisée dans le cadre de ce travail). On note que le nombre de paires est ici, malgré un seuil plus sensible, beaucoup moins élevé que le nombre théorique de paires à calculer qui est de 2 423 301 pour les 2 202 data elements de biologie numérique disponibles. Cela traduit le fait que l'élément le plus important de la *blocking strategy* est l'utilisation des unités normalisées en UCUM comme critère de séparation des data elements en différents clusters.

Ainsi, si on utilise uniquement les unités normalisés en UCUM comme critère pour créer des clusters de data elements, le nombre de paires à calculer est de 254 177, soit 89,51% de réduction par rapport au nombre maximal théorique de paires à calculer. De plus, en l'absence de normalisation des unités de mesure en UCUM, le nombre de paires à calculer sur la base d'un clustering basé sur les unités telles que disponibles serait plus faible, de 214 002 paires (différence de 40 175 paires). Cela traduit le fait que la normalisation permet de regrouper entre eux des data elements qui présentent en réalité la même unité de mesure alors que cette dernière était disponible avec des verbatims différents.

#### *2.4.1.2 Classification des paires candidates*

L'étape de classification des paires candidates retrouve de bonnes performances avec une F-mesure de 0,861 sur l'échantillon d'évaluation pour la forêt aléatoire, qui est le modèle ayant donné les meilleurs résultats pour notre tâche. Une limite importante de l'évaluation est que cette dernière a été réalisée sur la base d'un jeu de données de validation interne. Les performances obtenues sont donc très probablement sur-évaluées par rapport à celles qu'on aurait pu obtenir dans le cadre d'une validation externe, qui devra donc être réalisée.

Également, cette méthode d'alignement ne concerne que les data elements de biologie numérique avec plus de 10 000 observations disponibles (soit, au CHU de Bordeaux, uniquement 2 202 data elements). Le seuil de 10 000 observations a été choisi arbitrairement pour limiter le risque qu'une feature agrégée calculée ne représente un trop faible nombre d'individus (et donc qu'elle ne soit plus, à proprement parler, anonyme).

### 2.4.2 Comparaison avec des travaux antérieurs

De nombreuses études se sont intéressées aux mappings semi-automatiques des data elements de biologie vers la LOINC, sur la base des libellés des data elements.

Nikiema *et al.* [156, 157] proposent une modélisation de la LOINC sous forme de graphe à partir duquel les auteurs ont réalisé des alignements automatiques par décomposition / recomposition des libellés des data elements disponibles. Pour le sous-ensemble des data elements pour lequel un mapping 1-1 vers la LOINC était retrouvé, la précision évaluée était de 0,59.

Michel-Pique *et al.* [158] tirent partie de la structuration de la LOINC pour réaliser des alignements sur la base de *classifier chains* [159]. Les résultats de la prédiction d'un attribut du code LOINC étaient utilisés en entrée de la prédiction d'un autre attribut. Cette méthode de classification multi-labels a résulté en une F-mesure de 0,87, comparable à la nôtre.

Kelly *et al.* [160] utilisent des approches d'apprentissage supervisé (classification multi-labels) pour réaliser des alignements sur la base des chaînes de caractères représentant les data elements. Les chaînes de caractères correspondant aux data elements à aligner sont encodées en vecteurs sur la base d'une table de mapping token-LOINC, constituée à partir du jeu de données d'entraînement. Les meilleures performances retrouvaient une F-mesure de 0,943, dans un sous-ensemble restreint de codes LOINC utilisés dans le domaine de la cancérologie.

A la différence des approches décrites ci-dessus, la méthode proposée ici ne fonctionne que pour le sous-ensemble des data elements numériques de biologie. Cependant, elle offre

comme avantage de permettre l'alignement entre différents data elements, sans forcément chercher à réaliser un alignement vers un niveau conceptuel LOINC. De plus, la méthode que nous avons mise en œuvre pourrait être étendue à des data elements numériques hors du champ de la biologie (température, poids, taille, etc.).

Plus que de les opposer, il semble intéressant d'envisager de combiner l'approche basée sur les instances développées ici avec les approches basées sur le verbatim des data elements et d'en évaluer l'apport.

### 2.4.3 Perspectives

L'intérêt de méthodes basées sur des features agrégées totalement anonymes est qu'elles permettent de préserver la protection des données personnelles : les features agrégées ainsi que le modèle de classification pourraient être partagés sans risque du point de vue de la protection des données à caractère personnel. Le partage de ces éléments en *open-data* pourrait permettre d'aider d'autres établissements de santé à aligner leurs data elements de biologie numérique locaux à la LOINC.

Par exemple, nous pourrions partager les features agrégées rattachés aux concepts LOINC connus dans notre gold standard. Des établissements de santé pourraient ainsi calculer localement ces mêmes features puis, après une étape de *blocking strategy*, ils pourraient calculer les mesures de similarité entre leurs data elements locaux et les features partagées agrégées au niveau conceptuel de la LOINC. Ils seraient alors en mesure de conduire un alignement vers la LOINC sur la base du modèle de classification partagé ou sur la base d'un modèle de classification ré-entraîné localement (via les mappings déjà disponibles localement avec la LOINC qui seraient communs avec ceux partagés).

Plus largement, le partage des features agrégées, indépendamment de la disponibilité d'un alignement vers la LOINC, reste intéressant (ex : clustering non supervisé pour identifier des data elements communs entre établissements dans le cadre d'une étude multicentrique en utilisation secondaire).

# CHAPITRE 3

## Persistance du modèle de données d'i2b2 dans une base de données Elasticsearch

## Sommaire

## 3.1 Introduction

L'un des obstacles à la réutilisation des données de santé issues des EHRs est leur grande hétérogénéité [161, 162]. Comme détaillé dans la section 1.2.2.3, on distingue généralement deux types d'hétérogénéité :

— L'**hétérogénéité syntaxique** qui correspond au fait que les données de santé peuvent être stockées dans différents formats (base de données, fichier texte, image, etc.) ou selon différents modèles de données (c'est-à-dire dans divers logiciels spécialisés).

— L'**hétérogénéité sémantique** qui fait référence à la manière dont les informations sont stockées. Par exemple, la natrémie d'un patient peut être stockée de manière structurée en utilisant un code LOINC [87] (ex : `2947-0`) ou un code de terminologie locale (ex : `BIO:NaSg`), ou bien en texte libre dans un compte rendu d'examen complémentaire.

Ces deux types d'hétérogénéité entraînent des difficultés dans l'utilisation secondaire des données de santé. Diverses stratégies ont été mises en place pour réduire l'hétérogénéité des données de santé, notamment via des méthodes de ML et des méthodes à base de règles [163] au moment de l'exploitation des données.

L'une des stratégies largement utilisée dans le monde pour cela consiste à intégrer les données avant toute analyse, en mettant en place des EDS [100, 102, 106, 107, 164]. Une description détaillée des EDS est proposée section 1.3

D'un point de vue technique, un EDS correspond à une base de données dédiée à l'utilisation secondaire des données de santé. Les EDS sont alimentés par un processus appelé ETL (*Extract, Transform and Load*) appliqué aux données, qui sont tour à tour :

— Extraites de différentes bases de données de production (base de données utilisées par les logiciels métiers au sein des systèmes d'information hospitaliers) ou répliquées ;

— Transformées dans un format compatible avec le modèle de données cible (incluant les étapes de pseudonymisation et de standardisation) ;

— Chargées dans un EDS dédié.

Plus qu'une simple base de données, les EDS doivent prendre en compte toutes les problématiques sous-jacentes à la réutilisation des données de santé, en particulier les aspects liés à la gouvernance, les aspects éthiques, la transparence, la confidentialité et la sécurité. Ces aspects seront décrits au travers du chapitre 4, qui décrit la plateforme EDS mise en place au niveau du CHU de Bordeaux.

Différents modèles d'intégration sont utilisés dans les EDS [106, 107, 164, 165, 166, 167, 168, 169]. Quel que soit le modèle d'intégration choisi, l'un des premiers cas d'usage dans l'utilisation secondaire des données de santé est de réaliser des études de

faisabilité [170] et de phénotypage [171, 172]. Les premières consistent à dénombrer, et les secondes à identifier, les patients éligibles aux études sur la base de critères cliniques et/ou biologiques. Pour réaliser cette tâche, les chercheurs interrogent les données des EDS afin d'identifier la population d'intérêt le plus précisément possible. La recherche d'information dans les EDS doit donc présenter les caractéristiques suivantes :

— Un bon rappel, c'est-à-dire que les requêtes effectuées doivent identifier la population d'intérêt de la manière la plus exhaustive possible ;
— Une bonne précision, c'est-à-dire que les patients identifiés doivent effectivement correspondre aux critères de recherche.

La définition de ces requêtes est un **processus itératif** : les résultats des premières requêtes de dénombrement ou la visualisation des premiers dossiers patients permettent d'identifier les éléments à prendre en compte pour affiner la requête. Ainsi, les **temps de réponse** aux requêtes doivent être très bons, de l'ordre de la seconde, pour que ce processus itératif soit possible et acceptable.

De plus, une part importante de l'information médicale est disponible en texte libre [102, 173], comme les comptes rendus d'hospitalisation ou de consultation, les comptes rendus d'imagerie ou les ordonnances de sortie. Des méthodes spécifiques de TAL sont dans ce cas nécessaires pour extraire les informations pertinentes à partir du texte libre. Après une première étape de nettoyage (pré-traitement) du texte (suppression des caractères spéciaux, normalisation, racinisation, etc.), des méthodes basées sur le ML et des méthodes à base de règles sont généralement utilisées pour cela.

Il est ainsi nécessaire d'utiliser des systèmes de stockage de données pour les EDS qui permettent une interrogation rapide des données, et si possible qui incluent des méthodes efficaces d'interrogation du texte libre. Les principaux modèles de données utilisés dans les EDS reposent sur une base de données relationnelle pour conserver les données. Les systèmes de gestion de bases de données relationnelles (*Relational DataBase Management System* ou RDBMS en anglais) sont particulièrement adaptés pour garantir l'intégrité des données [174] et offrent un langage d'interrogation simple et flexible, le SQL (*Structured Query Language*). Toutefois, les performances des RDBMS diminuent à mesure que le volume de données augmente [76]. En outre, les RDBMS sont très efficaces pour le stockage de données structurées mais sont moins adaptés au stockage de données semi-structurées ou non structurées telles que le texte libre [76].

Les bases de données NoSQL sont apparues dans les années 2000 dans le contexte du Web 2.0 et ont été largement utilisées dans des applications manipulant de gros volumes de données (*Big Data*) et pour les usages temps réel. Les bases de données NoSQL sont des bases de données dans lesquelles les données ne sont pas stockées de manière relationnelle. Il existe différents types de bases de données NoSQL, telles que les bases

de données clé/valeur, les bases de données orientées documents, les bases de données orientées colonnes et les bases de données graphes [175, 176].

Une différence essentielle entre les bases de données NoSQL et les bases de données relationnelles traditionnelles est que dans les bases de données NoSQL, la structure des tables n'est pas fixe (*schema-free* en anglais) : la structure des données stockées peut donc différer d'une instance à une autre [177]. Également, ces systèmes ne permettent pas de réaliser des jointures entre les structures de données [178, 179]. En contrepartie, les bases de données NoSQL sont facilement déployables de manière distribuée, permettant un *scaling* horizontal [1] en cas d'augmentation du volume de données, et offrent souvent de très bonnes performances à l'interrogation [75].

Au CHU de Bordeaux, un EDS au format i2b2 [106] a été mis en place à partir de 2018 (il fait l'objet d'une présentation détaillée dans le chapitre 4). Les données intégrées dans cet EDS sont persistées dans un RDBMS. Elasticsearch [180] étant une base de données orientée documents (NoSQL) spécialisée dans l'indexation et la recherche en texte libre, c'est un bon candidat pour la persistance des données des EDS, qui contiennent des données volumineuses dont beaucoup sont des données en texte libre.

Ainsi, l'objectif de ce chapitre est :

1. De décrire les modifications nécessaires pour persister des données cliniques intégrées au format i2b2 dans Elasticsearch ;
2. D'évaluer les performances d'une telle implémentation, sur la base de l'EDS du CHU de Bordeaux.

## 3.2 Matériel et méthode

### 3.2.1 Matériel

#### *3.2.1.1 i2b2 - Informatics for Integrating Biology and the Bedside*

Une description générale de la plateforme i2b2 est présentée dans l'introduction (section 1.3.1.1). Comme précisé précédemment, l'application i2b2 est développée selon une architecture de micro-services, appelés cellules (*i2b2 cell* en anglais). La Figure 3.1 décrit en détail deux des principales cellules d'i2b2 :

— Le ***Data Repository Cell*** (également appelé *Clinical Research Chart* ou CRC ; zone en bleu clair sur la Figure 3.1) ;
— L'***Ontology Management Cell*** ou ONT (zone en vert clair sur la Figure 3.1).

1. Quand les performances d'une base de données sont diminuées, on peut soit ajouter plus de puissance au serveur de base de données (CPU, mémoire, etc.) - on parle alors de *scaling* vertical - soit ajouter des serveurs de bases de données supplémentaires pour répartir la charge - on parle alors de *scaling* horizontal.

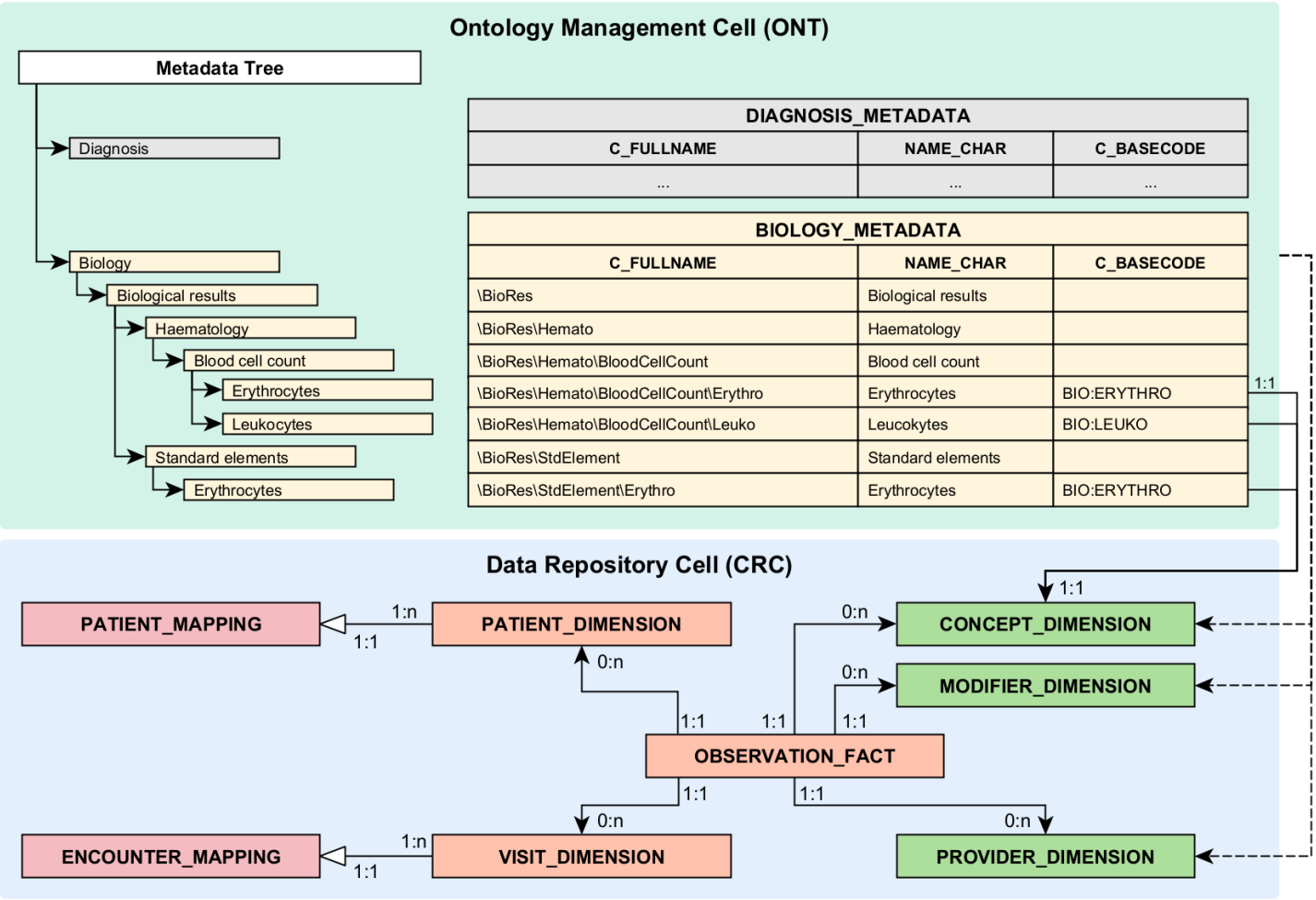


FIGURE 3.1 – *Data Repository Cell* (CRC) sur fond bleu et *Ontology Management Cell* sur fond vert (ONT) du modèle de données i2b2

Chacune des lignes de la table `BIOLOGY_METADATA` pour lesquelles un `C_BASECODE` est disponible est intégrée dans la table `CONCEPT_DIMENSION` (idem pour les tables de métadonnées qui alimentent les tables `MODIFIER_DIMENSION` et `PROVIDER_DIMENSION`). Les données contenues dans la table `OBSERVATION_FACT` sont rattachées à 1 patient décrit dans la table `PATIENT_DIMENSION`, 1 venue décrite dans la table `VISIT_DIMENSION`, 1 concept décrit dans la table `CONCEPT_DIMENSION`, 1 modifier de concept décrit dans la table `MODIFIER_DIMENSION` et 1 structure de soins productrice de l'observation décrite dans la table `PROVIDER_DIMENSION`. Le lien entre les données pseudonymisées et les données du Système d'Information Hospitalier se fait via les tables de mappings (`ENCOUNTER_MAPPING` et `PATIENT_MAPPING`).

### *Data Repository Cell* - CRC

Le CRC correspond à la cellule d'i2b2 en charge :

— D'exposer un service de requêtage des données intégrées dans i2b2. Ainsi, le CRC expose des fonctions permettant de dénombrer des patients / séjours correspondant à des critères clinico-biologiques ou encore des fonctions permettant de générer la liste de ces patients / séjours.

— De la persistance des données intégrées dans i2b2, ainsi que des données générées par le requêteur (requêtes et résultats des requêtes). Le CRC est basé sur un modèle de données en étoile [77] et est composé de :

1. Tables contenant les **données cliniques individuelles pseudonymisées** (en orange sur la Figure 3.1). Les données cliniques intégrées dans i2b2 sont persistées dans la table `OBSERVATION_FACT` sous forme d'observations. La table `OBSERVATION_FACT` est liée à deux tables de dimension, `PATIENT_DIMENSION` et `VISIT_DIMENSION`, décrivant respectivement les informations relatives aux patients (sexe biologique, date de naissance, etc.) et aux venues (date d'entrée, date de sortie, etc.).

2. Tables contenant la description des **valeurs possibles des colonnes de recherche** (*lookup column*) de la table `OBSERVATION_FACT` (en vert sur la Figure 3.1) :
   — La table `CONCEPT_DIMENSION` décrit les concepts utilisés pour caractériser le sens principal d'une observation. Par exemple, une observation rattachée à un `CONCEPT_CD = 'BIO:ERYTHRO'` fait référence à un résultat de biologie « Érythrocyte (en g/L) » ;
   — La table `MODIFIER_DIMENSION` décrit les valeurs possibles d'un niveau conceptuel secondaire, modifiant le niveau conceptuel principal. Par exemple, une observation rattachée au `CONCEPT_CD = 'BIO:ERYTHRO'` pourra être modifiée avec un `MODIFIER_CD = 'BIO:VALID'` (indiquant un « Résultat validé biologiquement ») ou avec un `MODIFIER_CD = 'BIO:INVALID'` (indiquant un « Résultat invalide biologiquement ») ;
   — La table `PROVIDER_DIMENSION` qui décrit les valeurs possibles des entités (soignants, unité de soins, etc.) qui sont à l'origine de l'observation. Par exemple, une observation rattachée à un `PROVIDER_ID = 'UNIT:RHUMATO'` a été produite par le service de « Rhumatologie ».

3. **Tables de mapping** permettant de gérer la correspondance entre les identifiants patients (`PATIENT_MAPPING`) et séjours (`VISIT_MAPPING`) utilisés dans i2b2 avec les identifiants dans la ou les sources intégrées dans i2b2 (en rouge sur la Figure 3.1).

### *Ontology Management Cell* - ONT

L'ONT correspond à la cellule d'i2b2 permettant l'intégration des « métadonnées » issues de différentes sources de données sous forme d'arbre hiérarchique. Les métadonnées sont intégrées dans différentes tables de métadonnées au sein de l'ONT, une pour chaque nœud racine de l'arbre de métadonnées (« Biologie », « Diagnostic », « Acte », etc.). Dans ces tables de métadonnées, chaque tuple correspond à un nœud de l'arbre hiérarchique. La clé primaire (`C_FULLNAME`) de ces tables correspond au chemin entre le nœud racine et le nœud courant (pour l'ensemble des nœuds intermédiaires et nœuds feuilles).

Cette organisation permet un accès multi-hiérarchique aux différents nœuds feuilles : un même nœud feuille de l'arbre peut être accessible par différents chemins hiérarchiques (au sein de la même table de métadonnées, ou bien au travers de différentes tables de métadonnées). Dans les tables de métadonnées, pour les nœuds feuilles uniquement, la colonne `C_BASECODE` contient le code utilisé comme valeur possible dans l'une des colonnes de recherche (ex : le `CONCEPT_CD`) de la table `OBSERVATION_FACT`. Un exemple de tuples pour la table de métadonnées `C_BASECODE` est fourni dans le Tableau 3.1.

Tableau 3.1 – Exemple de tuples présents dans la table `C_BASECODE` (dans le tableau à droite), en lien avec l'arbre hiérarchique correspondant (à gauche).

| **Arbre hiérarchique** | `C_FULLNAME` | `NAME_CHAR` | `C_BASECODE` |
|---|---|---|---|
| Résultat de biologie | \BioRes | Résultat de biologie | |
| ↪Hématologie | \BioRes\Hemato | Hématologie | |
| ↪NFS | \BioRes\Hemato\BloodCellCount | NFS | |
| ↪Erythrocyte | \BioRes\Hemato\BloodCellCount\Erythro | Erythrocyte | BIO:ERYTHRO |
| ↪Biologie standard | \BioRes\StdElements | Biologie Standard | |
| ↪Erythrocyte | \BioRes\StdElements\Erythro | Erythrocyte | BIO:ERYTHRO |

Le même nœud feuille `C_BASECODE = 'BIO:ERYTHRO'` est accessible au travers de différentes entrées dans l'arbre hiérarchique (fils de « NFS » d'une part, et fils de « Biologie standard » d'autre part). Chacun des chemins d'accès possibles est représenté par un `C_FULLNAME` différent. Dans la table `OBSERVATION_FACT`, c'est la colonne `C_BASECODE` qui est utilisée comme valeur possible dans les colonnes de recherche (ex : `CONCEPT_CD`).

En plus de donner un sens aux observations intégrées dans la table `OBSERVATION_FACT` du CRC, les métadonnées permettent de définir une vue de ces données grâce aux différents chemins d'accès aux valeurs possibles des colonnes de recherche (ex : `CONCEPT_CD`) qui contextualisent chaque observation. En outre, la séparation des données et des métadonnées dans deux cellules différentes permet de dissocier le chargement des données cliniques individuelles pseudonymisées du chargement des métadonnées décrivant ces données. Ainsi, en modifiant les métadonnées contenues dans l'ONT, il est possible de modifier la vue des données sans modifier les données elles-mêmes.

#### *3.2.1.2 Implémentation d'i2b2 au sein du CHU de Bordeaux*

Depuis 2018, le CHU de Bordeaux a mis en place un EDS s'appuyant sur i2b2. Cet EDS intègre les données de santé des patients venus au moins une fois à l'hôpital depuis 2010. Pour ces patients, les données disponibles depuis 2005 sont intégrées. Une description détaillée de l'EDS du CHU de Bordeaux est présenté dans le chapitre 4.

**Périmètre des données intégrées**

Les données provenant de plus de 20 sources de données sont chargées quotidiennement dans l'EDS du CHU de Bordeaux (le détail des sources est présenté dans la Figure 4.1 du chapitre 4). En août 2024, ces données représentent un total de :

— 2 502 063 patients distincts ;
— 20 982 497 venues distinctes ;
— 3 474 264 570 observations, dont 72 580 022 documents textuels.

**Stratégie de partitionnement de la table `OBSERVATION_FACT`**

Les données de l'EDS du CHU de Bordeaux sont persistées dans une base de données relationnelle Oracle. La table `OBSERVATION_FACT`, qui contient l'ensemble des données individuelles des patients pseudonymisées, a été partitionnée par source et par année (Figure 3.2). L'objectif de cette stratégie de partitionnement est de faciliter la mise à jour quotidienne des données et d'optimiser l'exécution des requêtes sur une grande quantité de données. Ainsi, plus de 600 partitions ont été créées. Ce partitionnement permet de charger les données de tout ou partie des sources de données, pour une ou plusieurs années. Grâce à cette organisation, les données de l'année en cours peuvent être rechargées quotidiennement tandis que les données historiques sont rechargées moins fréquemment.

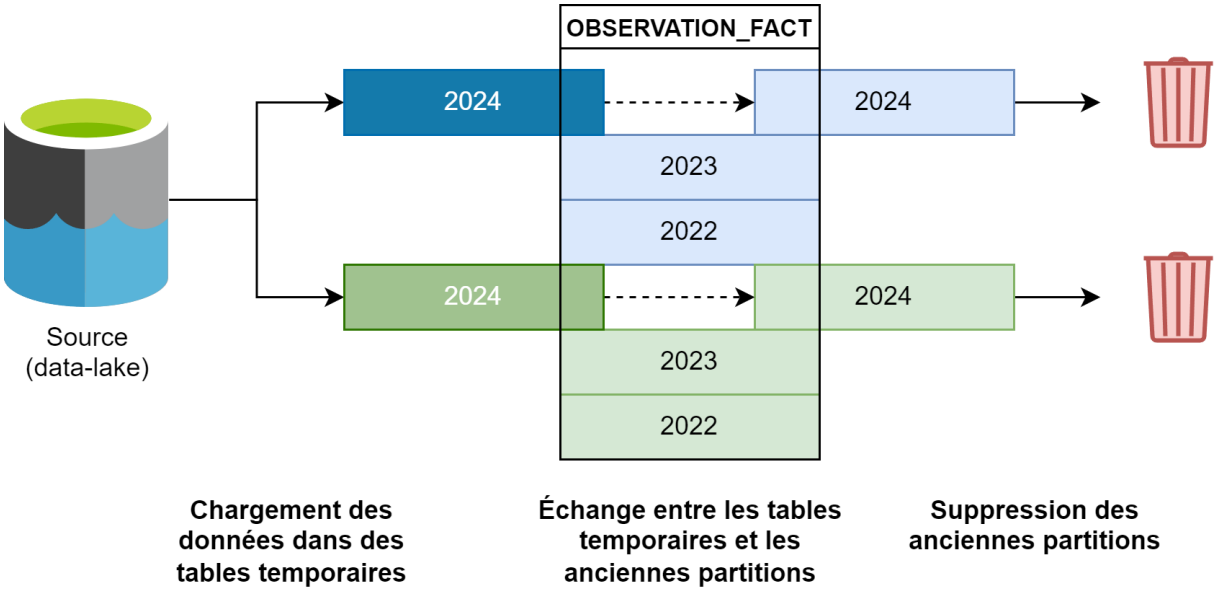


FIGURE 3.2 – Stratégie de partitionnement mise en œuvre au CHU de Bordeaux.

Les données de différentes sources (représentées par les différentes couleurs) sont extraites quotidiennement depuis les bases de données répliquées (*data-lake*) dans différentes tables intermédiaires (couleurs sombres). Au cours de l'ETL, seul un sous-ensemble des sources de données est extrait, sur la base de la temporalité des observations. Une fois que toutes les tables intermédiaires sont prêtes, les anciennes partitions sont successivement remplacées par les nouvelles, puis détruites. Ainsi, ce processus permet de recharger une sous-partie des données de la table `OBSERVATION_FACT`, tout en limitant le temps d'indisponibilité des données.

#### *3.2.1.3 Elasticsearch*

Elasticsearch [181] est une base de données NoSQL distribuée orientée documents, basée sur l'indexation Lucene [182]. La structure principale dans Elasticsearch est l'index.

### Organisation d'un index Elasticsearch

Dans Elasticsearch, un **index** (équivalent d'une table dans un RDMBS) permet de stocker des **documents** (chacun étant équivalent à un tuple dans un RDMBS) contenant des **champs** (équivalents aux colonnes d'une table dans un RDMBS). La structure d'un index est appelée **mapping** (équivalent à la description de la structure d'une table en termes de colonnes, de type de données et d'index dans un RDMBS).

Chaque index dans Elasticsearch est composé d'un ou plusieurs **shards** (Figure 3.3), correspondant à un index inversé Lucene [183]. Un index inversé Lucene est composé de deux sous-structures (voir un exemple dans le Tableau 3.2) :

— Le dictionnaire des termes, qui stocke tous les termes inclus dans les documents indexés au sein d'une liste ordonnée ;

— Le *postings list* [2], qui stocke pour chaque terme la liste des identifiants des documents contenant le terme courant, ainsi que la ou les positions du terme dans le document.

Tableau 3.2 – Exemple d'un index inversé Lucene

| **Terme** | **Postings List** |
|---|---|
| rhumatologie | 1[(56, 67)] |
| hospitalisation | 1[(12, 26), (157, 171)], 2[(18, 32)] |
| cœur | 2[(46, 54)] |

L'index inversé Lucene contient trois entrées : « rhumatologie », « hospitalisation » et « cœur ». Le terme « rhumatologie » est uniquement présent dans le document 1, avec une seule occurrence retrouvée à la position (56, 67) dans le document (la position correspond aux indexes des caractères du document). Le terme « hospitalisation » est retrouvé dans le document 1 (deux occurrences) et le document 2 (une occurrence). Le terme « cœur » est uniquement retrouvé dans le document 2, avec une seule occurrence. On note que la position du terme « cœur » dans le document 2 est associée à une chaîne de caractères plus longue que le terme « cœur » (neuf caractères dans le document 2 vs. cinq caractères pour le terme « cœur »). Cela illustre qu'un traitement du texte a eu lieu au moment de l'indexation (ex : racinisation du terme « cardiaque » en « cœur »).

Dans Elasticsearch, les requêtes sont parallélisées entre les différents *shards* d'un index, ce qui améliore les performances de recherche lorsqu'un index est divisé en plusieurs *shards*. Les *shards* primaires peuvent être répliqués sur différents nœuds au sein d'un **cluster** Elasticsearch (Figure 3.3), permettant une haute disponibilité : si un nœud du cluster est non fonctionnel (déconnecté du réseau, espace disque saturé, etc.), un autre nœud encore disponible peut être utilisé pour répondre à la requête.

Il est important de noter que les systèmes NoSQL tels qu'Elasticsearch ne prennent pas efficacement en charge les opérations de jointure [178]. Par conséquent, une étape de dé-normalisation consistant à dupliquer les données de manière à ce que toutes les informations à interroger se trouvent dans une structure unique est souvent nécessaire pour permettre une interrogation similaire à celle des bases de données relationnelles [184].

### Indexation du texte dans Elasticsearch

Elasticsearch prend en charge différents types de données. En particulier, le type de données *text* st associé à un processus d'indexation spécifique, impliquant des **analyseurs**. Au cours du processus d'indexation des données textuelles, un analyseur construit un index inversé Lucene en trois étapes successives :

1. Une étape de **filtrage des caractères**, au cours de laquelle le texte à indexer est nettoyé. Il peut s'agir de supprimer des caractères spéciaux ou de remplacer certains caractères par d'autres (ex : œ →oe).

2. La notion de *postings list* est détaillée dans `https://www.elastic.co/fr/blog/frame-of-reference-and-roaring-bitmaps`

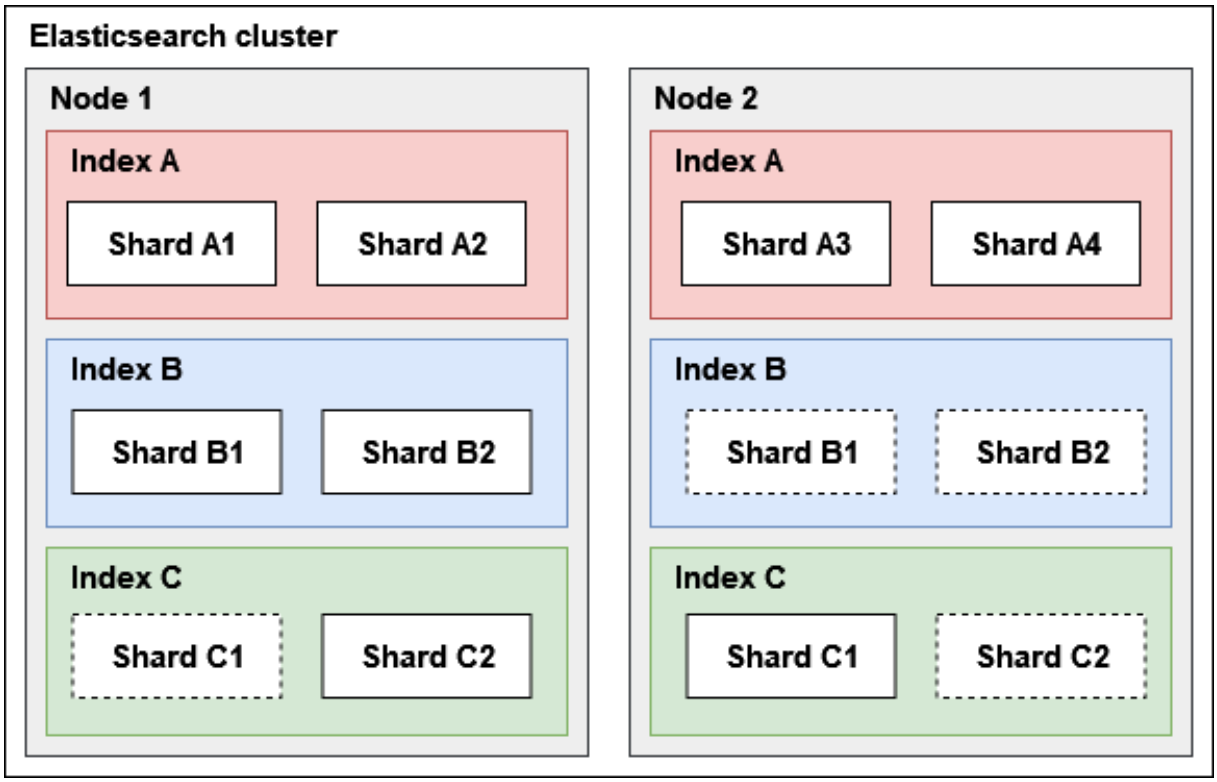


FIGURE 3.3 – Organisation et composant d'Elasticsearch

Les *shards* avec une bordure solide correspondent à des *shards* primaires, les *shards* avec une bordure en pointillé correspondent à des *shards* répliqués. L'index A est divisé en quatre *shards* primaires, sans réplication, répartis sur deux nœuds dans le cluster Elasticsearch. L'index B est divisé en deux *shards* primaires stockés sur le nœud 1 et deux *shards* répliqués stockés sur le nœud 2. L'index C est également divisé en deux *shards* primaires, mais répartis sur les deux nœuds du cluster Elasticsearch.

2. Une étape de **tokenisation**, au cours de laquelle le texte est découpé en tokens. Par défaut, Elasticsearch utilise le standard *Unicode Text Segmentation*[3], qui supprime la ponctuation et utilise le caractère espace comme séparateur.
3. Une étape de **filtrage des tokens**, au cours de laquelle des tokens peuvent être modifiés, supprimés ou ajoutés par rapport aux tokens générés à l'étape précédente. Par exemple, c'est à ce stade que les *stop words* (mots vides) et les synonymes sont appliqués.

Le résultat de cette chaîne de traitement est finalement stocké sous la forme d'un index inversé Lucene.

#### Le type de données *keyword* dans Elasticsearch

Le type de données ***keyword*** est dédié aux données structurées. Il s'agit d'un type de données particulier, qui peut être utilisé pour les données numériques et textuelles.

Lorsqu'un champ est mappé en tant que *keyword*, il est stocké tel quel dans l'index inversé de Lucene, sans aucun traitement avant l'indexation. Par exemple, le texte `'BIO:ERYTHRO'` sera stocké comme `'BIO:ERYTHRO'` s'il a pour type de données *keyword*, alors qu'il sera stocké comme `'bio'` (première entrée dans l'index inversé) et `'erythro'` (deuxième entrée dans l'index inversé) s'il est associé au type de données *text*.

Les champs de type *keyword* sont très efficaces pour l'agrégation, le tri ou la recherche de termes précis.

### 3.2.2 Méthodes

Une partie du modèle de données d'i2b2 a dû être adaptée pour que la persistance des données dans Elasticsearch soit possible[4]. Cette adaptation a été réalisée en deux étapes successives :

1. Mapping de la table `OBSERVATION_FACT` vers un index *observation_fact* Elasticsearch. Il s'agit de traduire les commandes SQL du langage de définition de données (*Data Definition Language* ou DDL) pour la création de la table `OBSERVATION_FACT` en une structure d'index *observation_fact* (principalement le nom et le type des champs, ainsi que le mode d'indexation).
2. Modification de la structure de l'index *observation_fact* :
   (a) En ajoutant à l'index des informations contenues dans les tables de dimension ;

3. La norme *Unicode Text Segmentation* est disponible via : `https://www.unicode.org/reports/tr29/tr29-17.html`
4. Dans la suite de ce document, toutes les notions relatives aux RDBMS sont écrites en `MAJUSCULES`, et toutes les notions relatives à Elasticsearch sont écrites en *italique*.

(b) En ajoutant à l'index des éléments servant à contextualiser les observations, extraites depuis l'ONT. Ces éléments sont ajoutés à l'index *observation_fact* en tant que facettes [5] [185] des observations. Cette étape a été nécessaire pour pallier l'impossibilité d'effectuer des jointures dans Elasticsearch.

Après ces deux étapes, la persistance du modèle de données d'i2b2 dans Elasticsearch a été évaluée.

#### *3.2.2.1 Mapping de la table `OBSERVATION_FACT` en index observation_fact*

Cette première étape a consisté à mapper l'ensemble des champs de la table `OBSERVATION_FACT` en tant qu'index Elasticsearch. Les mappings Elasticsearch permettent de définir deux éléments principaux :

— Le type de données d'un champ dans l'index Elasticsearch : numérique (long ou double), texte, date ou *keyword*.
— La manière dont un champ du document est indexé. Par exemple, pour un champ de type texte, le mapping définira le(s) analyseur(s) à utiliser pour indexer le contenu texte de ce champ.

Les deux critères considérés pour mapper la table `OBSERVATION_FACT` étaient :

1. La nécessité d'effectuer des requêtes de dénombrement ou d'agrégation sur le champ considéré. Dans ce cas, le type de données Elasticsearch attribué à ce champ était *keyword*. Dans la table `OBSERVATION_FACT`, les champs utilisés comme critères d'agrégation ou dans des requêtes de dénombrement sont des champs correspondant à des pseudo-clés étrangères des tables de dimension :
   — `ENCOUNTER_NUM` et `PATIENT_NUM` : colonnes utilisées pour dénombrer les patients ou les venues, agréger les informations par patient ou par venue, etc.
   — `CONCEPT_CD`, `MODIFIER_CD` et `PROVIDER_ID` : colonnes utilisées pour dénombrer les éléments par critère clinico-biologique, agréger les informations par fournisseur de données, etc.
   — `INSTANCE_NUM` : colonne qui correspond à la clé de regroupement d'une observation principale avec ses versions modifiées.
   — `TVAL_CHAR`, `VALUEFLAG_CD` et `UNITS_CD` : colonnes utilisées dans le cas d'observations numériques pour spécifier des éléments concernant l'opérateur numérique ($=$, $>$, $\geq$, etc.), l'interprétation par rapport à des valeurs limites ou l'unité associée aux résultats numériques.
2. La nécessité d'interroger un champ à l'aide de requêtes en texte libre. Dans ce cas, la colonne a été mappée au type de données *text*, associé à un analyseur pour

5. Les facettes de recherche sont des filtres qui permettent aux utilisateurs de restreindre ou d'affiner les résultats de recherche en fonction de plusieurs critères, comme la catégorie, la date ou le type de contenu.

permettre une gestion du texte libre. La colonne concernée dans la table `OBSERVATION_FACT` correspond au champ `OBSERVATION_BLOB`[6].

Le résultat du mapping de la table `OBSERVATION_FACT` en un index Elasticsearch est présenté dans le Tableau 3.3.

Tableau 3.3 – Mapping des colonnes de la table `OBSERVATION_FACT` aux types de données Elasticsearch

| Nom de colonne | Type de données RDMBS | Type de données Elasticsearch | Analyseur |
|---|---|---|---|
| ENCOUNTER_NUM | NUMBER | *keyword* | No |
| PATIENT_NUM | NUMBER | *keyword* | No |
| CONCEPT_CD | VARCHAR2 | *keyword* | No |
| PROVIDER_ID | VARCHAR2 | *keyword* | No |
| START_DATE | DATE | *date* | No |
| MODIFIER_CD | VARCHAR2 | *keyword* | No |
| INSTANCE_NUM | NUMBER | *keyword* | No |
| VALTYPE_CD | VARCHAR2 | *keyword* | No |
| TVAL_CHAR | VARCHAR2 | *keyword* | No |
| NVAL_NUM | NUMBER | *double* | No |
| VALUEFLAG_CD | VARCHAR2 | *keyword* | No |
| QUANTITY_NUM | NUMBER | *double* | No |
| UNITS_CD | VARCHAR2 | *keyword* | No |
| END_DATE | DATE | *date* | No |
| LOCATION_CD | VARCHAR2 | *keyword* | No |
| OBSERVATION_BLOB | TEXT | *text* | Yes |
| CONFIDENCE_NUM | NUMBER | *double* | No |
| UPDATE_DATE | DATE | *date* | No |
| DOWNLOAD_DATE | DATE | *date* | No |
| IMPORT_DATE | DATE | *date* | No |
| SOURCESYSTEM_CD | VARCHAR2 | *keyword* | No |
| UPLOAD_ID | NUMBER | long | No |

#### *3.2.2.2 Modification de la structure de l'index observation_fact*

Comme il n'est pas possible d'effectuer des jointures entre index dans Elasticsearch, deux types de modification ont été apportées à l'index *observation_fact* obtenu après l'étape de mapping pour contourner cette limitation (Figure 3.4) :

1. Certaines des colonnes des tables `PATIENT_DIMENSION` et `VISIT_DIMENSION` ont été ajoutées à l'index *observation_fact*. Pour des raisons de parcimonie, seules les colonnes qui étaient utilisées pour des besoins de requêtage localement au CHU de Bordeaux ont été intégrées dans l'index *observation_fact*. La colonne `PATIENT_DIMENSION.SEX_CD` a ainsi été ajoutée à l'index *observation_fact* tandis que `PATIENT_DIMENSION.BIRTH_DATE` et `VISIT_DIMENSION.START_DATE` ont été utilisées pour ajouter à l'index *observation_fact* le champ calculé

6. Dans `OBSERVATION_FACT`, la colonne `TVAL_CHAR` est également utilisée pour stocker du texte court. Dans notre implémentation au CHU de Bordeaux, tous les textes (courts ou longs) sont stockés dans `OBSERVATION_BLOB`.

*age_in_month_at_start_visit* correspondant à l'âge en mois au début de la venue à laquelle est rattachée chaque observation.

2. Afin de rendre les documents de l'index *observation_fact* accessibles au travers des différents niveaux hiérarchiques décrits dans la partie ONT du modèle i2b2, tous les `C_FULLNAME` associés aux `CONCEPT_PATH`, `MODIFIER_PATH` et `PROVIDER_PATH` ont également été intégrés dans l'index *observation_fact*. Il a été nécessaire de créer des tables intermédiaires supplémentaires dans l'ONT afin de permettre l'ajout, par jointure, de tous les chemins (intermédiaires et finaux) pour chaque valeur possible des colonnes de recherche de `OBSERVATION_FACT`. Dans l'index *observation_fact*, les chemins ont été stockés dans trois nouveaux champs appelés *c_fullname_concept*, *c_fullname_modifier* et *c_fullname_provider* en tant que listes de "*keywords*".

   Le pseudo-code décrivant la génération de la table `I2B2_PATH_CONCEPT`, utilisée pour l'ajout des `C_FULLNAME` associés aux `CONCEPT_PATH` dans l'index *observation_fact*, est proposé dans l'Algorithme 1.

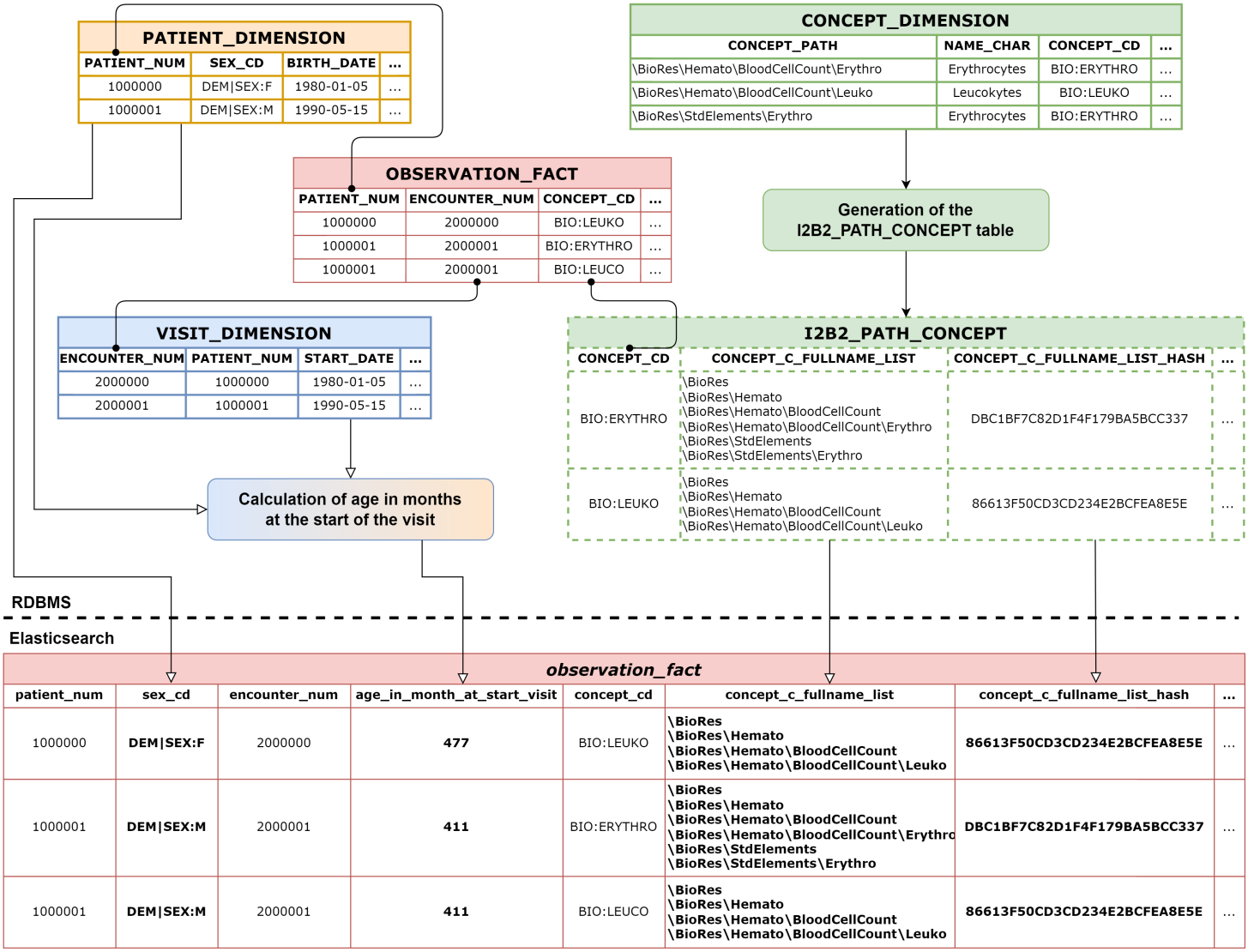


FIGURE 3.4 – Modification de la structure de l'index *observation_fact*.

La table `I2B2_PATH_CONCEPT` (bordure en pointillé) est créée sur la base de `CONCEPT_DIMENSION` : pour chaque `CONCEPT_CD`, la table contient la liste ordonnée des `CONCEPT_PATH`, y compris les chemins intermédiaires. Les données pertinentes des tables `PATIENT_DIMENSION`, `VISIT_DIMENSION` (tables i2b2 natives, bordure pleine) et `I2B2_PATH_CONCEPT` sont ajoutées à l'index *observation_fact* en tant que facette. La date de naissance est utilisée avec la date de début de visite pour calculer la facette *age_in_month_at_start_visit*. Les flèches avec les extrémités circulaires représentent les jointures entre les tables.

**Algorithm 1** CONSTRUCT_I2B2_PATH_CONCEPT

```
1: Variables
2:     dict_concept_fullnames: Dictionary of (String, List of Strings)
3:     I2B2_PATH_CONCEPT: Table of (CONCEPT_CD: String, C_FULLNAME_LIST: String)
4:     CONCEPT_DIMENSION: Table of (CONCEPT_CD: String, C_FULLNAME: String)
5:     row: Record of (CONCEPT_CD: String, C_FULLNAME: String)
6:     concept_cd, c_fullname, path, concept_c_fullname_list: String
7:     all_paths: List of Strings

8: Function get_all_paths(c_fullname: String): List of Strings
9:     Variables
10:        component: String
11:        components: List of Strings
12:        path: String
13:        paths: List of Strings
14:    Begin
15:        paths ← []
16:        components ← SPLIT(c_fullname, '\')
17:        path ← ""
18:        For each component in components do
19:            If component ≠ "" then
20:                If path = "" then
21:                    path ← "\" + component
22:                Else
23:                    path ← path + "\" + component
24:                End If
25:                ADD(paths, path)
26:            End If
27:        End For
28:        Return paths
29:    End Function
30: Begin
31:    dict_concept_fullnames ← Dictionary()
32:    I2B2_PATH_CONCEPT ← Table(CONCEPT_CD: String, C_FULLNAME_LIST: String)

33:    For each row in CONCEPT_DIMENSION do
34:        concept_cd ← row.CONCEPT_CD
35:        c_fullname ← row.C_FULLNAME
36:        all_paths ← get_all_paths(c_fullname)
37:        If NOT CONTAINS_KEY(dict_concept_fullnames, concept_cd) then
38:            dict_concept_fullnames[concept_cd] ← []
39:        End If
40:        For each path in all_paths do
41:            If NOT CONTAINS(dict_concept_fullnames[concept_cd], path) then
42:                ADD(dict_concept_fullnames[concept_cd], path)
43:            End If
44:        End For
45:    End For
```

```
31:    For each key_value_pair in dict_concept_fullnames do
32:      concept_cd ← key_value_pair.KEY
33:      SORT(key_value_pair.VALUE)
34:      concept_c_fullname_list ← CONCATENATE(key_value_pair.VALUE, '\n')
35:       INSERT INTO I2B2_PATH_CONCEPT (CONCEPT_CD, C_FULLNAME_LIST)
   VALUES (concept_cd, concept_c_fullname_list)
36:    End For
```

#### *3.2.2.3 Évaluation de la persistance des données dans Elasticsearch*

L'évaluation de l'implémentation proposée du modèle de données d'i2b2 pour la persistance des données cliniques dans Elasticsearch vise à évaluer :

1. **La capacité à charger un sous-ensemble des données dans l'EDS** (chargement / mise à jour de la partie CRC) sur une base quotidienne. Les mécanismes spécifiques utilisés pour charger les données ont été décrits. Les performances de chargement ont été évaluées en termes de temps de chargement médian et de nombre médian de documents chargés par minute.

2. **La capacité à mettre à jour les métadonnées** (chargement / mise à jour de la partie ONT). De même, les mécanismes spécifiques mis en œuvre pour mettre à jour les éléments de métadonnées intégrés dans l'index *observation_fact* ont été décrits. Les performances de mise à jour des métadonnées ont été évaluées en termes de temps médian de mise à jour par rapport au nombre moyen de *concept_cd* devant être mis à jour (mise à jour hebdomadaire).

3. **La possibilité d'effectuer des requêtes sur l'EDS persisté avec Elasticsearch**. Les résultats des requêtes de dénombrement (patients distincts ou venues distinctes) ont été comparés entre l'implémentation standard sur Oracle et l'implémentation avec Elasticsearch, sur la base d'un même jeu de données. L'adéquation des résultats des requêtes, ainsi que le temps d'exécution (temps d'exécution médian basé sur 20 exécutions), ont été évalués. Différents types de requêtes ont été effectués, avec une évaluation des critères au sein de la requête au niveau du patient (survenue de l'ensemble des critères à l'échelle de l'ensemble du dossier du patient) ou de la venue (survenue de l'ensemble des critères de la requête au sein d'une même venue). Les différentes requêtes qui ont été exécutées sont les suivantes :
   - Requêtes de dénombrement basées sur des données structurées (requêtes n° 1 et n° 2), correspondant à une requête renvoyant un faible nombre de patients ($\approx$ 1 500 patients) et une requête renvoyant un grand nombre de patients ($\approx$ 150 000 patients).
   - Requête de dénombrement basée sur des données structurées associées à un filtre numérique (requête n°3).
   - Requête de dénombrement basée sur des données en texte libre (requête n°4).

— Requête de dénombrement basée sur des données structurées associées à un filtre temporel (requête n°5).
— Combinaison des critères précédents, dans des requêtes booléennes ET/OU, avec des ordres d'évaluation de critères multiples (ex : [(1) ET (2)] vs [(2) ET (1)]).

## 3.3 Résultats

Un ETL spécifique a été développé à l'aide de Spring Boot [7] pour convertir les données d'i2b2 disponibles dans une base de données relationnelle Oracle en Elasticsearch.

### 3.3.1 Chargement des données dans Elasticsearch

Les données des tables `PATIENT_DIMENSION` et `VISIT_DIMENSION` ont été ajoutées à la table `OBSERVATION_FACT` à l'aide de jointures externes dans l'ETL. Pour les données contenues dans l'ONT, des tables intermédiaires contenant la liste ordonnée de tous les `CONCEPT_PATH` (respectivement `MODIFIER_PATH` ou `PROVIDER_PATH`) pour chaque `CONCEPT_CD` (respectivement `MODIFIER_CD` ou `PROVIDER_ID`) ont été créées. Un exemple pour les `CONCEPT_CD` est présenté dans le Tableau 3.4.

Tableau 3.4 – `I2B2_PATH_CONCEPT` : table de métadonnées intermédiaire (non disponible dans le modèle i2b2)

| CONCEPT_CD | CONCEPT_C_FULLNAME_LIST | CONCEPT_C_FULLNAME_LIST_HASH |
|---|---|---|
| BIO:ERYTHRO | \BioRes<br>\BioRes\Hemato<br>\BioRes\Hemato\BloodCellCount<br>\BioRes\Hemato\BloodCellCount\Erythro<br>\BioRes\StdElement<br>\BioRes\StdElement\Erythro | DBC1BF7C82D1F4F179BA5BCC337D6DE3 |
| BIO:LEUKO | \BioRes<br>\BioRes\Hemato<br>\BioRes\Hemato\BloodCellCount<br>\BioRes\Hemato\BloodCellCount\Leuko | 86613F50CD3CD234E2BCFEA8E5EB098A |

`I2B2_PATH_CONCEPT` contient, pour chaque `CONCEPT_CD` décrit dans l'ONT, tous les chemins (`C_FULLNAME`) disponibles. Il existe une table équivalente pour le `MODIFIER_CD` et `PROVIDER_ID`.

La requête SQL utilisée dans le processus ETL pour charger l'index *observation_fact* est disponible en Annexe B.1.

7. Framework basé sur Java, la documentation est disponible via `https://spring.io/projects/spring-boot`

#### *3.3.1.1 Partitionnement de l'index observation_fact*

Dans l'EDS du CHU de Bordeaux, les données contenues dans la table `OBSERVATION_FACT` sont partitionnées par source de données et par année (Figure 3.2).

Une organisation similaire a été mise en place au niveau de la persistance des données dans Elasticsearch, avec la création d'un index pour chaque partition de la table source : *observation_fact* est ainsi composée de plusieurs index (ex : *observation_fact-pmsi-2024*, *observation_fact-pmsi-2023*, *observation_fact-pmsi-2022*, etc.).

Pour s'assurer que la structure est la même pour chaque index, tous les index ont été créés sur la base d'un Elastisearch *index-template* [8]. Un *index-template* contient la définition de la structure d'un index Elasticsearch. Il peut être utilisé pour définir la structure d'un nouvel index au moment de sa création (équivalent à de l'héritage).

#### *3.3.1.2 Rolling strategy pour le chargement des données dans l'index observation_fact*

Pour limiter le temps pendant lequel les observations contenues dans les index d'*observation_fact* ne sont pas accessibles aux requêtes, la *rolling strategy* suivante a été mise en œuvre pour le chargement des données :

1. Création du nouvel index basé sur l'*index-template* d'*observation_fact*. Le nom de l'index nouvellement créé est défini selon la convention suivante : *observation-[nom_source]-[année]-[timestamp_création_index]*. La partie *[timestamp_création_index]* est utilisée pour garantir l'unicité de l'index dans Elasticsearch ;
2. Chargement des données dans le nouvel index à l'aide de la *Bulk API* [9] d'Elasticsearch. Cette dernière permet le chargement des documents dans Elasticsearch par lot, ici de 500 observations ;
3. Suppression des index existants pour lesquels un alias *observation-[nom_source]-[année]* est disponible [10]. Dans notre cas, l'alias utilisé correspond au nom de l'index sans la partie horodatée (date de création de l'index), permettant d'identifier les index existants correspondant au(x) chargement(s) précédent(s) d'une source pour une année spécifique ;
4. Ajout de deux alias à l'index nouvellement créé :
   (a) L'alias *observation-[nom_source]-[année]*, qui sera utilisé pour trouver l'index à supprimer la prochaine fois que la source de données sera chargée pour une année spécifique ;
   (b) L'alias *observation*, identique pour tous les index constituant *observation_fact*. L'alias *observation* est utilisé comme source dans les requêtes Elasticsearch,

8. La documentation des *index-template* est disponible via : `https://www.elastic.co/guide/en/elasticsearch/reference/current/index-templates.html`
9. La documentation de la *Bulk API* est disponible via : `https://www.elastic.co/guide/en/elasticsearch/reference/8.15/docs-bulk.html`
10. Dans Elasticsearch, un alias est utilisé pour donner un nom alternatif à un index.

permettant d'interroger l'ensemble des index constituant *observation_fact* de manière transversale.

#### *3.3.1.3 Évaluation technique de la persistance des données dans Elasticsearch*

Toutes les données contenues dans la table `OBSERVATION_FACT` ont été chargées dans un cluster Elasticsearch, composé de trois nœuds. Chaque index constituant *observation_fact* a été scindé en trois *shards* primaires (un *shard* par nœud) pour permettre l'exécution de requêtes en parallèle. Aucun *shard* répliqué n'a été créé.

Cela représente 2 502 063 patients, 20 982 497 visites et 3 474 264 570 observations. Le nombre de données disponibles dans l'index *observation_fact* est le même que celui disponible dans la table `OBSERVATION_FACT`.

La persistance des données avec Oracle nécessite 1 030 To pour les données et 2 640 To pour les index (y compris 420 Go pour l'index *full-text* [186] d'Oracle). La persistance des données avec Elasticsearch nécessite 1 635 To, soit 44 % de l'utilisation du disque avec Oracle.

Les durées de chargement des données dans Elasticsearch par source sont données dans le Tableau 3.5. Pour l'année 2023 complète (correspondant à environ 340 000 000 observations), la durée médiane de chargement des données de 10 sources en parallèle était de 7,41 heures, correspondant à un nombre médian de 46 546 000 documents indexés par heure. La vitesse d'indexation semble dépendre du type de données, avec une indexation des données numériques structurées (médiane de 171 600 documents/min) environ quatre fois plus rapide que pour les données en texte libre (médiane de 48 100 documents/min).

### 3.3.2 Stratégie de mise à jour des métadonnées Elasticsearch

En cas de modification de la partie ONT du modèle i2b2, une stratégie de mise à jour des métadonnées incluses dans l'index *observation_fact*, visant à ne pas recharger l'ensemble des données intégrées, a été mise en place. La stratégie de mise à jour du champ *concept_c_fullname_list* est la suivante :

1. Au moment du chargement des données dans *observation_fact*, en plus de stocker la liste des *c_fullname* rattachés à chaque *concept_cd* dans le champ *concept_c_fullname_list*, nous stockons également un hash de cette liste ordonnée dans le champ *concept_c_fullname_list_hash* (dernière colonne du Tableau 3.4).

2. Au cours du processus de mise à jour des métadonnées contenues dans *observation_fact*, pour chaque index :
   (a) On récupère la liste distincte des tuples (*concept_cd* ; *concept_c_fullname_hash*) disponibles dans l'ensemble des index

Tableau 3.5 – Métriques de chargement des données pour l'ensemble de l'année 2023 *(métriques calculées sur la base de huit chargements)*.

| Source | Nombre d'observations | Documents (x1000/min) Médiane [Q1 ; Q3] | Durée de chargement (min) Médiane [Q1 ; Q3][a] |
|---|---|---|---|
| Chimiothérapie | 478 438 | 185,3 [132,1 ; 222,8] | 2,5 [2,1 ; 4,9] |
| Formulaires | 77 460 099 | 181,6 [164,4 ; 188,9] | 429,1 [407,3 ; 490,1] |
| Biologie | 158 384 678 | 171,6 [162,4 ; 176,5] | 916,4 [893,1 ; 968,3] |
| Prescription / Administration de médicaments | 32 736 762 | 155,6 [148,6 ; 199,3] | 204,5 [187,4 ; 214,6] |
| Prescription / Réalisation de soins infirmiers | 16 523 635 | 107,4 [87,2 ; 138,8] | 149,2 [113,7 ; 180,2] |
| Radiologie | 9 438 970 | 76,6 [37,8 ; 91,7] | 130,8 [79,9 ; 201,4] |
| PMSI (diagnostics et actes) | 6 690 589 | 64,6 [57,3 ; 105,7] | 105,2 [64,2 ; 116,8] |
| Mouvements | 6 181 496 | 59,8 [54,3 ; 79,2] | 102,8 [77,6 ; 118,3] |
| Données démographiques | 2 142 053 | 59,6 [55,0 ; 122,6] | 34,7 [17,3 ; 37,8] |
| Notes infirmières | 12 984 197 | 58,8 [52,4 ; 66,7] | 223,6 [199,0 ; 248,7] |
| Prescription / Administration de soins | 4 930 320 | 56,5 [30,0 ; 68,6] | 86,9 [71,7 ; 162,6] |
| Documents | 6 066 806 | 48,1 [39,8 ; 78,3] | 125,9 [80,2 ; 151,3] |
| Chirurgie | 518 998 | 13,9 [6,5 ; 248,9] | 37,4 [2,1 ; 80,3] |
| Microbiologie | 3 364 431 | 13,2 [11,5 ; 17,6] | 254,3 [189,7 ; 290,2] |
| Anatomo-pathologie | 678 289 | 10,7 [7,6 ; 18,8] | 65,6 [38,8 ; 89,4] |
| Autre | 284 220 | 4,9 [3,1 ; 190,7] | 59,1 [1,5 ; 90,6] |
| Transfusion | 467 166 | 2,7 [2,1 ; 3,1] | 165,3 [152,4 ; 205,4] |
| **Total (10 sources en parallèle)** | **339 331 147** | **775,8 [749,2 ; 792,1]** | **444,8 [429,0 ; 461,2]** |

[a] Les durées médianes de chargement ont été calculées sans tenir compte de la parallélisation. Certaines sources sont composées de plusieurs index, comme la biologie (qui est composée de cinq index) ou les formulaires (qui sont composés de quatre index). Cela explique pourquoi la durée médiane maximale de chargement (916,4 minutes) est plus élevée que la durée médiane de chargement totale (444,8 minutes).

*observation_fact*, y compris les *concept_cd* pour lesquels *concept_c_fullname_hash* est vide (c'est-à-dire les `CONCEPT_CD` non décrits dans l'ONT).

(b) On récupère le contenu de la table `I2B2_CONCEPT_PATH`, c'est-à-dire l'ensemble des tuples (`CONCEPT_CD` ; `CONCEPT_C_FULLNAME_LIST` ; `CONCEPT_C_FULLNAME_LIST_HASH`). Il est alors possible d'identifier tous les *concept_cd* qui doivent être mis à jour en identifiant ceux pour lesquels le hash disponible dans les index d'*observation_fact* n'est pas le même que celui obtenu dans la table `I2B2_CONCEPT_PATH` de l'ONT.

(c) Pour chaque *concept_cd* au sein de chaque index constituant *observation_fact*, on met finalement à jour l'ensemble des documents en utilisant la fonctionnalité d'Elasticsearch *Update by query*[11]. Cette fonctionnalité d'Elasticsearch permet de mettre à jour les documents correspondant à une requête spécifique :

— L'attribut « script » de cette méthode permet de remplacer la valeur d'un champ par une autre valeur. Ici, nous remplaçons l'ancienne valeur du champ *concept_c_fullname_list* par la nouvelle liste des chemins mise à jour. Le même processus est réalisé pour le champ *concept_c_fullname_hash*.

— L'attribut « recherche » de cette méthode permet à une requête de mise à jour de ne cibler que les documents qui répondent à une requête, ici les documents dont le *concept_cd* a été identifié comme nécessitant une mise à jour à l'étape 2.(a).

Les métadonnées ajoutées aux index composant *observation_fact* sont mises à jour chaque semaine dans l'EDS du CHU de Bordeaux. Le nombre moyen de `CONCEPT_CD` nécessitant une mise à jour chaque semaine est de 9 324 `CONCEPT_CD`, correspondant à 51 904 783 observations. Pour ce processus de mise à jour, la durée médiane de mise à jour est de 119,5 minutes (IQR [74,5 ; 146,3]).

### 3.3.3 Requêtage des données dans Elasticsearch

Les résultats des requêtes obtenues sur la base d'une persistance via Oracle ont été comparées avec les résultats obtenus via Elasticsearch. Pour chaque requête, les critères ont été évalués au niveau patient (requête non temporelle) et au niveau venue (requête temporelle). Les requêtes ont été exécutées 20 fois pour chaque moteur afin d'évaluer les durées d'exécution.

La base de données Oracle (version 19.17) était déployée sous la forme d'un cluster sur deux serveurs utilisant 80 CPUs and 1 To de mémoire vive chacun. La base de

11. La documentation de *Update by query* est disponible via : `https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-update-by-query.html`

données Elasticsearch (version 7.17.3) était configurée sous la forme d'un cluster de trois nœuds, déployés via Docker, chaque nœud utilisait 25 CPUs et 30 Go de mémoire vive.

Dans le cas de requêtes nécessitant des jointures (requêtes de type « ET », nécessitant l'intersection entre deux listes de patients ou de venues), les jointures ont été réalisées au niveau de la couche applicative à la fois pour Oracle et pour Elasticsearch. Un index *full-text* était disponible dans la base de données Oracle. Les index ont été calculés dans la base de données Oracle pour l'optimisation des requêtes, et les statistiques des tables et des index étaient à jour.

Un total de onze requêtes différentes ont été comparées (Tableau 3.6) :

— Les requêtes ont généré les mêmes résultats que ce soit avec le moteur Oracle ou avec le moteur Elasticsearch, à l'exception des requêtes impliquant de la recherche *full-text*, qui ont systématiquement renvoyées un dénombrement plus important via Elasticsearch (ces différences sont dues à des implémentations différentes au niveau des deux moteurs pour l'analyse du texte libre).

— Les temps de réponses étaient généralement plus courts avec Elasticsearch qu'avec Oracle, à l'exception de la requête « Pembrolizumab » qui générait un petit nombre de résultats. Les requêtes temporelles (évaluant les critères au niveau de la même venue) étaient systématiquement plus lentes que les requêtes non temporelles avec les deux moteurs.

Tableau 3.6 – Comparaison des requêtes entre Oracle et Elasticsearch. IQR : intervalle interquartile

| Requête | Type de requête | Type de résultat | Durée d'exécution (médiane [IQR]) | | Effectif[a] |
|---|---|---|---|---|---|
| | | | Elasticsearch | Oracle | |
| (1) Pembrolizumab | Niveau patient | Patients | 0,22 [0,21 ; 0,23] | 0,15 [0,14 ; 0,16] | 1 701 |
| (1) Pembrolizumab | Niveau séjour | Patients | 0,43 [0,41 ; 0,48] | 0,18 [0,18 ; 0,19] | 1 701 |
| (1) Pembrolizumab | Niveau séjour | Venues | 0,43 [0,41 ; 0,49] | 0,20 [0,19 ; 0,20] | 3 986 |
| (2) Cancer | Niveau patient | Patients | 0,96 [0,95 ; 0,98] | 8,99 [8,66 ; 9,51] | 153 583 |
| (2) Cancer | Niveau séjour | Patients | 4,81 [4,74 ; 4,87] | 16,07 [15,55 ; 16,63] | 153 583 |
| (2) Cancer | Niveau séjour | Venues | 4,79 [4,76 ; 4,85] | 15,97 [15,08 ; 16,76] | 516 682 |
| (3) Sodium $\geq$ 145 mmol/L | Niveau patient | Patients | 0,38 [0,36 ; 0,41] | 9,42 [8,16 ; 10,05] | 26 426 |
| (3) Sodium $\geq$ 145 mmol/L | Niveau séjour | Patients | 0,75 [0,71 ; 0,80] | 9,39 [8,33 ; 9,92] | 26 426 |
| (3) Sodium $\geq$ 145 mmol/L | Niveau séjour | Venues | 0,77 [0,72 ; 0,82] | 9,41 [8,38 ; 9,90] | 33 070 |
| (4) Adenocarcinoma in free-text | Niveau patient | Patients | 0,60 [0,59 ; 0,62] | 6,13 [5,97 ; 6,51] | 69 022 / 25 547 |
| (4) Adenocarcinoma in free-text | Niveau séjour | Patients | 3,27 [3,20 ; 3,33] | 7,27 [7,01 ; 7,50] | 69 022 / 25 547 |
| (4) Adenocarcinoma in free-text | Niveau séjour | Venues | 3,27 [3,24 ; 3,33] | 7,26 [7,04 ; 7,53] | 369 953 / 92 638 |
| (5) Cancer in 2023 | Niveau patient | Patients | 0,31 [0,29 ; 0,32] | 6,72 [6,53 ; 7,25] | 19 251 |
| (5) Cancer in 2023 | Niveau séjour | Patients | 0,66 [0,63 ; 0,71] | 6,88 [6,62 ; 7,52] | 19 251 |
| (5) Cancer in 2023 | Niveau séjour | Venues | 0,66 [0,64 ; 0,70] | 7,03 [6,88 ; 7,41] | 3 606 |
| (1) or (2) | Niveau patient | Patients | 1,00 [0,97 ; 1,02] | 9,15 [8,79 ; 9,68] | 153 589 |
| (1) or (2) | Niveau séjour | Patients | 4,95 [4,91 ; 5,02] | 15,72 [15,41 ; 16,39] | 153 589 |
| (1) or (2) | Niveau séjour | Venues | 4,99 [4,91 ; 5,05] | 16,3 [15,74 ; 16,81] | 516 853 |
| (2) or (1) | Niveau patient | Patients | 0,99 [0,97 ; 1,02] | 9,04 [8,79 ; 9,32] | 153 589 |
| (2) or (1) | Niveau séjour | Patients | 4,96 [4,89 ; 5,02] | 15,85 [15,38 ; 16,76] | 153 589 |
| (2) or (1) | Niveau séjour | Venues | 4,95 [4,91 ; 5,04] | 16,51 [15,44 ; 16,91] | 516 853 |
| (1) and (2) | Niveau patient | Patients | 1,07 [1,04 ; 1,10] | 6,68 [6,45 ; 6,95] | 1 695 |
| (1) and (2) | Niveau séjour | Patients | 4,81 [4,77 ; 4,90] | 6,83 [6,52 ; 7,33] | 1 671 |
| (1) and (2) | Niveau séjour | Venues | 4,81 [4,74 ; 4,91] | 6,94 [6,61 ; 7,12] | 3 815 |
| (2) and (1) | Niveau patient | Patients | 1,09 [1,06 ; 1,10] | 10,49 [10,17 ; 11,12] | 1 695 |
| (2) and (1) | Niveau séjour | Patients | 4,96 [4,85 ; 5,02] | 21,38 [20,46 ; 22,62] | 1 671 |
| (2) and (1) | Niveau séjour | Venues | 4,94 [4,89 ; 5,01] | 21,38 [20,75 ; 22,08] | 3 815 |
| (5) and (3) | Niveau patient | Patients | 0,64 [0,59 ; 0,69] | 15,54 [14,41 ; 17,51] | 1 917 |
| (5) and (3) | Niveau séjour | Patients | 1,45 [1,37 ; 1,51] | 16,15 [14,96 ; 18,42] | 1 005 |
| (5) and (3) | Niveau séjour | Venues | 1,47 [1,4 ; 1,57] | 16,42 [14,95 ; 18,05] | 1 135 |
| (1) and (5) and (3) | Niveau patient | Patients | 0,89 [0,82 ; 0,93] | 15,68 [14,8 ; 17,29] | 78 |
| (1) and (5) and (3) | Niveau séjour | Patients | 2,05 [1,85 ; 2,08] | 15,76 [14,59 ; 17,98] | 11 |
| (1) and (5) and (3) | Niveau séjour | Venues | 1,95 [1,91 ; 2,12] | 15,61 [13,84 ; 17,23] | 11 |

[a] Si Oracle et Elasticsearch renvoient le même résultat, un seul décompte est fourni, sinon le décompte obtenu avec Elasticsearch est fourni avant celui obtenu avec Oracle.

## 3.4 Discussion

À notre connaissance, ce travail est le premier à évaluer la faisabilité technique et les performances de la persistance d'une partie du modèle d'i2b2 dans une base de données Elasticsearch. En plus de démontrer la faisabilité d'une telle persistance, l'évaluation montre une amélioration significative des temps de requêtes globaux, ainsi qu'une diminution des ressources matérielles requises, en particulier en termes de stockage, avec une réduction de 66% de l'espace disque requis. En outre, la mise en œuvre a été testée sur un grand volume de données, démontrant un passage à l'échelle réussi. L'implémentation proposée est utilisée en production au CHU de Bordeaux, et plus de 200 projets de recherche utilisent l'EDS du CHU de Bordeaux depuis 2022.

### 3.4.1 Comparaison avec d'autres implémentations d'EDS

Bien qu'aucune implémentation d'i2b2 utilisant un base de données NoSQL ne soit décrite à notre connaissance, la littérature contient des exemples d'EDS prenant en compte spécifiquement le besoin en analyse du texte libre.

OpenMRS [168] est un projet open-source de dossier patient informatisé ; il contient un module de recherche en texte libre basé sur Apache Lucene intégré à Hibernate Search [12]. Roogle [100] est un EDS orienté document où les données en texte libre sont pré-indexées sur la base de terminologies externes (ex : Medical Subject Headings ou MeSH) et sont stockées dans un index Lucene. Dr. Warehouse [102, 187] propose une implémentation incluant un découpage des documents en texte libre en « propositions », un stockage de ces « propositions » dans un index *full-text* Oracle associé à des éléments de contexte (antécédent, personnel, familial, etc.) et lié à la négation. Doc'EDS [167] propose un moteur de recherche pour le phénotypage basé sur des données structurées et non structurées, indexées via Apache Lucene.

L'ensemble des implémentations d'EDS qui utilisent des bases de données NoSQL [100, 167, 168] est basé sur des librairies bas niveau utilisant Apache Lucene. Elasticsearch est également basé sur Apache Lucene mais correspond à une librairie de haut niveau. Bien que cela limite la possibilité de paramétrer très finement les index pour répondre aux besoins les plus spécifiques, utiliser une librairie de haut niveau comme Elasticsearch permet de bénéficier de l'ensemble de l'environnement disponible autour de cette librairie, comme par exemple le moteur graphique Kibana [13] ou encore les clients d'interrogation disponibles en Java via Spring Boot [14].

12. Le documentation d'OpenMRS est disponible via `https://openmrs.atlassian.net/wiki/spaces/projects/pages/27009030/Universal+Search+Box+Design+Page`
13. La documentation de Kibana est disponible via : `https://www.elastic.co/kibana`
14. La documentation de Spring Data Elasticsearch est disponible via : `https://spring.io/projects/spring-data-elasticsearch`

### 3.4.2 Justification des choix techniques

#### *3.4.2.1 Justification des modifications du modèle d'intégration des données cliniques d'i2b2 pour une persistance dans Elasticsearch*

Dans notre implémentation, nous avons choisi d'ajouter les colonnes de recherche situées dans les tables de dimension ainsi que les chemins d'accès au `CONCEPT_CD` (ainsi qu'au `MODIFIER_CD` et au `PROVIDER_ID`) en tant que facettes des index constituant *observation_fact*. Combinée à des champs de type *keyword*, cette implémentation permet d'optimiser les temps d'interrogation, en particulier pour les requêtes d'énumération, au détriment du temps d'indexation. L'inconvénient d'une telle implémentation est que les données (CRC) et les métadonnées (ONT) ne sont ici plus séparées, ce qui signifie que si les métadonnées sont modifiées, les données doivent être rechargées ou mises à jour. Pour remédier à ce problème, nous avons mis en œuvre un processus de mise à jour des métadonnées intégrées aux index constituant *observation_fact* et avons démontré la faisabilité d'une telle mise à jour sans recharger l'ensemble des données.

Une implémentation alternative consisterait à créer un index *observation_fact* sans y ajouter d'autres champs, et à créer des index supplémentaires pour les différentes tables de dimension et la table de métadonnées ONT. De la même manière que dans l'implémentation proposée ici, des jointures au niveau de la couche applicative (c'est-à-dire la réalisation d'intersections de listes de patients ou de venues au niveau de l'application) seraient utilisées pour croiser les résultats obtenus sur la base de l'interrogation des différents index dans le cadre de requêtes booléennes « ET ». Cependant, avec cette approche, l'intersection du sexe (ex : « femme ») persisté dans l'index *patient_dimension* avec un critère persisté au niveau des index composant *observation_fact* impliquerait de monter en mémoire la liste de l'ensemble des clés patients correspondant à des femmes dans l'EDS (soit environ 1,2 million de femmes, pour l'EDS du CHU de Bordeaux), ce qui pourrait poser des problèmes de performance. Le problème serait encore plus important pour les données persistées au niveau de l'index *visit_dimension*.

Une autre solution pour Elasticsearch aurait été d'utiliser le type de données *nested* [15]. Ce type permet d'indexer le contenu du champ imbriqué en tant que document Lucene distinct du document principal. De cette manière, il aurait été possible d'indexer un document « patient », contenant des documents imbriqués « venues », contenant eux-mêmes des documents imbriqués « observations ». Cependant, avec cette implémentation, il aurait été nécessaire de recharger en même temps l'ensemble des données pour un patient donné, ce qui n'était pas compatible avec notre objectif de recharger les données récentes sur une base quotidienne (et les données plus anciennes avec une fréquence plus faible).

15. La documentation du type de données *nested* est disponible via : `https://www.elastic.co/guide/en/elasticsearch/reference/current/nested.html`

En outre, les performances des requêtes impliquant des documents imbriqués ne sont pas aussi bonnes que celles des requêtes des champs standards.

#### *3.4.2.2 Justification de l'organisation du cluster Elasticsearch*

Notre cluster Elasticsearch se compose de trois nœuds déployés via Docker. Chaque index est divisé en trois *shards* primaires répartis sur chacun des trois nœuds. En divisant chaque index en trois *shards*, les requêtes peuvent être exécutées en parallèle. Par ailleurs, aucune réplication des *shards* n'a été mise en place. En effet, notre objectif ici n'est pas la haute disponibilité ; nous ne voulons pas une réponse systématique à partir de la base de données, mais plutôt que l'infrastructure ne réponde que si l'ensemble des données sont disponibles - notre objectif final étant d'avoir une réponse valide au regard de l'ensemble des données.

### 3.4.3 Limites du travail actuel

Dans notre implémentation, nous avons choisi de mapper la colonne `TVAL_CHAR` de la table `OBSERVATION_FACT` en tant que *keyword* dans les index constituant *observation_fact*. Dans le modèle de données proposé par i2b2, la colonne `TVAL_CHAR` est utilisée dans deux cas de figure, en conjonction avec le contenu de la colonne `VALTYPE_CD` :

- `VALTYPE_CD = 'T'` : la colonne `TVAL_CHAR` contient le texte court associé au `CONCEPT_CD` ;
- `VALTYPE_CD = 'N'` : la colonne `TVAL_CHAR` contient l'opérateur mathématique associé à la valeur numérique contenue dans la colonne `NVAL_NUM` ('E' pour « Equal », 'NE' pour « Not equal », 'L' pour « Less than », etc.).

Dans l'implémentation d'i2b2 mise en place au CHU de Bordeaux, l'ensemble des données en texte libre, incluant les textes courts, est stocké dans la colonne `OBSERVATION_BLOB` ; la colonne `TVAL_CHAR` ne contient donc que les mot-clés correspondant aux opérateurs mathématiques des valeurs numériques. Ainsi, dans le cas d'une implémentation respectant le standard i2b2, il serait nécessaire de mapper la colonne `TVAL_CHAR` en *keyword* mais également d'y associer un analyseur afin de pouvoir y effectuer des recherches *full-text* en plus des requêtes d'agrégation.

Afin de pouvoir évaluer le requêtage au-dessus d'une base de données Elasticsearch, une partie du moteur de requêtes d'i2b2 a été ré-implémenté. Ce dernier permet d'effectuer des requêtes de dénombrement, multi-critères (requêtes booléennes « ET », « OU » et « NOT », incluant les recherches en texte libre), évaluées au niveau patient ou au niveau venue. Cependant, toutes les fonctionnalités disponibles dans le moteur de requêtes d'i2b2 n'ont pas été implémentées.

En particulier, la possibilité d'effectuer des requêtes sur la base du nombre d'occurrences des observations (patients venus au moins trois fois à l'hôpital, patients

avec au moins deux dosages de l'hémoglobine inférieurs à 10 g/dL, etc.) ou des requêtes prenant en compte une temporalité entre des évènements (patients avec un dosage d'hémoglobine inférieur à 10g/L dans les cinq jours après une chirurgie du colon, etc.) n'a pas été implémentée. Ces deux types de requêtes sont plus complexes à mettre en œuvre que celles déjà implémentées. En effet, elles nécessitent la réalisation d'étapes supplémentaires d'agrégation d'observations et de filtrage des agrégats, avant l'agrégation des résultats au niveau patient ou venue. Cependant, l'organisation des index constituant *observation_fact* proposée ici semble compatible avec la réalisation de telles requêtes.

L'ETL proposé pour le chargement de la base de données Elasticsearch s'appuie sur des données déjà intégrées dans un modèle i2b2 persisté dans un RDBMS (ici Oracle). Outre le fait de permettre un ETL simple entre i2b2 RDBMS et i2b2 Elasticsearch, le maintien de cette base de données relationnelle semble indispensable pour répondre à des usages spécifiques qui ne seraient pas possible avec la seule implémentation d'Elasticsearch (jointures avec le système d'information hospitalier contenant les données sources, préparation de données complexes avec des jointures entre `OBSERVATION_FACT` et elle-même, etc.). Ainsi, les améliorations de performance observées avec Elasticsearch doivent être mises en balance avec la nécessité de maintenir deux infrastructures en parallèle.

# CHAPITRE 4

## Les Entrepôts de Données de Santé du CHU de Bordeaux : de l'usage local aux réseaux fédérés

## Sommaire



# 4.1 Introduction

En France, le rapport Villani [188] a identifié le secteur de la santé comme l'un des cinq secteurs où les investissements financiers et humains devaient être les plus importants en matière d'IA. Le rapport mettait en particulier en avant la nécessité

d'« installer une coordination nationale » autour de l'IA en santé. Suite à ce rapport, la mission de préfiguration du HDH [19] a posé les bases de la stratégie nationale en matière d'utilisation secondaire des données de santé.

En parallèle de ces initiatives nationales, de nombreux établissements de santé ont mis en place localement des EDS pour tenter de lever les verrous de l'utilisation des données de santé [101, 167, 169, 189, 190, 60].

Dans ce contexte, l'appel à projets (AAP) national « accompagnement et soutien à la constitution d'entrepôts de données de santé hospitaliers » a été lancé par la Direction Générale de l'Offre de Soins (DGOS) du ministère de la santé et de la prévention, en partenariat avec la Banque Publique d'Investissement (BPI) et le HDH.

Sous l'égide du Groupement de Coopération Sanitaire (GCS) NOVA, les trois CHU de la Nouvelle-Aquitaine (CHU de Bordeaux, CHU de Limoges et CHU de Poitiers) ont été lauréats de la première vague de cet AAP. Fort de son expérience en matière d'EDS depuis 2018, le CHU de Bordeaux est le chef de file de ce consortium.

L'objectif de ce chapitre est de décrire le projet EDS tel qu'il est décliné au CHU de Bordeaux. Dans cette partie, nous montrons comment les travaux présentés dans les chapitres précédents contribuent au déploiement d'un EDS au service de la communauté hospitalière, mais également comment ils contribuent à permettre l'exploitation de ces mêmes données au sein de réseaux fédérés internationaux. Par ailleurs, nous décrivons les choix stratégiques qui nous semblent essentiels pour déployer un EDS répondant aux cas d'usage locaux et fédérés.

Après une présentation des éléments relatifs à la gouvernance de ce projet (section 4.2), les deux entrepôts différents, pensés pour des usages complémentaires, qui ont été mis en place au CHU de Bordeaux sont décrits :

— Un EDS basé sur le modèle de données i2b2. Ce premier entrepôt, mis en place dès 2018, est dédié aux usages locaux (section 4.3) ;
— Un EDS basé sur le modèle de données OMOP-CDM. Ce second entrepôt, mis en œuvre dans le contexte de EHDEN en 2020, est utilisé dans le cadre de la participation aux réseaux fédérés (section 4.4).

## 4.2 Gouvernance

Le projet EDS est piloté par le Pôle de Santé Publique, en lien avec la Direction de la Recherche Clinique de de l'Innovation (DRCI) et la Direction du Numérique (DNUM) du CHU de Bordeaux. Au sein du Service d'Information Médicale, c'est l'unité d'Informatique et Archivistique Médicales (IAM ; unité médicale) qui est en charge de la maîtrise d'ouvrage.

Une gouvernance a été prévue dès la phase initiale du projet, à partir de 2018, avec la mise en place d'un comité stratégique, d'un comité de pilotage et de deux groupes de travail (GT) : un GT technique et un GT sur les aspects réglementaires et éthiques.

A partir de 2023, la gouvernance a été revue afin d'être en conformité avec le référentiel EDS de la CNIL (le référentiel a été présenté section 1.3.2). Ont ainsi été constitués :

— Deux comités équivalents au comité de pilotage décrit dans le référentiel sur les EDS de la CNIL :
  1. Le **comité stratégique**, qui définit les grandes orientations stratégiques et scientifiques du projet ;
  2. Le **comité de coordination**, qui définit la mise en œuvre de l'entrepôt, veille à la nature et justifie les données intégrées dans l'entrepôt ;
— Un comité équivalent au CSE décrit dans le référentiel sur les EDS de la CNIL : le **Centre Éthique et Recherche en Santé** (CERS), qui rend un avis systématique préalable et motivé sur les projets nécessitant la réutilisation des données de l'entrepôt. Le CERS se réunit tous les 15 jours, permettant une réactivité dans les temps d'instruction des demandes.

## 4.3 L'EDS i2b2 : un entrepôt dédié aux usages locaux

Le premier EDS mis en place au CHU de Bordeaux, à partir de 2018, est l'entrepôt basé sur le modèle i2b2. Le modèle d'intégration i2b2 a été choisi pour sa grande souplesse à intégrer des données hétérogènes de santé hospitalières. De plus, la gestion fine des métadonnées dans l'ONT avait à l'époque été identifiée comme un facteur clé pour l'intégration des données tout en conservant un niveau de précision important.

### 4.3.1 Intégration des données dans l'EDS i2b2

#### *4.3.1.1 Périmètre des données intégrées dans l'EDS i2b2*

Deux types de données sont intégrées dans l'EDS i2b2 :

— Des **données à caractère personnel relatives aux patients** ;
— Des **métadonnées**, correspondant aux données des vocabulaires locaux, nationaux et internationaux utilisés pour décrire les données.

**Données à caractère personnel relatives aux patients**

L'EDS i2b2 du CHU de Bordeaux intègre les données issues du soin de l'ensemble des patients venus au moins une fois au CHU de Bordeaux, en hospitalisation ou en

consultation, depuis 2010. Pour ces patients, l'ensemble des données disponibles depuis 2005 [1] sont intégrées.

Au total, les données provenant de plus de 20 sources sont chargées quotidiennement dans l'EDS i2b2 du CHU de Bordeaux. La liste des sources de données intégrées dans l'EDS du CHU de Bordeaux est détaillée dans la Figure 4.1.

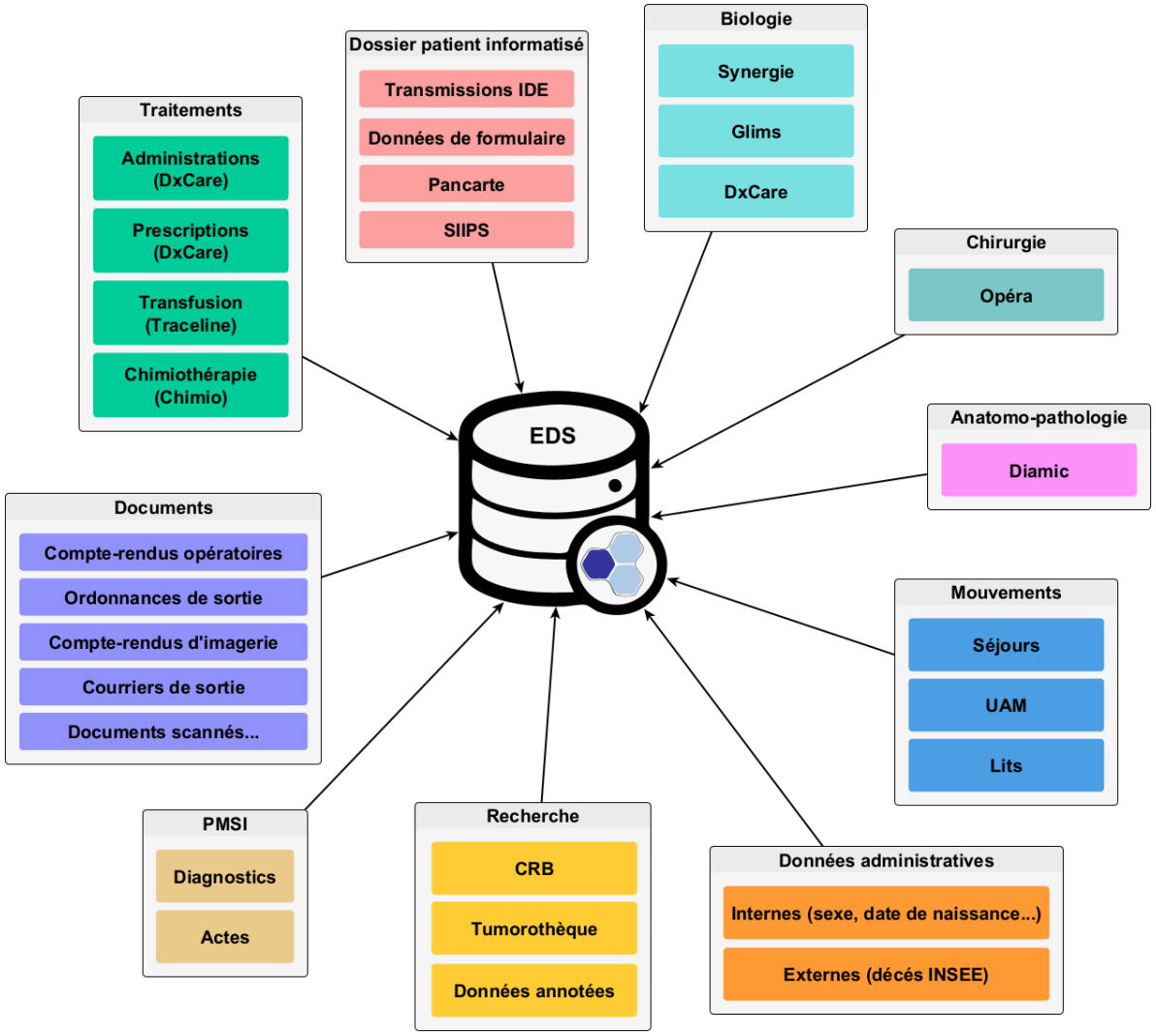


FIGURE 4.1 – Sources de données intégrées dans l'EDS du CHU de Bordeaux

En plus des données cliniques issues du soin, des données disponibles en *open-data* sont également intégrées dans l'EDS i2b2 du CHU de Bordeaux.

Dans le cadre de travaux précédents [191], un algorithme d'appariement des identités patients du CHU de Bordeaux avec celles contenues dans le fichier des personnes décédées mis à disposition par l'Institut national de la statistique et des études

1. 2005 correspond à l'année à partir de laquelle les premières données médicales numériques sont disponibles dans le SIH du CHU de Bordeaux

économiques (Insee) en *open-data*[2] a été mis en place localement. Cet appariement permet de mettre à disposition dans l'environnement EDS du CHU de Bordeaux les données relatives aux décès extra-hospitaliers. Au total, 77,2% des données de décès disponibles dans l'EDS i2b2 du CHU de Bordeaux sont issues de l'appariement avec le fichier des personnes décédées de l'Insee.

Également, un travail est en cours pour intégrer dans l'EDS du CHU de Bordeaux des données socio-économiques et environnementales disponibles en *open-data* rattachées à des informations géographiques (Système d'Information Géographique ou SIG). Il s'agit ici de mettre en place une chaîne de traitements permettant le géocodage en coordonnées géographiques (x, y) puis le transcodage en Îlots Regroupés pour l'Information Statistique (IRIS) des adresses de résidence connues des patients au CHU de Bordeaux.

Au total, en août 2024, les données intégrées dans l'EDS i2b2 du CHU de Bordeaux représentaient un total de :

— **2 502 063 patients** distincts ;
— **20 982 497 venues** distinctes ;
— **3 474 264 570 observations**, dont 72 580 022 documents textuels.

Le détail du nombre de patients distincts, du nombre de venues distinctes et du nombre d'observations par source est présenté dans le Tableau 4.1. Les deux sources d'information les plus importantes sont la biologie et les données issues des formulaires du DPI (respectivement 38,1 % et 23,9 % des observations disponibles). On note également le grand nombre de sources contenant des données en texte libre (532 239 705 observations soit 15,3 % des observations disponibles) Les données issues des soins critiques, en cours d'intégration, représentent encore un volume faible de données.

2. Le fichier des personnes décédées mis à disposition par l'Insee est disponible via : `https://www.data.gouv.fr/fr/datasets/fichier-des-personnes-decedees/`

Tableau 4.1 – Nombre d'observations, de patients distincts, de venues distinctes au sein de l'entrepôt de données de santé du CHU de Bordeaux en août 2024

| Source | Patients distincts | Venues distinctes | Observations |
|---|---|---|---|
| Biologie | 1 286 459 | 4 157 738 | 1 323 818 025 |
| Formulaires | 1 641 829 | 8 130 391 | 829 328 415 |
| Prescription / Administration de médicaments | 744 611 | 1 775 078 | 469 279 982 |
| Prescription / Réalisation de soins infirmiers | 688 450 | 1 875 347 | 263 813 687 |
| Radiologie | 1 095 002 | 3 690 225 | 127 437 286 |
| Mouvements | 2 315 342 | 20 259 380 | 86 453 388 |
| Documents | 2 045 379 | 12 202 309 | 72 580 022 |
| Diagnostics (CIM-10) | 947 024 | 2 650 212 | 63 135 010 |
| Prescription / Administration de soins | 725 741 | 1 582 745 | 54 581 222 |
| Notes infirmières | 619 767 | 1 729 145 | 54 558 694 |
| Microbiologie | 474 997 | 961 850 | 39 277 538 |
| Données démographiques | 2 502 063 | 20 982 497 | 23 841 233 |
| Actes (CCAM) | 1 396 469 | 5 720 634 | 16 503 052 |
| Chimiothérapie | 47 325 | 238 341 | 12 690 644 |
| Notes infirmières | 733 021 | 2 143 927 | 12 002 868 |
| Anatomo-pathologie | 397 492 | 658 679 | 10 371 426 |
| Chirurgie | 365 077 | 582 092 | 5 802 650 |
| Autres | 435 078 | 748 627 | 4 112 204 |
| Transfusion | 217 389 | 491 076 | 2 647 208 |
| Soins critiques | 8 733 | 9 349 | 2 030 016 |
| **Total** | **2 502 063** | **20 982 497** | **3 474 264 570** |

### Métadonnées

En plus de l'intégration des données individuelles relatives aux patients, un travail spécifique a été réalisé pour l'intégration des vocabulaires locaux, nationaux et internationaux utilisés pour décrire ces données patients. Chacune des sources de données intégrées dans l'EDS i2b2 a son propre dictionnaire, correspondant à des terminologies standards (CIM-10, ATC, etc.) ou des terminologies locales. Ces terminologies locales sont développées spécifiquement à la demande des soignants pour répondre au besoin de saisie de l'information dans les logiciels métiers sources.

Le dictionnaire de données intégré dans l'EDS i2b2 correspond à **2 022 478 entrées**, traduisant une grande hétérogénéité sémantique dans les données (cf. section 1.2.2.3). Ces entrées de dictionnaires se répartissent de la façon suivante :

- 979 176 data elements ;
  - 918 649 data elements à réponses ouvertes (93,82 %);
  - 60 527 data elements à réponses fermées (6,18 %) ;
- 601 427 modalités de réponses (issues des liste fermées de réponses possibles) ;
- 201 382 concepts issus d'une terminologie standards.

Les métadonnées sont organisées dans une base de données graphe sous la forme de triplets RDF (*Resource Description Framework*)[3]. Un méta-modèle, basé sur la norme ISO-11179-3 [143], est utilisé pour fournir une description de haut niveau des différentes métadonnées.

C'est au niveau de cette base de données graphe que sont persistés les alignements obtenus avec la méthode décrite dans le chapitre 2. Les métadonnées intégrées dans cette base de données graphe sont ensuite consommées par un programme pour alimenter l'ONT (cf. section 3.2.1.1) de l'application i2b2. C'est ce mécanisme qui permet de mettre à disposition, dans le contexte d'i2b2, les alignements vers la LOINC (et plus largement vers d'autre terminologies standards) pour faciliter les recherches au niveau du requêteur qui sera présenté dans la suite du document.

#### *4.3.1.2 Méthodes utilisées pour l'alimentation des données de l'EDS i2b2*

Le processus global d'alimentation de l'EDS i2b2 est présenté dans la Figure 4.2.

La source principale des données de l'EDS i2b2 du CHU de Bordeaux correspond aux données de soins répliquées dans un INFOCENTRE décisionnel, maintenu par l'équipe « Data Innovation » de la DNUM. Au niveau de l'INFOCENTRE, la fraîcheur des données disponibles par rapport à la production est de l'ordre de la seconde (pour les données du DPI) ou de l'heure (pour la plupart des autres sources de données). En ce qui concerne les données issues du fichier des personnes décédées mis à disposition par l'Insee, l'appariement avec les identités des patients du CHU de Bordeaux est mis à jour chaque trimestre[4].

Les données patients sont alimentées quotidiennement dans l'EDS au CHU de Bordeaux. Le processus d'ETL en charge de l'alimentation de l'EDS i2b2 prend en compte la date de production des données. Comme évoqué au chapitre 3, les données récentes (c'est-à-dire celles qui sont rattachées aux séjours de l'année en cours) sont rechargées quotidiennement, tandis que les données historiques, moins fréquemment modifiées dans le système de production, sont rechargées mensuellement ou pluri-annuellement.

Au cours du processus d'intégration des données patients dans l'EDS i2b2 du CHU de Bordeaux, une étape de **pseudonymisation** des données a été implémentée. Cette pseudonymisation concerne :

— Les **clés patients** (Numéro Identifiant Patient (NIP) ; également appelé Identifiant Permanent du Patient (IPP)) et clés séjours (Numéro de Dossier Administratif (NDA) ; également appelé Identifiant d'Épisode du Patient (IEP)). Afin de permettre la ré-identification des patients (ex : si l'on souhaite proposer à

3. Un triplet RDF est une structure de données composée de trois éléments : un sujet, un prédicat et un objet, qui ensemble décrivent une relation entre deux ressources.

4. L'Insee met à disposition mensuellement une nouvelle version du fichier des personnes décédées ; les données sont en général consolidées sous 3 mois.

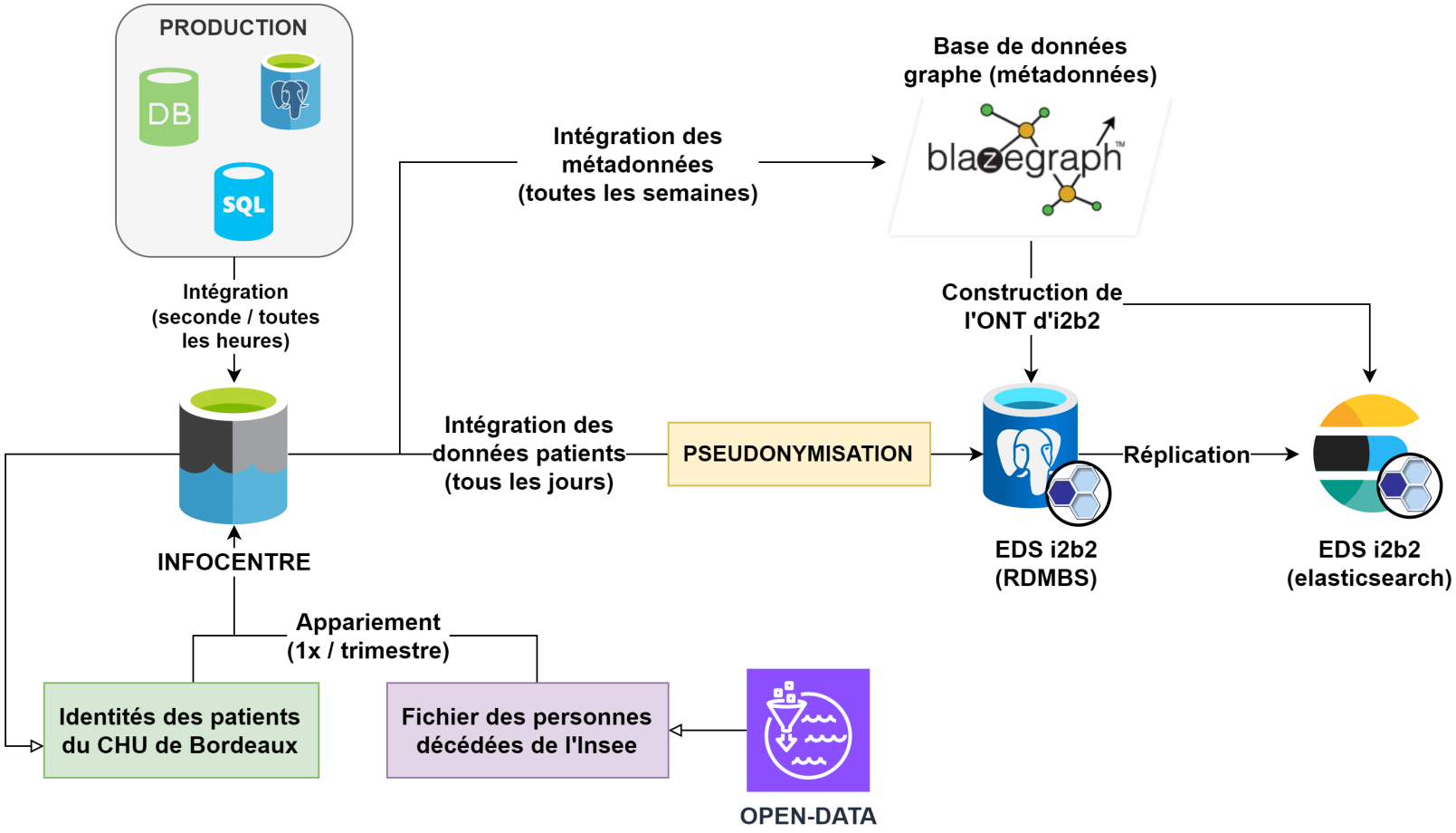


FIGURE 4.2 – Alimentation de l'EDS i2b2 du CHU de Bordeaux

un patient son inclusion dans une étude prospective, si un patient veut exercer ses droits quant aux données le concernant, etc.), des tables de correspondance spécifiques entre ces identifiants et leur équivalent pseudonymisé dans l'EDS i2b2 ont été mises en place ;

— Les **données en texte libre** contenant des données directement identifiantes (nom, prénom, date de naissance, adresse, numéro de sécurité sociale, etc.). Un algorithme de pseudonymisation à base de règles a été développé, permettant une pseudonymisation des documents après extraction du contenu texte des fichiers PDF et des documents « Word » (.doc, .docx, .odt, etc.).

L'ETL en charge de l'alimentation de l'EDS i2b2 du CHU de Bordeaux est développé via l'outil Talend Open Studio [5] et orchestré par Apache Airflow. [6]

En plus de l'EDS i2b2 persisté dans une base de données relationnelle (EDS i2b2 RDBMS), une version persistée dans Elasticsearch est également alimentée lors du processus d'ETL. Les travaux spécifiques ayant conduit à cette persistance des données du modèle i2b2 dans Elasticsearch sont détaillés dans le chapitre 3.

### 4.3.2 Architecture de l'EDS i2b2

Outre les deux bases de données qui constituent l'EDS i2b2 du CHU de Bordeaux, une application spécifique a été développée par l'équipe en charge de la mise en œuvre de l'EDS afin de faciliter les interactions avec les bases de données. L'ensemble des développements réalisés sont basés sur des technologies **open-source**. Le code produit sera prochainement mis à disposition, sous licence Apache 2 [7] :

— Côté **serveur** (*backend*), il s'agit d'une architecture en micro-services développée en Java, avec le *framework* « Spring Boot » [8], exposant des services au travers d'une couche d'API RESTfull. Les API sont sécurisées au travers d'un serveur d'authentification et d'habilitation, Keycloak [9], connecté à l'annuaire des professionnels du CHU de Bordeaux (*Lightweight Directory Access Protocol* ou LDAP) et implémentant le standard OIDC (*OpenID Connect*) [10].

5. La documentation de Talend est disponible via `https://www.talend.com/products/talend-open-studio/`
6. La documentation d'Apache Airflow est disponible via `https://airflow.apache.org/`
7. La licence Apache 2 est disponible via `https://www.apache.org/licenses/LICENSE-2.0`
8. La documentation de Spring Boot est disponible via `https://spring.io/projects/spring-boot`
9. La documentation de Keycloak est disponible via `https://www.keycloak.org/documentation`.
10. OIDC correspond à une couche d'identification basée sur le protocole OAuth 2.0. Dans OIDC, le client (couche d'API) vérifie l'identité d'un utilisateur final en se basant sur l'authentification fournie par un serveur d'autorisation (dans notre cas Keycloak)

— Côté **interface utilisateur graphique** (*Graphic User Interface* (GUI), également appelée *front-end*), une application Web à destination des utilisateurs finaux (médecin, infirmier, attaché de recherche clinique, etc.) est développée en Javascript au travers du *framework* React [11].

Les interactions réalisées côté *front-end* correspondent en réalité à des appels de la couche d'API exposant les services du *back-end*. De la même manière, cette couche d'API peut être interrogée de manière programmatique (via différents langages de programmation comme R ou Python) pour des utilisations plus avancées, dans des environnements dédiés et maîtrisés. C'est la même couche d'API, sécurisée par les mêmes mécanismes, qui est appelée lors de ces deux types d'interactions.

### 4.3.3 Accès aux données contenues dans l'EDS i2b2

Plus qu'une simple base de données, un EDS correspond à une plateforme complète favorisant l'utilisation secondaire des données de santé dans un environnement sécurisé, conforme à la réglementation en vigueur sur les données à caractère personnel.

L'EDS i2b2 du CHU de Bordeaux est dédié aux **usages locaux**, pour la recherche, les études et l'évaluation, mais aussi en support des activités de routine de l'établissement (ex : vigilances, codage PMSI, etc.). L'ensemble des services disponibles au-dessus de la base de données de l'EDS i2b2 est proposé au travers d'un portail web, disponible uniquement sur le réseau interne du CHU de Bordeaux, pour des utilisateurs authentifiés dans l'annuaire des personnels.

Pour la partie « recherche, étude et évaluation », en particulier dans le cadre de recherches n'impliquant pas la personne humaine (RNIPH), le portail de l'EDS du CHU de Bordeaux permet (Figure 4.3) :

— La déclaration d'un projet : déclaration institutionnelle, enregistrement auprès du DPO et demande d'avis au CSE local de l'établissement ;
— L'identification de patients / séjours éligibles pour participer à une étude (phénotypage) et inclusion de ces patients / séjours dans la cohorte du projet ;
— L'accès aux données individuelles pseudonymisées relatives aux patients / séjours de la cohorte d'un projet.
— L'accès à des environnements d'analyse spécifiques à chaque projet.

11. La documentation de React est disponible via `https://react.dev`

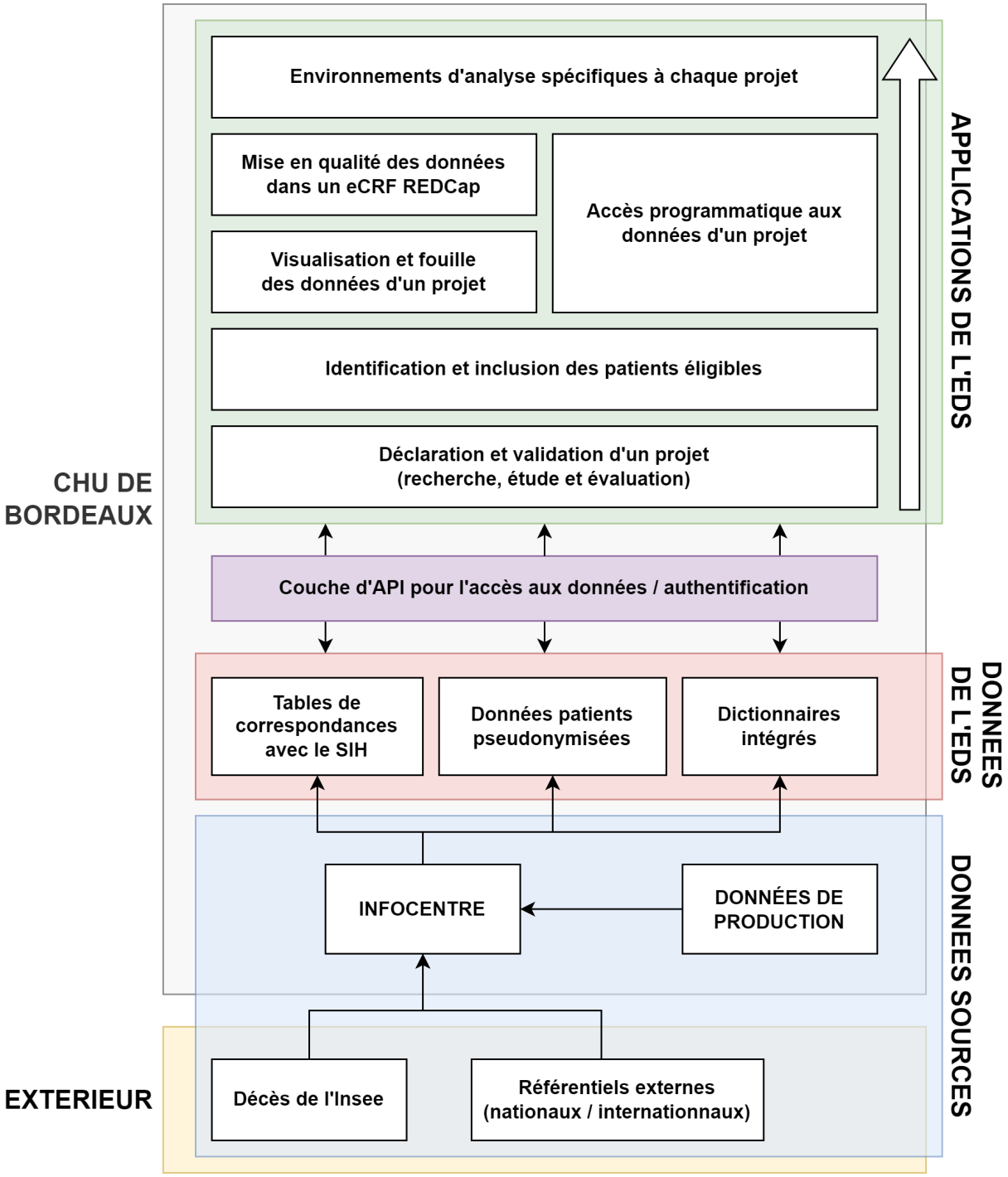


FIGURE 4.3 – Vue macroscopique des services mis à disposition sur le portail de l'EDS i2b2 du CHU de Bordeaux

#### *4.3.3.1 Circuit d'accès aux données à caractère personnel contenues dans l'EDS i2b2*

Dans le cas des RNIPH, le portail de l'EDS permet aux investigateurs de centraliser l'ensemble des formalités technico-réglementaires nécessaires pour la réalisation de leur étude. Deux types de RNIPH différentes peuvent être déclarées sur le portail de l'EDS i2b2 :

1. Les **RNIPH mono-service**, correspondant à des études internes. Ces études sont réglementées par l'article 65-2 de la loi informatique et libertés [89] ;
2. Les **RNIPH multi-services**, correspondant à des études impliquant des patients pris en charge par une autre équipe de soins que celle de l'investigateur principal. Pour ces études, la méthodologie de référence MR-004 s'applique.

En plus de la déclaration institutionnelle, de l'enregistrement auprès du DPO et de la demande d'avis auprès du CSE, le portail de l'EDS permet la saisie d'un résumé « grand public » du projet à destination des usagers. Ce résumé, avec certains éléments saisis dans le portail de l'EDS, alimente le **portail de transparence du CHU de Bordeaux** [12]. Ce dernier permet :

— Une **information individuelle par finalité**. Cela permet aux usagers de prendre connaissance de la liste des études en cours ou terminées au CHU de Bordeaux ;

— L'**exercice des droits**. En effet, les usagers peuvent, depuis ce portail de transparence, contacter l'investigateur principal ou le DPO, notamment pour exercer leur droit d'opposition à la participation à une étude.

Eu plus du portail de transparence du CHU de Bordeaux qui permet d'informer les usagers par finalités, un socle d'information générale a été mis en place au CHU de Bordeaux. Il est détaillé en annexe C.

Une fois l'ensemble des démarches techino-réglementaires validé, le projet de recherche peut démarrer.

#### *4.3.3.2 Identification des patients éligibles à la recherche*

La première étape d'un projet de recherche est l'identification des patients éligibles. Deux méthodes sont proposées pour identifier les patients susceptibles de participer à une étude :

— Si l'équipe investigatrice dispose déjà d'une liste de patients ou de séjours d'intérêt pour son étude, il est possible d'alimenter la cohorte du projet sur la base de cette liste ;

12. Le portail de transparence du CHU de Bordeaux est disponible via : `https://www.chu-bordeaux.fr/Professionnels-recherche/Recherche-clinique-et-Innovation/Participer-Ãă-une-recherche-clinique/Portail-de-transparence/`

— Si l'équipe investigatrice ne dispose pas d'une telle liste, le portail de l'EDS i2b2 met à disposition un outil pour le phénotypage des patients : le « requêteur ».

L'identification de patients ou de séjours sur la base de critères clinico-biologiques dans l'EDS i2b2 est un processus itératif, en particulier du fait de la très grande taille du vocabulaire à explorer (pour rappel, on retrouve 979 176 data element dans l'EDS i2b2 du CHU de Bordeaux) et de la multiplicité des sources de données pouvant contenir de l'information pertinente. Dans l'EDS i2b2, le nombre de requêtes moyen par projet est de 6,21 (sd = 8,91) et le nombre de requêtes médian par projet de 3 (IQR de [2 ; 8]) avec un maximum de 68 requêtes.

C'est le caractère itératif de cette étape de phénotypage, couplé avec notre objectif d'autonomisation des utilisateurs finaux, qui nous a poussé à chercher à améliorer les performances de l'interrogation de la base de données i2b2, et donc de se poser la question d'une persistance des données en NoSQL. Notre objectif était d'avoir des temps de réponse de l'ordre de la seconde au niveau du requêteur. Ce travail a été décrit dans le chapitre 3.

Le requêteur est un outil similaire au client Web proposé avec l'application i2b2 et présenté dans l'introduction (Figure 1.3 introduite lors de la présentation générale d'i2b2 dans la section 1.3.1.1). Le requêteur a été développé afin de répondre aux usages spécifiques locaux, et interroge les données de l'EDS i2b2 persistées dans Elasticsearch.

Le requêteur permet l'identification de patients / séjours sur la base de critères clinico-biologiques. Plus précisément, il permet de réaliser des requêtes booléennes (« ET » / « OU ») interrogeant les données structurées, numériques ou en texte libre intégrées dans l'EDS i2b2. Il est également possible d'appliquer des filtres sur l'âge ou le sexe du patient, ainsi que des filtres liés à la temporalité (date de début ou date de fin) des observations.

Que les patients aient été importés depuis une liste ou depuis le requêteur, une couche de sécurité s'applique si le projet correspond à une RNIPH mono-service, filtrant automatiquement les patients n'appartenant pas au périmètre de l'équipe de soins de l'investigateur principal du projet. Cette couche de sécurité dépend des droits de l'investigateur principal dans le DPI. Quel que soit le membre de l'équipe investigatrice habilité au sein du projet à exécuter des requêtes, ce sont les droits de l'investigateur principal qui sont utilisés pour la couche de sécurité. Un exemple de requête avec application de la couche de sécurité sur la base des droits de l'investigateur principal est présenté dans la Figure 4.4.

Après cette étape de phénotypage, l'équipe investigatrice peut alimenter la cohorte du projet avec la liste des patients / séjours identifiés. Seuls les nouveaux patients / séjours

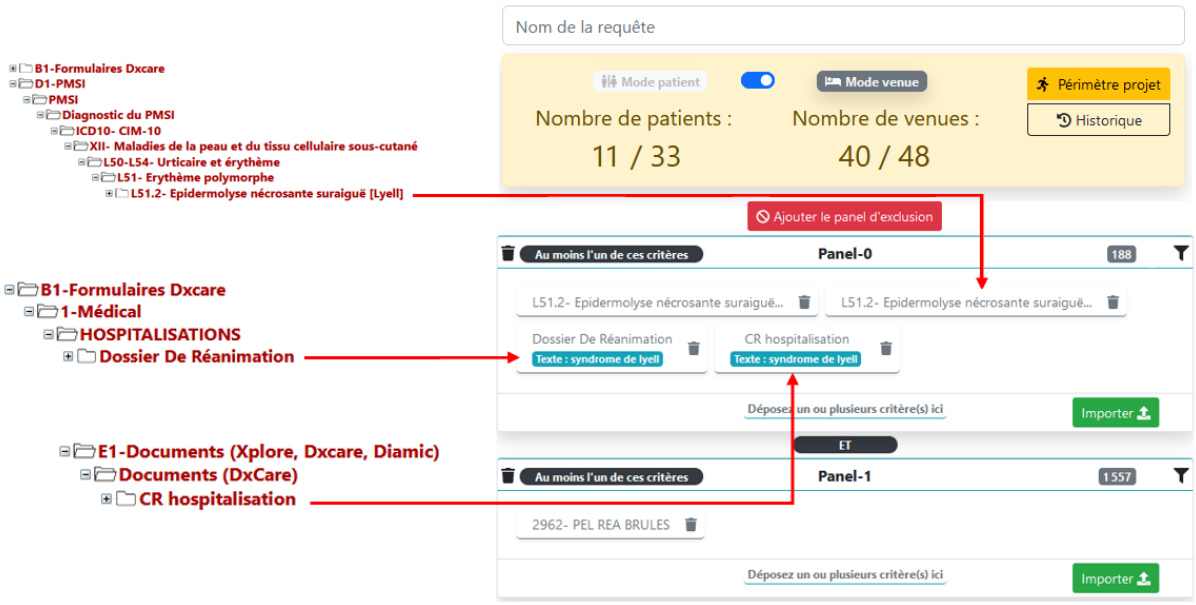


FIGURE 4.4 – Requêteur de l'EDS i2b2, pour le phénotypage des patients / séjours sur la base de critères clinico-biologiques

Requête combinant deux panneaux (étiquetés « Panel-0 » et « Panel-1 ») : i) le premier panneau permet de retrouver les patients ayant présenté un syndrome de Lyell au travers de données structurées codées en CIM-10 (dans les formulaires du DPI ou dans le PMSI) ou ayant une mention de « syndrome de Lyell » dans un compte rendu d'hospitalisation ou dans le formulaire spécifique à la réanimation (188 patients distincts identifiés) ; ii) le second panneau recherche les patients passés dans l'unité de réanimation des grands brûlés (1 557 patients distincts identifiés). La requête est exécutée en « Mode venue », indiquant que les critères des différents panneaux doivent être évalués au cours de la même venue. Au total dans l'EDS i2b2, 33 patients correspondant à 48 venues ont été identifiés, parmi lesquels respectivement 11 patients et 40 venues sont accessibles dans le périmètre mono-service de l'équipe de soins de l'investigateur principal.

sont importés dans la cohorte (si un patient était déjà présent dans la cohorte du fait d'un précédent import, il n'est pas ajouté une seconde fois).

Lors de ce processus d'inclusion des patients dans un projet, une pseudonymisation propre au projet est réalisée, permettant d'assurer qu'un même patient, inclus dans deux projets différents, se verra attribuer deux identifiants pseudonymes différents pour ces deux projets. Une table de correspondance entre les identifiants pseudonymes de l'EDS i2b2 et les identifiants pseudonymes internes au projet est conservée, permettant la ré-identification des patients si nécessaire.

#### *4.3.3.3 Accès aux données individuelles pseudonymisées des patients de la cohorte d'un projet*

Dans le contexte d'un projet déclaré et validé, l'accès aux données pseudonymisées individuelles patients peut se faire de deux manières pour les personnels habilités dans le projet :

1. Au travers d'une **visionneuse** du dossier médical d'un patient intégré dans l'EDS i2b2. Ce sont les modalités d'accès privilégiées pour les utilisateurs métiers issus du domaine de la santé (médecin, infirmier, attaché de recherche clinique, etc.) ;
2. **De manière programmatique**, via l'interrogation de la couche d'API disponible au-dessus des bases de données de l'EDS i2b2, pour les utilisateurs avancés (bio-informaticien, data-manager, data-scientist, etc.).

**Accès aux données patients au travers d'une visionneuse du dossier médical**

La visionneuse qui a été déployée au-dessus de l'EDS i2b2 est basée sur SmartCRF [192], qui a été développé lors de travaux antérieurs. Il s'agit d'un outil de visualisation du dossier médical des patients (ou des séjours) inclus dans une cohorte de l'EDS. La visualisation se fait patient par patient (respectivement séjour par séjour).

Côté **serveur** (*back-end*), la visionneuse interroge des services en charge de la récupération et de la pseudonymisation (spécifique dans le contexte du projet) des données du dossier médical d'un patient (ou d'un séjour) inclus dans une cohorte. Le service interrogé par la visionneuse, en plus de récupérer les observations contenues dans l'EDS i2b2, les expose au format HL7 FHIR (*Fast Healthcare Interoperability Resources*) [13]. Ainsi, la visionneuse consomme des données dans un format standard, permettant de la déployer sur des infrastructures différentes de celles disponibles au CHU de Bordeaux. Également, le fait que le service de récupération des données les expose en FHIR pourrait permettre dans le futur de brancher un autre outil que celui développé localement.

13. HL7 FHIR est un standard d'interopérabilité pour les échanges d'informations médicales. La spécification du standard est disponible via : `https://www.hl7.org/fhir/`.

Côté **interface utilisateur graphique** (*front-end*), la visionneuse est composée de cinq parties :

1. Un **moteur de recherche** (Figure 4.5), au niveau duquel il est possible d'effectuer une recherche plein texte dans le dossier du patient. Le moteur de recherche propose une auto-complétion basée sur les données disponibles, ainsi qu'un auto-complétion basée sur les métadonnées intégrées dans l'EDS i2b2. Le moteur de recherche peut être alimenté par des « termes de recommandation », qui correspondent à des termes pré-enregistrés au niveau de l'ensemble de la cohorte. Lorsqu'un terme est recherché dans le moteur de recherche, un extrait des différentes ressources FHIR contenant ce terme est affiché pour aider l'utilisateur final à identifier la ressource d'intérêt.

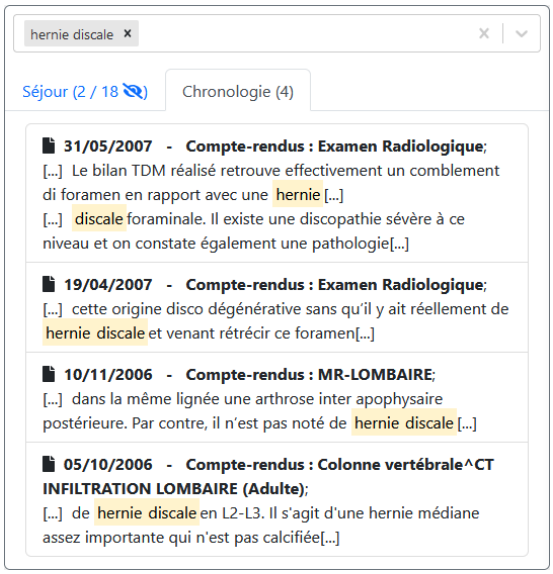


FIGURE 4.5 – Visionneuse du dossier médical intégré dans l'EDS i2b2 - Recherche plein texte

Le terme « hernie discale » a été recherché dans le moteur de recherche. Un extrait des documents qui contiennent ce terme est présenté dans la vue « Chronologie ». La vue « Séjour » est détaillée plus loin.

2. Un **explorateur des données du dossier patient** (séjour), **sous la forme d'un arbre** (Figure 4.6, bloc inférieur gauche). Ce dernier reprend l'organisation proposée dans le DPI, les données étant organisées par venue et par ressource FHIR (documents, formulaires, etc.). Il est possible de filtrer l'arbre par ressource FHIR ou par niveau hiérarchique, issu de l'intégration des métadonnées faite dans i2b2.

3. Le **bloc de consultation des données patients** (Figure 4.6, bloc inférieur droit), au niveau duquel les utilisateurs finaux consultent les différentes ressources FHIR. Des vues différentes sont proposées en fonction des types de ressources

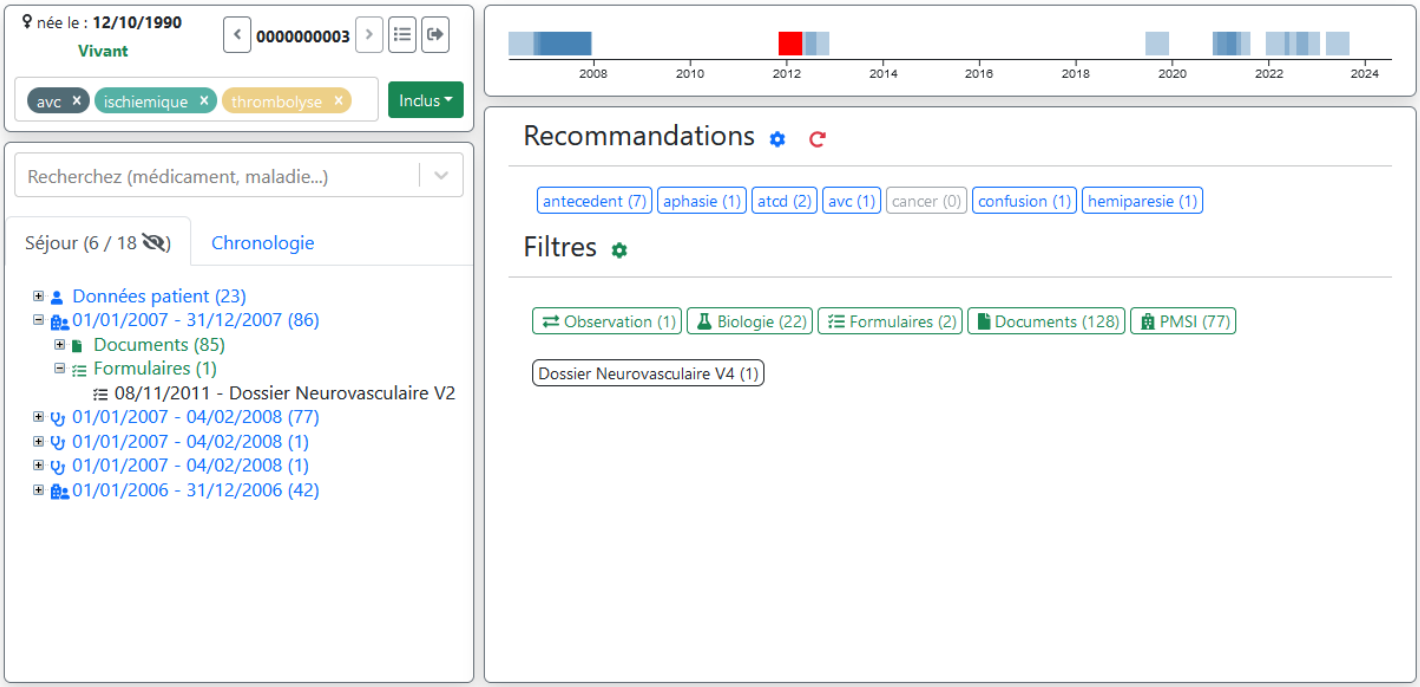


FIGURE 4.6 – Visionneuse du dossier médical intégré dans l'EDS i2b2

FHIR (ces vues sont disponibles en annexe D : Figure D.1, Figure D.2, Figure D.3 et Figure D.4).

4. Une **ligne de vie** (Figure 4.6, bloc supérieur droit), indicateur visuel de la temporalité des données disponibles d'une part et des données respectant les critères de recherche d'autre part. L'élément en cours de visualisation est affiché en rouge tandis que l'ensemble des éléments correspondant aux critères de recherche est affiché en bleu. A la différence de SmartCRF, cette ligne de vie n'est pas un outil pour la navigation dans le dossier médical mais uniquement un indicateur visuel.

5. Un **bloc d'annotation** du patient / séjour (Figure 4.6, bloc supérieur gauche). Trois types d'annotation sont possibles :
   (a) Une classification du dossier parmi : « Inclus », « Exclus », « A revoir » et « Non vu ». Ces annotations sont utilisées pour classer de manière macroscopique les patients / séjours.
   (b) Des tags du dossier (étiquettes). Les tags correspondent à des annotations libres du dossier du patient / séjour. Une liste de tags est gérée au niveau de la cohorte pour permettre aux utilisateurs d'utiliser des tags identiques pour l'ensemble de la cohorte. Ces tags sont utilisés pour classer les patients / séjours de manière plus microscopique que la classification précédente.

(c) La possibilité de saisir de l'information détaillée sur les patients / séjours dans un e-CRF adossé à la visionneuse. Le moteur d'e-CRF adossé à l'EDS i2b2 est REDCap [14]. Ce dernier permet de constituer, dans le cas où les données du dossier médical doivent être consolidées manuellement, les datamarts [15] des projets. L'e-CRF REDCap est accessible depuis la visionneuse en cliquant sur le lien correspondant à l'identifiant pseudonyme du patient dans le projet.

**Accès au données patients de manière programmatique**

Un accès aux données intégrées dans l'EDS i2b2 est également possible de manière programmatique, dans des environnements d'analyse dédiés, au travers de *Jupyter Notebook*. Des clients R et Python ont été développés pour permettre une interrogation facilitée de la couche d'API développée au-dessus des bases de données de l'EDS i2b2. Ces clients permettent de récupérer des données individuelles pseudonymisées dans le contexte d'un projet.

#### *4.3.3.4 Environnements d'analyse*

Les environnements d'analyse proposés pour l'EDS i2b2 du CHU de Bordeaux sont des *Jupyter Notebook* segmentés par projet. À partir de ces environnements d'analyse, il est possible :

— D'**interroger les données contenues dans l'e-CRF REDCap adossé au projet**. Les environnements d'analyse permettent ici de réaliser les analyses statistiques du projet dans des environnements internes à l'EDS et conformes à la réglementation. Seuls les résultats de ces analyses, correspondant à des résultats agrégés anonymes, peuvent être exportés de ces environnements ;

— D'**interroger l'EDS i2b2 de manière programmatique**, centré sur le projet en cours. Les données pseudonymisées reçues peuvent alors être traitées directement dans l'environnement d'analyse, ou bien après une étape intermédiaire de stockage dans l'e-CRF REDCap.

Ce processus est souvent réalisé pour les données de biologie, pour laquelle l'étape de *data-management* consiste la plupart du temps à filtrer les données pour ne garder que les données d'intérêt (ex : la biologie avant et après une chirurgie). Après cette étape de data-management, la biologie d'intérêt peut être stockée dans l'e-CRF REDCap adossé au projet ou bien directement dans l'environnement d'analyse (base de données, fichier .csv, etc.) ;

14. La documentation de REDCap est accessibe via `https://www.project-redcap.org/`
15. Les datamarts sont des sous-ensembles spécialisés d'un entrepôt de données. Il s'agit de la base de données qui sera accessible par les équipes investigatrices pour réaliser les analyses.

— D'**effectuer des analyses plus complexes**, comme des analyses d'images ou du TAL.

Les environnements d'analyse sont en cours de construction et ne sont pas encore en production au CHU de Bordeaux.

### 4.3.4 Utilisation de l'EDS i2b2 du CHU de Bordeaux

Les métriques d'utilisation de la plateforme ont été relevées en septembre 2024.

493 utilisateurs distincts s'étaient connectés au moins une fois sur le portail de l'EDS i2b2 du CHU de Bordeaux. Parmi eux, 113 ont été positionnés comme investigateur principal d'un projet.

231 projets sont en cours sur la plateforme, et 13 sont terminés. Le nombre de projets terminés est largement sous-estimé, les utilisateurs devant effectuer une action manuelle pour enregistrer le fait qu'un projet est terminé. La distribution du nombre de projets créés au cours du temps est présentée en Figure 4.7.

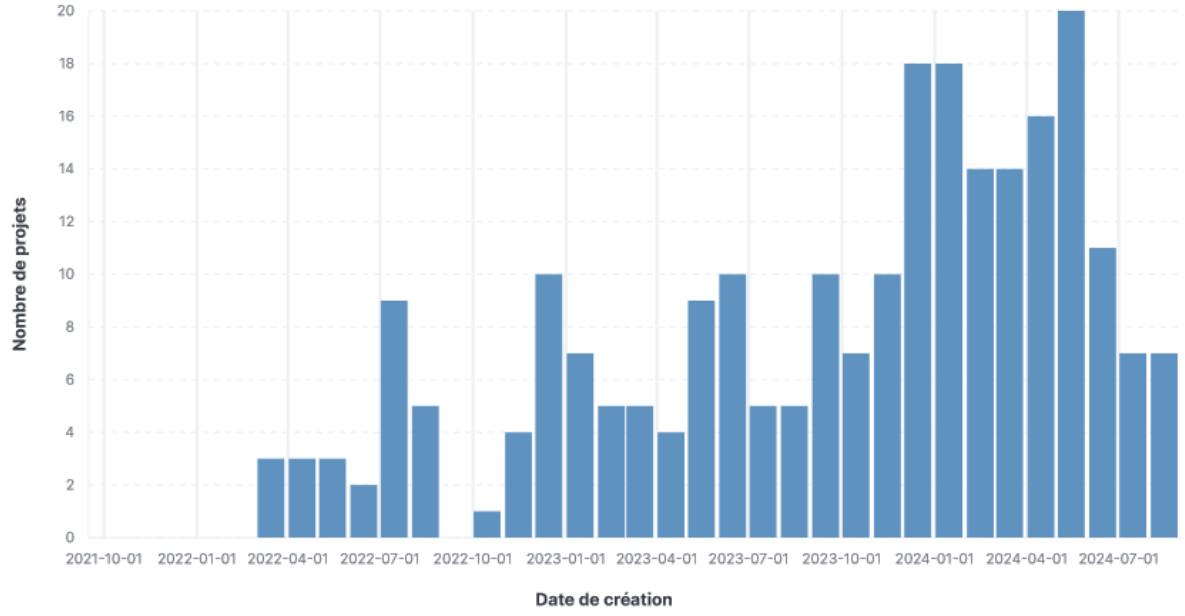


FIGURE 4.7 – Nombre de projets (en cours et terminés) déclarés sur la plateforme de l'EDS i2b2 du CHU de Bordeaux

Parmi les 244 projets en cours ou terminés, 156 ont inclus au moins 1 patient ou 1 séjour (60%) et 43 sont liés à un e-CRF REDCap (17,5%).

Si on s'intéresse aux types de projets déclarés :

— 166 sont des RNIPH mono-service (68%) ;

— 56 sont des RNIPH multi-services ou multi-établissements (23%) ;
— 17 sont des études en support de la pertinence des soins ou de vigilance (7%).

Le 10 principaux domaines médicaux couverts par les projets de recherche en cours ou terminés sur l'EDS i2b2 sont présentés dans le Tableau 4.2, la cardiologie et la santé publique étant les spécialités les plus représentées.

Tableau 4.2 – TOP-10 des domaines couverts par les projets de recherche (en cours et terminés) déclarés sur le portail de l'EDS i2b2

| Spécialité | Nombre de projets |
|---|---|
| Cardiologie | 57 |
| Santé publique | 34 |
| Cancérologie | 18 |
| Pédiatrie | 18 |
| Anesthésie | 16 |
| Dermatologie | 12 |
| Neurologie | 12 |
| Chirurgie orthopédique | 10 |
| Médecine interne | 7 |
| Endocrinologie | 6 |

En ce qui concerne les études portée par la cardiologie, le nombre important peut être expliqué par la présence de l'IHU (Institut Hospitalo-Universitaire) Lyric, dédié aux maladies du rythme cardiaque, au niveau duquel des compétences techniques (data scientist) existent et ont permis une appropriation rapide des outils.

En ce qui concerne les études portées par la Santé Publique, le nombre est en grande partie expliqué par la participation du CHU de Bordeaux au Réseau DARWIN, au travers de son second EDS au format OMOP.

## 4.4 L'EDS OMOP : l'entrepôt pour les réseaux fédérés

En plus de l'EDS i2b2, un second EDS au format OMOP a été implémenté au CHU de Bordeaux.

OMOP-CDM étant à la fois un modèle d'intégration syntaxique et sémantique, il est beaucoup plus adapté qu'i2b2 pour la participation à des réseaux fédérés. Une présentation du modèle de données OMOP-CDM a été effectuée dans la section 1.3.1.2. Une présentation des réseaux fédérés a été réalisée dans la section 1.3.3.

L'EDS au format OMOP a été mis en place dans le contexte d'un financement EHDEN [16] et a été validé au sein du réseau EHDEN en 2022.

16. EHDEN à pour objectif de mettre en oeuvre une fédération d'EDS standardisés suivant le modèle de données OMOP-CDM

### 4.4.1 Intégration des données dans l'EDS OMOP

#### *4.4.1.1 Périmètre des données intégrées dans l'EDS OMOP*

L'EDS OMOP intègre le même périmètre de patients que l'EDS i2b2 : patients venus au moins une fois au CHU de Bordeaux, en hospitalisation ou en consultation, depuis 2010.

Les données individuelles patients intégrées concernent les dimensions suivantes :

— **Données patients** (tables `PERSON` et `DEATH`) ;
— **Données de venues** (tables `VISIT_OCCURRENCE` et `VISIT_DETAIL`), incluant les types de venues (hospitalisation, consultation, etc.) et les mouvements ;
— **Données relatives aux diagnostics** (table `CONDITION_OCCURRENCE`), incluant les diagnostics issus du PMSI mais aussi ceux issus de l'anatomo-pathologie ;
— **Données relatives aux actes** (table `PROCEDURE_OCCURRENCE`), correspondant aux actes issus du PMSI ;
— **Données relatives aux prescriptions et aux administrations médicamenteuses** (table `DRUG_EXPOSURE`), incluant les prescriptions standards mais également les traitements de chimiothérapie ;
— **Données relatives aux résultats de biologie** (table `MEASUREMENT`).

Le Tableau 4.3 présente les volumes de données disponibles dans l'EDS OMOP. Tout comme l'EDS i2b2, la source de données occupant le plus gros volume de données est la biologie. La seconde source la plus importante en terme de volume de données correspond aux médicaments.

Tableau 4.3 – Données intégrées dans l'EDS OMOP du CHU de Bordeaux

| Table | Nombre d'observations |
|---|---|
| `PERSON` | 2 296 220 |
| `DEATH` | 222 169 |
| `VISIT_OCCURRENCE` | 17 448 437 |
| `VISIT_DETAIL` | 20 142 056 |
| `CONDITION_OCCURRENCE` | 16 959 245 |
| `DRUG_EXPOSURE` | 120 849 521 |
| `PROCEDURE_OCCURRENCE` | 5 459 313 |
| `MEASUREMENT` | 607 355 508 |

En plus des données individuelles patients, les métadonnées (appelées « vocabulaires standards » ou *standardized vocabularies* dans OMOP) décrivant les données intégrées dans OMOP-CDM sont chargées dans l'EDS OMOP. On distingue :

— Les vocabulaires proposés par OHDSI [109], qui peuvent être explorés et téléchargés via le moteur de recherche Athena. Dans l'EDS OMOP du CHU de Bordeaux,

9 154 810 concepts sont intégrés depuis les vocabulaires proposés par OHDSI, dont 3 368 683 (36,8%) sont considérés comme standards.

— Les vocabulaires décrivant les données individuelles intégrées dans OMOP, correspondant aux métadonnées des sources de données intégrées dans l'EDS OMOP. Dans l'EDS OMOP, 576 969 concepts issus des différentes sources de données sont intégrés (6% des concepts disponibles).

Le résultat d'une analyse descriptive détaillée de l'EDS au format OMOP est présenté en annexe E (résultat du CdmOnboarding).

#### *4.4.1.2 Méthodes utilisées pour l'alimentation de l'EDS OMOP*

L'EDS OMOP du CHU de Bordeaux est alimenté depuis l'EDS i2b2 (Figure 4.8). En effet, une partie importante du travail d'intégration a déjà été réalisée au niveau de l'EDS i2b2 (dénormalisation des données, pseudonymisation, etc.). Ces éléments d'intégration étant communs avec les besoins pour l'EDS OMOP, il s'agissait donc ici de capitaliser sur le travail déjà réalisé précédemment.

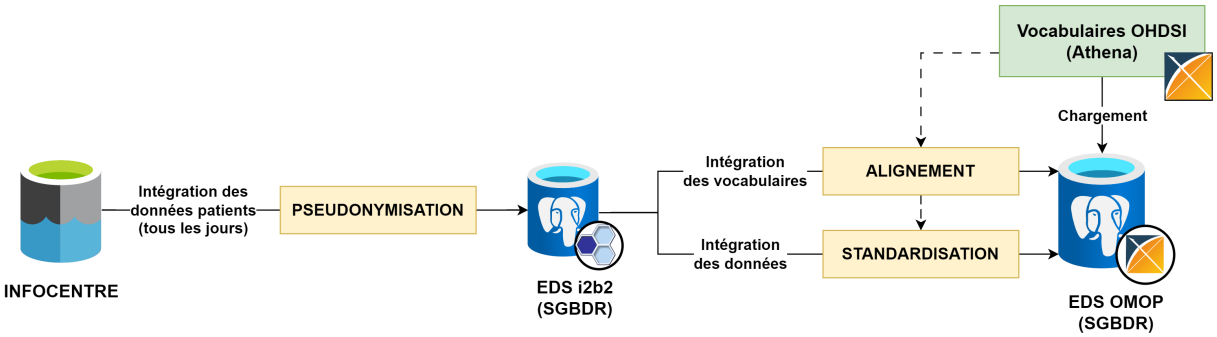


FIGURE 4.8 – Alimentation de l'EDS OMOP du CHU de Bordeaux

Dans l'EDS OMOP, en plus de l'étape d'intégration syntaxique, une étape d'intégration sémantique en lien avec le vocabulaire standard proposé par OHDSI est réalisée. L'objectif est ici d'**aligner 80% des observations intégrées dans le modèle OMOP-CDM vers des concepts standards d'OMOP**. Ce travail d'alignement est réalisé :

— Automatiquement sur la base des alignements disponibles dans le vocabulaire d'OHDSI. Par exemple, OHDSI propose un alignement entre les concepts de la CIM-10 internationnale et ceux disponibles dans la SNOMED-CT.

— Automatiquement sur la base d'alignements disponibles par ailleurs. En particulier en France, de nombreuses initiatives locales ont cherché à générer et mettre à disposition des alignements vers les concepts standards d'OMOP pour des vocabulaires standards utilisés en France. C'est le cas par exemple au niveau des médicaments, où des travaux précédents ont permis de mettre à disposition des

alignements entre des référentiels français et des référentiels internationaux [193]. Des travaux au niveau national, en lien avec l'« OMOP-isation du SNDS » [17], sont également en cours et devraient permettre à terme le partage d'alignements pour les terminologies standards françaises, en particulier pour la CCAM.

— Manuellement, en lien avec des besoins spécifiques au cours d'une étude. Il s'agit ici de réaliser des alignements *de novo* ou d'aligner plus précisément des concepts déjà alignés à un niveau plus précis, en lien avec une étude spécifique.

Les données individuelles patients ainsi que les métadonnées sont chargées 2 à 4 fois par an dans l'EDS OMOP, au sein d'une base de données relationnelle (PostgreSQL). Comme pour l'EDS i2b2, l'ETL utilisé pour le chargement de l'EDS OMOP a été développé avec Talend Open Studio.

### 4.4.2 Accès aux données contenues dans l'EDS OMOP

Le circuit de demande d'accès aux données de l'EDS OMOP est le même que celui pour la demande d'accès aux données contenues dans l'EDS i2b2 (déclaration des études sur le portail de l'EDS et validation des études par le CSE).

Cependant, à la différence de l'EDS i2b2 où les données sont exploitées en autonomie par les équipes investigatrices, les données contenues dans l'EDS OMOP sont exploitées par l'unité IAM (l'équipe en charge de la mise en œuvre de l'EDS du CHU de Bordeaux), en lien avec des cliniciens du CHU de Bordeaux, dans le cadre de la participation à des réseaux fédérés. L'exploitation des données se fait directement au niveau de la base de données PostgreSQL, avec exécution de scripts d'analyse développés de manière centralisée au niveau des réseaux fédérés. En particulier, le CHU de Bordeaux participe au réseau DARWIN-EU.

### 4.4.3 Utilisation de l'EDS OMOP du CHU de Bordeaux dans le cadre du réseau fédéré DARWIN-EU

Le réseau DARWIN-EU est un réseau fédéré d'EDS dont les données sont intégrées au format OMOP. Ce réseau est présenté en détail dans la section 1.3.3.2.

#### *4.4.3.1 Description du processus de recrutement d'un data partner (DP) dans le réseau DARWIN-EU*

Pour intégrer le réseau DARWIN-EU, une étape initiale de recrutement du CHU de Bordeaux en tant que DP a été réalisée. Cette étape correspond à la validation de la qualité

17. La documentation de transformation du SNDS au format OMOP est disponible via : `https://documentation-snds.health-data-hub.fr/omop/introduction/snds_omop.html`

des données intégrées dans l'EDS OMOP par la réalisation d'une analyse descriptive détaillée de la base de données. Cette analyse est réalisée au travers d'un package R, appelé CdmOnboarding [18]. Le résultat de la dernière analyse du CDM Onboarding est disponible en annexe E. Il s'agit de :

— **Décrire les données cliniques intégrées dans OMOP**, en terme de nombre d'observations par domaine, de nombre de patients concernés, de distribution des données au cours du temps et par source de données, etc. ;
— **Décrire les alignements vers les concepts standards d'OMOP** pour les différents domaines intégrés : pourcentage d'observations alignées, pourcentage de concepts alignés. Pour les médicaments, une description plus détaillée est réalisée, incluant en particulier la granularité des alignements disponibles ;
— **Décrire les résultats d'évaluation de la qualité des données** au travers de différents packages R. La qualité des données est évaluée en terme de complétude, de plausibilité et de conformité.

Une fois la qualité des données intégrées dans l'EDS OMOP du CHU de Bordeaux validée par DARWIN-CC et l'EMA, le processus de recrutement a abouti à la signature d'un accord cadre ("*Framework Agreement*" en anglais) entre le CHU de Bordeaux et le DARWIN-CC. Le CHU de Bordeaux a été sélectionné comme DP lors de la première vague de recrutement de ce réseau, en novembre 2022.

#### *4.4.3.2 Description du déroulé d'une étude dans DARWIN-EU*

La réalisation des études au sein du réseau DARWIN-EU estest composée de 4 étapes :

1. **Sélection des DP pour l'étude et la gestion des aspects technico-réglementaires** :

   — DARWIN-CC évalue, pour chaque étude, quels DP seraient pertinents pour participer à l'étude. Cette évaluation est en particulier basée sur des données agrégées décrivant la base de données source que le CHU de Bordeaux (et plus largement l'ensemble des DP) met à disposition du DARWIN-CC après chaque mise à jour de l'EDS OMOP ;
   — Si l'EDS OMOP du CHU de Bordeaux est identifié par le DARWIN-CC comme base de données pertinente pour une étude, DARWIN-CC propose au CHU de Bordeaux de participer à l'étude. Le CHU de Bordeaux est libre, pour chaque étude, de participer ou non (en particulier en fonction de la disponibilité des données nécessaires à la conduite de l'étude, en fonction de l'intérêt des cliniciens sur la thématique) ;
   — Si le CHU de Bordeaux accepte de participer à une étude, DARWIN-CC transmet le protocole de l'étude, validé par l'EMA. Ce protocole est transmis

18. Le package CdmOnboarding est disponible via : `https://github.com/darwin-eu/CdmOnboarding`

au CSE du CHU de Bordeaux pour avis. En parallèle, un *Work-Order* de l'étude est signé entre le CHU de Bordeaux et DARWIN-CC.

2. **Exécution locale des scripts d'analyse de l'EDS OMOP**. Une fois l'ensemble des étapes technico-réglementaires locales réalisées, l'étude peut démarrer. DARWIN-CC met à disposition des différents DP participant à l'étude les scripts d'analyse de l'étude au sein d'un package R dédié à l'étude. En règle générale, deux types de scripts différents sont exécutés au cours du temps :
   — Au cours d'une phase initiale, des scripts de diagnostics sont exécutés au niveau des base de données de chaque DP. Ces derniers permettent de décrire finement les données disponibles localement afin de guider la programmation du script principal d'analyse ;
   — Dans une seconde phase, le script principal d'analyse est exécuté sur la base de données de l'EDS OMOP. Ce script est en charge de la production des résultats agrégés correspondant aux objectifs principaux et secondaires de l'étude. Les résultats produits sont évalués localement (DARWIN-CC met à disposition, dans le package R dédié à l'étude, un outil permettant une visualisation des résultats produits) et, s'ils sont jugés corrects, sont envoyés au niveau du DARWIN-CC.
3. **Agrégation et interprétation des résultats au niveau du DARWIN-CC** :
   — Les résultats agrégés des différents DP produits localement sont transmis au DARWIN-CC. Ces données sont intégrées dans des outils graphiques permettant l'exploration des données de tout ou partie des DP. L'exploration centralisée des résultats au niveau du DARWIN-CC permet souvent d'identifier des erreurs au niveau des scripts d'analyse. Dans ce cas, une nouvelle version de ces scripts est alors produite au niveau du DARWIN-CC et ré-exécutée localement (processus itératif) ;
   — Au cours de cette étape, les DP sont impliqués dans l'interprétation des résultats (ex : dans le cas où les résultats d'un DP sont différents des autres, le DP est sollicité pour identifier la cause de cette différence (différence de pratique, biais dans les données, etc.)).
4. **Rédaction d'un rapport et diffusion des résultats** : à l'issue de l'exécution des scripts d'analyse et de l'interprétation des résultats, un rapport d'analyse est rédigé et transmis à l'EMA. L'étude, ainsi que le rapport, sont également mis à disposition du grand public au niveau du catalogue des études sur les RWD (*Catalogue of RWD studies*) de l'EMA [19].

La liste des études dans lesquelles le CHU de Bordeaux est impliqué est fournie dans le Tableau 4.4. Parmi les 30 études qui ont été réalisées par DARWIN-EU depuis sa mise

19. Le *Catalogue of RWD studies* est disponible via : `https://catalogues.ema.europa.eu/catalogue-rwd-studies`

en place en 2021, **14 études ont impliquée l'EDS OMOP du CHU de Bordeaux**, parmi lesquelles 9 sont terminées. L'ensemble des études réalisées étaient des études « prêtes à l'emploi » (*Off-The-Shelf Studies* en anglais)[20], hormis une étude qui a été caractérisée comme complexe par DARWIN-CC (« P2-C3-003 » dans le Tableau 4.4).

Tableau 4.4 – Liste des études impliquant le CHU de Bordeaux dans DARWIN-EU

| Code | Nom | Statut |
|---|---|---|
| P1-C1-003 | Drug utilization study of Antibiotics in the 'Watch' category of the WHO AWaRe classification of antibiotics for evaluation and monitoring of use | Terminée |
| P2-C1-001 | Multiple myeloma: patient characterisation, treatments and survival in the period 2012-2022 | Terminée |
| P2-C1-002 | Drug utilization study of prescription opioids | Terminée |
| P2-C1-003 | Co-prescribing of endothelin receptor antagonists (ERAs) and phosphodiesterate-5 inhibitors (PDE-5is) in pulmonary arterial hypertension (PAH) | Terminée |
| P2-C1-005 | Drug utilisation study of medicines with prokinetic properties in children and adults diagnosed with gastroparesis | Terminée |
| P2-C1-006 | Treatment patterns of drugs used in adult and paediatric population with systemic lupus erythematosus | Terminée |
| P2-C1-007 | Natural history of dermatomyositis (DM) and polymyositis (PM) in adults and paediatric populations | Terminée |
| P2-C1-010 | Characterization of patients with chronic hepatitis B and C | Terminée |
| P2-C1-011 | Age specific incidence rates of RSV related disease in Europe | Terminée |
| P2-C1-014 | Monitoring prescription of essential medicines administered in ICU | En cours |
| P2-C3-003 | Overall survival in patients with locally advanced or metastatic non-small cell lung cancer treated with selected immunotherapies as first line of treatment | En cours |
| P3-C1-003 | Chondrosarcoma: patient demographics, treatments, and survival in the period 2010-2023 | En cours |
| P3-C1-005 | Characterising interstitial lung disease in Europe | En cours |
| P3-C1-007 | Paracetamol prescribing and paracetamol overdose in Europe: a descriptive analysis of trends and patient characteristics | En cours |

## 4.5 Discussion

Le CHU de Bordeaux a mis en place depuis 2018 un EDS pour les recherches, études et évaluations, mais aussi en support des activités de routine de l'établissement (vigilance, pertinence et qualité des soins, etc.). Les efforts initiaux ont consisté en l'intégration des données des différentes applications métiers du SIH, conduisant à l'intégration de plus de 20 sources de données différentes dans l'EDS i2b2.

20. Il s'agit principalement de questions de caractérisation qui peuvent être adressées à l'aide de protocoles d'analyse génériques. C'est le cas notamment d'études sur l'épidémiologie des maladies (ex : estimation de la prévalence ou de l'incidence des maladies dans des périodes et des groupes de population définis) ou d'études sur la consommation de médicaments à l'échelle d'une population.

La maîtrise complète des flux est un élément essentiel dans la mise en place d'un EDS. Cette maîtrise permet :

1. D'**éviter les « boîtes-noires »**, garantissant l'explicabilité des résultats obtenus à partir des EDS. La maîtrise des flux permet de revenir facilement au niveau SIH source pour valider la qualité des données intégrées (*data-lineage*).

2. D'être **agile sur l'intégration de nouvelles sources de données**, ou sur l'intégration de données complémentaires à partir de sources déjà intégrées. Par exemple, en lien avec les besoins spécifiques d'un projet de recherche, les données de biologie internes aux laboratoires du CHU de Bordeaux (données non disponibles dans le DPI) ont été intégrées dans l'EDS i2b2 en quelques jours [21].

3. De **changer facilement de modèle d'intégration / d'exposition des données** :

   — Dans le cas du CHU de Bordeaux, cette maîtrise des flux pour alimenter l'EDS i2b2 a permis une mise en place rapide de l'EDS au format OMOP, faisant du CHU de Bordeaux le premier CHU en France à avoir été validé dans le cadre de EHDEN. Le choix a été fait d'utiliser l'EDS i2b2 comme source de données pour l'EDS OMOP, permettant de capitaliser sur le premier niveau d'intégration déjà réalisé dans l'EDS i2b2 ;

   — Également, cette maîtrise a permis au CHU de Bordeaux de participer au réseau fédéré 4CE, au niveau duquel un format spécifique d'intégration des données était proposé. Plus que le modèle d'intégration initial (i2b2), c'est la réactivité de l'équipe en charge de la mise en œuvre de l'EDS pour passer de ce modèle d'intégration à un modèle d'exposition différent qui a été primordiale dans la capacité du CHU de Bordeaux à intégrer ce réseau. Cette réactivité a permis au CHU de Bordeaux de partager des données agrégées dès la première phase de 4CE. *In fine*, les données du CHU de Bordeaux ont été impliquées dans 10 études au sein de ce réseau [5, 3, 119, 121, 122, 123, 124, 125, 127, 129].

Par ailleurs, le pilotage et la mise en œuvre de l'EDS par une **équipe médicale spécialisée en informatique médicale** est un facteur important dans le développement des EDS au CHU de Bordeaux. En plus de faciliter les échanges entre différents métiers (professionnel de santé, ingénieur informatique, data scientist, etc.), le pilotage médical du projet permet la mise en œuvre d'une plateforme centrée sur les besoins métiers (domaine de la santé), au service des professionnels de santé acteurs de l'utilisation secondaire des données de santé.

21. Il s'agissait des Ct (cycles de seuil) pour les PCR COVID-19.

À partir de 2021, après que le socle des données issues du soin aient été intégrées dans l'EDS i2b2, un virage a été pris avec le développement *de novo* d'une application au-dessus de la base de données de l'EDS. L'objectif de cette application est de permettre à des utilisateurs finaux, non experts en informatique médicale mais experts du domaine de la santé (médecin, infirmier, etc.), de conduire leur projet de **manière autonome**. C'est dans ce contexte que le travail détaillé dans le chapitre 3 a été réalisé, avec pour objectif de mettre à disposition des utilisateurs finaux une application dans laquelle les temps de requêtes de phénotypage (processus itératifs) étaient compatibles avec un usage en autonomie.

Dans un contexte où les EDS correspondent à des environnements complets favorisant l'utilisation secondaire des données, éviter que l'équipe en charge de la mise en œuvre de l'EDS ne soit le « goulot d'étranglement » des demandes est primordial pour éviter les « cimetières de données ».

Malgré les étapes d'intégration et de standardisation des données comme présenté dans le chapitre 2, les données de santé hospitalières intégrées dans les EDS restent complexes de par leur hétérogénéité, en particulier sémantique. De plus, les dictionnaires de données à explorer sont de très grande dimension (le CHU de Bordeaux intègre près d'un million de data elements dans l'EDS i2b2), complexifiant d'autant la fouille et l'accès aux données. Bien que l'accent soit mis sur les développements réalisés pour rendre les utilisateurs finaux autonomes, une formation de ces derniers à ces nouveaux outils et à l'organisation interne des données au sein des EDS est indispensable. Ainsi, les interactions entre les experts du domaine médical (les « cliniciens » ou les chercheurs) et les experts en informatique médicale semblent essentielles pour assurer le bon déroulement des études observationnelles basées sur les EDS.

À ce jour, l'application développée au-dessus de l'EDS i2b2 permet :

— La réalisation des démarches technico-réglementaires des projets ;
— L'identification des patients sur la base de critères clinico-biologiques ;
— L'accès aux données individuelles pseudonymisées (visionneuse / de manière programmatique) ;
— L'accès à un eCRF intégré, permettant une consolidation manuelle des données ;
— L'accès à des environnements d'analyse dédiés aux projets.

L'application n'intègre pas encore d'outils d'aide à la réalisation d'analyses statistiques simples (analyses descriptives, analyses comparatives simples), ce qui permettrait de couvrir l'ensemble des étapes d'un projet de recherche (hors diffusion des résultats).

L'ensemble des outils utilisés et développés sont basés sur des solutions **open-source** (PostgreSQL, Elasticsearch, SpringBoot, etc.). Avec la maîtrise complète des flux d'alimentation des EDS, l'utilisation de ces technologies open-source contribue à

éviter les « boîtes noires ».

À partir de 2021, un second EDS au format OMOP-CDM a été déployé dans le contexte d'un financement EHDEN. Cet EDS OMOP permet au CHU de Bordeaux de participer à des réseaux fédérés basés sur le modèle OMOP, en particulier au réseau fédéré DARWIN-EU. L'intégration des données dans OMOP pour la participation à des **réseaux fédérés** requiert une standardisation forte vers les concepts standards proposés via la communauté OMOP au travers d'OHDSI. Le travail présenté dans le chapitre 2 s'inscrit complètement dans cette démarche. Cette standardisation, qui réduit grandement l'hétérogénéité sémantique des données, implique une perte dans la finesse de l'information, incompatible avec les usages locaux nécessitant souvent un accès à des données précises et détaillées. Dans notre contexte, l'EDS OMOP peut être vu comme un datamart standardisé et spécialisé pour l'usage fédéré.

Plus que de simples bases de données, les EDS du CHU de Bordeaux sont des environnements complets qui intègrent l'ensemble des aspects de gouvernance, réglementaires, éthiques, de transparence vis-à-vis des usagers et techniques pour favoriser l'utilisation secondaire des données de santé. Ils font intervenir différents types de structures et d'acteurs hospitaliers (DRCI, DNUM, DPO, Responsable de la Sécurité des Systèmes d'Information (RSSI), équipes médicales, etc.) qui collectivement organisent et développent des outils au service de la recherche, des études, de l'évaluation et des activités de routine des établissements hospitaliers.

Les usagers doivent être impliqués fortement dans ces organisations nouvelles, en particulier pour permettre l'exercice de leurs droits (notamment le droit d'opposition à l'utilisation secondaire des données de santé). Également, la transparence vis-à-vis des usagers des traitements réalisés sur les données les concernant est un facteur majeur dans la mise au cœur du dispositif de recherche des patients.

Enfin, si de nombreuses solutions existent aujourd'hui pour déployer des EDS hospitaliers, il reste encore beaucoup à faire pour faciliter l'utilisation secondaire des données de santé par des utilisateurs finaux. À cet égard, les EDS font toujours l'objet de recherches continues de la part de la communauté scientifique d'informatique médicale, dans le but de développer des méthodes d'intégration et d'utilisation des données de santé.

# CHAPITRE 5

# Conclusion et perspectives

Cette thèse portait sur l'utilisation secondaire des données de santé issues du soin, et plus précisément dans le contexte des EDS hospitaliers.

Dans le chapitre 1, nous avons décrit le principe de l'utilisation secondaire des données de santé, en particulier pour des usages dans le cadre de la recherche, des études et de l'évaluation de la qualité et de la sécurité des soins. Nous avons en particulier identifié les principaux verrous liés à l'utilisation secondaire des données de santé, impliquant les données d'une part, et des aspects organisationnels d'autre part. Nous avons enfin introduit la notion d'EDS comme outil technique, organisationnel et réglementaire pour favoriser l'utilisation secondaire des données de santé dans le contexte hospitalier.

Dans le chapitre 2, nous avons décrit une méthode d'alignement de la biologie numérique basée sur les instances, qui garantit le respect de la protection des données à caractère personnel. Les résultats de cette méthode, mis en œuvre dans l'environnement EDS du CHU de Bordeaux, permettent de réduire l'hétérogénéité sémantique des données intégrées dans l'EDS. Cela favorise non seulement l'utilisation autonome des données de santé par les utilisateurs finaux en réduisant la taille du dictionnaire à fouiller, mais aussi l'intégration de l'EDS dans des réseaux fédérés internationaux.

Dans le chapitre 3, nous avons détaillé les adaptations nécessaires du modèle d'intégration des données cliniques d'i2b2 pour assurer la persistance des données dans une base de données NoSQL. Une implémentation concrète a été réalisée au sein de l'EDS du CHU de Bordeaux, permettant une évaluation de la persistance en NoSQL des données intégrées dans i2b2 en termes de ressources techniques nécessaires, de passage à l'échelle mais aussi de performances de requêtage. Nous avons montré une amélioration du temps de réponse pour les requêtes de phénotypage, en particulier pour les requêtes impliquant du texte libre. Cette implémenttaion est aujourd'hui en production au CHU de Bordeaux.

Enfin, dans le chapitre 4, nous avons présenté l'environnement EDS du CHU de Bordeaux, qui est composé de deux EDS distincts. En premier lieu, nous avons décrit l'EDS i2b2, dédié aux usages locaux réalisés en autonomie par les utilisateurs finaux. En second lieu, nous avons exposé l'EDS OMOP qui est dédié à une exploitation des données dans le cadre des réseaux fédérés. En particulier, nous avons décrit la participation du CHU de Bordeaux dans le réseau DARWIN-EU.

Les EDS sont des environnements techniques et organisationnels favorisant largement l'utilisation secondaire des données de santé au sein des établissements de santé hospitaliers.

Cependant, s'ils permettent de « désiloter » les données en les intégrant dans des modèles de données communs, l'**hétérogénéité sémantique** persiste largement au sein des « silos sémantiques » dans les EDS. De plus, la dimension des métadonnées à explorer et les informations médicales disponibles en texte libre dans les EHRs contreviennent encore à une **exploitation efficiente en autonomie par les utilisateurs finaux** des

données issues du soin contenues dans les EDS. Enfin, la production d'information fiable et valide à partir des RWD, pour permettre la production de RWE, est encore soumise à de nombreux biais en lien avec la **qualité des données** disponibles dans les EDS. Parfois présentés comme des objets pour la recherche, les EDS sont aussi des objets de recherche en informatique médicale.

Également, des organisations doivent encore être trouvées pour **ouvrir les données en dehors des établissements hospitaliers**, en particulier en les liant avec des données issues de la ville pour former de véritables **EDS de territoires**. L'articulation avec des initiatives comme P4DP (*Platform for Data in Primary care* ; [194]) pourrait permettre à terme la mise en place de tels EDS territoriaux. La mise en réseau des EDS hospitaliers, au niveau régional (ex : réseau fédéré EDS@NOVA avec les CHU de Bordeaux, Limoges et Poitiers), national ou supra-national revêt également un intérêt majeur. Enfin, la mise en place de l'**Espace Européen des Données de Santé** (EEDS ; [195]), tant en termes d'utilisation primaire que secondaire des données de santé, va remodeler le paysage des données de santé dans l'UE.

# Bibliographie


1. **Griffier, R**, Cossin, S, Konschelle, F, Mougin, F, Jouhet, V, Data Element Mapping in the Data Privacy Era. Studies in Health Technology and Informatics. 2022 May 25 ; 294:332-6. DOI : 10.3233/SHTI220469
2. **Griffier, R**, Mougin, F, Jouhet, V, Integrating healthcare data in an i2b2 model persisted through Elasticsearch. 2024 Aug 25. DOI : 10.2196/preprints.65753
3. Brat, GA, Weber, GM, Gehlenborg, N, Avillach, P, Palmer, NP, Chiovato, L, International electronic health record-derived COVID-19 clinical course profiles : the 4CE consortium. NPJ digital medicine. 2020 ; 3:109. DOI : 10.1038/s41746-020-00308-0
4. Weber, GM, Hong, C, Palmer, NP, Avillach, P, Murphy, SN, Gutiérrez-Sacristán, A, International Comparisons of Harmonized Laboratory Value Trajectories to Predict Severe COVID-19 : Leveraging the 4CE Collaborative Across 342 Hospitals and 6 Countries : A Retrospective Cohort Study. medRxiv : The Preprint Server for Health Sciences. 2021 Feb 5 :2020.12.16.20247684. DOI : 10.1101/2020.12.16.20247684
5. Moal, B, Orieux, A, Ferté, T, Neuraz, A, Brat, GA, Avillach, P, Acute respiratory distress syndrome after SARS-CoV-2 infection on young adult population : International observational federated study based on electronic health records through the 4CE consortium. PloS One. 2023 ; 18:e0266985. DOI : 10.1371/journal.pone.0266985
6. Du, M, Dernie, F, Català, M, Delmestri, A, Man, WY, Brash, JT, Treatment of systemic lupus erythematosus : Analysis of treatment patterns in adult and paediatric patients across four European countries. European Journal of Internal Medicine. 2024 Aug 11 :S0953–6205(24)00344-3. DOI : 10.1016/j.ejim.2024.08.008

7. Dernie, F, Corby, G, Robinson, A, Bezer, J, Mercade-Besora, N, **Griffier, R,** Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU). Pharmacoepidemiology and Drug Safety. 2024 Nov 12 ; 33:e70042. DOI : 10.1002/pds.70042

8. Thomason, J. Big tech, big data and the new world of digital health. Global Health Journal. Special issue on Intelligent Medicine Leads the New Development of Human Health 2021 Dec 1 ; 5:165-8. DOI : 10.1016/j.glohj.2021.11.003

9. Becker, R, Chokoshvili, D, Comandé, G, Dove, ES, Hall, A, Mitchell, C, Secondary Use of Personal Health Data : When Is It "Further Processing" Under the GDPR, and What Are the Implications for Data Controllers ? European Journal of Health Law. 2022 Aug 1 ; 30:129-57. DOI : 10.1163/15718093-bja10094

10. Dicuonzo, G, Galeone, G, Shini, M, Massari, A, Towards the Use of Big Data in Healthcare : A Literature Review. Healthcare (Basel, Switzerland). 2022 Jul 1 ; 10:1232. DOI : 10.3390/healthcare10071232

11. Meng, W, Ou, W, Chandwani, S, Chen, X, Black, W, Cai, Z, Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer. Journal of Biomedical Informatics. 2019 Dec ; 100:103335. DOI : 10.1016/j.jbi.2019.103335

12. Essay, P, Fisher, JM, Mosier, JM, Subbian, V, Validation of an Electronic Phenotyping Algorithm for Patients With Acute Respiratory Failure. Critical Care Explorations. 2022 Mar ; 4:e0645. DOI : 10.1097/CCE.0000000000000645

13. Geva, A, Gronsbell, JL, Cai, T, Cai, T, Murphy, SN, Lyons, JC, A Computable Phenotype Improves Cohort Ascertainment in a Pediatric Pulmonary Hypertension Registry. The Journal of pediatrics. 2017 Sep ; 188:224-231.e5. DOI : 10.1016/j.jpeds.2017.05.037

14. Toh, S, Avorn, J, D'Agostino, RB, Gurwitz, JH, Psaty, BM, Rothman, KJ, Reusing Mini-Sentinel data following rapid assessments of potential safety signals via modular analytic programs. Pharmacoepidemiology and Drug Safety. 2013 Oct ; 22:1036-45. DOI : 10.1002/pds.3478

15. Kruse, CS, Stein, A, Thomas, H, Kaur, H, The use of Electronic Health Records to Support Population Health : A Systematic Review of the Literature. Journal of Medical Systems. 2018 Sep 29 ; 42:214. DOI : 10.1007/s10916-018-1075-6

16. Ferté, T, Jouhet, V, **Griffier, R**, Hejblum, BP, Thiébaut, R, Bordeaux University Hospital Covid-19 Crisis Task Force, The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France. JAMIA open. 2022 Dec ; 5:ooac086. DOI : 10.1093/jamiaopen/ooac086

17. FDA, Real-World Evidence. 2023 Apr 25. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [Accessed on: 2024 Oct 24]

18. Swift, B, Jain, L, White, C, Chandrasekaran, V, Bhandari, A, Hughes, DA, Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care. Clinical and Translational Science. 2018 Sep ; 11:450-60. DOI : 10.1111/cts.12559

19. Cuggia, M, Polton, D, Wainrib, G, Combes, S, Health Data Hub - Mission de prefiguration. 2018 Oct 12. Available from: https://www.vie-publique.fr/rapport/37719-health-data-hub-mission-de-prefiguration [Accessed on: 2024 Sep 28]

20. Loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé. 2016 Jan 26. Available from: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000031912641 [Accessed on: 2024 Sep 28]

21. Loi n° 98-1194 du 23 décembre 1998 de financement de la sécurité sociale pour 1999. 1998 Dec 23. Available from: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000392993 [Accessed on: 2024 Sep 28]

22. Loi n° 2007-127 du 30 janvier 2007 ratifiant l'ordonnance n° 2005-1040 du 26 août 2005 relative à l'organisation de certaines professions de santé et à la répression de l'usurpation de titres et de l'exercice illégal de ces professions et modifiant le code de la santé publique. 2007 Jan 30. Available from: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000822417 [Accessed on: 2024 Sep 28]

23. Loi n° 2004-810 du 13 août 2004 relative à l'assurance maladie. 2004 Aug 13. Available from: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000625158 [Accessed on: 2024 Sep 28]

24. Elbaum, M, Morelle, DA, Minault, B, Ribieras, H, Les cohortes pour les études et la recherche en santé. 2020 Feb. Available from: https://www.enseignementsup-recherche.gouv.fr/fr/les-cohortes-pour-les-etudes-et-la-recherche-en-sante-47633 [Accessed on: 2024 Oct 24]

25. Goldberg, M, Carton, M, Descatha, A, Leclerc, A, Roquelaure, Y, Santin, G, CONSTANCES : a general prospective population-based cohort for occupational and environmental epidemiology : cohort profile. Occupational and Environmental Medicine. 2017 Jan ; 74:66-71. DOI : 10.1136/oemed-2016-103678

26. Arrêté du 6 novembre 1995 relatif au Comité national des registres - Article 2. 1995 Nov 6. Available from: https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000006731725/2024-10-08 [Accessed on: 2024 Oct 8]

27. Francis, F, Terroba, C, Persoz, C, Gagliolo, J, Alla, F, Quelle place pour les registres de morbidité à l'ère des données massives de santé ? Revue d'Épidémiologie et de Santé Publique. 2020 Apr 1 ; 68:117-23. DOI : 10.1016/j.respe.2019.11.018

28. Gardy, J, Wilson, S, Guizard, AV, Bouvier, V, Launay, L, Alves, A, Access to primary care and mortality in excess for patients with cancer in France : Results from 21 French Cancer Registries. Cancer. 2024 Aug 20. DOI : 10.1002/cncr.35519

29. Robert, E, Guibaud, P, Maternal valproic acid and congenital neural tube defects. Lancet (London, England). 1982 Oct 23 ; 2:937. DOI : 10.1016/s0140-6736(82)90908-4

30. Kelly, JT, Campbell, KL, Gong, E, Scuffham, P, The Internet of Things : Impact and Implications for Health Care Delivery. Journal of Medical Internet Research. 2020 Nov 10 ; 22:e20135. DOI : 10.2196/20135

31. Kynoch, K, Ameen, M, Ramis, MA, Khalil, H, Use of Patient-Reported Data within the Acute Healthcare Context : A Scoping Review. International Journal of Environmental Research and Public Health. 2022 Sep 6 ; 19:11160. DOI : 10.3390/ijerph191811160

32. Patrick, DL, Burke, LB, Powers, JH, Scott, JA, Rock, EP, Dawisha, S, Patient-reported outcomes to support medical product labeling claims : FDA perspective. Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research. 2007 ; 10 Suppl 2:S125-137. DOI : 10.1111/j.1524-4733.2007.00275.x

33. Gunter, TD, Terry, NP, The Emergence of National Electronic Health Record Architectures in the United States and Australia : Models, Costs, and Questions. Journal of Medical Internet Research. 2005 Mar 14 ; 7:e383. DOI : 10.2196/jmir.7.1.e3

34. Safran, C, Bloomrosen, M, Hammond, WE, Labkoff, S, Markel-Fox, S, Tang, PC, Toward a National Framework for the Secondary Use of Health Data : An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association : JAMIA. 2007 Feb ; 14:1. DOI : 10.1197/jamia.M2273

35. MacKenzie, SL, Wyatt, MC, Schuff, R, Tenenbaum, JD, Anderson, N, Practices and perspectives on building integrated data repositories : results from a 2010 CTSA survey. Journal of the American Medical Informatics Association : JAMIA. 2012 Jun ; 19(e1):e119-124. DOI : 10.1136/amiajnl-2011-000508

36. Kim, HS, Kim, JH, Proceed with Caution When Using Real World Data and Real World Evidence. Journal of Korean Medical Science. 2019 Jan 16 ; 34. DOI : 10.3346/jkms.2019.34.e28

37. Bian, J, Lyu, T, Loiacono, A, Viramontes, TM, Lipori, G, Guo, Y, Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. Journal of the American Medical Informatics Association. 2020 Dec 9 ; 27:1999-2010. DOI : 10.1093/jamia/ocaa245

38. Liu, F, Panagiotakos, D, Real-world data : a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology. 2022 Nov 5 ; 22:287. DOI : 10.1186/s12874-022-01768-6

39. Weiner, MG, Embi, PJ, Toward reuse of clinical data for research and quality improvement : the end of the beginning ? Annals of Internal Medicine. 2009 Sep 1 ; 151:359-60. DOI : 10.7326/0003-4819-151-5-200909010-00141

40. Hersh, WR, Weiner, MG, Embi, PJ, Logan, JR, Payne, PR, Bernstam, EV, Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Medical care. 2013 Aug ; 51:S30-S37. DOI : 10.1097/MLR.0b013e31829b1dbd

41. Bastarache, L, Brown, JS, Cimino, JJ, Dorr, DA, Embi, PJ, Payne, PRO, Developing real-world evidence from real-world data : Transforming raw data into analytical datasets. Learning Health Systems. 2022 Jan ; 6:e10293. DOI : 10.1002/lrh2.10293

42. Curtis, MD, Griffith, SD, Tucker, M, Taylor, MD, Capra, WB, Carrigan, G, Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. Health Services Research. 2018 Dec ; 53:4460-76. DOI : 10.1111/1475-6773.12872

43. Miksad, RA, Abernethy, AP, Harnessing the Power of Real-World Evidence (RWE) : A Checklist to Ensure Regulatory-Grade Data Quality. Clinical Pharmacology & Therapeutics. 2018 ; 103:202-5. DOI : 10.1002/cpt.946

44. Bhatt, A. Data quality – The foundation of real-world studies. Perspectives in Clinical Research. 2023 Jun ; 14:92. DOI : 10.4103/picr.picr_12_23

45. Zins, C. Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology. 2007 ; 58:479-93. DOI : 10.1002/asi.20508

46. Blumentritt, R, Johnston, R, Towards a Strategy for Knowledge Management. Technology Analysis & Strategic Management. 1999 Sep 1 ; 11:287-300. DOI : 10.1080/095373299107366

47. Dang, A. Real-World Evidence : A Primer. Pharmaceutical Medicine. 2023 Jan 1 ; 37:25-36. DOI : 10.1007/s40290-022-00456-6

48. Kim, HS, Lee, S, Kim, JH, Real-world Evidence versus Randomized Controlled Trial : Clinical Research Based on Electronic Medical Records. Journal of Korean Medical Science. 2018 Jun 26 ; 33. DOI : 10.3346/jkms.2018.33.e213

49. Tang, AS, Woldemariam, SR, Miramontes, S, Norgeot, B, Oskotsky, TT, Sirota, M, Harnessing EHR data for health research. Nature Medicine. 2024 Jul ; 30:1847-55. DOI : 10.1038/s41591-024-03074-8

50. Wang, Z, Craven, C, Syed, M, Greer, M, Seker, E, Syed, S, Clinical Data Warehousing : A Scoping Review. Journal of the Society for Clinical Data Management. 2024 Aug 28 ; 5. DOI : 10.47912/jscdm.320

51. Khalaf Hamoud, A, Salah Hashim, A, Akeel Awadh, W, Clinical data warehouse : a review. Iraqi Journal for Computers and Informatics. 2018 Dec 31 ; 44. DOI : 10.25195/2017/4424

52. Shin, SY, Kim, WS, Lee, JH, Characteristics Desired in Clinical Data Warehouse for Biomedical Research. Healthcare Informatics Research. 2014 Apr 30 ; 20. Publisher : Korean Society of Medical Informatics:109-16. DOI : 10.4258/hir.2014.20.2.109

53. Safran, C. Reuse of clinical data. Yearbook of Medical Informatics. 2014 Aug 15 ; 9:52-4. DOI : 10.15265/IY-2014-0013

54. Shah, SM, Khan, RA, Secondary Use of Electronic Health Record : Opportunities and Challenges. IEEE Access. 2020 ; 8:136947-65. DOI : 10.1109/ACCESS.2020.3011099

55. Liman, L, May, B, Fette, G, Krebs, J, Puppe, F, Using a Clinical Data Warehouse to Calculate and Present Key Metrics for the Radiology Department : Implementation and Performance Evaluation. JMIR medical informatics. 2023 May 22 ; 11:e41808. DOI : 10.2196/41808

56. Windsor, C, Hua, C, De Roux, Q, Harrois, A, Anguel, N, Montravers, P, Healthcare trajectory of critically ill patients with necrotizing soft tissue infections : a multicenter retrospective cohort study using the clinical data warehouse of Greater Paris University Hospitals. Annals of Intensive Care. 2022 Dec 20 ; 12:115. DOI : 10.1186/s13613-022-01087-5

57. Keum, N, Yoo, J, Hur, S, Shin, SY, Dykes, PC, Kang, MJ, The potential for drug incompatibility and its drivers - A hospital wide retrospective descriptive study. International Journal of Medical Informatics. 2024 Nov ; 191:105584. DOI : 10.1016/j.ijmedinf.2024.105584

58. Meijden, SL, Boekel, AM, Goor, H, Nelissen, RG, Schoones, JW, Steyerberg, EW, Automated Identification of Postoperative Infections to Allow Prediction and Surveillance Based on Electronic Health Record Data : Scoping Review. JMIR medical informatics. 2024 Sep 10 ; 12:e57195. DOI : 10.2196/57195

59. Hoffmann, JA, Corboy, JB, Liu, L, Cieslak, K, Pergjika, A, Patel, TR, Use of Electronic Health Record-Based Measures to Assess Quality of Care for Pediatric Agitation. Hospital Pediatrics. 2024 Apr 15 ; 14:319-27. DOI : 10.1542/hpeds.2023-007532

60. Doutreligne, M, Degremont, A, Jachiet, PA, Lamer, A, Tannier, X, Good practices for clinical data warehouse implementation : A case study in France. PLOS Digital Health. 2023 Jul 6 ; 2:e0000298. DOI : 10.1371/journal.pdig.0000298

61. Giordano, C, Brennan, M, Mohamed, B, Rashidi, P, Modave, F, Tighe, P, Accessing Artificial Intelligence for Clinical Decision-Making. Frontiers in Digital Health. 2021 Jun 25 ; 3. DOI : 10.3389/fdgth.2021.645232

62. Johnson, KB, Wei, WQ, Weeraratne, D, Frisse, ME, Misulis, K, Rhee, K, Precision Medicine, AI, and the Future of Personalized Health Care. Clinical and Translational Science. 2021 ; 14:86-93. DOI : 10.1111/cts.12884

63. Ferté, T, Cossin, S, Schaeverbeke, T, Barnetche, T, Jouhet, V, Hejblum, BP, Automatic phenotyping of electronical health record : PheVis algorithm. Journal of Biomedical Informatics. 2021 May ; 117:103746. DOI : 10.1016/j.jbi.2021.103746

64. Rohr, O, Priou, S, Chatellier, G, Babai, S, Gallien, S, Flicoteaux, R, Prevalence and risks of intravenous chemotherapy-induced severe neutropenia in solid cancers : a multicenter retrospective cohort study on real-life data. Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer. 2024 Sep 13 ; 32. DOI : 10.1007/s00520-024-08817-4

65. Madhavan, S, Bastarache, L, Brown, JS, Butte, AJ, Dorr, DA, Embi, PJ, Use of electronic health records to support a public health response to the COVID-19 pandemic in the United States : a perspective from 15 academic medical centers. Journal of the American Medical Informatics Association : JAMIA. 2021 Feb 15 ; 28:393-401. DOI : 10.1093/jamia/ocaa287

66. Dagliati, A, Malovini, A, Tibollo, V, Bellazzi, R, Health informatics and EHR to support clinical research in the COVID-19 pandemic : an overview. Briefings in Bioinformatics. 2021 Mar 1 ; 22:812-22. DOI : 10.1093/bib/bbaa418

67. Obeid, JS, Beskow, LM, Rape, M, Gouripeddi, R, Black, RA, Cimino, JJ, A survey of practices for the use of electronic health records to support research recruitment. Journal of Clinical and Translational Science. 2017 Nov 22 ; 1:246. DOI : 10.1017/cts.2017.301

68. Nashwan, AJ, Hani, SB, Transforming cancer clinical trials : The integral role of artificial intelligence in electronic health records for efficient patient recruitment. Contemporary Clinical Trials Communications. 2023 Dec ; 36:101223. DOI : 10.1016/j.conctc.2023.101223

69. Yu, S, Ma, Y, Gronsbell, J, Cai, T, Ananthakrishnan, AN, Gainer, VS, Enabling phenotypic big data with PheNorm. Journal of the American Medical Informatics Association. 2018 Jan 1 ; 25:54-60. DOI : 10.1093/jamia/ocx111

70. Prince, K, Jones, M, Blackwell, A, Simpson, A, Meakins, S, Vuylsteke, A, Barriers to the secondary use of data in critical care. Journal of the Intensive Care Society. 2018 May 1 ; 19:127-31. DOI : 10.1177/1751143717741082

71. Vukovic, J, Ivankovic, D, Habl, C, Dimnjakovic, J, Enablers and barriers to the secondary use of health data in Europe : general data protection regulation perspective. Archives of Public Health. 2022 Apr 9 ; 80:115. DOI : 10.1186/s13690-022-00866-7

72. Kohane, IS, Aronow, BJ, Avillach, P, Beaulieu-Jones, BK, Bellazzi, R, Bradford, RL, What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. Journal of Medical Internet Research. 2021 Mar 2 ; 23:e22219. DOI : 10.2196/22219

73. Lamer, A, Saint-Dizier, C, Paris, N, Chazard, E, Data Lake, Data Warehouse, Datamart, and Feature Store : Their Contributions to the Complete Data Reuse Pipeline. JMIR medical informatics. 2024 Jul 17 ; 12:e54590. DOI : 10.2196/54590

74. Cios, KJ, William Moore, G, Uniqueness of medical data mining. Artificial Intelligence in Medicine. Medical Data Mining and Knowledge Discovery 2002 Sep 1 ; 26:1-24. DOI : 10.1016/S0933-3657(02)00049-0

75. Lee, KKY, Tang, WC, Choi, KS, Alternatives to relational database : Comparison of NoSQL and XML approaches for clinical data storage. Computer Methods and Programs in Biomedicine. 2013 Apr 1 ; 110:99-109. DOI : 10.1016/j.cmpb.2012.10.018

76. Gamal, A, Barakat, S, Rezk, A, Standardized electronic health record data modeling and persistence : A comparative review. Journal of Biomedical Informatics. 2021 Feb 1 ; 114:103670. DOI : 10.1016/j.jbi.2020.103670

77. Quintas, R. The Data Warehouse Toolkit, 3rd Edition. Wiley. 2013 Jul. Available from: https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-toolkit/ [Accessed on: 2024 May 7]

78. Horgan, D, Hajduch, M, Vrana, M, Soderberg, J, Hughes, N, Omar, MI, European Health Data Space—An Opportunity Now to Grasp the Future of Data-Driven Healthcare. Healthcare. 2022 Sep ; 10:1629. DOI : 10.3390/healthcare10091629

79. Poggiani, A, Cortesi, A, Challenges in Health Information Systems : Health Data Management and Access for Life Sciences Research. Sous la dir. de Saeed K et Dvorský J. Cham : Springer Nature Switzerland, 2024 :195-211. DOI : 10.1007/978-3-031-71115-2_14

80. Umberfield, EE, Staes, CJ, Morgan, TP, Grout, RW, Mamlin, BW, Dixon, BE, Chapter 9 - Syntactic interoperability and the role of syntactic standards in health information exchange. Health Information Exchange (Second Edition). Sous la dir. de Dixon BE. Academic Press, 2023 Jan 1:217-36. DOI : 10.1016/B978-0-323-90802-3.00004-6

81. Brämer, GR. International statistical classification of diseases and related health problems. Tenth revision. World Health Statistics Quarterly Rapport Trimestriel De Statistiques Sanitaires Mondiales. 1988 ; 41:32-6

82. WHO, Anatomical Therapeutic Chemical (ATC) Classification. 1976. Available from: https://www.who.int/tools/atc-ddd-toolkit/atc-classification [Accessed on: 2024 Aug 24]

83. Schulz, S, Rodrigues, JM, Rector, A, Chute, CG, Interface Terminologies, Reference Terminologies and Aggregation Terminologies : A Strategy for Better Integration. Studies in Health Technology and Informatics. 2017 ; 245:940-4. DOI : 10.3233/978-1-61499-830-3-940

84. Bidgood, WD, Horii, SC, Introduction to the ACR-NEMA DICOM standard. Radiographics : A Review Publication of the Radiological Society of North America, Inc. 1992 Mar ; 12:345-55. DOI : 10.1148/radiographics.12.2.1561424

85. Assurance Maladie, NABM - Nomenclature des Actes de Biologie Médicale. 2024 Sep 11. Available from: https://smt.esante.gouv.fr/terminologie-nabm/ [Accessed on: 2024 Sep 28]

86. Huff, SM, Rocha, RA, McDonald, CJ, De Moor, GJ, Fiers, T, Bidgood, WD, Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. Journal of the American Medical Informatics Association : JAMIA. 1998 ; 5:276-92. DOI : 10.1136/jamia.1998.0050276

87. McDonald, CJ, Huff, SM, Suico, JG, Hill, G, Leavelle, D, Aller, R, LOINC, a universal standard for identifying laboratory observations : a 5-year update. Clinical Chemistry. 2003 Apr ; 49:624-33. DOI : 10.1373/49.4.624

88. Lee, MK, Park, HA, Min, YH, Kim, Y, Min, HK, Ham, SW, Evaluation of the Clinical Data Dictionary (CiDD). Healthcare Informatics Research. 2010 Jun ; 16:82-8. DOI : 10.4258/hir.2010.16.2.82

89. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés - Article 65. 1978 Jan 6. Available from: https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000038888757 [Accessed on: 2024 Sep 10]

90. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE. Doc ID : 02016R0679-20160504 Doc Sector : 0 Doc Title : Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) Doc Type : R Usr_lan : en. 2016 Apr 27. Available from: https://eur-lex.europa.eu/eli/reg/2016/679 [Accessed on: 2024 Sep 30]

91. Article 4 EU règlement général sur la protection des données (EU-RGPD). 2016 Apr 25. Available from: https://www.privacy-regulation.eu/fr/4.htm [Accessed on: 2024 Sep 28]

92. Article 35 EU règlement général sur la protection des données (EU-RGPD). 2016 Apr 25. Available from: https://www.privacy-regulation.eu/fr/35.htm [Accessed on: 2024 Sep 28]

93. Aamot, H, Kohl, CD, Richter, D, Knaup-Gregori, P, Pseudonymization of patient identifiers for translational research. BMC Medical Informatics and Decision Making. 2013 Jul 24; 13:75. DOI : 10.1186/1472-6947-13-75

94. Ciampi, M, Sicuranza, M, Silvestri, S, A Privacy-Preserving and Standard-Based Architecture for Secondary Use of Clinical Data. Information. 2022 Feb; 13:87. DOI : 10.3390/info13020087

95. Tannier, X, Wajsbürt, P, Calliger, A, Dura, B, Mouchet, A, Hilka, M, Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse. Methods of Information in Medicine. 2024 Mar 5. Publisher : Georg Thieme Verlag KG. DOI : 10.1055/s-0044-1778693

96. Délibération n° 2018-155 du 3 mai 2018 portant homologation de la méthodologie de référence relative aux traitements de données à caractère personnel mis en œuvre dans le cadre des recherches n'impliquant pas la personne humaine, des études et évaluations dans le domaine de la santé (MR-004). 2018 May 3. Available from: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037187498 [Accessed on: 2024 Oct 10]

97. Marchand-Arvier, J, Allassonniere, PS, Hoang, A, Jannot, DAS, Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé. 2023 Dec 5. Available from: https://igas.gouv.fr/federer-les-acteurs [Accessed on: 2024 Oct 24]

98. Conseil d'Etat, Etude annuelle 2014 du Conseil d'Etat - Le numerique et les droits fondamentaux. Le numérique et les droits fondamentaux. Conseil d'État, 2014 Sep 9:446. Available from: https://www.vie-publique.fr/files/rapport/pdf/144000541.pdf [Accessed on: 2024 Oct 1]

99. Ganslandt, T, Mate, S, Helbing, K, Sax, U, Prokosch, H, Unlocking Data for Clinical Research – The German i2b2 Experience. Applied Clinical Informatics. 2011 Mar 30; 2:116-27. DOI : 10.4338/ACI-2010-09-CR-0051

100. Cuggia, M, Garcelon, N, Campillo-Gimenez, B, Bernicot, T, Laurent, JF, Garin, E, Roogle : an information retrieval engine for clinical data warehouse. Studies in Health Technology and Informatics. 2011 Jan; 169:584-8. DOI : 10.3233/978-1-60750-806-9-584

101. Karakachoff, M, Goronflot, T, Coudol, S, Toublant, D, Bazoge, A, Constant Dit Beaufils, P, Implementing a Biomedical Data Warehouse From Blueprint to Bedside in a Regional French University Hospital Setting : Unveiling Processes, Overcoming Challenges, and Extracting Clinical Insight. JMIR medical informatics. 2024 Jun 24; 12:e50194. DOI : 10.2196/50194

102. Garcelon, N, Neuraz, A, Salomon, R, Faour, H, Benoit, V, Delapalme, A, A clinician friendly data warehouse oriented toward narrative reports : Dr. Warehouse. Journal of Biomedical Informatics. 2018 Apr 1; 80:52-63. DOI : 10.1016/j.jbi.2018.02.019

103. Zapletal, E, Rodon, N, Grabar, N, Degoulet, P, Methodology of integration of a clinical data warehouse with a clinical information system : the HEGP case. Studies in health technology and informatics. 2010; 160(Pt 1):193-7

104. Maier, C, Lang, L, Storf, H, Vormstein, P, Bieber, R, Bernarding, J, Towards Implementation of OMOP in a German University Hospital Consortium. Applied Clinical Informatics. 2018 Jan; 09:54-61. DOI : 10.1055/s-0037-1617452

105. Gagalova, KK, Elizalde, MAL, Portales-Casamar, E, Görges, M, What You Need to Know Before Implementing a Clinical Research Data Warehouse : Comparative Review of Integrated Data Repositories in Health Care Institutions. JMIR Formative Research. 2020 Aug 27; 4:e17687. DOI : 10.2196/17687

106. Murphy, SN, Weber, G, Mendis, M, Gainer, V, Chueh, HC, Churchill, S, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association : JAMIA. 2010; 17:124-30. DOI : 10.1136/jamia.2009.000893

107. Stang, PE, Ryan, PB, Racoosin, JA, Overhage, JM, Hartzema, AG, Reich, C, Advancing the science for active surveillance : rationale and design for the Observational Medical Outcomes Partnership. Annals of Internal Medicine. 2010 Nov 2; 153:600-6. DOI : 10.7326/0003-4819-153-9-201011020-00010

108. Voss, EA, Blacketer, C, Sandijk, S, Moinat, M, Kallfelz, M, Speybroeck, M, European Health Data & Evidence Network-learnings from building out a standardized international health data network. Journal of the American Medical Informatics Association : JAMIA. 2023 Dec 22; 31:209-19. DOI : 10.1093/jamia/ocad214

109. Reich, C, Ostropolets, A, Ryan, P, Rijnbeek, P, Schuemie, M, Davydov, A, OHDSI Standardized Vocabularies-a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association : JAMIA. 2024 Feb 16; 31:583-90. DOI : 10.1093/jamia/ocad247

110. Klann, JG, Joss, MAH, Embree, K, Murphy, SN, Data model harmonization for the All Of Us Research Program : Transforming i2b2 data into the OMOP common data model. PLOS ONE. 2019 Feb 19; 14:e0212463. DOI : 10.1371/journal.pone.0212463

111. Phillips, L. NCBO Extraction Tool version 2.0. Available from: https://community.i2b2.org/wiki/display/NCBO/NCBO+Extraction+Tool+version+2.0 [Accessed on: 2024 Oct 10]

112. Weber, GM, Murphy, SN, McMurry, AJ, MacFadden, D, Nigrin, DJ, Churchill, S, The Shared Health Research Information Network (SHRINE) : A Prototype Federated Query Tool for Clinical Data Repositories. Journal of the American Medical Informatics Association. 2009 Sep 1 ; 16:624-30. DOI : 10.1197/jamia.M3191

113. CNIL, La CNIL adopte un référentiel sur les entrepôts de données de santé. 2021 Oct 26. Available from: https://www.cnil.fr/fr/la-cnil-adopte-un-referentiel-sur-les-entrepots-de-donnees-de-sante [Accessed on: 2024 Oct 10]

114. Hallock, H, Marshall, SE, 't Hoen, PAC, Nygård, JF, Hoorne, B, Fox, C, Federated Networks for Distributed Analysis of Health Data. Frontiers in Public Health. 2021 Sep 30 ; 9. DOI : 10.3389/fpubh.2021.712569

115. Weber, GM. Federated queries of clinical data repositories : Scaling to a national network. Journal of Biomedical Informatics. 2015 Jun 1 ; 55:231-6. DOI : 10.1016/j.jbi.2015.04.012

116. Palchuk, MB, London, JW, Perez-Rey, D, Drebert, ZJ, Winer-Jones, JP, Thompson, CN, A global federated real-world data and analytics platform for research. JAMIA Open. 2023 Jul 1 ; 6:ooad035. DOI : 10.1093/jamiaopen/ooad035

117. Mateus, P, Moonen, J, Beran, M, Jaarsma, E, Landen, SM, Heuvelink, J, Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model : A Netherlands consortium of dementia cohorts case study. Journal of Biomedical Informatics. 2024 Jul 1 ; 155:104661. DOI : 10.1016/j.jbi.2024.104661

118. Le, TT, Gutiérrez-Sacristán, A, Son, J, Hong, C, South, AM, Beaulieu-Jones, BK, Multinational characterization of neurological phenotypes in patients hospitalized with COVID-19. Scientific Reports. 2021 Oct 12 ; 11:20238. DOI : 10.1038/s41598-021-99481-9

119. Weber, GM, Zhang, HG, L'Yi, S, Bonzel, CL, Hong, C, Avillach, P, International Changes in COVID-19 Clinical Trajectories Across 315 Hospitals and 6 Countries : Retrospective Cohort Study. Journal of Medical Internet Research. 2021 Oct 11 ; 23:e31400. DOI : 10.2196/31400

120. Klann, JG, Strasser, ZH, Hutch, MR, Kennedy, CJ, Marwaha, JS, Morris, M, Distinguishing Admissions Specifically for COVID-19 From Incidental SARS-CoV-2 Admissions : National Retrospective Electronic Health Record Study. Journal of Medical Internet Research. 2022 May 18 ; 24:e37931. DOI : 10.2196/37931

121. Weber, GM, Hong, C, Xia, Z, Palmer, NP, Avillach, P, L'Yi, S, International comparisons of laboratory values from the 4CE collaborative to predict COVID-19 mortality. NPJ digital medicine. 2022 Jun 13 ; 5:74. DOI : 10.1038/s41746-022-00601-0

122. Tan, BWL, Tan, BWQ, Tan, ALM, Schriver, ER, Gutiérrez-Sacristán, A, Das, P, Long-term kidney function recovery and mortality after COVID-19-associated acute kidney injury : An international multi-centre observational cohort study. EClinicalMedicine. 2023 Jan ; 55:101724. DOI : 10.1016/j.eclinm.2022.101724

123. Bourgeois, FT, Gutiérrez-Sacristán, A, Keller, MS, Liu, M, Hong, C, Bonzel, CL, International Analysis of Electronic Health Records of Children and Youth Hospitalized With COVID-19 Infection in 6 Countries. JAMA network open. 2021 Jun 1 ; 4:e2112596. DOI : 10.1001/jamanetworkopen.2021.12596

124. Gutiérrez-Sacristán, A, Serret-Larmande, A, Hutch, MR, Sáez, C, Aronow, BJ, Bhatnagar, S, Hospitalizations Associated With Mental Health Conditions Among Adolescents in the US and France During the COVID-19 Pandemic. JAMA network open. 2022 Dec 1 ; 5:e2246548. DOI : 10.1001/jamanetworkopen.2022.46548

125. Sperotto, F, Gutiérrez-Sacristán, A, Makwana, S, Li, X, Rofeberg, VN, Cai, T, Clinical phenotypes and outcomes in children with multisystem inflammatory syndrome across SARS-CoV-2 variant eras : a multinational study from the 4CE consortium. EClinicalMedicine. 2023 Oct ; 64:102212. DOI : 10.1016/j.eclinm.2023.102212

126. Estiri, H, Strasser, ZH, Brat, GA, Semenov, YR, Consortium for Characterization of COVID-19 by EHR (4CE), Patel, CJ, Evolving phenotypes of non-hospitalized patients that indicate long COVID. BMC medicine. 2021 Sep 27 ; 19:249. DOI : 10.1186/s12916-021-02115-0

127. Dagliati, A, Strasser, ZH, Hossein Abad, ZS, Klann, JG, Wagholikar, KB, Mesa, R, Characterization of long COVID temporal sub-phenotypes by distributed representation learning from electronic health record data : a cohort study. EClinicalMedicine. 2023 Oct ; 64:102210. DOI : 10.1016/j.eclinm.2023.102210

128. Zhang, HG, Honerlaw, JP, Maripuri, M, Samayamuthu, MJ, Beaulieu-Jones, BR, Baig, HS, Potential pitfalls in the use of real-world data for studying long COVID. Nature Medicine. 2023 May ; 29:1040-3. DOI : 10.1038/s41591-023-02274-y

129. Wang, X, Zhang, HG, Xiong, X, Hong, C, Weber, GM, Brat, GA, SurvMaximin : Robust federated approach to transporting survival risk prediction models. Journal of Biomedical Informatics. 2022 Oct ; 134:104176. DOI : 10.1016/j.jbi.2022.104176

130. Arlett, P, Kjær, J, Broich, K, Cooke, E, Real-World Evidence in EU Medicines Regulation : Enabling Use and Establishing Value. Clinical Pharmacology and Therapeutics. 2022 Jan ; 111:21-3. DOI : 10.1002/cpt.2479

131. European Medicines Agency, Data Analysis and Real World Interrogation Network (DARWIN EU). 2021 Jun 4. Available from: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/real-world-evidence/data-analysis-real-world-interrogation-network-darwin-eu [Accessed on: 2024 Sep 29]

132. Lea, NC, Nicholls, J, Fitzpatrick, NK, Between Scylla and Charybdis : Charting the Wicked Problem of Reusing Health Data for Clinical Research Informatics. Yearbook of Medical Informatics. 2018 Aug ; 27:170-6. DOI : 10.1055/s-0038-1641219

133. Blasini, R, Strantz, C, Gulden, C, Helfer, S, Lidke, J, Prokosch, HU, Evaluation of Eligibility Criteria Relevance for the Purpose of IT-Supported Trial Recruitment : Descriptive Quantitative Analysis. JMIR formative research. 2024 Jan 31 ; 8:e49347. DOI : 10.2196/49347

134. Safran, C. Update on Data Reuse in Health Care. Yearbook of Medical Informatics. 2017 Aug ; 26:24-7. DOI : 10.15265/IY-2017-013

135. Tripathy, JP. Secondary Data Analysis : Ethical Issues and Challenges. Iranian Journal of Public Health. 2013 Dec ; 42:1478-9

136. Sarwar, F, Tassawar, F, Naeem, F, Shafaq, F, Khan, HKH, Yaseen, H, Ethical Dilemmas in Using Electronic Medical Records. Journal of Society of Prevention, Advocacy and Research KEMU. 2022 Sep 20 ; 1. Number : 2. Available from: https://journalofspark.com/journal/index.php/JSpark/article/view/112 [Accessed on: 2024 Sep 3]

137. Fleischer, NJ, Khalil, A, Limitations and recommendations for use of secondary data analysis in pediatric research. Children's Health Care. 2023 Nov 9 ; 0:1-17. DOI : 10.1080/02739615.2023.2279064

138. Chute, CG, Pathak, J, Savova, GK, Bailey, KR, Schor, MI, Hart, LA, The SHARPn project on secondary use of Electronic Medical Record data : progress, plans, and possibilities. AMIA Annual Symposium proceedings AMIA Symposium. 2011 ; 2011:248-56

139. Rosenbloom, ST, Miller, RA, Johnson, KB, Elkin, PL, Brown, SH, Interface Terminologies : Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. Journal of the American Medical Informatics Association : JAMIA. 2006 ; 13:277-88. DOI : 10.1197/jamia.M1957

140. Byun, A, Sung, S, Yu, J, Chang, E, Park, HA, Harmonization of Data Across Cohorts Using Standard Terminologies. Studies in Health Technology and Informatics. 2024 Aug 22 ; 316:1943-4. DOI : 10.3233/SHTI240813

141. Ranegger, R, Baumberger, D, Grgic, P, Jagfeld, G, Semantic Interoperability of Nursing Data - Mapping an Interface Terminology to SNOMED CT. Studies in Health Technology and Informatics. 2024 Aug 22 ; 316:1297-301. DOI : 10.3233/SHTI240650

142. Fung, KW, Xu, J, Brear, H, Lane, A, Lau, M, Wong, A, Promoting interoperability between SNOMED CT and ICD-11 : lessons learned from the pilot project mapping between SNOMED CT and the ICD-11 Foundation. Journal of the American Medical Informatics Association : JAMIA. 2024 Aug 1 ; 31:1631-7. DOI : 10.1093/jamia/ocae143

143. Ngouongo, SMN, Löbe, M, Stausberg, J, The ISO/IEC 11179 norm for metadata registries : Does it cover healthcare standards in empirical research ? Journal of Biomedical Informatics. 2013 Apr 1 ; 46:318-27. DOI : 10.1016/j.jbi.2012.11.008

144. Parr, SK, Shotwell, MS, Jeffery, AD, Lasko, TA, Matheny, ME, Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. Journal of the American Medical Informatics Association : JAMIA. 2018 Oct 1 ; 25:1292-300. DOI : 10.1093/jamia/ocy110

145. Otero-Cerdeira, L, Rodríguez-Martínez, FJ, Gómez-Rodríguez, A, Ontology matching : A literature review. Expert Systems with Applications. 2015 Feb 1 ; 42:949-71. DOI : 10.1016/j.eswa.2014.08.032

146. Ochieng, P, Kyanda, S, Large-Scale Ontology Matching : State-of-the-Art Analysis. ACM Comput Surv. 2018 Jul 25 ; 51:75 :1-75 :35. DOI : 10.1145/3211871

147. Euzenat, J, Shvaiko, P, Ontology Matching. Berlin, Heidelberg : Springer Berlin Heidelberg, 2013. DOI : 10.1007/978-3-642-38721-0

148. Murdoch, B. Privacy and artificial intelligence : challenges for protecting health information in a new era. BMC medical ethics. 2021 Sep 15 ; 22:122. DOI : 10.1186/s12910-021-00687-3

149. Wolfien, M, Ahmadi, N, Fitzer, K, Grummt, S, Heine, KL, Jung, IC, Ten Topics to Get Started in Medical Informatics Research. Journal of Medical Internet Research. 2023 Jul 24 ; 25:e45948. DOI : 10.2196/45948

150. Yadav, N, Pandey, S, Gupta, A, Dudani, P, Gupta, S, Rangarajan, K, Data Privacy in Healthcare : In the Era of Artificial Intelligence. Indian Dermatology Online Journal. 2023 Oct 27 ; 14:788-92. DOI : 10.4103/idoj.idoj_543_23

151. Abubakar, M, Hamdan, H, Mustapha, N, Aris, TNM, Instance-Based Ontology Matching : A Literature Review. Recent Advances on Soft Computing and Data Mining. Sous la dir. de Ghazali R, Deris MM, Nawi NM et Abawajy JH. Cham : Springer International Publishing, 2018 :455-69. DOI : 10.1007/978-3-319-72550-5_44

152. Winkler, WE. Matching and record linkage. WIREs Computational Statistics. 2014 ; 6:313-25. DOI : 10.1002/wics.1317

153. Azzalini, F, Jin, S, Renzi, M, Tanca, L, Blocking Techniques for Entity Linkage : A Semantics-Based Approach. Data Science and Engineering. 2021 Mar 1 ; 6:20-38. DOI : 10.1007/s41019-020-00146-w

154. UCUM. Available from: https://ucum.org/ [Accessed on: 2024 Oct 10]

155. Arrêté du 11 août 2021 relatif à un programme de financement destiné à encourager l'équipement numérique des laboratoires de biologie médicale - Fonction « Transcodeur LOINC ». 2021 Aug 11. Available from: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043934053 [Accessed on: 2024 Sep 2]

156. Nikiema, JN, Mougin, F, Jouhet, V, Building a Graph Representation of LOINC® to Facilitate its Alignment to French Terminologies. AMIA Annual Symposium proceedings AMIA Symposium. 2020 ; 2020:933-42

157. Nikiema, JN, **Griffier, R**, Jouhet, V, Mougin, F, Aligning an interface terminology to the Logical Observation Identifiers Names and Codes (LOINC®). JAMIA open. 2021 Apr ; 4:ooab035. DOI : 10.1093/jamiaopen/ooab035

158. Michel-Picque, T, Bringay, S, Poncelet, P, Patel, N, Mayoral, G, Classifier Chains for LOINC Transcoding. Studies in Health Technology and Informatics. 2024 Aug 22 ; 316:1314-8. DOI : 10.3233/SHTI240654

159. Read, J, Pfahringer, B, Holmes, G, Frank, E, Classifier chains for multi-label classification. Machine Learning. 2011 Dec 1 ; 85:333-59. DOI : 10.1007/s10994-011-5256-5

160. Kelly, J, Wang, C, Zhang, J, Das, S, Ren, A, Warnekar, P, Automated Mapping of Real-world Oncology Laboratory Data to LOINC. AMIA Annual Symposium Proceedings. 2022 Feb 21 ; 2021:611-20

161. Sujansky, W. Heterogeneous database integration in biomedicine. Journal of Biomedical Informatics. 2001 Aug ; 34:285-98. DOI : 10.1006/jbin.2001.1024

162. Wollersheim, D, Sari, A, Rahayu, W, Archetype-based electronic health records : a literature review and evaluation of their applicability to health data interoperability and access. Health Information Management : Journal of the Health Information Management Association of Australia. 2009 ; 38:7-17. DOI : 10.1177/183335830903800202

163. Müller, A, Christmann, LS, Kohler, S, Eils, R, Prasser, F, Machine Learning for Medical Data Integration. Studies in Health Technology and Informatics. 2023 May 18 ; 302:691-5. DOI : 10.3233/SHTI230241

164. De Moor, G, Sundgren, M, Kalra, D, Schmidt, A, Dugas, M, Claerhout, B, Using electronic health records for clinical research : the case of the EHR4CR project. Journal of Biomedical Informatics. 2015 Feb ; 53:162-73. DOI : 10.1016/j.jbi.2014.10.006

165. Guérin, J, Laizet, Y, Le Texier, V, Chanas, L, Rance, B, Koeppel, F, OSIRIS : A Minimum Data Set for Data Sharing and Interoperability in Oncology. JCO Clinical Cancer Informatics. 2021 Mar 15 ; 5:CCI.20.00094. DOI : 10.1200/CCI.20.00094

166. Forrest, CB, McTigue, KM, Hernandez, AF, Cohen, LW, Cruz, H, Haynes, K, PCORnet® 2020 : current state, accomplishments, and future directions. Journal of Clinical Epidemiology. 2021 Jan ; 129:60-7. DOI : 10.1016/j.jclinepi.2020.09.036

167. Pressat-Laffouilhère, T, Balayé, P, Dahamna, B, Lelong, R, Billey, K, Darmoni, SJ, Evaluation of Doc'EDS : a French semantic search tool to query health documents from a clinical data warehouse. BMC Medical Informatics and Decision Making. 2022 Feb 8 ; 22:34. DOI : 10.1186/s12911-022-01762-4

168. Verma, N, Mamlin, B, Flowers, J, Acharya, S, Labrique, A, Cullen, T, OpenMRS as a global good : Impact, opportunities, challenges, and lessons learned from fifteen years of implementation. International Journal of Medical Informatics. 2021 May ; 149:104405. DOI : 10.1016/j.ijmedinf.2021.104405

169. Madec, J, Bouzillé, G, Riou, C, Van Hille, P, Merour, C, Artigny, ML, eHOP Clinical Data Warehouse : From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Studies in Health Technology and Informatics. 2019 Aug 21 ; 264:1536-7. DOI : 10.3233/SHTI190522

170. Jungkunz, M, Köngeter, A, Mehlis, K, Winkler, EC, Schickhardt, C, Secondary Use of Clinical Data in Data-Gathering, Non-Interventional Research or Learning Activities : Definition, Types, and a Framework for Risk Assessment. Journal of Medical Internet Research. 2021 Jun 8 ; 23:e26631. DOI : 10.2196/26631

171. Shivade, C, Raghavan, P, Fosler-Lussier, E, Embi, PJ, Elhadad, N, Johnson, SB, A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association : JAMIA. 2014 ; 21:221-30. DOI : 10.1136/amiajnl-2013-001935

172. Schlegel, DR, Ficheur, G, Secondary Use of Patient Data : Review of the Literature Published in 2016. Yearbook of Medical Informatics. 2017 Aug ; 26:68-71. DOI : 10.15265/IY-2017-032

173. Murdoch, TB, Detsky, AS, The inevitable application of big data to health care. JAMA. 2013 Apr 3 ; 309:1351-2. DOI : 10.1001/jama.2013.393

174. Sahatqija, K, Ajdari, J, Zenuni, X, Raufi, B, Ismaili, F, Comparison between relational and NOSQL databases. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2018 May :216-221. DOI : 10.23919/MIPRO.2018.8400041

175. Cattell, R. Scalable SQL and NoSQL data stores. SIGMOD Rec. 2011 May 6 ; 39:12-27. DOI : 10.1145/1978915.1978919

176. Davoudian, A, Chen, L, Liu, M, A Survey on NoSQL Stores. ACM Comput Surv. 2018 Apr 17 ; 51:40 :1-40 :43. DOI : 10.1145/3158661

177. Ercan, M, Lane, M, An evaluation of the suitability of NoSQL databases for distributed EHR systems. ACIS, 2014. Available from: https://openrepository.aut.ac.nz/handle/10292/8134 [Accessed on: 2024 Oct 13]

178. Gudivada, VN, Rao, D, Raghavan, VV, Renaissance in Database Management : Navigating the Landscape of Candidate Systems. Computer. 2016 Apr ; 49. Conference Name : Computer:31-42. DOI : 10.1109/MC.2016.115

179. Sánchez-de-Madariaga, R, Muñoz, A, Lozano-Rubí, R, Serrano-Balazote, P, Castro, AL, Moreno, O, Examining database persistence of ISO/EN 13606 standardized electronic health record extracts : relational vs. NoSQL approaches. BMC Medical Informatics and Decision Making. 2017 Aug 18 ; 17:123. DOI : 10.1186/s12911-017-0515-4

180. Kononenko, O, Baysal, O, Holmes, R, Godfrey, MW, Mining modern repositories with elasticsearch. Proceedings of the 11th Working Conference on Mining Software Repositories. MSR 2014. New York, NY, USA : Association for Computing Machinery, 2014 May 31:328-31. DOI : 10.1145/2597073.2597091

181. Gormley, C, Tong, Z, Elasticsearch : The Definitive Guide : A Distributed Real-Time Search and Analytics Engine. O'Reilly Media, Inc., 2015 Jan 23. 719 p.

182. Białecki, A, Muir, R, Ingersoll, G, Apache Lucene 4. OSIR@SIGIR. 2012. Available from: https://www.semanticscholar.org/paper/Apache-Lucene-4-Bia%C5%82ecki-Muir/2795d9d165607b5ad6d8b9718373b82e55f41606 [Accessed on: 2024 Aug 19]

183. Blair, DC. Information Retrieval, 2nd ed. C.J. Van Rijsbergen. London : Butterworths ; 1979 : 208 pp. Price : $32.50. Journal of the American Society for Information Science. 1979 ; 30:374-5. DOI : 10.1002/asi.4630300621

184. Kim, HJ, Ko, EJ, Jeon, YH, Lee, KH, Migration from RDBMS to Column-Oriented NoSQL : Lessons Learned and Open Problems. Proceedings of the 7th International Conference on Emerging Databases. Sous la dir. de Lee W, Choi W, Jung S et Song M. Singapore : Springer, 2018 :25-33. DOI : 10.1007/978-981-10-6520-0_3

185. Tunkelang, D. Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Cham : Springer International Publishing, 2009. DOI : 10.1007/978-3-031-02262-3. [Accessed on: 2024 Jun 27]

186. Oracle, Oracle Text. 2013. Available from: https://docs.oracle.com/database/121/DFSIG/oracle-text.htm#DFSIG269 [Accessed on: 2024 Aug 19]

187. Garcelon, N, Neuraz, A, Benoit, V, Salomon, R, Burgun, A, Improving a full-text search engine : the importance of negation detection and family history context to identify cases in a biomedical data warehouse. Journal of the American Medical Informatics Association : JAMIA. 2017 May 1 ; 24:607-13. DOI : 10.1093/jamia/ocw144

188. Villani, C, Schoenauer, M, Bonnet, Y, Berthet, C, Cornut, AC, Levin, F, Donner un sens a l'intelligence artificielle : pour une strategie nationale et européenne. Donner un sens a l'intelligence artificielle : pour une strategie nationale et européenne. 2018 Mar 28. Available from: https://www.vie-publique.fr/files/rapport/pdf/184000159.pdf [Accessed on: 2024 Oct 1]

189. Jannot, AS, Zapletal, E, Avillach, P, Mamzer, MF, Burgun, A, Degoulet, P, The Georges Pompidou University Hospital Clinical Data Warehouse : A 8-years follow-up experience. International Journal of Medical Informatics. 2017 Jun 1 ; 102:21-8. DOI : 10.1016/j.ijmedinf.2017.02.006

190. Artemova, S, Madiot, PE, Caporossi, A, PREDIMED group, Mossuz, P, Moreau-Gaudry, A, PREDIMED : Clinical Data Warehouse of Grenoble Alpes University Hospital. Studies in Health Technology and Informatics. 2019 Aug 21 ; 264:1421-2. DOI : 10.3233/SHTI190464

191. Cossin, S, Diouf, S, **Griffier, R**, Le Barrois d'Orgeval, P, Diallo, G, Jouhet, V, Linkage of Hospital Records and Death Certificates by a Search Engine and Machine Learning. JAMIA open. 2021 Jan ; 4:ooab005. DOI : 10.1093/jamiaopen/ooab005

192. Cossin, S, Lebrun, L, Aymeric, N, Mougin, F, Lambert, M, Diallo, G, SmartCRF : A Prototype to Visualize, Search and Annotate an Electronic Health Record from an i2b2 Clinical Data Warehouse. Studies in Health Technology and Informatics. 2019 Aug 21 ; 264:1445-6. DOI : 10.3233/SHTI190476

193. Cossin, S, Lebrun, L, Lobre, G, Loustau, R, Jouhet, V, **Griffier, R,** Romedi : An Open Data Source About French Drugs on the Semantic Web. Studies in Health Technology and Informatics. 2019 Aug 21 ; 264:79-82. DOI : 10.3233/SHTI190187

194. Platform for Data in Primary care - P4DP. https ://www.p4dp.fr/. Available from: https://www.p4dp.fr/ [Accessed on: 2024 Oct 13]

195. European Commission, European Health Data Space. 2024 Apr 24. Available from: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [Accessed on: 2024 Aug 24]

196. Article 13 EU règlement général sur la protection des données (EU-RGPD). 2016 Apr 25. Available from: https://www.privacy-regulation.eu/fr/13.htm [Accessed on: 2024 Sep 28]

197. Article 14 EU règlement général sur la protection des données (EU-RGPD). 2016 Apr 25. Available from: https://www.privacy-regulation.eu/fr/14.htm [Accessed on: 2024 Sep 28]

198. Article 21 EU règlement général sur la protection des données (EU-RGPD). 2016 Apr 25. Available from: https://www.privacy-regulation.eu/fr/21.htm [Accessed on: 2024 Sep 28]

# Annexes

# ANNEXE A

# DARWIN-EU

| Country | Data Partner | Data Source |
|---|---|---|
| Belgium | IQVIA LPD Belgium | IQVIA Longitudinal Patient Database Belgium |
| Croatia | CIPH | Croatian National Public Health Information System |
| Denmark | DKMA | Danish Data Health Registries |
| Estonia | EBB | Estonian Biobank |
| Finland | FinOMOP | Tampere University Hospital patient cohort |
| Finland | FinOMOP | Hospital District of Helsinki and Uusimaa |
| Finland | FinOMOP | Finnish Care Register for Health Care |
| Finland | FinOMOP | Auria Clinical Informatics |
| France | CDW Bordeaux | Clinical Data Warehouse of Bordeaux University Hospital |
| France | HDH | Système National des Données de Santé |
| Germany | IQVIA DA Germany | IQVIA Disease Analyzer Germany |
| Hungary | SUCD | Semmelweis University Clinical Data |
| Netherlands | IPCI | Integrated Primary Care Information |
| Netherlands | NCR | Netherlands Cancer Registry |
| Norway | NLHR | Norwegian Linked Health Registry data |
| Portugal | EMDB | Egas Moniz Health Alliance database - Baixo Vouga (Região de Aveiro) |
| Portugal | EMDB | Egas Moniz Health Alliance database - Gaia E Espinho |
| Portugal | EMDB | Egas Moniz Health Alliance database - Entre o Douro e Vouga |
| Portugal | ULSM-RT | Unidade Local de Saúde de Matosinhos Realtime Database |
| Spain | BIFAP | BIFAP |

Tableau A.1 – Liste des *data partners* inclus dans le réseau DARWIN-EU

Extrait de `https://www.darwin-eu.org/index.php/data/data-network`

# ANNEXE B

## Persistance du modèle de données d'i2b2 dans une base de données Elasticsearch

Listing B.1 – Requête SQL utilisée pour charger l'index observation_fact

```
SELECT
    o.*,
    c.CONCEPT_C_FULLNAME_LIST,
    c.CONCEPT_C_FULLNAME_LIST_HASH,
    m.MODIFIER_C_FULLNAME_LIST,
    m.MODIFIER_C_FULLNAME_LIST_HASH,
    MONTHS_BETWEEN(v.START_DATE, p.BIRTH_DATE) as
        AGE_IN_MONTH_AT_START_VISIT,
    p.SEX_CD
FROM
    OBSERVATION_FACT o
    LEFT OUTER JOIN I2B2_PATH_CONCEPT c
        ON o.CONCEPT_CD = c.CONCEPT_CD
    LEFT OUTER JOIN I2B2_PATH_MODIFIER m
        ON o.MODIFIER_CD = m.MODIFIER_CD
    LEFT OUTER JOIN VISIT_DIMENSION v
        ON o.ENCOUNTER_NUM = v.ENCOUNTER_NUM
        AND o.PATIENT_NUM = v.PATIENT_NUM
    LEFT OUTER JOIN PATIENT_DIMENSION p
        o.PATIENT_NUM = p.PATIENT_NUM
```

# ANNEXE C

## Information patient, transparence et exercice des droits au CHU de Bordeaux

L'utilisation secondaire des données de santé issues du soin est basée sur les données individuelles de santé relatives aux patients. Une information et une transparence vis-à-vis des patients est un élément fondamental dans la confiance entre les usagers et les acteurs de la recherche.

Au sein de l'UE, les articles 13 et 14 du RGPD [196, 197] indiquent que : « *Lorsqu'il a l'intention d'effectuer un traitement ultérieur des données à caractère personnel pour une finalité autre que celle pour laquelle les données à caractère personnel ont été collectées, le responsable du traitement fournit au préalable à la personne concernée des informations au sujet de cette autre finalité et toute autre information pertinente visée au paragraphe 2.* ». Dans un contexte de recherche, cela implique une information individuelle et préalable à chaque finalité, c'est-à-dire à chaque étude, recherche ou évaluation.

Dans le contexte des EDS hospitaliers, la CNIL autorise la réalisation d'une information générale sur l'utilisation secondaire des données de santé associée à la mise à disposition, au travers d'un portail de transparence, de la liste des études réalisées. Ce portail de transparence correspond ici à une information par finalité.

Au CHU de Bordeaux, la stratégie d'information implique :

1. Une **information collective générale** : il s'agit ici d'informer de manière collective les patients sur le fait que les données de santé les concernant peuvent être utilisées pour des finalités autres que le soin. Cette information collective générale implique notamment des campagnes d'affichage dans les lieux accueillant du public, des mentions sur l'utilisation secondaire des données de santé dans le

livret d'accueil ou encore une page dédiée sur le site Internet du CHU de Bordeaux. Hors les murs, cette information collective générale implique également de la communication dans les médias locaux ainsi que sur les réseaux sociaux.

2. Une **information individuelle générale** : il s'agit également ici d'informer les patients de manière générale sur l'utilisation secondaire des données de santé, mais cette fois au travers d'une information individuelle. Au CHU de Bordeaux, cette information individuelle générale passe par :

   — L'information individuelle unique et préalable (dite « One-Shot »), au travers d'un échange dédié entre un professionnel de santé et un patient. Au cours de cet échange, une note d'information est remise au patient. En cas d'opposition du patient, une traçabilité dans le DPI est assurée par les professionnels de santé.

   — L'envoi de mails ciblés, à destination des patients sortis du CHU de Bordeaux. Ce mail contient les pointeurs vers les différentes pages du site Internet relatives à l'utilisation secondaire des données de santé, ainsi qu'un lien vers la note d'information.

3. Une **information collective par finalité**, au travers du portail de transparence du CHU de Bordeaux. Ce dernier liste l'ensemble des études réalisées à partir de l'EDS du CHU de Bordeaux. Il permet également à un patient de contacter l'investigateur principal d'une étude, ou le DPO, afin d'exercer ses droits le cas échéant.

Associé à cette information et cette transparence, l'exercice des droits des patients est un élément socle fondamental. Différents droits peuvent être exercés par les usagers pour des traitements basés sur les données qui les concernent, et notamment le droit d'opposition. L'article 21 du RGPD [198] définit : « *La personne concernée a le droit de s'opposer à tout moment, pour des raisons tenant à sa situation particulière, à un traitement des données à caractère personnel la concernant [...].* ».

Au CHU de Bordeaux, ce droit d'opposition peut être exercé au moment de l'information, avec traçabilité de l'opposition dans le DPI par le professionnel de santé, ou après l'information via le portail de transparence.

# ANNEXE D

## La visionneuse de l'EDS i2b2 du CHU de Bordeaux

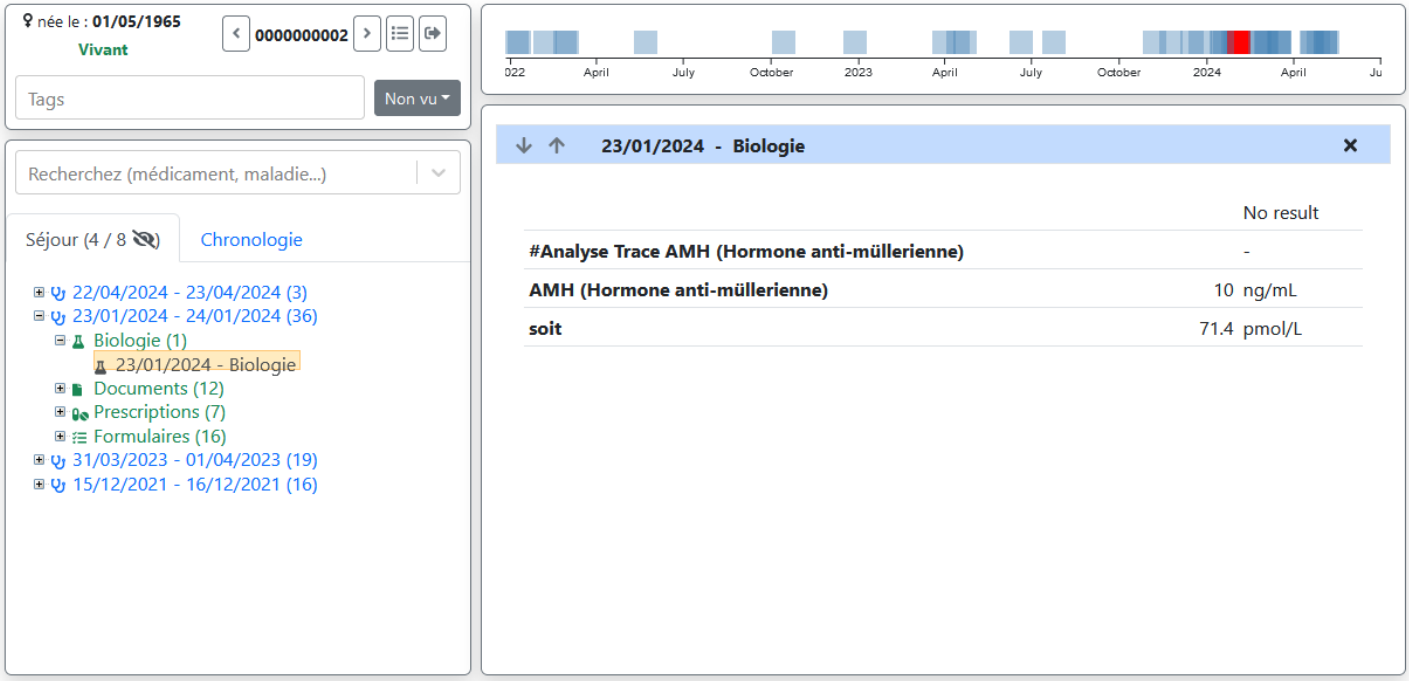


FIGURE D.1 – Visionneuse du dossier médical intégré dans l'EDS i2b2 - Vue biologie

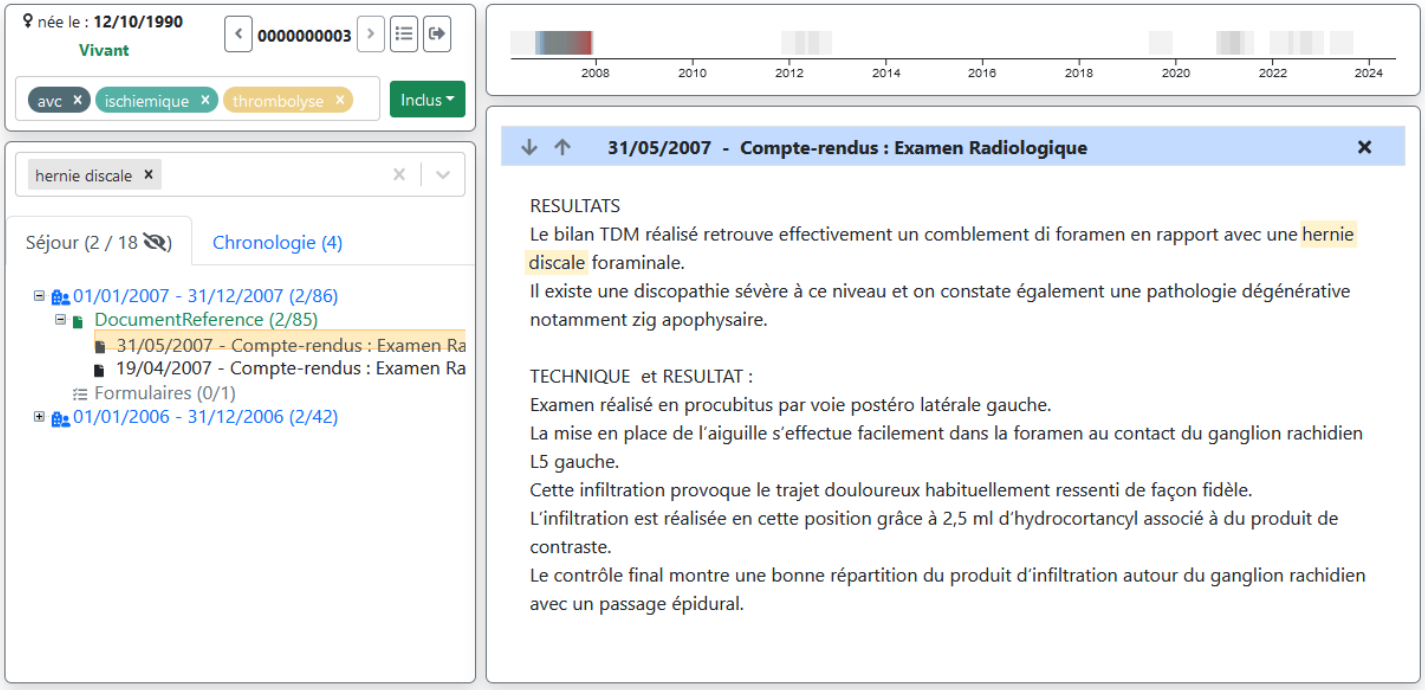


FIGURE D.2 – Visionneuse du dossier médical intégré dans l'EDS i2b2 - Vue documents

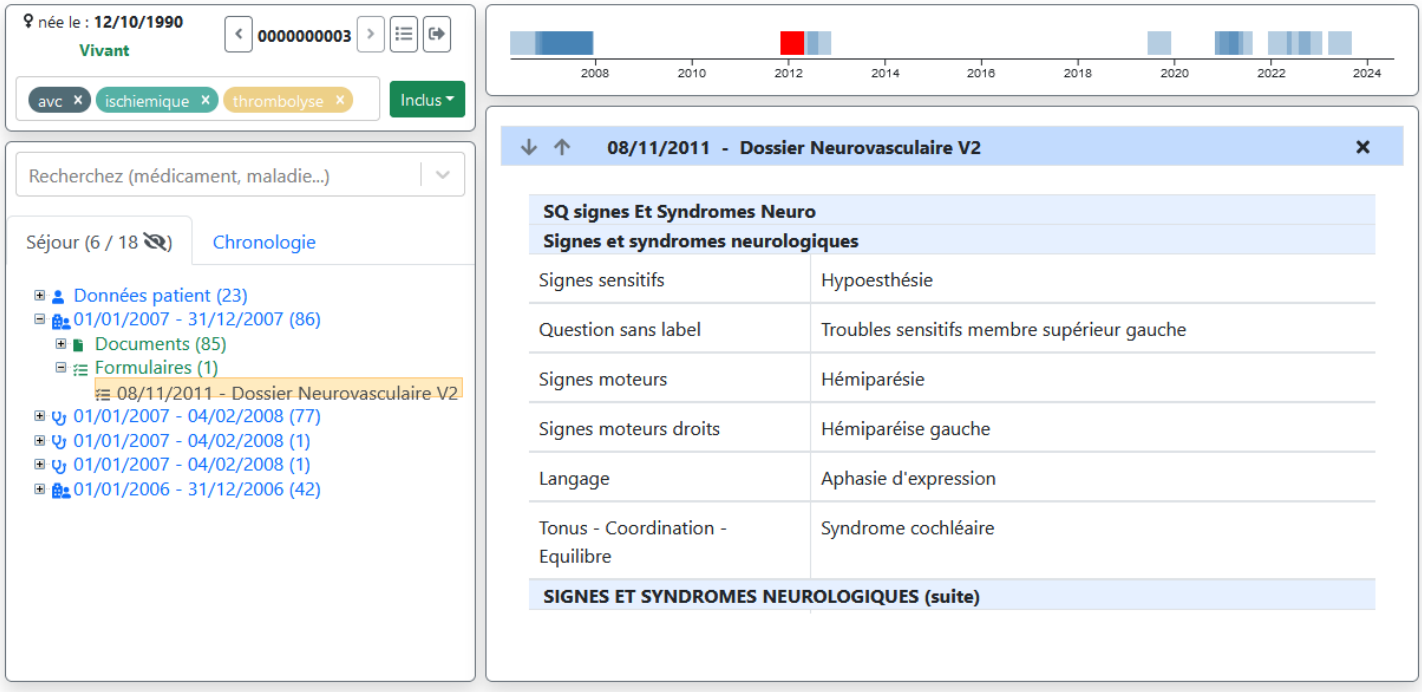


FIGURE D.3 – Visionneuse du dossier médical intégré dans l'EDS i2b2 - Vue formulaires

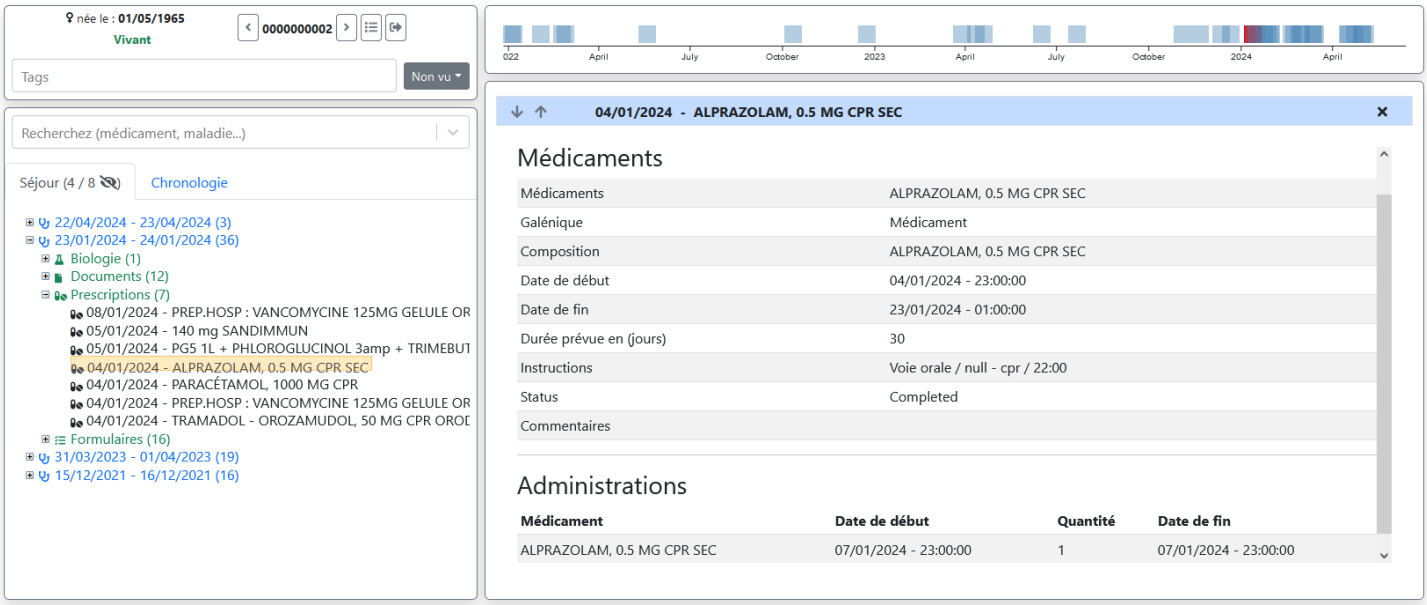


FIGURE D.4 – Visionneuse du dossier médical intégré dans l'EDS i2b2 - Vue prescription

# ANNEXE E

## Résultat du CDMOnboarding de l'EDS OMOP du CHU de Bordeaux

# CDM Onboarding report for the OMOP database

Romain GRIFFIER, Vianney JOUHET, Guillaume VERDY

## 1. Execution details

| Detail | Value |
|---|---|
| CdmOnboarding package version | 3.1.0 |
| Database | postgresql |
| CDM version | v5.4 |
| Execution date | 2024-10-04 |
| Execution duration | 58m 31s |
| Achilles version | 1.7.2 |
| Achilles execution date | 2024-09-30 |

Symbols used in table/figure captions: Ⓞ=Computed directly from OMOP CDM data, Ⓐ=Computed from Achilles results, Ⓢ=Estimated from system tables.

### *1.1. CDM Source Table*

**Table 1.** Content of the OMOP cdm_source table Ⓞ

| field | Values |
|---|---|
| CDM_SOURCE_NAME | Clinical DataWarehouse Bordeaux University Hospital |
| CDM_SOURCE_ABBREVIATION | CDWBordeaux |
| CDM_HOLDER | CHU de Bordeaux |
| SOURCE_DESCRIPTION | Electronic Health Records of the Bordeaux University Hospital, France |
| SOURCE_DOCUMENTATION_REFERENCE | https://gitub.u-bordeaux.fr/scossi910e/ehden-bordeaux/-/wikis/home |
| CDM_ETL_REFERENCE | https://gitub.u-bordeaux.fr/scossi910e/ehden-bordeaux-etl-omop/-/wikis/home |
| SOURCE_RELEASE_DATE | 2024-02-22 |
| CDM_RELEASE_DATE | 2024-02-22 |
| CDM_VERSION | v5.4 |
| CDM_VERSION_CONCEPT_ID | 756265 |
| VOCABULARY_VERSION | v5.0 31-AUG-23 |

# 2. Clinical data

## *2.1. Record counts per OMOP CDM table*

**Table 2.** The number of records in all clinical data tables ©

| Table | #Records | #Persons | %Persons |
|---|---|---|---|
| measurement | 589,494,508 | 1,208,600 | 52.4% |
| drug_exposure | 121,131,921 | 717,602 | 31.1% |
| condition_occurrence | 27,645,929 | 987,700 | 42.8% |
| visit_detail | 20,249,963 | 2,298,554 | 99.7% |
| procedure_occurrence | 18,068,539 | 1,405,772 | 61.0% |
| visit_occurrence | 17,539,096 | 2,298,974 | 99.7% |
| condition_era | 15,370,834 | 964,792 | 41.8% |
| drug_era | 11,991,638 | 698,840 | 30.3% |
| person | 2,306,342 | 2,306,342 | 100% |
| observation_period | 2,298,975 | 2,298,975 | 99.7% |
| death | 222,428 | 222,428 | 9.6% |
| care_site | 3,278 | NA | NA |
| location | 14 | NA | NA |
| provider | 1 | NA | NA |
| metadata | 1 | NA | NA |
| cdm_source | 1 | NA | NA |
| device_exposure | 0 | 0 | 0% |
| dose_era | 0 | 0 | 0% |
| note | 0 | 0 | 0% |
| observation | 0 | 0 | 0% |
| payer_plan_period | 0 | 0 | 0% |
| specimen | 0 | 0 | 0% |
| episode | 0 | 0 | 0% |
| cost | 0 | NA | NA |
| note_nlp | 0 | NA | NA |
| fact_relationship | 0 | NA | NA |
| episode_event | 0 | NA | NA |

Query executed in 662.03 seconds

### 2.2. *Data density plots*

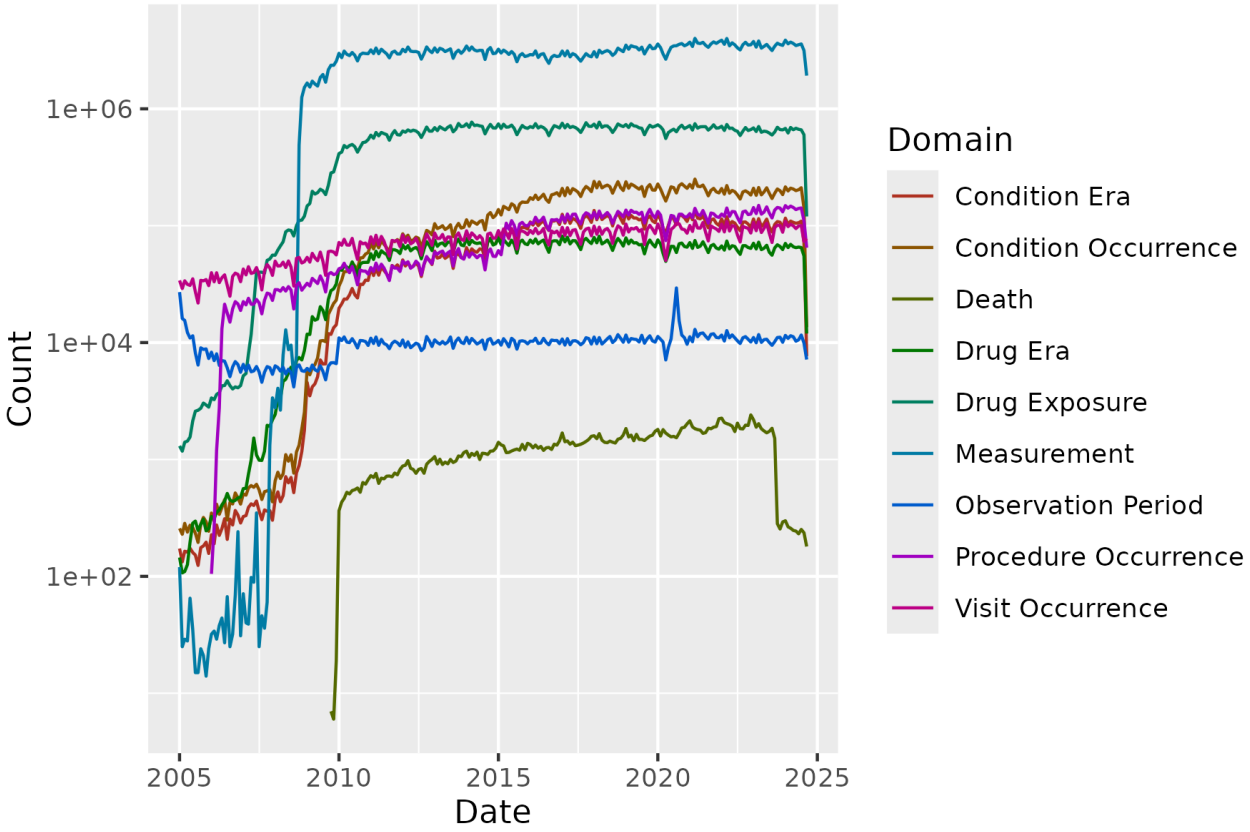


**Figure 1.** Total record count over time per OMOP data domain. Ⓐ

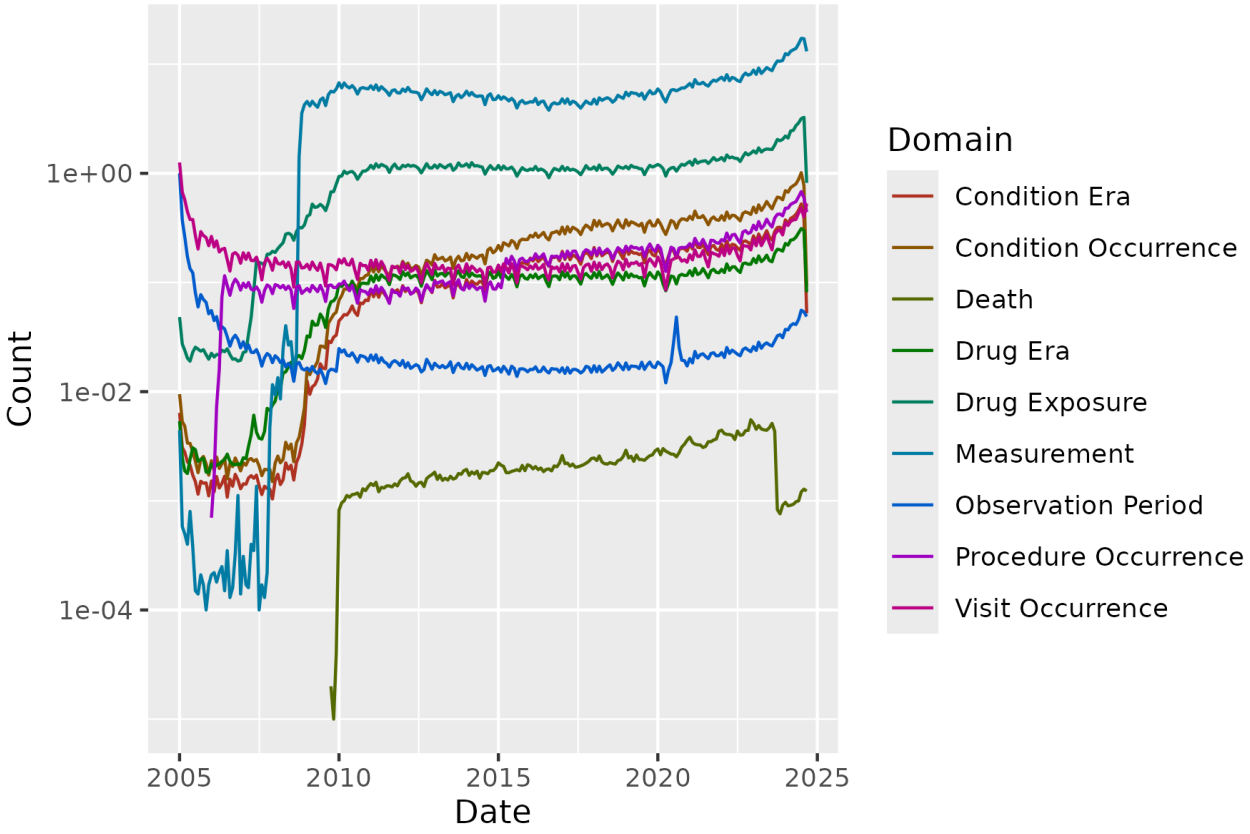


**Figure 2.** Number of records per person over time per OMOP data domain. Ⓐ

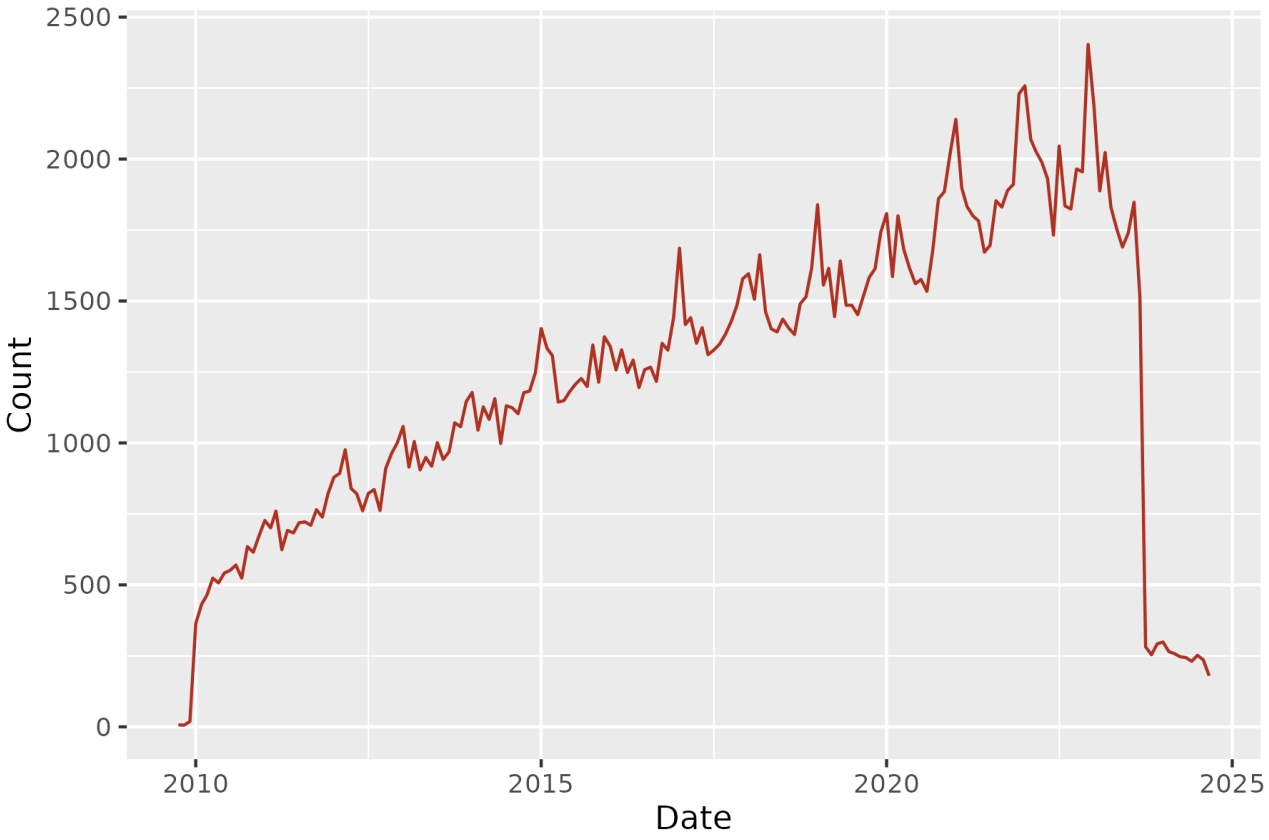


**Figure 3.** Number of deaths in each month. Overall mortality: 9.64%. Ⓐ

## 2.3. Distinct concepts per person

**Table 3.** The number of distinct concepts per person per OMOP data domains. Only persons with at least one record in that domain are included in the calculation. Ⓐ

| DOMAIN | N_PERSONS | MIN | P10 | P25 | MEDIAN | P75 | P90 | MAX |
|---|---|---|---|---|---|---|---|---|
| Visit Occurrence | 2,298,974 | 1 | 1 | 1 | 1 | 2 | 3 | 4 |
| Condition Occurrence | 987,231 | 1 | 1 | 2 | 6 | 13 | 28 | 333 |
| Procedure Occurrence | 1,403,962 | 1 | 1 | 2 | 3 | 6 | 12 | 91 |
| Drug exposure | 717,544 | 1 | 1 | 3 | 8 | 16 | 30 | 268 |
| Drug era | 698,777 | 1 | 1 | 3 | 8 | 16 | 28 | 198 |
| Condition era | 964,548 | 1 | 1 | 2 | 5 | 13 | 28 | 332 |
| Measurement | 1,205,352 | 1 | 1 | 7 | 42 | 64 | 94 | 477 |

Query executed in 0.14 seconds

## 2.4. Observation Period

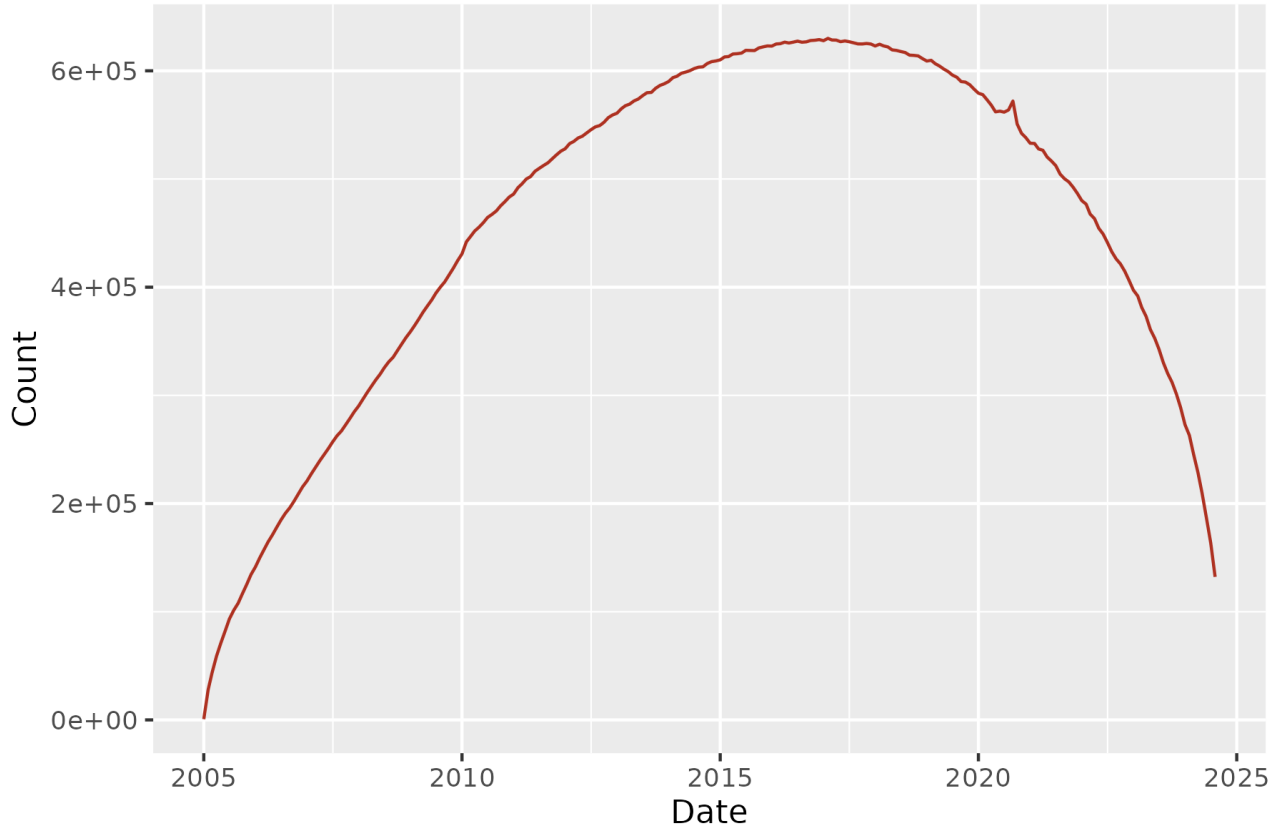


**Figure 4.** Persons with continuous observation by month.Ⓐ In the last 6 months (before 2024-02-22), there are 482,460 persons with an active observation period.Ⓞ

**Table 4.** Length of first observation period (years). Ⓐ

| AVG | STDEV | MIN | P10 | P25 | MEDIAN | P75 | P90 | MAX |
|---|---|---|---|---|---|---|---|---|
| 4 | 5.2 | 0 | 0.2 | 0.2 | 1.1 | 6.7 | 12.6 | 19.7 |

Query executed in 0.19 seconds

**Table 5.**

| Field | Value | %Persons |
|---|---|---|
| Persons with 1 observation period(s) | 2,298,975 | 99.68058 |
| Persons with overlapping observation periods | 0 | NA |
| Number of overlapping observation periods | 0 | NA |

Queries executed in 0.17 seconds and 2.01 seconds

## *2.5. Date Range*

**Table 6.** Minimum and maximum event start date in each table, within an observation period and at least 5 records. Floored to the nearest month. Ⓐ

| Domain | N | Type | Start date [Min, Max] | End date [Min, Max] |
|---|---|---|---|---|
| Condition | 27,206,912 | Claim discharge record (S) | [2006-01, 2024-09] | [2006-01, 2024-09] |
| Condition | 439,017 | EHR Pathology report (S) | [2001-08, 2024-10] | [2001-08, 2024-10] |
| Death | 134,808 | EHR (S) | [2005-04, 2024-09] | [NA, NA] |
| Death | 87,620 | Government report (S) | [2005-02, 2023-09] | [NA, NA] |
| Drug | 1,115,981 | EHR prescription (S) | [2020-11, 2024-09] | [2020-11, 2024-11] |
| Drug | 120,015,940 | EHR administration record (S) | [1899-12, 2202-04] | [1899-12, 2202-04] |
| Measurement | 589,351,171 | Lab result (NA) | [1201-01, 3022-03] | [NA, NA] |
| Measurement | 143,337 | EHR Pathology report (S) | [2014-01, 2024-09] | [NA, NA] |
| Obs. Period | 2,298,975 | Period covering healthcare encounters (NA) | [2005-01, 2024-09] | [2005-03, 2024-09] |
| Procedure | 18,068,539 | Hospitalization Cost Record (NA) | [2006-01, 2024-12] | [NA, NA] |
| Visit | 11,768 | Still patient (NA) | [2005-01, 2024-09] | [2024-09, 2024-09] |
| Visit | 17,527,328 | EHR administration record (S) | [2005-01, 2024-09] | [2005-01, 2024-09] |
| Visit Detail | 20,249,963 | EHR administration record (S) | [2000-01, 2024-09] | [2000-01, 2202-12] |

Query executed in 331.70 seconds

**Table 7.** Length of stay by visit concept. The length should be interpreted as number of nights, meaning a length of 0 is a same-day visit. Ⓐ

| Domain | Concept Name | AVG | STDEV | MIN | P10 | P25 | MEDIAN | P75 | P90 | MAX |
|---|---|---|---|---|---|---|---|---|---|---|
| Visit | Emergency Room Visit | 0.2 | 4.8 | 0 | 0 | 0 | 0 | 0 | 1 | 4,396 |
| Visit | Emergency Room and Inpatient Visit | 4.3 | 9.9 | 0 | 0 | 0 | 1 | 5 | 11 | 690 |
| Visit | Inpatient Visit | 10.6 | 53.0 | -730 | 0 | 0 | 2 | 5 | 15 | 7,208 |
| Visit | Outpatient Visit | 2.1 | 26.3 | 0 | 0 | 0 | 0 | 0 | 0 | 4,100 |
| Visit Detail | Emergency Room Visit | 0.1 | 1.4 | -1 | 0 | 0 | 0 | 0 | 1 | 961 |
| Visit Detail | Emergency Room and Inpatient Visit | 1.5 | 4.8 | -242 | 0 | 0 | 0 | 1 | 4 | 640 |
| Visit Detail | Inpatient Visit | 3.3 | 60.4 | -342 | 0 | 0 | 0 | 2 | 5 | 6,408 |
| Visit Detail | Intensive Care | 4.0 | 8.4 | -394 | 0 | 1 | 1 | 4 | 9 | 418 |
| Visit Detail | Outpatient Visit | 4.8 | 98.3 | -3,064 | 0 | 0 | 0 | 0 | 0 | 65,743 |

Query executed in 4.66 seconds

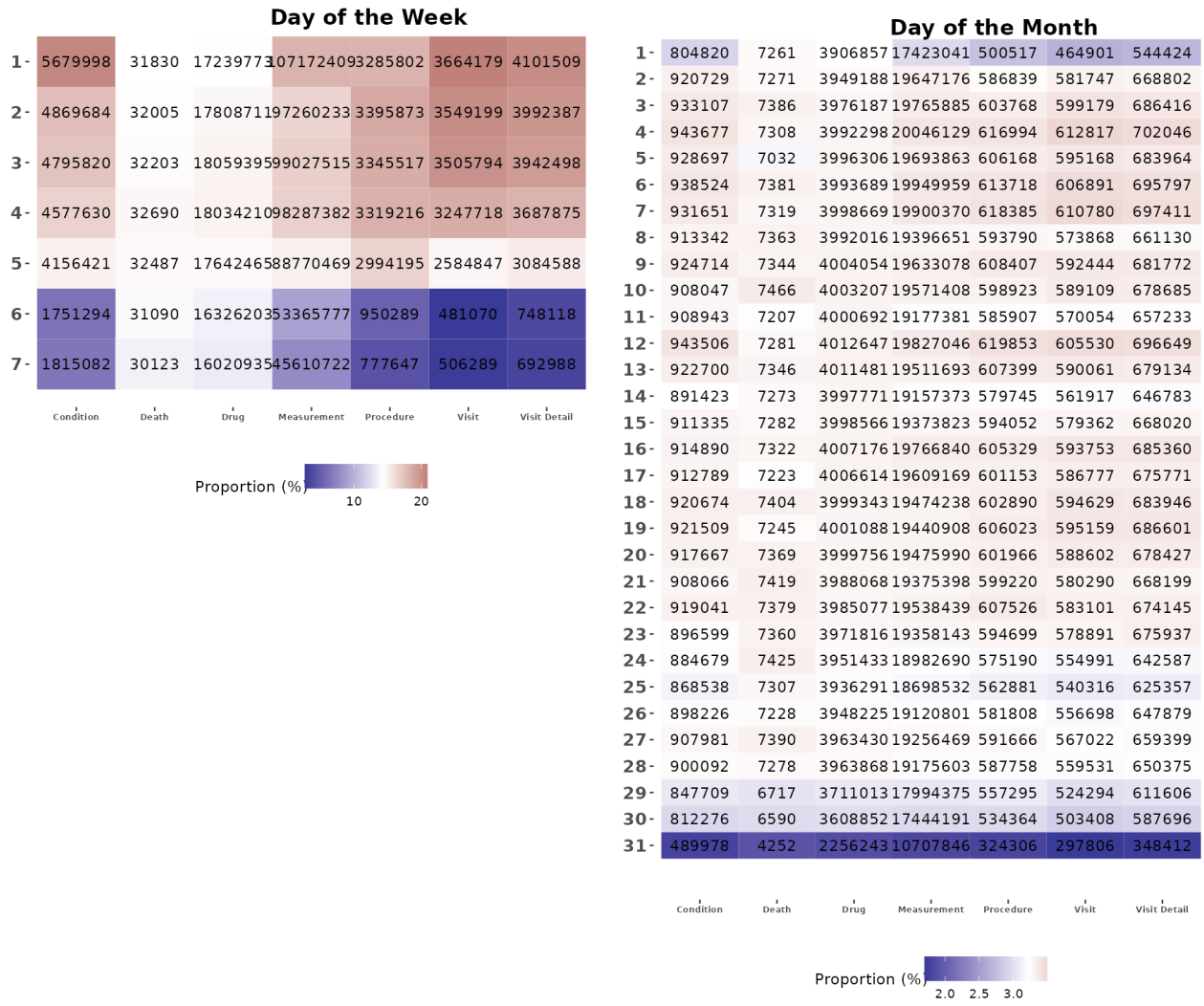


**Figure 5.** Day of the Week and Day of the Month distribution of event start dates after 1900-01-01 per domain. 1 = Monday ... 7 = Sunday. ⓪

Queries executed in 351.36 seconds and 418.29 seconds

## *2.6. Day, Month, Year of Birth*

**Table 8.** Distribution of day, month and year of birth of persons. ⓪

| | %Missing | MIN | P10 | P25 | MEDIAN | P75 | P90 | MAX |
|---|---|---|---|---|---|---|---|---|
| Year | 0% | 1900 | 1942 | 1956 | 1979 | 1997 | 2012 | 2024 |
| Month | 0% | 1 | 2 | 3 | 6 | 9 | 11 | 12 |
| Day | 0% | 1 | 3 | 8 | 16 | 23 | 28 | 31 |

Query executed in 5.67 seconds

# 3. Vocabulary mappings

Vocabulary version: v5.0 31-AUG-23
Pre-processing query executed in 734.41 seconds

## *3.1. Mapping Completeness*

**Table 9.** The number and percentage of codes and records that are mapped to an OMOP concept (not 0 and <2B). Note: for one-to-many mappings, the source codes will be counted multiple times so the reported total source codes could be bigger than actual number of unique source codes. ◎

| Domain | #Codes Source | #Codes Mapped | %Codes Mapped | #Records Source | #Records Mapped | %Records Mapped |
|---|---|---|---|---|---|---|
| Condition | 34,173 | 21,671 | 63.4% | 27,645,929 | 25,244,108 | 91.3% |
| Condition status | 993 | 993 | 100% | 27,645,929 | 27,645,929 | 100% |
| Death cause | 0 | NA | NA | NA | NA | NA |
| Device | 0 | NA | NA | NA | NA | NA |
| Drug | 22,307 | 14,095 | 63.2% | 121,131,921 | 115,648,066 | 95.5% |
| Drug Route | 48 | 47 | 97.9% | 99,481,689 | 95,895,662 | 96.4% |
| Measurement | 85,459 | 4,527 | 5.3% | 589,494,508 | 515,931,346 | 87.5% |
| Measurement Unit | 68 | 68 | 100% | 252,026,902 | 252,026,902 | 100% |
| Measurement value | 3,565 | 91 | 2.6% | 11,616,071 | 3,116,250 | 26.8% |
| Observation | 0 | NA | NA | NA | NA | NA |
| Observation Unit | 0 | NA | NA | NA | NA | NA |
| Observation value | NA | NA | NA | NA | NA | NA |
| Procedure | 8,420 | 1,028 | 12.2% | 18,068,539 | 12,448,429 | 68.9% |
| Provider Specialty | 0 | NA | NA | NA | NA | NA |
| Specimen | 0 | NA | NA | NA | NA | NA |
| Visit | 5 | 5 | 100% | 17,539,096 | 17,539,096 | 100% |
| Visit detail | 5 | 5 | 100% | 20,249,963 | 20,249,963 | 100% |

Query executed in 0.09 seconds

## *3.2. Drug Mappings*

**Table 10.** The level of the drug mappings ◎

| Class | #Records | #Patients | #Codes | %Records |
|---|---|---|---|---|
| Clinical Drug | 106,625,293 | 8,882,044 | 7,960 | 88.0% |
| Ingredient | 6,902,323 | 860,255 | 4,259 | 5.7% |
| Undefined | 5,483,855 | 1,048,085 | 8,212 | 4.5% |
| Clinical Drug Form | 1,949,469 | 163,453 | 236 | 1.6% |
| Quant Clinical Drug | 70,790 | 13,902 | 165 | 0.1% |
| Branded Drug | 63,432 | 6,638 | 31 | 0.1% |
| Clinical Drug Comp | 20,945 | 1,048 | 16 | 0.0% |
| Branded Drug Form | 8,818 | 920 | 2 | 0.0% |
| Branded Drug Comp | 5,127 | 968 | 5 | 0.0% |
| Quant Branded Drug | 1,869 | 335 | 17 | 0.0% |

Query executed in 1.03 seconds

## *3.3. Unmapped Codes*

**Table 11.** Top 25 unmapped drugs. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Source Value | #Records | %Records |
|---|---|---|---|
| 1 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|22347 | 599,500 | 0.5% |
| 2 | TRL:DRUG_PRESC-PSL_VALUEDOMAIN\|140 | 328,500 | 0.3% |
| 3 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|26311 | 228,200 | 0.2% |
| 4 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|20981 | 193,000 | 0.2% |
| 5 | ICCA\|PRESC:DRUG_COMPONENT-MATERIAL_VALUEDOMAIN\|42134 | 176,700 | 0.1% |
| 6 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|21234 | 167,900 | 0.1% |
| 7 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|25822 | 121,100 | 0.1% |
| 8 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|30464 | 120,900 | 0.1% |
| 9 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|22146 | 89,000 | 0.1% |
| 10 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|21068 | 70,600 | 0.1% |
| 11 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|20982 | 62,000 | 0.1% |
| 12 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|20980 | 61,000 | 0.1% |
| 13 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|30147 | 59,900 | 0.0% |
| 14 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|36175 | 58,500 | 0.0% |
| 15 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|20979 | 46,900 | 0.0% |
| 16 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|36004 | 46,800 | 0.0% |
| 17 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|21213 | 39,700 | 0.0% |
| 18 | TRL:DRUG_PRESC-PSL_VALUEDOMAIN\|252 | 39,300 | 0.0% |
| 19 | ICCA\|PRESC:DRUG_COMPONENT-MATERIAL_VALUEDOMAIN\|32881 | 38,700 | 0.0% |
| 20 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|32845 | 36,400 | 0.0% |
| 21 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|31803 | 34,800 | 0.0% |
| 22 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|32181 | 33,700 | 0.0% |
| 23 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|21690 | 33,300 | 0.0% |
| 24 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|34813 | 31,800 | 0.0% |
| 25 | ADM:DRUG_ADM-DRUG_VALUEDOMAIN\|437 | 30,900 | 0.0% |

Query executed in 0.04 seconds

**Table 12.** Top 25 unmapped conditions. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Source Value | #Records | %Records |
|---|---|---|---|
| 1 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.94 | 165,400 | 0.6% |
| 2 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.93 | 109,800 | 0.4% |
| 3 | PMSI:DIAGNOSTIC-ICD10CHU\|C78.7 | 93,600 | 0.3% |
| 4 | PMSI:DIAGNOSTIC-ICD10CHU\|C78.0 | 84,700 | 0.3% |
| 5 | PMSI:DIAGNOSTIC-ICD10CHU\|C79.5 | 68,100 | 0.2% |
| 6 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.95 | 60,200 | 0.2% |
| 7 | PMSI:DIAGNOSTIC-ICD10CHU\|R26.8 | 59,800 | 0.2% |
| 8 | PMSI:DIAGNOSTIC-ICD10CHU\|R40.0 | 52,500 | 0.2% |
| 9 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.04 | 51,400 | 0.2% |
| 10 | PMSI:DIAGNOSTIC-ICD10CHU\|R63.4 | 48,400 | 0.2% |
| 11 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.00 | 48,000 | 0.2% |
| 12 | PMSI:DIAGNOSTIC-ICD10CHU\|C79.3 | 46,900 | 0.2% |
| 13 | PMSI:DIAGNOSTIC-ICD10CHU\|C78.6 | 44,400 | 0.2% |
| 14 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.90 | 30,700 | 0.1% |
| 15 | PMSI:DIAGNOSTIC-ICD10CHU\|U08.9 | 30,300 | 0.1% |
| 16 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.05 | 28,600 | 0.1% |
| 17 | PMSI:DIAGNOSTIC-ICD10CHU\|M8144/3 | 26,100 | 0.1% |
| 18 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.06 | 23,400 | 0.1% |
| 19 | PMSI:DIAGNOSTIC-ICD10CHU\|C79.2 | 21,900 | 0.1% |
| 20 | PMSI:DIAGNOSTIC-ICD10CHU\|I50.12 | 20,100 | 0.1% |
| 21 | PMSI:DIAGNOSTIC-ICD10CHU\|Y42.0 | 20,100 | 0.1% |
| 22 | PMSI:DIAGNOSTIC-ICD10CHU\|C79.7 | 19,700 | 0.1% |
| 23 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.03 | 18,500 | 0.1% |
| 24 | PMSI:DIAGNOSTIC-ICD10CHU\|E66.01 | 17,800 | 0.1% |
| 25 | PMSI:DIAGNOSTIC-ICD10CHU\|W19.0 | 17,700 | 0.1% |

Query executed in 0.05 seconds

**Table 13.** Top 25 unmapped measurements. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ◎

| # | Source Value | #Records | %Records |
|---|---|---|---|
| 1 | DXC\|RESULTAT\|NUM:352860 | 1,923,500 | 0.3% |
| 2 | DXC\|RESULTAT\|NUM:1158430 | 1,813,200 | 0.3% |
| 3 | DXC\|RESULTAT\|NUM:1158431 | 1,810,500 | 0.3% |
| 4 | DXC\|RESULTAT\|NUM:1158433 | 1,810,500 | 0.3% |
| 5 | GLI:75213 | 1,803,500 | 0.3% |
| 6 | DXC\|RESULTAT\|NUM:1179667 | 1,277,000 | 0.2% |
| 7 | GLI:90908-90908\|90853 | 1,110,400 | 0.2% |
| 8 | DXC\|RESULTAT\|NUM:1158587 | 954,200 | 0.2% |
| 9 | DXC\|RESULTAT\|NUM:1158588 | 952,900 | 0.2% |
| 10 | GLI:75702 | 948,800 | 0.2% |
| 11 | GLI:75713 | 947,600 | 0.2% |
| 12 | DXC\|RESULTAT\|NUM:1158592 | 697,600 | 0.1% |
| 13 | GLI:64062 | 591,400 | 0.1% |
| 14 | DXC\|RESULTAT\|NUM:1158590 | 459,400 | 0.1% |
| 15 | DXC\|RESULTAT\|NUM:1184150 | 433,400 | 0.1% |
| 16 | GLI:255805-255805\|418356 | 415,500 | 0.1% |
| 17 | DXC\|RESULTAT\|NUM:1143912 | 402,300 | 0.1% |
| 18 | SYN\|ANA:375453 | 398,900 | 0.1% |
| 19 | DXC\|RESULTAT\|NUM:1144183 | 348,900 | 0.1% |
| 20 | DXC\|RESULTAT\|NUM:1144347 | 339,700 | 0.1% |
| 21 | DXC\|RESULTAT\|NUM:1146187 | 320,200 | 0.1% |
| 22 | SYN\|ANA:359007 | 314,700 | 0.1% |
| 23 | GLI:90913-90913\|90853 | 273,600 | 0.0% |
| 24 | SYN\|ANA:11752 | 272,100 | 0.0% |
| 25 | SYN\|ANA:9229 | 267,800 | 0.0% |

Query executed in 0.06 seconds

**Table 14.** Omitted because no unmapped observations were found with a count >5. ◎

Query executed in 0.06 seconds

**Table 15.** Top 25 unmapped procedures. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Source Value | #Records | %Records |
|---|---|---|---|
| 1 | PMSI:ACTE-CCAM\|YYYY600 | 913,500 | 5.1% |
| 2 | PMSI:ACTE-CCAM\|GLHF001 | 369,900 | 2.0% |
| 3 | PMSI:ACTE-CCAM\|ZZML003 | 164,700 | 0.9% |
| 4 | PMSI:ACTE-CCAM\|GLQP005 | 148,300 | 0.8% |
| 5 | PMSI:ACTE-CCAM\|GLQP002 | 137,400 | 0.8% |
| 6 | PMSI:ACTE-CCAM\|EQLF003 | 90,800 | 0.5% |
| 7 | PMSI:ACTE-CCAM\|YYYY011 | 68,300 | 0.4% |
| 8 | PMSI:ACTE-CCAM\|GLQD007 | 63,200 | 0.3% |
| 9 | PMSI:ACTE-CCAM\|GLQD001 | 60,800 | 0.3% |
| 10 | PMSI:ACTE-CCAM\|YYYY232 | 59,500 | 0.3% |
| 11 | PMSI:ACTE-CCAM\|ZZQP001 | 58,000 | 0.3% |
| 12 | PMSI:ACTE-CCAM\|HSLD002 | 54,700 | 0.3% |
| 13 | PMSI:ACTE-CCAM\|ZZQP003 | 45,700 | 0.3% |
| 14 | PMSI:ACTE-CCAM\|GLQP012 | 40,800 | 0.2% |
| 15 | PMSI:ACTE-CCAM\|ZZQX188 | 40,200 | 0.2% |
| 16 | PMSI:ACTE-CCAM\|ZZQX027 | 34,800 | 0.2% |
| 17 | PMSI:ACTE-CCAM\|ECQH011 | 34,100 | 0.2% |
| 18 | PMSI:ACTE-CCAM\|GLQP016 | 33,700 | 0.2% |
| 19 | PMSI:ACTE-CCAM\|DEMP002 | 31,200 | 0.2% |
| 20 | PMSI:ACTE-CCAM\|ZZML001 | 29,700 | 0.2% |
| 21 | PMSI:ACTE-CCAM\|YYYY041 | 24,600 | 0.1% |
| 22 | PMSI:ACTE-CCAM\|ZZQL900 | 23,200 | 0.1% |
| 23 | PMSI:ACTE-CCAM\|ZZQL018 | 22,700 | 0.1% |
| 24 | PMSI:ACTE-CCAM\|ZZQX077 | 21,500 | 0.1% |
| 25 | PMSI:ACTE-CCAM\|ZZMP017 | 21,300 | 0.1% |

Query executed in 0.06 seconds

**Table 16.** Omitted because no unmapped devices were found with a count >5. ⓪

Query executed in 0.04 seconds

**Table 17.** Omitted because no unmapped visits were found with a count >5. ⓪

Query executed in 0.04 seconds

**Table 18.** Omitted because no unmapped visit details were found with a count >5. ⓪

Query executed in 0.04 seconds

**Table 19.** Omitted because no unmapped measurement units were found with a count >5. ⓪

Query executed in 0.03 seconds

**Table 20.** Omitted because no unmapped observation units were found with a count >5. ⓪

Query executed in 0.03 seconds

**Table 21.** Top 25 unmapped measurement values. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Source Value | #Records | %Records |
|---|---|---|---|
| 1 | 90853 | 1,592,900 | 13.7% |
| 2 | 418356 | 493,200 | 4.2% |
| 3 | 6691887 | 432,400 | 3.7% |
| 4 | 3777005 | 319,100 | 2.7% |
| 5 | 6691886 | 256,100 | 2.2% |
| 6 | 7099489 | 221,100 | 1.9% |
| 7 | 6988334 | 198,200 | 1.7% |
| 8 | 6984086 | 178,500 | 1.5% |
| 9 | 1309921 | 169,200 | 1.5% |
| 10 | 418475 | 145,000 | 1.2% |
| 11 | 6984044 | 142,600 | 1.2% |
| 12 | 418438 | 132,200 | 1.1% |
| 13 | 6984054 | 117,900 | 1.0% |
| 14 | 6988336 | 111,300 | 1.0% |
| 15 | 83866 | 108,000 | 0.9% |
| 16 | 6988363 | 97,400 | 0.8% |
| 17 | 83847 | 94,600 | 0.8% |
| 18 | 673520 | 91,300 | 0.8% |
| 19 | 18205460 | 89,200 | 0.8% |
| 20 | 74409 | 88,100 | 0.8% |
| 21 | 418361 | 79,400 | 0.7% |
| 22 | 6983950 | 68,500 | 0.6% |
| 23 | 3831033 | 65,400 | 0.6% |
| 24 | 74410 | 63,000 | 0.5% |
| 25 | 83867 | 61,500 | 0.5% |

Query executed in 0.04 seconds

**Table 22.** Omitted because no unmapped observation values were found with a count >5. ⓪

Query executed in 0.02 seconds

**Table 23.** All 1 unmapped drug route. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Source Value | #Records | %Records |
|---|---|---|---|
| 1 | | 3,586,100 | 3.6% |

Query executed in 0.04 seconds

**Table 24.** Omitted because no unmapped specialty were found with a count >5. ⓪

Query executed in 0.04 seconds

## *3.4. Mapped Codes*

**Table 25.** Top 25 mapped drugs. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ◎

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 19107244 | acetaminophen 1000 MG Oral Tablet | 6,366,300 | 5.3% |
| 2 | 41083792 | Sodium 9 MG/ML Injectable Solution | 6,285,900 | 5.2% |
| 3 | 40160948 | enoxaparin sodium 100 MG/ML Injectable Solution | 2,644,500 | 2.2% |
| 4 | 36890652 | Potassium Chloride 600 MG Oral Capsule | 2,050,100 | 1.7% |
| 5 | 19076324 | glucose 50 MG/ML Injectable Solution | 1,662,200 | 1.4% |
| 6 | 40220876 | acetaminophen 10 MG/ML Injection | 1,659,300 | 1.4% |
| 7 | 19004043 | aspirin 75 MG Oral Tablet | 1,533,600 | 1.3% |
| 8 | 2911857 | insulin aspart 100 UNT/ML Pen Injector | 1,042,800 | 0.9% |
| 9 | 19044885 | zopiclone 7.5 MG Oral Tablet | 1,034,800 | 0.9% |
| 10 | 19072933 | alprazolam 0.25 MG Oral Tablet | 1,020,000 | 0.8% |
| 11 | 41205450 | lauromacrogols / Potassium / Sodium Oral Powder | 969,000 | 0.8% |
| 12 | 43195340 | POLYETHYLENE GLYCOL 3350 5900 MG Powder for Oral Solution | 951,300 | 0.8% |
| 13 | 19122503 | acetaminophen 500 MG Disintegrating Oral Tablet | 902,300 | 0.7% |
| 14 | 904458 | esomeprazole 40 MG Delayed Release Oral Tablet | 890,600 | 0.7% |
| 15 | 904457 | esomeprazole 20 MG Delayed Release Oral Tablet | 871,800 | 0.7% |
| 16 | 1125360 | acetaminophen 500 MG Oral Capsule | 848,400 | 0.7% |
| 17 | 43011843 | heparin sodium, porcine 25000 UNT/ML Injectable Solution | 826,700 | 0.7% |
| 18 | 724855 | oxazepam 10 MG Oral Tablet | 816,500 | 0.7% |
| 19 | 957136 | furosemide 40 MG Oral Tablet | 801,700 | 0.7% |
| 20 | 36886766 | Glucose 100 MG/ML / Potassium Chloride 2 MG/ML / Sodium Chloride 4 MG/ML Prefilled Syringe | 776,200 | 0.6% |
| 21 | 19104216 | nefopam 10 MG/ML Injectable Solution | 763,200 | 0.6% |
| 22 | 19106809 | oxazepam 50 MG Oral Tablet | 741,900 | 0.6% |
| 23 | 19098167 | lansoprazole 30 MG Disintegrating Oral Tablet | 688,300 | 0.6% |
| 24 | 19060972 | albuterol 2 MG/ML Inhalation Solution | 640,600 | 0.5% |
| 25 | 43146166 | Acetaminophen 325 MG / Caffeine 30 MG / Opium 10 MG Oral Capsule | 632,400 | 0.5% |

Query executed in 0.05 seconds

**Table 26.** Top 25 mapped conditions. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 320128 | Essential hypertension | 923,100 | 3.3% |
| 2 | 444094 | Finding related to pregnancy | 295,700 | 1.1% |
| 3 | 433736 | Obesity | 265,800 | 1.0% |
| 4 | 437827 | Pure hypercholesterolemia | 229,500 | 0.8% |
| 5 | 4103418 | Tobacco dependence in remission | 221,700 | 0.8% |
| 6 | 4223659 | Fatigue | 216,200 | 0.8% |
| 7 | 312437 | Dyspnea | 189,400 | 0.7% |
| 8 | 201826 | Type 2 diabetes mellitus | 186,400 | 0.7% |
| 9 | 437530 | Disorder of lipid metabolism | 181,600 | 0.7% |
| 10 | 27674 | Nausea and vomiting | 174,700 | 0.6% |
| 11 | 437264 | Tobacco dependence syndrome | 163,000 | 0.6% |
| 12 | 4171917 | Localized edema | 148,600 | 0.5% |
| 13 | 313459 | Sleep apnea | 143,600 | 0.5% |
| 14 | 4110961 | Generalized ischemic myocardial dysfunction | 143,200 | 0.5% |
| 15 | 435243 | Alcohol dependence | 141,100 | 0.5% |
| 16 | 442019 | Complication of procedure | 132,400 | 0.5% |
| 17 | 4130162 | Insulin treated type 2 diabetes mellitus | 131,200 | 0.5% |
| 18 | 439777 | Anemia | 130,200 | 0.5% |
| 19 | 140673 | Hypothyroidism | 128,600 | 0.5% |
| 20 | 434547 | Complication of surgical procedure | 128,600 | 0.5% |
| 21 | 378253 | Headache | 128,000 | 0.5% |
| 22 | 437663 | Fever | 126,300 | 0.5% |
| 23 | 75860 | Constipation | 122,700 | 0.4% |
| 24 | 317576 | Coronary arteriosclerosis | 121,700 | 0.4% |
| 25 | 4099811 | Tobacco dependence, continuous | 121,700 | 0.4% |

Query executed in 1.17 seconds

**Table 27.** Top 25 mapped measurements. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ◎

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 3015377 | Calcium [Moles/volume] in Serum or Plasma | 13,936,400 | 2.4% |
| 2 | 3000963 | Hemoglobin [Mass/volume] in Blood | 12,068,300 | 2.0% |
| 3 | 3020416 | Erythrocytes [#/volume] in Blood by Automated count | 12,013,200 | 2.0% |
| 4 | 3023599 | MCV [Entitic volume] by Automated count | 12,012,300 | 2.0% |
| 5 | 3012030 | MCH [Entitic mass] by Automated count | 12,005,600 | 2.0% |
| 6 | 3051314 | MCHC [Mass/volume] in Cord blood | 12,005,400 | 2.0% |
| 7 | 3019897 | Erythrocyte distribution width [Ratio] by Automated count | 11,926,100 | 2.0% |
| 8 | 3039827 | Platelets [#/volume] in Body fluid by Automated count | 11,805,900 | 2.0% |
| 9 | 3020779 | Urea [Moles/volume] in Serum or Plasma | 11,705,100 | 2.0% |
| 10 | 3020564 | Creatinine [Moles/volume] in Serum or Plasma | 11,630,600 | 2.0% |
| 11 | 3019550 | Sodium [Moles/volume] in Serum or Plasma | 11,565,200 | 2.0% |
| 12 | 3023103 | Potassium [Moles/volume] in Serum or Plasma | 11,465,500 | 1.9% |
| 13 | 3020630 | Protein [Mass/volume] in Serum or Plasma | 11,414,500 | 1.9% |
| 14 | 3043111 | Platelet mean volume [Entitic volume] in Blood by Automated count | 11,409,100 | 1.9% |
| 15 | 3014576 | Chloride [Moles/volume] in Serum or Plasma | 11,171,500 | 1.9% |
| 16 | 3004077 | Glucose [Mass/volume] in Capillary blood | 11,140,100 | 1.9% |
| 17 | 3019909 | Hematocrit [Volume Fraction] of Blood by Centrifugation | 9,587,600 | 1.6% |
| 18 | 3018010 | Neutrophils/100 leukocytes in Blood | 9,471,700 | 1.6% |
| 19 | 3019069 | Monocytes/100 leukocytes in Blood | 9,441,500 | 1.6% |
| 20 | 3006504 | Eosinophils/100 leukocytes in Blood | 9,441,400 | 1.6% |
| 21 | 3022096 | Basophils/100 leukocytes in Blood | 9,440,900 | 1.6% |
| 22 | 3000905 | Leukocytes [#/volume] in Blood by Automated count | 8,532,300 | 1.4% |
| 23 | 3020460 | C reactive protein [Mass/volume] in Serum or Plasma | 8,067,400 | 1.4% |
| 24 | 3012323 | Lymphocytes/100 leukocytes in Blood by Flow cytometry (FC) | 7,661,300 | 1.3% |
| 25 | 3003458 | Phosphate [Moles/volume] in Serum or Plasma | 7,590,300 | 1.3% |

Query executed in 0.12 seconds

**Table 28.** Omitted because no mapped observations were found with a count >5. ◎

Query executed in 0.02 seconds

**Table 29.** Top 25 mapped procedures. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 4163872 | Plain chest X-ray | 739,700 | 4.1% |
| 2 | 4145308 | 12 lead ECG | 659,200 | 3.6% |
| 3 | 4335825 | Transthoracic echocardiography | 608,200 | 3.4% |
| 4 | 36713734 | Doppler ultrasonography of intrathoracic vascular structure | 572,400 | 3.2% |
| 5 | 4141651 | Continuous ECG monitoring | 514,100 | 2.8% |
| 6 | 4120120 | Hemodialysis | 283,800 | 1.6% |
| 7 | 4242373 | External beam radiation therapy by linear accelerator | 265,600 | 1.5% |
| 8 | 4259495 | Slit lamp fundus examination | 246,000 | 1.4% |
| 9 | 4239130 | Oxygen therapy | 231,700 | 1.3% |
| 10 | 4144452 | Optical coherence tomography of eye region | 191,500 | 1.1% |
| 11 | 4213288 | Insertion of catheter into artery | 169,200 | 0.9% |
| 12 | 4296597 | Cardiovascular stress testing | 165,100 | 0.9% |
| 13 | 4164571 | Positive end expiratory pressure ventilation therapy, initiation and management | 162,200 | 0.9% |
| 14 | 4081708 | Interventional radiology | 155,300 | 0.9% |
| 15 | 4132859 | Immunocytochemical procedure | 142,200 | 0.8% |
| 16 | 4244986 | CT of brain without contrast | 137,900 | 0.8% |
| 17 | 4322479 | Insertion of catheter for central venous pressure monitoring | 137,200 | 0.8% |
| 18 | 4335400 | Computed tomography of abdomen and pelvis with contrast | 127,200 | 0.7% |
| 19 | 4330453 | CT, 3 dimensional reconstruction | 117,700 | 0.7% |
| 20 | 4091138 | Speech audiometry | 115,000 | 0.6% |
| 21 | 4181917 | Electroencephalogram | 114,000 | 0.6% |
| 22 | 4174669 | General anesthesia | 113,400 | 0.6% |
| 23 | 4100052 | Regional anesthesia | 113,400 | 0.6% |
| 24 | 44806445 | Cytopathology test | 112,700 | 0.6% |
| 25 | 4029715 | Radiation oncology AND/OR radiotherapy | 102,800 | 0.6% |

Query executed in 0.05 seconds

**Table 30.** Omitted because no mapped devices were found with a count >5. ⓪

Query executed in 0.03 seconds

**Table 31.** All 4 mapped visits. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 9202 | Outpatient Visit | 13,053,800 | 74.4% |
| 2 | 9201 | Inpatient Visit | 2,489,500 | 14.2% |
| 3 | 9203 | Emergency Room Visit | 1,180,300 | 6.7% |
| 4 | 262 | Emergency Room and Inpatient Visit | 815,800 | 4.7% |

Query executed in 0.03 seconds

**Table 32.** All 5 mapped visit details. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 9202 | Outpatient Visit | 12,740,600 | 62.9% |
| 2 | 9201 | Inpatient Visit | 4,024,200 | 19.9% |
| 3 | 262 | Emergency Room and Inpatient Visit | 1,902,300 | 9.4% |
| 4 | 9203 | Emergency Room Visit | 1,141,900 | 5.6% |
| 5 | 32037 | Intensive Care | 441,300 | 2.2% |

Query executed in 0.05 seconds

**Table 33.** Top 25 mapped measurement units. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 8554 | percent | 107,771,000 | 42.8% |
| 2 | 8753 | millimole per liter | 62,564,900 | 24.8% |
| 3 | 8713 | gram per deciliter | 12,470,300 | 4.9% |
| 4 | 8564 | picogram | 12,261,900 | 4.9% |
| 5 | 8636 | gram per liter | 11,804,200 | 4.7% |
| 6 | 8583 | femtoliter | 9,705,700 | 3.9% |
| 7 | 44777602 | kilopascal | 9,220,000 | 3.7% |
| 8 | 8840 | milligram per deciliter | 4,817,300 | 1.9% |

| | | | | |
|---|---|---|---|---|
| 9 | 8795 | milliliter per minute | 4,795,900 | 1.9% |
| 10 | 8555 | second | 4,174,500 | 1.7% |
| 11 | 8751 | milligram per liter | 3,788,400 | 1.5% |
| 12 | 8842 | nanogram per milliliter | 1,853,100 | 0.7% |
| 13 | 9570 | milliliter per deciliter | 1,043,600 | 0.4% |
| 14 | 8845 | picogram per milliliter | 1,026,200 | 0.4% |
| 15 | 8725 | nanogram per liter | 879,300 | 0.3% |
| 16 | 8588 | millimeter | 674,500 | 0.3% |
| 17 | 44777612 | milligram per millimole | 653,700 | 0.3% |
| 18 | 8587 | milliliter | 621,300 | 0.2% |
| 19 | 8698 | liter per minute | 388,000 | 0.2% |
| 20 | 8729 | picomole per liter | 378,100 | 0.2% |
| 21 | 8736 | nanomole per liter | 256,100 | 0.1% |
| 22 | 9436 | thousand per milliliter | 222,900 | 0.1% |
| 23 | 8505 | hour | 151,500 | 0.1% |
| 24 | 8550 | minute | 143,500 | 0.1% |
| 25 | 8518 | Henry | 120,500 | 0.0% |

Query executed in 0.03 seconds

**Table 34.** Omitted because no mapped observation units were found with a count >5. ◎

Query executed in 0.03 seconds

**Table 35.** All 6 mapped measurement values. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ◎

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 45878583 | Negative | 2,582,500 | 22.2% |
| 2 | 45884084 | Positive | 446,100 | 3.8% |
| 3 | 45878239 | Few | 39,700 | 0.3% |
| 4 | 45880448 | Colorless | 24,100 | 0.2% |
| 5 | 45881248 | Schistocytes | 17,800 | 0.2% |
| 6 | 36308775 | Bloody | 6,400 | 0.1% |

Query executed in 0.04 seconds

**Table 36.** Omitted because no mapped observation values were found with a count >5. ◎

Query executed in 0.03 seconds

**Table 37.** Top 25 mapped drug route. Counts are rounded up to the nearest hundred. Values with a record count <=5 are omitted. ⓪

| # | Concept id | Concept Name | #Records | %Records |
|---|---|---|---|---|
| 1 | 4132161 | Oral route | 75,229,300 | 75.6% |
| 2 | 4171047 | Intravenous route | 7,780,000 | 7.8% |
| 3 | 4142048 | Subcutaneous route | 6,241,700 | 6.3% |
| 4 | 40486444 | Conjunctival route | 2,977,200 | 3.0% |
| 5 | 40490507 | Cutaneous route | 1,048,600 | 1.1% |
| 6 | 4006860 | Intra-articular route | 1,036,700 | 1.0% |
| 7 | 4290759 | Rectal route | 389,500 | 0.4% |
| 8 | 4262099 | Transdermal route | 247,100 | 0.2% |
| 9 | 40486069 | Respiratory tract route | 189,400 | 0.2% |
| 10 | 4292110 | Sublingual route | 189,400 | 0.2% |
| 11 | 4302612 | Intramuscular route | 169,600 | 0.2% |
| 12 | 4262914 | Nasal route | 130,300 | 0.1% |
| 13 | 4156705 | Intracardiac route | 73,400 | 0.1% |
| 14 | 4023156 | Otic route | 48,600 | 0.0% |
| 15 | 4156704 | Gingival route | 35,600 | 0.0% |
| 16 | 4132254 | Gastrostomy route | 27,500 | 0.0% |
| 17 | 4132711 | Nasogastric route | 21,300 | 0.0% |
| 18 | 4057765 | Vaginal route | 12,400 | 0.0% |
| 19 | 37018288 | Extracorporeal route | 10,700 | 0.0% |
| 20 | 4157761 | Perineural route | 8,000 | 0.0% |
| 21 | 4186832 | Endotracheopulmonary route | 7,600 | 0.0% |
| 22 | 40490896 | Peridural route | 6,000 | 0.0% |
| 23 | 4243022 | Intraperitoneal route | 5,900 | 0.0% |
| 24 | 4156706 | Intradermal route | 4,000 | 0.0% |
| 25 | 4302788 | Intraspinal route | 1,700 | 0.0% |

Query executed in 0.03 seconds

**Table 38.** Omitted because no mapped specialty were found with a count >5. ⓪

Query executed in 0.03 seconds

## *3.5. Source to concept map*

**Table 39.** Source to concept map breakdown ⓪

| SOURCE_VOCABULARY_ID | TARGET_VOCABULARY_ID | COUNT |
|---|---|---|
| i2b2 | Cancer Modifier | 2,566 |
| i2b2 | CPT4 | 10 |
| i2b2 | Gender | 2 |
| i2b2 | HCPCS | 10 |
| i2b2 | ICD10PCS | 1 |
| i2b2 | ICD9Proc | 2 |
| i2b2 | ICDO3 | 139,596 |
| i2b2 | LOINC | 5,777 |
| i2b2 | None | 7 |
| i2b2 | OMOP Extension | 4 |
| i2b2 | RxNorm | 34,502 |
| i2b2 | RxNorm Extension | 5,241 |
| i2b2 | SNOMED | 32,959 |
| i2b2 | UK Biobank | 1 |

Query executed in 0.30 seconds

# 4. Data Quality Dashboard

DataQualityDashboard executed at 2024-10-01 13:30:47 in 13 hours.

**Table 40.** Number of passed, failed and total DQD checks per category. For DQD v2, the checks with status 'NA' are not included.

| Category | Pass | Fail | Total | %Pass |
|---|---|---|---|---|
| **Plausibility** | 368 | 36 | 724 | 50.8% |
| **Conformance** | 748 | 7 | 823 | 90.9% |
| **Completeness** | 256 | 9 | 293 | 87.4% |
| **Total** | **1,788** | **52** | **1,840** | **97.2%** |

# 5. Drug Exposure Diagnostics

**Table 41.** Drug Exposure Diagnostics results for selected ingredients. Executed with minCellCount = 5, sample = 1e+06, earliestStartDate = 2010-01-01. # = Number of records. Type (n,%) = Frequency and percentage of available drug types. Route (n,%) = Frequency and percentage of available routes. Dose Form present n (%) = Frequency and percentage with dose form present. Fixed amount dose form n (%) = Frequency and percentage of missing denominator unit concept id. Amount distrib. [null or missing] = Distribution of amount (median q05-q95), frequency and percentage of null or missing amount. Quantity distrib. [null or missing] = Distribution of quantity (median q05-q95), frequency and percentage of null or missing quantity. Exposure days distrib. [null or missing] = Distribution of exposure days (median q05-q95), frequency and percentage of null days_supply or missing exposure dates. Neg. Days n (%) = Frequency and percentage of negative exposure days. ◎

| Ingredient | Concept ID | # | Type (n,%) | Route (n,%) | Dose Form present n (%) | Fixed amount dose form n (%) | Amount distrib. [null or missing] | Quantity distrib. [null or missing] | Exposure days distrib. [null or missing] | Neg. Days n (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| hepatitis B surface antigen vaccine | 528323 | 4035 | EHR administration record (4035, 100%) | NA (25, 0.6%); Intramuscular route (2113, 52.4%); Subcutaneous route (1897, 47%) | 0 (0%) | 4035 (100%) | NA (NA-NA) [4035, 100%] | NA (NA-NA) [4035, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |
| latanoprost | 954688 | 140313 | EHR prescription (5, 0%); EHR administration record (140308, 100%) | NA (5416, 3.9%); Conjunctival route (134897, 96.1%) | 0 (0%) | 220 (0.2%) | NA (NA-NA) [140313, 100%] | NA (NA-NA) [140313, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |
| mesalamine | 968426 | 17348 | EHR prescription (29, 0.2%); EHR administration record (17319, 99.8%) | NA (396, 2.3%); Rectal route (1941, 11.2%); Oral route (15011, 86.5%) | 0 (0%) | 17348 (100%) | 800 (500-2000) [1216, 7%] | NA (NA-NA) [17348, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |
| adalimumab | 1119119 | 711 | EHR administration record (711, 100%) | NA (114, 16%); No matching concept (1, 0.1%); Subcutaneous route (596, 83.8%) | 0 (0%) | 128 (18%) | NA (NA-NA) [711, 100%] | NA (NA-NA) [711, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| acetaminophen | 1125315 | 11527892 | EHR administration record (999999, 100%); EHR prescription (1, 0%) | Rectal route (4298, 0.4%); No matching concept (566, 0.1%); Intravenous route (52422, 5.2%); Oral route (809830, 81%); NA (132884, 13.3%) | 0 (0%) | 985638 (85.6%) | 1000 (10-1000) [166167, 14.4%] | 1 (1-1.75) [1151365, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |
| acetylcysteine | 1139042 | 62341 | EHR administration record (62341, 100%) | NA (396, 0.6%); Nasal route (24278, 38.9%); Oral route (37667, 60.4%) | 0 (0%) | 37832 (34%) | 200 (200-200) [73527, 66%] | NA (NA-NA) [111359, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |
| sumatriptan | 1140643 | 933 | EHR administration record (933, 100%) | Nasal route (419, 44.9%); Oral route (10, 1.1%); Subcutaneous route (502, 53.8%); NA (2, 0.2%) | 0 (0%) | 40 (4.3%) | 50 (50-50) [923, 98.9%] | NA (NA-NA) [933, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |
| albuterol | 1154343 | 1027876 | EHR prescription (269, 0%); EHR administration record (999731, 100%) | Oral route (25980, 2.6%); Rectal route (94, 0%); Subcutaneous route (264, 0%); Intravenous route (396, 0%); Intramuscular route (76, 0%);NA (820560, 82.1%); Intra-articular route (152630, 15.3%) | 0 (0%) | 190011 (19%) | 0.100000001490116 (0.100000001490116-0.100000001490116) [811471, 81.1%] | NA (NA-NA) [1000000, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |
| prednisolone | 1550557 | 367081 | EHR prescription (444, 0.1%); EHR administration | Intra-articular route (64, 0%); Rectal route (264, 0.1%); NA (1393, | 0 (0%) | 365684 (99.4%) | 20 (5-20) [10288, 2.8%] | NA (NA-NA) [367850, 100%] | 1 (1-1) [NA, NA%] | 0 (0%) |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | record (366637, 99.9%) | 0.4%); Periarticular route (28, 0%); Peridural route (106, 0%); Nasal route (505, 0.1%); Oral route (356258, 97.1%); No matching concept (8463, 2.3%) | | | | | | |
| acyclovir | 1703687 | 69594 | EHR administration record (69594, 100%) | No matching concept (38439, 55.2%); Conjunctival route (888, 1.3%); NA (9021, 13%); Intravenous route (678, 1%); Oral route (7482, 10.8%); Cutaneous route (13086, 18.8%) | 0 (0%) | 38644 (55.5%) | 500 (500-500) [30950, 44.5%] | 1 (1-1) [69593, 100%] | 1 (1-2) [NA, NA%] | 0 (0%) |
| ulipristal | 40225722 | 311 | EHR administration record (311, 100%) | Oral route (311, 100%) | 0 (0%) | 311 (100%) | 30 (5-30) [0, 0%] | NA (NA-NA) [311, 100%] | 1 (1-1) [0, 0%] | 0 (0%) |

Query executed in 892.30 seconds

# 6. Technical Infrastructure

## *6.1. R packages*

**Table 42.** Versions of all installed R packages from DARWIN EU® and the OHDSI Health Analytics Data-to-Evidence Suite (HADES). Packages can be installed from CRAN (install.packages("<package_name>")) or Github (remotes::install_github("<organisation>/<package>")) Ⓢ

| Organisation | Package | Version |
|---|---|---|
| DARWIN EU® | CDMConnector | 1.5.0 |
| DARWIN EU® | CdmOnboarding | 3.1.0 |
| DARWIN EU® | CohortSurvival | 0.5.2 |
| DARWIN EU® | DashboardExport | 1.2.0 |
| DARWIN EU® | DrugExposureDiagnostics | 1.0.1 |
| DARWIN EU® | PatientProfiles | 1.1.1 |
| DARWIN EU® | TreatmentPatterns | 2.6.9 |
| DARWIN EU® | CodelistGenerator | Not installed |
| DARWIN EU® | DrugUtilisation | Not installed |
| DARWIN EU® | IncidencePrevalence | Not installed |
| DARWIN EU® | OMOPGenerics | Not installed |
| DARWIN EU® | PaRe | Not installed |
| DARWIN EU® | ReportGenerator | Not installed |
| DARWIN EU® | deckR | Not installed |
| OHDSI HADES | Achilles | 1.7 |
| OHDSI HADES | Andromeda | 0.6.7 |
| OHDSI HADES | CirceR | 1.3.2 |
| OHDSI HADES | DataQualityDashboard | 2.6.0 |
| OHDSI HADES | DatabaseConnector | 6.2.4 |
| OHDSI HADES | ParallelLogger | 3.3.0 |
| OHDSI HADES | ROhdsiWebApi | 1.3.3 |
| OHDSI HADES | ResultModelManager | 0.5.9 |
| OHDSI HADES | SqlRender | 1.16.1 |
| OHDSI HADES | BigKnn | Not installed |
| OHDSI HADES | BrokenAdaptiveRidge | Not installed |
| OHDSI HADES | Capr | Not installed |
| OHDSI HADES | Characterization | Not installed |
| OHDSI HADES | CohortDiagnostics | Not installed |
| OHDSI HADES | CohortExplorer | Not installed |
| OHDSI HADES | CohortGenerator | Not installed |
| OHDSI HADES | CohortMethod | Not installed |
| OHDSI HADES | Cyclops | Not installed |
| OHDSI HADES | DeepPatientLevelPrediction | Not installed |
| OHDSI HADES | EmpiricalCalibration | Not installed |
| OHDSI HADES | EnsemblePatientLevelPrediction | Not installed |
| OHDSI HADES | Eunomia | Not installed |
| OHDSI HADES | EvidenceSynthesis | Not installed |
| OHDSI HADES | FeatureExtraction | Not installed |
| OHDSI HADES | Hydra | Not installed |
| OHDSI HADES | IterativeHardThresholding | Not installed |
| OHDSI HADES | MethodEvaluation | Not installed |
| OHDSI HADES | OhdsiSharing | Not installed |
| OHDSI HADES | OhdsiShinyModules | Not installed |
| OHDSI HADES | PatientLevelPrediction | Not installed |
| OHDSI HADES | PheValuator | Not installed |
| OHDSI HADES | PhenotypeLibrary | Not installed |
| OHDSI HADES | SelfControlledCaseSeries | Not installed |
| OHDSI HADES | SelfControlledCohort | Not installed |
| OHDSI HADES | ShinyAppBuilder | Not installed |

## 6.2. System Information

Installed R version: R version 4.2.3 (2023-03-15)
System CPU vendor: GenuineIntel
System CPU model: Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz
System CPU number of cores: 4
System RAM: 16.82 GB
DBMS: PostgreSQL 15.4 (Debian 15.4-2.pgdg120+1) on x86_64-pc-linux-gnu, compiled by gcc (Debian 12.2.0-14) 12.2.0, 64-bit
WebAPI version:

## 6.3. Vocabulary Query Performance

The number of 'Maps To' relations is equal to 6,675,774 and queried in 33.34 seconds (4.99481e-06 s/#).

## 6.4. Applied indexes

**Table 43.** The indexes applied on the OMOP CDM tables Ⓢ

| TABLENAME | INDEXNAMES |
|---|---|
| care_site | idx_care_site_id_1,xpk_care_site |
| condition_era | idx_condition_era_concept_id_1,idx_condition_era_person_id_1,xpk_condition_era |
| condition_occurrence | idx_condition_concept_id_1,idx_condition_person_id_1,idx_condition_visit_id_1,xpk_condition_occurrence |
| cost | idx_cost_event_id,xpk_cost |
| death | idx_death_person_id_1 |
| device_exposure | idx_device_concept_id_1,idx_device_person_id_1,idx_device_visit_id_1,xpk_device_exposure |
| dose_era | idx_dose_era_concept_id_1,idx_dose_era_person_id_1,xpk_dose_era |
| drug_era | idx_drug_era_concept_id_1,idx_drug_era_person_id_1,xpk_drug_era |
| drug_exposure | idx_drug_concept_id_1,idx_drug_person_id_1,idx_drug_visit_id_1,xpk_drug_exposure |
| episode | xpk_episode |
| fact_relationship | idx_fact_relationship_id1,idx_fact_relationship_id2,idx_fact_relationship_id3 |
| location | idx_location_id_1,xpk_location |
| measurement | idx_measurement_concept_id_1,idx_measurement_person_id_1,idx_measurement_visit_id_1,xpk_measurement |
| metadata | idx_metadata_concept_id_1,xpk_metadata |
| note | idx_note_concept_id_1,idx_note_person_id_1,idx_note_visit_id_1,xpk_note |
| note_nlp | idx_note_nlp_concept_id_1,idx_note_nlp_note_id_1,xpk_note_nlp |
| observation | idx_observation_concept_id_1,idx_observation_person_id_1,idx_observation_visit_id_1,xpk_observation |
| observation_period | idx_observation_period_id_1,xpk_observation_period |
| payer_plan_period | idx_period_person_id_1,xpk_payer_plan_period |
| person | idx_gender,idx_person_id,xpk_person |
| procedure_occurrence | idx_procedure_concept_id_1,idx_procedure_person_id_1,idx_procedure_visit_id_1,xpk_procedure_occurrence |
| provider | idx_provider_id_1,xpk_provider |
| specimen | idx_specimen_concept_id_1,idx_specimen_person_id_1,xpk_specimen |
| visit_detail | idx_visit_det_concept_id_1,idx_visit_det_occ_id,idx_visit_det_person_id_1,xpk_visit_detail |
| visit_occurrence | idx_visit_concept_id_1,idx_visit_person_id_1,xpk_visit_occurrence |

Query executed in 0.13 seconds
The following expected indexes are missing:

xpk_concept, xpk_vocabulary, xpk_domain, xpk_concept_class, xpk_relationship, idx_concept_concept_id, idx_concept_code, idx_concept_vocabluary_id, idx_concept_domain_id, idx_concept_class_id, idx_vocabulary_vocabulary_id, idx_domain_domain_id, idx_concept_class_class_id, idx_concept_relationship_id_1, idx_concept_relationship_id_2, idx_concept_relationship_id_3, idx_relationship_rel_id, idx_concept_synonym_id, idx_concept_ancestor_id_1, idx_concept_ancestor_id_2, idx_source_to_concept_map_3, idx_source_to_concept_map_1, idx_source_to_concept_map_2, idx_source_to_concept_map_c, idx_drug_strength_id_1, idx_drug_strength_id_2, idx_episode_person_id_1, idx_episode_concept_id_1, idx_episode_event_id_1, idx_ee_field_concept_id_1

## *6.5. Achilles Query Performance*

**Table 44.** Execution time of Achilles analyses. Total: 6h 13m 54s. Median: 2.5s. Longest duration: 1h 5m 24s (analysis 424). Ⓐ

| ID | NAME | DURATION |
|---|---|---|
| 0 | Source name | 16.75 |
| 1 | Number of persons | 0.83 |
| 2 | Number of persons by gender | 66.6 |
| 3 | Number of persons by year of birth | 2.06 |
| 4 | Number of persons by race | 1.74 |
| 5 | Number of persons by ethnicity | 1.76 |
| 7 | Number of persons with invalid provider_id | 0.15 |
| 8 | Number of persons with invalid location_id | 0.14 |
| 9 | Number of persons with invalid care_site_id | 0.14 |
| 10 | Number of all persons by year of birth by gender | 1.68 |
| 11 | Number of non-deceased persons by year of birth by gender | 1.71 |
| 12 | Number of persons by race and ethnicity | 1.91 |
| 101 | Number of persons by age, with age at first observation period | 30.09 |
| 102 | Number of persons by gender by age, with age at first observation period | 1.92 |
| 108 | Number of persons by length of observation period, in 30d increments | 6.46 |
| 109 | Number of persons with continuous observation in each year | 55.28 |
| 110 | Number of persons with continuous observation in each month | 78.6 |
| 111 | Number of persons by observation period start month | 1.58 |
| 112 | Number of persons by observation period end month | 1.61 |
| 113 | Number of persons by number of observation periods | 2.62 |
| 114 | Number of persons with observation period before year-of-birth | 2.88 |
| 115 | Number of persons with observation period end < observation period start | 0.12 |
| 116 | Number of persons with at least one day of observation in each year by gender and age decile | 45.01 |
| 117 | Number of persons with at least one day of observation in each month | 3636 |
| 118 | Number of observation periods with invalid person_id | 0.52 |
| 119 | Number of observation period records by period_type_concept_id | 0.16 |
| 200 | Number of persons with at least one visit occurrence, by visit_concept_id | 29.53 |
| 201 | Number of visit occurrence records, by visit_concept_id | 1.96 |
| 202 | Number of persons by visit occurrence start month, by visit_concept_id | 14.22 |
| 204 | Number of persons with at least one visit occurrence, by visit_concept_id by calendar year by gender by age decile | 20.1 |
| 207 | Number of visit records with invalid person_id | 17.19 |
| 209 | Number of visit records with invalid care_site_id | 0.78 |
| 210 | Number of visit_occurrence records outside a valid observation period | 1.57 |
| 211 | Number of visit records with end date < start date | 0.51 |
| 212 | Number of persons with at least one visit occurrence, by calendar year by gender by age decile | 18.09 |
| 220 | Number of visit occurrence records by visit occurrence start month | 3.27 |
| 221 | Number of persons by visit start year | 12.55 |
| 225 | Number of visit_occurrence records by visit_source_concept_id | 1.95 |
| 226 | Number of records by domain by visit_concept_id | 390 |
| 230 | Number of visit_occurrence records inside valid observation period | 1.72 |
| 231 | Proportion of people with at least one visit_occurrence record outside a valid observation period | 13.92 |

| | | |
|---|---|---|
| 232 | Proportion of visit_occurrence records outside a valid observation period | 6.19 |
| 300 | Number of providers | 0.02 |
| 301 | Number of providers by specialty concept_id | 0.02 |
| 303 | Number of providers records by specialty_concept_id and visit_concept_id | 1.09 |
| 325 | Number of provider records by specialty_source_concept_id | 0.02 |
| 400 | Number of persons with at least one condition occurrence, by condition_concept_id | 23.12 |
| 401 | Number of condition occurrence records, by condition_concept_id | 2.73 |
| 402 | Number of persons by condition occurrence start month, by condition_concept_id | 22.51 |
| 404 | Number of persons with at least one condition occurrence, by condition_concept_id by calendar year by gender by age decile | 32.07 |
| 405 | Number of condition occurrence records, by condition_concept_id by condition_type_concept_id | 2.91 |
| 409 | Number of condition occurrence records with invalid person_id | 12.44 |
| 410 | Number of condition occurrence records outside valid observation period | 2.41 |
| 411 | Number of condition occurrence records with end date < start date | 0.61 |
| 412 | Number of condition occurrence records with invalid provider_id | 0.93 |
| 413 | Number of condition occurrence records with invalid visit_id | 15.7 |
| 414 | Number of condition occurrence records, by condition_status_concept_id | 2.55 |
| 415 | Number of condition occurrence records, by condition_type_concept_id | 2.57 |
| 416 | Number of condition occurrence records, by condition_status_concept_id, condition_type_concept_id | 2.82 |
| 420 | Number of condition occurrence records by condition occurrence start month | 4.54 |
| 424 | Number of distinct people with co-occurring condition_occurrence condition_concept_id pairs | 3924 |
| 425 | Number of condition_occurrence records by condition_source_concept_id | 2.73 |
| 430 | Number of condition_occurrence records inside a valid observation period | 2.33 |
| 431 | Proportion of people with at least one condition_occurrence record outside a valid observation period | 18.27 |
| 432 | Proportion of condition_occurrence records outside a valid observation period | 2.62 |
| 500 | Number of persons with death, by cause_concept_id | 0.39 |
| 501 | Number of records of death, by cause_concept_id | 0.27 |
| 502 | Number of persons by death month | 0.34 |
| 504 | Number of persons with a death, by calendar year by gender by age decile | 0.77 |
| 505 | Number of death records, by death_type_concept_id | 0.3 |
| 509 | Number of death records with invalid person_id | 0.24 |
| 510 | Number of death records outside valid observation period | 0.48 |
| 525 | Number of death records by cause_source_concept_id | 0.05 |
| 530 | Number of death records inside a valid observation period | 0.19 |
| 531 | Proportion of people with at least one death record outside a valid observation period | 0.52 |
| 532 | Proportion of death records that occur outside a valid observation period | 0.49 |
| 600 | Number of persons with at least one procedure occurrence, by procedure_concept_id | 9 |
| 601 | Number of procedure occurrence records, by procedure_concept_id | 1.91 |
| 602 | Number of persons by procedure occurrence start month, by procedure_concept_id | 13.91 |
| 604 | Number of persons with at least one procedure occurrence, by procedure_concept_id by calendar year by gender by age decile | 20.08 |
| 605 | Number of procedure occurrence records, by procedure_concept_id by procedure_type_concept_id | 2.11 |
| 609 | Number of procedure occurrence records with invalid person_id | 5.82 |
| 610 | Number of procedure occurrence records outside valid observation period | 1.48 |
| 612 | Number of procedure occurrence records with invalid provider_id | 0.63 |
| 613 | Number of procedure occurrence records with invalid visit_id | 24.47 |
| 620 | Number of procedure occurrence records  by procedure occurrence start month | 3.19 |
| 624 | Number of distinct people with co-occurring procedure_occurrence procedure_concept_id pairs | 449.4 |
| 625 | Number of procedure_occurrence records by procedure_source_concept_id | 2.02 |

| | | |
|---|---|---|
| 630 | Number of procedure_occurrence records inside a valid observation period | 1.67 |
| 631 | Proportion of people with at least one procedure_occurrence record outside a valid observation period | 12.77 |
| 632 | Proportion of procedure_occurrence records outside a valid observation period | 6.67 |
| 691 | Percentage of total persons that have at least x procedures | 7.23 |
| 700 | Number of persons with at least one drug exposure, by drug_concept_id | 99.6 |
| 701 | Number of drug exposure records, by drug_concept_id | 52.59 |
| 702 | Number of persons by drug exposure start month, by drug_concept_id | 130.8 |
| 704 | Number of persons with at least one drug exposure, by drug_concept_id by calendar year by gender by age decile | 133.2 |
| 705 | Number of drug exposure records, by drug_concept_id by drug_type_concept_id | 30.93 |
| 709 | Number of drug exposure records with invalid person_id | 64.2 |
| 710 | Number of drug exposure records outside valid observation period | 20.03 |
| 711 | Number of drug exposure records with end date < start date | 8.51 |
| 712 | Number of drug exposure records with invalid provider_id | 3.87 |
| 713 | Number of drug exposure records with invalid visit_id | 47.8 |
| 720 | Number of drug exposure records  by drug exposure start month | 18.84 |
| 725 | Number of drug_exposure records by drug_source_concept_id | 12.84 |
| 730 | Number of drug_exposure records inside a valid observation period | 8.66 |
| 731 | Proportion of people with at least one drug_exposure record outside a valid observation period | 106.8 |
| 732 | Proportion of drug_exposure records outside a valid observation period | 153 |
| 791 | Percentage of total persons that have at least x drug exposures | 52.36 |
| 800 | Number of persons with at least one observation occurrence, by observation_concept_id | 0.03 |
| 801 | Number of observation occurrence records, by observation_concept_id | 0.01 |
| 802 | Number of persons by observation occurrence start month, by observation_concept_id | 0.02 |
| 804 | Number of persons with at least one observation occurrence, by observation_concept_id by calendar year by gender by age decile | 0.02 |
| 805 | Number of observation occurrence records, by observation_concept_id by observation_type_concept_id | 0.02 |
| 807 | Number of observation occurrence records, by observation_concept_id and unit_concept_id | 0.01 |
| 809 | Number of observation records with invalid person_id | 0.02 |
| 810 | Number of observation records outside valid observation period | 0.02 |
| 812 | Number of observation records with invalid provider_id | 0.02 |
| 813 | Number of observation records with invalid visit_id | 0.02 |
| 814 | Number of observation records with no value (numeric, string, or concept) | 0.02 |
| 820 | Number of observation records  by observation start month | 0.02 |
| 822 | Number of observation records, by observation_concept_id and value_as_concept_id | 0.02 |
| 823 | Number of observation records, by observation_concept_id and qualifier_concept_id | 0.01 |
| 824 | Number of distinct people with co-occurring observation observation_concept_id pairs | 0.01 |
| 825 | Number of observation records by observation_source_concept_id | 0.01 |
| 826 | Number of observation records by value_as_concept_id | 0.02 |
| 827 | Number of observation records by unit_concept_id | 0.02 |
| 830 | Number of observation records inside a valid observation period | 0.02 |
| 831 | Proportion of people with at least one observation record outside a valid observation period | 0.02 |
| 832 | Proportion of observation records outside a valid observation period | 0.02 |
| 891 | Percentage of total persons that have at least x observations | 0.02 |
| 900 | Number of persons with at least one drug era, by drug_concept_id | 8.44 |
| 901 | Number of drug era records, by drug_concept_id | 1.44 |
| 902 | Number of persons by drug era start month, by drug_concept_id | 10.82 |
| 904 | Number of persons with at least one drug era, by drug_concept_id by calendar year by gender by age decile | 17.51 |
| 908 | Number of drug eras without valid person | 5.93 |

| | | |
|---|---|---|
| 910 | Number of drug_era records outside valid observation period | 1.08 |
| 911 | Number of drug eras with end date < start date | 0.23 |
| 920 | Number of drug era records by drug era start month | 2.5 |
| 930 | Number of drug_era records inside a valid observation period | 1.18 |
| 931 | Proportion of people with at least one drug_era record outside a valid observation period | 7.6 |
| 932 | Proportion of drug_era records outside a valid observation period | 5.2 |
| 1,000 | Number of persons with at least one condition era, by condition_concept_id | 8.85 |
| 1,001 | Number of condition era records, by condition_concept_id | 1.6 |
| 1,002 | Number of persons by condition era start month, by condition_concept_id | 13.17 |
| 1,004 | Number of persons with at least one condition era, by condition_concept_id by calendar year by gender by age decile | 18.81 |
| 1,008 | Number of condition eras without valid person | 1.07 |
| 1,010 | Number of condition_era records outside a valid observation period | 1.18 |
| 1,011 | Number of condition eras with end date < start date | 0.26 |
| 1,020 | Number of condition era records by condition era start month | 2.64 |
| 1,030 | Number of condition_era records inside a valid observation period | 1.28 |
| 1,031 | Proportion of people with at least one condition_era record outside a valid observation period | 10.39 |
| 1,032 | Proportion of condition_era records outside a valid observation period | 1.38 |
| 1,100 | Number of persons by location 3-digit zip | 0.13 |
| 1,101 | Number of persons by location state | 0.13 |
| 1,102 | Number of care sites by location 3-digit zip | 0.02 |
| 1,103 | Number of care sites by location state | 0.01 |
| 1,200 | Number of persons by place of service | 0.12 |
| 1,201 | Number of visits by place of service | 7.09 |
| 1,202 | Number of care sites by place of service | 0.02 |
| 1,203 | Number of visits by place of service discharge type | 1.1 |
| 1,300 | Number of persons with at least one visit detail, by visit_detail_concept_id | 16.21 |
| 1,301 | Number of visit detail records, by visit_detail_concept_id | 2.15 |
| 1,302 | Number of persons by visit detail start month, by visit_detail_concept_id | 16.36 |
| 1,304 | Number of persons with at least one visit detail, by visit_detail_concept_id by calendar year by gender by age decile | 24.13 |
| 1,307 | Number of visit records with invalid person_id | 11.05 |
| 1,309 | Number of visit_detail records with invalid care_site_id | 0.67 |
| 1,310 | Number of visit_detail records outside a valid observation period | 1.64 |
| 1,311 | Number of visit_detail records with end date < start date | 0.39 |
| 1,312 | Number of persons with at least one visit detail, by calendar year by gender by age decile | 22.24 |
| 1,320 | Number of visit detail records by visit detail start month | 3.62 |
| 1,321 | Number of persons by visit start year | 14.93 |
| 1,325 | Number of visit_detail records by visit_detail_source_concept_id | 2.2 |
| 1,326 | Number of records by domain by visit_detail_concept_id | 381 |
| 1,330 | Number of visit_detail records inside a valid observation period | 2.33 |
| 1,331 | Proportion of people with at least one visit_detail record outside a valid observation period | 25.67 |
| 1,332 | Proportion of visit_detail records outside a valid observation period | 3.88 |
| 1,408 | Number of persons by length of payer plan period, in 30d increments | 0.02 |
| 1,409 | Number of persons with continuous payer plan in each year | 0.04 |
| 1,410 | Number of persons with continuous payer plan in each month | 0.04 |
| 1,411 | Number of persons by payer plan period start month | 0.01 |
| 1,412 | Number of persons by payer plan period end month | 0.02 |
| 1,413 | Number of persons by number of payer plan periods | 0.02 |
| 1,414 | Number of persons with payer plan period before year-of-birth | 0.02 |
| 1,415 | Number of persons with payer plan period end < payer plan period start | 0.04 |
| 1,425 | Number of payer_plan_period records by payer_source_concept_id | 0.02 |
| 1,610 | Number of records by revenue_code_concept_id | 0.02 |
| 1,800 | Number of persons with at least one measurement occurrence, by measurement_concept_id | 590.4 |
| 1,801 | Number of measurement occurrence records, by measurement_concept_id | 211.8 |

| | | |
|---|---|---|
| 1,802 | Number of persons by measurement occurrence start month, by measurement_concept_id | 795 |
| 1,804 | Number of persons with at least one mesurement occurrence, by measurement_concept_id by calendar year by gender by age decile | 963.6 |
| 1,805 | Number of measurement occurrence records, by measurement_concept_id by measurement_type_concept_id | 210.6 |
| 1,807 | Number of measurement occurrence records, by measurement_concept_id and unit_concept_id | 423.6 |
| 1,809 | Number of measurement records with invalid person_id | 356.4 |
| 1,810 | Number of measurement records outside valid observation period | 442.8 |
| 1,811 | Number of measurement records with a valid value (with a mapped, non-null value_as_number) | 217.8 |
| 1,812 | Number of measurement records with invalid provider_id | 171 |
| 1,813 | Number of measurement records with invalid visit_id | 201.6 |
| 1,814 | Number of measurement records with no value (numeric, string, or concept) | 184.2 |
| 1,818 | Number of measurement records below/within/above normal range, by measurement_concept_id and unit_concept_id | 205.2 |
| 1,819 | Number of measurement records, by measurement_concept_id, with a valid value (with a mapped, non-null value_as_number) | 202.8 |
| 1,820 | Number of measurement records by measurement start month | 234.6 |
| 1,821 | Number of measurement records with no numeric value | 162.6 |
| 1,822 | Number of measurement records, by measurement_concept_id and value_as_concept_id | 202.2 |
| 1,823 | Number of measurement records, by measurement_concept_id and operator_concept_id | 205.2 |
| 1,825 | Number of measurement records by measurement_source_concept_id | 201.6 |
| 1,826 | Number of measurement records by value_as_concept_id | 196.2 |
| 1,827 | Number of measurement records by unit_concept_id | 199.2 |
| 1,830 | Number of visit_detail records inside a valid observation period | 190.8 |
| 1,831 | Proportion of people with at least one measurement record outside a valid observation period | 1012.2 |
| 1,832 | Proportion of measurement records outside a valid observation period | 354.6 |
| 1,833 | Proportion of measurement records inside a valid observation period and without a value | 211.2 |
| 1,891 | Percentage of total persons that have at least x measurements | 456 |
| 1,900 | Source values mapped to concept_id 0 by table, by column, by source_value | 442.2 |
| 2,000 | Number of patients with at least 1 Dx and 1 Rx | 418.2 |
| 2,001 | Number of patients with at least 1 Dx and 1 Proc | 846 |
| 2,002 | Number of patients with at least 1 Meas, 1 Dx and 1 Rx | 660.6 |
| 2,003 | Number of patients with at least 1 Visit | 15.48 |
| 2,004 | Number of distinct patients that overlap between specific domains | 506.4 |
| 2,100 | Number of persons with at least one device exposure, by device_concept_id | 0.02 |
| 2,101 | Number of device exposure records, by device_concept_id | 0.02 |
| 2,102 | Number of persons by device records start month, by device_concept_id | 0.02 |
| 2,104 | Number of persons with at least one device exposure, by device_concept_id by calendar year by gender by age decile | 0.02 |
| 2,105 | Number of device exposure records, by device_concept_id by device_type_concept_id | 0.02 |
| 2,110 | Number of device_exposure records outside valid observation period | 0.03 |
| 2,120 | Number of device_exposure records by device_exposure start month | 0.02 |
| 2,125 | Number of device_exposure records by device_source_concept_id | 0.02 |
| 2,130 | Number of device_exposure records inside a valid observation period | 0.02 |
| 2,131 | Proportion of people with at least one device_exposure record outside a valid observation period | 0.03 |
| 2,132 | Proportion of device_exposure records outside a valid observation period | 0.02 |
| 2,191 | Percentage of total persons that have at least x device exposures | 0.02 |
| 2,200 | Number of persons with at least one note by note_type_concept_id | 0.03 |
| 2,201 | Number of note records, by note_type_concept_id | 0.02 |

Query executed in 0.44 seconds

# 7. Appendix

## 7.1. Vocabulary table counts

**Table 45.** The number of records in all vocabulary tables. ◎

| TABLENAME | COUNT |
|---|---|
| concept_ancestor | 81,765,717 |
| concept_relationship | 60,512,700 |
| concept | 9,732,666 |
| concept_synonym | 4,031,142 |
| drug_strength | 2,980,115 |
| relationship | 690 |
| concept_class | 417 |
| vocabulary | 127 |
| domain | 50 |

Query executed in 29.91 seconds

## 7.2. Vocabulary concept counts

Vocabulary version: v5.0 31-AUG-23

**Table 46.** The vocabularies available in the CDM with concept count. Note that this does not reflect which concepts are actually used in the clinical CDM tables. S=Standard, C=Classification and '-'=Non-standard ◎

| ID | Name | Version | S | C | - |
|---|---|---|---|---|---|
| ABMS | Provider Specialty (American Board of Medical Specialties) | 2018-06-26 ABMS | 85 | 0 | 13 |
| AMT | Australian Medicines Terminology (NEHTA) | Clinical Terminology v20210630 | 6,839 | 0 | 130,011 |
| APC | Ambulatory Payment Classification (CMS) | 2018-January-Addendum-A | 715 | 0 | 1,195 |
| ATC | WHO Anatomic Therapeutic Chemical Classification | RxNorm 20210907 | 0 | 6,509 | 231 |
| BDPM | Public Database of Medications (Social-Sante) | BDPM 20191006 | 1,106 | 0 | 43,270 |
| Cancer Modifier | Diagnostic Modifiers of Cancer (OMOP) | Cancer Modifier 20220909 | 5,317 | 38 | 688 |
| CCAM | Common Classification of Medical Acts (ATIH) | CCAM version 64 | 0 | 0 | 10,206 |
| CDM | OMOP Common DataModel | CDM v6.0.0 | 1,060 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| CGI | Cancer Genome Interpreter (Pompeu Fabra University) | CGI20180216 | 0 | 0 | 5,351 |
| CIEL | Columbia International eHealth Laboratory (Columbia University) | Openmrs 1.11.0 20150227 | 0 | 0 | 50,881 |
| CIM10 | International Classification of Diseases, Tenth Revision, French Edition (ATIH) | CIM10 2022 | 0 | 0 | 12,226 |
| CIViC | Clinical Interpretation of Variants in Cancer (civicdb.org) | CIViC 2022-10-01 | 0 | 0 | 1,386 |
| ClinVar | ClinVar (NCBI) | ClinVar v20200901 | 0 | 0 | 8,072 |
| CMS Place of Service | Place of Service Codes for Professional Claims (CMS) | 2009-01-11 | 51 | 0 | 9 |
| CO-CONNECT | CO-CONNECT (University of Dundee) | CO-CONNECT 2023-05-31 | 0 | 0 | 537 |
| CO-CONNECT MIABIS | CO-CONNECT MIABIS (University of Dundee) | CO-CONNECT MIABIS 2023-05-31 | 0 | 0 | 60 |
| CO-CONNECT TWINS | CO-CONNECT TWINS (University of Dundee) | CO-CONNECT TWINS 2023-05-31 | 0 | 0 | 3,875 |
| Cohort | Legacy OMOP HOI or DOI cohort | NA | 0 | 78 | 0 |
| Cohort Type | OMOP Cohort Type | NA | 0 | 0 | 1 |
| Concept Class | OMOP Concept Class | NA | 0 | 0 | 417 |
| Condition Status | OMOP Condition Status | NA | 22 | 0 | 0 |
| Condition Type | OMOP Condition Occurrence Type | NA | 0 | 0 | 118 |
| Cost | OMOP Cost | NA | 51 | 0 | 0 |
| Cost Type | OMOP Cost Type | NA | 0 | 0 | 8 |
| CPT4 | Current Procedural Terminology version 4 (AMA) | 2023 Release | 10,489 | 3,605 | 3,078 |

| | | | | | |
|---|---|---|---|---|---|
| CTD | Comparative Toxicogenomic Database (NCSU) | CTD 2020-02-19 | 0 | 0 | 8,698 |
| Currency | International Currency Symbol (ISO 4217) | 2008 | 180 | 0 | 0 |
| CVX | CDC Vaccine Administered CVX (NCIRD) | CVX 20230418 | 223 | 0 | 33 |
| Death Type | OMOP Death Type | NA | 0 | 0 | 14 |
| Device Type | OMOP Device Type | NA | 0 | 0 | 4 |
| dm+d | Dictionary of Medicines and Devices (NHS) | DMD 2023-05-22 | 26,607 | 0 | 378,824 |
| Domain | OMOP Domain | NA | 0 | 0 | 65 |
| DPD | Drug Product Database (Health Canada) | DPD 25-JUN-17 | 0 | 0 | 193,647 |
| DRG | Diagnosis-related group (CMS) | 2011-18-02 | 752 | 0 | 610 |
| Drug Type | OMOP Drug Exposure Type | NA | 0 | 0 | 16 |
| EDI | Korean Electronic Data Interchange code system (HIRA) | EDI 2019.10.01 | 0 | 0 | 313,431 |
| EphMRA ATC | Anatomical Classification of Pharmaceutical Products (EphMRA) | EphMRA ATC 2016 | 0 | 895 | 0 |
| Episode | OMOP Episode | Episode 20201014 | 14 | 0 | 4 |
| Episode Type | OMOP Episode Type | NA | 0 | 0 | 5 |
| Ethnicity | OMOP Ethnicity | NA | 2 | 0 | 0 |
| GCN_SEQNO | Clinical Formulation ID (FDB) | 20151119 Release | 0 | 0 | 29,659 |
| Gender | OMOP Gender | NA | 2 | 0 | 3 |
| GGR | Commented Drug Directory (BCFI) | GGR 20210901 | 751 | 0 | 26,457 |
| HCPCS | Healthcare Common Procedure Coding System (CMS) | 20230701 Alpha Numeric HCPCS File | 7,741 | 0 | 3,816 |
| HemOnc | HemOnc | HemOnc 2022-11-29 | 2,185 | 378 | 5,465 |
| HES Specialty | Hospital Episode Statistics Specialty (NHS) | 2018-06-26 HES Specialty | 57 | 0 | 108 |

| | | | | | |
|---|---|---|---|---|---|
| i2b2 | i2b2 Vocabulary | NA | 0 | 0 | 577,854 |
| ICD10 | International Classification of Diseases, Tenth Revision (WHO) | 2021 Release | 0 | 0 | 16,519 |
| ICD10CM | International Classification of Diseases, Tenth Revision, Clinical Modification (NCHS) | ICD10CM FY2023 code descriptions | 0 | 0 | 98,583 |
| ICD10CN | International Classification of Diseases, Tenth Revision, Chinese Edition (CAMS) | 2016 Release | 0 | 0 | 34,491 |
| ICD10GM | International Classification of Diseases, Tenth Revision, German Edition | ICD10GM 2022 | 0 | 0 | 17,213 |
| ICD10PCS | ICD-10 Procedure Coding System (CMS) | ICD10PCS 2021 | 194,874 | 0 | 107 |
| ICD9CM | International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 1 and 2 (NCHS) | test | 0 | 0 | 17,564 |
| ICD9Proc | International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 3 (NCHS) | ICD9CM v32 master descriptions | 2,223 | 0 | 2,434 |
| ICD9ProcCN | International Classification of Diseases, Ninth Revision, Chinese Edition, Procedures (CAMS) | 2017 Release | 0 | 0 | 13,385 |
| ICDO3 | International Classification | ICDO3 SEER Site/Histology Released 06/2020 | 56,972 | 0 | 7,499 |

| | | | | | |
|---|---|---|---|---|---|
| | of Diseases for Oncology, Third Edition (WHO) | | | | |
| JAX | The Clinical Knowledgebase (The Jackson Laboratory) | JAX v20200824 | 0 | 0 | 7,855 |
| JMDC | Japan Medical Data Center Drug Code (JMDC) | JMDC 2020-04-30 | 1,313 | 0 | 37,485 |
| KCD7 | Korean Standard Classfication of Diseases and Causes of Death, 7th Revision (STATISTICS KOREA) | 7th revision | 0 | 0 | 22,508 |
| KDC | Korean Drug Code (HIRA) | KDC 2020-07-31 | 112 | 0 | 63,749 |
| KNHIS | Korean Payer (KNHIS) | NA | 3 | 0 | 0 |
| Korean Revenue Code | Korean Revenue Code (KNHIS) | NA | 7 | 0 | 0 |
| Language | OMOP Language | Language 20221030 | 1 | 0 | 0 |
| LOINC | Logical Observation Identifiers Names and Codes (Regenstrief Institute) | LOINC 2.75 | 115,614 | 49,729 | 104,134 |
| MDC | Major Diagnostic Categories (CMS) | 2013-01-06 | 26 | 0 | 0 |
| Meas Type | OMOP Measurement Type | NA | 0 | 0 | 12 |
| Medicare Specialty | Medicare provider/supplier specialty codes (CMS) | 2018-06-26 Specialty | 112 | 0 | 8 |
| MeSH | Medical Subject Headings (NLM) | 2023 Release | 0 | 0 | 352,735 |
| Metadata | OMOP Metadata | NA | 1 | 0 | 2 |
| MMI | Modernizing Medicine (MMI) | NA | 4 | 0 | 0 |
| Multum | Cerner Multum (Cerner) | 2013-07-10 | 0 | 0 | 9,770 |
| NAACCR | Data Standards & | NAACCR v18 | 22,807 | 0 | 11,666 |

| | | | | | |
|---|---|---|---|---|---|
| | Data Dictionary Volume II (NAACCR) | | | | |
| NCCD | Normalized Chinese Clinical Drug knowledge base (UTHealth) | NCCD_v02_2020 | 0 | 0 | 51,583 |
| NCIt | NCI Thesaurus (National Cancer Institute) | NCIt 20220509 | 0 | 0 | 2,426 |
| NDC | National Drug Code (FDA and manufacturers) | NDC 20230827 | 11,507 | 0 | 1,160,387 |
| NDFRT | National Drug File - Reference Terminology (VA) | RXNORM 2018-08-12 | 0 | 0 | 69,567 |
| Nebraska Lexicon | Nebraska Lexicon (UNMC) | Nebraska Lexicon 20190816 | 4,187 | 0 | 461,614 |
| NFC | New Form Code (EphMRA) | NFC 20160704 | 0 | 692 | 0 |
| NHS Ethnic Category | NHS Ethnic Category | 2023 | 0 | 0 | 15 |
| NHS Place of Service | NHS Admission Source and Discharge Destination | 2023 | 0 | 0 | 15 |
| None | OMOP Standardized Vocabularies | v5.0 31-AUG-23 | 0 | 0 | 1 |
| Note Type | OMOP Note Type | NA | 0 | 0 | 10 |
| NUCC | National Uniform Claim Committee Health Care Provider Taxonomy Code Set (NUCC) | 2018-06-26 NUCC | 674 | 0 | 181 |
| Observation Type | OMOP Observation Type | NA | 0 | 0 | 29 |
| Obs Period Type | OMOP Observation Period Type | NA | 0 | 0 | 6 |
| OMOP Extension | OMOP Extension (OHDSI) | OMOP Extension 20230531 | 1,219 | 0 | 53 |
| OMOP Genomic | OMOP Genomic vocabulary | OMOP Genomic 20210727 | 79,791 | 0 | 41,200 |

| | | | | | |
|---|---|---|---|---|---|
| OMOP Invest Drug | OMOP Investigational Drugs | OMOP Invest Drug version 2022-05-12 | 0 | 0 | 29,727 |
| OncoKB | Oncology Knowledge Base (MSK) | OncoKB v20210502 | 0 | 0 | 5,569 |
| OncoTree | OncoTree (MSK) | OncoTree version 2021-11-02 | 0 | 0 | 885 |
| OPCS4 | OPCS Classification of Interventions and Procedures version 4 (NHS) | 2021 Release | 2,373 | 0 | 8,627 |
| OPS | Operations and Procedures Classification (OPS) | OPS Version 2022 | 0 | 0 | 42,959 |
| OSM | OpenStreetMap (OSMF) | OSM Release 2019-02-21 | 203,339 | 0 | 0 |
| OXMIS | Oxford Medical Information System (OCHP) | NA | 0 | 0 | 8,118 |
| PCORNet | National Patient-Centered Clinical Research Network (PCORI) | NA | 2 | 0 | 79 |
| Plan | OMOP Health Plan | NA | 11 | 0 | 0 |
| Plan Stop Reason | OMOP Plan Stop Reason | NA | 13 | 0 | 0 |
| PPI | AllOfUs_PPI (Columbia) | Codebook Version 0.4.43 + COVID + MHWB + SDOH + PFH | 2,275 | 0 | 4,254 |
| Procedure Type | OMOP Procedure Occurrence Type | NA | 0 | 0 | 97 |
| Provider | OMOP Provider | NA | 6 | 0 | 0 |
| Race | Race and Ethnicity Code Set (USBC) | Version 1.0 | 50 | 0 | 3 |
| Read | NHS UK Read Codes Version 2 (HSCIC) | NHS READV2 21.0.0 20160401000001 + DATAMIGRATION_25.0.0_20180403000001 | 0 | 0 | 108,945 |
| Relationship | OMOP Relationship | NA | 14 | 0 | 698 |
| Revenue Code | UB04/CMS1450 Revenue Codes (CMS) | 2010 Release | 538 | 0 | 0 |
| RxNorm | RxNorm (NLM) | RxNorm 20230703 | 151,257 | 35,543 | 119,827 |
| RxNorm Extension | OMOP RxNorm Extension | RxNorm Extension 2023-08-24 | 1,867,581 | 0 | 277,744 |
| SMQ | Standardised MedDRA | Version 14.0 | 0 | 318 | 6 |

| | | | | | |
|---|---|---|---|---|---|
| | Queries (MSSO) | | | | |
| SNOMED | Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO) | 2021-07-31 SNOMED CT International Edition; 2021-09-01 SNOMED CT US Edition; 2021-11-24 SNOMED CT UK Edition | 538,088 | 0 | 516,847 |
| SNOMED Veterinary | SNOMED Veterinary Extension (VTSL) | SNOMED Veterinary 20190401 | 31,994 | 0 | 1,690 |
| SOPT | Source of Payment Typology (PHDSC) | SOPT Version 9.2 | 162 | 0 | 6 |
| Specimen Type | OMOP Specimen Type | NA | 0 | 0 | 1 |
| SPL | Structured Product Labeling (FDA) | NDC 20230827 | 0 | 661,892 | 14,973 |
| Sponsor | OMOP Sponsor | NA | 6 | 0 | 0 |
| Supplier | OMOP Supplier | NA | 0 | 0 | 1 |
| Type Concept | OMOP Type Concept | Type Concept 20221030 | 80 | 0 | 0 |
| UB04 Point of Origin | UB04 Claim Source Inpatient Admission Code (CMS) | NA | 0 | 0 | 23 |
| UB04 Pri Typ of Adm | UB04 Claim Inpatient Admission Type Code (CMS) | NA | 6 | 0 | 0 |
| UB04 Pt dis status | UB04 Patient Discharge Status Code (CMS) | NA | 0 | 0 | 55 |
| UB04 Typ bill | UB04 Type of Bill - Institutional (USHIK) | NA | 4 | 0 | 294 |
| UCUM | Unified Code for Units of Measure (Regenstrief Institute) | Version 1.8.2 | 1,029 | 0 | 89 |
| UK Biobank | UK Biobank (UK Biobank) | version 2021-03-18 | 3,837 | 292 | 15,208 |
| US Census | Census regions of the United States (USCB) | US Census 2017 Release | 13 | 0 | 0 |
| VA Class | VA National Drug File Class (VA) | RxNorm 20230807 | 0 | 0 | 576 |
| VANDF | Veterans Health Administration | RxNorm 20230807 | 10,290 | 0 | 31,395 |

| | | | | | |
|---|---|---|---|---|---|
| | National Drug File (VA)) | | | | |
| Visit | OMOP Visit | Visit 20211216 | 19 | 0 | 0 |
| Visit Type | OMOP Visit Type | NA | 0 | 0 | 18 |
| Vocabulary | OMOP Vocabulary | NA | 0 | 0 | 149 |

Query executed in 5.08 seconds

**Intégration et utilisation secondaire des données de santé hospitalières hétérogènes : des usages locaux à l'analyse fédérée**

**Résumé :** Les données issues du soin peuvent être utilisées pour des finalités autres que celles pour lesquelles elles ont été collectées initialement : c'est l'utilisation secondaire des données de santé. Dans le contexte hospitalier, afin de lever les verrous de l'utilisation secondaire des données de santé (verrous liés aux données et verrous organisationnels), une stratégie classique consiste à mettre en place un Entrepôt de Données de Santé (EDS). Dans le cadre de cette thèse, trois contributions à l'EDS du CHU de Bordeaux sont décrites. Premièrement, une méthode d'alignement des data elements de biologie numérique basée sur les instances et conforme aux règles de protection des données à caractère personnel est présentée, avec une F-mesure à 0,850, permettant de réduire l'hétérogénéité sémantique des données. Ensuite, une adaptation du modèle d'intégration des données cliniques d'i2b2 est proposée pour assurer la persistance des données d'un EDS dans une base de données NoSQL, Elasticsearch. Cette implémentation a été évaluée sur la base de données de l'EDS du CHU de Bordeaux et retrouve des performances améliorées en termes de stockage et de temps de requêtage, par rapport à une base de données relationnelle. Enfin, une présentation de l'environnement EDS du CHU de Bordeaux est réalisée, avec la description d'un premier EDS dédié aux usages locaux et qui peut être exploité en autonomie par les utilisateurs finaux (i2b2), et d'un second EDS, dédié aux réseaux fédérés (OMOP) permettant notamment la participation au réseau fédéré DARWIN-EU.
**Mots-clés :** données de santé hétérogènes, entrepôt de données de santé, alignement de data elements, biologie, elasticsearch, i2b2, réseau fédéré, OMOP

---

**Integration and secondary use of heterogeneous hospital health data: from local uses to federated analysis**

**Abstract:** Healthcare data can be used for purposes other than those for which it was initially collected: this is the secondary use of health data. In the hospital context, to overcome the obstacles to secondary use of healthcaree data (data and organizational barriers), a classic strategy is to set up Clinical Data Warehouses (CDWs). This thesis describes three contributions to the Bordeaux University Hospital's CDW. Firstly, an instance-based, privacy-preserving, method for mapping numerical biology data elements is presented, with an F-measure of 0,850, making it possible to reduce the semantic heterogeneity of data. Next, an adaptation of the i2b2 clinical data integration model is proposed to enable CDW data persistence in a NoSQL database, Elasticsearch. This implementation has been evaluated on the Bordeaux University Hospital's CDW, showing improved performance in terms of storage and query time, compared with a relational database. Finally, the Bordeaux University Hospital's CDW environment is presented, with the description of a first CDW dedicated to local uses that can be used autonomously by end users (i2b2), and a second CDW dedicated to federated networks (OMOP) enabling participation in the DARWIN-EU federated network.
**Keywords:** heterogeneous health data, clinical data warehouse, data element mapping, biology, elasticsearch, i2b2, federated network, OMOP

---

**AHeaD - Bordeaux Population Health**
Inserm U1219, Université de Bordeaux, 33000 Bordeaux, France.