UNIVERSITE DE LILLE

UFR3S, FACULTE DE MEDECINE HENRI WAREMBOURG

Année 2023

THESE POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE

Réutilisation de données en médecine générale

Présentée et soutenue publiquement le 13 Octobre 2023
à 18h00 au Pôle Formation

Par Eole Nyangwile

JURY

Président :

Monsieur le Professeur Philippe AMOUYEL

Assesseurs :

Monsieur le Docteur Matthieu CALAFIORE
Monsieur le Docteur Bertrand LEGRAND

Directeur de thèse :

Monsieur le Professeur Emmanuel CHAZARD

Avertissement

La Faculté n’entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Sommaire

Table des matières

Résumé

Contexte : Avec l'avènement du numérique, de volumineuses données s'accumulent, permettant de constituer des entrepôts de données de plus en plus importants. La réutilisation des données, et notamment celles de santé produites à l'occasion d'un épisode de soin, est décrite dans de nombreuses études. Cependant, très peu d'entre elles décrivent les modalités de réalisation de ce travail sur des données issues de cabinets de médecine générale, et aucune n'inclut de données françaises.

Notre objectif est de réaliser ce procédé de transformation vers le modèle de données OMOP avec des données semblables à celles issues de cabinet de médecine générale, puis d'en réaliser une rapide analyse descriptive.

Matériel et Méthode : Nous utilisons un jeu de données simulées dans le cadre du CPER Tec'Santé. Ce jeu de données inclut15000 patients fictifs, et respecte la structure et les distributions observées dans une base de données réelle. Nous avons réalisé un processus d'ETL vers le modèle de données commun OMOP. Nous avons dans un premier temps réalisé et détaillé la cartographie syntaxique puis la cartographie sémantique avant de présenter une rapide analyse descriptive des résultats.

Résultats : Grâce au modèle de données commun OMOP, les 7 jeux de données fictives ont pu être transformées dans format standardisé. Concernant le vocabulaire dans chaque base, 51% à 100% des terminologies disponibles ont pu être mises en correspondance avec un concept normalisé lors de la cartographie sémantique.

Discussion : Ce travail montre la faisabilité du processus avec des données issues des cabinets de médecine générale français. Les résultats descriptifs illustrent le potentiel scientifique de telles données. Un tel processus doit être réalisé en équipe.

Liste des abréviations

ALD	Affection Longue Durée
AM	Assurance Maladie
ATC	*Anatomical Therapeutic Chemical-code*
ATIH	Agence Technique de l'Information Hospitalière
CCAM	Classification Commune des Actes Médicaux
CDA	*Clinical Document Architecture*
CIM	Classification Internationale des Maladies
CISP	Classification Internationale des Soins Primaires
CNIL	Commission Nationale de l'Informatique et des Libertés
CNSA	Caisse Nationale de Solidarité pour l'Autonomie
CPER	Contrat Plan Etat Région
DMP	Dossier Médical Partagé
DPI	Dossier Patient Informatisé
ECG	ElectroCardioGramme
EDS	Entrepôt de Données de Santé
ETL	*Extract Transformation and Loading* / Extraction Transformation et Chargement
FDA	*Food and Drug Administration*
FHIR	*Fast Healthcare Interoperability Resources*
GAM	Gestion Administrative des Malades
HL7	*Health Level Seven*
HPRIM	Harmonie et Promotion de l'Informatique Médicale
HTTP	*Hypertext Transfer Protocol* / Protocole de Transfert Hypertexte
LOINC	*Logical Observation Identifiers Names and Codes*
MCO	Médecine, Chirurgie, Obstétrique

MDC/CDM *Common Data Model* / Modèle de Données Commun

MDPH Maison Départementale des Personnes Handicapées

OHDSI *Observational Health Data Sciences and Informatics*

OMOP *Obersvational Medical Outcomes Partenership*

P4DP *Platform For Data in Primary care*

PCORNet *United States Patient Centered Outcomes Research Network*

PMSI Programme de Médicalisation des Systèmes d'Information

RIM *Reference Information Model*

SIH Système d'information hospitalier

SNDS Système National des Données de Santé

SNIIRAM Système National d'Identification Inter-Régimes de l'Assurance Maladie

SNOMED *Systematized Nomenclature of Medicine Clinical Terms*

SOAP *Simple Object Access Protocol*

SQL *Structured Query Language*

SSR Soins de Suite et Rédaptation

UCD Unité Commune de Dispensation

XML *Extensible Markup Language*

Introduction

Ce travail de thèse s'intéresse à la réutilisation de bases de données de santé issues des cabinets de médecine générale. Dans un premier temps nous définirons ce que sont les données massives (*Big Data*) et la réutilisation de données (*Data Reuse*).

Puis nous donnerons quelques éléments de contexte sur la méthodologie décrite dans la littérature pour extraire et caractériser ces données ainsi que sur les méthodes généralement employées pour réutiliser ces données.

Enfin, nous définirons l'objectif de ce travail.

1 Données massives ou Big data

Avec les années et l'avènement du numérique, un grand nombre de données ont été accumulées permettant de constituer des bases d'informations de plus en plus grandes dans de nombreux domaines. Dans le domaine de la santé l'informatisation des dossiers médicaux a notamment permis l'augmentation de l'information facilement mobilisable concernant les personnes bénéficiant de soins allant jusque la création de bases de données dites massives [1]. Ces **bases de données massives** ou *Big Data* sont définies comme étant des données de grande dimension aussi bien en termes de volume d'individus, de variables [2], de modalités ou nombre de mesures pour ces variables ainsi que de nombre de tables et relations.

Plusieurs auteurs s'accordent à caractériser [2–5] ces données massives en trois grands aspects (Les 3V) :

- Leur **volume** (taille des bases de données de grande importance)
- Leur **variété**, diversité (différentes provenances, différentes présentations des données : données structurées ou non) pouvant avoir pour conséquence la complexité de leur exploitation
- Leur **vélocité** (conservation des données et fréquence des mises à jour des bases avec possible remplacement)

Auxquelles s'ajoutent parfois d'autres caractérisations avec notamment

- Des enjeux de **véracité** : (fiabilité des données notamment dans le domaine de la santé où limité codage paiement) [2,6].
- Leur **valorisation et impact** (gestion, traitement et analyse des données différents selon leur finalité) [5].

Dans le domaine de la santé, ces données sont souvent collectées lors des soins pour une finalité première qui est le soin et la continuité des soins, mais nous allons voir que ces données peuvent être réutilisées.

2 Réutilisation de données ou *data reuse*

La réutilisation de données, aussi appelée « utilisation secondaire » de données correspond à l'utilisation de données à des fins autres que celles prévues initialement

[7]. En termes informatiques, il est usuel de qualifier la première utilisation de « transactionnelle », et les réutilisations de « décisionnelles ».

Dans le domaine de la santé, la réutilisation de données se définit comme l'usage de données initialement collectées lors de la production de soins ou la facturation des soins, à des fins secondaires de recherche, d'analyse de l'activité, d'analyse médico-économique, de calcul d'indicateurs de qualité et sécurité des soins, de santé publique, de pharmacovigilance, etc… [7,8]

Cette réutilisation des données présente plusieurs avantages :

- Liées à la **volumétrie** des données (permettant d'avoir des données relativement exhaustives sur une grande population non sélectionnée)
- Liées à la **disponibilité** des données (permettant une analyse à faible coût par exemple dans le cadre de la recherche)
- Liées à la **variété** des données (permettant aussi l'étude de nombreux paramètres de santé et leur interaction)

Mais aussi des contraintes [7–9]:

- Liées aux **modalités de recueil** qui se doivent adaptées aux soins du patient (données concernant le diagnostic, les actes effectués, les traitements administrés, les comptes rendus d'examen paracliniques, les données non structurées (images et texte libre)) traduisant une richesse de l'information qui peut parfois être trop importante et impacter la fouille des données
- Liées à la **collection et au stockage** des données massives qui permettent souvent uniquement l'utilisation de données structurées relatives au remboursement ou la régulation
- Liées à la **perte d'information** (consécutive à plusieurs données notamment celles sus citées)
- Liées à la nécessité **d'agréger** les données compte tenu de leur volume
- Liées aux **considérations éthiques** (pas de consentement explicite pour l'utilisation des données nécessitant une anonymisation ou une dé-identification)
- Liées à la **complexité** de la méthode d'extraction et de réutilisation de ces données

La réutilisation de données structurées (voir le chapitre 3 : Interopérabilité) est habituellement réalisée en 5 étapes [10] (Figure 1) :

- Le **pré-processing** des données (transformation des données d'une base de données dite « native » en un entrepôt de données standardisé).
- **L'extraction** de **caractéristiques** (afin de passer d'un entrepôt de données contenant des dizaines ou centaines de tables, à une seule table de données pouvant être analysées par des méthodes traditionnelles : ces données ressemblent alors à des données issues d'un questionnaire).
- **Analyse graphique et statistique** (de ces données individuelles afin d'identifier des éventuelles association)
- **Réorganisation et filtrage** des résultats de l'analyse
- **Prise de décision**

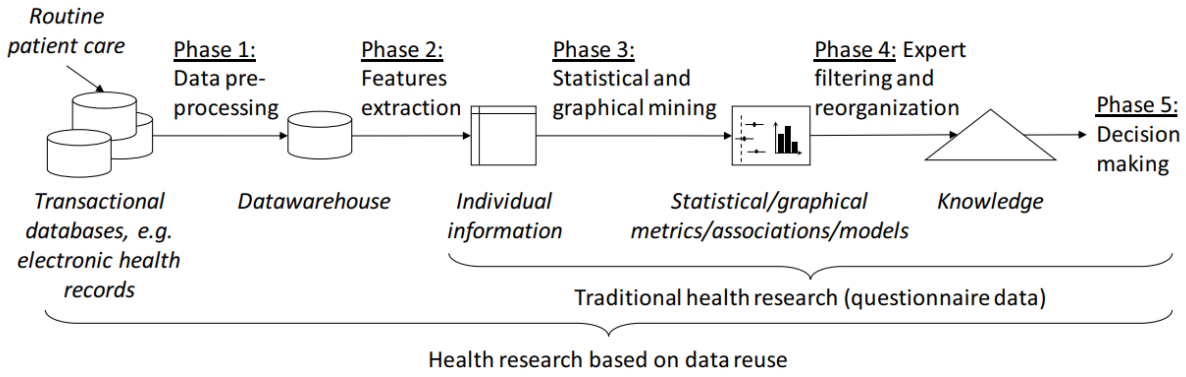

Figure 1. Processus complet de réutilisation de données [10])

Ce travail de thèse s'intéressera principalement à la phase 1 dite de « pré-traitement », qui permet de constituer un entrepôt de données de santé (EDS). Nous nous y intéresserons en particulier dans le cadre des données de médecine ambulatoire. Nous allons définir ce qu'est l'interopérabilité, concept essentiel à la mécanique de réalisation du « pré-traitement »

3 Interopérabilité

3.1 Généralités

L'interopérabilité est la capacité de différents produits ou systèmes à fonctionner ensemble, à savoir communiquer, échanger et utiliser les informations échangées [11].

Dans un contexte transactionnel, l'interopérabilité est essentielle pour faire cohabiter des logiciels traitant des données d'un même patient. Ainsi par exemple, dans un système d'information hospitalier (SIH), il est essentiel que le logiciel du laboratoire de biologie médicale connaisse l'identité des patients telle qu'elle est définie dans le logiciel de gestion administrative des malades (GAM). L'interopérabilité s'entend alors essentiellement à travers **l'échange de messages standardisés**. C'est le cas par exemple du standard HL7 CDM (pour *Charge Description Master Segment*) aussi nommé HL7 Version 2 [12].

Dans un contexte décisionnel, l'interopérabilité permet d'une part d'**agréger des données**, au sujet des mêmes patients, issues de logiciels différents (ex : résultats d'analyse de biologie médicale et diagnostics codés d'un même patient), mais également d'agréger des données au sujet de patients différents issues de logiciels similaires (ex : résultats d'analyse de biologie médicale issus de deux établissements différents, au sujet de patients différents). L'interopérabilité s'entend alors essentiellement à travers la définition de **modèles de données standardisés** et de **terminologies partagées** pour les extractions de données en masse. C'est le cas par exemple du standard HL7 RIM (pour *Reference Information Model*) [13].

L'interopérabilité est un concept important car ce travail de thèse reposera en partie sur la possibilité de pouvoir faire communiquer nos différentes bases de données et sources de données ensemble afin de pouvoir en sortir une table de donnée uniforme exploitable.

On définit 3 champs pour l'interopérabilité, détaillés ci-dessous.

3.2 Interopérabilité physique

Aussi appelée interopérabilité technique, elle correspond aux applications et **infrastructures** pouvant relier entre eux différents **systèmes ou services technologiques**. Cela inclut les spécificités d'interface, la possibilité d'interconnexion des services, les services d'intégration des données, les modalités de présentation et d'échange des données et les protocoles de communication sécurisés [14,15].

Les standards d'interopérabilité physique relèvent de l'informatique générale. On peut citer par exemple SOAP (*Simple Object Access Protocol*), qui définit l'envoi de fichiers XML (*Extensible Markup Language*) par le protocole HTTP (*Hypertext Transfer Protocol* ou Protocole de transfert hypertexte). Ces standards ne définissent toutefois pas le contenu des messages.

Contrexemple : si un système envoie automatiquement un email alors qu'un autre système attend une connexion HTTP, alors ces systèmes ne sont pas interopérables physiquement. Leur exploitation jointe nécessiterait la mise en place d'un programme de conversion de format de données.

3.3 Interopérabilité syntaxique

L'interopérabilité syntaxique ou structurelle correspond à la **structure ou le format** d'échange de données entre différentes **bases de données** à savoir leur nature, leur type et leur format [15]. Cela permet une cohérence de syntaxe des différentes données et d'assurer l'échange de données non altérées. Plus simplement, dans un schéma relationnel, l'interopérabilité syntaxique suppose que les tables de données soient définies de la même manière, et comportent exactement les mêmes colonnes.

Contrexemples : si deux systèmes prennent en chargent le même format de données ou une même syntaxe mais ne représentent pas les données de la même manière (ex : l'un représente l'âge du patient, tandis que l'autre représente sa date de naissance et sa date de consultation), alors ces systèmes ne sont pas interopérables syntaxiquement. Leur exploitation jointe nécessiterait la mise en place d'un programme de transformation de données, au sens communément admis dans les ETL (*Extract Transformation & Loading* ou Collecte des données, transformation et implémentation en français).

Les principaux standards étant : HL7 FHIR CDA RIM, OMOP, PCORnet [16].

3.4 Interopérabilité sémantique (terminologies)

L'interopérabilité sémantique est la capacité de plusieurs systèmes d'information à échanger et à réussir à utiliser ces données de manière compréhensible. L'interopérabilité sémantique se place donc à la fois dans **l'échange** mais également dans le **codage de la donnée** et son vocabulaire afin que le système qui la recevra puisse l'interpréter. On utilise souvent un **vocabulaire commun**, aussi appelé une norme terminologique (un standard, obtenu à partir du contenu local par normalisation puis une transformation sémantique de la data dans une norme terminologique définie à une échelle locale ou internationale afin d'éliminer les ambiguïtés de compréhension) [15,17].

Contrexemples : si un système de données code le sexe d'une personne en « 0 » / « 1 » et un autre en « Homme » / « Femme » alors ces deux systèmes ne sont pas

interopérables sémantiquement. Leur exploitation jointe nécessiterait la mise en place de tables de correspondance, ou *mappings*.

En France, les terminologies les plus connues pour les données de santé sont la **CIM** (Classification Internationale des Maladies pour les diagnostics et affections), la **CCAM** (Classification Commune des Actes Médicaux pour les actes), **ATC** (*Anatomical Therapeutic Clinical*, classification internationale regroupant de nombreuses substances actives et médicaments), **UCD** (Unités Communes de Dispensations pour les substances actives délivrées en établissement de santé), et **LOINC** (*Logical Observations Identifiers Names and Codes*, une classification internationale pour référencer l'expression des résultats de biologie).

Après avoir défini les différents types d'interopérabilité et leurs champs d'action, nous allons décrire les principaux standards d'interopérabilité utilisés dans le domaine de la santé.

4 Principaux standards d'interopérabilité en santé

4.1 Généralités

Il existe de nombreux standards d'interopérabilité pour transmettre des données de santé et permettre à des logiciels de communiquer. Ces données sont alors généralement des petits ensembles d'informations relatives à un seul patient, qu'on appelle volontiers des « messages ». Il peut par exemple s'agir de transmettre un résultat d'analyse de biologie médicale (avec HPRIM), un ensemble d'images (avec DICOM), ou encore d'informer les différents logiciels d'un SIH de la création d'un séjour pour un patient donné par le logiciel de GAP (gestion administrative des patients) (avec HL7-RIM) [10].

Tous ces **standards**, aussi **appelés normes**, ont un rôle essentiel pour les données de santé en permettant l'échange et le partage d'informations sur les patients, et séjours entre les différents systèmes de soins. Ces standards garantissent alors que toutes les données échangées peuvent être interprétées et exploitées de manière **cohérente** par les différents systèmes.

Certains standards, moins nombreux, peuvent être utilisés pour décrire des bases de données entières, comportant des informations hétérogènes au sujet de plusieurs patients. Sous la forme de modèles de données commun (MDC ou *CDM Common Data model*) ils permettent pour les données de santé de normaliser la structure ou le vocabulaire afin de permettre leur exploitation à grande échelle [16,18–20].

Les plus connus sont HL7-RIM, PCORnet, et OMOP. Nous décrirons chacun d'entre eux ci-dessous.

4.2 HL7-RIM (ou v3) et HL7 FHIR (ou v5)

HL7 (*Health Level Seven)* est un **ensemble de normes** [21] largement utilisé, développé par l'organisation internationale *Health Level Seven International*. Il définit des normes de messagerie et de documentation pour l'échange, l'intégration, le partage et la recherche d'informations électroniques sur la santé [REF]. HL7icu, axé sur les données cliniques et administratives et comporte différentes versions, telles

que HL7 v2 (ou CDA) [12] , v3 (ou RIM) [13] et la norme plus récente *Fast Healthcare Interoperability Resources* (FHIR ou v5) [22].

HL7 v3, basée sur un modèle d'information de référence (RIM) est une norme utilisée pour traiter et échanger des données de santé complexes. Viangteeravat et al. [23] démontrent d'ailleurs son utilité pour l'intégration des informations dans le processus de cartographie de données de santé.

Par ailleurs, **HL7-RIM** est ce jour le modèle utilisé en France pour le **Dossier Médical Partagé** (DMP). Et on lui distingue, dans sa forme la plus développée 4 types d'entité de base : (cf Figure 2)

- Sujet (toutes les informations liées aux soins dont par exemple : les informations sur la personne, sa localisation, son statut vivant/décédé ….)
- Rôle (des personnes intervenant dans la prise en charge)
- Participation
- Actes (toute action réalisée impliquant le patient avec par exemple : les observations constatées, mesures, diagnostics, régime alimentaire, …)

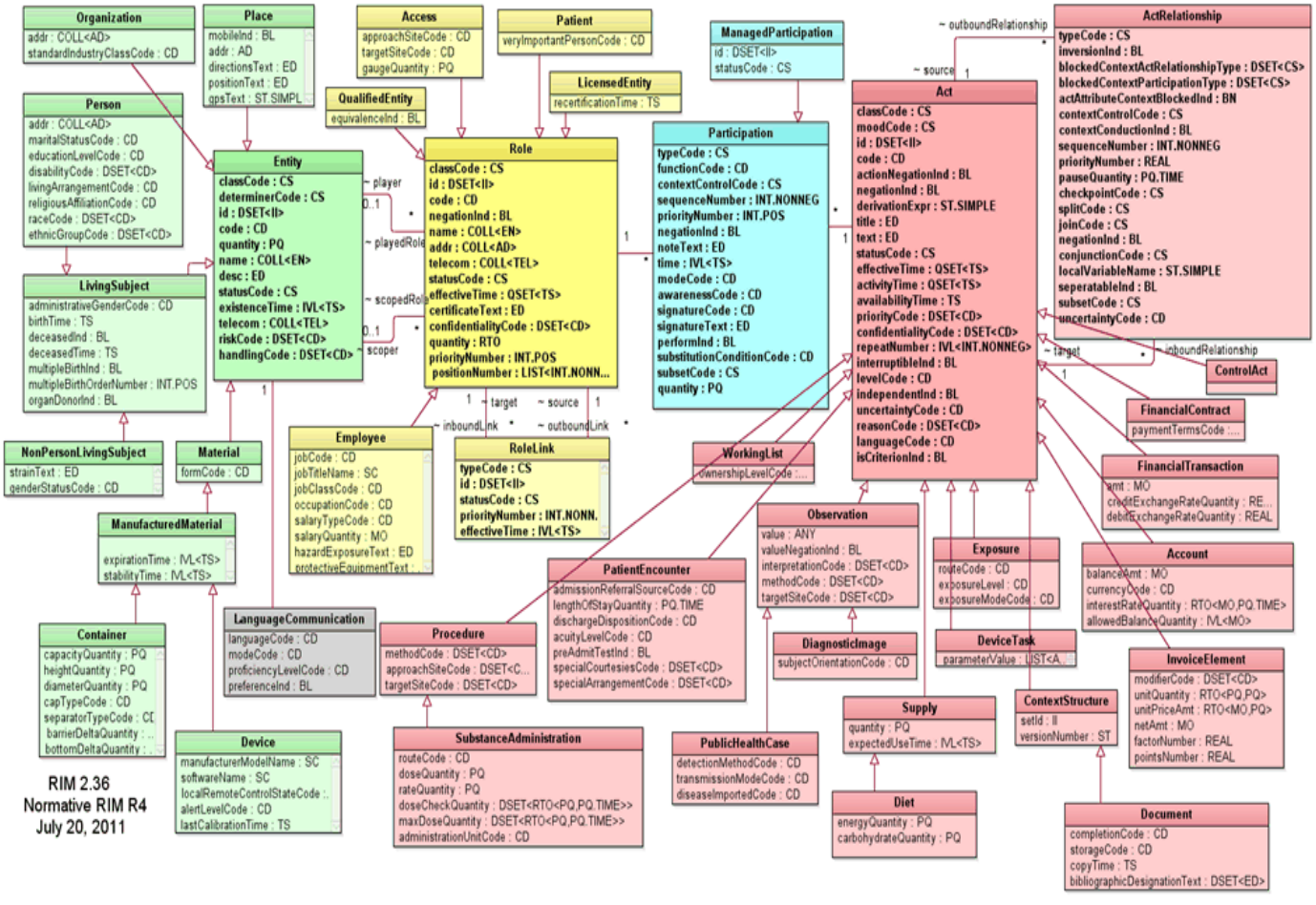

Figure 2. Représentation graphique de l'utilisation de la norme HL7-RIM pour les données de santé [24]

HL7-FHIR, autre norme internationale pour l'échange et le traitement de données de santé a vu son utilisation et son exploitation augmenter de plus en plus dans le traitement des données de santé [25,26]. De par son utilisation, il a permis de rendre

ces données plus accessibles que ce soit pour les échanges (via l'interopérabilité), la sécurité, la mise en place de procédés, etc … [27]. Ayaz et al. [27] précisait dans son étude les quelques domaines dans lesquels HL7-FHIR avait pu être utile et mis en place : notamment la **recherche**, la **revue de dossier** électroniques et de **pratiques** médicales, l'**exploitation et la gestion de bases** de données, mais également dans un domaine plus large que la santé avec les applications mobiles, le traitement d'images, etc ...

4.3 PCORnet

PCORnet (*The National Patient Centeer Clinical Research Network*) est un outil national, étasunien, créé en 2014 par PCORI (*Patient-Centered Outcomes Research Institute*), un partenariat entre les instituts de santé nationaux (NIH) et plusieurs instituts de recherche, réseaux de soins de santé et organisations de patients.

Le premier but de ce modèle de données commun est d'améliorer la qualité de la prise en charge des patients en agissant sur la **qualité des données de santé**. Il permet également d'encourager la **recherche** impliquant les données patient en fournissant un accès plus rapide et efficace aux données de santé [28].

Dans cette optique, le modèle de données commun, PCORnet standardise l'architecture et le vocabulaire de données de santé issues de diverses sources (dossiers médicaux électroniques, registres, laboratoires, ...).

Ce modèle a montré son potentiel lors de la réutilisation de données de santé lors d'une étude multicentrique à des fin de recherche [29].

4.4 OMOP

Plus récemment et dans la même optique mais avec une plus grande échelle, le partenariat OMOP [30] (*Observational Medical Outcomes Partnership*), entre diverses institutions publiques et privées a été créé aux Etats Unis en 2008. Il est présidé par la FDA (*Food and Drug Administration*) et financé par plusieurs groupements pharmaceutiques. Le consortium OHDSI (*Observational Health Data Sciences and Informatics*) assure aujourd'hui son développement et sa maintenance. Il avait pour ambition initiale **d'améliorer la qualité et sécurité des produits médicaux** en développant et promouvant l'utilisation de modèles de données de santé avec une organisation commune et a donc collaboré avec différents chercheurs et partenaires de données de santé (hôpitaux, assurances, …).

Le modèle de données OMOP-CDM *(Common Data Model,* ou modèle de données commun) est un **modèle de données standardisé** permettant d'uniformiser les différentes bases fournissant des données de santé sur leur structure, leur contenu et la sémantique. Cela s'est traduit par la création d'un modèle de données commun avec l'utilisation d'un **vocabulaire normalisé** pour permettre de réaliser des analyses fiables sur les différentes bases de données.

Le CDM permet ainsi, après transformation de réaliser des analyses normalisées sur les différentes sources de données notamment pour les études scientifiques.

Les principaux éléments recensés et normalisés dans le CDM sont les informations concernant [31]:

- La personne : informations démographiques et dates de naissances, décès sur les individus
- La période d'observation
- L'occurrence des visites et rencontres médicales
- Les pathologies : avec diagnostics spécifiques
- L'exposition aux médicaments : informations de prescription et/ou administration
- Les procédures : actes et interventions réalisées
- Les observations non liées aux pathologies ou procédures
- L'exposition aux dispositifs médicaux
- Les notes : en texte libre

Les différentes sources de données sont ainsi collectées puis soumises à un **processus d'ETL** (*Extraction Transformation and Loading*, ou Extraction Transformation et Chargement). Ce processus en 3 étapes permet dans un premier temps d'extraire les données des différentes sources puis de les transformer au format standardisé OMOP-CDM. Cela pouvant impliquer la mise en correspondance des éléments de données de la source avec divers **concepts normés** du CDM ou même le changement de format des données. Enfin ces données sont chargées dans la base de données CDM de l'OMOP où elles deviennent accessibles à tous les chercheurs et analystes pour diverses **études** ou **analyses** visant **à améliorer la gestion de la santé**, la **surveillance sanitaire**, ou encore la **qualité des soins** fournis.

Disponibles en open-data, les différentes données issues du CDM-OMOP sont aussi bien utilisées par les chercheurs et scientifiques que par les sociétés pharmaceutiques, organismes de régulation des produits de santé, établissements prestataires de soins, …

Bien que OMOP et son modèle de données commun ne soient pas interchangeables avec les normes d'interopérabilité traditionnelles telles que HL7 (dont RIM), les deux normes de standardisation des données de santé constituent une base importante pour l'organisation des données de santé dans un format adapté à la recherche.

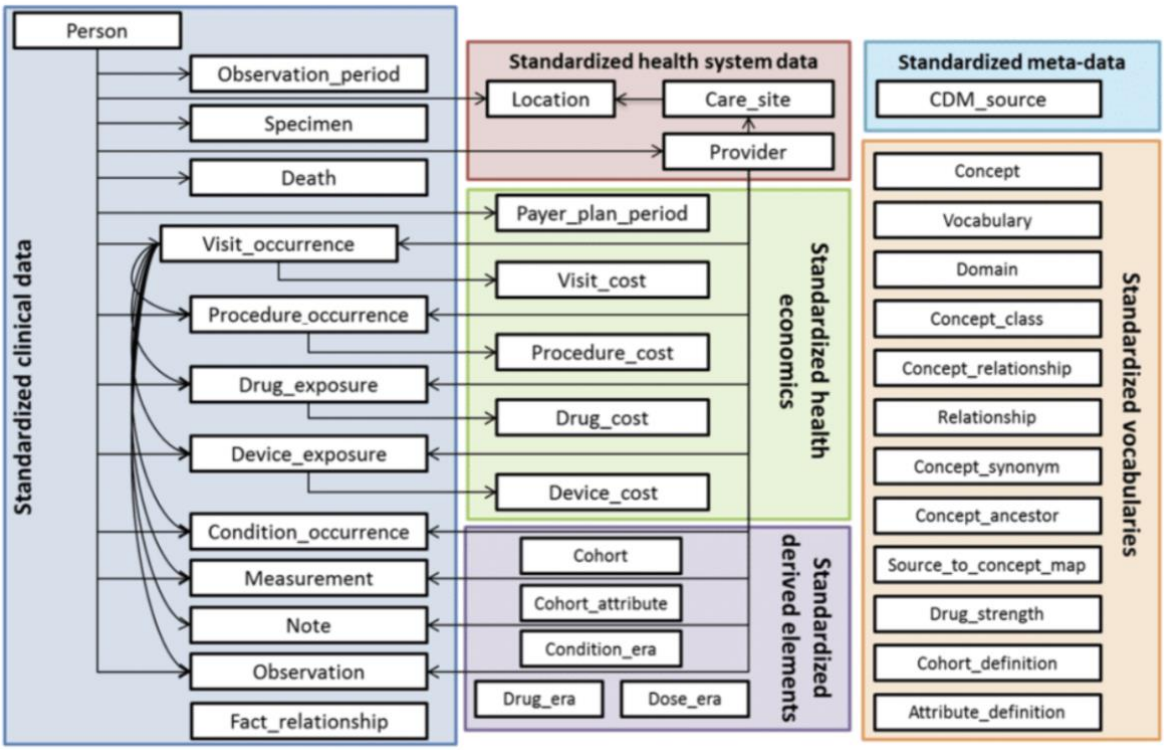

Figure 3. Représentation graphique de l'utilisation du CDM-OMOP pour les données de santé [31]

Le modèle de données commun OMOP a par ailleurs déjà fait ses preuves dans la standardisation de nombreuses bases de données [32–40] dont des bases de données de santé européennes telles que le SNDS ou d'autres internationales [41].

De par la richesse de sa documentation et de la multiplicité des outils fournis c'est ce modèle de données commun qui sera utilisé dans notre travail.

D'une part, quelques études ont prouvé une meilleure qualité et complétude dans les processus utilisés dans le modèle de données commun OMOP par rapport à la norme HL7-FHIR ou encore au modèle de données commmun PCORnet [16,20,42].

D'autre part, et bien que potentiellement utilisable en synergie avec les autres modèles de données commun [34,43], il nous semble plus pertinent au vu de la richesse de ce modèle, de réaliser ce travail en utilisant le MDC-OMOP.

Maintenant que nous avons défini les concepts de réutilisation de données, d'interopérabilité et les différents modèles de données communs principalement utilisés dans le domaine de la santé, nous allons préciser les grandes caractéristiques des données nous souhaitons étudier à savoir les données de santé.

5 Données structurées de santé

5.1 Généralités

Les données de santé sont de plusieurs nature. Ces données sont très souvent **recueillies pour assurer la continuité des soins**, mais leur usage peut parfois être différent.

Il est nécessaire de différencier la typologie de ces données et donc l'usage qu'il est possible d'en faire.

On retrouve des **données structurées** [44], renseignées sous un format unique et souvent visibles sous forme de tableaux ou listes de données ; ainsi que des **données non structurées ou semi-structurées** (images, texte libre, signaux, …). Nous nous intéresserons uniquement aux données structurées dans ce travail.

Les données de santé peuvent parfois être produites automatiquement (comme c'est le cas avec les résultats de biologie produits par les automates et pour certains relus par les humains), parfois par des humains dans le flux normal (données administratives, courriers), parfois encodées par des humains a posteriori en plus (données concernant les diagnostics encodés en CIM10 et données concernant les actes codés en CCAM).

5.2 Données structurés fréquemment rencontrées

Les principaux types de données structurées recueillis sont :

- Données **d'identité** (nom, prénom, date de naissance, adresse)
- Données de **couverture assurantielle**, de **facturation** (régime ou caisses d'assurance maladie obligatoire et complémentaire, identité, droits et motifs de ces droits…)
- Données de **mouvement** (dates de consultation, dates d'entrée ou de sortie, lieux de consultation, lieux d'hospitalisation, etc ..)
- **Diagnostics** codés en CIM10 [45] ou dans d'autres terminologies (en hospitalisation, ces informations sont obligatoirement recueillies dans le cadre du PMSI)
- **Actes** codés en CCAM [46] ou dans d'autres terminologies (en hospitalisation, ces informations sont obligatoirement recueillies dans le cadre du PMSI ; en ambulatoire, certains actes peuvent être facturés en CCAM)
- **Analyses de biologie** médicale codées en LOINC (on distinguera les données de prescriptions et les résultats témoignant de l'analyse des prélèvements, qu'ils soient réalisés par des automates ou des professionnels de santé) [47]
- **Médicaments** prescrits / délivrés / administrés codés en UCD ou ATC [48]
- **Données médicales** (produites par les appareils de suivi en routine : mesures de tension, suivi de dispositifs implantables ou externes)

5.3 Exemples de contextes de réutilisation de données structurées

5.3.1 Bases nationales assurantielles

En France, on identifie le système national d'identification inter-régimes de l'assurance maladie (SNIIRAM) [49]. Le SNIIRAM intègre les deux principales bases de données existantes du système de santé français : celle des **soins de ville** produite par l'Assurance Maladie à partir des données de remboursement et celle des **hospitalisations**, le PMSI [50].

A titre plus global, toutes ces données ainsi que celles concernant les causes médicales de décès (CépiDC), celles relatives au handicap (MDPH données de la CNSA) ou encore quelques données provenant des organismes d'Assurance Maladie complémentaire sont rassemblées dans le **SNDS** (système National des Données de Santé) [51].

Toutes ces données sont reliées entre elle par un **identifiant unique** pour chaque **assuré**. Il s'agit d'une des bases de données de santé les plus complètes permettant de suivre le parcours du patient en combinant les données de santé de différents systèmes.

Concernant les données de soins de ville produites et collectées par l'assurance maladie [49] ; ce sont principalement des données liées à des coûts en termes de santé associés aux soins et notamment aux dépenses réalisées par l'Assurance Maladie pour la santé en France. Il s'agit d'une volonté politique historique de rendre la santé accessible à tous dans le pays et donc à permettre aux citoyens d'avoir une prise en charge de leurs dépenses de santé. Cette prise en charge étant contrebalancée par le respect d'un parcours optimal de santé afin de recevoir des soins adaptés pour tous. Ainsi le gouvernement établit un budget annuel qui sera alloué à la santé et permettra prendre en charge une partie des dépenses des usagers via le système de l'Assurance Maladie. Cette couverture assurantielle médicale est obligatoire pour tous les citoyens du territoire. Concernant les données, on retrouve [52] :

- La **date** de la prestation
- Le **type** de **prescripteur**
- Le **type** de **prestation** (soins et prestations en espèces)
- Le **type d'exécutant** (médecins par spécialité, chirurgiens-dentistes, auxiliaires médicaux, laboratoires d'analyses, pharmaciens, etc.)
- Le **type de poste de dépenses**
- Le **taux de remboursement**
- Le **type de couverture**

Nous détaillerons les données du PMSI dans le paragraphe suivant.

5.3.2 Bases intra-hospitalières

En France, les données de santé intra-hospitalières sont regroupées et accessibles de manière globale via le **Programme de médicalisation des systèmes d'information** (PMSI). On y retrouve principalement les données de MCO (médecine, chirurgie et obstétrique), chirurgie, SSR et psychiatrie.

Ces données sont gérées au sein de chaque établissement lors des différentes hospitalisations de patients puis envoyées sous un format unique chaque année à

l'ATIH (Agence technique de l'information sur l'hospitalisation) [53] qui collige l'ensemble des informations de santé. Ces données déclarées permettent aux établissements de **déclarer leur activité** et ainsi d'**obtenir une dotation** liée à l'activité de soins.

Il s'agit de données structurées et non structurées pouvant correspondre à (liste non exhaustive) :

- Des données **socio-démographiques**
- Des données **assurantielles**
- Des données de **facturation** (tarification séjour qui sera envoyé ensuite à l'assurance maladie)
- Des données **cliniques** (diagnostics, actes, ..)
- Des données **paracliniques** (résultats de biologie, imagerie, ..)
- Des **notes**

Enfin, avec les données assurantielles, ces données peuvent être accessibles par des personnes extérieures à l'établissement sur demande et avec une autorisation CNIL via le site du SNDS (Système National des Données de Santé). Les données du **PMSI** font déjà l'objet d'études de pratiques au sein même des hôpitaux (études de revue de pratique, évaluation de l'activité) notamment via les services DIM (Département d'information médicale).

Ces données, souvent réutilisées en interne au sein de la structure sont présentes sous plusieurs formats : liées l'entrepôt de données et aux logiciels métiers choisis mais également transformés pour une partie en simultanée dans un format standardisé entre tous les établissements de santé de France qui est celui utilisé pour la transmission des données à l'AM. En effet, les données PMSI permettent aux structures de faire **rembourser leurs dépenses à l'assurance maladie** sur la base de l'activité déclarée, celle-ci se doit d'être sous un format standard afin de garantir une équité dans le traitement des données.

La réutilisation possible de ces données, hors la partie économique permet également d'assurer la **qualité des soins** et des bonnes pratiques en **réévaluant l'activité** au sein de l'établissement de santé mais est aussi parfois utilisée en **recherche** compte tenu de la richesse des données patient collectées [54].

5.3.3 Données en médecine générale

Contrairement aux données assurantielles et hospitalières, les données de médecine générale ne sont quant à elles que peu exploitées. En effet, en dehors des **réseaux de surveillance** [55] **ou observatoires** [56], implémentés à partir de données des médecins partenaires et dédiés à quelques pathologies spécifiques souvent épidémiques, ces données n'ont pour le moment que peu d'applications en dehors du domaine du soin. C'est pourquoi nous en avons fait l'objet de notre étude, la caractérisation des données structurées en médecine générale ainsi que le potentiel de réutilisation de celles-ci seront développés ci-dessous.

6 Les données de médecine générale ambulatoire

6.1 Types de données

Obligatoire depuis la loi de 1970 [57], le dossier patient est un outil indispensable pour la continuité et la traçabilité des soins de chaque médecin généraliste. Initialement format papier, on a vu se développer avec l'avènement du numérique une **Informatisation de ce Dossier Patient (DPI)**. Mis à disposition depuis 2018, le Dossier Médical Partagé (DMP) va même au-delà avec un objectif de centralisation de certaines données de santé du patient afin de les rendre accessibles par tous les professionnels de santé qui prennent en charge cette personne.

En médecine générale ambulatoire, nous retrouvons également des données structurées ainsi que non structurées. Nous traiterons ici des **données structurées** à savoir (ces informations sont inconstamment disponibles) :

- Données **d'identité** (identité patient, date de naissance, lieu de résidence, de naissance, coordonnées, …)
- Données de **couverture assurantielle**, de **facturation** (numéro sécurité sociale, mutuelle, ..)
- Données de **mouvement** (dates de consultations, téléconsultations, ...)
- Diagnostics en texte libre, ou codés selon les préférences logiciels : terminologies liées à l'éditeur, ou souvent une forme simplifiée de la CIM10 avec pour certains logiciels la possibilité d'un transcodage CIM10 ou CISP (Classification Internationale Soins Primaires) comme c'est le cas chez Cegedim [58], etc.
- **Actes** facturés et codés en CCAM
- **Résultats** d'analyses de biologie médicale éventuellement codés en LOINC
- **Médicaments** prescrits/délivrés/administrés codés en UCD ou ATC

A ces données peuvent s'ajouter des **données non structurées** : imageries, données en texte libre produits par le médecin au cours de la consultation, courriers (parfois scannés ou télétransmis) de confrères ou collègues prenant en charge le patient, etc.

En général utilisées pour assurer la continuité des soins du patient, les données produites n'ont très souvent que peu d'autres finalités. Certains logiciels médicaux permettent également d'utiliser ces données comme indicateurs épidémiologiques concernant la patientèle.

6.2 Potentiel de réutilisation, applications proposées dans la littérature

Dans la littérature Pubmed, nous ne retrouvons à ce jour que **5 articles** [59–63] associant les concepts de « réutilisation de données » et « médecine générale ». La méthodologie de la recherche est disponible en Annexe 1 de même que l'abstract de chacun des articles.

Aucun de ces articles n'est Français ou ne contient de données de médecine en France.

Bien que la plupart s'accordent sur l'intérêt de réutiliser les données de médecine générale, qui sont une **grande source de richesse et de diversité** [59,60], seule l'équipe de Monhagan et al [63] s'est intéressée de manière qualitative à l'intérêt porté

sur la réutilisation des données de médecine générale auprès de personnes effectuant des soins de ville de premier recours. Dans cette étude, pratiquement toutes les personnes interrogées s'accordaient sur l'importance de pouvoir réutiliser ces données à des fins autres que le soin.

Une équipe Australienne a réalisé une **étude de faisabilité** sur quelques données patient [61] de cabinets de médecine générale au sein de communautés médicales.

Enfin, Agrawal et al [59] proposent une **méthodologie de travail** pour la réutilisation de données et rejoignent Mohaghan et Zulutela [61,63] dans la nécessité de porter une grande attention à la **protection des données** ainsi qu'au respect du consentement des patients dans ce cadre.

6.3 Défis technologiques (interopérabilité)

Nous pouvons donc constater que la réutilisation de données issues de la médecine générale n'est que très peu décrite et réalisée dans la littérature.

Cela s'explique par plusieurs facteurs notamment :

- Le manque d'interopérabilité physique et syntaxique : pas de standards partagés entre logiciels
- Peu d'interopérabilité sémantique :
    - **Pas de terminologies normées** pour les **diagnostics** (en général une centaine de libellés simplifiés propres à chaque logiciel, potentiellement transcodables en CIM10), ni pour la biologie (libellés propres à chaque laboratoire)
    - Mais par contre les **terminologies sont identiques** pour les **actes** CCAM qui sont soumis à facturation de même que pour les prescriptions de médicaments (si réalisées avec le logiciel médical).

Cependant, de notre expérience et à la lecture de la bibliographie citée précédemment, les données qui ont un potentiel de réutilisation apparaissent très similaires, et on imagine assez facilement comment les représenter dans un schéma de données simplifié :

- Données patient (identité, sexe, date naissance)
- Mouvements (modalité et dates de contacts : consultations, visites, appels, téléconsultations, etc.)
- Résultats de biologie (date, marqueur étudié, milieu prélevé, valeur, unité, bornes d'interprétations spécifiques du laboratoire)
- Prescriptions de médicaments : dates, médicament, posologie, unité, voie d'administration
- Autres prescriptions (consultation, intervention, imagerie, vaccin, etc.)
- Diagnostics (ou motifs de recours au soin, ou symptômes) : dates (en général un par consultation), code, libellé
- Actes réalisés : date, opérateur, code ou libellé de l'acte, résultat s'ils sont disponibles

Nous pouvons remarquer que pour les données avec des terminologies non normées, il s'agit en général de paramètres **limités en nombres** et **intelligibles** par un humain permettant un *mapping* (mise au point d'une table de correspondance) relativement facilement.

Par ailleurs, sur le terrain, **15 éditeurs** [64] de logiciels médicaux se partagent **80% des professionnels de santé libéraux** (Panorama des éditeurs pour les Professionnels de Santé libéraux). Parmi ceux-ci, 4 prennent en charge plus de 50% des médecins généralistes à savoir Compugroup (31.4%), Cegedim-Crossway (15.2%), Sephira (15.1%) et Prokov(7.7%).

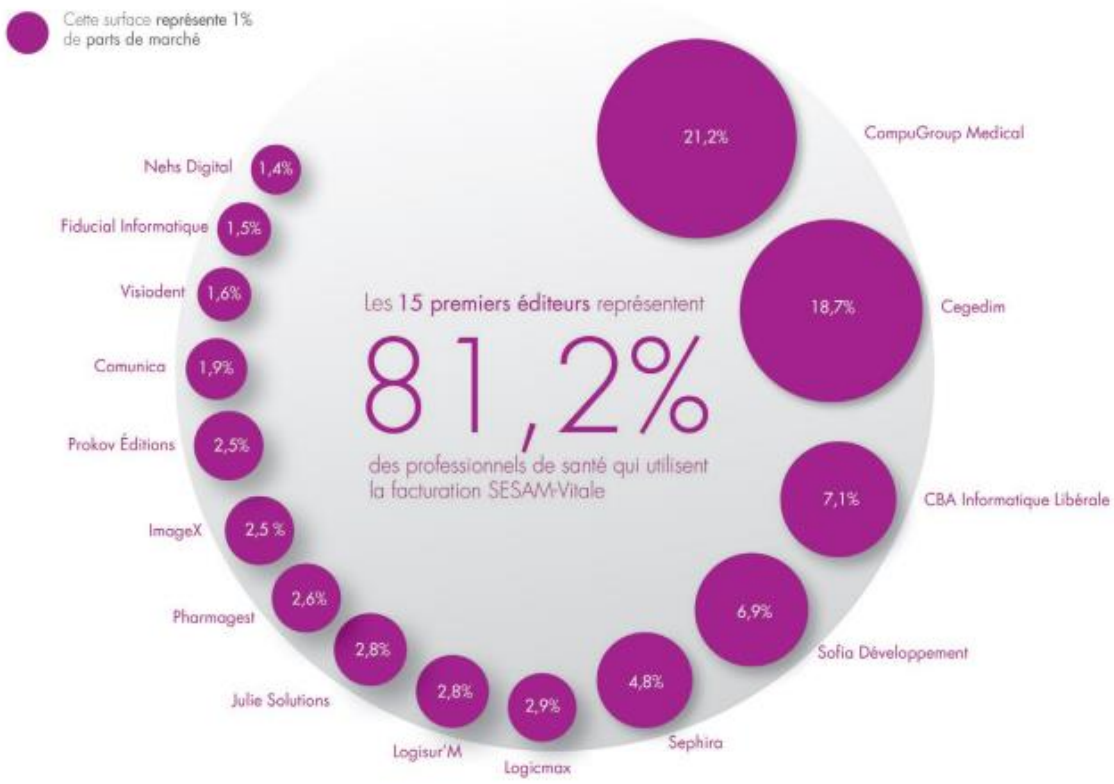

Figure 4. Répartition des parts de marché des différents logiciels métiers pour les professionnels de santé libéraux [64]

Il est donc probable que de nombreuses similitudes soient observables dans les différentes données issues de cabinets médicaux et que l'interopérabilité soit peu contraignante.

7 Objectif

L'objectif de ce travail est double :

1. « Partie 1 : **Mise en place d'un ETL** via le modèle de données commun OMOP» :
   Mettre en place une procédure d'ETL dans un cabinet de groupe de médecine générale, alimentant le standard OMOP
2. « Partie 2 : Analyse descriptive » :
   Réaliser une rapide **analyse descriptive** pour démontrer le fonctionnement de la procédure d'ETL, et illustrer le potentiel de réutilisation des données ainsi recueillies

Matériel et méthodes

1 Données étudiées

Le travail a été réalisé sur plusieurs bases de **données fictives** librement inspirées des données patients issues d'un des cabinets de médecine générale de la maison médicale de la Bourgogne à Tourcoing. Il s'agit d'un cabinet de médecine générale composé de deux médecins généralistes titulaires situé dans la ville de Tourcoing.

Dans le cadre des travaux préparatoires à la mise en place de la ***plateforme eSanté du CPER Tec'Santé***, des algorithmes de simulation de données ont été développés. S'appuyant sur des distributions et un formalisme observé dans une base de données réelles, ces algorithmes permettent de générer une base de données fictive suivant le même format, avec l'ambition de respecter les distributions ainsi que les associations statistiques entre variables. Les données étant simulées, il ne s'agit pas de données réelles, et il n'est donc **pas possible d'identifier un patient.** Les résultats de l'analyse statistique d'une telle base de données sont donc inexacts, mais ont l'ambition d'être très proches de ce qu'aurait donné l'analyse de la base de données source.

Les jeux de données étudiés représentent les différents types de données que l'on peut retrouver dans un logiciel métier. A savoir des données concernant les patients, les visites, les éventuels diagnostics posés à l'issue des visites, les traitements prescrits, les actes réalisés (ici vaccinations) et les différentes mesures de paramètres réalisables ou récupérables dans les cabinets de médecine générale.

Les données suivent un modèle de données « propriétaire ». Plus précisément, il s'agit des données telles qu'on peut les extraire sur le logiciel HelloDoc, c'est-à-dire dans un format résultant du modèle de données initialement choisi par l'éditeur, et altéré par les différentes étapes de l'extraction de données. Ce schéma de données sera décrit au début de la section Résultats (voir Réalisation de l'ETL).

2 Réalisation de l'ETL

La Figure 5 illustre l'architecture de notre procédure d'ETL et les grandes étapes qui sont réalisées et que nous allons décrire.

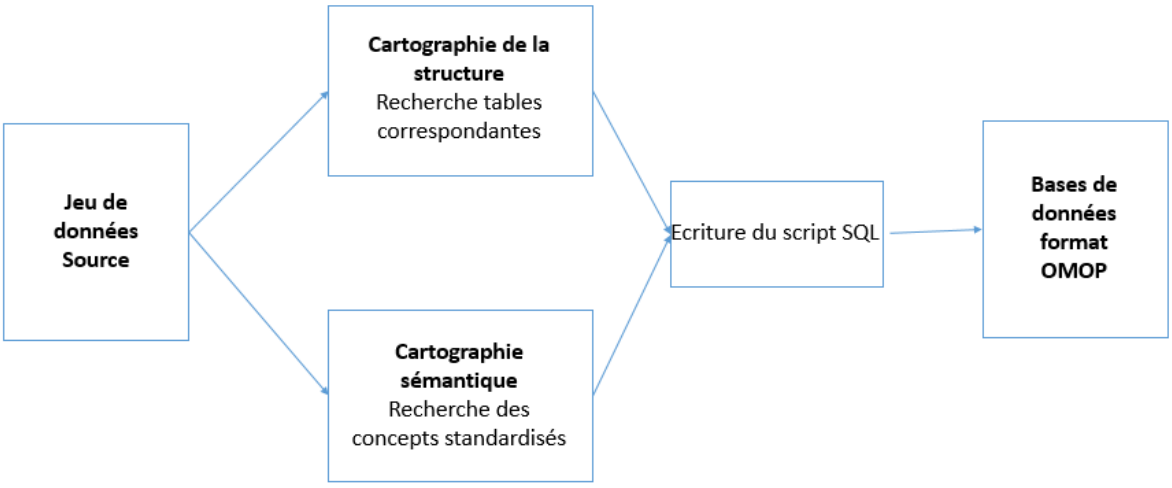

Figure 5. Schématisation du processus d'ETL

Les bases de données sont dans un premier temps importées dans PostgreSQL (Version 15) [65], un système de gestion de bases de données relationnelles *open source* et gratuit pour ce type d'utilisation.

La transformation en modèle de données commun vise le modèle OMOP dans sa version 5 (dernière version pour laquelle les données de vocabulaire sont disponibles).

2.1 Cartographie syntaxique ou structurelle

Le **modèle OMOP** définit **13 tables de données cliniques standardisées** (Figure 3) ainsi que 12 tables en lien avec des concepts de vocabulaire standardisées. Il comporte également des tables de données de coût, en lien avec le système de santé, les éléments dérivés standardisés ou encore les métadonnées standardisées bien que ces éléments ne constituent pas l'objet principal de notre étude. La documentation concernant les différentes tables OMOP est disponible via sur le site de l'OHDSI [31] et nous servira de base pour la compréhension de notre cartographie.

Notre jeu de données simulées contient quant à lui ***7 tables de données cliniques***.

Afin de pouvoir mettre en correspondance ces deux entités et réaliser leur cartographie, nous allons dans un premier temps identifier pour chacun de nos jeux de données le type d'information qu'il contient (données chiffrées, textuelles, …). Après avoir parcouru chaque tableau, le travail a été de choisir un équivalent dans le modèle OMOP. Les tables du modèle OMOP et son contenu étant bien documentés, l'enjeu sera essentiellement de bien **identifier** la typologie des données dans les bases fictives. En cas de doute, l'aide via le forum de la communauté OMOP ou le groupe de travail THEMIS reprenant en diverses conventions de cartographies établies avec la communauté seront consultés [66].

L'objectif de ce travail n'étant pas de réaliser une étude épidémiologique (donc centrée sur la personne), les tables seront par la suite analysées chacune par rapport à l'entité qu'elle représente (ex : des personnes pour la table de patients, des mesures pour la table de biométrie, etc.). Les fréquences observées ne tiendront donc pas compte du fait que certains patients ont plus de mesures que d'autres, et ne devront pas être commentées en termes épidémiologiques.

La cartographie est ensuite réalisée via PostgreSQL pour réaliser dans un premier temps l'extraction (*Extract*) puis les transformation (*Transform*) nécessaires et enfin le l'implémentation des données (*Loading*) dans la table OMOP souhaitée.

Les **outils OMOP WhiteRabbit** [67] et **RabbitInAHat** [68] sont utilisés au cours du processus afin de réaliser une présentation graphique des cartographies ainsi qu'une base pour les scripts SQL qui seront par la suite modifiés et adaptés aux besoins de chaque bas OMOP.

```
--- Extraction des données

DROP TABLE IF EXISTS etl_person_extract
create table if not exists etl_person_extract
as
select "No dossier" as person_source_value
             , "Date de naissance" as birth_date
             , Sexe as gender_source_value
from patient p
group by "No dossier", "Date de naissance", Sexe ;



--- Transformation et chargement des données

drop table if exists etl_person_tr ;
create table if not exists etl_person_tr
      as
select person_source_value as person_id
             , CASE
--- Ici insertion direct du gender_concept_id (car 3 codes)
             WHEN gender_source_value = "Masculin" THEN 8507 -- male
             WHEN gender_source_value = "Féminin" THEN 8532 -- female
             ELSE 8551 -- unknown
             END AS gender_concept_id
             , extract(year from birth_date) as year_of_birth
             , extract(month from birth_date) as month_of_birth
             , extract(day from birth_date) as day_of_birth
             , birth_date as birth_datetime
             , null::date as death_datetime -- todo
             , 0::integer as race_concept_id
             , 0::integer as ethnicity_concept_id
             , 0::integer as location_id -- todo
             , 0::integer as provider_id
             , person_source_value
             , gender_source_value
             , 0::integer as race_source_value
             , 0::integer as race_source_concept_id
             , 0::integer as ethnicity_source_value
             , 0::integer as ethnicity_source_concept_id
from etl_person_extract epe

--- Sinon jointure pour relier aux concepts, ici pour l'exemple le code présenté
ci-dessus mais non utilisé dans ce cas

leftjoin omop.concept c on epe.gender_source_value = c.concept_code
                        and c.domain_id= « Gender »
;
```

Figure 6. Structure des scripts SQL permettant la cartographie, exemple de la table patient

Les scripts SQL incluront toutes les informations requises pour qualifier table OMOP, cependant si aucune des informations de la table source ne correspond, alors les données n'y seront pas implémentées. A contrario, les colonnes de la table source ne

trouvant pas de correspondance dans la table OMOP seront exclues dès l'extraction des données.

Nous présenterons dans les **résultats la cartographie** réalisée à partir des bases de données ainsi que la correspondance entre les informations présentes dans les deux bases.

2.2 Cartographie sémantique

En parallèle de la cartographie syntaxique, les données de chaque jeu de données de la base fictive seront analysées afin de réaliser une cartographie sémantique.

Le modèle de données commun OMOP reposant sur **un vocabulaire standardisé**, il est nécessaire, si l'on veut transformer nos données de pouvoir également avoir accès à ces données standardisées.

L'outil Athena (CDM version 5) [69], compile et met à disposition les différentes bases de vocabulaire standardisées existantes reprenant plusieurs terminologies notamment SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) [70], LOINC (Logical Observation Identifiers, Names and Codes pour l'expression des résultats de laboratoire) [47] ou encore RxNorm [71] (pour les médicaments aux Etats Unis, avec l'existence de RxNorm Extension incluant certains des médicaments disponibles à dans d'autres pays). Ces bases internationales reconnues permettent de favoriser l'interopérabilité pour divers médicaux.

Cependant l'outil Athena comprenant majoritairement des bases de terminologies internationales, ces dernières ne comprennent pratiquement pas terminologies en Français (excepté la CIM 10 [45]) ce qui va nécessiter une adaptation et une traduction pour nos terminologies locales.

Une fois les bases de vocabulaire standardisées chargées dans **l'outil Usagi** [72], une première cartographie sémantique est réalisée sur les données et l'utilisateur doit par la suite confirmer ou infirmer le choix de « mapping » réalisé par l'application. Il s'agit d'une relecture manuelle qui doit par la suite être réitérée avec une recherche pour chaque terme non associé à un concept standard via la fonction de recherche de l'outil Athena afin de déterminer si d'autres bases de vocabulaire standardisées pourraient correspondre.

A l'issue de ce travail on recensera alors le volume global d'entités pour chaque tables (nombre de différents médicaments ou de diagnostics par exemple) avec leur fréquence. Puis une normalisation sera réalisée, après ***traduction*** des termes ***en Anglais***, sur les 100 entités avec la fréquence la plus élevée. Enfin, une révision manuelle de cette cartographie sera réalisée et présentée sous forme d'un tableau reprenant le volume d'entité, le volume d'entité cartographiées parmi celles avec la fréquence la plus élevée, le volume d'entités pour lesquelles la cartographie a échoué en nombre et en proportion.

Une fois la cartographie sémantique et l'association entre les terminologies des jeux de données source et celles standardisées du modèle OMOP, les concepts normalisés retrouvés seront également ***implémentés dans le script SQL***. L'ensemble des scripts ne sera pas présenté ici mais un exemple sera disponible en annexe.

2.3 Référencement dans la base OMOP

Habituellement, lorsqu'une équipe réaliser le *mapping* d'une de ses bases de données vers le format OMOP, elle enregistre ce procédé sur le site web du consortium OMOP, pour permettre la documentation du travail réalisé et le partage des procédés développés [31]. Le jeu de données initial de notre étude étant un jeu de données fictif, le travail fourni sur la base et notamment les différents résultats issus de sa transformation ne seront pas partagés sur le site web du consortium OMOP.

3 Analyse descriptive

3.1 Design

Il s'agit d'une étude observationnelle, descriptive, réalisée de manière systématique sur les données fictives transformées et implémentées dans le modèle OMOP. Ces données sont stockées dans les différentes tables ne pouvant être reliées que par les numéros de dossier (numéros de patients). Chaque table correspond à une entité (individu statistique), qui peut correspondre à la personne physique, ou à un concept potentiellement répété qui lui est relatif.

3.2 Analyse statistique

Pour chaque jeu de données, après la transformation et l'implémentation dans le modèle OMOP, les différentes informations seront décrites de la manière suivante, pour chaque colonne de la table OMOP contenant des informations :

- Par **concept standardisé** (ou l'absence de concept) pour lequel seront précisés :
    - La ou les entités cartographies avec une **terminologie correspondante**
    - La **fréquence** de chaque entité cartographie ainsi que celle, globale du concept standardisé
- Pour les 20 concepts les plus fréquemment retrouvés de chaque jeu de données

Une **représentation graphique** de la distribution des différentes entités pourra également réalisée.

Les variables quantitatives sont agrégées en présentant leur **moyenne et leur écart type (SD)** en cas de distribution symétrique, ou de **médiane et ses quartiles** en cas de distribution asymétrique. Les variables qualitatives ou binaires sont présentées en listant, pour toutes ou certaines modalités, leur **effectif brut et leur pourcentage**.

Aucun test statistique n'est réalisé, aucun intervalle de confiance n'est calculé.

Pour l'ensemble des résultats on nommera et considérera comme étant une entité : les terminologies des variables issues des jeux de données sources agrégées ou non ; et comme un concept les terminologies standardisées issues des bases de vocabulaire normalisées.

3.3 Logiciels de travail

Les logiciels utilisés pour les processus d'extraction, transformation et modification des données sont **PostgreSQL** (version 15)[65], **R et Rstudio** (version 2023.06.2+561).

Les outils fournis par l'OHDSI afin de nous accompagner dans l'ETL au format OMOP sont WhiteRabbit [67], RabbitInAHat [68] pour nous aider dans les présentations graphiques de la cartographie, et Usagi [72] couplé à l'outil ATHENA [69] pour nous aider à identifier et à cartographier tous les concepts via le vocabulaire standardisé. Enfin, les résultats du groupe de travail THEMIS [66] ont également été utilisés afin d'aider à la réflexion lors de la construction de la cartographie.

4 Respect de la réglementation en matière de protection des données

La base de données fournie étant générée à partir de données simulées, il ne s'agit ici ni d'une recherche impliquant la personne humaine ni d'une étude sur données réelles. L'utilisation du jeu de données ne nécessite pas de demande d'autorisation CNIL ni de Comité de Protection des Personnes.

5 Informations quant aux terminologies utilisées dans les résultats

On considérera comme étant un(e) :

- **Enregistrement** : chaque ligne issue de la base de donnée source correspondant à une ligne d'information (un numéro de dossier avec une date avec un diagnostic…)
- **Entité** : idée générale, souvent libellé associé à une donnée de la table source (diagnostic ou nom d'un traitement, d'un résultat de biologie par exemple)
- **Concept** : terminologie standardisée de la table cible OMOP associée à une ou plusieurs entités à l'issue de la cartographie (exprimée sous forme de nombre)

Résultats

1 Réalisation de l'ETL

1.1 Interopérabilité Syntaxique (structure)

Après lecture des différents jeux de données ainsi que des informations qu'ils contenaient, une cartographie globale a été réalisée permettant de mettre en lien nos tables avec les tables correspondant dans le modèle de données commun OMOP. Les **7 jeux de données source** disponibles ont pu trouver une correspondance avec **5 tables du modèle OMOP**, cartographie présentée dans la Figure 2.

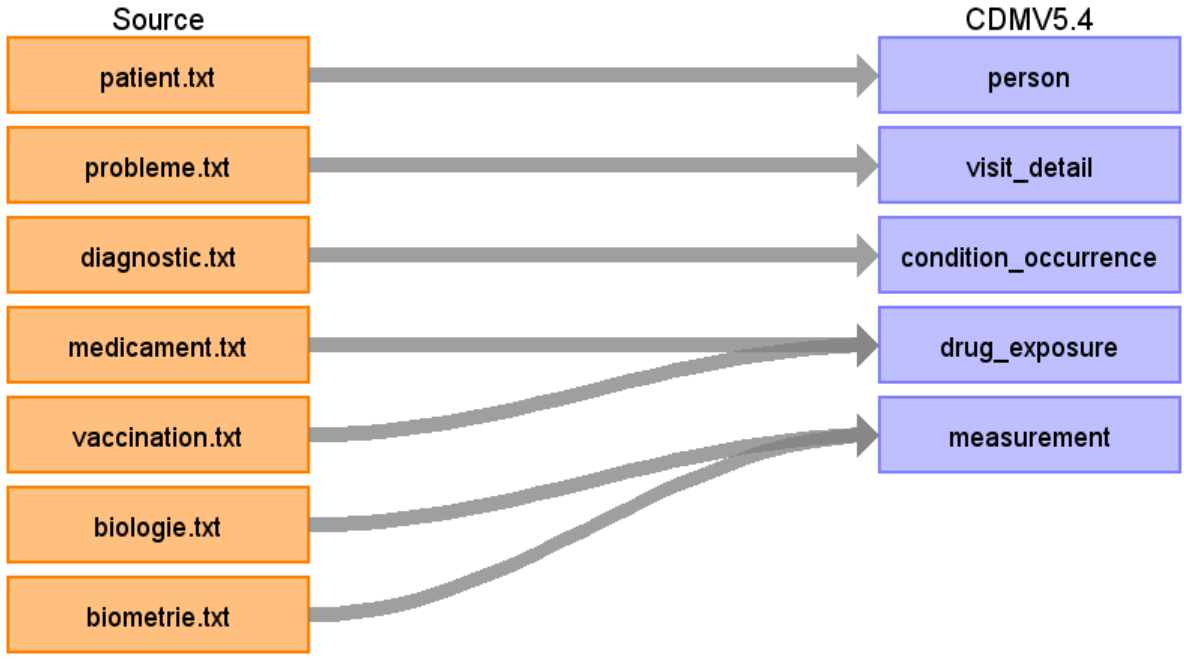

Figure 7. Cartographie de l'ensemble des données source au modèle OMOP

Le modèle final présente un nombre de tables inférieur au jeu de données sources ce qui s'explique par le fait que nos jeux de données contenaient un type d'information par table alors que le modèle commun se veut plus généraliste et ses bases de données moins spécifiques.

La table **OMOP médicaments (*drug_exposure***) par exemple regroupe deux tables du jeu de données à savoir la table « médicament » (et donc des **données de prescription**) ainsi que la table « vaccination » (et donc des données à la fois de prescription mais aussi **d'administration**). Cette table regroupant tous ces différents types d'informations.

De la même façon, la table OMOP **mesures (*measurment*)** rassemble **toutes les mesures** (quantitatives ou catégorielles) effectuées sur un patient obtenues à l'issue d'un examen, ou d'un test réalisé sur les personnes. Cela pouvant aussi bien correspondre à un résultat de mesure pris par le praticien au cabinet (pression artérielle, taille) que via un appareil (Holter, ECG, …) ou encore même un résultat de laboratoire (kaliémie, etc.).

Une fois, la cartographie globale effectuée, celle-ci a été précisée pour chaque jeu de données source en lien avec les données du modèle OMOP. La figure ci-dessous présente d'ailleurs la mise en relation du jeu de données patient selon les informations contenues dans le jeu source et nécessaires dans le modèle de données commun.

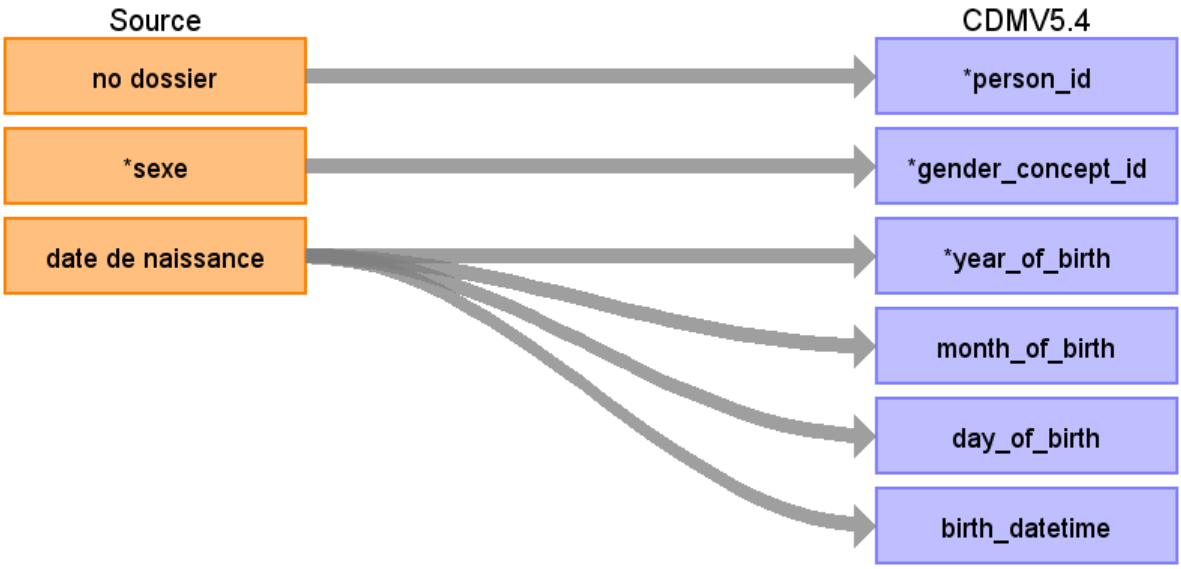

Figure 8. Cartographie de la table source patient vers la table OMOP personne

Pour compléter cette Figure 8, le Tableau 1 suivant apporte les précisions quant à **l'association des données issues de chaque table** (source et OMOP). Celles non présentées sur la Figure 8 étant celles pour lesquelles la cartographie au format standardisé n'a pas pu trouver de correspondance et ne sont donc pas exploitées pour le fichier source ni disponibles pour le modèle transformé. Les **informations en vert sont celles qui ont pu être associées** dans les deux tables, et celles marquées d'un astérisque (*) sont celles définies comme nécessaires pour la création de la table OMOP dans son format le plus aboutit.

Les figures et tableaux représentant la cartographie des autres tables de données ne sont pas présentées ici mais sont disponibles dans l'Annexe 2.

Le Tableau 2 résume la cartographie des différentes tables de manière succincte ; on observe une **proportion non négligeable de données « perdues »** (parfois des colonnes vides) initialement contenues dans les tables source. Concernant les données dites « requises » dans les tables transformées au modèle OMOP, on observe la perte de 1 à 2 colonnes d'informations dont pour 6 tables sur 7, l'absence **d'identifiant de visite**.

Celui-ci a pu être recréé pour chacune des bases concernées en concaténant le numéro de dossier patient et la date de consultation mais n'est toutefois, en l'état, **pas transposable entre les différentes bases de données** car généré via les informations présentes dans chaque table. En effet, les informations relatives à un diagnostic posé en consultation par exemple (donc à un numéro de patient et une date de consultation) ne pourront pas être rattachées aux résultats de biologie (associés à un numéro de patient également mais à une date de réalisation du bilan en dehors de toute consultation).

Table source (6 colonnes)	Table OMOP (18 colonnes dont 5 requises)
no dossier	person_id*
Sexe	gender_concept_id*
date de naissance	year_of_birth*
date de naissance	month_of_birth
date de naissance	day_of_birth
date de naissance	birth_datetime
ALD	
Code régime	
	race_concept_id*
	ethnicity_concept_id*
	location_id
	provider_id
	care_site_id
	person_source_value
	gender_source_value
	gender_source_concept_id
	race_source_value
	race_source_concept_id
	ethnicity_source_value
	ethnicity_source_concept_id

Tableau 1. Cartographie complète de la table source patient vers la table OMOP person, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de la table

La Figure 9 nous montre le code utilisé pour la table diagnostic et adapté dans chaque base de données.

```
-- Création d'une colonne reprenant les informations inhérentes à chaque visite : l'identifiant patient "No dossier" et la date de consultation

ALTER TABLE diagnostic
ADD COLUMN visit_id varchar(500)

INSERT INTO diagnostic(visit_id)
    SELECT ("No dossier" || '' || "Date de consultation")
    FROM diagnostic d
```

Figure 9. Structure des scripts SQL permettant la création de la colonne d'identifiant de la visite, exemple de la table diagnostic

On peut noter que, dans les bases de données décrivant des **séjours hospitaliers**, le **séjour** est considéré comme **l'entité principale**, et que quasiment toutes les informations (hormis celles de la personne) sont rattachées au séjour. Inversement, dans le **champ ambulatoire**, la consultation n'est qu'une information parmi d'autres et toutes les informations sont **directement rattachées au patient**. Ainsi par exemple, les résultats d'analyse de biologie médicale sont télétransmis, et correspondent à des prélèvements réalisés en ville, en-dehors de toute consultation au cabinet.

Table Source	Table OMOP	Colonnes non cartographiées de la source / Nombre de colonnes total	Type d'informations contenues	Colonnes requises non cartographiées du modèle de données commun / Nombre de colonnes requises total	Type d'informations contenues
Patient	Personne (PERSON)	3/5	ALD (oui/non), Régime (code)	2/5	Concept d'ethnicité, Concept de race
Problème	Visite détail (VISIT DETAIL)	1/4	Date de début (vide)	3/6	Concept de visite (consultation, domicile, hospitalisation, ...) Date de fin de visite **Identifiant visite**
Diagnostic	Diagnostic (CONDITION_OCCURENCE)	2/5	Code (code propre du cabinet) Caractéristique (précisions si information)	2/5	Type de concept selon la pathologie **Identifiant visite**
Médicament	Médicaments (DRUG_EXPOSURE)	9/12	Date de consultation Motif de prescription Code CIP Coût ALD Note Code UCD, Code ATC, Code CIS	2/5	Type de concept selon le traitement **Identifiant visite**
Vaccin	Médicaments (DRUG_EXPOSURE)	8/15	Code CIP,Rappel Zone d'injection Date de lecture, Lecture du test, Test positif Résultat Commentaires	2/5	Type de concept selon le traitement **Identifiant visite**
Biologie	Mesure (MEASUREMENT)	1/6	Valeur Entière (vide)	¼	**identifiant visite**
Biométrie	Mesure (MEASUREMENT)	1/6	Valeur Entière (vide)	¼	**identifiant visite**

Tableau 2.Correspondance en termes de perte d'information de la cartographie structurelle entre les données sources vs OMOP
Les colonnes identifiant de visite et donc marquées en gras ont été recrées manuellement après cartographie mais ne sont pas transposables

1.2 Interopérabilité Sémantique (vocabulaire)

Le Tableau 3 présente la perte d'information liée à la cartographie sémantique et permettant la transformation de notre base de données source contenant beaucoup d'informations dont du texte libre, en une table standardisée. A noter que les données cartographiées présentées pour la cartographie concernent uniquement les 100 entités avec la plus grande fréquence pour chaque table et ne sont donc pas exhaustives.

On observe une **perte de données comprise entre 0%** pour la table patient (concept de genre uniquement) et **49%** pour la table OMOP médicaments après la première cartographie automatisée associée à une relecture et modification manuelle. Une correspondance et un mapping complet n'ont pas pu être atteints pour nos différentes bases de données notamment celles comprenant des traitements ou vaccinations.

A noter que les résultats présentés pour la cartographie initiale tiennent compte de la **traduction** des différentes **terminologies en anglais** à l'exception de la table diagnostic, pour laquelle nous avons travaillé en première intention sur les terminologies françaises (la CIM10 étant disponible dans les terminologies standardisées de la base Athena), ainsi que des tables médicaments et vaccination puis sur la traduction anglaise pour la relecture.

Pour certains jeux de données, des choix ont dû être faits concernant la cartographie. Comme c'est le cas par exemple pour la table source biométrie associée à la table OMOP mesure. Le Tableau 4 en est un exemple, pour lequel nous avons décidé de traiter l'ajout de données informatives comme une information associée au concept et non comme étant le concept de la mesure. Ainsi la **pression artérielle** mesurée au bras droit et celle mesurée au bras gauche (précisions dans la terminologie de l'entité qui relevait plutôt d'informations ajoutées ne modifiant pas la valeur de la mesure réalisée ni sa significativité) ont pu être cartographiées comme **un seul et même concept.** Alors que les entités associées à la mesure de la pression artérielle pour lesquelles plusieurs concepts différents existent dans le vocabulaire standard associé à la table mesure ont été **associées à chacun de ces différents concepts** (exemple : pression artérielle systolique, diastolique, mesurée sur patient est assis ou couché).

Un concept lié à la latéralité existe dans la table OMOP procédure cependant celle-ci ne prenant pas en compte les paramètres mesurés ou mesurables à l'issue d'un geste ou d'une intervention nous avions choisi de lier de manière syntaxique la table source biométrie à la table mesures qui met en avant les résultats obtenus.

Lors de la validation de la cartographie et des termes rencontrés, un grand soin a été porté à avoir un **concept le plus proche possible de la typologie de l'information utilisée**. Ainsi, le concept choisi pour la mesure du débit expiratoire de pointe aura été celle liée au *Peak Expiratory flow rate* (donc une valeur, ce qui correspond à nos données) et non au concept de *Peak Flow Expiratory measurement* (qui permet plus de qualifier la réalisation du geste mais sans intégrer la mesure. Ce concept a été ici assimilé à une procédure pour laquelle on pourrait répondre par oui ou non quant à sa réalisation ; contrairement au taux pour lequel on attend la valeur de la mesure, donnée fournie dans notre table source.

Table OMOP	Table	Nombre d'enregistrements	Nombre total d'entités référenciées	Cartographiés des 100 entités les plus fréquentes	Nombre de concepts à l'issue du mapping	Terminologie automatiquement référencée
Personnes (*Person*)	Patient	15000	2	100 %	2	OMOP Gender
Visites (*Visit Detail*)	Problème	14 711	2036	97%	83	LOINC SNOMED
Diagnostics (*Condition Occurrence*)	Diagnostic	241 215	4908	98%	85	CIM 10 LOINC SNOMED
Expositions à des médicaments (*Drug Exposure*)	Médicaments	780 220	12842	51%	45	SNOMED Rx Norm (et Rx Norm Extension)
	Vaccins	12002	129	54%	33	Rx Norm (et son Extension) SNOMED
Mesures (*Measurement*)	Biologie	376 840	375	74%	65	SNOMED LOINC
	Biométrie	384 852	37	81 %	18	SNOMED

Tableau 3. Résultats de la cartographie standard sur les concepts. Perte de données à l'issue de de la cartographie sémantique des 100 entités les plus fréquentes

Concept OMOP	Libellé (table initiale)	Fréquence	Pourcentage
Pression artérielle (Blood Pressure) 4326744 (SNOMED)	TA Max	94181	24,39
	TA Min	93880	24,47
Pression artérielle diastolique 4154790 (SNOMED)	PAD assis BD	6	0,0
	PAD assis BG	1	0,0
	PAD couché BD	39	0,01
	PAD couché BG	44	0,01
Pression artérielle diastolique couché 4236281 (SNOMED)	PAD couché BD	39	0,01
	PAD couché BG	44	0,01
Pression artérielle systolique 4154794 (SNOMED)	PAS assis BD	6	0,0
	PAS assis BG	1	0,0
	PAS debout BD	2	0,0
	PAS debout BG	1	0,0
Pression artérielle systolique couché 4248525 (SNOMED)	PAS couché BD	38	0,01
	PAS couché BG	95	0,02

Tableau 4. Table de cartographie pour les enregistrements de pression artérielle de la table biometrie – Focus sur les concepts de pression artérielle OMOP mesures

2 Analyse descriptive

2.1 Table Patient – *OMOP personne*

On a recensé **15 000 identifiants** patients différents dans la base de données. Les patients étaient nés entre 1900 et 2020 avec une **médiane en 1988** et un **intervalle interquartile de [1972 ; 2004].** L'année de naissance est manquante dans 175 cas sur 15 000. Les dates de décès n'étant pas disponibles, l'âge n'a pas été calculé.

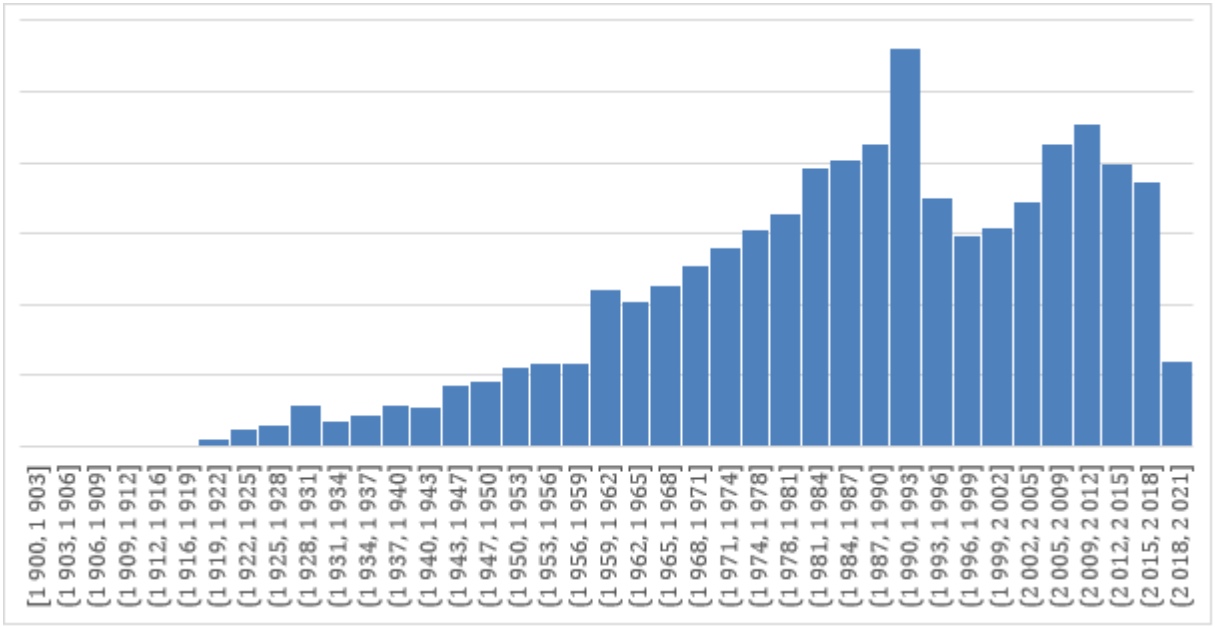

Figure 10. Répartition graphique de l'année de naissance des patients

Ci-dessus, une représentation graphique de la répartition du nombre de patients par année de naissance. Le **3ème quartile débute en 2004** donc pour un jeu simulant des données des années 2000 à 2020 environ, cela correspond à **au moins ¼ des consultations** réalisées s'intéressant à une **population pédiatrique.**

Concept OMOP	Entité (table initiale)	Fréquence	Pourcentage
Sexe			
8532	Féminin	7854	52.7 %
8507	Masculin	7049	47.3 %
8551	Inconnu	97	
Race	-	-	-
Ethnicité	-	-	-

Tableau 5. Caractéristiques sociales issues de la table OMOP personne

Dans notre jeu de données, on comptait **7049 hommes** soit **47.3%** du nombre de sujets.

Aucune donnée en France ne prenant en compte les considérations raciales ou ethniques n'est recueillie en routine et ces deux colonnes ne contiennent aucune donnée.

2.2 Table Problème – OMOP *visit detail*

La table problème reprenait les principaux motifs évoqués par le patient lors de la prise de rendez-vous. Il ne s'agit pas de diagnostic à proprement parlé. Ces données n'ont pas été renseignées de manière systématique ce qui explique des effectifs plus faibles que pour les autres tables.

Sur les 14 771 enregistrements, on pouvait identifier 5 315 patients distincts et **11 603 visites** soit une moyenne de **1.27 problèmes évoqués** pour chaque consultation.

Concept OMOP	Libellé initial	Fréquence d'enregistrements	Pourcentage
4299535	Grossesse	771	5,2
37109809	Diagnostic non encore établi	458	3,1
320128	Hypertension artérielle (HTA) essentielle	401	2,7
42538827	Contraception par oestro-progestatif	329	2,2
316866	Hypertension artérielle	283	1,9
4085799	A la demande du patient	264	1,8
201826	Diabète de type 2	257	1,7
4191716	Syndrome dépressif	247	1,7
4189941	En prévention de Rachitisme	231	1,6
42538827	Contraception Oestro-Progestative	179	1,2
4029305	Hypercholestérolémie	178	1,2
201254	Diabète type II (Diabète gras ou non-insulino-dépendant ou DNID)	177	1,2
4103574	Dépression	173	1,2
45769852	Suivi gynécologique	165	1,1
40766945	Tabagisme	161	1,1
4318982	Autres troubles de la vision	153	1,0
436962	Insomnie	147	1,0
441542	Anxiété	143	1,0
44803659	Frottis vaginal	142	1,0
4338031	Syndrome anxio-dépressif chronique	140	1,0

Tableau 6. Table de cartographie sémantique des 20 premières entités de la table probleme – OMOP visit detail

La cartographie de la table problème s'est révélée relativement simple en terme de recherche de vocabulaire associé et a été réalisée **pour 97 entités /100.** La principale difficulté rencontrée était liée à la cartographie structurelle puisque les informations concernant le **type de visite étaient manquantes** et n'ont donc pas pu être liées à ces concepts (pas de précisions s'il s'agissait de consultations ou de visites à domicile par exemple).

Ces 97 entités mappées en 83 concepts différents représentent 61,36% des 14 771 enregistrements.

2.3 Table Diagnostic – OMOP diagnostic

La table diagnostic reprend le ou les diagnostics principaux établis par le médecin généraliste à l'issue de la consultation.

Sur les 240 215 enregistrements, on pouvait identifier 12 547 patients différents et 169 523 visites soit une moyenne de **13,51 visites** par patient.

Concept OMOP	Libellé initial	Fréquence d'enregistrements	Pourcentage
321318	Angine aigue (avec test de diagnostic rapide ou TDR négatif)	23703	9,8
Non mappé	A la demande du patient	15751	6,5
4144375	Vaccination	8781	3,6
4189941	Prévention du rachitisme	7531	3,1
4101468	Gastro-entérite	6357	2,6
380731	Otite aigue	4849	2,0
4270490	Rhinopharyngite	4746	2,0
4016388	Renouvellement d'ordonnance*	4383	1,8
4283893	Sinusite	4160	1,7
321318	Angine aigue (avec test de diagnostic rapide ou TDR positif)	3683	1,5
4057824	Constipation	3413	1,4
194133	Lumbago	2871	1,2
4270490	Rhinopharyngite aigue	2369	1,0
4208264	Contracture musculaire	2143	0,9
4101468	Gastro-entérite virale	2028	0,8
257007	Allergie naso-sinusienne	1925	0,8
321318	Angine érythémateuse	1907	0,8
321318	Angine aigue	1849	0,8
4147145	Tendinite	1792	0,7
437113	Asthénie	1766	0,7

Tableau 7. Table de cartographie sémantique des 20 premières entités de la table diagnostic – OMOP diagnostic

La principale difficulté lors de la recherche de correspondance a été de pouvoir combiner la grande richesse descriptive de notre jeu de données fictif avec les concepts standardisés de vocabulaire. En effet, dans la section diagnostic (condition occurrence), il n'y a que peu de place pour les observations associées. Ainsi, **l'angine,** par exemple, présente 4 fois dans les 20 premières occurrences a toujours été associée au **même concept 321318**, puisque dans les diagnostics la distinction concernant la réalisation d'un test associé (une mesure donc) ou son caractère érythémateux (une observation) ne sont pas décrites. A elle seule, en cumulant les concepts, elle **représente 12.9 % des diagnostics enregistrés** dans le jeu de données.

2.4 Table Médicament – *OMOP médicament*

La table médicament reprend les traitements prescrits à l'issue des consultations.

Sur les 780 220 enregistrements, on pouvait identifier 12 696 patients différents et 210 828 visites soit une moyenne de **3,7 médicaments prescrits** par consultation.

Concept OMOP	Libellé initial	Fréquence d'enregistrements	Pourcentage
43183424	PIVALONE SUSP NAS PULV 10ML	29303	3,76
40952951	DOLIPRANE 1 000MG CPR 8	15954	2,04
43179126	SPIFEN 400MG CPR 20	13097	1,68
1125315	DOLIPRANE 2,4% SUSP BUV 100ML	12849	1,65
Non mappé	FLECTOR 1% GEL FL 100G	11417	1,46
Non mappé	MAXILASE 200U/ML SP 125ML	11178	1,43
Non mappé	PHARYNDOL ADULTE SPRAY GORGE 30ML	10069	1,29
40952951	DAFALGAN 1G CPR 8	8758	1,12
19095164	ZYMAD 80 000UI AMP BUV 2ML 1	7400	0,95
23497631	SPASFON CPR 30	6304	0,81
Non mappé	RHINOTROPHYL 3 % Pulvérisations nasales Flacon de 20 ml	6127	0,79
19041813	THIOCOLCHICOSIDE 4MG ACTAVIS CPR24	5494	0,70
Non mappé	RHINOTROPHYL PULV NAS 20ML	5229	0,67
19033719	ADVIL 20MG/ML ENF SUSP BUV 200ML	5070	0,65
19018403	TIORFANOR 175MG CPR 12	4882	0,63
3474419	KARDEGIC 75MG SACHET 30	4743	0,61
Non mappé	PHARYNDOL ENFANT SPRAY GORGE 20ML	4193	0,54
40952951	PARACETAMOL 1G ALMUS CPR 8	4147	0,53
Non mappé	VELMETIA 50MG/1 000MG CPR 56	4000	0,51
19033719	ADVILMED 20MG/ML ENF SUSP BUV200ML	3936	0,50

Tableau 8. Table de cartographie sémantique des 20 premières entités de la table medicament- OMOP medicament

La principale difficulté lors de la recherche de correspondance a été, malgré la non-nécessité systématique de traduction vers la langue anglaise des entités, de pouvoir **identifier des correspondances avec les concepts en termes de posologie et principe actif**. Ainsi, pour de nombreux médicaments résultant du mélange de plusieurs principes actifs ou même selon certaines formes galéniques (crèmes notamment avec le Dexeryl, Flector, …) certains concepts n'ont pas pu être cartographiés. On recense un total de 51 entités mappées / 100.

2.5 Table Vaccination – *OMOP médicament*

La table vaccination compile les différents vaccins réalisés lors d'un passage au cabinet ou en visite à domicile.

Sur les 12 002 enregistrements, on pouvait identifier 3 444 patients différents.

Concept OMOP	Libellé initial	Fréquence d'enregistrements	Pourcentage
36952682	VACCIN PREVENAR13 SER A/A 0,5ML	928	7,7
Non mappé	VACCIN PRIORIX PDR+SOLV SER 0,5ML 1	848	7,1
Non mappé	VACCIN INFANRIX HEXA SER 0,5ML+2AIG 1	760	6,3
43714854	VACCIN BOOSTRIXTETRA SER 0,5ML 1	705	5,9
36269232	VACCIN NEISVAC SER 0,5ML 1	652	5,4
43589219	VACCIN VAXIGRIP SER 0,5ML 1	629	5,2
43661170	VACCIN REPEVAX SER 0,5ML 1	550	4,6
528323	VACCIN ENGERIX B 10MCG/0,5ML SER 1	487	4,1
Non mappé	VACCIN INFLUVAC SER 0,5ML 1	418	3,5
Non mappé	VACCIN INFANRIXQUINTA INJ FL+SRG 1	351	2,9
40857355	VACCIN MENINGITEC 0,5ML SER 1	311	2,6
528323	VACCIN ENGERIX B 20MCG/1ML SER 1	303	2,5
43661170	VACCIN REPEVAX SER S/A 0,5ML 1	289	2,4
36952682	VACCIN PREVENAR SER A/A 0,5ML	288	2,4
Non mappé	VACCIN INFANRIXQUINTA SER 0,5ML 1	270	2,2
528323	VACCIN ENGERIX B 10Y SRG.BACK.0.5ML 1	260	2,2
35408091	VACCIN HEXYON SER IM 0,5ML 1	233	1,9
36952682	VACCIN PREVENAR Injectable Boîte de 1 Flacon (+ seringue + 2 aiguilles) de ½ ml	233	1,9
43291286	VACCIN BCG SSI PDR+SOLV+SER 1ML 1	232	1,9
36405748	VACCIN GARDASIL SER+2 AIG 0,5ML 1	230	1,9

Tableau 9. Table de cartographie sémantique des 20 premières entités de la table vaccination - OMOP vaccination

De la même façon que pour la table médicaments, la difficulté lors de la recherche de correspondance a été, malgré la non-nécessité systématique de traduction vers la langue anglaise des entités, de pouvoir identifier des correspondances avec les concepts en termes de **posologie et principe actif.** De nombreux vaccins associent

plusieurs principes actifs et trouvent difficilement une correspondance ; ainsi seuls 54 entités /100 ont été mappés.

2.6 Table Biologie – OMOP Mesure

La table biologie reprend les principaux résultats de biologies reçus par le médecin généraliste et implémentés dans son logiciel patient. Ceux-ci n'étant pas rattachables à une visite mais à une date de réception des données.

Sur les 376 840 enregistrements, on pouvait identifier 4 177 patients différents et 20 245 envois soit une moyenne de **18,61 paramètres mesurées** dans chaque envoie de résultats de biologie.

Concept OMOP	Libellé initial	Fréquence d'enregistrements	Pourcentage
36310193	Plaquettes	11918	3,2
4054726	Hématies	11773	3,1
Non mappé	VGM	11509	3,1
Non mappé	TGMH	11506	3,1
45878743	Leucocytes	11505	3,1
Non mappé	Index d'anisocytose	11502	3,1
Non mappé	CGMH	9327	2,5
4156660	Glycémie à jeun	8812	2,3
37394438	Créatininémie	7716	2,0
37205200	Clairance de la créatinine	7701	2,0
4151358	Hématocrite (Ht)	7651	2,0
4015178	Hémoglobine (Hb)	7647	2,0
4261202	Urémie	7510	2,0
36308680	Polynucléaires neutrophiles %	7326	1,9
4208938	Natrémie	7244	1,9
36309122	Lymphocytes %	7228	1,9
36310855	Polynucléaires basophiles %	7228	1,9
4216098	Polynucléaires éosinophiles %	7228	1,9
36310669	Monocytes %	7227	1,9
4019545	Chlorémie	7215	1,9

Tableau 10. Table de cartographie sémantique des 20 premières entités de la table biologie - OMOP mesure

74 entités / 100 de la table de données biologie ont été cartographiées. Les principales difficultés rencontrées étaient liées à la traduction des terminologies en anglais.

Parmi les entités les plus fréquemment codées, on retrouve un nombre moyen de **plaquettes à 235 216** (sd 111 294,66) **x10^9/L** et un taux de **leucocytes moyen** à **6 348** (3 505,67**) x 10^9/L.**

Concept OMOP	Moyenne	Déviation standard
36310193 – Plaquettes	235 216,17	111 294,66
4054726 – Hématies	4 014 492, 94	1 503 507,76
Non mappé – VGM	89,46	7,43
Non mappé – TGMH	29,29	2,76
45878743 – Leucocytes	6 348, 59	3 505,67

Tableau 11. Caractéristiques des 5 concepts de la table biologie avec la prévalence la plus importante

2.7 Table Biométrie – OMOP mesure

La table biométrie reprend les principaux résultats des paramètres mesurés au cabinet. Il ne s'agissait pas ici de la prescription de ces examens mais bien du résultat de leur réalisation.

Sur les 384 852 enregistrements, on pouvait identifier 11 665 patients différents soit une moyenne de **32,99 paramètres mesurés** par patient.

Le Tableau 12 présente le résultat de la cartographie sémantique réalisée. Les 5 premières entités représentent entre 32.59% et 5.14% du nombre total de valeurs présentes dans le jeu de données.

Concept OMOP	Libellé initial	Fréquence d'enregistrements	Pourcentage
4099154	Poids	125435	32,59
4326744	TA Max	94181	24,47
4326744	TA Min	93880	24,39
607590	Taille	43405	11,28
4239411	Pouls	19767	5,14
Non mappé	Périmètre crânien	4600	1,20
4209432	Température	2118	0,55
Non mappé	MADRS	336	0,09
4121509	Tour de taille	313	0,08
45455173	SpO2	275	0,07
4248525	PAS couché BG	95	0,02
4106714	Débit expiratoire de pointe	70	0,02
Non mappé	Périmètre ombilical	70	0,02
4111665	Tour de hanches	59	0,02
4236281	PAD couché BG	44	0,01
4236281	PAD couché BD	39	0,01
4248525	PAS couché BD	38	0,01
1576151	IMC	33	0,01
4171375	Tour du cou	25	0,01
4154790	PAD debout BD	17	0,00
Non mappé	MMSE	15	0,00
4313591	Fréquence respiratoire	6	0,00
4154790	PAD assis BD	6	0,00
4154794	PAS assis BD	6	0,00
Non mappé	IPS	4	0,00
Non mappé	CO expiré	3	0,00
4154794	PAS debout BD	2	0,00
4313591	5-Mots	1	0,00
4092028	Axe QRS	1	0,00
4088512	Complexe QRS	1	0,00
Non mappé	GradePodo	1	0,00
4116637	Intervalle PQ	1	0,00
4116637	Intervalle QT	1	0,00
4154790	PAD assis BG	1	0,00
4154790	PAD debout BG	1	0,00
4154794	PAS assis BG	1	0,00
4154794	PAS debout BG	1	0,00

Tableau 12. Standardisation des données de la table biométrie au modèle OMOP. Présentation des concepts

Associé à la Figure 11 représentant la répartition du nombre de concepts selon leur fréquence cumulée, on peut voir que **97.8% des informations implémentées** dans la base de données source biométrie sont représentées **par les 5 premiers entités** présentées. On observe également le caractère anecdotique des 30 dernières entités, donc l'effectif cumulé n'atteint pas 1% de l'effectif total.

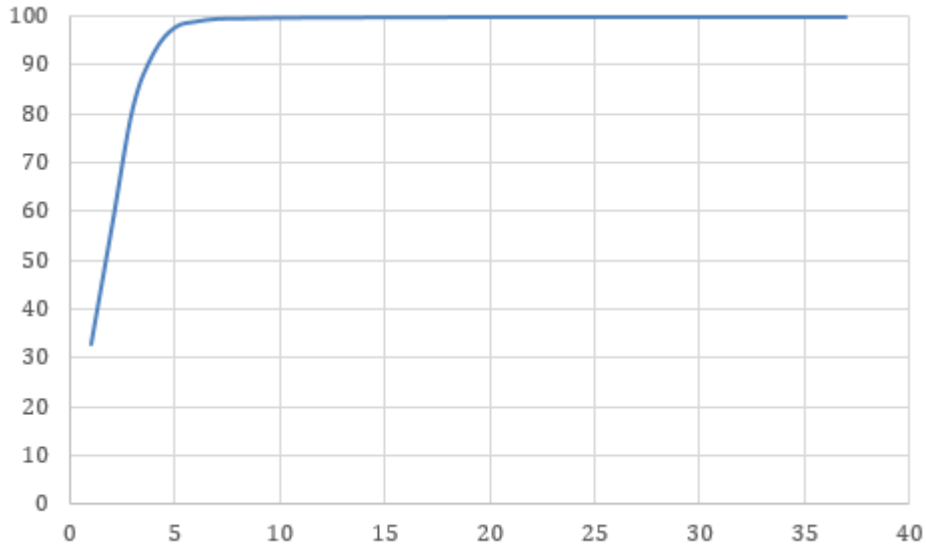

Figure 11. Fréquence cumulée de la répartition de la fréquence d'apparition des concepts OMOP selon le nombre d'entités

Si l'on s'intéresse aux données les plus représentées, on observe un **poids moyen à 58,3 kilogrammes** (sd 52,7) et une mesure de tension artérielle avec une valeur moyenne à 85,5 mmHg (sd 48,5). Pour la **taille** des individus on avait une taille **moyenne de 140,7 cm** (sd 38,8) et pour le **pouls de 78,2** (sd 13,6) battements par minute en moyenne.

Concept OMOP	Moyenne	Déviation standard
4099154 – Poids (kilogrammes)	58,3	52,7
4326744 - Pression artérielle (mmHg)	85,5	47,5
607590 - Taille (cm)	140,7	38,8
4239411 – Pouls (bpm)	78,3	13,6
Non mappé – Périmètre crânien (cm)	43,1	4,7

Tableau 13. Caractéristiques des 5 concepts de la table biométrie avec la prévalence la plus importante

Discussion

1 Principaux résultats

Dans cette étude de faisabilité, nous avons réalisé la standardisation d'une partie d'un jeu de données fictif inspiré des données d'un cabinet de médecine générale. Bien qu'ayant été réalisé sur une fraction du jeu de données sources, il s'agit là d'un résultat encourageant quant à la transposition de la méthode sur des données non fictives issues d'un ou plusieurs cabinets de médecine générale.

Cela nous a permis de comprendre que malgré la grande disparité des données et la diversité liée à leur structure initiale (notes, texte semi-structuré), de nombreuses similarités existent avec les bases de vocabulaire standardisées existantes et disponibles.

Bien qu'il s'agisse de données inspirées d'un seul cabinet, leur exhaustivité dans chaque table est assez importante pour permettre de supposer des similarités avec d'autres cabinets de médecine générale et donc une reproductibilité du travail.

2 Discussion sur la méthode d'ETL et le modèle OMOP

Le modèle OMOP a déjà fait ses preuves dans la transformation de nombreuses typologies de bases de données au niveau international [15–17,19,20] mais également en France [39]. Un grand nombre de ressources sont disponibles via la documentation de l'OHDSI et nous permet de nous accompagner dans ce parcours. Il s'agit cependant d'un **travail souvent long** et nécessitant un certain niveau de **technicité**. Kahn et al [73] soulignait par ailleurs l'importance de réaliser nombreuses vérifications pour contrôler la qualité tout au long du processus d'ETL. Le livre de l'OHDSI [74] apporte des précisions spécifiques quant aux contrôles qualité en nous précisant que le concept de preuve apporté lors de la génération des données doit pouvoir être répétable, reproductible, généralisable, robuste et calibré. A cela doit s'ajouter la possibilité de pouvoir vérifier la qualité des données (de par leur cohérence, en les utilisant dans des études, etc), leur validité clinique, celle du logiciel et de la méthode aussi bien par les outils mis à disposition par l'OHDSI que via des analyses statistiques, ou tout autre procédé permettant de s'assurer de la conformité des données aux standards, leur validité, la recherche et gestion d'erreurs, etc.

2.1 Points forts du travail

L'un des points forts de cette étude est le **fort taux de transformation** des entités du jeu de données de base en **concepts standardisés (entre 51% et 100%).** Le travail ayant été réalisé par une seule personne ici et sans pouvoir échanger sur un consensus il a parfois été décidé, faute de mieux de ne pas « mapper » certaines entités et entrainant une perte de données qui reste cependant correcte au vu du volume de données cartographiées.

L'utilisation de vocabulaire standardisé a également permis de rendre la base de données **plus lisible et plus facilement exploitable**, dans la mesure où les nombreux

diagnostics ou motifs de visite, mesures de biométrie, pouvant être recensées, selon la source, sous plusieurs dénomination (tel que nous l'avons vu pour la pression artérielle par exemple), une éventuelle requête et une étude sur données sera facilité par l'utilisation de **vocabulaire commun.**

Un autre des points forts réside dans l'absence de difficultés rencontrées dans la réalisation de la cartographie structurelle. L'essentiel des bases de donnée OMOP étant gratuites, bien documentées et en libre accès[31], le « **mapping structurel** » s'est révélé accessible avec le soutien de la communauté OMOP et de ses forums de discussion pour les points de divergence [66]. Les informations et différentes conventions établies ont pu répondre aux interrogations potentielles concernant l'association de certaines entités.

Pour finir, concernant le modèle de données commun en lui-même, l'existence de plusieurs **études** utilisant le modèle associé aux **nombreux outils** fournis par OHDSI ont permis de grandement simplifier les premières étapes du travail aussi bien en termes de cartographie que pour le recensement du vocabulaire ; faisant du modèle OMOP un modèle globalement accessible pour une personne souhaitant effectuer ce travail tel que cela a été mon cas.

2.2 Limites du travail

Le modèle OMOP étant un **modèle international**, il nous a paru difficile de trouver une correspondance pertinente pour des données dites « requises » mais non existantes en France (données d'ethnicité ou de race par exemple).

De plus, nous ne disposions pas d'un **numéro unique de séjour** ou de contact dans chaque table de données (exceptée la table des contacts). Cette particularité est propre à la médecine ambulatoire, mais tranche par rapport à la culture hospitalière, dans laquelle quasiment toutes les informations sont rattachables à un séjour, lequel séjour est rattaché au patient. Bien que cela aboutisse à la génération d'un champ « identifiant du contact » vide, nous avons pu observer que cette absence ne posait pas de problème technique dans la mise en œuvre du modèle OMOP. Une question reste cependant ouverte : les logiciels tirant parti du modèle OMOP sont-ils tous compatibles avec l'absence d'identifiant de la visite ? Par ailleurs lors de la recréation de cet identifiant de visite nous avons rapidement été confronté aux **limites inhérentes à la médecine généra**le pour cette donnée avec les informations non associées à une consultation par exemple.

Concernant la cartographie **sémantique**, nous nous sommes heurtés ici à un travail colossal. D'une part de par la volumétrie en termes de diversité des données disponibles mais également en termes de **choix** et travail. En effet, chaque entité, d'abord mappée de manière automatique, doit ensuite voir son concept doit être relu afin d'être validé. Si ce n'est pas le cas, une fouille manuelle des bases de vocabulaire doit alors être effectuée pour essayer d'identifier un concept correspondant et pertinent dans les bases de vocabulaire standardisées. Par ailleurs, le travail étant effectué par un seul lecteur, cela n'exclue pas la possibilité de réaliser des erreurs dans la validation de la cartographie suggérée, de cartographier avec un terme non pertinent ou encore la non possibilité de tout cartographier faute de temps. De nombreuses décisions ayant été prises sur le choix des concepts pertinents ou non, un travail avec un ou plusieurs relecteurs aurait beaucoup gagner en termes de qualité [18].

Toute erreur ayant un impact non négligeable sur la réutilisation des données a postériori, un **travail minutieux** et le plus qualitatif possible est donc de rigueur. L'existence d'imprécisions étant inhérent à la transformation des données, il convient malgré tout de les limiter le plus possible.

Cependant, la majorité du volume des données recueillies étant souvent limitée à un nombre restreint d'entités, et bien que dans notre cas nous n'ayons pas réalisé une cartographie complète de chacune des tables fournies, ce travail semble suffisant pour estimer de la faisabilité de transformer des données issues de cabinet de médecine générale dans des terminologies standardisées.

L'autre point faible de la cartographie sémantique est l'absence de correspondance linguistique entre les bases de données source (en Français) et cible (en Anglais pour la plupart). Les équipes de recherche de Lamer et al. [39] ou encore celle de Yoon et al. [75] déclaraient également avoir été confronté) cette **problématique de la langue,** et des **habitudes culturelles associées** à un pays, qui s'est également posée lors de ce travail. L'outil Athena recense une seule base de vocabulaire standardisé en français (CIM10) ou et quelques autres comprenant du vocabulaire français (Rx Norm et RxNorm Extension), la grande majorité des autres données de vocabulaire sont en anglais et issues de différents pays. La nécessité pour de nombreuses entités de réaliser **une traduction préalable** est également un des potentiels freins à la qualité de ce travail puisqu'il constitue une potentielle source d'erreurs.

Le modèle OMOP et les bases de données standardisées fournies étant malgré tout en **constante amélioration et expansion (**création de l'extension de RxNorm afin d'intégrer des médicaments non uniquement disponibles aux Etats Unis par exemple), nous n'excluons pas la possibilité de voir un jour apparaître des bases de vocabulaire en Français pouvant faciliter le travail.

Certaines autres difficultés inhérentes à la forme des **données de vocabulaire standard** ont également été recensées avec notamment une trop grande exhaustivité dans certaines bases de données. C'est notamment le cas avec la base de données médicaments qui était disponible selon les posologies existantes à l'international. Ces dernières n'étant pas toujours recensées dans la base de vocabulaire standardisé, cela nous obligeait parfois à choisir le concept le plus simplifié (souvent la molécule plutôt que le médicament si la distinction existait) pour ne pas insérer de données erronées. De même certains traitements (et notamment les vaccins) contenant plusieurs molécules ont également été complexes à cartographier compte tenu de la non existence de correspondances ou de la non mise à jour de la base de vocabulaire. Ces choix auraient également nécessité de réaliser le travail à plusieurs pour pouvoir discuter de la pertinence entre les concepts et surtout d'avoir une autre lecture après cartographie afin de l'uniformiser. Overhage et al.[76] de même que Rosenbloom et al. [16] mentionnaient déjà cette difficulté de se complier à un langage commun, nécessitant parfois d'accepter l'existence de **limites dans la correspondance** entre les concepts source et cible pouvant être liée aux différences de définition par exemple (et donc une certaine inexactitude) ou alors de se résoudre, dans certains cas, à accepter la **perte d'information** (et la non exhaustivité de la réutilisation des données pour la recherche par exemple).

Enfin, quelques limites peuvent également se poser quant aux **sources de données** utilisées. D'une part pour les données fictives du jeu de données source qui, fidèles aux données sources, puisqu'elles sont **codées manuellement** dans la table de base,

ne sont pas exemptes d'erreurs ou même d'omissions. Le dossier médical ayant souvent vocation à être utilisé par un praticien (au plus par un petit collectif s'il existe un remplaçant ou que la patientèle est mutualisée dans un cabinet), il est possible d'y retrouver quelques **habitudes personnelles** en matière de **codage** (aussi appelé codages locaux) ne permettant ni l'exhaustivité des données source ni la qualité optimale du remplissage du dossier patient. D'autre part, nous avons travaillé avec la version 5 du modèle OMOP, qui n'est pas la plus récente sauf pour l'outil Athena[69] et ses bases de vocabulaire standardisé. Ainsi, il n'est pas exclu que le même travail réalisé quelques mois après aurait pu permettre de cartographier quelques concepts supplémentaires si les bases de vocabulaires sont à jour ou nouvellement implémentées. La décision de créer une **base de vocabulaire personnalisée** n'a pas été envisagée dans notre cas mais reste également une option [76].

Dans les deux cas, les données mériteraient à être mises à jour et implémentées régulièrement nécessitant donc un travail régulier sur les bases de données source et cible.

3 Perspectives

A notre connaissance, à ce jour aucune étude attestant de la transformation de données médecine générale en France n'a été publiée et peu dans le monde entier s'intéressent à la réutilisation de données issues de médecine générale [59–63]. Par ailleurs, aucune étude n'a été recensées comme réalisant la transformation d'une base de données (fictive ou réelle) issue de cabinets de médecine générale au modèle OMOP.

Cette **étude de faisabilité** prouve que de nombreuses données issues de la médecine générale peuvent finalement être transformées dans un format standardisé, ce qui s'accorde avec la recherche actuelle. Ce modèle standard, permettant un accès facilité aux données pourrait avoir de nombreuses applications de grande envergure dans les domaines tels que la recherche, l'analyse des données, l'amélioration des soins et pratiques, la surveillance épidémiologique, …

En effet, dans la continuité directe de ce travail, un projet de plus grande envergure réunissant les **données réelles de 4 cabinets de médecine générale** partenaire du CHU de Lille vont également être extraites, transformées dans un modèle de données commun afin de pouvoir être réutilisées. Ce travail réunit une équipe composée de plusieurs médecins, ingénieurs, doctorants, et étudiants en master et est actuellement en cours de réalisation. Le travail **pluridisciplinaire** avec les équipes et la possibilité d'un **espace de concertation** permettra notamment de combler plusieurs des limites évoquées dans la présente étude et ainsi améliorer sa qualité.

De plus, à un niveau national, le **projet P4DP** (Platform For Data in Primary care), portée par un consortium regroupant le Collège national des Généralistes Enseignants, l'université Côte d'Azur, l'université Rouen Normandie, le CHU de Rouen, Loamics et le Health Data Hub est financé par l'Etat Français. L'objectif étant de créer l'un des premiers **entrepôts de données de santé de médecine générale** à des fins de recherche et d'innovation [77].

Conclusion

Ce travail a permis de montrer que, même s'il s'agit souvent de données non formalisées, écrites en texte libres et propres à chaque cabinet, on retrouve dans les données issues de la médecine générale de nombreux concepts pouvant être assimilables à des données que l'on peut standardiser.

Bien que le processus d'ETL ici n'ait pas porté sur l'ensemble des données mais seulement une fraction, il peut être considéré comme preuve de concept dans la possibilité de mettre les données issues de la médecine générale dans un format de données commun.

Ce travail s'inscrit par ailleurs dans la continuité de ceux réalisés en France aussi bien au CHU de Lille que via le projet D4DP tous deux ayant pour but de réutiliser les données de médecine générale afin que celles-ci aient un potentiel d'exploitation supérieur à celui uniquement concentré sur le recueil et le suivi lors des soins prodigués au patient.

Liste des tables

Liste des figures

Références

[1] Sessler DI. Big Data–and its contributions to peri-operative medicine. Anaesthesia 2014;69:100–5. https://doi.org/10.1111/anae.12537.

[2] Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int 2015;2015:639021. https://doi.org/10.1155/2015/639021.

[3] Berger ML, Doban V. Big data, advanced analytics and the future of comparative effectiveness research. J Comp Eff Res 2014. https://becarispublishing.com/doi/10.2217/cer.14.2 (accessed March 23, 2023).

[4] Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genet Med Off J Am Coll Med Genet 2013;15:802–9. https://doi.org/10.1038/gim.2013.121.

[5] Favaretto M, De Clercq E, Schneble CO, Elger BS. What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. PloS One 2020;15:e0228987. https://doi.org/10.1371/journal.pone.0228987.

[6] Ward JS, Barker A. Undefined By Data: A Survey of Big Data Definitions 2013.

[7] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017;26:38–52. https://doi.org/10.15265/IY-2017-007.

[8] Safran C. Reuse of clinical data. Yearb Med Inform 2014;9:52–4. https://doi.org/10.15265/IY-2014-0013.

[9] Safran C. Update on Data Reuse in Health Care. Yearb Med Inform 2017;26:24–7. https://doi.org/10.15265/IY-2017-013.

[10] Chazard E, Ficheur G, Caron A, Lamer A, Labreuche J, Cuggia M, et al. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. Stud Health Technol Inform 2018;255:15–9.

[11] Larousse É. Définitions : interopérabilité - Dictionnaire de français Larousse n.d. https://www.larousse.fr/dictionnaires/francais/interop%C3%A9rabilit%C3%A9/43787 (accessed September 1, 2023).

[12] HL7 Standards Product Brief - CDA® (HL7 Clinical Document Architecture) | HL7 International n.d. https://www.hl7.org/implement/standards/product_brief.cfm?product_id=496 (accessed September 24, 2023).

[13] Reference Information Model (RIM) Downloads | HL7 International n.d. https://www.hl7.org/implement/standards/rim.cfm (accessed September 24, 2023).

[14] European Commission. Directorate General for Informatics. New European interoperability framework: promoting seamless services and data flows for European public administrations. LU: Publications Office; 2017.

[15] Interoperability in Healthcare | HIMSS 2020. https://www.himss.org/resources/interoperability-healthcare (accessed September 24, 2023).

[16] Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing Knowledge Consistently Across Health Systems. Yearb Med Inform 2017;26:139–47. https://doi.org/10.15265/IY-2017-018.

[17] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41:391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[18] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS Wash DC 2016;4:1244. https://doi.org/10.13063/2327-9214.1244.

[19] Weeks J, Pardee R. Learning to Share Health Care Data: A Brief Timeline of Influential Common Data Models and Distributed Health Data Networks in U.S. Health Care Research. eGEMs 2023;7:4. https://doi.org/10.5334/egems.279.

[20] Liyanage H, Liaw S-T, Jonnagaddala J, Hinton W, de Lusignan S. Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs. Stud Health Technol Inform 2018;255:60–4.

[21] Introduction to HL7 Standards | HL7 International n.d. http://www.hl7.org/implement/standards/ (accessed September 26, 2023).

[22] Index - FHIR v5.0.0 n.d. https://www.hl7.org/fhir/ (accessed September 24, 2023).

[23] Viangteeravat T, Anyanwu MN, Nagisetty VR, Kuscu E, Sakauye ME, Wu D. Clinical data integration of distributed data sources using Health Level Seven (HL7) v3-RIM mapping. J Clin Bioinforma 2011;1:32. https://doi.org/10.1186/2043-9113-1-32.

[24] Reference Information Model - TD 2023. https://vico.org/HL7_RIM/index.html#introduction (accessed September 28, 2023).

[25] Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The Use of FHIR in Digital Health - A Review of the Scientific Literature. Stud Health Technol Inform 2019;267:52–8. https://doi.org/10.3233/SHTI190805.

[26] Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019;94:103188. https://doi.org/10.1016/j.jbi.2019.103188.

[27] Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. JMIR Med Inform 2021;9:e21929. https://doi.org/10.2196/21929.

[28] PCORnet. Natl Patient-Centered Clin Res Netw n.d. https://pcornet.org/ (accessed September 26, 2023).

[29] Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc JAMIA 2016;23:909–15. https://doi.org/10.1093/jamia/ocv188.

[30] Who We Are – OHDSI n.d. https://ohdsi.org/who-we-are/ (accessed September 26, 2023).

[31] documentation:overview [Observational Health Data Sciences and Informatics] n.d. https://www.ohdsi.org/web/wiki/doku.php?id=documentation:overview (accessed September 25, 2023).

[32] Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review. Int J Mol Sci 2022;23:11834. https://doi.org/10.3390/ijms231911834.

[33] Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol 2021;21:238. https://doi.org/10.1186/s12874-021-01434-3.

[34] Rinner C, Gezgin D, Wendl C, Gall W. A Clinical Data Warehouse Based on OMOP and i2b2 for Austrian Health Claims Data. Stud Health Technol Inform 2018;248:94–9.

[35] Sathappan SMK, Jeon YS, Dang TK, Lim SC, Shao Y-M, Tai ES, et al. Transformation of Electronic Health Records and Questionnaire Data to OMOP CDM: A Feasibility Study Using SG_T2DM Dataset. Appl Clin Inform 2021;12:757–67. https://doi.org/10.1055/s-0041-1732301.

[36] You SC, Lee S, Cho S-Y, Park H, Jung S, Cho J, et al. Conversion of National Health Insurance Service-National Sample Cohort (NHIS-NSC) Database into Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM). Stud Health Technol Inform 2017;245:467–70.

[37] OMOP CDM: An approach for standardizing health data | BMD Software 2022. https://www.bmd-software.com/news/omop-cdm-an-approach-for-standardizing-health-data/ (accessed March 31, 2023).

[38] Usability of OMOP Common Data Model for Detailed Lab Microbiology Results - PubMed 2023. https://pubmed.ncbi.nlm.nih.gov/35612079/ (accessed June 16, 2023).

[39] Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre M-M, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. Appl Clin Inform 2020;11:13–22. https://doi.org/10.1055/s-0039-3402754.

[40] Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333–41. https://doi.org/10.1016/j.jbi.2016.10.016.

[41] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574–8.

[42] Peng Y, Nassirian A, Ahmadi N, Sedlmayr M, Bathelt F. Towards the Representation of Genomic Data in HL7 FHIR and OMOP CDM. Stud Health Technol Inform 2021;283:86–94. https://doi.org/10.3233/SHTI210545.

[43] Klann JG, Phillips LC, Herrick C, Joss MAH, Wagholikar KB, Murphy SN. Web services for data warehouses: OMOP and PCORnet on i2b2. J Am Med Inform Assoc JAMIA 2018;25:1331–8. https://doi.org/10.1093/jamia/ocy093.

[44] Différences entre les types de données structurées et non structurées n.d. https://www.oracle.com/fr/big-data/structured-vs-unstructured-data/ (accessed September 26, 2023).

[45] ICD-10 Version:2008 n.d. https://icd.who.int/browse10/2008/fr (accessed September 26, 2023).

[46] Codage des actes médicaux - CCAM n.d. https://www.ameli.fr/medecin/exercice-liberal/facturation-remuneration/consultations-actes/nomenclatures-codage/codage-actes-medicaux-ccam (accessed September 26, 2023).

[47] Home – LOINC 2023. https://loinc.org/ (accessed June 16, 2023).

[48] WHOCC - ATC/DDD Index 2023. https://www.whocc.no/atc_ddd_index/ (accessed June 16, 2023).

[49] Données de santé en France : état des lieux et enjeux des bases de données, rôle de l'Assurance Maladie 2023. https://assurance-maladie.ameli.fr/presse/2023-04-20-dp-donnees-de-sante (accessed July 5, 2023).

[50] Prévention M de la S et de la, Prévention M de la S et de la. Programme de médicalisation des systèmes d'information (PMSI). Ministère Santé Prév 2023. https://sante.gouv.fr/professionnels/gerer-un-etablissement-de-sante-medico-social/financement/financement-des-etablissements-de-sante-10795/financement-des-etablissements-de-sante-glossaire/article/programme-de-medicalisation-des-systemes-d-information-pmsi (accessed September 26, 2023).

[51] Qu'est-ce que le SNDS ? | SNDS n.d. https://www.snds.gouv.fr/SNDS/Qu-est-ce-que-le-SNDS (accessed September 26, 2023).

[52] Bases de données (Open Data) n.d. https://assurance-maladie.ameli.fr/etudes-et-donnees/donnees/liste-bases-de-donnees-open-data (accessed September 26, 2023).

[53] ATIH : Agence technique de l'information sur l'hospitalisation n.d. https://www.atih.sante.fr/ (accessed September 26, 2023).

[54] Cren P-Y, Bertrand N, Le Deley M-C, Génin M, Mortier L, Odou P, et al. Is the survival of patients treated with ipilimumab affected by antibiotics? An analysis of 1585 patients from the French National hospital discharge summary database (PMSI). Oncoimmunology 2020;9:1846914. https://doi.org/10.1080/2162402X.2020.1846914.

[55] Réseau Sentinelles > France > Le réseau Sentinelles n.d. https://www.sentiweb.fr/?page=presentation (accessed September 28, 2023).

[56] OMG - Observatoire de la Médecine Générale n.d. http://omg.sfmg.org/ (accessed September 28, 2023).

[57] Article 16 de la loi n° 76-616 du 9 juillet 1976 relative à la lutte contre le tabagisme Modifié par Loi n°91-32 du 10 janvier 1991 - art. 4 () JORF 12 janvier 1991 n.d.

[58] Définir une liste de diagnostic pour une zone de traumatisme n.d. https://www.cegedim-logiciels.com/dyn/espace_client/Aide_en_ligne/MediClick/5.17/content/ch06s07s02s02s01.html (accessed September 1, 2023).

[59] Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity 2020;124:525–34. https://doi.org/10.1038/s41437-020-0303-2.

[60] Waschkau A, Wilfling D, Steinhäuser J. Are big data analytics helpful in caring for multimorbid patients in general practice? - A scoping review. BMC Fam Pract 2019;20:37. https://doi.org/10.1186/s12875-019-0928-5.

[61] Monaghan T, Manski-Nankervis J-A, Canaway R. Big data or big risk: general practitioner, practice nurse and practice manager attitudes to providing de-identified patient health data from electronic medical records to researchers. Aust J Prim Health 2020;26:466–71. https://doi.org/10.1071/PY20153.

[62] Tran B, Straka P, Falster MO, Douglas KA, Britz T, Jorm LR. Overcoming the data drought: exploring general practice in Australia by network analysis of big data. Med J Aust 2018;209:68–73. https://doi.org/10.5694/mja17.01236.

[63] de Zulueta P. Confidentiality, privacy, and general practice: GPDPR and the brave new world of “big data.” Br J Gen Pract J R Coll Gen Pract 2021;71:420–1. https://doi.org/10.3399/bjgp21X717017.

[64] SESAM-Vitale G. Étude de marché 2021 des industriels pour les professionnels de santé libéraux n.d.

[65] PostgreSQL: The world’s most advanced open source database n.d. https://www.postgresql.org/ (accessed September 24, 2023).

[66] THEMIS 2023.

[67] White Rabbit 2023.

[68] Rabbit in a Hat n.d. https://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html (accessed September 24, 2023).

[69] Athena n.d. https://athena.ohdsi.org/ (accessed September 24, 2023).

[70] Home. SNOMED Int n.d. https://www.snomed.org (accessed September 26, 2023).

[71] RxNorm | The Measures Management System n.d. https://mmshub.cms.gov/measure-lifecycle/measure-specification/specify-code/RxNorm (accessed September 26, 2023).

[72] USAGI for vocabulary mapping – OHDSI n.d. https://www.ohdsi.org/analytic-tools/usagi/ (accessed September 24, 2023).

[73] Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care 2012;50 Suppl:S60-67. https://doi.org/10.1097/MLR.0b013e318259bff4.

[74] Chapter 14 Evidence Quality | The Book of OHDSI n.d. https://ohdsi.github.io/TheBookOfOhdsi/EvidenceQuality.html (accessed September 28, 2023).

[75] Yoon D, Ahn EK, Park MY, Cho SY, Ryan P, Schuemie MJ, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research. Healthc Inform Res 2016;22:54–8. https://doi.org/10.4258/hir.2016.22.1.54.

[76] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc JAMIA 2012;19:54–60. https://doi.org/10.1136/amiajnl-2011-000376.

[77] P4DP, un consortium pour créer le premier entrepôt de données de santé pour la médecine générale. Health Data Hub n.d. https://www.health-data-hub.fr/actualites/p4dp-un-consortium-pour-creer-le-premier-entrepot-de-donnees-de-sante-pour-la-medecine (accessed September 25, 2023).

Annexe 1 : articles traitant de la réutilisation de données en médecine générale

Afin d'identifier dans la littérature le potentiel de réutilisation des données de médecine générale nous avons effectué la recherche suivante :

> (("data reuse"[Title]) OR ("second utilisation of data"[Title]) OR ("big data"[Title]) OR ("réutilisation de données"[Title])) AND (("general practitioner"[Title]) OR ("general practitionner"[Title]) OR ("general practitioners"[Title]) OR ("general practitionners"[Title]) OR ("GP"[Title]) OR ("general practice"[Title]) OR ("general practices "[Title]) OR ("family practitionner"[Title]) OR ("family practitioner"[Title]) OR ("family practitionners"[Title]) OR ("family practitionners"[Title]) OR ("family practice"[Title]) OR ("médecin généraliste"[Title]) OR ("médecine générale "[Title])) AND (("2010"[Date - Publication] : "2021"[Date - Publication]))

Cela nous a permis d'identifier les 5 articles cités ci-dessous [28–32].

Big data in digital healthcare: lessons learnt and recommendations for general practice [59]

Raag Agrawal 1 2, Sudhakaran Prabakaran 3 4 5

PMID: 32139886 PMCID: PMC7080757 DOI: 10.1038/s41437-020-0303-2

Big Data will be an integral part of the next generation of technological developments-allowing us to gain new insights from the vast quantities of data being produced by modern life. There is significant potential for the application of Big Data to healthcare, but there are still some impediments to overcome, such as fragmentation, high costs, and questions around data ownership. Envisioning a future role for Big Data within the digital healthcare context means balancing the benefits of improving patient outcomes with the potential pitfalls of increasing physician burnout due to poor implementation leading to added complexity. Oncology, the field where Big Data collection and utilization got a heard start with programs like TCGA and the Cancer Moon Shot, provides an instructive example as we see different perspectives provided by the United States (US), the United Kingdom (UK) and other nations in the implementation of Big Data in patient care with regards to their centralization and regulatory approach to data. By drawing upon global approaches, we propose recommendations for guidelines and regulations of data use in healthcare centering on the creation of a unique global patient ID that can integrate data from a variety of healthcare providers. In addition, we expand upon the topic by discussing potential pitfalls to Big Data such as the lack of diversity in Big Data research, and the security and transparency risks posed by machine learning algorithms.

Are big data analytics helpful in caring for multimorbid patients in general practice? - A scoping review [60]

Alexander Waschkau [1], Denise Wilfling [2], Jost Steinhäuser [2]

Affiliations expand

- PMID: 30813904
- PMCID: PMC6394098
- DOI: 10.1186/s12875-019-0928-5

Free PMC article

Abstract

Background: The treatment of multimorbid patients is one crucial task in general practice as multimorbidity is highly prevalent in this setting. However, there is little evidence how to treat these patients and consequently there are but a few guidelines that focus primarily on multimorbidity. Big data analytics are defined as a method that obtains results for high volume data with high variety generated at high velocity. Yet, the explanatory power of these results is not completely understood. Nevertheless, addressing multimorbidity as a complex condition might be a promising field for big data analytics. The aim of this scoping review was to evaluate whether applying big data analytics on patient data does already contribute to the treatment of multimorbid patients in general practice.

Methods: In January 2018, a review searching the databases PubMed, The Cochrane Library, and Web of Science, using defined search terms for "big data analytics" and "multimorbidity", supplemented by a search of grey literature with Google Scholar, was conducted. Studies were not filtered by type of study, publication year or language. Validity of studies was evaluated independently by two researchers.

Results: In total, 2392 records were identified for screening. After title and abstract screening, six articles were included in the full-text analysis. Of those articles, one reported on a model generated with big data techniques to help caring for one group of multimorbid patients. The other five articles dealt with the analysis of multimorbidity clusters. No article defined big data analytics explicitly.

Conclusions: Although the usage of the phrase "Big Data" is growing rapidly, there is nearly no practical use case for big data analysis techniques in the treatment of multimorbidity in general practice yet. Furthermore, in publications addressing big data analytics, the term is rarely defined. However, possible models and algorithms to address multimorbidity in the future are already published.

Keywords: Big data analytics; General practice; Multimorbidity; eHealth.

Overcoming the data drought: exploring general practice in Australia by network analysis of big data [62]

Bich Tran [1], Peter Straka [2], Michael O Falster [3], Kirsty A Douglas [4], Thomas Britz [2], Louisa R Jorm [3]

Affiliations expand

- PMID: 29976132

- DOI: 10.5694/mja17.01236

Abstract

Objectives: To investigate the organisation and characteristics of general practice in Australia by applying novel network analysis methods to national Medicare claims data.

Design: We analysed Medicare claims for general practitioner consultations during 1994-2014 for a random 10% sample of Australian residents, and applied hierarchical block modelling to identify provider practice communities (PPCs).

Participants: About 1.7 million patients per year.

Main outcome measures: Numbers and characteristics of PPCs (including numbers of providers, patients and claims), proportion of bulk-billed claims, continuity of care, patient loyalty, patient sharing.

Results: The number of PPCs fluctuated during the 21-year period; there were 7747 PPCs in 2014. The proportion of larger PPCs (six or more providers) increased from 32% in 1994 to 43% in 2014, while that of sole provider PPCs declined from 50% to 39%. The median annual number of claims per PPC increased from 5000 (IQR, 40-19 940) in 1994 to 9980 (190-23 800) in 2014; the proportion of PPCs that bulk-billed all patients was lowest in 2004 (21%) and highest in 2014 (29%). Continuity of care and patient loyalty were stable; in 2014, 50% of patients saw the same provider and 78% saw a provider in the same PPC for at least 75% of consultations. Density of patient sharing in a PPC was correlated with patient loyalty to that PPC.

Conclusions: During 1994-2014, Australian GP practice communities have generally increased in size, but continuity of care and patient loyalty have remained stable. Our novel approach to the analysis of routinely collected data allows continuous monitoring of the characteristics of Australian general practices and their influence on patient care.

Confidentiality, privacy, and general practice: GPDPR and the brave new world of 'big data' [63]

Paquita de Zulueta [1]

Affiliations expand

- PMID: 34446416
- PMCID: PMC8378569
- DOI: 10.3399/bjgp21X717017

Free PMC article

No abstract available

Big data or big risk: general practitioner, practice nurse and practice manager attitudes to providing de-identified patient health data from electronic medical records to researchers [61]

Timothy Monaghan [1], Jo-Anne Manski-Nankervis [2], Rachel Canaway [2]

Affiliations expand

- PMID: 33292925
- DOI: 10.1071/PY20153

Abstract

Research utilising de-identified patient health information extracted from electronic medical records (EMRs) from general practices has steadily grown in recent years in response to calls to increase use of health data for research and other secondary purposes in Australia. Little is known about the views of key primary care personnel on this issue, which are important, as they may influence whether practices agree to provide EMR data for research. This exploratory qualitative study investigated the attitudes and beliefs of general practitioners (GPs), practice managers (PMs) and practice nurses (PNs) around sharing de-identified EMR patient health information with researchers. Semi-structured interviews were conducted with 11 participants (6 GPs, 3 PMs and 2 PNs) recruited via purposive sampling from general practices in Victoria, Australia. Transcripts were coded and thematically analysed. Participants were generally enthusiastic about research utilising de-identified health information extracted from EMRs for altruistic reasons, including: positive effects on primary care research, clinical practice and population health outcomes. Concerns raised included patient privacy and data breaches, third-party use of extracted data and patient consent. These findings can provide guidance to researchers and policymakers in designing and implementing projects involving de-identified health information extracted from EMRs.

Annexe 2 : Cartographie OMOP pour chaque jeu de données source

1 Table patient – OMOP personne

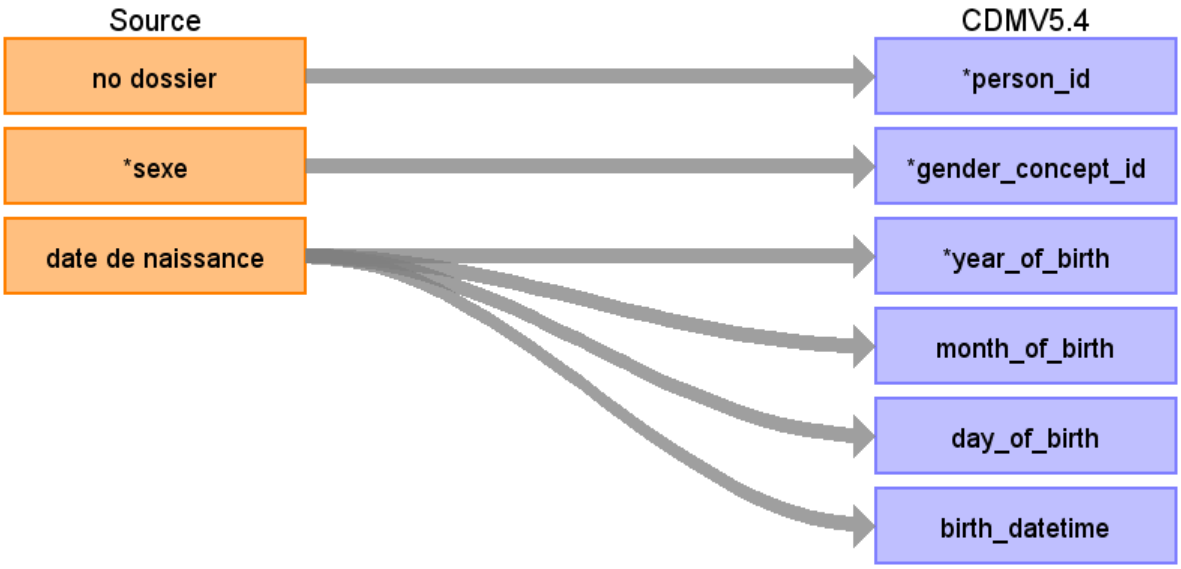

Figure 12. Représentation de la cartographie structurelle - table patient

Table source (5 colonnes)	Table OMOP (18 colonnes dont 5 requises)
no dossier Sexe date de naissance date de naissance date de naissance date de naissance	person_id* gender_concept_id* year_of_birth* month_of_birth day_of_birth birth_datetime
ALD Code régime	*Pas de correspondance* *Pas de correspondance*
Information non disponible *Information non disponible*	race_concept_id* ethnicity_concept_id* location_id provider_id care_site_id person_source_value gender_source_value gender_source_concept_id race_source_value race_source_concept_id ethnicity_source_value ethnicity_source_concept_id

Tableau 14. Cartographie complète de la table source patient vers la table OMOP person, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de

2 Table problème – OMOP visite_détail

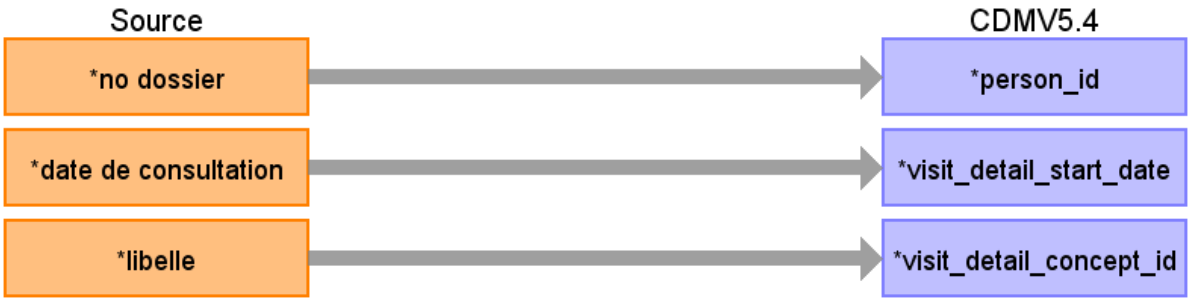

Figure 13. Représentation de la cartographie structurelle - table problème

Table source (4 colonnes)	Table OMOP (19 colonnes dont 6 requises)
no dossier libelle date de consultation	person_id* visit_detail_id* visit_detail_start_date*
-- Création manuelle	visit_occurrence_id*
date de début (identique date consultation	*-- Pas de correspondance*
	visit_detail_concept_id* visit_detail_end_date* visit_detail_start_datetime visit_detail_end_datetime visit_detail_type_concept_id provider_id care_site_id visit_detail_source_value visit_detail_source_concept_id admitted_from_concept_id admitted_from_source_value discharged_to_source_value discharged_to_concept_id preceding_visit_detail_id parent_visit_detail_id

Tableau 15. Cartographie complète de la table source problème vers la table OMOP visite_détails, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de la table

3 Table diagnostic – OMOP diagnostic (occurrence_condition)

Figure 14. Représentation de la cartographie structurelle - table probleme

Table source (5 colonnes)	Table OMOP (17 colonnes dont 5 requises)
no dossier date de consultation libelle code (partiellement rempli, non utilisé)	person_id* condition_start_date* condition_concept_id* condition_concept_id*
-- Création manuelle	visit_occurrence_id*
caractéristique	
Information non disponible	condition_type_concept_id* visit_detail_id condition_start_datetime condition_end_date condition_end_datetime visit_detail_id condition_status_concept_id stop_reason provider_id condition_occurrence_id condition_source_value condition_source_concept_id condition_status_source_value

Tableau 16. Cartographie complète de la table source diagnostic vers la table OMOP diagnostic, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de la table

4 Tables médicaments / vaccination – OMOP médicament (drug exposture)

4.1 Table médicaments

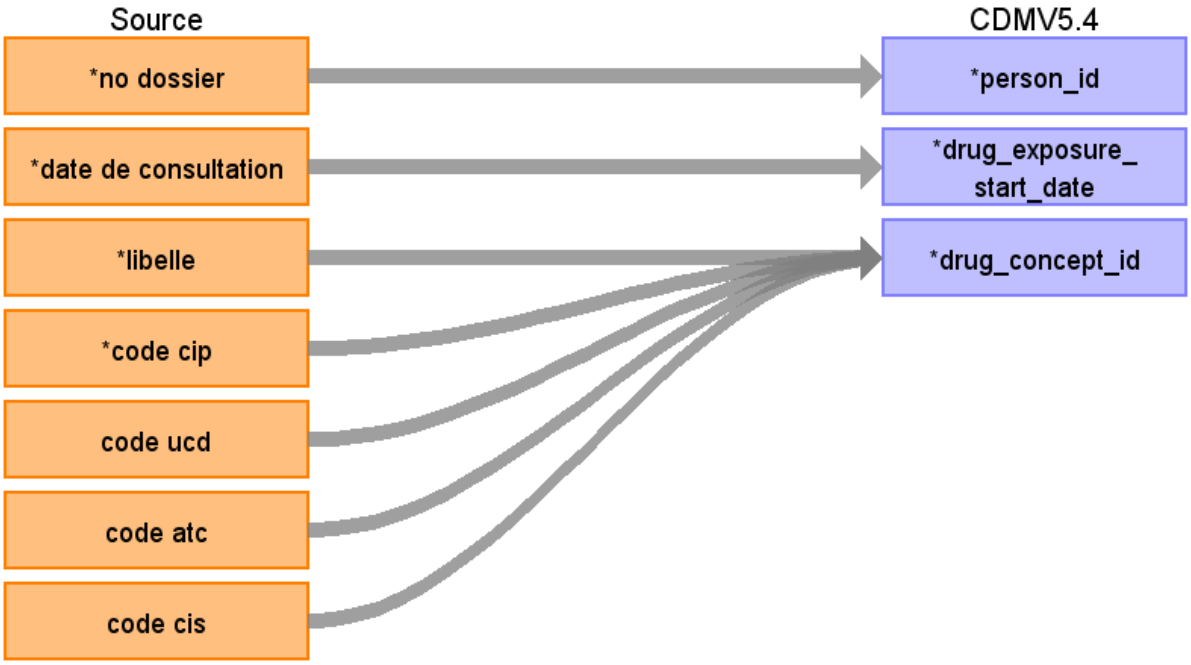

Figure 15. Représentation de la cartographie structurelle - table médicaments

Table source (9 colonnes)	Table OMOP (23 colonnes dont 5 requises)
no dossier date de vaccination libelle code cip code ucd code atc code cis	person_id* drug_exposure_start_date* drug_concept_id* drug_concept_id* drug_concept_id* drug_concept_id* drug_concept_id*
--Création manuelle *--Création manuelle*	visit_occurrence_id drug_exposure_id*
note motif de prescription	
	drug_type_concept_id* drug_exposure_start_datetime drug_exposure_end_date drug_exposure_end_datetime verbatim_end_date visit_detail_id stop_reason Refills Quantity days_supply Sig route_concept_id visit_occurrence_id drug_source_value drug_source_concept_id route_source_value dose_unit_source_value

Tableau 17. Cartographie complète de la table source médicament vers la table OMOP médicaments, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de la table

4.2 Table vaccination

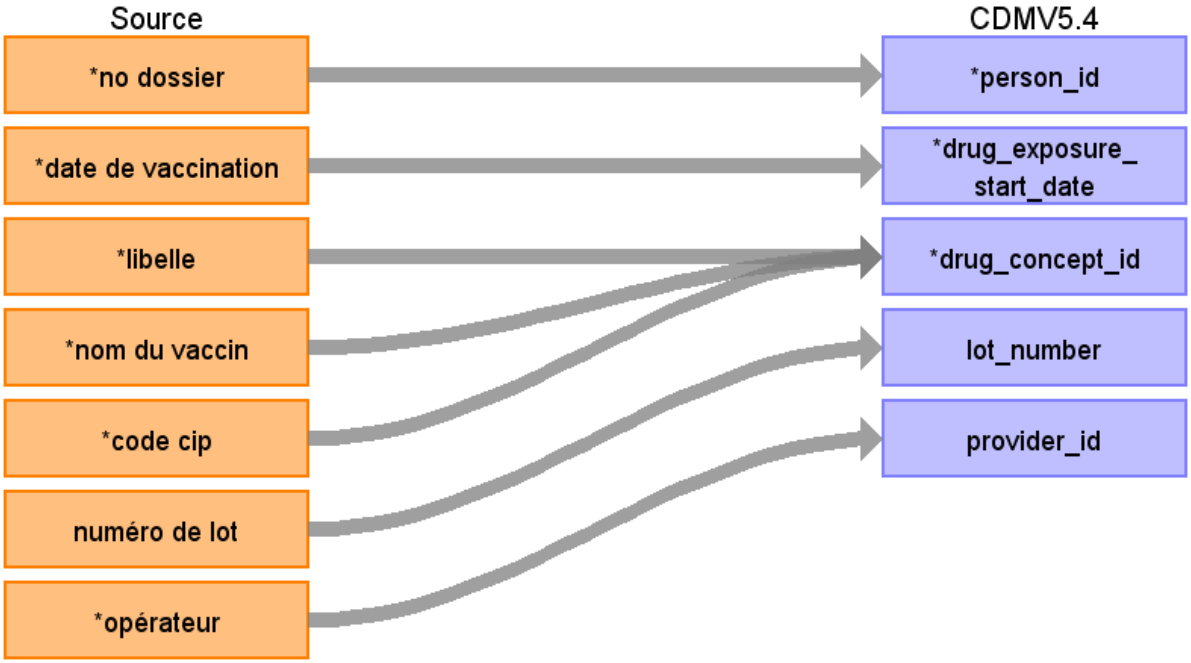

Figure 16. Représentation de la cartographie structurelle - table vaccination

Table source (6 colonnes)	Table OMOP (23 colonnes dont 5 requises)
no dossier date de vaccination libelle nom du vaccin code cip numéro de lot opérateur	person_id* drug_exposure_start_date* drug_concept_id* drug_concept_id* drug_concept_id* birth_datetime provider_id
--Création manuelle *--Création manuelle*	visit_occurrence_id drug_exposure_id*
rappel zone d'injection effet secondaire date de lecture lecture du test test positif résultat Commentaires	
	drug_type_concept_id* drug_exposure_start_datetime drug_exposure_end_date drug_exposure_end_datetime verbatim_end_date visit_detail_id stop_reason Refills Quantity days_supply Sig route_concept_id visit_occurrence_id drug_source_value drug_source_concept_id route_source_value dose_unit_source_value

Tableau 18. Cartographie complète de la table source vaccination vers la table OMOP médicaments, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de la table

5 Tables biologie / biométrie – OMOP mesures (measurement)

5.1 Table biologie

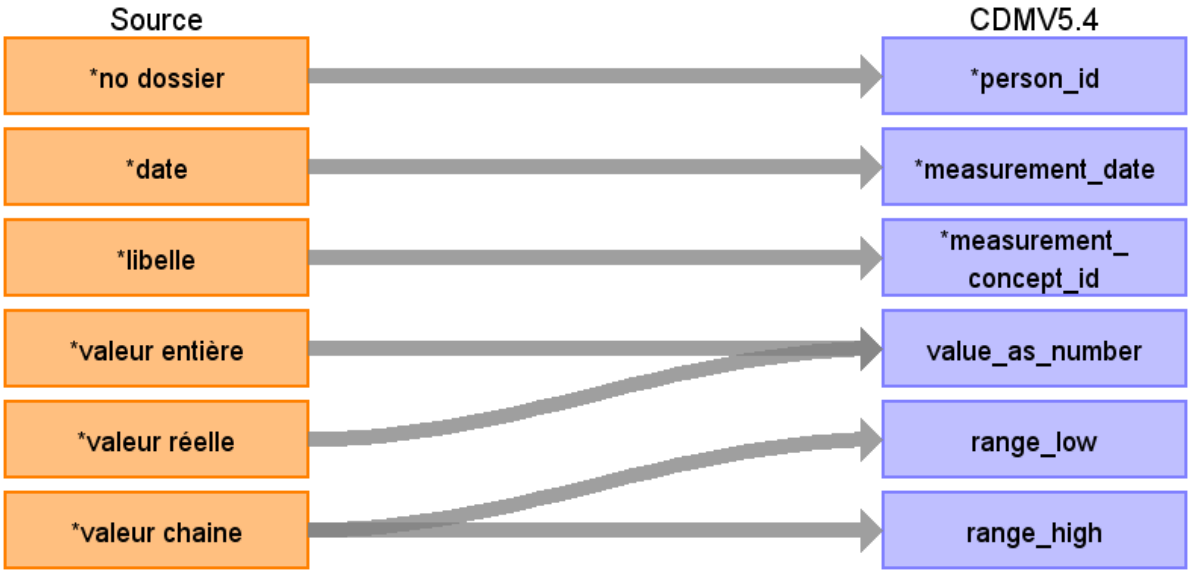

Figure 17. Représentation de la cartographie structurelle - table biologie

Table source (6 colonnes)	Table OMOP (23 colonnes dont 4 requises)
no dossier date libelle valeur réelle valeur chaine valeur chaine	person_id* measurement_date* measurement_concept_id* value_as_number range_low range_high
--Création manuelle *--Création manuelle*	visit_occurrence_id measurement_id*
	value_source_value measurement_datetime measurement_time measurement_type_concept_id operator_concept_id value_as_concept_id unit_concept_id provider_id visit_occurrence_id visit_detail_id measurement_source_value measurement_source_concept_id unit_source_value unit_source_concept_id measurement_event_id meas_event_field_concept_id

Tableau 19. Cartographie complète de la table source biologie vers la table OMOP mesures, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de la table

5.2 Table biométrie

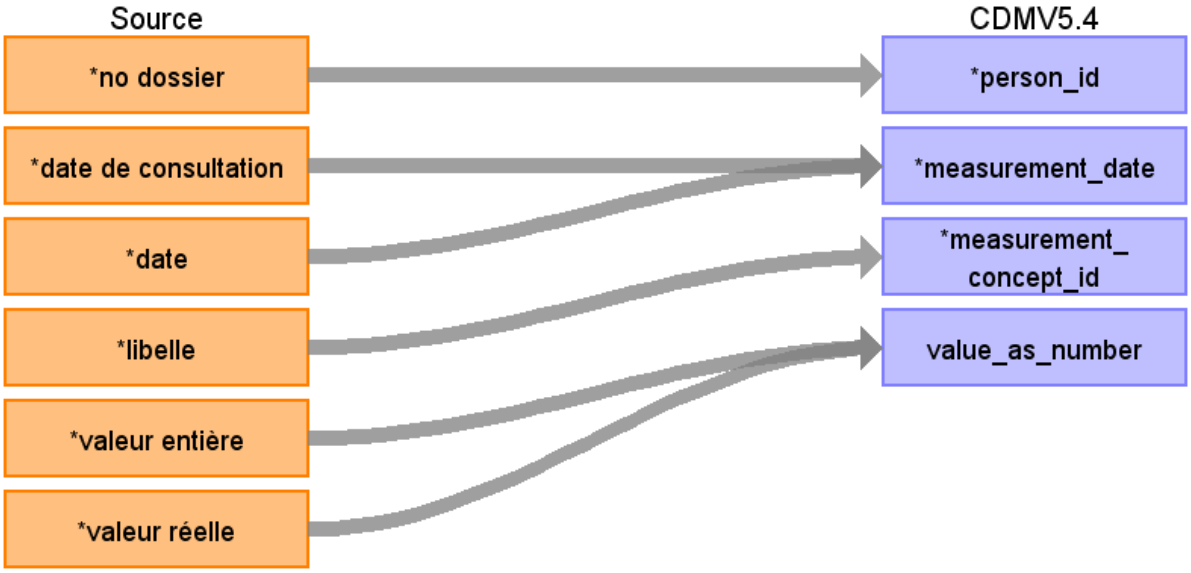

Figure 18. Représentation de la cartographie structurelle - table biométrie

Table source (6 colonnes)	Table OMOP (23 colonnes dont 4 requises)
no dossier date de consultation libelle valeur entière valeur réelle	person_id* measurement_date* measurement_concept_id* value_as_number value_as_number
date (identique date consultation)	*Pas de correspondance*
--Création manuelle *--Création manuelle*	visit_occurrence_id measurement_id*
	value_source_value measurement_datetime measurement_time range_low range_high measurement_type_concept_id operator_concept_id value_as_concept_id unit_concept_id provider_id visit_occurrence_id visit_detail_id measurement_source_value measurement_source_concept_id unit_source_value unit_source_concept_id measurement_event_id meas_event_field_concept_id

Tableau 20. Cartographie complète de la table source biométrie vers la table OMOP mesures, données non implémentées incluses. En vert les colonnes de la table ayant pu être associées, marqués d'une astérisque les colonnes de la table OMOP requises pour la création de la table

AUTEUR : Nom : NYANGWILE **Prénom :** Eole

Date de Soutenance : 13/10/2023

Titre de la Thèse : Réutilisation de données en Médecine Générale

Thèse - Médecine - Lille 2023

Cadre de classement : Médecine

DES + FST ou option : DES de Santé Publique et Médecine Sociale

Mots-clés : Réutilisation de données, Médecine Générale, OMOP, ETL

Résumé :

Contexte : Avec l'avènement du numérique, de volumineuses données s'accumulent permettant de constituer des entrepôts de données de plus en plus importants. La réutilisation des données, et notamment celles de santé produites à l'occasion d'un épisode de soin, est décrite dans de nombreuses études. Cependant très peu décrivent les modalités de réalisation de ce travail sur des données issues de cabinets de médecine générale et aucune n'inclut des données françaises.

Notre objectif est de réaliser ce procédé de transformation vers le modèle de données OMOP avec des données semblables à celles issues de cabinet de médecine générale, puis d'en réaliser une rapide analyse descriptive.

Matériel et Méthode : Nous utilisons un jeu de données simulées dans le cadre du CPER Tec'Santé. Ce jeu de données décrit 15000 patients, et respecte la structure et les distributions observées dans une base de données réelle. Nous avons réalisé un processus d'ETL vers le modèle de données commun OMOP. Nous avons dans un premier temps réalisé et détaillé la cartographie syntaxique puis la cartographie sémantique avant de présenter une rapide analyse descriptive des résultats.

Résultats : Grâce au modèle de données commun OMOP, les 7 jeux de données fictives ont pu être transformés dans format standardisé. Concernant le vocabulaire dans chaque base, 51% à 100% des terminologies disponibles ont pu être mises en correspondance avec un concept normalisé lors de la cartographie sémantique.

Discussion : Ce travail montre la faisabilité du processus avec des données issues des cabinets de médecine générale français. Les résultats descriptifs illustrent le potentiel scientifique de telles données. Un tel processus doit être réalisé en équipe.

Composition du Jury :

Président : Monsieur le Professeur Philippe AMOUYEL

Assesseurs : Monsieur le Docteur Matthieu CALAFIORE
Monsieur le Docteur Bertrand LEGRAND

Directeur : Monsieur le Professeur Emmanuel Chazard