THÈSE DE DOCTORAT

Discipline : Recherche clinique, innovation technologique, santé publique

Réutilisation des données de soins premiers : spécificités, standardisation et suivi de la prise en charge dans les Maisons de Santé Pluridisciplinaires

Par MATHILDE FRUCHART

Université de Lille

École Doctorale Biologie Santé de Lille
ULR 2694 METRICS (Université de Lille, CHU de Lille)

Dirigée par ANTOINE LAMER

Présentée le **18 novembre 2024** devant le jury composé de :

Docteur Vianney JOUHET	Université de Bordeaux	Rapporteur
Docteure Rosy TSOPRA	Université Paris Cité	Rapporteur
Professeur Jean-Baptiste BEUSCART	Université de Lille	Président du jury
Professeure Béatrice LOGNOS	Université de Montpellier	Examinateur
Professeure Fleur MOUGIN	Université de Bordeaux	Examinateur
Professeure Sylvia PELAYO	Université de Lille	Examinateur
Docteur Bastien RANCE	Université Paris Descartes	Examinateur
Docteur Antoine LAMER	Université de Lille	Directeur de thèse
Docteur Paul QUINDROIT	Université de Lille	Invité

Remerciements

Aux membres du jury,

Madame la Docteure Rosy TSOPRA, et
Monsieur le Docteur Vianney JOUHET

Je vous remercie pour le temps accordé à l'évaluation de ce manuscrit et pour avoir accepté d'être les rapporteurs de cette thèse. Je suis reconnaissante de votre participation dans mon jury de thèse et de pouvoir compter sur vos expertises.

Madame la Professeure Béatrice LOGNOS, et
Monsieur le Docteur Bastien RANCE

Je vous remercie pour votre accompagnement tout au long de mon doctorat. Je suis reconnaissante pour les recommandations sur le plan médical et sur l'analyse des données. Vos expertises m'ont permis de m'améliorer et d'aboutir à ce projet. Je vous remercie également pour votre soutien et vos conseils.

Madame la Professeure Fleur MOUGIN,
Madame la Professeure Sylvia PELAYO, et
Monsieur le Professeur Jean-Baptiste BEUSCART

Je vous remercie d'avoir accepté de faire partie du jury de ma thèse et de pouvoir compter sur vos recommandations.

Monsieur le Docteur Antoine LAMER, directeur de ma thèse,

Je te remercie de m'avoir proposé ce sujet de thèse, qui m'a permis d'approfondir mes connaissances et de m'épanouir sur le plan professionnel. Merci pour le partage de ton expertise et pour m'avoir incluse dans de nombreux projets de recherche, ce qui m'a permis d'apprendre énormément. Je suis très reconnaissante pour la confiance que tu m'as accordée dans le projet PriCaDa, pour les recommandations de participation aux congrès, et pour le soutien que tu m'as apporté durant mes moments de doute et d'anxiété.

À toutes les personnes avec qui j'ai travaillé,

À toute l'équipe du CERIM,
Je remercie profondément toute l'équipe du CERIM pour sa bienveillance, son accueil et son soutien. Merci pour les moments de convivialité, de partage et de conseils.
Merci, Monsieur **Emmanuel Chazard**, pour votre accueil au sein du CERIM. Un grand merci **Paul Quindroit** pour ton encadrement, ton soutien et ton apport dans chacune des tâches de mon travail. Merci **Anaïs Payen** pour ton aide, tes critiques constructives, tes conseils si précieux pour la préparation de thèse. Merci pour nos moments d'entraide post-réunion. Merci **Eiya Ayed** et **Erwin Gerard** pour vos conseils sur l'aspect pharmaceutique. Merci d'avoir apporté la bonne humeur dans notre bureau. Merci **Renaud Périchon**, **Sophie Quenton** et **Julien Soula** pour votre expertise en informatique et votre secours lors de mes nombreux problèmes techniques. Enfin, merci **Mélanie Steffe** pour ton aide dans mes démarches pour les congrès et pour tes réponses à mes nombreuses questions.

Aux membres du projet PriCaDa,
Je remercie **Matthieu Calafiore**, **Jean-Baptiste Beuscart**, **Paul Quindroit**, **Anaïs Payen** et **Antoine Lamer** de m'avoir fait confiance sur la partie technique du projet PriCaDa. Merci d'avoir répondu à mes nombreuses questions lors des réunions mensuelles et de m'avoir familiarisée avec le versant clinique du projet. Merci aux stagiaires **Antoine Teston**, **Ezechiel Djohi** et **Elliot Houdant** d'avoir fourni un travail considérable dans le projet et merci pour l'aide apportée dans les diverses tâches techniques.

À mes relectrices, Nathalie, Célestine et ma mère Myriam,
Merci pour le temps précieux consacré à la relecture de mon travail. Merci pour les critiques qui m'ont permis d'avancer et de m'améliorer.

À ma famille,

Merci à toute ma famille pour ces moments de partage et de convivialité.

À mes parents,
Je suis très reconnaissante pour votre soutien inestimable durant mes études depuis le premier jour. Merci pour les valeurs que vous m'avez transmises depuis toute petite. Merci pour vos encouragements constants, votre compréhension et votre amour qui m'ont donné la force et la motivation nécessaires pour surmonter les épreuves de la vie. Je vous remercie d'avoir toujours cru en moi, d'être une source d'inspiration au quotidien car sans vous je ne serais pas celle que je suis aujourd'hui. Merci à mon père de me rappeler constamment de penser à moi en priorité. Merci à ma mère pour toutes nos pauses de midi passées ensemble et pour ton soutien infini.

À mon grand frère Pierre, à mon petit frère Valentin et à Célestine,
Merci d'avoir été présents dans chaque moment important de ma vie. Merci à vous de m'avoir supportée, vous êtes mes piliers et mes repères. Les moments difficiles me rappellent à quel point nous sommes soudés et forts ensemble. Votre présence a été une source de réconfort et de motivation dans tous mes projets.

À mes cousins, à Philippe et Catherine,
Un grand merci à mes cousins **Adrien**, **Quentin** et **Flavien** pour leur joie de vivre m'apportant une bouffée d'air frais et de bonne humeur. Les moments chaleureux des cousinades m'ont apporté beaucoup de plaisir. Merci à mon oncle **Philippe** et ma tante **Catherine** de m'avoir accueillie à Ars-sur-Forman lorsque j'avais besoin de paix, de déconnexions et de rigolades.

À mes ami.e.s,

À Pierre,
Le mot "merci" est trop simple pour exprimer toute ma gratitude envers toi. Je suis reconnaissante pour ta douceur, ton calme et la plénitude que tu apportes dans ma vie. Merci de me supporter et de me soutenir dans mes humeurs de tous les jours (ce qui ne doit pas être facile parfois, je le reconnais...). Tu es mon pilier. J'apprends tellement de ta légèreté, de ton optimisme et de ta manière de voir les choses avec simplicité. Merci pour tous ces voyages, ces évasions, ces balades, ces discussions, ces nombreux fous rires... Bref, merci de partager mon quotidien.

Aux filles,
Marion, **Naomi**, **Faustine**, **Mathilde**, **Célestine**, **Alice**, je suis profondément reconnaissante pour nos années d'amitiés inconditionnelles. Merci pour les restaurants, les cinémas, les sorties, les voyages, les soirées et les fous rires indénombrables. Merci pour vos soutiens dans tous les moments d'échanges, de doutes et de remises en question.
À ma meilleure amie depuis 15 ans, **Naomi**, mon Dupont, ma physionomie, je te remercie du fond du coeur pour tout ce que tu m'apportes au quotidien. Tu es mon modèle de force et de courage, merci de me rassurer en permanence et d'être toujours présente.

À Marouane et Fabio,
Merci pour les pauses cafés, les restaurants et les sorties qui m'ont permis de me divertir mais aussi de me remettre en question et d'évoquer mes doutes et mes craintes. Merci de m'avoir motivée et rassurée quand j'en avais besoin.

À la team glaces,
Erwin, **Martin**, **Malik**, **Eiya**, merci pour ces nombreux afterworks à la Canopée, ces pauses goûter à manger les magnums de la cafétéria du CHU. Merci pour ces moments de rires, de soutien entre doctorants et de partage. Merci de m'avoir écoutée et épaulée lorsque j'en avais besoin.

Résumé

Contexte : La réutilisation des données de santé, au-delà de leur usage initial, permet d'améliorer la prise en charge des patients, de faciliter la recherche et d'optimiser le pilotage des établissements de santé. Pour cela, les données sont extraites des logiciels de santé, transformées et stockées dans un entrepôt de données grâce à un processus *extract-transform-load* (ETL). Des modèles de données communs, comme le modèle OMOP, existent pour stocker les données dans un format homogène, indépendant de la source. Les données de facturation des soins centralisées dans la base nationale (SNDS), les données hospitalières, les données des réseaux sociaux et des forums, et les données de ville sont des sources de données représentatives du parcours de soins des patients. La dernière source de données est encore peu exploitée.

Objectif : L'objectif de cette thèse a été d'intégrer les spécificités de la réutilisation des soins premiers pour implémenter un entrepôt de données, tout en montrant la contribution des soins premiers au domaine de la recherche.

Méthodes : Dans un premier temps, les données de soins premiers d'une maison de santé ont été extraites du logiciel de soins WEDA. Un entrepôt de données de soins premiers a été implémenté à l'aide d'un processus ETL. La transformation structurelle (harmonisation de la structure de la base de données) et sémantique (harmonisation du vocabulaire utilisé dans les données) ont été mises en place pour aligner les données avec le modèle de données commun OMOP. Pour intégrer les données des médecins généralistes de plusieurs maisons de santé, un outil de généralisation des processus ETL a été développé et testé sur quatre maisons de santé. Par la suite, un algorithme d'évaluation de la persistance à un traitement prescrit et des tableaux de bord ont été développés. Grâce à l'utilisation du modèle OMOP, ces outils sont partageables avec d'autres maisons de santé. Enfin, des études rétrospectives ont été réalisées sur la population de patients diabétiques des quatre maisons de santé.

Résultats : Sur plus de 20 ans, les données des 117 005 patients de quatre maisons de santé ont été chargées dans le modèle OMOP, grâce à notre outil d'optimisation des processus ETL. Ces données couvrent les résultats de biologie des laboratoires de ville et les données relatives aux consultations de médecins généralistes. Le vocabulaire propre aux soins premiers a été aligné avec les concepts standards du modèle. Un algorithme pour évaluer la persistance à un traitement prescrit par le médecin généraliste, ainsi qu'un tableau de bord pour le suivi des indicateurs de performance (ROSP) et de l'activité du cabinet ont été développés. Basés sur les entrepôt de données des quatre maisons de santé, nous avons décrit le suivi des patients diabétiques. Ces études utilisent les données de résultats de biologie, les données de consultation et les prescriptions médicamenteuses, au format OMOP. Les scripts de ces études et les outils développés pourront être partagés.

Conclusion : Les données de soins premiers représentent un potentiel pour la réutilisation des données à des fins de recherche et d'amélioration de la qualité des soins. Elles complètent les bases de données existantes (hospitalières, nationales et réseaux sociaux) en intégrant les données cliniques de ville. L'utilisation d'un modèle de données commun facilite le développement d'outils et la conduite d'études, tout en permettant leur partage. Les études pourront être répliquées dans différents centres, afin de comparer les résultats.

Mots-clés : réutilisation des données, extract-transform-load (ETL), modèle de données commun OMOP, soins premiers, Maison de Santé Pluridisciplinaire (MSP), tableau de bord

Abstract

Context : Reusing healthcare data beyond its initial use helps to improve patient care, facilitate research, and optimize the management of healthcare organizations. To achieve this, data is extracted from healthcare software, transformed and stored in a data warehouse through an extract-transform-load (ETL) process. Common data models, such as the OMOP model, exist to store data in a homogeneous, source-independent format. Data from healthcare claims centralized in the national database (SNDS), hospital, social networks and forums, and primary care are different data sources representative of the patient care pathway. The last data source has not been fully exploited.

Objective : The aim of this thesis was to incorporate the specificities of primary care data reuse to implement a data warehouse while highlighting the contribution of primary care to the field of research.

Methods : The first step was to extract the primary care data of a multidisciplinary health center (MHC) from the WEDA care software. A primary care data warehouse was implemented using an ETL process. Structural transformation (harmonization of the database structure) and semantic transformation (harmonization of the vocabulary used in the data) were implemented to align the data with the common OMOP data model. A process generalization tool was developed to integrate general practitioners (GP) data from multiple care structures and tested on four MHCs. Subsequently, algorithm for assessing the persistence of a prescribed treatment and dashboards were developed. Thanks to the use of the OMOP model, these tools can be shared with other MHCs. Finally, retrospective studies were conducted on the diabetic population of the four MHCs.

Results : Over a period of more than 20 years, data of 117,005 patients from four MHCs were loaded into the OMOP model using our ETL process optimization tool. These data include biological results from laboratories and GP consultation data. The vocabulary specific to primary care was aligned with the standard concepts of the model. An algorithm for assessing persistence with treatment prescribed by the GP and also a dashboard for monitoring performance indicators (ROSP) and practice activity have been developed. Based on the data warehouses of four MHCs, we described the follow-up of diabetic patients. These studies use biological results, consultation and drug prescriptions data in OMOP format. The scripts of these studies and the tools developed can be shared.

Conclusion : Primary care data represent a potential for reusing data for research purposes and improving the quality of care. They complement existing databases (hospital, national and social networks) by integrating clinical data from the city. The use of a common data model facilitates the development of tools and the conduct of studies, while enabling their sharing. Studies can be replicated in different centers to compare results.

Keywords : data reuse, extract-transform-load (ETL), OMOP common data model, primary care, multidisciplinary health center (MHC), dashboard

Valorisation scientifique

Publications de la thèse

[1] **Fruchart M**, Quindroit P, Jacquemont C, Beuscart JB, Calafiore M, Lamer A, et al. Transforming Primary Care Data into the Observational Medical Outcomes Partnership Common Data Model : Development and Usability Study. JMIR Medical Informatics. 2024. doi : 10.2196/49542.

[2] **Fruchart M**, Lamer A, Lemaitre M, Beuscart JB, Calafiore M, Quindroit P, et al. Description of a French Population of Diabetics Treated Followed up by General Practitioners. Stud Health Technol Inform. 2023. doi : 10.3233/SHTI230289.

[3] **Fruchart M**, Quindroit P, Patel H, Beuscart JB, Calafiore M, Lamer A. Implementation of a Data Warehouse in Primary Care : First Analyses with Elderly Patients. Stud Health Technol Inform. 2022. doi : 10.3233/SHTI220510.

Contributions en lien avec la thèse

[1] A. Lamer, **M. Fruchart**, N. Paris, B. Popoff, A. Payen, T. Balcaen, W. Gacquer, M. Cuggia, M. Doutreligne, E. Chazard. Description standardisée du processus d'extraction de caractéristiques afin d'améliorer la réutilisation des données. Revue d'Épidémiologie et de Santé Publique. 2023. doi : 10.1016/j.respe.2023.101465.

[2] Quindroit P, **Fruchart M**, Degoul S, Perichon R, Martignène N, Soula J, Marcilly R, Lamer A. Definition of a Practical Taxonomy for Referencing Data Quality Problems in Health Care Databases. Methods Inf Med. 2023. doi : 10.1055/a-1976-2371.

[3] Lamer A, **Fruchart M**, Paris N, Popoff B, Payen A, Balcaen T, Gacquer W, Bouzillé G, Cuggia M, Doutreligne M, Chazard E. Standardized Description of the Feature Extraction Process to Transform Raw Data Into Meaningful Information for Enhancing Data Reuse : Consensus Study. JMIR Med Inform. 2022. doi : 10.2196/38936.

Publications en dehors de la thèse

[1] **Fruchart M**, El Idrissi F, Lamer A, Belarbi K, Lemdani M, Zitouni D, Guinhouya BC. Identification of early symptoms of endometriosis through the analysis of online social networks : A social media study. Digit Health. 2023. doi : 10.1177/20552076231176114.

[2] **Fruchart M**, Verdier L, Beuscart JB, Lamer A. Publication Dynamics on Social Media During the Orpea Nursing Homes Scandal : A Twitter Analysis. Stud Health Technol Inform. 2023. doi : 10.3233/SHTI230191.

[3] **Fruchart M**, Guinhouya B, Pelayo S, Vilhelm C, Lamer A. Jupyter Notebooks for Introducing Data Science to Novice Users. Stud Health Technol Inform. 2022. doi : 10.3233/SHTI220598.

Autres contributions

[1] M. Mammar, C. Saint-Dizier, **M. Fruchart**, A. Lamer. Étude automatisée du subreddit 'besoindeparler' - Comparaison des habitudes de publication entre les discussions sur la santé mentale et des sujets à connotations positives. Journal of Epidemiology and Population Health. 2024. doi : 10.1016/j.jeph.2024.202375.

[2] El Idrissi F, **Fruchart M**, Belarbi K, Lamer A, Dubois-Deruy E, Lemdani M, N'Guessan AL, Guinhouya BC, Zitouni D. Exploration of the core protein network under endometriosis

symptomatology using a computational approach. Front Endocrinol (Lausanne). 2022. doi : 10.3389/fendo.2022.869053.

[3] Lamer A, **Fruchart M**, Paris N, Popoff B, Payen A, Balcaen T, Gacquer W, Bouzillé G, Cuggia M, Doutreligne M, Chazard E. Standardized Description of the Feature Extraction Process to Transform Raw Data Into Meaningful Information for Enhancing Data Reuse : Consensus Study. JMIR Med Inform. 2022. doi : 10.2196/38936.

[4] Lamer A, Oubenali N, Marcilly R, **Fruchart M**, Guinhouya B. Master's Degree in Health Data Science : Implementation and Assessment After Five Years. Stud Health Technol Inform. 2022. doi : 10.3233/SHTI220906.

[5] Lamer A, Al Massati S, Saint-Dizier C, Fares E, Chazard E, **Fruchart M.** Data Management for Health Data Reuse : Proposal of a Standard Workflow and a R Tutorial with Jupyter Notebook. Stud Health Technol Inform. 2022. doi : 10.3233/SHTI220912.

[6] Patel H, Patel R, Zitouni D, Guinhouya B, **Fruchart M**, Lamer A. Automated Twitter Extraction and Visual Analytics with Dashboards : Development and First Experimentations. Stud Health Technol Inform. 2022. doi : 10.3233/SHTI220562.

Communications orales

[1] Description of a French Population of Diabetics Treated Followed up by General Practitioners, *Medical Informatics Europe*, Göteborg, Suède, Mai 2023.

[2] Implementation of a Data Warehouse in Primary Care : First Analyses with Elderly Patients, *Medical Informatics Europe*, Nice, France, Mai 2022.

Poster

[1] Publication Dynamics on Social Media During the Orpea Nursing Homes Scandal : A Twitter Analysis, *Medical Informatics Europe*, Göteborg, Suède, Mai 2023.

[2] Identification of early symptoms of endometriosis through the analysis of online social networks : a social media study, *Congrès clinique et scientifique annuelle*, Ottawa, Canada, Juin 2023.

ChatGPT a été utilisé pour faciliter la formulation et la correction de ce manuscrit.

Table des matières

Acronymes

Achilles Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems
ADO Anti-Diabétique Oral
ALD Affection à Longue Durée
ANS Agence du Numérique en Santé
API Application Programming Interface (interface de programmation d'application)
ATC Anatomique, Thérapeutique, Chimique
Athena Automated Terminology Harmonization, Extraction, and Normalization for Analytics

CCAM Classification commune des Actes Médicaux
CDM Common Data Model (Modèle de données commun)
CIC-IT Centre d'Investigation Clinique - Innovation Technologique
CIM-10 Classification internationale des maladies-10ème révision
CIP Code Identifiant de Présentation
CNAM Caisse Nationale de l'Assurance Maladie
CNIL Commission nationale de l'informatique et des libertés

DGOS Direction Générale de l'Offre de Soins
DMP Dossier Médical Partagé
DPI Dossier Patient Informatisé
DREES Direction de la Recherche, des Études, de l'Évaluation et des Statistiques
DSE Dossier de Santé Électronique

EDS Entrepôt de Données de Santé
EDSH Entrepôt de Données de Santé Hospitalier
ETL Extract-Transform-Load

HAS Haute Autorité de Santé
HDS Hébergeur de Données de Santé

LOINC Logical Observation Identifiers Names & Codes

MG Médecin généraliste
MSP Maison de Santé Pluridisciplinaire

NABM Nomenclature des Actes de Biologie Médicale
NLP Natural Language Processing

OHDSI Observational Health Data Sciences and Informatics
OMOP Observational Medical Outcomes Partnership
OMS Organisation Mondiale de la Santé

PMSI Programme de Médicalisation des Systèmes d'Information

RegEx Regular Expression
RGPD Règlement Général sur la Protection des Données
ROSP Rémunération sur Objectif de Santé Publique
RxNorm normalized medical prescription

SI Système d'Information

SNDS Système National des Données de Santé
SNIIRAM Système National d'Information Inter-régimes de l'Assurance Maladie
SNOMED-CT Systematized Nomenclature of Medecine - Clinical Terms

UCUM Unified Code for Units of Measure

XML eXtensible Markup Language

Liste des figures

Liste des tableaux

Introduction

Chapitre 1

Introduction

Sommaire

1.1 Réutilisation des données de santé

1.1.1 - Informatisation et protection des données sensibles

Au cœur du système de santé, le patient génère des données tout au long de sa vie dans des environnements variés : hôpitaux, cabinets médicaux, laboratoires d'analyses, pharmacies, voire sur internet. Ces données, complémentaires, permettent de tracer son parcours de soins (Figure 1.1).

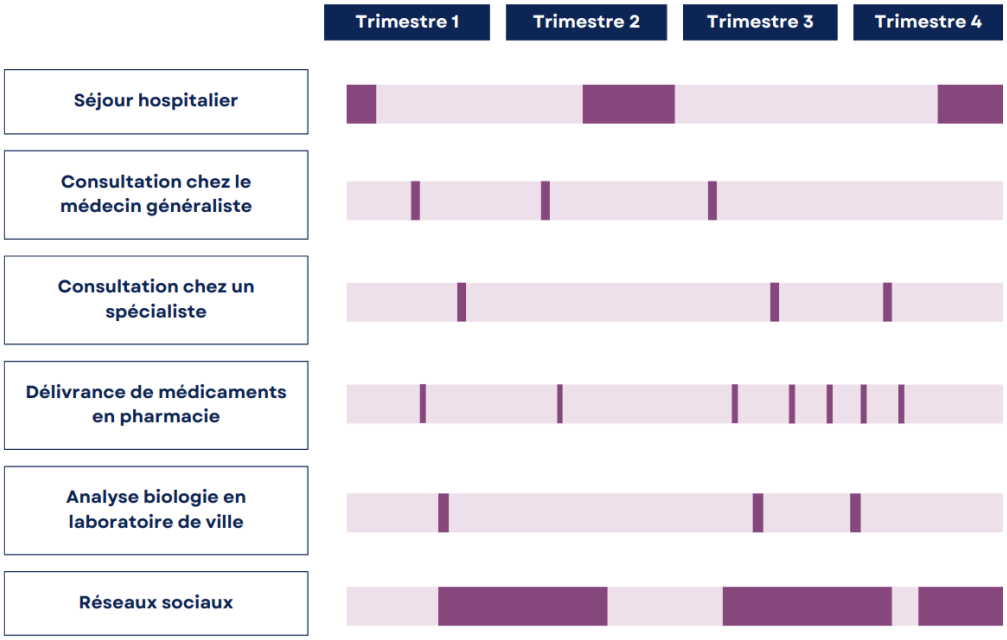


FIGURE 1.1 – Collecte des données de santé sur différents milieux au cours d'une année pour un patient. *En violet foncé : moment de collecte d'information de santé pour le milieu correspondant.*

Dans le contexte actuel de numérisation et de dématérialisation des informations, les données de santé ont connu une transformation majeure, tant au niveau de leur mode de saisie que dans les possibilités d'analyse [1]. En effet, la loi 2016-41 du 26 janvier 2016 visant à moderniser le système de santé, marque un tournant pour les professionnels de santé dans le suivi des patients car elle encadre les conditions d'archivage numérique des dossiers médicaux papiers [2]. Cette loi permet de renforcer la prévention et la promotion de la santé, de faciliter les parcours de santé, d'innover pour garantir la pérennité du système de santé et de renforcer l'efficacité des politiques. Dans ce cadre, les Dossiers Patients Informatisés (DPI) ont été développés pour faciliter la saisie numérique des données médico-administratives et le stockage des données de patients dans les Systèmes d'Information (SI) [3-5]. Ces DPI permettent également aux médecins d'accéder aux informations personnelles de leurs patients. Cependant, ces DPI sont spécifiques à chaque établissement et peuvent varier d'un service de soins à un autre. Pour la contribution de plusieurs professionnels de santé dans un seul SI, les Dossiers

Médicaux Partagés (DMP) sont apparus. Chaque patient dispose de son propre DMP centralisé [4, 6]. Le DMP contribue à l'amélioration de la qualité des soins et à la coordination entre les professionnels de santé. Il est utilisé par l'ensemble des professionnels de santé impliqués dans le parcours de soins d'un patient, que ce soit lors de consultations en ville, d'hospitalisations ou de délivrances de médicaments en pharmacie [6].

Les données de santé archivées dans ces DPI sont sensibles car elles permettent l'identification directe d'un individu et de ses conditions médicales [7]. Chaque éditeur de logiciel doit donc en assurer la protection, en les stockant dans une infrastructure sécurisée garantissant leur intégrité, leur confidentialité et leur disponibilité. Pour cela, les Hébergeurs de Données de Santé (HDS) sont certifiés et assurent le stockage, la protection et la conformité des données aux réglementations du Règlement Général sur la Protection des Données (RGPD) [8, 9]. Le RGPD regroupe les législations sur la protection des données en respectant les droits de chaque individu et en restreignant les traitements de données aux organismes [10, 11]. Le RGPD est entré en vigueur en 2018 et s'assure du respect confidentiel lors du traitement des données personnelles de chaque pays membre de l'Union Européenne [12]. Les organismes ont le devoir de mettre en place des mesures de sécurité et de recueillir le consentement d'individus lors d'utilisation de données personnelles. Le RGPD offre des règles permettant d'assurer un cadre juridique dans le développement d'activités numériques par des professionnels. Ces derniers doivent garantir la transparence sur l'utilisation des données en détaillant les méthodes appliquées et en donnant la possibilité de limiter ou d'effacer les données non autorisées [11].
En France, la Commission nationale de l'informatique et des libertés (CNIL), créée à partir de la loi Informatique et Libertés du 6 janvier 1978, garantit le respect et la protection des données personnelles, qu'elles soient informatiques ou sur papier [13]. Cette commission veille au respect du RGPD au niveau national. Elle assure que l'analyse de données personnelles ne porte atteinte ni à l'individu, ni à sa vie privée, et met à disposition des recommandations et des guides pour aider les organismes à respecter les exigences de la RGPD. La CNIL a le pouvoir d'effectuer des contrôles et de sanctionner les organismes ne respectant pas la RGPD [10].

1.1.2 - Entrepôts de données de santé

Les logiciels recensant des données médicales peuvent être nombreux au sein d'un même établissement. Au sein d'un hôpital, il y a généralement un logiciel par service, adapté à la prise en charge du patient et à son suivi au cours de l'hospitalisation. Un logiciel de soins peut être utilisé simultanément par plusieurs professionnels de santé.
Les données enregistrées dans ces logiciels sont hétérogènes et reposent sur des technologies et des nomenclatures différentes, selon les choix des éditeurs [14]. Cette diversité rend l'interopérabilité entre logiciels difficile et complique la réutilisation des données de santé. De

plus, les schémas et structures des bases de données sont également la propriété des éditeurs. Récupérer et réutiliser directement les données de ces logiciels est critique, car cela pourrait compromettre leur utilisation en routine [15]. Cependant, les données collectées ne sont pas nécessairement conçues pour répondre à des questions de recherche, ces logiciels ont été initialement développés pour la gestion des soins ou des tâches administratives. Par conséquent, il est impossible de réaliser des analyses directement via ces outils [5].
Pour lever ces différentes barrières, les Entrepôts de Données de Santé (EDS) ont été développés pour stocker, analyser, et partager les données enregistrées initialement par des logiciels différents [16].

1.1.2.1 - Données hospitalières

Les données hospitalières contiennent des informations réutilisables sur les patients, la durée de leur séjour, leurs antécédents, les médicaments administrés et les résultats de biologie effectuées à l'hôpital.

Les premiers EDS sont apparus dans le milieu hospitalier [17], initialement à l'échelle d'un seul établissement [18, 19]. Ces Entrepôts de Données de Santé Hospitaliers (EDSH) incluent une minorité de la population nationale, étant donné qu'ils se limitent aux patients d'un seul hôpital. En revanche, ces EDSH regroupent les données de plusieurs logiciels d'un établissement [16]. Au niveau national, certains pays ont développé des EDSH pour étudier des évènements comme les sorties hospitalières [20] ou les risques de réadmissions [21-23] sur les données de la population du pays.
Cependant, ces bases de données ne contiennent que les informations des patients ayant séjourné dans l'hôpital, et les données collectées reflètent une période limitée de leur vie. Ainsi, il est impossible de savoir si ces patients ont été ré-hospitalisés dans d'autres établissements. Les soins reçus en dehors du milieu hospitalier, comme ceux en ville, ne sont également pas inclus dans ces EDSH.

1.1.2.2 - Données de la population nationale

Les données des bases de données nationales regroupent les données de remboursements d'actes, de facturations, de médicaments délivrés en pharmacie et de décès. Le Système National des Données de Santé (SNDS) est une base de données exhaustive qui couvre presque toute la population française [24, 25]. Géré par la Caisse Nationale de l'Assurance Maladie (CNAM), le SNDS regroupe les données de facturation de l'assurance maladie (extraites du Système National d'Information Inter-régimes de l'Assurance Maladie (SNIIRAM)), les données de facturation des hôpitaux (base du Programme de Médicalisation des Systèmes d'Information (PMSI)), les données de décès (base de l'Inserm CépiDC) et les données relatives au handicap [26]. Les

données du SNDS sont mises à disposition à des fins de recherches et d'études évaluant la prise en charge médico-sociale, les trajectoires de soins ou encore les tendances épidémiologiques [27].
Des EDS de plusieurs établissements ou à l'échelle nationale ont été implémentés, offrant la possibilité d'étudier une population plus large [28, 29]. Les pays nordiques bénéficient de registres nationaux de données de remboursements [30], de décès [31], de données spécifiques à une pathologie comme le diabète [32], le cancer [33] ou les maladies infectieuses [34].

1.1.2.3 - Données issues des réseaux sociaux

Les réseaux sociaux et les forums ont été conçus pour faciliter la collaboration et l'interaction des utilisateurs d'internet. Les réseaux sociaux, comme Twitter, Facebook et Instagram, se distinguent par des applications de création de contenus, d'échange de messages et de vidéos [35]. Les forums, comme Doctissimo.fr ou Reddit.com, sont des plateformes en ligne composées d'un post principal, sur un sujet spécifique, et des réponses associées publiées par plusieurs utilisateurs. Le *web scraping* permet d'extraire les données en ligne en utilisant des programmes pour collecter des posts relatifs à un thème donné et des données qui peuvent être exploitées directement après extraction [36]. Les données sont directement réutilisées après l'extraction et un "*pre-processing*" (nettoyage) des données [37].

Sur les réseaux sociaux, les informations proviennent directement des utilisateurs, qui partagent librement leurs inquiétudes ou avis [38], leurs parcours de soins à venir ou mal vécus [39, 40], leurs symptômes ou conditions médicales [37, 41-44], ou encore les effets liés à la prise d'un traitement [45-47]. Les études de veille sur internet, comme l'infovigilance, permettent de surveiller la progression d'une maladie [40, 48] et de détecter les mauvaises informations [43, 49]. La pharmacovigilance est utilisée pour identifier les effets d'un médicament [50].

1.1.3 - Cas d'usages à partir des EDS

Les différents EDS décrits précédemment facilitent la réutilisation des données dans divers contextes tels que la recherche, l'évaluation de la qualité des soins, ou l'aide au pilotage de l'activité.

En recherche, ces données peuvent être exploitées pour analyser les conséquences postopératoires d'une intervention chirurgicale majeure [51] ou pour vérifier si l'interrogation d'un EDS permet une amélioration de la détection de l'anaphylaxie causée par des médicaments [52]. Les données des EDS peuvent concerner toutes les spécialités médicales, comme le cancer [53-55], les pathologies chroniques [56-60] ou la COVID-19 [61-64].

Les EDS contribuent également à la médecine personnalisée et à l'amélioration de la qualité des soins en permettant de prédire l'évolution de l'état de santé des patients. Cela inclut, par

exemple, une évaluation de l'efficacité d'une dose de médicament chez des patients atteints de maladies chroniques [65] ou de l'efficacité des traitements de deuxième intention visant à réduire la glycémie [66]. L'analyse des ressources disponibles pour la prise en charge permet aussi d'identifier des risques potentiels, tels que le risque suicidaire dans les services de santé mentale [67, 68], le risque de réadmission après une hospitalisation [69], ou encore le risque de chutes chez les patients hospitalisés [69].

Dans le domaine de la santé publique, les EDS assurent une surveillance épidémiologique grâce à la détection de signaux en temps réel ou à l'évaluation de l'impact des interventions lors d'une épidémie [70, 71]. Ils permettent également une surveillance pharmacologique, garantissant le suivi de la sécurité et de l'efficacité des traitements de première et deuxième intention chez les patients atteints de maladies chroniques [66, 72].

La réutilisation des données implémentée dans les EDS offre, par ailleurs, la possibilité de suivre l'activité d'une structure par le biais d'indicateurs spécifiques (par exemple, le nombre de consultations, le nombre de patients diabétiques, le nombre de médicaments prescrits ou le nombre de cas liés à une épidémie) [73, 74]. Ce suivi est souvent réalisé grâce à des tableaux de bord connectés aux EDS [75], qui présentent des indicateurs sous forme graphique (courbes, histogrammes) pour faciliter la visualisation des données et la prise de décision [74]. Les tableaux de bord sont généralement thématiques et permettent à l'utilisateur de filtrer les informations par population, par professionnel de santé ou par période. À chaque nouvelle intégration de données dans l'EDS, les graphiques et indicateurs sont automatiquement mis à jour [75]. Une interface ergonomique et une utilisation intuitive sont essentielles pour garantir l'efficacité de ces outils [76].

Néanmoins, la réutilisation des données de santé se heurte à de nombreux obstacles en termes de structure et de qualité de la donnée (i.e., extraction d'informations pour les données non structurées) [77-80], de sécurité et de droit d'utilisation des données [78], de volume et de disponibilité des données [81], ou d'extraction complexe et d'identification de pathologies [82].

À retenir : La réutilisation des données de santé

- Le processus de **collecte et de numérisation des données** de santé est réalisé tout au long de la vie du patient.
- Ces données sont issues dans **différentes sources** : les hôpitaux, les cabinets de ville, les bases de données nationales et sur internet (réseaux sociaux, forums).
- **Les EDS** collectent et stockent les données de plusieurs logiciels dans un **format homogène**.
- La réutilisation des données des EDS permet d'**évaluer l'activité** et la **qualité des soins** d'un établissement, d'aider au **pilotage de l'activité** et d'analyser les données à des **fins de recherche**.

1.2 Standardisation des données

1.2.1 - Processus Extract-Transform-Load

Les EDS sont alimentés par un processus Extract-Transform-Load (ETL). Ce dernier permet de convertir les données brutes et hétérogènes en informations exploitables et documentées avec un vocabulaire commun facilitant leur exploitation [83-85]. Les ETL sont composés de trois phases (Figure 1.2) [86, 87] :

1. *extract* pour l'extraction des données,
2. *transform* pour la transformation,
3. *load* pour le chargement des données.

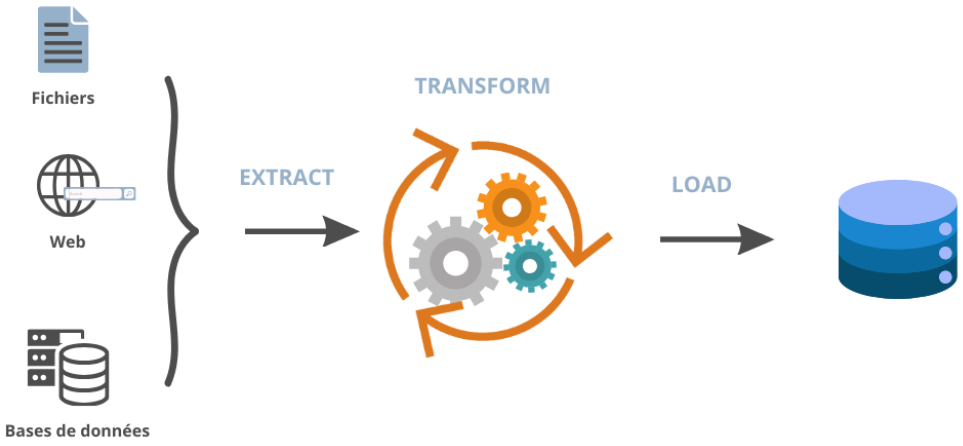


FIGURE 1.2 – Pipeline de la définition d'un processus ETL.

1.2.1.1 - Extraction

L'extraction des données consiste à récupérer les données de diverses sources, comme un logiciel, une Application Programming Interface (interface de programmation d'application)s (APIs), un ensemble de fichiers plats ou une application. L'intégralité des données ou les données pertinentes sélectionnées peuvent être extraites.

Les données extraites sont structurées ou non structurées. Les données structurées sont organisées selon un format et un modèle prédéfini [88] et sont généralement stockées dans des bases de données relationnelles ou des tableurs, sous forme de lignes et de colonnes [88, 89]. Chaque colonne représente un attribut (ou variable) et chaque ligne correspond à un enregistrement unique. Ce type de données se présente souvent sous forme de valeurs numériques, pour indiquer un poids par exemple, ou de codes, pour identifier un médicament et en documenter l'administration au patient.

En revanche, les données non structurées ne suivent pas de format ou modèle prédéfini, ce qui les rend plus difficiles à collecter, stocker et analyser avec des méthodes classiques [88, 89]. Contrairement aux données structurées, elles ne respectent pas de schéma spécifique, offrant ainsi plus de flexibilité, mais compliquant leur gestion et leur traitement. Les comptes-rendus d'hospitalisation en sont un exemple : ils contiennent du texte libre, consignant les informations de l'hospitalisation, avec des données cliniques telles que les diagnostics principaux, les antécédents médicaux et les médicaments administrés [83, 90]. Les données structurées et non structurées peuvent être conservés dans le processus ETL, mais nécessitent des transformations adaptées à leur format.

1.2.1.2 - Transformation

Une fois extraites, les données sont transformées, nettoyées et formatées, en garantissant leur qualité et leur homogénéité. Les transformations peuvent être de deux types : (1) transformation sémantique, (2) transformation structurelle.
La transformation sémantique consiste à transformer le vocabulaire utilisé localement (i.e., utilisé par le logiciel ou la source d'extraction des données) en un vocabulaire standard (i.e., utilisé par une communauté plus large) [91]. Lors de cette étape, les variations sémantiques d'un même terme (par exemple, les synonymes, les fautes de frappes, les traductions) seront alignées à un seul terme standard, ce qui facilitera par la suite l'interrogation de la base de données (Figure 1.3).

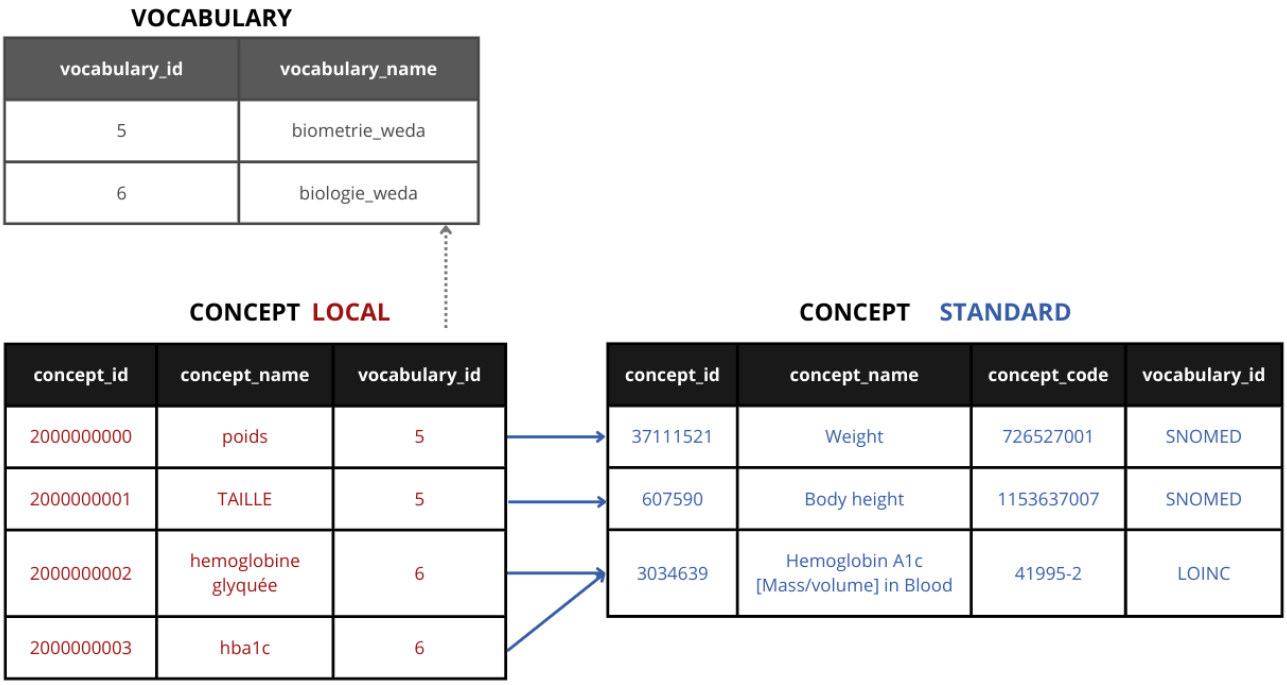


FIGURE 1.3 – Exemple d'alignements sémantiques de concepts locaux aux concepts standards. *Regroupement des termes liés à l'hémoglobine glyquée en un seul concept.*

Les opérations de transformation sémantique portent également sur les enregistrements eux-mêmes, comme pour la suppression des doublons, la correction des valeurs aberrantes ou la normalisation des valeurs numériques [92].
La transformation structurelle concerne l'organisation de la base de données (i.e., les noms des tables et les relations qui les lient), ainsi que les intitulés des tables et des colonnes. Elle permet d'aligner la structure d'une base de données à la convention d'un modèle commun.

1.2.1.3 - Chargement

Le chargement des données archive les données transformées dans un EDS, en appliquant des contraintes à respecter pour s'assurer de la véracité et de la plausibilité des données (par exemple, une date de prise de tension doit être comprise dans les dates de début et de fin de consultation du patient). Les contraintes de clés étrangères sont également appliquées pour s'assurer de la cohérence de la base de données (i.e., les mesures prises lors d'une consultation doivent être liées à une consultation existante).

1.2.2 - Modèle de Données Commun OMOP

Les Common Data Model (Modèle de données commun) (CDM) sont apparus pour pallier aux problèmes d'hétérogénéité des données. Parmi eux, le modèle Observational Medical Outcomes Partnership (OMOP) est un modèle de bases de données standard développé par le consortium Observational Health Data Sciences and Informatics (OHDSI) [93, 94]. Ce modèle impose une structure commune et un vocabulaire standard. La communauté OHDSI comporte plus de 3 000 collaborateurs répartis dans 80 pays et couvre les données de plus de 810 millions de patients [94].

Le modèle OMOP utilise des vocabulaires et des terminologies reconnus à l'international. Les concepts médicaux, tels que les diagnostics, les procédures ou les médicaments sont codés de manière uniforme et selon un vocabulaire commun. Cela permet de comparer et d'analyser ces concepts de manière cohérente, même lorsqu'ils proviennent de différents pays. Systematized Nomenclature of Medecine - Clinical Terms (SNOMED-CT) est la terminologie utilisée pour les diagnostics, normalized medical prescription (RxNorm) pour les médicaments et Logical Observation Identifiers Names & Codes (LOINC) pour les résultats de biologie.

Le modèle OMOP comporte 39 tables (Figure 1.4). Ces principales tables incluent les informations démographiques, les données de consultations des patients, de diagnostics, d'expositions aux médicaments, de procédures médicales, de mesures cliniques (par exemple, la créatinine ou l'hémoglobine glyquée) ou biométriques (par exemple, le poids, la taille ou la fréquence cardiaque) ainsi que d'observations ou de symptômes [94, 95]. Il existe également des tables

pour les métadonnées des établissements et des professionnels de santé, ainsi que des tables de vocabulaire pour standardiser les termes médicaux utilisés (i.e., pour l'alignement sémantique).

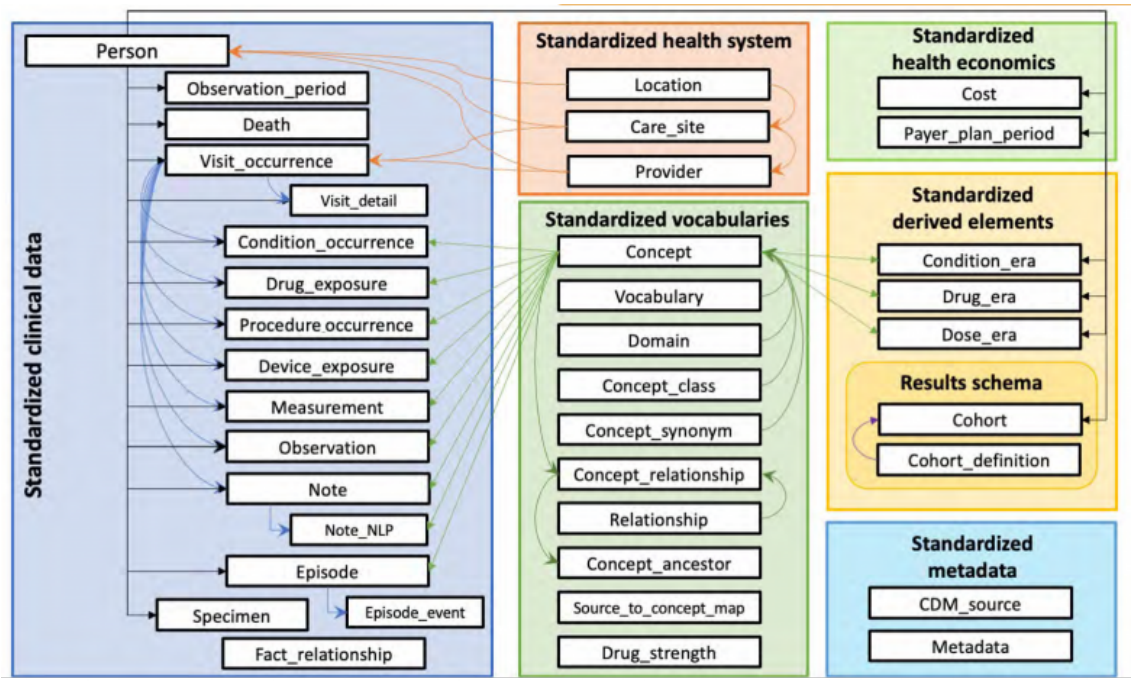


FIGURE 1.4 – Schéma des tables relationnelles du modèle commun OMOP [95].

Le modèle OMOP est composé de 394 variables (ou colonnes) dont 193 permettent d'identifier la table (nommées *xx_id* et composées d'identifiants uniques, comme la colonne *person_id*, qui affecte un identifiant unique à chaque ligne de la table *PERSON*). 101 variables permettent de standardiser les termes contenus dans les données (nommées *xx_concept_id* et liées à la table *CONCEPT* qui regroupe les vocabulaires utilisés). 43 variables conservent les données brutes (nommées *xx_source_value* pour conserver les termes locaux) [95].

En outre, le modèle OMOP est accompagné d'un ensemble d'outils et de méthodologies développés par la communauté OHDSI. Ces outils permettent d'avoir accès à des bases de données référençant le vocabulaire commun du modèle (i.e., Automated Terminology Harmonization, Extraction, and Normalization for Analytics (Athena)) [96], d'analyser les données brutes pour faciliter le développement de l'ETL (i.e., Rabbit-in-a-hat et White Rabbit) [97, 98] et d'évaluer la qualité des données implémentées dans le modèle (i.e., Achilles, Atlas) [99]. White Rabbit scanne la base de données source et génère un rapport sur chaque donnée (incluant le format et des statistiques de distribution). Rabbit-in-a-hat utilise ce rapport pour créer une visualisation des alignements entre les tables sources et celles du modèle, produisant ensuite un script de transformation. Achilles fournit un rapport récapitulatif des données intégrées dans le modèle, affichant visuellement des statistiques sur la population, les traitements et les mesures.
Des packages en langage R ont été développés à partir de la structure du modèle OMOP pour

favoriser les études statistiques. Le package *PatientLevelPredictor* permet de construire des modèles prédictifs autour du patient. Le package *Treatment Patterns* permet d'analyser des voies de traitement d'une population d'intérêt. Enfin *CohortSurvival* analyse les données de survie de manière fiable et reproductible [100].

L'objectif initial du modèle OMOP est de répondre à des questions pharmaco-épidémiologiques en se référant à des bases de données hospitalières ou médico-administratives. Depuis, de nouveaux types de données ont été intégrés, notamment dans les domaines de l'oncologie, de la microbiologie et de l'anesthésie-réanimation [75, 101, 102]. Des analyses observationnelles, réalisées à partir du modèle OMOP, ont étudié l'efficacité et les effets des traitements [103-105], ainsi que la faisabilité d'intégration de vocabulaires spécifiques à une pathologie dans le modèle [106]. Certaines études proposent des outils, développés sur ce modèle, comme un outil de désidentification des données [107], une application pour la visualisation des antécédents cliniques des patients [108] ou un package R pour calculer les taux d'incidence et la prévalence d'une pathologie [109].

À retenir : La standardisation des données

- L'ETL est un processus d'**extraction, de transformation et de chargement des données** dans un modèle de données commun (CDM).
- La **standardisation des données** est un processus permettant, à partir d'une donnée dans un format brut (format directement issu du logiciel), de **rendre homogène** la **structure** de la base de données, le vocabulaire utilisé et le **format** de la donnée.
- Le modèle de données commun OMOP est utilisé par **80 pays** et propose une structure et un vocabulaire **standards**, compréhensibles par toute la communauté.
- Le modèle OMOP permet de réutiliser les données pour répondre à des **questions scientifiques**, **fédérer** les analyses et **favoriser l'interopérabilité** des outils et des méthodes développés dans plusieurs pays.

1.3 Soins premiers

1.3.1 - Définition

Les soins premiers sont définis comme des soins personnalisés et centrés sur le patient. Un article du code de la santé publique (article L.1411-11) définit les soins premiers comme comprenant [110] :

— La prévention, le dépistage, le diagnostic, le traitement et le suivi des patients ;

— La dispensation et l'administration des médicaments, produits et dispositifs médicaux, ainsi que le conseil pharmaceutique ;

— L'orientation dans le système de soins et le secteur médico-social ;

— L'éducation pour la santé.

Les soins premiers représentent les soins de premiers recours assurant l'accessibilité, la continuité des soins, la globalité de la prise en charge, la coordination et la proximité des soins (i.e., les soins curatifs, préventifs, palliatifs et de réadaptation) [111-113]. Ces soins se font généralement en cabinet de ville, en Maison de Santé Pluridisciplinaire (MSP), en officine ou en pôle santé où de nombreux professionnels de santé exercent leurs fonctions (par exemple, les kinésithérapeutes, les infirmières, les médecins généralistes, les sage-femmes ou les pharmaciens). Les structures de soins premiers sont les premiers contacts avec le système national de santé pour chaque individu [114]. Depuis la mise en place de la stratégie nationale de santé en 2013, le gouvernement a placé les soins premiers en priorité des thématiques d'appels à projets de recherche sur les soins et offres de soins de la Direction Générale de l'Offre de Soins (DGOS) [115]. L'Organisation Mondiale de la Santé (OMS) définit les objectifs des soins premiers comme permettant l'intégration de la santé pour tous, la diminution de la disparité sociale en termes de santé et l'organisation du système par rapport aux besoins de la population [116].

1.3.2 - Médecine générale

La discipline de premier contact avec le système de soins est la médecine générale. Elle permet un accès non limité à tous les profils de patients. Selon la Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (DREES), la France comptait plus de 99 500 médecins généralistes au 1er janvier 2023, soit 43 % de la totalité des médecins [117, 118]. La médecine générale est en collaboration avec d'autres disciplines de soins premiers afin d'utiliser la globalité des ressources du système de soins et d'orienter les patients vers les spécialités les plus adaptées à leur problématique. Selon les besoins des patients, le médecin généraliste assure la continuité des soins longitudinaux [119]. Afin de coordonner au mieux les soins premiers, l'accès aux soins en France est centré autour d'un professionnel de santé, le médecin

traitant. Cette coordination permet un suivi médical et une prévention personnalisée de la prise en charge de chaque patient sur une longue période. Le médecin traitant est le médecin généraliste référent du patient qui doit être désigné à 16 ans et déclaré sur le DPI du patient [112]. Le médecin traitant a l'obligation d'orienter et de suivre son patient tout au long de son parcours de soins. Il garantit une gestion efficace du système de soins en délivrant des soins de qualité [112, 120]. Le médecin traitant doit également initier un protocole de soins pour les patients atteints d'Affection à Longue Durée (ALD) [121]. Lors d'une consultation, le Médecin généraliste (MG) doit archiver plusieurs informations médicales dans le dossier de son patient afin d'assurer un suivi de qualité. Une consultation se déroule en quatre étapes :

— L'interrogatoire : le patient rapporte le motif de consultation, les symptômes potentiels et son mode de vie ;
— L'examen : le MG examine le patient et prend quelques mesures biométriques (poids, taille, fréquence cardiaque ou autre en fonction du motif de la consultation) ;
— La conclusion : le MG établit un diagnostic et une conclusion sur la condition de son patient ;
— La prescription : le MG peut prescrire une ordonnance contenant un ou plusieurs médicaments selon la condition de son patient. Il peut également prescrire des soins infirmiers ou des rendez-vous chez d'autres spécialistes.

La Rémunération sur Objectif de Santé Publique (ROSP) est un dispositif entré en vigueur en France, en 2017, selon les articles 27-1 et 27-6 et à l'annexe 16 de la Convention médicale du 25 août 2016 [122]. Le but est d'inciter les MG à améliorer la qualité des soins donnés à leurs patients et à optimiser leur prise en charge [123]. Cette rémunération complémentaire repose sur des indicateurs définis par la CNAM incluant la prévention et le dépistage, le suivi des pathologies chroniques, l'efficience des prescriptions et la coordination des soins [124]. Les indicateurs de la ROSP concernent les médecins traitants des enfants, des adultes et les médecins spécialistes (en cardiologie, gastro-entérologie et hépatologie, endocrinologie et diabétologie) [125]. Pour les indicateurs de la ROSP adultes, 29 indicateurs sont répartis en sous-thèmes [124].
Les MG reçoivent une rémunération proportionnelle à leur performance, calculée à partir d'indicateurs ayant chacun un objectif et un seuil intermédiaire à atteindre. Plus un indicateur est proche de l'objectif, plus la rémunération sera élevée. Les indicateurs sont calculés sur la base des données de remboursement de l'assurance maladie, et concernent les patients ayant déclaré un médecin généraliste comme médecin traitant avant le 31 décembre de l'année passée. Les indicateurs de la ROSP se calculent une fois l'année terminée. La rémunération est perçue à la fin du 1er semestre de l'année N+1 [124].
Cependant, la déclaration des indicateurs de la ROSP auprès de l'assurance maladie impose

une charge administrative considérable pour les MG et les met sous pression pour atteindre les objectifs. Les MG n'ont pas la possibilité de vérifier les chiffres de la ROSP car ils n'ont pas accès aux données de remboursements et sont pénalisés lorsque leurs patients ne réalisent pas un acte prescrit. La ROSP est vouée à changer dans les années à venir et sera remplacée par une liste d'indicateurs et d'objectifs à atteindre.

1.3.3 - Maisons de santé et systèmes d'information des soins premiers

Les médecins de ville, en application de l'Article L.1111-15 du Code de la Santé Publique, ont désormais l'obligation d'alimentation du DMP et d'envoi par messagerie sécurisée des données sensibles des patients [126]. Les MSP sont des structures de soins de proximité regroupant plusieurs professionnels de santé de soins premiers (par exemple, MG, pharmaciens, sages-femmes, dentistes, infirmiers ou spécialistes) [127-129]. Ces structures assurent l'exercice coordonné grâce à une collaboration pluridisciplinaire et des échanges sur les parcours de soins des patients suivis. Les patients bénéficient d'une prise en charge globale et continue [128]. Chaque MSP regroupe environ 10 à 15 professionnels. En France, 2 501 MSP sont en fonction, dont 231 dans les Hauts-de-France (Figure 1.5) [127, 129]. Plus de 3 000 professionnels de santé de la région Hauts-de-France exercent dans une MSP dont 857 MG [127].

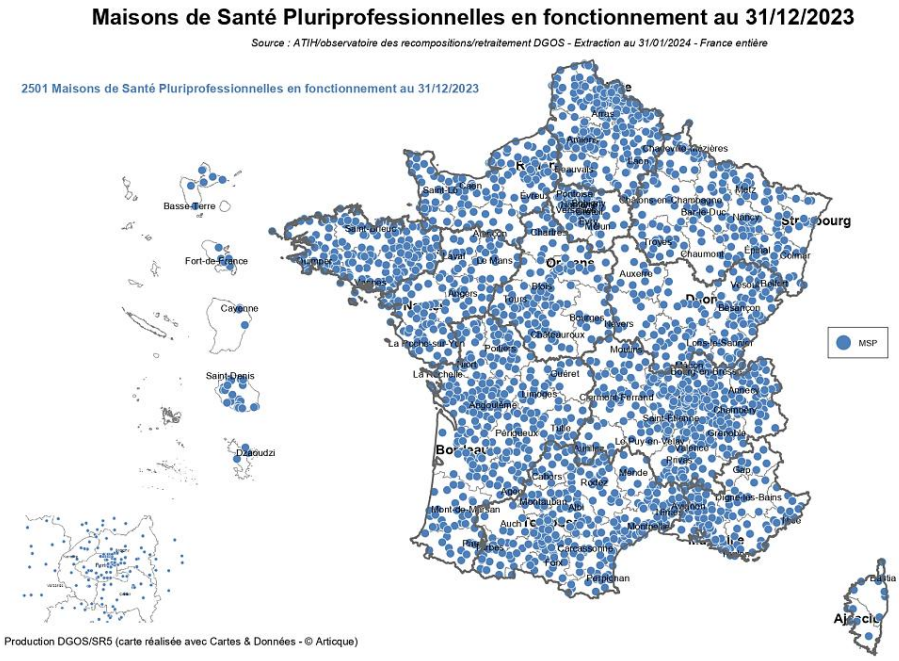


FIGURE 1.5 – Répartition des MSP en France au 31/12/2023 [129].

La numérisation des données de dossiers médicaux et l'apparition des DPI ont conduit au développement de logiciels spécialisés dans la saisie des informations de patients dans les MSP, couvrant essentiellement les données des soins premiers. Dans le logiciel, le MG saisit les informations cliniques des différentes étapes de la consultation (i.e., interrogatoires, examens, conclusions et prescriptions).

En 2012, le label « e-santé Logiciel Maisons et Centres de santé » mis en place par l'Agence du Numérique en Santé (ANS) permet aux MSP d'identifier les logiciels adaptés à leur besoin [130]. En France, 25 logiciels sont labellisés e-santé [131]. Ce label rassemble un ensemble de critères assurant que le SI réponde aux besoins des professionnels de santé en exercice coordonné. Cette labellisation est facultative pour les éditeurs mais permet le choix à de nombreux utilisateurs. Les critères se caractérisent par [131] :

- Adéquation aux besoins des professionnels par les fonctionnalités suivantes :
  - exercice individuel de chaque professionnel ;
  - coordination pluriprofessionnelle (accès aux dossiers patients, communications et réunions pluriprofessionnelles) ;
  - pilotage de l'activité (suivi de l'activité de la structure et report vers les institutionnels) ;
  - gestion de la structure (logistique, ressources).
- Conformité à la réglementation : conditions d'hébergement et sécurité des données (agrément HDS), numérisation des feuilles de soins, outil d'aide à la prescription.
- DMP-Compatibilité : capacité de création, de consultation et d'alimentation du DMP.

En 2020, le Ségur de la santé lance un plan visant à moderniser le système de santé et à rendre les SI interopérables [132]. L'un des objectifs principaux du Ségur de la santé est de garantir que chaque professionnel de santé dispose des outils numériques nécessaires pour échanger et saisir de manière sécurisée les données de santé [132-134]. Le Ségur de la santé encourage également l'innovation dans le domaine de la e-santé en facilitant la mise en place de nouveaux services numériques et en soutenant la recherche et le développement. Ainsi, les logiciels e-santé deviennent des outils de gestion et de coordination [134]. En 2024, pour faciliter aux MG la consultation des informations du DPI, le Ségur a indiqué certaines mises à jour à implémenter dans les logiciels, comme la simplification de la consultation des informations et le renforcement de la sécurité des données [135].

1.3.4 - Réutilisation des données de soins premiers

La réutilisation des données de soins premiers permet d'étudier les informations de patients atteints de maladies neuromusculaires [136], de diabète [137, 138], de maladies dermatologiques [139, 140], de maladies pulmonaires [141], de cancers [142], d'infections urinaires [143] ou des patients âgés [144].

Pour faciliter la réutilisation, des principaux projets nationaux d'EDS en soins premiers sont apparus comme le Clinical Practice Research Datalink et le Health Improvement Network au Royaume-Uni [145, 146], les entrepôts de données de la Veterans Administration aux États-Unis [147] et le Canadian Primary Care Sentinel Surveillance Network au Canada [148]. Ces EDS ne suivent pas les conventions d'un modèle de données commun. Cependant, un projet britannique exploite une base de données nationale (BioBank) pour normaliser les données de vaccination dans un modèle de données commun [149]. Un second projet, australien, utilise les données de trois logiciels de soins premiers pour les implémenter au format OMOP. Les données de prescriptions de millions de patients ont été intégrées au modèle [150]. Ces EDS spécifiques à un type de données permettent néanmoins leur réutilisation.

1.3.5 - Les objectifs du projet PriCaDa

Le projet PriCaDa (primary care data warehouse), financé par la DGOS par le biais de l'appel à projet ReSP-Ir 2021 a l'ambition d'implémenter un EDS en ambulatoire (ou soins de ville regroupant les consultations en cabinet de ville, hors hôpital). Le projet est porté par le Département de médecine générale de l'université de Lille (Dr Matthieu Calafiore), et concerne quatre MSP (Métropole Européenne de Lille), l'ULR2694 METRICS (Pr Jean-Baptiste Beuscart, Pr Emmanuel Chazard, Dr Antoine Lamer et Dr Paul Quindroit), le CHU de Lille (Pr Grégoire Ficheur) et un partenaire industriel (WEDA, éditeur de logiciel médical labellisé e-santé) [151]. Les MSP de Wattrelos, Lille-Moulins, Guesnain et Tourcoing sont les quatre MSP du projet et contiennent respectivement les dossiers de 20 445, 26 146, 7 058 et 63 356 patients qui ont consulté un MG entre 1997 et 2023. Pendant cette période, le site de Lille-Moulins comptait dix MG, Guesnain quatre MG, Tourcoing et Wattrelos en comptaient huit chacun. Ce projet permet aux professionnels de santé des MSP d'obtenir une vue d'ensemble de l'activité de leur structure ainsi qu'un suivi personnalisé et détaillé de leurs patients. L'intégration de plusieurs MSP au projet permet de comparer les conditions médicales et les prises en charge des patients des différents milieux. Ma thèse a fait parti de ce projet et a été financée par le CHU de Lille (i.e., PriCaDa) et par le Centre National pour la Médecine de Précision des Diabètes (PreciDIAB), également partenaire du projet.

À retenir : Les soins premiers

- Les soins premiers sont des **soins de premiers recours**, centrés sur le patient et regroupent les disciplines exercées en cabinet de ville.
- Le MG est le professionnel référant de soins premiers et **suit ses patients sur une longue période** (plusieurs années). Suivant le respect des recommandations de suivi des patients chroniques, il peut percevoir **une ROSP** versée par l'assurance maladie.
- Plusieurs professionnels de soins premiers peuvent exercer dans **une MSP** et utiliser un logiciel commun **labellisé e-santé**. Ce logiciel permet de stocker les données des patients de la MSP.
- Les données de soins premiers peuvent être extraites des logiciels et stockées dans des **entrepôts de données** pour être réutilisées.
- Le projet français **PriCaDa** vise à créer un entrepôt de données de soins premiers afin de proposer aux professionnels des outils d'évaluation de leurs activités et de faciliter les projets de recherche.

1.4 Question de recherche

L'introduction donne une vue globale sur la réutilisation des données de santé. Les données de santé proviennent de diverses sources dont l'exploitation peut varier selon l'origine. Le milieu hospitalier est le plus avancé dans ce domaine, avec des logiciels qui stockent quotidiennement les données de milliers de patients. Les données de soins premiers sont collectées à chaque consultation et sont également stockées dans des logiciels de soins. Cependant, ces logiciels utilisés par les médecins dans leur pratique quotidienne ne permettent pas de réutiliser directement les données. De ce fait, les données de soins premiers, comme les données de remboursements françaises (i.e., le SNDS), sont principalement stockées dans des bases de données locales et souveraines, ce qui peut entraîner des difficultés pour l'interopérabilité des analyses et des outils en raison des variations dans leur structure.

1.4.1 - Questions

La question principale de cette thèse est :
Quelles sont les spécificités de la réutilisation des données de soins premiers et quelles stratégies permettent de gérer ces spécificités ?

Les questions secondaires de cette thèse sont :

1. Les données de soins premiers sont-elles compatibles avec un modèle de données commun (i.e., OMOP) ?
2. Comment faciliter l'implémentation des EDS dans plusieurs MSP ?
3. Quels outils peuvent être proposés aux professionnels de santé pour les aider dans la prise en charge de leurs patients ?
4. À quelles questions de recherche peut-on répondre en réutilisant les données de soins premiers ?

1.4.2 - Objectifs

Cette thèse répond aux quatre objectifs suivants :

Objectif 1 : Définir les spécificités de la réutilisation des données de soins premiers en comparaison de trois autres sources de données (EDSH, SNDS, réseaux sociaux).

Objectif 2 : Réussir à implémenter les données de soins premiers dans un CDM et s'affranchir du volume des MSP.

Objectif 3 : Permettre une aide à la prise en charge et une analyse des profils patients pour les professionnels de santé de soins premiers.

Objectif 4 : Montrer l'apport de la réutilisation des données de soins premiers dans le spectre de la recherche.

Standardisation des données de soins premiers vers le modèle de données commun OMOP

Chapitre 2

Standardisation des données de soins premiers vers OMOP

Sommaire

2.1 Contexte

L'objectif du projet PriCaDa est de réutiliser les données de soins premiers pour fournir aux professionnels de santé, une vue d'ensemble de leurs pratiques. Les données de la MSP de Wattrelos (Nord, France) ont été extraites du logiciel WEDA (WEDA, Montpellier, France). WEDA occupe la troisième place en terme de volume de ventes et est utilisé par plus de 20 000 professionnels de santé en France [152]. Un processus d'ETL a été mis en place pour alimenter un EDS de soins premiers (Figure 2.1) et proposer aux professionnels de santé un retour sur leur activité grâce à des tableaux de bord. Le développement de ce processus sera détaillé dans ce chapitre.

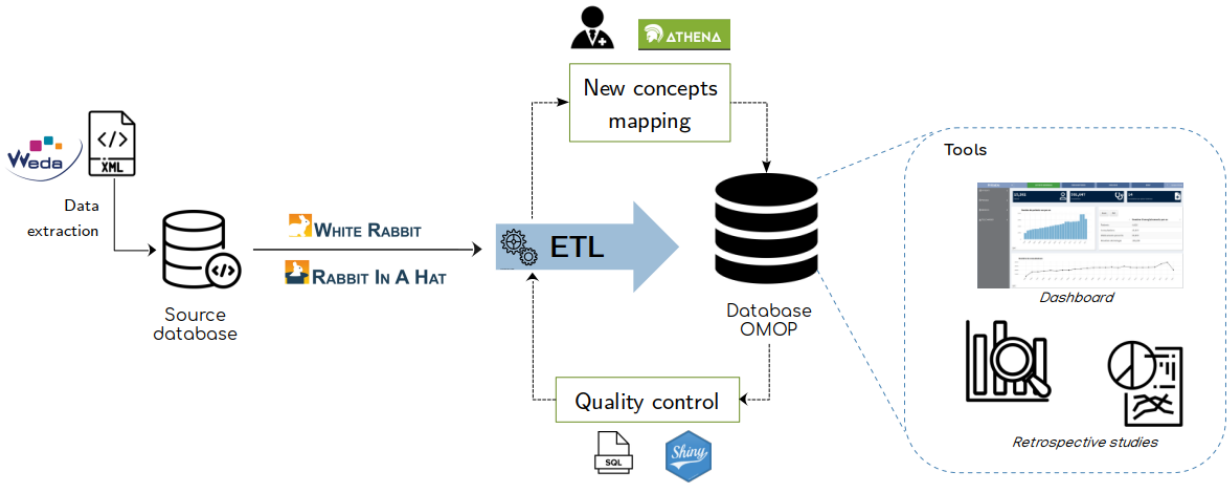


FIGURE 2.1 – Pipeline de transformation des données de soins premiers vers OMOP. *ETL : Extract-Transform-Load (extraction, transformation, chargement)*

La première étape du processus ETL consistait à extraire, classer et normaliser les attributs des fichiers bruts. Suite à l'extraction, les données ont suivi plusieurs transformations pour être intégrées au modèle OMOP. Les transformations concernaient la structure de la base de données, le vocabulaire utilisé pour caractériser les concepts médicaux, la résolution des problèmes de valeurs manquantes et de doublons. Une fois les données transformées, le chargement dans le modèle final s'est effectué en appliquant les contraintes et les relations entre les tables. Les données devaient respecter les conditions de la vie réelle (par exemple, les identifiants des patients de la table de consultation devaient exister dans la table des patients ou, une date de mesures prises lors de la consultation devait correspondre à la date de la consultation). Des métriques d'évaluation de la qualité des données ont été appliquées pour justifier le bon fonctionnement de l'ETL.

2.2 Matériels et méthodes

Les fichiers issus du logiciel WEDA étaient au format eXtensible Markup Language (XML) (Figure 2.2) [153]. Les XML sont composés d'éléments (par exemple, la borne <Patient>), d'attributs qui qualifient l'élément (par exemple, pour l'élément <Patient>, on peut avoir <Sexe> et <Date de naissance>) puis chaque attribut contient du texte (par exemple, l'attribut <Sexe> peut contenir "Femme"). Chaque composant avait sa place dans la base de données. L'élément désignait le nom d'une table, l'attribut le nom d'une colonne et le texte la valeur de la donnée. Chaque élément correspondait à un type d'enregistrement (i.e., un patient, une consultation, ou une prescription) ou à un type d'attribut qui caractérisait un enregistrement (i.e., une date, un poids, une valeur de biologie).

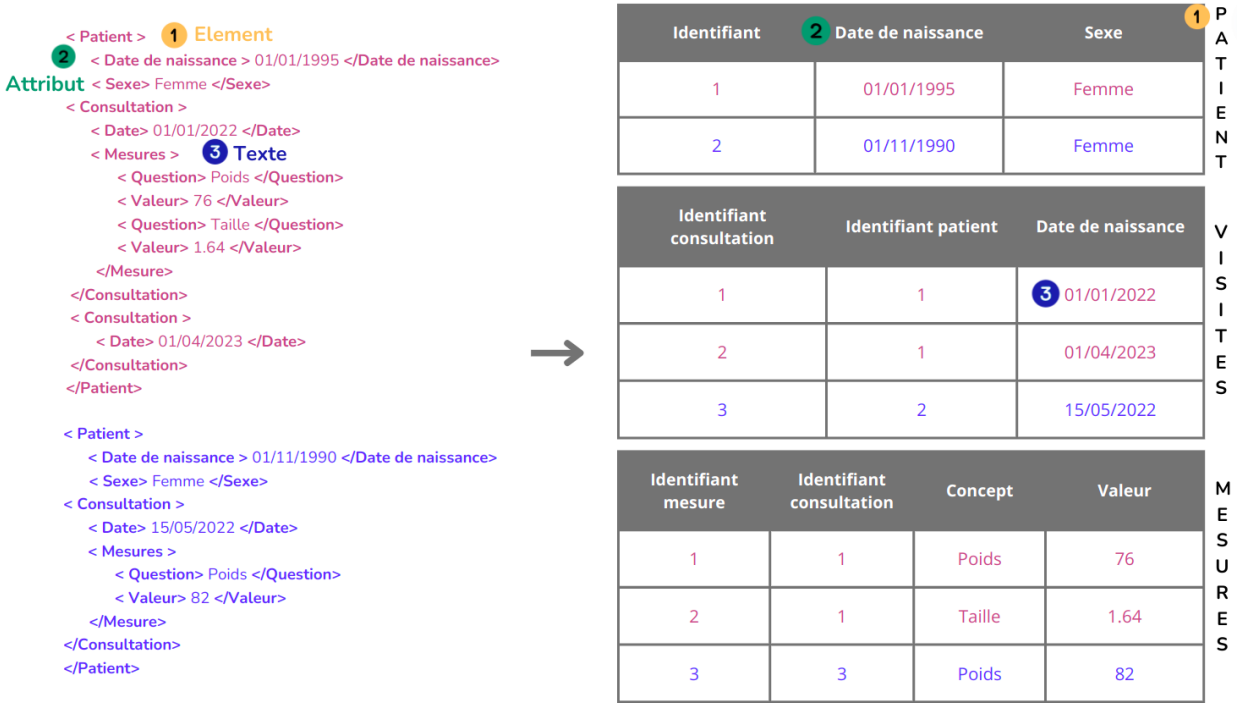


FIGURE 2.2 – Intégration des éléments, attributs et valeurs XML dans des tables. *1 : élément du XML correspondant aux noms des tables, 2 : Attributs du XML correspondant aux noms de colonnes, 3 : texte du XML correspondant aux valeurs.*

Après approbation par les professionnels de santé de la MSP, WEDA a exporté un fichier XML pour chaque patient, qui contenait toutes les informations saisies dans le logiciel.
Une consultation de soins premiers se déroule généralement en quatre temps (cf. Section 1.3.2 de l'introduction). À chaque étape, le MG recueillait des données sur l'état de santé du patient (Figure 2.3). Lors de l'interrogatoire du patient, le motif de la consultation était saisi sous forme de texte libre. Le MG enregistrait les signes cliniques et les symptômes également en texte libre. Durant l'examen, le MG pouvait réaliser des mesures biométriques et saisir les valeurs associées dans les champs appropriés du logiciel. Certains commentaires ou informations sur une vaccination étaient inscrits en texte libre. La documentation du diagnostic, s'il a été posé, se faisait aussi en texte libre. Concernant la prescription, chaque médicament était documenté avec un code Code Identifiant de Présentation (CIP), lié à la base de données Vidal [152].
Le dossier de chaque patient contenait des informations démographiques (i.e., le sexe, l'année de naissance, la ville de résidence et le pays de résidence), la date de la première consultation et le nom du MG référent du patient. Il contenait également les antécédents médicaux, documentés par des codes Classification internationale des maladies-10ème révision (CIM-10) ou en texte libre à la convenance du MG.
En dehors de la consultation, le MG pouvait également recevoir des informations cliniques de ses patients, comme des résultats d'analyses biologiques provenant des laboratoires de ville, des rapports cliniques de la part d'autres médecins spécialistes, ou des comptes rendus d'hospitalisation.

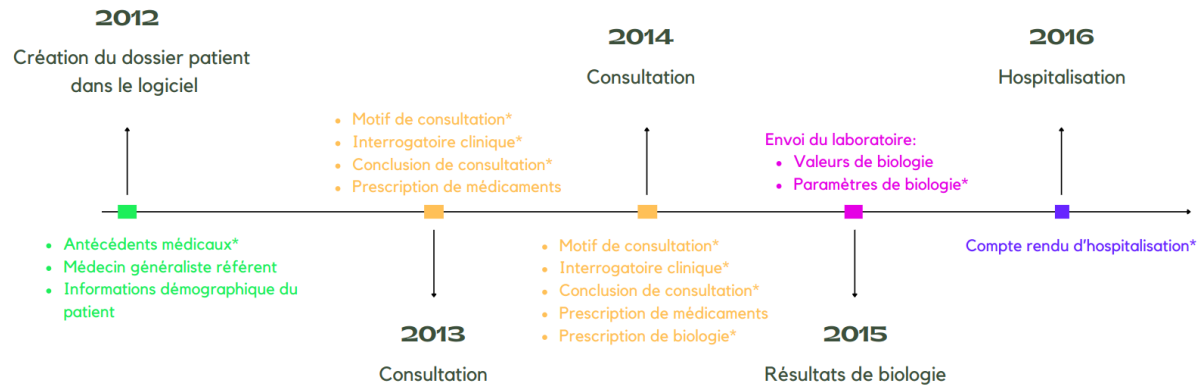


FIGURE 2.3 – Exemple des données du suivi d'un patient saisies dans le logiciel WEDA. ** : données saisies en texte libre (plusieurs informations consignées), en vert : informations du dossier du patient, en jaune : informations de la consultation, en violet : résultats du laboratoire, en bleu : compte-rendu extérieur au cabinet.*

La documentation des informations textuelles dépendait des habitudes d'utilisation du logiciel par chaque MG. Ainsi, certains MG pouvaient consigner toutes les observations dans un seul champ (par exemple, les concepts médicaux, le motif de la consultation, les symptômes ou les diagnostics), alors que d'autres utilisaient les champs du logiciel dédiés (Figure 2.4). En revanche, les données saisies dans les champs spécifiques du logiciel étaient structurées (par exemple, un champ dédié à la valeur du poids du patient contenait une information médicale directement accessible).

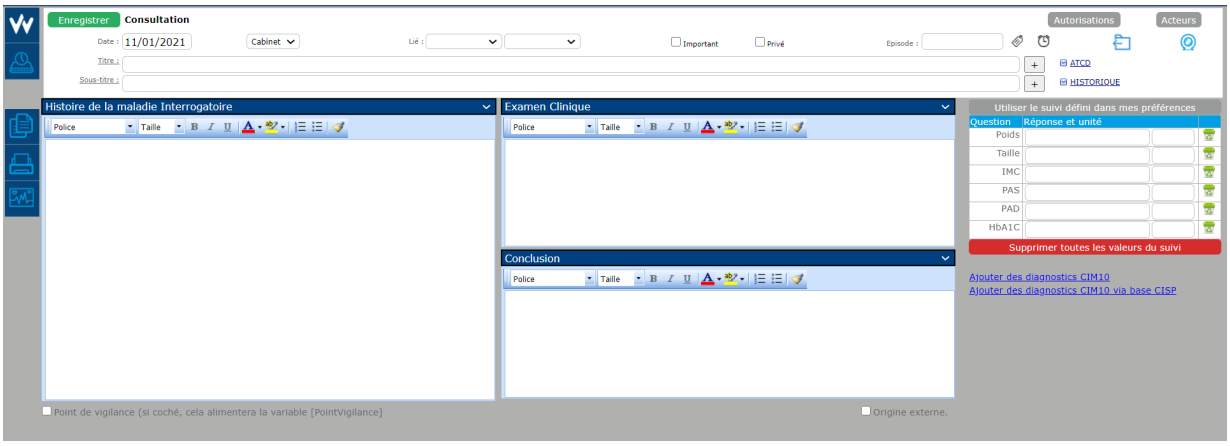


FIGURE 2.4 – Interface de consultation du le logiciel WEDA [152]. *La section "Titre" était initialement prévu pour renseigner le motif, "Interrogatoire" pour les symptômes, "Examen clinique" pour les mesures et commentaires et "Conclusion" pour le diagnostic. La section "Questions", à droite, collecte des données dans un format structurée pour les champs dédiés.*

2.2.1 - Extraction des données du logiciel WEDA

Une analyse exploratoire des fichiers a été réalisée pour sélectionner les données intéressantes et réutilisables. Les fichiers des patients ont été parcourus un à un par un programme en langage Python. Les données d'intérêt ont été extraites et intégrées dans une table dont les noms de colonnes correspondaient aux noms des attributs du fichier XML. Ces tables ont été intégrées dans un schéma PostgreSQL et constituaient une base de données source (i.e., base de données conservant le format et la nomenclature brute des données extraites du logiciel) (Figure 2.2).

2.2.2 - Transformation des données

Les étapes de transformation servaient à aligner la structure et le vocabulaire des données extraites pour s'adapter aux règles et à la nomenclature du modèle. Il existait deux types de transformations : la transformation sémantique et la transformation structurelle. La transformation sémantique concernait la donnée stockée dans chaque variable (i.e., donnée contenue dans une cellule d'un tableau). Dans ce type de transformation, les termes locaux ont été alignés aux vocabulaires standards disponibles dans l'outil Athena [96, 154]. Chaque MSP utilisait un vocabulaire local spécifique (i.e., dépendant du logiciel et du pays). La transformation structurelle concernait le nom des variables et le nom des tables de la base de données.

2.2.2.1 - Transformation sémantique

Les domaines médicaux utilisés en médecine générale regroupaient les médicaments et leur posologie prescrits (i.e., durée du traitement, nombre de renouvellements, dose et période de prise), les résultats de biologie des laboratoires de ville, les mesures biométriques prises lors de la consultation et les antécédents du patient. Chacun de ces domaines contenait des concepts décrivant l'expérience de soins d'un patient. Les concepts étaient les codes de chaque classification de vocabulaire, appelée terminologie, pour désigner un terme médical. Le type de concept était lié à chaque enregistrement pour indiquer la provenance de la donnée (par exemple, donnée issue de la consultation, d'une délivrance pharmaceutique ou d'un compte rendu par exemple). Par exemple, le concept "A10BX02" était utilisé dans la terminologie Anatomique, Thérapeutique, Chimique (ATC) pour désigner la Repaglinide.

Les terminologies suivantes sont intégrées dans le logiciel :

— CIM-10 pour les antécédents médicaux,

— CIP pour les médicaments prescrits,

— Termes biologiques en langue française pour les résultats de biologie dont le style grammatical dépendait de chaque laboratoire.

La transformation sémantique a permis l'alignement d'un concept issu d'une terminologie locale au concept correspondant d'une terminologie standard (Table 2.1).

Domaines	Terminologies locales	Terminologies standards
Biologie	Semi-structuré	LOINC
Biométrie	Semi-structuré	SNOMED-CT
Unités de mesure	Semi-structuré	UCUM
Médicaments	CIP	RxNorm
Antécédents médicaux	CIM-10	CIM-10

TABLE 2.1 – Terminologies standards du modèle OMOP associées aux domaines de concepts disponibles dans la base de données source

Quatre niveaux de difficulté concernant la transformation sémantique ont été définis :

1. Niveau 1 : pour les éléments d'une terminologie locale appartenant déjà à la terminologie standard OHDSI,
2. Niveau 2 : pour les éléments d'une terminologie locale pour lesquels les correspondances avec la terminologie standard existaient déjà dans le modèle,
3. Niveau 3 : pour les éléments d'une terminologie locale qui ne correspondaient pas à la terminologie standard et qui se présentaient sous un format structuré (un alignement manuelle était nécessaire),
4. Niveau 4 : pour les éléments d'une terminologie locale qui ne correspondaient pas à la terminologie standard et qui se présentaient sous un format non structuré.

Concernant les niveaux de difficulté 2 et 3, les concepts ont été associés manuellement et indépendamment par deux annotateurs. Un score de Kappa a été calculé et a mesuré le degré de consensus entre les annotateurs chargés de trouver un alignement à chaque concept [155]. Les éventuels désaccords dans les alignements ont été réglés par une troisième personne. Les concepts de soins premiers non alignés à une terminologie du modèle OMOP ont été chargés dans la table *CONCEPT*.
Le niveau 4, étant du texte libre, nécessitait une extraction des différents concepts consignés

dans une donnée. Le texte libre a été stocké sans transformation dans la table *NOTE*.

Le modèle OMOP propose plusieurs tables de vocabulaire contenant les terminologies standards et permettant de réaliser l'alignement sémantique. Ces tables sont *CONCEPT* (pour stocker tous les concepts locaux et standards), *CONCEPT_RELATIONSHIP* (pour l'alignement sémantique), *VOCABULARY* (pour identifier le type de chaque concept) et *RELATIONSHIP* (pour caractériser la relation entre les concepts) (Figure 2.5). À l'issue de l'alignement sémantique, les concepts standards ont été chargés dans la colonne *x_concept_id* de la table correspondante et associés aux concepts locaux (par exemple, dans la table *MEASUREMENT*, les concepts standards ont été chargés dans la colonne *measurement_concept_id*).

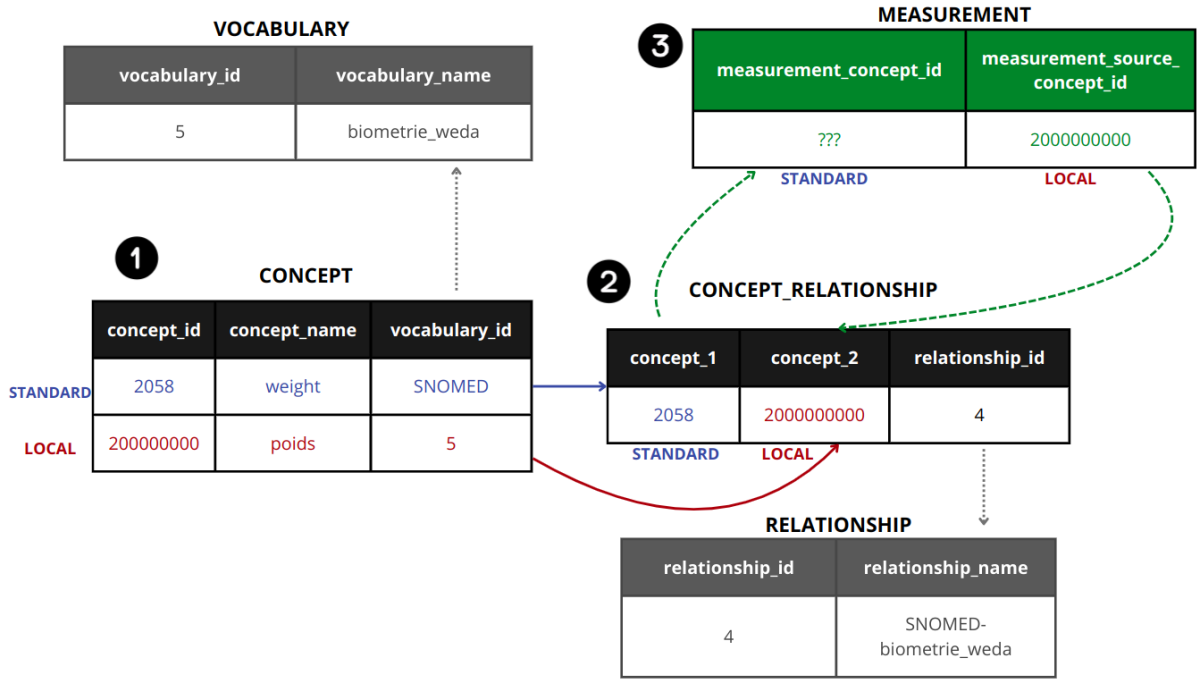


FIGURE 2.5 – Étapes de la standardisation du vocabulaire local avec le modèle OMOP. *1 : Chargement des concepts standards et des concepts locaux (identifiants supérieurs à 2000000000) dans la table CONCEPT. 2 : l'alignement sémantique entre les concepts locaux et standards a été chargé dans la table CONCEPT_RELATIONSHIP. 3 : le concept standard a été mis à jour dans la table source (colonne x_concept_id).*

2.2.2.2 - Transformation structurelle

La transformation structurelle a permis d'aligner les tables et les colonnes de la base de données source vers le modèle de données final, OMOP. Les variables et la structure des tableaux ont été transformées pour correspondre à la nomenclature du modèle OMOP.
La structure de la base de données source a été évaluée à l'aide de l'outil WhiteRabbit [98]. Cet outil a permis de créer un rapport détaillé sur chaque table, chaque attribut et chaque type

de données. L'outil Rabbit-in-a-hat a ensuite été utilisé pour rédiger les spécifications de la transformation structurelle et associer les tables et colonnes locales au modèle OMOP [97].

2.2.3 - Chargement et évaluation de la qualité des données

Après les transformations sémantique et structurelle, les données ont été chargées dans le modèle OMOP. Nous avons utilisé l'outil Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (Achilles), un outil d'évaluation et de visualisation de la qualité des données développé par la communauté OHDSI. Achilles a vérifié que les transformations respectaient les contraintes du modèle OMOP (i.e., conformités des liens entre les clés primaires et étrangères), les contraintes de vocabulaire (i.e., concepts standards complétés dans chaque table) et les règles de gestion (i.e., règles garantissant la cohérence et la chronologie des données en vie réelle).
Sur la base des tableaux fournis par Achilles, un tableau de bord a résumé l'évaluation de la qualité des données [156]. Ce tableau de bord a produit les distributions de données pour les concepts de chaque table. Kahn et al. ont mis au point un *framework* d'évaluation de la qualité des données pour leur réutilisation [157]. L'onglet "*overview*" du tableau de bord Atlas a fourni un résumé de la qualité des données du modèle final comportant le nombre total de réussites et d'échecs par catégorie de Kahn (Figure 2.7). La conformité décrivait « *la conformité de la représentation des données par rapport aux définitions de formatage, relationnelles ou informatiques internes ou externes* ». L'exhaustivité calculait « *les fréquences des attributs de données présents dans un ensemble de données sans référence aux valeurs des données* ». La plausibilité décrivait « *la crédibilité ou la véracité des valeurs des données* ». La vérification était une stratégie « *pour la source d'attentes ou de comparaisons des données des Dossier de Santé Électronique (DSE) sur la base de caractéristiques internes* » [157].

Les résultats de plusieurs requêtes, appliquées à la base de données, ont été comparés avec les données disponibles dans le logiciel WEDA pour tester la pertinence de l'utilisation des données de soins premiers dans le modèle OMOP. Les requêtes sur le logiciel ont été effectuées par un MG utilisant régulièrement le logiciel dans sa pratique clinique. Deux requêtes correspondaient à l'activité générale de la MSP, deux correspondaient aux données de prescription et une cinquième correspondait aux données des résultats de biologie.

L'EDS a été stocké dans une base de données PostgreSQL, en utilisant la version 5.4 du OMOP [95].

2.3 Résultats

2.3.1 - Extraction des données du logiciel WEDA

Les données ont été extraites en juillet 2021. Les profils de patients les plus anciens remontaient à 1997. Les données des mesures biométriques, des prescriptions de médicaments et des analyses de biologie étaient disponibles à partir de 2013. L'extraction des données de la MSP de Wattrelos contenait 18 395 dossiers de patients.

2.3.2 - Transformation des données

2.3.2.1 - Transformation sémantique

Au total, 17 domaines de concepts locaux ont été alignés aux concepts standards du modèle (Table 2.2). Les concepts spécifiques aux soins premiers ont été ajoutés à la table *CONCEPT* (n=10 221) et alignés aux concepts standards. Les alignements sémantiques ont été intégrés dans la table *CONCEPT_RELATIONSHIP* (n=9 432). Ces alignements représentaient plus de 80% des prescriptions de médicaments et plus de 90 % des antécédents médicaux. Seulement 4,9 % des mesures biométriques ont été alignées à des concepts standards. Ces alignements représentaient 96 % des enregistrements de la base de données. En effet, un petit nombre de concepts locaux étaient régulièrement utilisés (le poids, la taille et la fréquence cardiaque) et représentaient la majorité des enregistrements. Les autres concepts étaient saisis sous forme de texte libre. La forme grammaticale de ces concepts dépendait des habitudes de saisie des MG.

La terminologie LOINC a été utilisée pour l'alignement sémantique des résultats de biologie. Ces résultats étaient exprimés par des termes spécifiques aux laboratoires et par une unité de mesure. La ponctuation et les caractères spéciaux ont été supprimés et les abréviations ou les différences orthographiques ont été remplacées et regroupées sous le nom complet (par exemple, "CRP" a été remplacé par "C-reactive protein"). Ce nettoyage a permis de réduire le nombre de concepts locaux de 3 003 à 2 312. L'alignement sémantique était limité aux résultats de biologie les plus fréquents, dans le but de couvrir plus de 80% des enregistrements. Les désaccords sur l'alignement des résultats de biologie par les experts ont été résolus par consensus et par l'intervention d'un troisième annotateur. Les experts n'étaient pas d'accord sur 24 % des concepts associés, ce qui correspondait à un coefficient Kappa de 75 %.
La SNOMED-CT a été utilisée pour l'alignement sémantique des résultats biométriques, l'UCUM pour les unités de mesure et la CIM-10 pour les antécédents médicaux du patient.
Les prescriptions médicamenteuses ont été enregistrées à l'aide de la terminologie CIP. La terminologie standard des médicaments dans modèle est RxNorm. Pour l'alignement sémantique, la terminologie CIP a été alignée avec la terminologie ATC à l'aide d'une correspondance faite

par l'équipe de pharmaciens de l'ULR Metrics. Ensuite, les codes ATC ont été alignés aux codes RxNorm en utilisant les correspondances déjà présentes dans la table *CONCEPT_RELATIONSHIP*.

TABLE 2.2 – Transformation sémantique entre le vocabulaire local et le vocabulaire standard OMOP. *Lorsqu'un vocabulaire local est manquant, cela signifie que le concept a dû être créé. NA : non applicable.*

Domaines	**Terminologies locales**	**Terminologies standards**	**Niveaux de difficulté**	**Nombre de concepts locaux**	**Concepts standards alignés**	**Enregistrements associés**
Centre	-	Care site	Niveau 1	1	100%	100%
Antécédents médicaux	ICD-10	ICD-10	Niveau 1	705	96%	98%
Antécédents médicaux	Texte libre	-	Niveau 4	-	-	-
Consultations	-	Visit	Niveau 1	1	100%	100%
Médicaments	ATC	RxNorm	Niveau 2	9,946	100%	100%
Médicaments	CIP code	ATC	Niveau 3	9,946	91%	84%
Paramètres biométriques	Texte libre saisie dans des champs dédiés	SNOMED	Niveau 3	243	4.9%	96%
Paramètres de biologie	Texte libre saisie dans des champs dédiés	LOINC	Niveau 3	2,312	7%	88%
Unités	Texte libre saisie dans des champs dédiés	UCUM	Niveau 3	217	30%	77% (22% NA)
Informations démographiques	Texte libre saisie dans des champs dédiés	sex	Niveau 3	2	100%	100%
Détails des consultations (motif, interrogatoire, conclusion)	Texte libre	-	Niveau 4	-	-	-
Compte rendu clinique (extérieur)	Texte libre	-	Niveau 4	-	-	-
Prescriptions d'actes	Texte libre	-	Niveau 4	-	-	-
Informations supplémentaires	Texte libre	-	Niveau 4	-	-	-
Prescriptions de vaccins	Texte libre	-	Niveau 4	-	-	-

2.3.2.2 - Transformation structurelle

Les données extraites du logiciel étaient stockées dans 13 tables de la base de données source. À l'issue de l'ETL, 12 tables du modèle OMOP ont été chargées (Figure 2.6).
Plusieurs tables de la base de données source correspondaient à des tables différentes dans le OMOP. Les informations démographiques du patient, les noms des MG, les périodes de suivi des patients et le nom de la MSP ont été stockées, respectivement, dans les tables OMOP *PERSON*, *PROVIDER*, *OBSERVATION_PERIOD* et *CARE_SITE*.

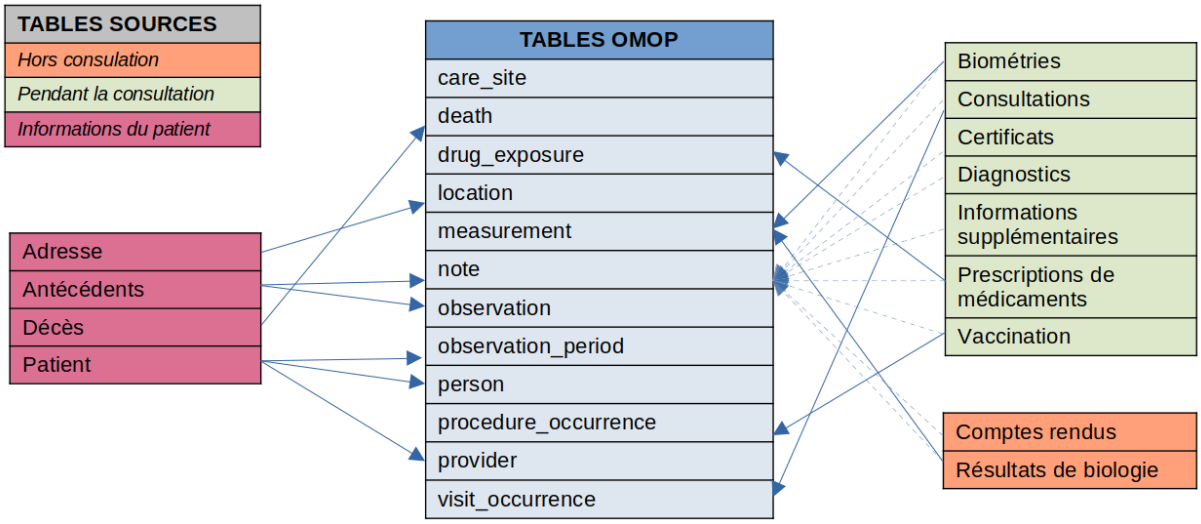


FIGURE 2.6 – Transformation structurelle de la base de données relationnelle locale au format OMOP.

La table de la base de données source *DIAGNOSTICS* pouvait contenir les informations issues de l'interrogatoire avec le patient (recensement des symptômes), de l'examen clinique (mesures et commentaires) et de la partie « résultat » (diagnostics) de la consultation. Cette table consignait, sous forme de texte, les données relatives à plusieurs types d'informations médicales différentes. La plupart des antécédents médicaux et des informations recueillies au cours de la consultation étaient aussi saisis dans le logiciel en texte libre. Ces données, en texte libre, ont été archivées dans la table *NOTE* du modèle final. Le champ *concept_type* de la table permettait de distinguer les notes sur les symptômes, le motif de la consultation, les antécédents médicaux, le résultat de la consultation et, dans certains cas, le diagnostic associé. Les rapports médicaux émanant d'un médecin spécialiste externe à la MSP étaient également conservés dans la table *NOTE*.
Les antécédents médicaux (uniquement ceux documentés par des codes CIM-10) ont été stockés dans la table du modèle OMOP *OBSERVATION*.

Chaque consultation correspondait à un enregistrement dans la table *VISIT_OCCURRENCE*, identifié par le type de concept d'une consultation ("Ambulatory Primary Care Clinic / Center" (*concept_id* = 38004247)). 12 091 enregistrements ont été créés dans la table *VISIT_OCCURRENCE* pour associer les résultats de biologie des laboratoires "Lab" (*concept_id* = 32856).
Les tables de la base de données source *RESULTATS DE BIOLOGIE* et *BIOMETRIES* contenaient, respectivement, des informations provenant des résultats de biologie et des données de mesures biométriques liées à la consultation. Dans le modèle final, ces mesures ont été implémentées dans la table *MEASUREMENT*. Le type de concept "Lab" (*concept_id* = 32856) était associé aux résultats de biologie et "Ambulatory Primary Care Clinic / Center" (*concept_id* = 38004247) aux mesures biométriques. La colonne *measurement_source_type_id* permettait de discriminer chaque type de données source (i.e., mesures de biologie des mesures biométriques prises lors d'une consultation).
Chaque prescription de médicament a été stockée dans la table *DRUG_EXPOSURE*.

Les clés primaires et étrangères ont été identifiées. De nouveaux identifiants uniques ont été créés pour la clé primaire de chaque table.

2.3.3 - Chargement et évaluation de la qualité des données

De 1997 à 2021, 592 227 consultations de 18 395 patients de la MSP de Wattrelos ainsi que 12 091 consultations en laboratoire de ville, ont été intégrées dans le OMOP. Les tables *NOTE* et *MEASUREMENT* contenaient plus d'un million d'enregistrements (i.e., 2 091 705 et 1 120 859 enregistrements respectivement) (Table 2.3). Les tables *CARE_SITE* et *DRUG_ERA* ont été implémentées une fois les données chargées dans le modèle OMOP.

TABLE 2.3 – Comparaison du nombre d'enregistrements par table dans la base de données source et dans le modèle final OMOP. *Export des données de WEDA jusque 2021*

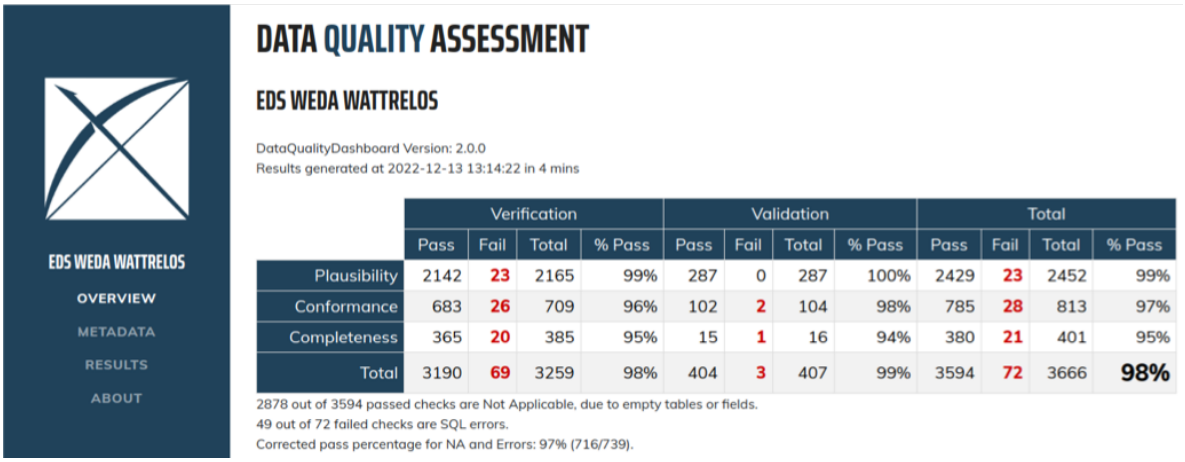
	Source (N)	**OMOP (N)**
CARE_SITE	-	1
DEATH	419	419
DRUG_ERA	-	1 084 012
DRUG_EXPOSURE	924 216	814 772
LOCATION	19 662	11 433
MEASUREMENT	1 120 859	1 120 859
NOTE	2 772 809	2 091 705
OBSERVATION	64 669	2 315
OBSERVATION_PERIOD	-	18 256
PERSON	18 395	18 395
PROCEDURE_OCCURRENCE	12 202	12 123
PROVIDER	8	8
VISIT_OCCURRENCE	592 227	604 318

Le temps de chargement total de l'ETL était de 78 minutes. L'étape la plus longue (55 minutes) était l'étape d'extraction des données des fichiers XML (Table 2.4).

TABLE 2.4 – Temps de calcul par étape de l'ETL.

	Temps (minutes)	**CPU (%)**
Extraction	55,30	95
Transformation structurelle	1,38	59
Transformation sémantique	5,67	48
Chargement	16,17	0

Sur le tableau de bord Atlas, nous avons trouvé 28 échecs de conformité qui ne respectaient pas les spécifications du OMOP CDM (i.e., mauvais format ou mauvaise saisie des données) (Figure 2.7). Il y avait 21 échecs de complétude liés à des données potentiellement manquantes et 23 échecs liés à la plausibilité pour des dates ou des valeurs de mesure non plausibles.



	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	2142	23	2165	99%	287	0	287	100%	2429	23	2452	99%
Conformance	683	26	709	96%	102	2	104	98%	785	28	813	97%
Completeness	365	20	385	95%	15	1	16	94%	380	21	401	95%
Total	3190	69	3259	98%	404	3	407	99%	3594	72	3666	**98%**

FIGURE 2.7 – Tableau de bord d'évaluation de la qualité des données Atlas OHDSI.

Des concordances entre les résultats obtenus à partir des requêtes du logiciel et les enregistrements de l'EDS ont été obtenues. En cas de différences dans les résultats des requêtes, nous avons vérifié les dossiers des patients dans le logiciel et identifié la raison de la discordance. En majorité, ces différences provenaient de la déclaration du MG en tant que MG référent du patient (i.e., médecin traitant). Cette déclaration était datée dans le logiciel WEDA. Cependant cette date n'était pas présente dans la base de données. Pour éviter ce décalage de dates, les patients ayant des dates de déclarations postérieures à la date d'extraction des données (2021) ont été retirés des résultats de requêtes sur le logiciel. De plus, les requêtes concernant les résultats de biologie ont été vérifiées manuellement par l'un des MG de la MSP. Certains résultats de biologie ont été enregistrés dans une section du logiciel nommée "rapports" et d'autres dans la section "consultations". Par conséquent, certains résultats n'étaient pas retrouvés dans la base de données car non extraits. Sur cinq requêtes, quatre avaient les mêmes résultats après ajustements (cf. Annexe A).

2.4 Discussion

2.4.1 - Principaux résultats

Considérant le modèle OMOP comme modèle de données final, les différentes étapes de l'ETL ont permis d'intégrer les données de soins premiers. Les données de soins premiers provenant d'une MSP ont été intégrées dans le OMOP CDM. Jusqu'à 2021, soit sur plus de 20 ans, 592 226 consultations et 10 221 concepts spécifiques aux soins premiers ont été intégrés, ainsi que 9 432 alignements sémantiques. L'intégration des données de soins premiers a pris en considération les données de plusieurs structures de soins : les cabinets de consultations médicales et les laboratoires de ville. L'EDS regroupait les données des médicaments prescrits et leur posologie (i.e., durée du traitement, nombre de renouvellements, doses, période de prise), les résultats de

biologie reçus des laboratoires de ville, les mesures biométriques prises lors de la consultation et les antécédents du patient.

L'intégration des données de soins premiers au modèle OMOP a nécessité l'alignement sémantique des concepts spécifiques aux soins de ville. 10 221 concepts locaux et de 12 091 visites en laboratoire ont été ajoutés.

2.4.2 - Forces

L'avantage de l'intégration des données de soins premiers dans des EDS était l'implémentation de données cliniques extra-hospitalières sur une longue période de suivi. La base de données de soins premiers a inclus des données cliniques de ville et des résultats de biologie extra-hospitaliers. En complément, la base de données a intégré les prescriptions de médicaments et les posologies associées.

La forte cohésion entre les scientifiques des données et les MG a permis de rendre la base de données plus proche de l'activité de la MSP. Chaque étape de la transformation des données a été approuvée par les MG de la MSP. De plus, collaborer avec le développeur du logiciel a permis de comprendre la structure du logiciel et le format de l'export des données ; ceci a accéléré le processus d'extraction.

2.4.3 - Limites

Une grande partie des données de soins premiers était saisie sous forme de texte libre et nécessitait l'utilisation de méthodes de traitement du langage naturel ou d'un examen manuel. Le texte libre était difficile à associer à des concepts standards. Les techniques de traitement du langage naturel (Natural Language Processing (NLP)), telles que les algorithmes de *fuzzy matching*, les outils *SpaCy* et les expressions régulières, pourraient être utilisées pour l'identification des concepts et l'extraction d'informations dans du texte [158, 159].
Bien que les concepts les plus fréquents aient été associés aux concepts standards du modèle OMOP, il est nécessaire de mettre à jour régulièrement les alignements des nouveaux concepts. De plus, le logiciel WEDA a fourni des informations sur les prescriptions de médicaments, mais pas sur leur délivrance ni sur l'observance du traitement par le patient.

2.4.4 - Comparaison avec les autres bases de données de santé

L'extraction des données implique une collaboration avec l'éditeur du logiciel pour avoir des extractions de données routinières, comme à l'hôpital, ou ponctuelles.
Un hôpital regroupe plusieurs centaines de logiciels ; l'implémentation d'un EDS se fait donc sur un logiciel à la fois et peut être redondante. En France, le détail des résultats de biologie extra-hospitaliers n'est ni documenté dans les bases de données nationales ni dans les bases de

données des hôpitaux. Le SNDS contient les données de facturation d'actes biologiques réalisés en ville et les données sur la délivrance des prescriptions mais ne contient pas les valeurs des mesures.
En soins premiers, on retrouve souvent un seul logiciel par cabinet qui contient le détail des données des MSP et des laboratoires sur plusieurs années.
La manière de récupérer les données est différente pour les données sur internet (forum, réseaux sociaux). En effet, l'extraction des données d'internet nécessite des méthodes d'aspiration telles que le *web scraping* [36]. Ces méthodes peuvent être réalisées par des API payantes développées par l'éditeur de la source [160] ou par des programmes se basant sur des packages Python comme BeautifulSoup [161] et Selenium [162].

Concernant les étapes de transformation de la donnée, les logiciels hospitaliers enregistrent les données sous forme de codes associés aux terminologies standards du modèle OMOP.
Dans d'autres bases de données, certains concepts sont à aligner aux terminologies du modèle OMOP, notamment si les concepts sont des termes locaux. Cela est valable pour les données de soins premiers et les données des bases de données nationales. Dans le SNDS, les actes sont directement codés selon la Classification commune des Actes Médicaux (CCAM), les médicaments délivrés codés en CIP, sans renseignement sur les posologies et les actes de biologie selon la Nomenclature des Actes de Biologie Médicale (NABM) [163]. Les différentes terminologies entre les bases de données sources doivent être standardisées pour se référencer à un langage commun dans le modèle.
Pour les réseaux sociaux, aucune transformation n'a été faite pour le stockage des données dans un modèle de données commun. Après un nettoyage du texte, les informations médicales extraites peuvent être réutilisées. De nombreux modèles de machine learning sont appliquées à l'analyse de ces données [164-166].

Stratégie d'optimisation pour l'implémentation et la réplication des EDS de soins premiers vers le modèle OMOP

Chapitre 3

Stratégie d'optimisation d'ETL orientés OMOP

Sommaire

3.1 Contexte

En mars 2023, la France comptait 2 251 MSP. L'objectif est d'atteindre 4 000 MSP d'ici 2027 [167]. En raison du grand nombre de MSP, le développement du processus ETL pour l'intégration des données d'une MSP nécessite des ressources de temps, d'outils et de connaissances techniques [28]. Cependant, en France, le nombre de logiciels labellisés e-santé pour les soins premiers est restreint à une vingtaine de solutions [131]. Avec ce nombre limité, on pouvait exploiter les ETL déjà développés pour intégrer les données d'autres MSP utilisant les mêmes logiciels.

La première étape d'un ETL, *extract*, consiste à extraire les données du logiciel source. Elle nécessite la maîtrise du modèle de données du logiciel. La seconde étape, *transform*, consiste à transformer la structure initiale pour correspondre à la structure finale de l'EDS. Les autres

étapes de transformation sont liées à la transformation sémantique et à la gestion de la qualité des données. Une fois ces transformations établies, les données sont chargées dans le modèle final, *load*.
Dans le cadre du projet PriCaDa, un processus ETL a été mis en œuvre dans la MSP de Wattrelos (cf. Chapitre 2) [168]. Nous nous sommes appuyés sur ce premier travail pour intégrer les données de trois MSP supplémentaires. L'intégration de ces sources de données a été l'occasion de proposer une stratégie pour implémenter et répliquer les ETL plus facilement. Cette stratégie a consisté à réutiliser un maximum d'étapes du processus ETL déjà développées, pour intégrer les données d'une MSP. La stratégie s'est appuyée sur le partage d'un environnement technique favorisant le déploiement de l'ETL.

3.2 Matériels et méthodes

3.2.1 - Développement initial des ETL

Lors du développement d'un ETL (cf. Chapitre 2), plusieurs étapes étaient fortement liées aux caractéristiques du logiciel, tandis que d'autres relevaient du modèle de données de l'EDS. Deux logiciels de soins premiers sont utilisés dans les MSP du projet PriCaDa : WEDA à Wattrelos, Guesnain et Lille-Moulins, et Crossway à Tourcoing [169, 170]. Les volumes de données couvrant la période de 1997 à 2023, relatifs aux consultations, résultats de biologie, mesures biométriques et prescriptions, ont été calculés pour chacune des quatre MSP du projet PriCaDa (cf. Figure 3.1). Les données brutes ont été recueillies auprès de l'éditeur du logiciel WEDA (MSP de Wattrelos, Guesnain et Lille-Moulins) et du MG responsable de la MSP de Tourcoing.

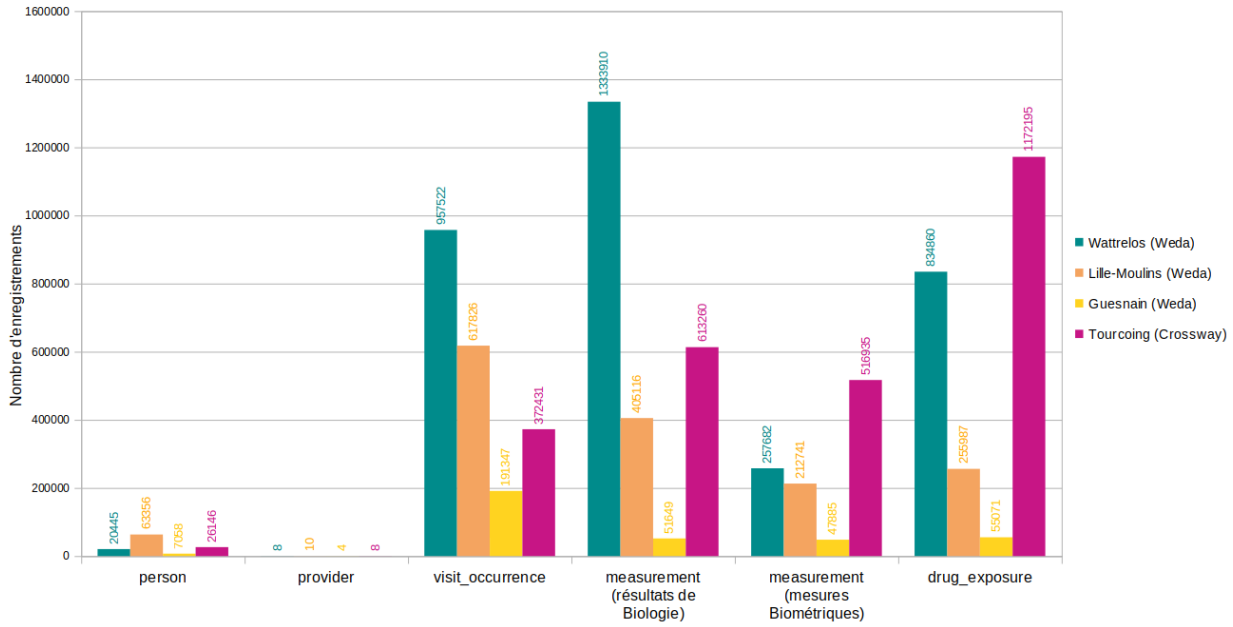


FIGURE 3.1 – Nombre d'enregistrements par table pour chaque MSP de 1997 à 2023.

Un groupe de travail a été constitué et était composé de trois data scientists spécialisés dans le développement d'ETL. Le groupe de travail s'est réuni pour mettre en place la stratégie de généralisation des ETL, et pour implémenter l'ETL dans trois MSP. Une personne était affectée au développement de l'ETL de Crossway et deux autres au déploiement de l'ETL de WEDA sur une autre MSP.

3.2.2 - Généralisation des ETL

À chaque étape du processus ETL, le groupe de travail a identifié les opérations spécifiques aux logiciels, et celles qui pouvaient être généralisées. Lors de l'implémentation, ces opérations ont été séparées par étapes de l'ETL (cf. paragraphe 3.2.1). À la suite de l'étape *extract*, les données ont été stockées dans une base de données relationnelle source conservant le format des données du logiciel. Une nomenclature de dossiers et de fichiers a été définie pour l'écriture et le stockage des scripts. Les opérations dépendantes des logiciels ont été stockées et séparées dans des dossiers propres à chaque logiciel intégré.

3.2.3 - Déploiement de l'ETL

La stratégie de déploiement des ETL avait pris en compte deux contextes :

1. les données provenaient d'un logiciel pour lequel un ETL avait déjà été implémenté.
2. les données provenaient d'un logiciel pour lequel aucun ETL n'était encore développé ;

Dans le premier contexte, un ETL avait été développé pour l'intégration des données provenant du même logiciel. Dans le second contexte, la partie dépendante du logiciel devait être développée et intégrée.

Un fichier de configuration a été créé pour lancer le processus ETL adapté au contexte. C'est-à-dire soit utiliser l'ETL existant, soit développer l'ETL. Ce fichier regroupait toutes les variables spécifiques au logiciel et à la MSP. Il contenait les chemins des répertoires et fichiers nécessaires au déploiement de l'ETL.

Après la transformation structurelle, un fichier plat a été intégré comme modèle pour stocker les alignements de concepts. Les vocabulaires standards utilisés pour l'alignement des concepts locaux étaient LOINC, SNOMED-CT, UCUM, RxNorm et CIM-10 dans modèle OMOP (cf. Chapitre 2, table 2.1). Ces alignements consistaient à créer des jointures complexes entre les tables de vocabulaires pour faire correspondre le vocabulaire local au vocabulaire standard du modèle OMOP (cf. processus détaillé dans la figure 2.5 du chapitre 2).

Des outils de la communauté OHDSI, tels que Rabbit-in-a-hat et WhiteRabbit, ont été utilisés pour la transformation structurelle, c'est-à-dire pour aligner les noms des tables et des colonnes de la base de données source avec le OMOP CDM [97, 98]. L'outil Athena [96] a été utilisé pour la transformation sémantique. Les scripts et les outils spécifiques à l'ETL ont été partagés sur un dépôt GitLab privé (une solution open-source qui simplifie la gestion des dépôts de code source et le suivi de versions) [171].
Les bases de données de chaque MSP sont stockées sur PostgreSQL, au format OMOP version 5.4 [95].

3.2.4 - Validation et test

Nous avons validé notre stratégie avec deux sources de données supplémentaires. Les données de la MSP de Lille-Moulins ont été utilisées pour la validation et l'ajustement des opérations de l'ETL. Lors de cette validation, les paramètres et opérations ont été adaptés et ajustés à l'ETL. Les données de la MSP de Guesnain et la nouvelle extraction, plus récente, des données de la MSP de Wattrelos ont permis de tester le processus finalisé. Le test évalue la compatibilité de l'ETL sur une nouvelle source de données.

Après avoir converti les données dans un format standardisé, une évaluation de la qualité a été effectuée selon les métriques établies par Kahn et al. [172] (cf. définitions des métriques 2.2.3 du chapitre 2). Les outils développés par la communauté OHDSI, tels qu'Achilles, et le tableau de bord sur la qualité des données ont permis de calculer ces métriques sur le modèle final [156, 173]. La conformité, l'exhaustivité et la plausibilité ont permis d'évaluer les données.

3.3 Résultats

3.3.1 - Développement initial des ETL

La consolidation des trois ETL a duré six mois et comprenait une réunion de présentation de l'ETL de WEDA, axée sur les enjeux et les objectifs, ainsi que cinq réunions de groupes de travail. Lors de la première réunion, nous avons défini l'organisation de l'environnement de travail. La structure des fichiers et dossiers, ainsi que la nomenclature ont été maintenues cohérentes tout au long du projet (cf. Annexe B). Une liste des opérations à suivre pour le développement d'un ETL a été partagée (cf. Annexe C). Lors des réunions de groupes de travail, le développement de l'ETL des données du logiciel Crossway et le déploiement de l'ETL du logiciel WEDA ont été effectués. Au cours des développements, les étapes et les opérations ont été mises en commun et généralisées.

3.3.2 - Généralisation des ETL

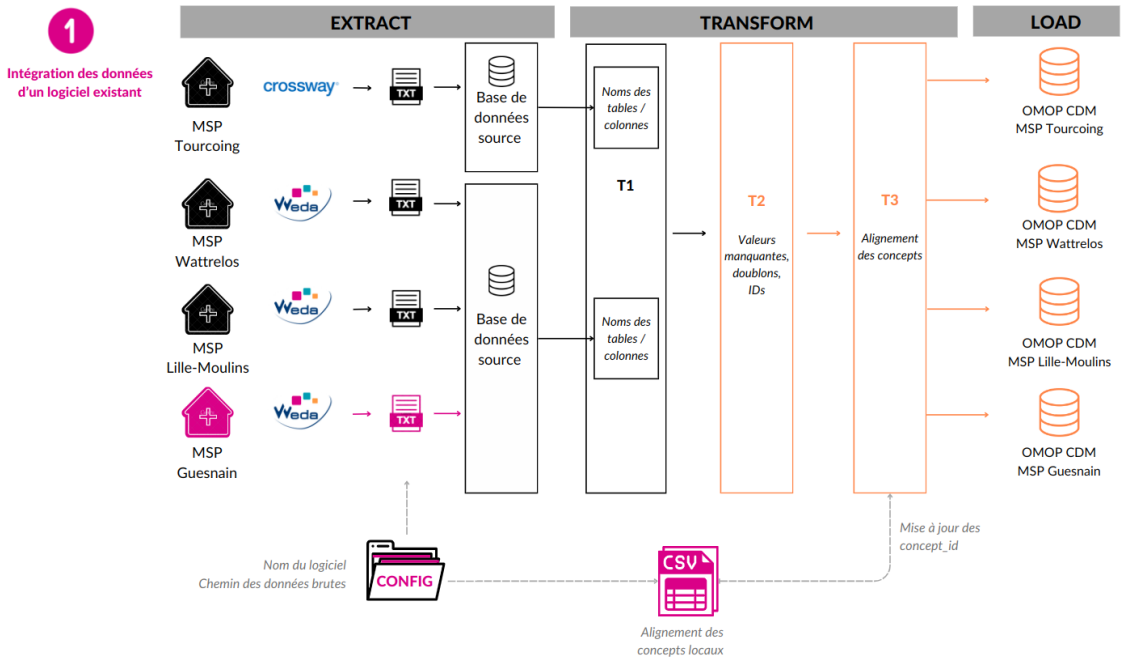
Durant les réunions de ces groupes de travail, le processus ETL a été subdivisé en plusieurs étapes distinctes. La première étape, *extract*, consistait à communiquer avec l'éditeur du logiciel pour définir les modalités d'extraction des données, et se terminait une fois les données stockées dans une base de données source, indépendante du logiciel. La première transformation, appelée *T1*, comprenait la transformation structurelle des données brutes au format OMOP. Les colonnes et les tables de la base de données source ont été renommées selon la convention OMOP. La deuxième transformation, *T2*, prenait en charge les opérations de gestion de la qualité des données telles que le dédoublonnage des enregistrements, le traitement des données manquantes, et l'attribution d'identifiants uniques. La troisième et dernière transformation, *T3*, concernait l'alignement sémantique. La dernière étape, *load*, consistait à charger les données dans le modèle final en appliquant des contraintes (clés primaires et clés étrangères). À la suite de *T1*, les étapes *T2*, *T3* et *load* devenaient indépendantes des logiciels et pouvaient être généralisées. Les opérations dépendantes et indépendantes des logiciels sont présentées dans le Tableau 3.1.

Étapes	WEDA	Crossway	Généralisation
extract	Transformation des XML par patient à une base de données source	Transformation du fichier plat en une base de données source	-
T1	Transformation structurelle de la base de données source au format OMOP	Transformation structurelle de la base de données source au format OMOP	-
T2	Dédoublonnage Gestion des valeurs manquantes Identifiants uniques	Dédoublonnage Gestion des valeurs manquantes Identifiants uniques	Opérations basées sur la structure du modèle OMOP Valeurs manquantes : NA Auto-incrémentation des identifiants
T3	Alignement sémantique des concepts liés au logiciel	Alignement sémantique des concepts liés au logiciel	Prise en compte d'un fichier plat et du nom du logiciel (vocabulaire local) en début de processus si de nouveaux concepts locaux doivent être alignés
load	Identification des clés primaires et étrangères	Identification des clés primaires et étrangères	Chargement des nouveaux concepts, des données et des clés primaires et étrangères dans le modèle final

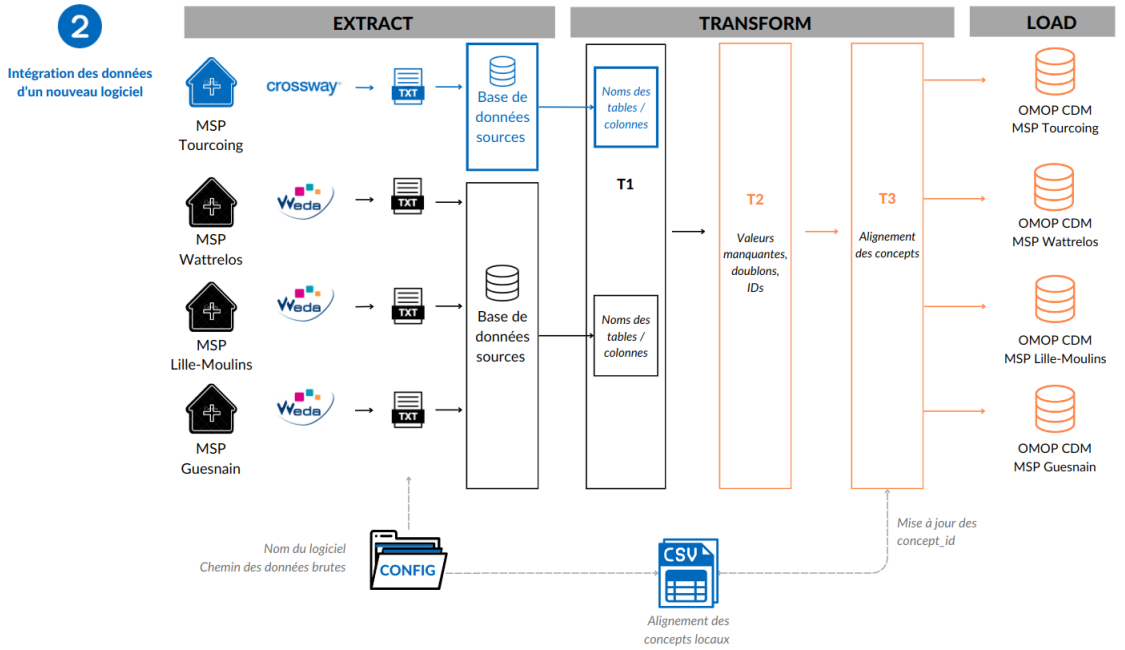
TABLE 3.1 – Résumé des opérations par étape de l'ETL pour chaque logiciel et identification des opérations indépendantes du logiciel. *NA : Non Applicable.*

3.3.3 - Déploiement de l'ETL

L'intégration des données d'un logiciel dont l'ETL était déjà développé n'a nécessité aucune modification ni ajout d'opérations (Figure 3.2a). Lors de l'intégration des données d'un nouveau logiciel, les opérations à ajouter ou à modifier dans l'ETL se trouvaient dans les étapes *extract* et *T1* (Figure 3.2b). Un dossier, nommé avec le nom du logiciel, a dû être ajouté en amont de l'ETL pour documenter les opérations d'extraction des données, d'implémentation dans une base de données source, ainsi que de transformation structurelle.



(a) Cas 1 : Étapes à modifier pour l'intégration des données d'un logiciel dont l'ETL a déjà été développé. *En rose : fichier de configuration à modifier, ajouter des alignements de concepts si nécessaire, en orange : étapes basées sur la structure du modèle OMOP.*



(b) Cas 2 : Étapes à ajouter et à modifier pour l'intégration des données d'un nouveau logiciel. *En bleu : fichier de configuration à modifier, ajouter des alignements de concepts si nécessaire, étape d'extraction et d'alimentation de la base de données source à ajouter, étape T1 à ajouter également, en orange : étapes basées sur la structure du modèle OMOP.*

FIGURE 3.2 – Intégration des données et déploiement selon le contexte de la stratégie d'optimisation d'ETL.

Le fichier de configuration définissait les variables propres à l'environnement de l'utilisateur et de la MSP. Ce fichier permettait de connaître le contexte d'intégration des nouvelles données. Ces variables incluaient le chemin des fichiers de données brutes, le nom du logiciel (pour lancer les dossiers des étapes d'extraction et la *T1* du même nom) et, si nécessaire, le chemin du fichier plat contenant les nouveaux alignements.
Une fois les concepts locaux liés aux concepts des soins premiers chargés pour un logiciel, cette étape n'était plus nécessaire et permettait un gain de temps dans l'exécution de l'ETL (Figure 3.2). La complétion de ce fichier et du fichier de configuration a automatisé la phase de jointures complexes entre les tables de vocabulaires du modèle OMOP.

3.3.4 - Validation et test

Lors de la validation de l'ETL, des opérations ont été ajustées, notamment pour l'étape *T3*. En effet, les valeurs manquantes dans WEDA étaient par défaut formatées à "01/01/0001", ce qui différait des données du logiciel Crossway. De plus, des divergences existaient entre les terminologies et concepts utilisés par Crossway et WEDA. Pour remédier à ces différences, il a été nécessaire de générer un *vocabulary_id* propre à chaque logiciel dans le fichier plat des aligments. Ce *vocabulary_id* a ensuite été inséré dans les tables *CONCEPT* et *VOCABULARY*.
Lors de la phase de test, en dehors du paramétrage des variables dans le fichier de configuration, aucune autre modification n'a été nécessaire.

Le tableau de bord sur la qualité des données au format OMOP a affiché les métriques de conformité, d'exhaustivité et de plausibilité. Les taux de réussite pour ces différentes métriques étaient supérieurs à 93 % (Tableau 3.2). Le taux de plausibilité était de 94 % pour les données de la MSP de Lille-Moulins et de 100 % pour la MSP de Tourcoing. Les échecs de plausibilité indiquaient des incohérences dans les valeurs, qui ne respectaient pas les règles métiers en conditions réelles. Le taux d'exhaustivité était de 93 % pour les données de la MSP de Guesnain et 95 % pour la MSP de Wattrelos. Les échecs liés à l'exhaustivité indiquaient des données manquantes. Enfin les échecs liés à la conformité pouvaient être causés par des erreurs de format ou des saisies de données qui ne respectaient pas les règles du modèle final.

	Wattrelos	Guesnain	Lille-Moulins	Tourcoing
Plausibilité (%)	99	95	94	100
Conformité (%)	97	97	97	97
Exhaustivité (%)	95	93	94	93
Total (%)	98	96	96	98

TABLE 3.2 – Évaluation de la qualité des données par métriques de Kahn et al.

3.4 Discussion

3.4.1 - Principaux résultats

Dans ce chapitre, nous avons proposé une stratégie d'optimisation des processus ETL pour éviter de recommencer la mise en œuvre d'un ETL à chaque intégration de données d'une MSP. Cette stratégie prenait en compte deux contextes : (1) l'intégration des données d'un logiciel dont l'ETL avait déjà été développé, (2) l'intégration des données d'un nouveau logiciel.
La mise en œuvre de cette stratégie a abouti à une généralisation des étapes et des opérations communes, une standardisation des données de soins premiers, et un environnement de travail optimisé pour l'exécution des scripts. Certaines étapes de l'ETL, spécifiques au logiciel, ont été identifiées, telles que l'extraction des données, la création d'une base de données relationnelle source adaptée au format du logiciel, ainsi que l'alignement structurel de cette base avec le modèle OMOP. À partir de cette étape, les opérations ont pu être généralisées, quel que soit le logiciel, notamment pour la gestion des données, la transformation sémantique, la gestion des clés primaires et étrangères, des valeurs manquantes et des doublons.
Pour l'intégration d'un nouveau logiciel, les étapes de l'ETL spécifiques à la structure du logiciel ont dû être développées. En revanche, pour l'intégration des données d'un logiciel dont l'ETL avait déjà été implémenté, les données pouvaient être intégrées sans modification ni ajout d'opérations. L'environnement de travail incluait également un fichier de configuration permettant de renseigner les variables spécifiques au logiciel et à l'utilisateur et un fichier d'aide au stockage des alignements de concepts.

Cette approche reposait sur le respect d'une nomenclature de scripts et de répertoires de travail, et l'adoption d'un modèle de données commun (i.e., OMOP). La méthode a été appliquée à deux logiciels de soins premiers, WEDA et Crossway, utilisés par quatre MSP. Le projet s'est déroulé sur une période de six mois ; la majeure partie du temps ayant été consacrée au développement des ETL pour WEDA et Crossway, ainsi qu'aux sessions de travail en groupe. Les données de deux MSP ont été utilisées pour valider et tester le processus ETL basé sur cette stratégie. L'évaluation qualitative des données a été réalisée à l'aide des métriques de Kahn [157].

3.4.2 - Forces

Le fichier de configuration modifiable par l'utilisateur a permis de stocker les variables nécessaires à l'intégration de nouvelles données sans modifier le code source.
La séparation des étapes à développer pour l'intégration d'un nouveau logiciel a réduit le nombre d'opérations à ajouter aux scripts existants, facilitant ainsi l'intégration des données de logiciels déjà pris en charge et permettant un gain de temps. La période de temps de travail a

été considérable pour la généralisation d'un ETL à un logiciel supplémentaire. De nombreuses MSP dépendant de ce logiciel pourront désormais être intégrées.
La stratégie n'est pas spécifique aux soins premiers et pourrait être utilisée dans d'autres contextes.

3.4.3 - Limites

La stabilité dans le temps des ETL dépend de l'éditeur et des mises à jour du logiciel. Si les données exportées par le logiciel subissent des changements de structure ou de format, l'étape d'extraction devra être ajustée. Néanmoins, les étapes indépendantes du logiciel seront conservées.
La stratégie d'optimisation a été réalisée sur trois MSP utilisant le logiciel WEDA mais n'a été appliquée et développée que sur une MSP utilisant le logiciel Crossway. L'inclusion des données d'une autre MSP utilisant Crossway ou un troisième logiciel aurait permis de valider la stratégie à plus grande échelle.

3.4.4 - Comparaisons aux travaux existants

Les communautés de chercheurs internationaux, comme OHDSI, proposent des *frameworks* pour rendre plus facile l'utilisation de leurs outils et de leur modèle. Certaines stratégies permettent l'intégration des données provenant de différentes sources, pour pallier aux difficultés telles que la diversité des structures et des terminologies. Les méthodologies pour la gestion de la mise en œuvre d'un ETL proposent des formats standards pour l'échange et l'intégration des données entre différents systèmes [84, 85, 174, 175]. Des méthodes ont été développées pour gérer ou éviter les défis rencontrés lors du développement d'un ETL [172] et pour évaluer la qualité des données [157, 176, 177]. La diversité des formats de données et des terminologies en soins premiers rend l'utilisation des méthodes ou outils existants difficile [84, 177, 178], en raison des habitudes de saisie des MG et des formats variés d'extraction de données selon les éditeurs de logiciels.

Une étude a abordé la réplication des processus ETL pour intégrer des données de plusieurs sources dans un modèle de données commun [28]. Cependant, dans cette étude, chaque ETL est développé indépendamment pour chaque source, sans mise en commun ou généralisation des étapes ; ce qui complique l'intégration d'autres sources. L'étude est descriptive et ne propose pas de méthodologie ou de stratégie d'optimisation. Notre travail pourrait favoriser la réplication de leur ETL pour intégrer les données de nouvelles cliniques. De plus, notre approche pourrait être partagée avec les hôpitaux, leur permettant d'intégrer les données de différents services en intégrant les concepts spécifiques au contexte clinique et les opérations de transformation structurelle propres à chaque logiciel utilisé.

Développement d’outils d’aide à la prise en charge des patients

Chapitre 4

Outils d'aide à la prise en charge des patients

Sommaire

4.1 Évaluation de la persévérance aux médicaments

4.1.1 - Contexte

L'adhérence à un traitement est la combinaison de la persévérance (i.e., poursuite sur la durée) et de la conformité (i.e., dose correcte prise au bon moment de la journée) à la prescription [179, 180].
Le manque d'adhérence est un enjeu de santé publique, il impacte les symptômes d'une pathologie, en particulier lorsque la persévérance d'un traitement n'est pas respectée. Un arrêt de plus de 60 jours dans la prise d'un traitement augmente l'apparition des symptômes, favorise la progression de la maladie et augmente les coûts de dépense en santé [179, 181]. Le manque d'adhérence est surtout observé chez les patients sous traitements antidouleur ou sous traitements prescrits par le MG, par rapport aux traitements prescrits par d'autres spécialistes [182]. Les personnes diabétiques respectent mieux la persévérance de leur traitement Anti-Diabétiques Oraux (ADO) par rapport aux traitements injectables (i.e., insulines) [183]. Pour les personnes

diabétiques, le manque de persévérance impacte et augmente les valeurs de l'hémoglobine glyquée et de la glycémie [183].

Il existe plusieurs méthodes pour évaluer la non-adhérence aux traitements. Les méthodes non directes consistent à s'appuyer sur les données exploitables dans les bases de données. Les données de délivrance en pharmacie sont les plus utilisées et prennent en compte le nombre de médicaments fournis et le nombre de jours d'observance prescrits [180, 184]. Ces mêmes méthodes ne prennent pas en compte le surplus ou le stockage de médicaments des patients [180, 184]. Enfin, les méthodes directes sont les plus représentatives de la réalité en terme d'administration du traitement, mais sont plus invasives et plus compliquées à mettre en œuvre. Ces méthodes peuvent être l'attribution de cachets digitaux, l'évaluation de taux biologiques (i.e., tests sanguins ou urinaires) ou l'observation de l'administration du médicament à l'hôpital [180].

Dans les bases de données de soins premiers, les informations de prescriptions et de posologies sont associées. Les posologies contiennent les informations de durée et de renouvellement des traitements, ce qui permet de suivre la chronologie du traitement prescrit. La séquence est la durée durant laquelle le patient est sous un traitement. Cependant, ces séquences ne sont pas calculées directement par le logiciel de soins premiers.
Dans cette partie, nous décrirons le développement et l'implémentation d'un algorithme pour la détection des séquences de traitement et pour évaluer la persévérance à partir des données de prescriptions médicamenteuses de la MSP de Wattrelos. L'algorithme a été appliqué à la population de patients diabétiques pour évaluer leur persévérance aux traitements antidiabétiques.

4.1.2 - Méthodes

4.1.2.1 - Données extraites

Le logiciel WEDA contenait une section dédiée à la prescription médicamenteuse. Le MG sélectionnait le médicament à prescrire à l'aide d'une liste déroulante liée à la base de données Vidal, et choisissait la posologie souhaitée (Figure 4.1). Le MG obtenait une posologie déjà rédigée en fonction de la sélection du traitement, de la durée de traitement, du moment de la prise et du nombre de prises. Les posologies ont ainsi été renseignées en texte, de manière structurée, et ont été stockées dans la table *NOTE* du modèle OMOP (cf. Chapitre 2). Nous avons stockées les prescriptions de médicaments dans la table *DRUG_EXPOSURE* dans un format structuré facilitant l'analyse.

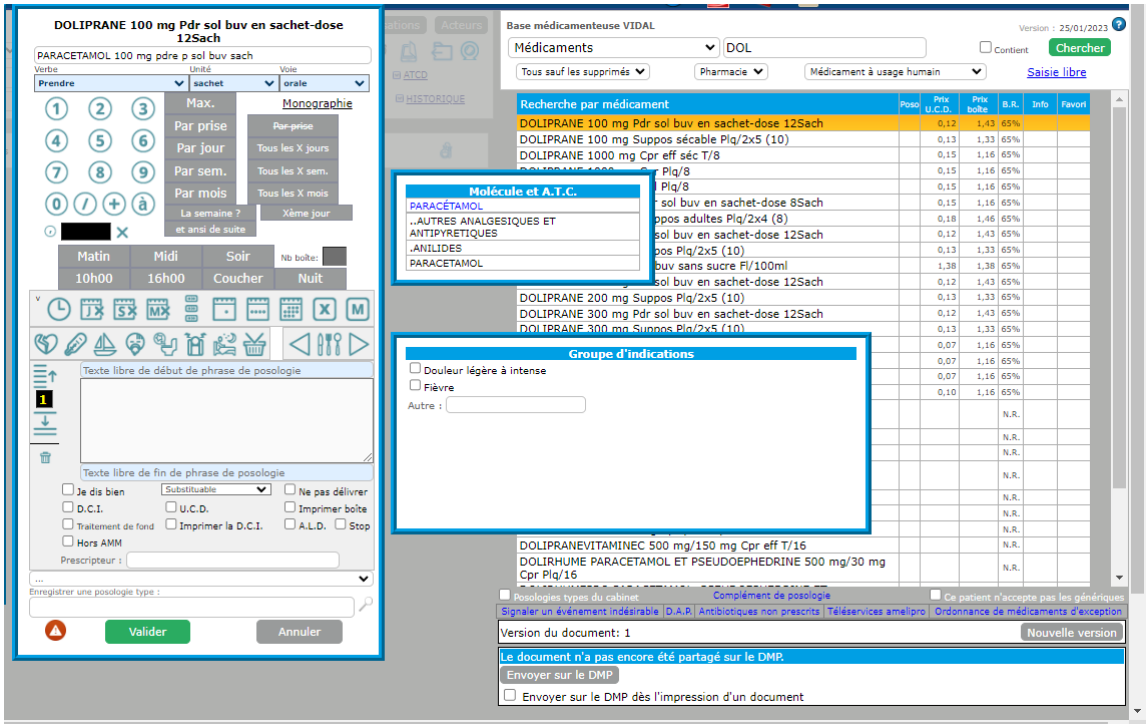


FIGURE 4.1 – Interface de prescription dans le logiciel WEDA. *La section de gauche : sélection de la posologie associée au médicament (durée de traitement et dose). Section de droite : sélection d'un médicament dans la base de données Vidal.* [152].

Pour développer l'algorithme de détection de séquences de traitements prescrits par le MG, nous avons utilisé les données de la MSP de Wattrelos. Cet algorithme s'est basé sur les médicaments prescrits et leurs posologies. Les prescriptions antidiabétiques, des patients adultes, ont été extraites par les codes ATC débutant par "A10". Les notes de posologies ont été associées à ces prescriptions.

4.1.2.2 - Calculs des durées

Deux informations étaient nécessaires pour évaluer la persévérance à partir des informations de prescription (Figure 4.2) :

— La durée de traitement prescrite, pour calculer une date de fin de la prescription ;

— La durée de l'arrêt entre deux prescriptions, à partir de la date de fin de la prescription initiale et la date de la prescription suivante.

La date initiale et la date de la prescription suivante sont associées à chaque prescription.

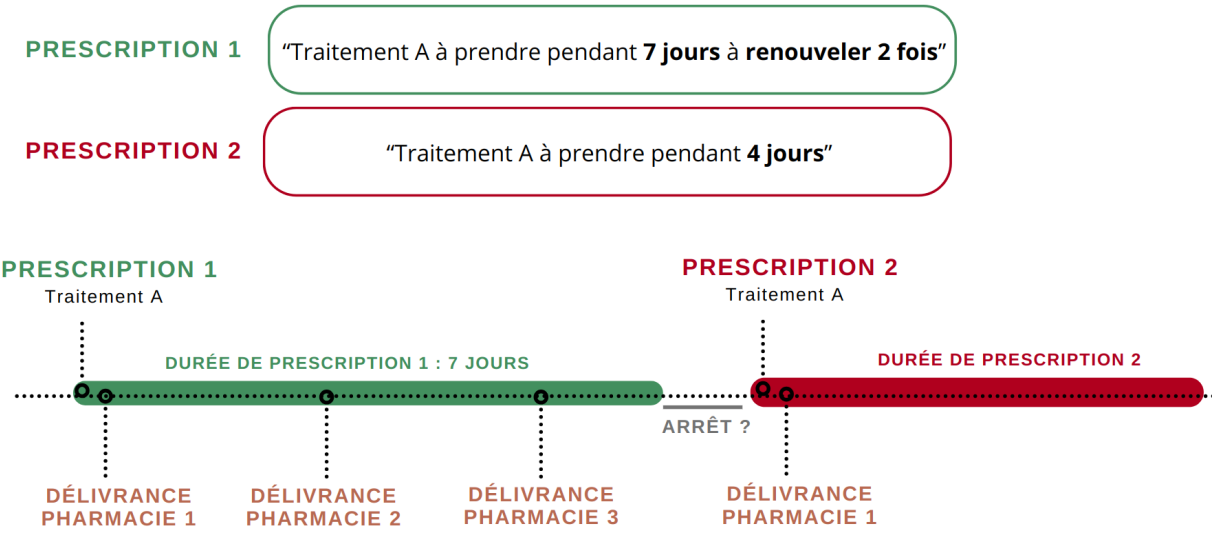


FIGURE 4.2 – Exemple de représentation des durées de prescription d'un traitement renouvelable. *Prescription 1 : à renouveler 2 fois signifie 3 délivrances (1ère délivrance liée à la prescription initiale + 2 renouvellements)*.

Tout d'abord, la durée du traitement prescrite et le nombre de renouvellements de l'ordonnance ont été extraits à partir de méthodes de Regular Expression (RegEx), c'est-à-dire des méthodes d'extraction d'informations contenues dans du texte.
La RegEx permettant d'extraire le nombre et l'unité de la durée inscrite dans la posologie était la suivante :

(?:pendant|pdt|qsp|qs|ttt pour|traitement pour) \s(\d+) \s*(mois|jours|j|semaine)*

Cette formule permettait de détecter trois groupes :

1. le terme annonceur de durée (pendant, traitement pour, etc.), ces expressions ont été identifiées après une analyse exploratoire ;
2. le nombre associé à la durée du traitement ;
3. l'unité de durée (i.e., jour, mois ou semaine).

Les nombres de mois et de semaines ont ensuite été convertis en nombre de jours.

La RegEx permettant d'extraire le nombre de renouvellements de l'ordonnance était la suivante :

(?:à renouveler|renouveler|ar) (\d+) (?:fois|x)

Cette formule permettait de détecter trois groupes :

1. le terme annonceur du renouvellement, identifié après une analyse exploratoire ;
2. le nombre associé au renouvellement ;
3. l'indicateur "fois" qui suit le nombre de renouvellements.

Ces RegEx ont permis de récupérer la durée de la prescription pour calculer la date à partir de laquelle la prescription n'était plus couverte (i.e., la date de fin de la prescription).
Avec ces deux informations, nous avons calculé la durée de la prescription (Tableau 4.1). Le nombre total de jours de traitement a été obtenu en multipliant la durée du traitement extraite par RegEx (en jours) par le nombre de renouvellements + 1 car la prescription initiale était considérée comme la première délivrance du traitement. Enfin, la durée de l'arrêt entre deux prescriptions a été calculée comme la différence entre la date de la prescription suivante et la date de la prescription initiale, à laquelle était ajoutée la durée de traitement prescrite.

Médicament	Patient	Date de prescription initiale	Date prescription suivante	Durée de la prescription	Temps d'arrêt
A	1	D1	D2	DT (jours) x (RO + 1)	D2 - (D1 + DF1)
A	1	D2	D3	DT (jours) x (RO + 1)	D3 - (D2 + DF2)

TABLE 4.1 – Calculs des durées de traitements et des temps d'arrêt entre deux prescriptions. *DF : date de fin de la prescription, calculée en ajoutant la durée de prescription à la date de prescription initiale ; RO : nombre de renouvellements inscrits dans la posologie ; DT : durée de traitement inscrit dans la posologie*.

Dans l'exemple de la Figure 4.2, en supposant que la première prescription a été faite le 07/01/2020 et la seconde le 11/02/2020, les calculs étaient les suivants :

— Durée de la prescription = 7 x (2 + 1) = 21 jours ;

— Durée de l'arrêt entre les deux traitements = 11/02/2020 - (07/01/2020 + 21) = 14 jours.

4.1.2.3 - Interprétation des résultats de l'algorithme

Lorsqu'une prescription est finie, le patient prend de nouveau rendez-vous avec son MG pour renouveler son traitement. Le patient peut avoir quelques jours de retard dans la prise de son rendez-vous selon les disponibilités du médecin ou son stock de médicaments. Pour

laisser une marge à l'interprétation de la persévérance, l'algorithme tenait compte d'un delta. Par défaut ce delta était paramétré à 30 jours.

— Arrêt de traitement < 0 jour : le patient consultait pour le renouvellement de son traitement avant la fin du traitement précédent. Cela montre une anticipation et donc une bonne persévérance.

— 0 jour <= arrêt de traitement <= delta paramétré : le patient consultait légèrement après la fin de sa dernière prescription, le retard était acceptable.

— Arrêt de traitement > delta paramétré : il y avait un arrêt considérable dans la prescription. Cela représentait un manque de persévérance.

4.1.2.4 - Statistiques

Les statistiques des résultats de l'algorithme ont été réalisées avec R Studio en utilisant le langage R (version 4.4.0).
Le test de normalité (Shapiro test) a été appliqué aux variables quantitatives. Les variables quantitatives suivant une distribution normale ont été décrites par la moyenne et l'écart-type, tandis que celles ne suivant pas une loi normale ont été exprimées par leur médiane, leur premier quartile (Q1) et leur troisième quartile (Q3).

4.1.3 - Résultats

4.1.3.1 - Données extraites

De janvier 2013 à janvier 2023, 34 694 prescriptions d'antidiabétiques ont été prescrites à 914 patients. Cela représentait une moyenne de 3 438 prescriptions (minimum : 2 140, maximum : 4 930) et une moyenne de 409,1 patients (minimum : 312, maximum : 600) par an. Le médecin prescrivait une médiane de 6 [4 ; 11] prescriptions d'antidiabétiques à ces patients par an.

4.1.3.2 - Calculs des durées

Pour cette même période, 26 420 durées entre prescriptions ont été calculées (76 %). Pour 28 259 prescriptions d'antidiabétiques (81 %), une date de fin de traitement a été identifiée à l'aide de l'algorithme (Tableau 4.2).
Lorsque les posologies étaient incomplètes ou indisponibles (9 539 prescriptions), la durée de la prescription n'a pas pu être récupérée. De plus, les prescriptions uniques ou les dernières prescriptions faites n'ont pas fait l'objet de prescription suivante (2 282 prescriptions), donc la durée d'arrêt de traitement n'a pas pu être calculée.

	Nombre de prescriptions	Proportion (%)
Posologie correcte	28 259	81
Pas de posologie	3 104	9
Manque d'informations dans la posologie	6 435	18,5
Manque la date de la prescription suivante	2 282	6,6
Posologie correcte + prescription suivante	26 420	76,1

TABLE 4.2 – Nombre et proportion des prescriptions par catégorie de problème lors du calcul de l'algorithme.

4.1.3.3 - Interprétations

La durée médiane [Q1 ; Q3] entre toutes les prescriptions était de -2,4 jours [-13,4 ; 6,8] (Figure 4.3). Il y avait un manque de persévérance pour 2 925 prescriptions d'antidiabétiques sur les 26 420 (11,1 %), c'est-à-dire avec une interruption de plus de 30 jours entre deux prescriptions successives. La durée médiane [Q1 ; Q3] entre ces prescriptions était de 61,6 jours [40,8 ; 102,8]. Une anticipation entre deux prescriptions a été retrouvée pour 16 378 prescriptions (62 %) et correspondait à une durée médiane de -8,2 jours [-29,8 ; -2,4] entre deux prescriptions.

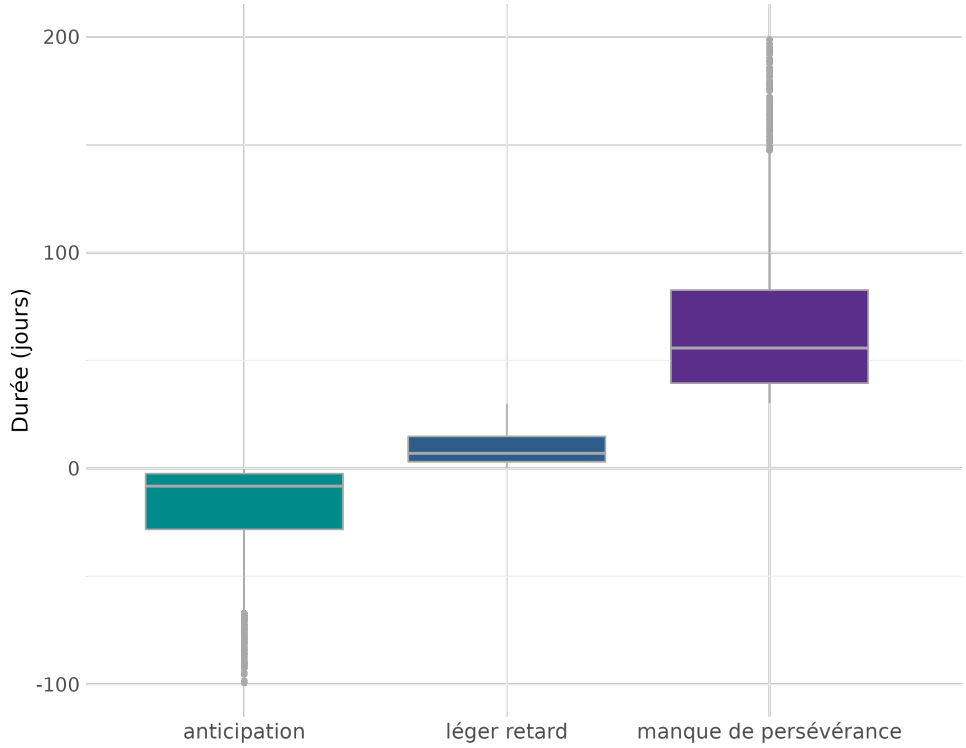


FIGURE 4.3 – Représentation des durées entre les prescriptions d'un traitement par catégorie d'interprétation. *Arrêt < 0 jours : anticipation ; 0 jours <= arrêt <= 30 jours : léger retard ; Arrêt > 30 jours : manque de persévérance.*

Le médicament antidiabétique le plus prescrit était la Metformine (code ATC : A10BA02), avec 10 947 prescriptions. Parmi ces prescriptions, 9,1 % étaient suivies d'une interruption de plus de 30 jours, indiquant un manque de persévérance (Tableau 4.3). Parmi les ADO, la durée médiane de l'arrêt entre les prescriptions de Metformine était de -2,4 [-14,2 ;7,75] jours. Le taux le plus élevé de manque de persévérance était observé pour le Glibenclamide (code ATC : A10BB01) et représentait 22 % des prescriptions (24 sur 109). Concernant les insulines, l'Umuline rapide (code ATC : A10AB01) présentait le taux de manque de persévérance le plus élevé, soit 13,5 % des prescriptions.

Codes ATC	Durée entre les prescriptions (médiane[Q1 ;Q3]) (en jours)	Proportion du manque de persévérance (%) (n/N)
A10BB01	9,6 [-3,2 ;28,2]	22.0 (24/109)
A10BF01	2 [-4,5 ;14,1]	14.3 (25/175)
A10AB01	0,3 [-9,25 ;-7,20]	13.5 (5/37)
A10BB12	0 [-7,2 ;8,15]	12.6 (52/414)
A10BD15	-11,2 [-18,8 ;1,8]	12.1 (4/33)
A10AE05	-2,4 [-8,4 ;20,6]	12.0 (46/382)
A10BD10	-0,2 [-4,8 ;12,2]	11.6 (21/181)
A10BD08	-0,2 [-8,8 ;12,8]	11.1 (83/745)
A10BD07	-2,4 [-12,8 ;7,2]	10.7 (206/1924)
A10BH02	-2,4 [-20,3 ;10,8]	9.6 (32/332)
A10AB05	-2,4 [-19,2 ;14,8]	9.4 (146/1543)
A10BH03	-1,3 [-9,4 ;10,85]	9.4 (23/244)
A10BA02	-2,4 [-14,2 ;7,75]	9.1 (1002/10947)
A10AC01	-2,4 [-28,2 ;10,8]	8.9 (20/225)
A10BB09	-2,4 [-9,4 ;6,8]	8.3 (373/4506)

TABLE 4.3 – Taux des prescriptions suivies d'un manque de persévérance pour les 15 codes ATC d'antidiabétiques ayant une proportion de non-persévérance élevée.

4.2 Visualisation du suivi de l'activité et des patients

4.2.1 - Contexte

Les techniques d'analyse visuelle, lorsqu'elles sont appliquées aux données, permettent de créer des tableaux de bord ainsi que d'autres outils de visualisation [185]. Le tableau de bord est un outil visuel et interactif relié à une base de données. Il permet de suivre des indicateurs de performance, de visualiser des graphiques ou de saisir des données. Grâce aux indicateurs de performance, le tableau de bord facilite la prise de décision concernant les patients, améliore les soins [186-188] et porte l'attention sur les aspects critiques dans la prise en charge des patients en fonction de la complexité de leur état de santé [189]. La mise à jour en temps réel permet de suivre des épidémies, de mettre en évidence de nouvelles informations et de faciliter la prise de décision [70, 189, 190]. Les tableaux de bord peuvent être appliqués à différentes sources de données pour permettre aux professionnels de santé d'avoir accès à des informations agrégées, et de visualiser les informations disponibles [189]. On retrouve des tableaux de bord dans des secteurs tels que les services hospitaliers [191], le suivi du diabète [192], la gestion des soins infirmiers [193] et le contrôle des infections hospitalières [194].

Afin de développer un tableau de bord, il convient de choisir les sources de données, les indicateurs de performance et les graphiques les plus adaptés aux besoins des utilisateurs [188, 195]. Pour éviter la complexité liée aux sources multiples et à la gestion des données, il est recommandé de développer son tableau de bord à partir d'EDS [196]. La convivialité et la facilité d'utilisation de l'interface doivent également être prise en compte. L'ergonomie de l'outil peut être évaluée avant le déploiement. Pour cela, le Centre d'Investigation Clinique - Innovation Technologique (CIC-IT) de Lille dispose d'ergonomes spécialisés dans l'évaluation et la conception de dispositifs médicaux innovants et/ou d'applications informatiques de santé (interfaces homme/machine) [197].

Un tableau de bord, fondé sur les données de soins premiers de la MSP de Wattrelos au format OMOP, a été conçu pour donner aux MG une vue d'ensemble de leur activité et du suivi de leurs patients. Il inclut également les indicateurs de la ROSP (définis dans la section 1.3.2 de l'introduction).

4.2.2 - Méthodes

Le développement des tableaux de bord s'est déroulé en plusieurs étapes. Dans un premier temps, des échanges avec les médecins de la MSP de Wattrelos ont permis d'identifier et de cibler les besoins des MG. Dans un second temps, une maquette (i.e., première version) d'un tableau de bord a été développée à partir des technologies R (version 4.4.0) et R Shiny. Enfin, nous avons développé une deuxième version du tableau de bord en tenant compte d'un rapport

d'évaluation du tableau de bord réalisé par les ergonomes du CIC-IT.

4.2.2.1 - Données

Les données de l'intégralité de la base de données au format OMOP de la MSP de Wattrelos ont été exploitées pour le tableau de bord. Les informations utilisées concernaient les patients adultes, couvrant la période de 1997 à 2023.

4.2.2.2 - Identification des besoins

Des réunions avec les MG de la MSP ont permis de déterminer les filtres permettant aux utilisateurs de mettre à jour l'interface en fonction de leurs choix, ainsi que les indicateurs et les thèmes souhaités. Ces thèmes ont conduit au déploiement de plusieurs panels (ou fenêtres) pour afficher les indicateurs et graphiques associés.

Les indicateurs évoqués par les MG étaient ceux contenus pour la ROSP. Ces indicateurs de la ROSP sont annuellement calculés par l'assurance maladie, à partir des données de soins facturés. Afin de les adapter aux soins premiers, ils ont été calculés sur notre base de données au format OMOP. Certains critères nécessitaient également une recherche textuelle dans la table *NOTE*, à l'aide de RegEx.
Les indicateurs que nous avons implémenté sur le tableau de bord étaient basés sur les thèmes Prévention et Suivi des pathologies chroniques de la ROSP adultes, issus du rapport de la CNAM [124]. Nous avons tenu un tableau de suivi du calcul de chaque indicateur pour comparer les résultats obtenus sur la base de données aux résultats du rapport annuel d'un MG. Ce tableau de suivi comportait également des renseignements sur les difficultés de calcul (par exemple, la recherche d'actes prescrits dans les données textuelles).

4.2.2.3 - Déploiement et évaluation

Une première version du tableau de bord a été déployée suivant les filtres, les thèmes et les indicateurs souhaités par les MG. Un ergonome a évalué le tableau de bord et a effectué un rapport heuristique. L'évaluation vérifiait que l'interface graphique utilisateur respectait une liste de critères d'utilisabilité. La liste était composée de 85 critères regroupés en 11 dimensions et associés à un niveau (i.e., majeure, mineure) [198]. Trois évaluateurs ont comparé chaque critère au tableau de bord indépendamment. Un consensus a permis de mettre en commun les résultats des critères des trois évaluations. Une deuxième version du tableau de bord a été déployée en y intégrant les indicateurs calculés de la ROSP.

4.2.3 - Résultats

4.2.3.1 - Identification des besoins

Les thèmes qui ont émergés des discussions avec les MG regroupaient l'activité générale du cabinet, les prescriptions, les valeurs biologiques et les indicateurs de la ROSP.

Dans le tableau de bord, ces thèmes étaient répartis dans différentes fenêtres. Dans chacune, les indicateurs associés aux thèmes étaient présentés :

— Activité générale : nombre de patients, nombre de consultations et nombre médian de consultations par patient ;

— Prescriptions : nombre de consultations avec au moins une prescription, nombre médian de médicaments par prescription ;

— Biologie : nombre de résultats de biologie ;

— ROSP : jauge avec l'objectif et le résultat de chaque critère.

Neuf indicateurs de la ROSP ont été calculés et comparés au rapport de la CNAM, à partir des données des patients du Docteur Calafiore, sur l'année 2021. Le tableau de suivi indiquait les données extraites pour chaque calcul, le dénominateur et le numérateur de l'indicateur, les difficultés et les critères non calculés (cf. Annexe D).
Les indicateurs proches des résultats du rapport de l'assurance maladie étaient :

— Nombre de dosage d'hémoglobine glyquée : Part des patients traités par antidiabétiques ayant bénéficié d'au moins deux dosages d'hémoglobine glyquée dans l'année ;

— Dépistage de maladies rénales chroniques : Part des patients de moins de 81 ans traités par antidiabétiques ayant bénéficié d'une recherche annuelle de microalbuminurie sur échantillon d'urines et d'un dosage annuel de la créatininémie avec estimation du débit de filtration glomérulaire ;

— Dépistage des maladies rénales chroniques chez les hypertendus : Part des patients traités par antihypertenseurs ayant bénéficié d'une recherche annuelle de protéinurie ou de microalbuminurie et d'un dosage annuel de la créatininémie avec estimation du débit de filtration glomérulaire ;

— Surveillance d'un traitement par anti-vitamine K : Part des patients traités par anti-vitamine K au long cours ayant bénéficié d'au moins autant de dosages de l'International Normalized Ratio dans l'année que de délivrances d'anti-vitamine K.

La principale difficulté rencontrée sur les autres critères était les problèmes d'extraction d'informations dans le texte (RegEx) et concernait les indicateurs :

— Traitement par benzodiazépine anxiolytique : Part des patients ayant initié un traitement par benzodiazépine anxiolytique et dont la durée de traitement est > à 12 semaines ;

— Fond d'oeil chez le diabétique : Part des patients traités par antidiabétiques ayant bénéficié d'une consultation ou d'un examen du fond d'oeil ou d'une rétinographie dans les deux ans et un trimestre.

Les critères non applicables étaient les critères concernant les enfants, les critères sur les ALD (information non disponible dans la base de données), les critères déclaratifs (critères que les MG doivent eux-même déclarer), les critères sur des prescriptions de dépistages (données textuelles) et les critères sur l'efficience (par manque de temps).

4.2.3.2 - Déploiement et évaluation

Sur la première version du tableau de bord, un panneau de filtres sur la gauche de l'interface permettait de sélectionner le sexe des patients, l'âge, la période souhaitée, et le médecin souhaité (Figure 4.4). Le panel des indicateurs de la ROSP n'avait pas été déployé sur la première version du tableau de bord.

FIGURE 4.4 – Maquette de la première version du tableau de bord.

L'évaluation initiale par les ergonomes a catégorisé 17 critères en non applicables, 42 respectés et 26 non respectés, dont 20 majeurs et 6 mineurs (Figure 4.5).

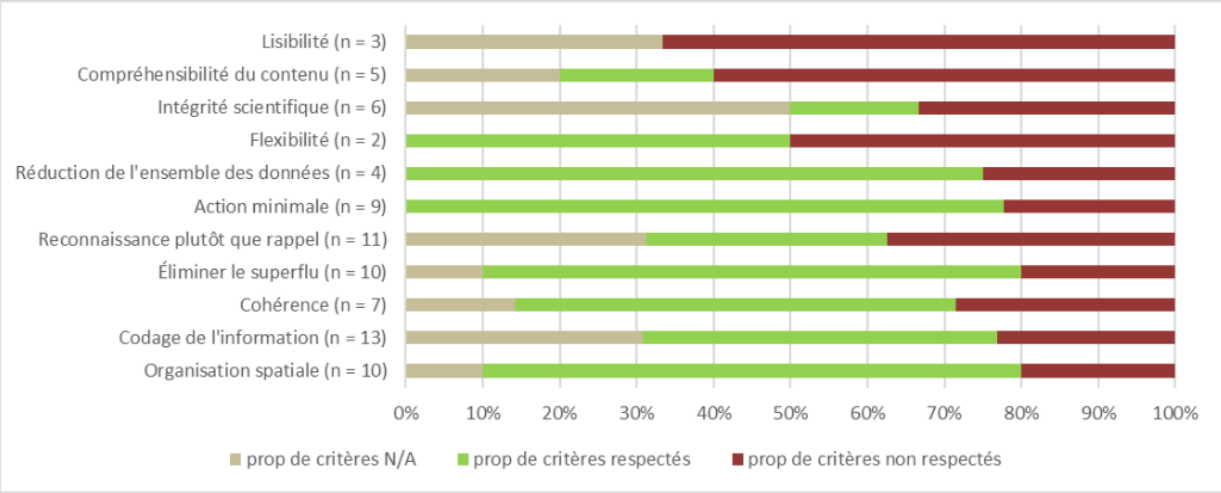


FIGURE 4.5 – Représentation graphique extraite du rapport des ergonomes sur la proportion de critères non applicables, respectés et non respectés.

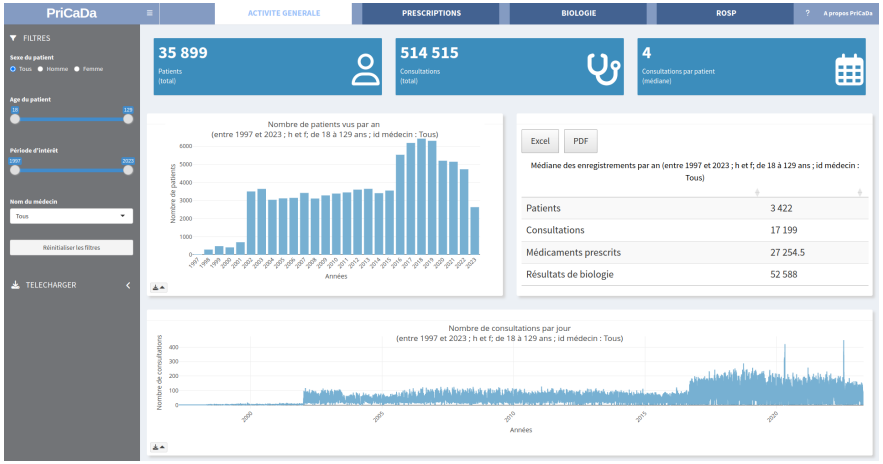
Les critères non respectés ont été séparés en deux catégories :

— L'apparence générale en panel (onglet actif lié visuellement à sa page de par sa couleur de fond) ;

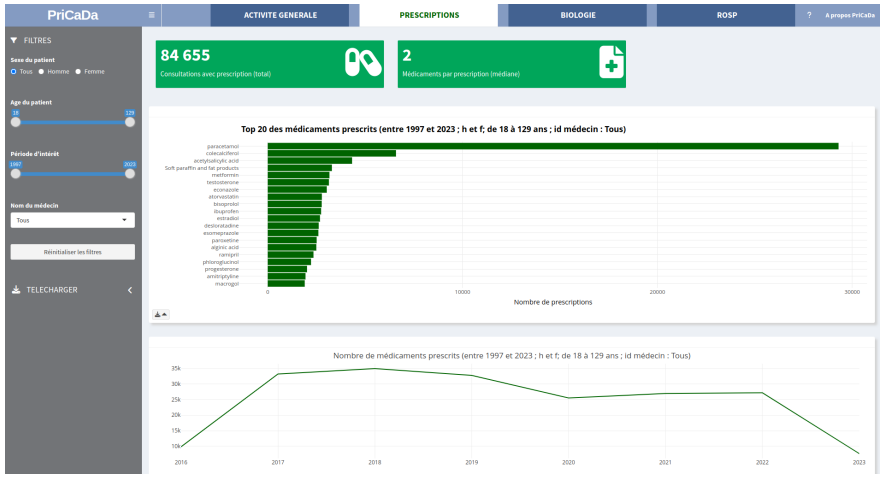
— La lecture des graphiques et tableaux (titre précis adapté aux filtres en cours, unités sur les axes, légendes, fonction zoom).

Des propositions et des recommandations ont été faites pour chacun des 26 critères non respectés par les ergonomes pour améliorer le tableau de bord (cf. Annexe E).

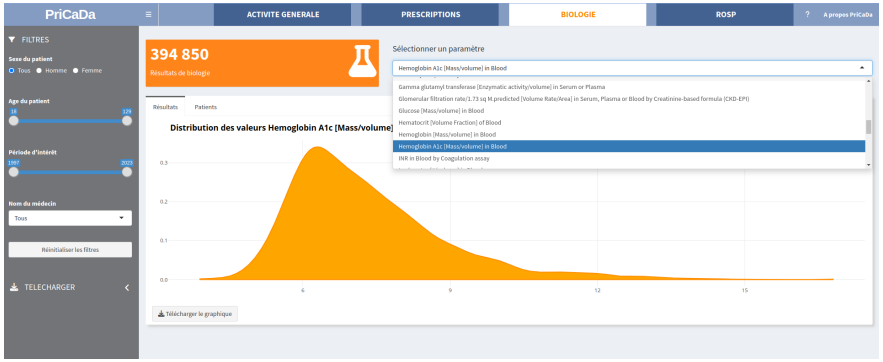
Ces recommandations ont permis le déploiement d'une seconde version du tableau de bord (Figure 4.6). Cette version intégrait une harmonisation des couleurs par thème, un bouton de réinitialisation pour mettre les filtres sur les positions de défaut, et le panel des indicateurs de la ROSP avec une représentation des critères sous forme de jauge (objectifs et résultats) (Figure 4.6d).



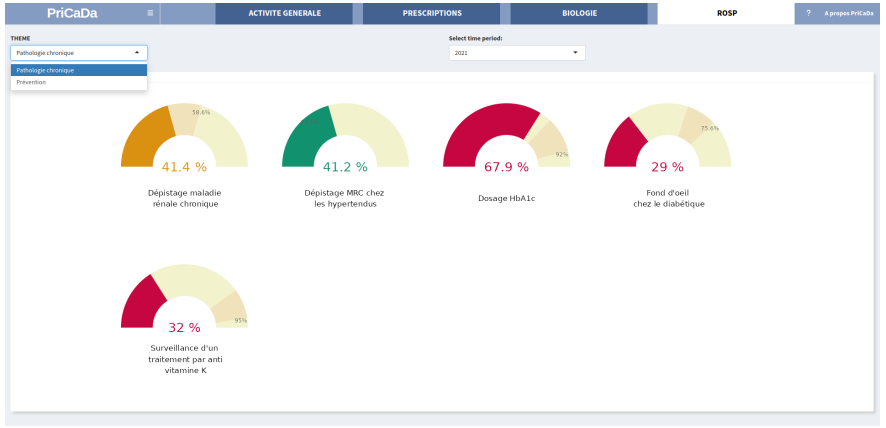
(a) Activité de la MSP



(b) Prescriptions de médicaments



(c) Résultats de biologie



(d) Indicateurs de la ROSP

FIGURE 4.6 – Différents panels de la deuxième version du tableau de bord.

4.3 Discussion

4.3.1 - Principaux résultats

Les deux outils développés à partir des données de soins premiers au format OMOP avaient pour objectif de donner aux MG un aperçu de la prise en charge des patients. Ces outils permettaient l'amélioration des pratiques et l'ajustement des prises en charge des patients par le MG.
L'algorithme de détection de séquences de traitements prescrits a permis d'aider le MG à identifier les patients atteints de pathologies chroniques, comme le diabète, présentant des difficultés à respecter de la persévérance de leur traitement. L'évaluation de la persévérance s'est faite en amont des délivrances de la pharmacie et a permis d'évaluer l'action du MG dans la prise en charge, indépendamment de la réponse du patient.

La majorité des renouvellements prescriptions d'antidiabétiques étaient anticipés ; c'est-à-dire lorsque la demande de represcription avait lieu avant la fin de la prescription initiale. La durée médiane entre toutes les prescriptions était de -2,4 jours. Cela indiquait une anticipation de plus de deux jours pour la moitié des prescriptions antidiabétiques. En outre, 11 % des prescriptions étaient suivies d'un arrêt de traitement. Ce résultat était probablement sous-estimé en raison de problèmes liés aux informations disponibles et à la posologie. La Metformine était le médicament le plus prescrit et représentait un taux de 9.1 % de prescriptions suivies d'un arrêt. La Glibenclamide et l'Umuline rapide avaient les taux les plus élevés de prescriptions suivies d'un arrêt de traitement.

Le tableau de bord a permis un aperçu global de l'activité du MG et de la prise en charge des patients. Le suivi des indicateurs de la ROSP a donné une indication aux MG sur leur progression pour atteindre l'objectif fixé par la CNAM. L'évaluation ergonomique de la première version du tableau de bord a identifié 26 critères, non respectés, à améliorer. Ces 26 critères ont facilité le développement d'une deuxième version du tableau de bord en y incluant neuf indicateurs de la ROSP concernant les patients adultes.

4.3.2 - Forces

Les données de posologie des traitements ont permis de calculer des durées de séquences de traitements. Les traitements avec un taux élevé de manque de persévérance ont ainsi pu être identifiés et analysés. L'identification d'un manque de persévérance peut pousser le professionnel de santé à discuter avec son patient et à lui proposer un traitement plus adapté.
Le tableau de bord a transformé les données de la MSP, enregistrées de 1997 à 2023, en visualisation directe et didactique. Il a également permis de filtrer les données, pour ajuster des indicateurs, et de mettre à jour automatiquement les graphiques et les indicateurs. Les données

prescrites ont été utilisées pour calculer les indicateurs de la ROSP et pour évaluer le travail des professionnels de santé. L'assurance maladie calcule les indicateurs de la ROSP à partir des données de facturation (i.e., des actions menées par le patient). Le suivi des indicateurs permettait aux professionnels d'améliorer la prise en charge des patients concernés par chaque indicateur et, à terme, de bénéficier d'un avantage financier grâce à ROSP.

4.3.3 - Limites

Le manque d'informations sur le suivi des patients en dehors de la MSP a pu biaiser l'identification du manque de persévérance d'un traitement, lors de l'application de l'algorithme de détection de séquences de traitements. Par exemple, un patient diabétique pouvait être suivi par un diabétologue tout en recevant des prescriptions d'antidiabétiques de la part de son MG pour des dépannages. Les bases de données de la MSP ne contenaient pas d'informations sur le suivi et les prescriptions faites par le diabétologue. En effet, l'algorithme de détection de séquences de traitements reposait sur les ordonnances prescrites par les médecins d'un même établissement sans avoir accès aux ordonnances prescrites par un autre spécialiste (par exemple, dans un autre cabinet ou à l'hôpital) (Figure 4.7). Par conséquent, les données d'une structure pouvaient refléter d'un manque de persévérance alors que le patient avait simplement consulté dans d'autres établissements. Ce changement de professionnels de santé pouvait donc entraîner des informations manquantes dans la prise en charge du diabète.

FIGURE 4.7 – Suivi de la persévérance à un traitement continu sur différents milieux médicaux. *MSP = maisons de santé dont les données des différents MG sont intégrées au format OMOP.*

Les données permettant d'évaluer la persévérance d'un traitement étaient uniquement basées sur la prescription du médecin. Les informations sur les délivrances de médicaments n'étaient pas disponibles dans les données des MG. De plus, la qualité des données saisies par les professionnels de santé pouvait biaiser la fiabilité des outils développés. Les informations saisies en texte libre nécessitaient un champ lexical couvrant les variations grammaticales des concepts à identifier. Par exemple, un indicateur de la ROSP calculait le nombre de fonds d'œil réalisés par le patient diabétique. Cependant, les prescriptions d'actes de fond d'œil pouvaient être faites à l'oral (information non saisie donc non disponible), écrites par le MG (information saisie en texte libre dont la forme grammaticale dépend du professionnel), ou prescrites mais non réalisées par le patient. Les actes ou traitements prescrits mais non réalisés par le patient pouvaient ainsi altérer et surestimer le calcul des indicateurs.

4.3.4 - Comparaison aux autres bases de données

La variation des données disponibles dans les différents domaines de santé (les EDSH, le SNDS, les bases de données de soins premiers et données de santé issues d'internet) limite leur capacité à répondre aux mêmes objectifs. Les données de médicaments, par exemple, sont utilisées différemment selon le contexte.
À l'hôpital, elles servent à identifier les réponses biologiques à l'administration d'un traitement [199]. Dans le SNDS, elles permettent d'évaluer les délivrances de médicaments ou la dose journalière de prise de traitement [200, 201]. Sur Internet, les forums sont souvent le lieu où les utilisateurs postent des déclarations d'effets secondaires suite à la prise d'un traitement [46, 202, 203]. Le profiling, qui consiste à associer différentes informations à un profil unique, peut retracer les événements vécus par un utilisateur à partir des données des forums [204]. Enfin, en soins premiers, plus spécifiquement dans les logiciels des MG, les données de prescriptions médicamenteuses incluaient la durée du traitement, le nombre de prises et la dose.
Cependant, sur ces quatre sources de données le chaînage des informations pour un même patient n'est pas encore possible et requiert une identification de chaque patient.

Les tableaux de bord sont relativement bien développés pour les EDSH [188, 196]. Ils sont utilisés pour le suivi d'indicateurs spécifiques à un thème ou un service précis [75]. En revanche, le SNDS n'étant pas intégré dans un modèle de données commun, reconnu à l'international, le développement d'outils ou de tableaux de bord réplicables sur différentes bases de données est compromis. En soins premiers, le tableau de bord est utile pour connaître l'activité du médecin sur une période choisie. Un suivi de l'activité aide le médecin à améliorer des indicateurs de performance pour obtenir la ROSP.

Analyses des données de soins premiers

Chapitre 5

Analyses des données de soins premiers

Sommaire

5.1 Contexte

En France, en 2022, la prévalence des personnes diabétiques était estimée à 5,6 %, soit plus de trois millions de patients [138, 205, 206]. En ce qui concerne le Nord de la France, cette prévalence s'élève à 6,5 % [206]. Il existe deux types de diabète. Le diabète de type 2, le plus fréquent, touche 92 % des personnes diabétiques et est lié à une baisse de sensibilité à l'insuline [206]. Ce type de diabète apparaît avec l'âge, la sédentarité, le surpoids et le manque d'activité physique ; il peut être traité par rééquilibrage alimentaire ou par traitement antidiabétique [205, 207]. Le diabète de type 1 est plus rare et apparaît dans l'enfance. Il se traduit par une destruction des cellules du pancréas qui produisent l'insuline [205]. Le traitement du diabète de type 1 se fait par injection d'insuline [205, 207]. Les facteurs de risques du diabète seraient liés à une prédisposition génétique ou environnementale [206]. Les deux types de diabètes sont associés à une surmortalité, à une morbidité importante et à une utilisation considérable des ressources de santé, et donc une augmentation des dépenses [208].

La Haute Autorité de Santé (HAS) recommande plusieurs critères de suivi des personnes diabétiques, comme l'évaluation de la créatinine au moins une fois par an, et propose un guide de parcours de soins [209].
Des études réalisées à partir de la réutilisation des données de personnes diabétiques ont permis d'évaluer l'impact biologique [138, 183, 210] ou les effets secondaires [211] suite à la prise d'antidiabétiques. Certaines de ces études ont été réalisées à partir de données de soins premiers [138, 210] ou de bases de données nationales [183, 208, 211]. Pour améliorer le suivi des personnes diabétiques, des modèles de prédiction sur les complications [59] ou des visualisations sous forme de tableaux de bord [189, 192] ont été développés pour le médecin. La majeure partie de la prise en charge d'une personne diabétique se fait en médecine de ville (i.e., en soins premiers) [212].

En France, les études sur les personnes diabétiques basées sur des données de ville se font à partir de programmes d'inclusion de patients [213], du SNDS [214, 215] ou de données des MG [138]. Les études sur les personnes diabétiques ne sont pas réalisées sur des bases de données standardisées et ne peuvent donc pas être partagées à d'autres centres ou à la communauté de chercheurs.
Nous avons étudié les personnes diabétiques sur les données de soins premiers. Pour cette étude, les bases de données des différentes MSP implémentées selon la stratégie définie précédemment (cf. Chapitre 3) ont été utilisées. Dans une première partie, l'objectif était de décrire le suivi des patients, de la MSP de Wattrelos, sous antidiabétiques et les réponses biologiques à la prise de traitements. Dans une seconde partie, nous avons étendu l'étude pour comparer le suivi des personnes diabétiques dans les quatre MSP.

5.2 Suivi des patients sous traitement antidiabétique dans la MSP de Wattrelos

5.2.1 - Matériels et méthodes

5.2.1.1 - Données

Nous avons travaillé sur la base de données de la MSP de Wattrelos qui regroupe les informations de huit médecins généralistes exerçant depuis 1997. Cette base de données a été structurée au format OMOP et l'implémentation de l'ETL est détaillée dans les chapitres 2 et 3. Les données couvraient la période de 1997 à 2023.

5.2.1.2 - Population

Nous avons sélectionné les patients de 18 ans et plus, actifs depuis 2018, c'est-à-dire ceux ayant eu au moins une consultation entre le 1er janvier 2018 et le 1er janvier 2023, et ayant

reçu au moins une prescription d'antidiabétiques (ADO ou insuline). Pour cela, nous avons utilisé les codes ATC et les concepts standards associés (cf. Annexe F).

5.2.1.3 - Variables

Pour chaque patient, nous avons extrait l'âge, le sexe, les dates des consultations, les prescriptions médicamenteuses et les résultats des analyses biologiques.
Les prescriptions d'antidiabétiques ont été extraites de la table *DRUG_EXPOSURE*, en utilisant les codes ATC commençant par "A10". Les prescriptions d'antidiabétiques ont été regroupées en trois classes en fonction de ces codes ATC :

— les traitements insuline (codes ATC commençant par "A10A"),

— les traitements ADO (codes ATC ne commençant pas par "A10A" et n'appartenant pas à la liste des traitements combinés suivante),

— les traitements ADO + insuline (codes ATC : A10BD01, A10BD02, A10BD03, A10BD04, A10BD05, A10BD06, A10BD07, A10BD08, A10BD09, A10BD10, A10BD11, A10BD12, A10BD13, A10BD14, A10BD15, A10BD16, A10BD17, A10BD18, A10BD19, A10BD20, A10BD21, A10BD22).

Nous avons extrait les résultats de l'hémoglobine glyquée de la table *MEASUREMENT* avec le concept standard "Hemoglobin A1c [Mass/volume] in Blood" de la classification LOINC (*concept_id* = 3034639). Nous n'avons conservé que les résultats de l'hémoglobine glyquée compris entre la date de début et la date de fin d'un traitement antidiabétique. Pour cela, nous avons appliqué l'algorithme de détection de séquences de traitements (voir chapitre 4), qui permettait de récupérer les dates de début et de fin de chaque séquence de traitement. Un diabète est considéré comme équilibré lorsque les valeurs de l'hémoglobine glyquée sont inférieures ou égales à 7 % [216]. Nous avons calculé le nombre de valeurs au-dessus de ce seuil pour chaque classe de traitement d'antidiabétiques. Enfin, les résultats biologiques ont été alignés à la date de la première prescription du traitement (T0).

5.2.1.4 - Statistiques

Les statistiques ont été réalisées avec R Studio et le langage R(4.4.0). Les variables qualitatives ont été exprimées en pourcentages. Le test de normalité (test de Shapiro) a été appliqué aux variables quantitatives. Les variables suivant une distribution normale ont été décrites par la moyenne et l'écart-type, tandis que celles ne suivant pas une loi normale ont été exprimées par leur médiane ainsi que leur premier quartile (Q1) et leur troisième quartile (Q3).
La comparaison des moyennes entre deux groupes a été réalisée à l'aide du test de Student, à condition que les variables suivaient une loi normale. En cas d'absence de normalité, le test de Mann-Whitney a été utilisé pour comparer les médianes de deux groupes. Pour la comparaison

des moyennes de plusieurs groupes, le test ANOVA a été appliqué. Ces tests généraient une p-value ; lorsque cette valeur est inférieure à 5% (0,05), la différence observée entre les groupes était significative.

5.2.2 - Résultats

5.2.2.1 - Population

Sur la période de janvier 2018 à janvier 2023, 769 patients adultes ont été traités par antidiabétiques dans la MSP de Wattrelos et 10 807 consultations comportaient des prescriptions d'antidiabétiques. Parmi ces consultations, 5 270 (48,8 %) concernaient les patients de sexe féminin et 5 528 (51,2 %) les patients de sexe masculin (p-value $< 0{,}001$). L'âge médian [Q1 ; Q3] à la consultation était de 69 ans [60 ; 76]. L'âge était plus élevé chez les femmes que chez les hommes, avec respectivement 70 ans [60 ; 78] et 68 ans [59,7 ; 75] (p-value $< 0{,}001$). La majorité des consultations de patients masculins concernait des patients âgés de 73 ans à 77 ans (n=1 057, 20 %) alors que les consultations de patientes âgées de 68 ans à 72 ans étaient plus fréquentes (n=945, 17 %) (Figure 5.1).

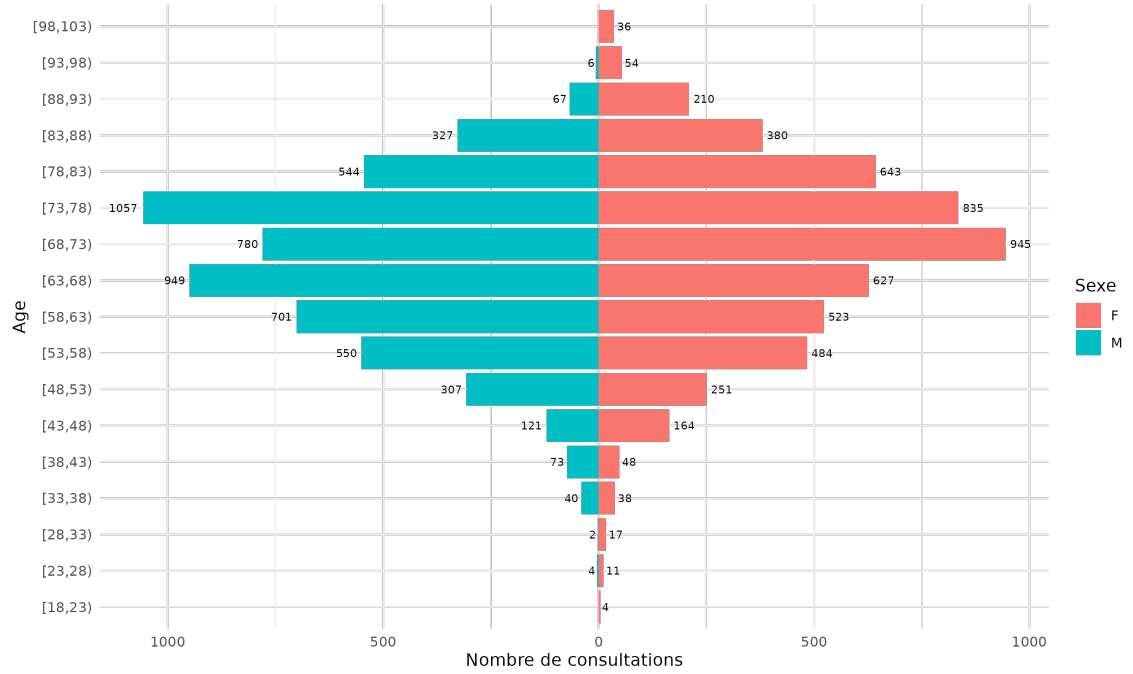


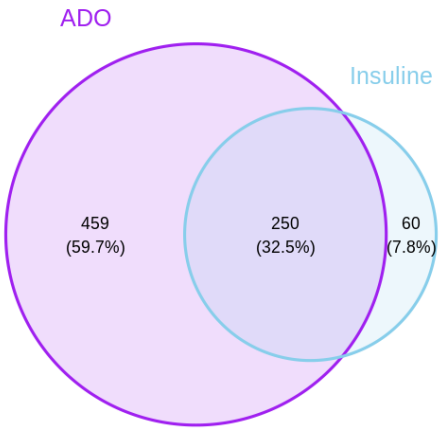
FIGURE 5.1 – Pyramide des âges du nombre de consultations avec prescription d'antidiabétiques. *F = féminin, M = masculin*.

5.2.2.2 - Prescriptions d'antidiabétiques

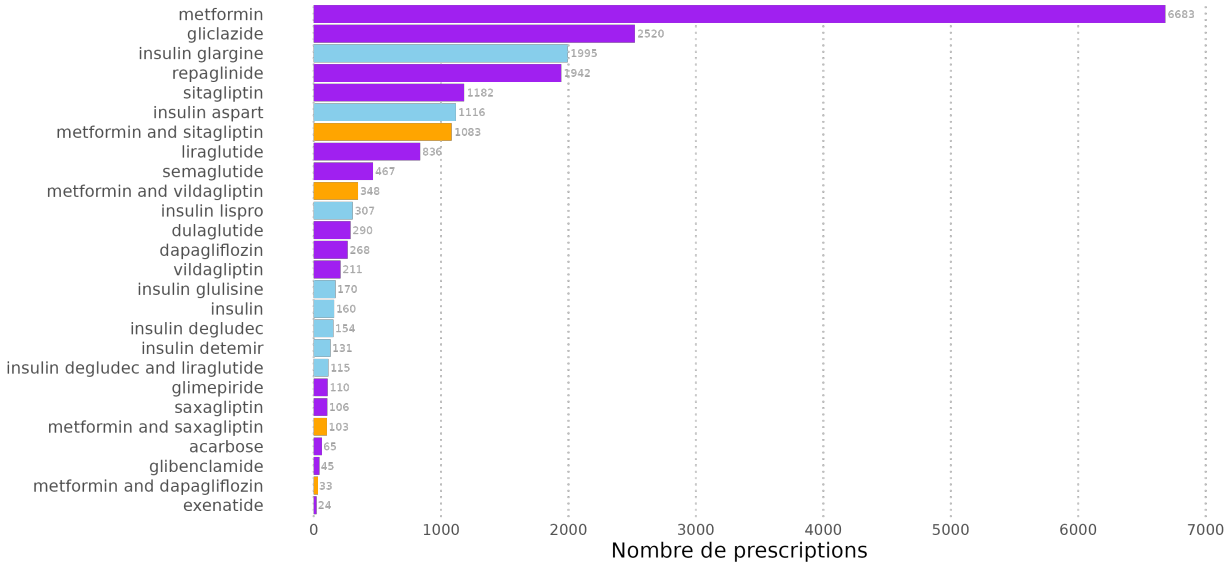
Sur la période de janvier 2018 à janvier 2023, 20 320 antidiabétiques ont été prescrits : 4 128 étaient des prescriptions d'insuline (20,3%), 14 641 des prescriptions d'ADO (72%) et 1

551 des prescriptions combinées d'ADO et d'insuline (7,7%). 459 patients (59,7%) ont reçu uniquement des traitements ADO et 60 patients (7,8%) des traitements insuline. 250 patients (32,5%) ont eu des traitements d'ADO et d'insuline (Figure 5.2a).

La Metformine était le traitement d'ADO le plus prescrit (n=6 683, 46%), suivi de la Gliclazide (n=2 520, 17%) (Figure 5.2b). Concernant les insulines, l'insuline glargine (insuline lente) était l'insuline la plus prescrite (n=1 995, 48%), suivie de l'insuline aspart (insuline rapide) (n=1 116, 27%). Le traitement combiné le plus prescrit était la Metformine associée à la Sitagliptine (n=1 083, 70%).



(a) Proportion de patients selon les classes de traitements antidiabétiques.

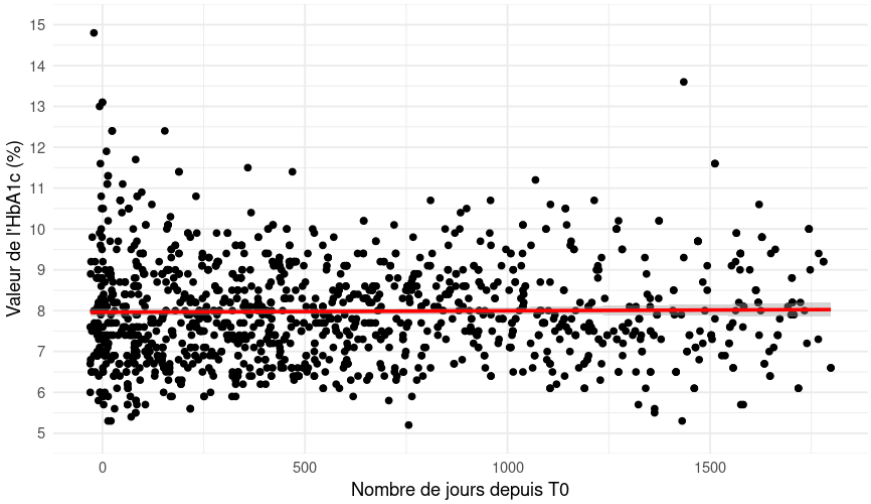


(b) Fréquence des prescriptions antidiabétiques en fonction des classes de traitement *(violet = ADO, bleu = insuline, orange = ADO + insuline)*.

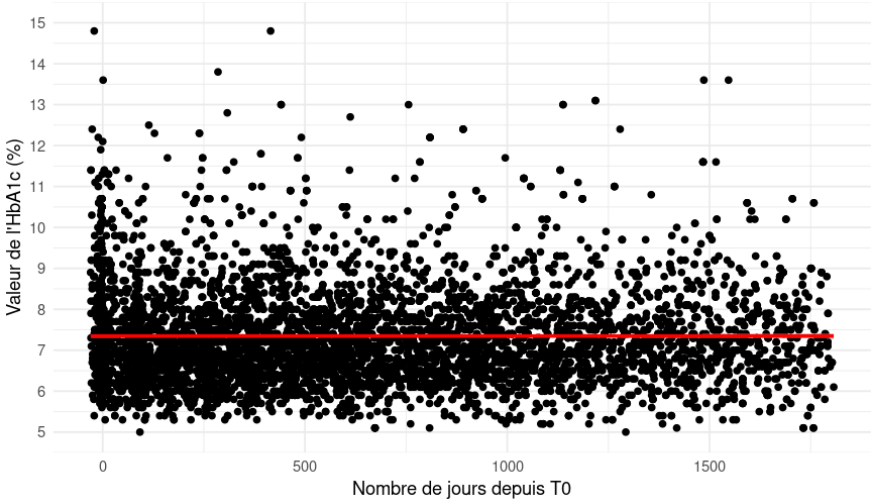
FIGURE 5.2 – Répartition des prescriptions par classes d'antidiabétiques.

5.2.2.3 - Suivi de l'hémoglobine glyquée

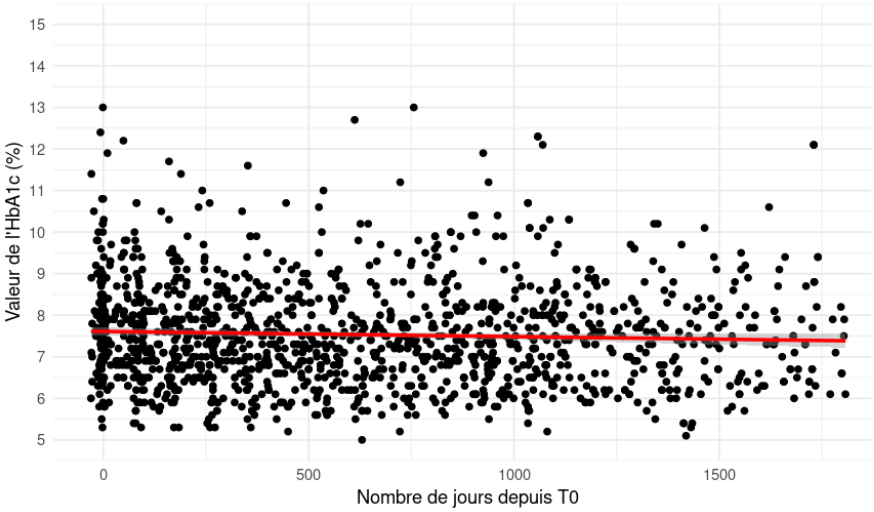
La valeur médiane [Q1 ; Q3] de l'hémoglobine glyquée était de 7,3 % [6,7 ; 8,1]. Pour les traitements d'insuline, d'ADO et les traitements combinés d'ADO et d'insuline, les valeurs médianes étaient, respectivement, de 7,8 % [7,1 ; 8,7], de 7,2 % [6,5 ; 7,9] et de 7,4 % [6,7 ; 8,2] (p-value $< 0{,}001$).
La tendance médiane des valeurs de l'hémoglobine glyquée était de 7,8 % [7,1 ; 8,7] au cours des traitements d'insuline, 7,2 % [6,5 ; 7,9] pour les traitements d'ADO et 7,5 % [6,7 ; 8,2] pour les traitements combinés (Figure 5.3).
Pour les patients sous insuline, sur 1 352, valeurs 805 étaient au-dessus de 7,5 % (59,5 %) (Figure 5.3a). 6 673 valeurs de l'hémoglobine glyquée ont été obtenues pour les patients sous ADO, 2 336 valeurs étaient supérieures à 7,5 % (35 %) (Figure 5.3b). Enfin, sur 1 533 valeurs, 670 étaient supérieures au seuil pour les patients sous traitements combinés (ADO + insuline) (43,7 %) (Figure 5.3c).



(a) Mesures de l'hémoglobine glyquée des patients sous insuline.



(b) Mesures de l'hémoglobine glyquée des patients sous ADO.



(c) Mesures de l'hémoglobine glyquée des patients sous traitement combiné (ADO + insuline).

FIGURE 5.3 – Évolution des valeurs de l'hémoglobine glyquée des patients sous séquences de traitements antidiabétiques.

5.3 Suivi des patients sous traitement antidiabétique dans quatre MSP

5.3.1 - Matériels et méthodes

5.3.1.1 - Population

Nous avons inclus les patients de plus de 55 ans, pris en charge dans les quatre MSP (Wattrelos, Lille-Moulins, Guesnain et Tourcoing) avec au moins une prescription d'antidiabétiques en 2022. Ces bases de données ont été implémentées au format OMOP. Le développement des ETL a été décrit dans le chapitre 3.

5.3.1.2 - Données extraites

Pour les patients répondant aux critères d'inclusion, nous avons extrait les données de leurs consultations, de leurs résultats biologiques et de leurs prescriptions médicamenteuses depuis leurs 55 ans. Les prescriptions d'antidiabétiques ont été classées selon les catégories d'âge suivantes : 55-64 ans, 65-74 ans, 75-84 ans, et >= 85 ans.

Une ordonnance est la combinaison de traitements dispensés à un patient au même moment et par le même prescripteur. Les ordonnances ont été catégorisées par les classes : ADO, insuline ou traitement combiné d'ADO + insuline. Lorsqu'une ordonnance contenait une prescription d'insuline et une prescription d'ADO, elle était affectée à la classe des traitements combinés ADO + insuline.

La table *DRUG_EXPOSURE* a été filtrée sur les code CIP correspondant aux codes ATC commençant par "A10" (antidiabétiques). Chaque code ATC a été affecté à une classe (ADO, insuline et ADO + insuline) et à une sous-famille :

— Les ADO : inhibiteurs des DPP-4 (dipeptidylpeptidase-4), Metformine, Répaglinide, Sulfamide, inhibiteurs des alpha-glucosidases, inhibiteurs des SGLT2,

— Les insulines : les antogonistes des récepteurs au GLP-1 (glucagon-like peptide-1), les insulines rapides, les insulines lentes,

— Les ADO + insuline.

Les résultats de l'hémoglobine glyquée ont été extraits de la table *MEASUREMENT* à l'aide du *concept_id* 3034639. Les résultats de la créatinine ont été extraits à partir du concept "Creatinine [Mass/volume] in Blood" de la classification LOINC (*concept_id*=3051825). Pour éviter les redondances, seules une valeur de créatinine dans le sang et une valeur de l'hémoglobine glyquée exprimées en pourcentage ont été prises en compte à chaque date d'analyse biologique. Pour évaluer les résultats de biologie sous traitement, l'algorithme de détection des séquences de traitements a été appliqué (voir Chapitre 4) pour récupérer les dates de début et de fin de

chaque traitement. Un arrêt de traitement tolérable a été fixé à 60 jours après la date de fin de prescription (ou date de début lorsque des informations étaient manquantes). Les valeurs biologiques par catégorie de traitement ont été récupérées par patient et par an. Enfin, les résultats ont été comparés entre les différentes MSP.

5.3.1.3 - Statistiques

La description des données incluaient l'âge des patients à la prescription, les proportions d'ordonnances par tranche d'âge, le nombre d'ordonnances annuelles, ainsi que les proportions d'ordonnances contenant uniquement des ADO, uniquement des insulines, ou une combinaison d'insuline et d'ADO. De plus, le nombre de résultats de biologie, ainsi que les proportions de résultats pour la créatinine et l'hémoglobine glyquée ont été également calculés.
Nous avons calculé les proportions de patients ayant au moins un résultat de créatinine ou d'hémoglobine glyquée et les proportions de patients ayant reçu au moins une ordonnance contenant exclusivement des traitements de chaque classe de traitements.

Les variables qualitatives ont été exprimées en effectifs et pourcentages. Le test de normalité (test de Shapiro) a été appliqué aux variables quantitatives. Les variables suivant une distribution normale ont été décrites par la moyenne et l'écart-type (moyenne(+/-ET)), tandis que celles ne suivant pas une loi normale ont été exprimées par la médiane ainsi que le premier quartile (Q1) et le troisième quartile (Q3) (médiane[Q1 ;Q3]). La comparaison des moyennes entre deux groupes a été réalisée avec le test de Student, lorsque les variables suivaient une loi normale. En cas d'absence de normalité, le test de Mann-Whitney a été utilisé pour comparer les médianes des deux groupes. Nous avons utilisé le test ANOVA pour comparer les moyennes de plusieurs groupes et le test Kruskal Wallis pour comparer les médianes. Les variables qualitatives ont été comparées par le test de Khi-deux. Nous avons considéré les variables comme différentes significativement lorsqu'une p-value était inférieure à 5% (0,05). Les statistiques ont été réalisées sur R Studio en utilisant le langage R(4.4.0).

5.3.2 - Résultats

Les MSP de Wattrelos, Lille-Moulins, Guesnain et Tourcoing comptabilisaient, respectivement, 12 753, 2 296, 953 et 8 485 ordonnances prescrites aux patients âgés de plus de 55 ans (Tableau 5.1). L'âge médian [Q1 ; Q3] était légèrement supérieur pour les patients de la MSP de Guesnain et Wattrelos, avec, respectivement, 70 ans [64 ; 75] et 70 ans [63 ; 76], contre 66 ans [62 ; 71] et 65 ans [60-71] pour Lille-Moulins et Tourcoing (p-value $< 0{,}001$). La majorité des ordonnances d'antidiabétiques étaient prescrites aux patients âgés de 65 à 74 ans dans la population générale (9 808 ordonnances, 39,5 %) et dans la MSP de Lille-Moulins (1 065 ordonnances, 46,4 %) (p-value $< 0{,}001$).

Les ordonnances d'ADO étaient les plus fréquentes dans la MSP de Guesnain avec 761 ordonnances prescrites contenant uniquement des ADO (79,8 %). Les ordonnances d'ADO étaient plus nombreuses que les ordonnances d'insuline ou de traitements combinés.
En moyenne (+/- écart-type), 1 484 (+/- 833,6) résultats de biologie ont été réalisés par année de suivi des patients de la MSP de Wattrelos sous traitements antidiabétiques. La moyenne de la population générale était de 314,5 [103,8 ; 633,2] résultats par an.
Les proportions de mesures de l'hémoglobine glyquée étaient plus élevées que les proportions de mesures de la créatinine (13 879 mesures de l'hémoglobine glyquée, 61 %, contre 8 853 mesures de la créatinine, 38 %, pour la population générale).

	Population générale (n=24 487)	**Wattrelos** (n=12 753)	**Lille-Moulins** (n=2 296)	**Guesnain** (n=953)	**Tourcoing** (n=8 485)	***p-value***
Âge à l'ordonnance, *années*	68,3 (+/- 8,5)	70 [63 ; 76]	66 [62 ; 71]	70 [64 ; 75]	65 [60-71]	*< 0,001*
55-64, % (n)	32,9 (8 153)	25,7 (3 277)	34,3 (789)	22 (210)	44,1 (3 740)	*< 0,001*
65-74, % (n)	39,5 (9 808)	38,7 (4 937)	46,4 (1 065)	44,3 (422)	38,2 (3 244)	*< 0,001*
75-84, % (n)	22,2 (5 509)	28,2 (3 591)	18,5 (425)	26,1 (249)	14,1 (1 198)	*< 0,001*
<=85, % (n)	5,4 (1 337)	7,4 (939)	0,7 (17)	7,5 (72)	3,6 (303)	*< 0,001*
ADO , *%*	57,4 (14 053)	64,9 (8 278)	59,5 (1 367)	79,8 (761)	43 (3 647)	*< 0,001*
Insuline, *%*	5,2 (1 284)	5,4 (689)	6,7 (155)	6,4 (61)	4,4 (373)	*< 0,001*
ADO + Insuline, *%*	37,4 (9 150)	29,7 (3 786)	33,7 (774)	13,1 (125)	52,6 (4 465)	*< 0,001*
Résultats de biologie, *n par an*	314,5 [103,8 ; 633,2]	1 485 (+/- 833,6)	265,4 (+/- 152,4)	75,7 (+/- 86,5)	233,6 (+/- 186,9)	*< 0,001*
Créatinine, *%*	39 (8 853)	41,1 (6 709)	37,2 (789)	43,2 (131)	30,8 (1 224)	*< 0,001*
Hémoglobine glyquée, *%*	61 (13 879)	58,9 (9 626)	62,8 (1 334)	56,8 (172)	69,2 (2 747)	*< 0,001*

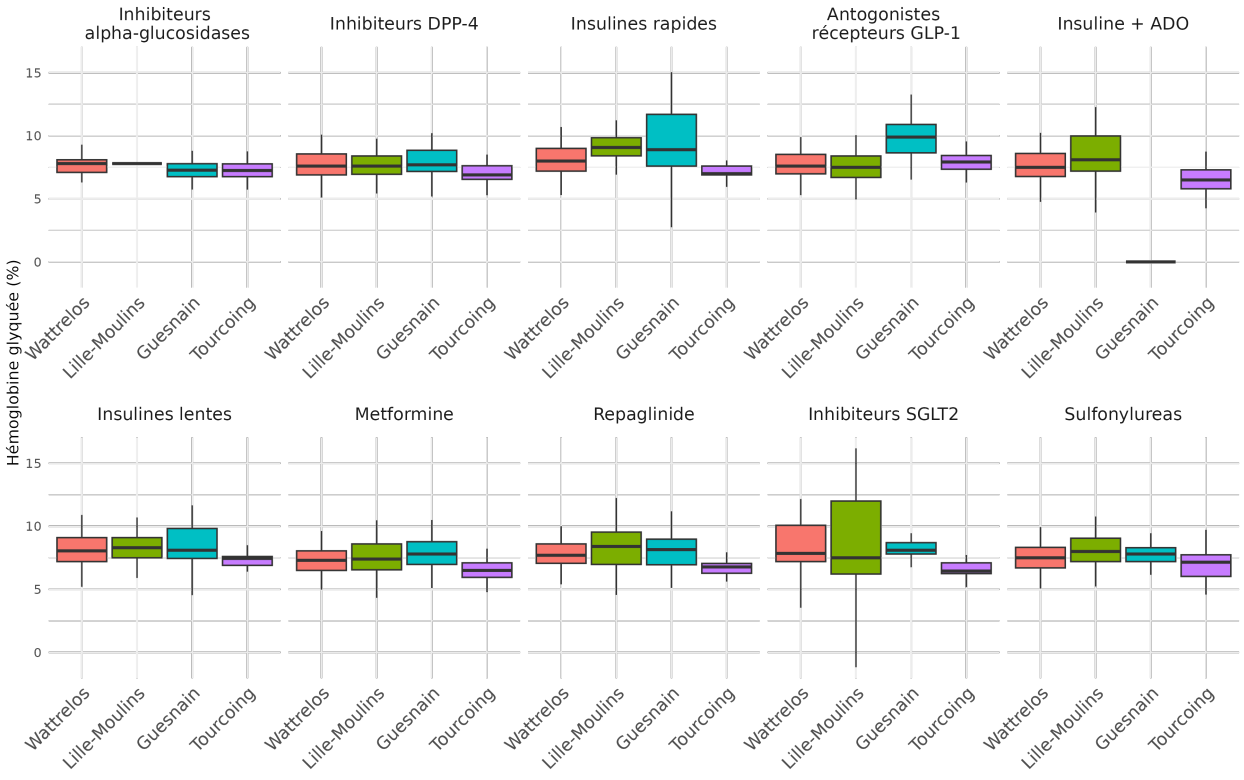
TABLE 5.1 – Description des ordonnances antidiabétiques des patients âgés de plus de 55 ans par MSP.

Dans les MSP de Wattrelos, Lille-Moulins, Guesnain et Tourcoing, 511, 134, 115 et 153 patients de plus de 55 ans ont eu respectivement des prescriptions d'antidiabétiques (Tableau 5.2). Les patients de la MSP de Wattrelos étaient majoritairement des hommes (281 patients, 55 %), le taux était plus élevé que dans la population générale (480 patients, 52,9 %) (p-value $< 0{,}001$).
Dans la MSP de Lille-Moulins, 54 patients (46,9 %) ont eu au moins une mesure de la créatinine. Ce taux était plus bas que pour la population générale (829 patients, 91,3 %). Cette différence a été également identifiée sur le taux de patients ayant eu au moins une mesure de l'hémoglobine glyquée (61 patients, 53 %, dans la MSP de Guesnain contre 838, 92,3 %, pour la population générale).
Concernant les ordonnances, les taux étaient similaires sur les quatre MSP pour le taux de patients ayant eu au moins une ordonnance d'ADO et au moins une ordonnance d'insuline (respectivement, p-value=0,5279 et p-value=0,5366)). 21 ordonnances de traitements combinés (12,3 %) ont été prescrites dans la MSP de Guesnain contre 91 (59,5 %) dans la MSP de Tourcoing (p-value $< 0{,}001$).

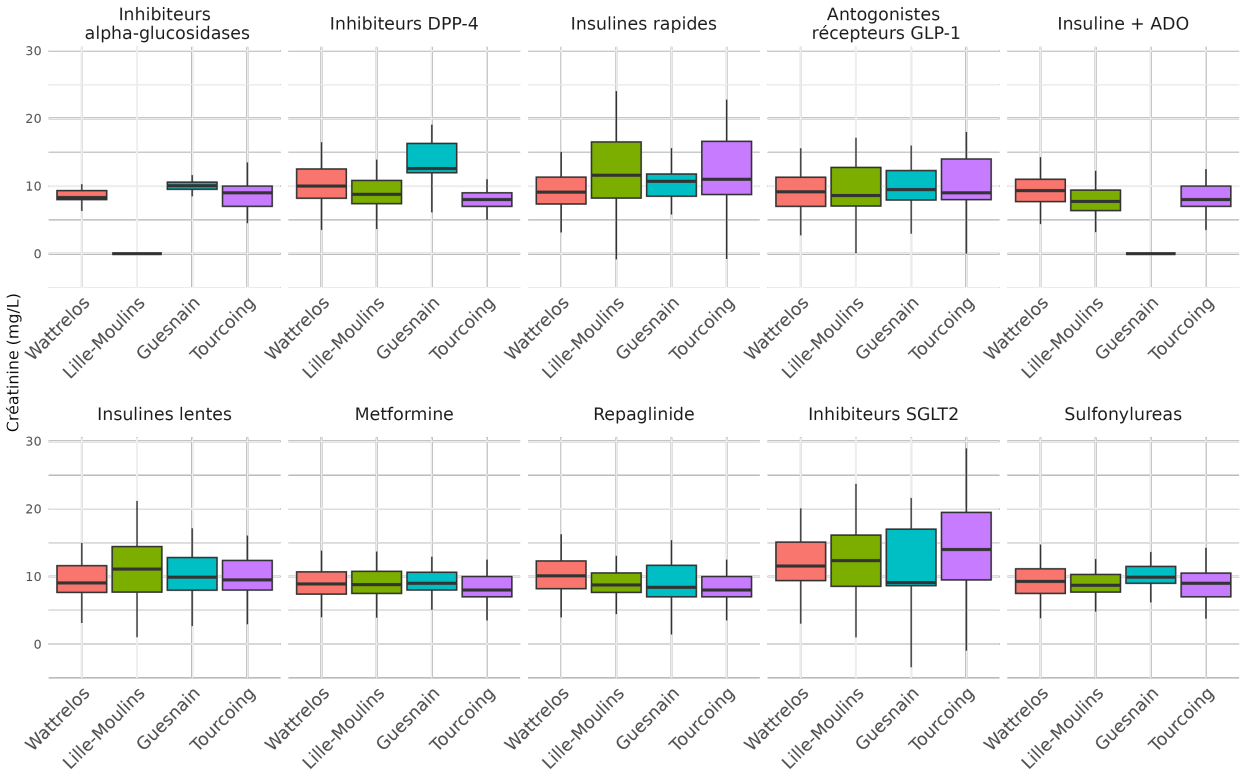
	Population générale (n=908)	**Wattrelos** (n=511)	**Lille-Moulins** (n=134)	**Guesnain** (n=115)	**Tourcoing** (n=153)	***p-value***
Hommes, *% (n)*	52,9 (480)	55 (281)	52 (70)	49,6 (57)	48 (74)	*< 0,001*
Résultats de biologie, *n patients par an*	215 [27 ; 450,5]	237 (+/- 135,3)	69,1 (+/- 28,9)	22,5 (+/- 24,6)	67 (+/- 39,6)	*< 0,001*
Créatinine, *% (n)*	91,3 (829)	98,2 (501)	95,5 (128)	46,9 (54)	98 (150)	*< 0,001*
Hémoglobine glyquée, *% (n)*	92,3 (838)	98,2 (502)	95,5 (128)	53 (61)	98,7 (151)	*< 0,001*
Ordonnances, *n patients par an*	265,1 (+/- 139,8)	265,1 (+/- 139,8)	70,4 (+/- 36,4)	82 (71,2 ; 99,2]	42,5 [9,5 ; 74]	*< 0,001*
ADO, *%(n)*	83,7 (760)	82,8 (423)	83,6 (112)	81,7 (94)	87,6 (134)	*0.5279*
Insuline, *% (n)*	13,8 (125)	15 (77)	11,9 (16)	10,4 (12)	13,7 (21)	*0.5366*
ADO + Insuline, *% (n)*	36,1 (328)	34,2 (175)	32,8 (44)	12,3 (21)	59,5 (91)	*< 0,001*

TABLE 5.2 – Description des patients sous antidiabétiques âgés de plus de 55 ans par MSP.

En comparant les valeurs biologiques entre les quatre MSP en fonction des familles de traitements, les valeurs de l'hémoglobine glyquée étaient différentes sur les MSP (p-value < 0,001) et identiques selon les traitements (p-value = 0,42) (Figure 5.4a). Les valeurs étaient plus élevées pour les patients sous antogonistes des récepteurs au GLP-1 dans la MSP de Guesnain (9,9 % [8,6 ; 10,9]). Les valeurs les plus basses ont été retrouvées pour les patients sous inhibiteurs de SGLT2 de la MSP de Tourcoing (6,4 % [6,2 ; 7,1]). Les valeurs les plus élevées dans la MSP de Tourcoing concernaient les patients sous antogonistes des récepteurs au GLP-1 (7,9 % [7,3 ; 8,4]) .
Les valeurs de la créatinine étaient identiques en fonction des MSP (p-value = 0,47) et différentes selon les traitements (p-value = 0,03). Les valeurs étaient plus élevées pour les patients sous inhibiteurs de SGLT2 de la MSP de Tourcoing (14 mg/L [9,5 ; 19,5]) (Figure 5.4b). Les valeurs les plus basses concernaient les patients sous traitement combiné (ADO + insuline) de la MSP de Lille-Moulins (7,7 mg/L [6,4 ; 9,4]). Dans les MSP de Lille-Moulins et Guesnain, les patients sous inhibiteurs des alpha-glucosidases n'avaient pas de mesures de la créatinine. Un patient était sous inhibiteurs des alpha-glucosidases dans la MSP de Lille-Moulins et n'avait pas de valeur de créatinine. Dans la MSP de Guesnain, les patients sous insuline + ADO (n=3, 2,6 %) n'avaient pas de mesure de créatinine ni d'hémoglobine glyquée durant leur période de traitement.



(a) Valeurs de l'hémoglobine glyquée des patients sous différentes familles d'antidiabétiques.



(b) Valeurs de la créatinine des patients sous différentes familles d'antidiabétiques.

FIGURE 5.4 – Valeurs des résultats de biologie sous différentes familles d'antidiabétiques.

5.4 Discussion

5.4.1 - Étude sur la MSP de Wattrelos

Dans la MSP de Wattrelos, la majorité des 10 807 consultations avec prescriptions d'antidiabétiques étaient faites à des hommes (51,2 %). 72 % des prescriptions d'antidiabétiques étaient des ADO, dont 46 % étaient de la Metformine. 20,3 % des prescriptions étaient des insulines, 48 % des prescriptions d'insuline étaient de l'insuline glargine. Plus de la moitié des valeurs de l'hémoglobine glyquée sous traitements d'insuline (59,5 % des mesures) et plus d'un tiers sous traitements d'ADO (35 % des mesures) étaient supérieures à 7,5 % .

Selon l'Inserm, la majorité des patients traités sous antidiabétiques sont des patients diabétiques de type 2 dont le diagnostic se fait en moyenne autour de 65 ans, avec une incidence maximale chez les hommes autour de 75 ans [217]. Cela correspond aux incidences maximales de patients âgés de 63 à 68 ans chez les hommes de la MSP de Wattrelos et de 73 à 78 ans chez les femmes. La Metformine est fréquemment prescrite car l'Assurance Maladie la recommande comme traitement de première intention [218, 219]. En effet, ce traitement est le plus prescrit dans la MSP de Wattrelos et représente 46 % des traitements antidiabétiques prescrits.

L'hémoglobine glyquée est un paramètre biologique indiquant la bonne prise en charge du diabète. Le taux d'hémoglobine glyquée doit être inférieur à 7 % (entre 6,5 % et 8 % selon les comorbidités et l'espérance de vie). Si le taux dépasse les 7 %, un traitement doit être instauré ou réévalué [216, 219]. La majorité des valeurs de l'hémoglobine glyquée prises au cours du traitement d'ADO était inférieure à 7,5 % (65 % des valeurs). Cela montre un maintien correct de l'équilibre de l'hémoglobine glyquée chez les patients sous ADO. Cependant, cet équilibre est à réévaluer pour les patients sous traitements d'insuline.

5.4.2 - Étude sur les quatre MSP

Dans les MSP de Wattrelos, Lille-Moulins, Guesnain et Tourcoing, respectivement, 511, 134, 115 et 153 patients de plus de 55 ans ont eu des prescriptions d'antidiabétiques. L'âge des patients était plus élevé dans la MSP de Guesnain et Wattrelos. Plus de 90 % des patients ont eu au moins une mesure de la créatinine et une mesure de l'hémoglobine glyquée durant leur suivi, sauf pour les patients de la MSP de Guesnain.
Sur les quatre MSP les valeurs de l'hémoglobine glyquée étaient différentes alors que les valeurs de la créatinine étaient les mêmes. Les valeurs de l'hémoglobine glyquée étaient identiques alors que les valeurs de la créatinines étaient différentes selon les familles de traitements d'antidiabétiques.

Selon les recommandations de la Haute Autorité de Santé, un patient diabétique doit bénéficier d'une évaluation de l'hémoglobine glyquée au moins deux fois par an et d'une

évaluation de la créatinine au moins une fois par an [209]. Presque la totalité des patients des MSP de Wattrelos, Lillle-Moulins et Tourcoing avaient plus d'une mesure de la créatinine et de l'hémoglobine glyquée. Les valeurs de l'hémoglobine glyquée sont recommandées à moins de 7 % pour un patient diabétique sans comorbidité et les valeurs de la créatinine à 13 mg/L [209]. Les valeurs de la créatinine étaient plus élevées pour les patients sous inhibiteurs de SGLT2 de la MSP de Tourcoing (i.e., médiane supérieure à 14 mg/L). Les valeurs de l'hémoglobine glyquée étaient plus élevées pour les patients de la MSP de Guesnain sous antogonistes des récepteurs au GLP-1 (i.e., médiane des valeurs de 9,9 %).

5.4.3 - Forces

L'étude a été appliquée à plusieurs bases de données au format OMOP. Elle pourrait être partagée et reproduite sur d'autres bases de données implémentées dans ce format, et contenant les mêmes informations cliniques en soins premiers. En étendant les études sur plusieurs bases de données, nous avons pu comparer les résultats obtenus sur les différentes régions.
Le suivi d'une pathologie sur plusieurs années a permis de vérifier la conformité des résultats avec les recommandations nationales et d'alerter les professionnels de santé sur les points de vigilance à observer. Les données disponibles offraient la possibilité de croiser les informations et de sélectionner précisément les éléments d'intérêt, tels que les résultats de biologie sous un traitement spécifique, pour suivre leurs évolutions sur plusieurs années. Ces analyses pourraient être entendues à d'autres pathologies que le diabète.

5.4.4 - Limites

Plusieurs limites liées à la collecte des données et au suivi du patient ont entraînées une sous-estimation des résultats. Dans certains logiciels, les données textuelles et les données de biologie ont être saisies dans différents champs et de plusieurs manières selon les habitudes du MG. Lors du processus ETL, ces informations pouvaient être stockées différemment. Une compréhension des habitudes de saisies de chaque MG pourrait être nécessaire pour résoudre ce problème.
Le suivi des patients devrait également être pris en compte. Un patient suivi par un diabétologue pourrait recevoir des prescriptions en dehors de la MSP, tout comme c'est le cas lors d'une hospitalisation. Lors du croisement des informations biologiques avec les traitements, le manque d'information concernant les traitements (i.e., absence de posologie, prescriptions chez un autre spécialiste) conduisait à des examens biologiques non pris en compte.

5.4.5 - Comparaison aux autres bases de données

Tout comme pour le SNDS, la réutilisation des données de soins premiers a l'avantage de couvrir plusieurs années. Cependant, les informations sont complémentaires dans ces deux types de bases de données. Le SNDS comprend les données de facturation de tous les citoyens du territoire français [27]. Cela peut donc pallier aux problèmes de suivi chez plusieurs spécialistes. Le SNDS contient les informations de diagnostics, d'actes médicaux, de délivrances de médicaments et de consultations médicales. Des études sur l'impact des traitements sur la santé des patients (par exemple, les conséquences ou efficacités) ont été réalisées [200, 220, 221]. En revanche, contrairement aux soins premiers, le détail des valeurs des examens biologiques et biométriques, les comptes rendus ou les posologies des médicaments prescrits ne sont pas disponibles [200].

Les bases de données hospitalières n'utilisent que les données du séjour hospitalier du patient. Les données qui sont collectées vont de la prise de mesures, à chaque seconde durant une opération, au dossier médical du patient [222]. Ces bases de données sont utilisées pour analyser les actes médicaux ou des variables biologiques à différents moments de la prise en charge (i.e., avant ou après opération) [223, 224]. Les données en dehors des séjours hospitaliers ne sont pas accessibles.

Bilan et conclusion

Chapitre 6

Bilan et conclusion

La réutilisation des données de soins premiers est l'opportunité de répondre à des problématiques différentes de celles traitées avec les bases de données hospitalières, nationales ou les données issues des réseaux sociaux. Les données de soins premiers, collectées dans les cabinets ou dans les maisons de santé de ville, couvrent un suivi longitudinal du parcours de soins des patients. Ces patients sont souvent pris en charge sur plusieurs années par le même professionnel de santé, et pour des affections parfois différentes de celles traitées à l'hôpital.

Dans ce travail, nous avons étudié les spécificités de la réutilisation des données de soins premiers par rapport aux autres sources de données, et nous avons présenté les stratégies mises en place pour gérer ces spécificités. Nous avons traité les axes suivants :

1. Standardisation des données : nous avons intégré les données d'une MSP dans un EDS pour faciliter leur réutilisation (cf. Chapitre 2). L'utilisation d'un modèle de données standard, OMOP, nous a permis d'utiliser les outils et méthodes de la communauté OHDSI, tout en permettant de partager nos propres développements.

2. Intégration des données de plusieurs structures de soins : le nombre important de structures de soins premiers en France nous a amené à proposer une stratégie afin d'optimiser et d'adapter les processus ETL. Cette stratégie permet d'alimenter et de faciliter les nouvelles implémentations d'entrepôts de données (cf. Chapitre 3).

3. Développement d'outils pour aider la prise en charge des patients : en tenant compte des spécificités des soins premiers, nous avons développé des outils pour les médecins généralistes (cf. Chapitre 4). Ces outils étaient partageable et permettaient d'avoir une vue d'ensemble de l'activité clinique et du suivi longitudinal des patients.

4. Analyse des données : en nous appuyant sur les données de plusieurs MSP, standardisées au format OMOP, nous avons étudié l'évolution des résultats de biologie sous différents traitements d'antidiabétiques (cf. Chapitre 5). L'utilisation d'un modèle de données commun a facilité la conduite d'études rétrospectives à partir de quatre bases de données différentes. Ce type d'étude ne pouvait être réalisé qu'à partir des données de soins premiers, se basant sur les prescriptions médicamenteuses et sur les résultats d'analyses biologiques sur plusieurs années.

6.1 Bilan

Les données de consultations en soins premiers regroupent le motif de consultation, les symptômes, le mode de vie du patient, les antécédents médicaux, les mesures biométriques, les vaccinations et les prescriptions médicamenteuses. En dehors des consultations, les résultats des analyses biologiques, réalisées en laboratoire de ville, sont également collectées par les logiciels de soins premiers. À l'hôpital, les séjours s'étalent sur plusieurs jours et permettent une collecte quotidienne de données. En soins premiers, la collecte des données se fait lors d'une consultation de quelques minutes seulement. Dans les EDS de chacune de ces sources, les informations cliniques et démographiques sont implémentées dans plusieurs tables. Enfin, la qualité des données pouvait varier selon les habitudes de saisie des professionnels de santé. Certains professionnels consignaient toutes les informations du patient dans un seul champ du logiciel, souvent en texte libre.

Malgré ces particularités, les données de soins premiers ont été standardisées et implémentées dans un EDS au format OMOP. Initialement conçu pour la réutilisations des bases médico-administratives et la conduite d'études pharmaco-épidémiologiques, ce modèle de données standard a ensuite été étendu aux données hospitalières. Afin de pouvoir utiliser ce modèle pour les données des médecins généralistes, les concepts propres aux soins premiers ont été alignés aux concepts standards et intégrés dans les tables de vocabulaire. La consultation alimentait la table de visites du modèle OMOP, mais aussi les tables des autres domaines (i.e., les mesures biométriques, les prescriptions, les notes, les actes médicaux). La réception de résultats de biologie par le MG a été modélisée comme un passage en laboratoire d'analyses médicales. Enfin, les données textuelles ont été conservées dans une table dédiée, permettant, a posteriori, une recherche textuelle pour retrouver les informations pertinentes dans le texte (cf. objectif 1 détaillé dans la section 1.4.2 de l'introduction).

Les opportunités de réutilisation des données et les processus à mettre en oeuvre, de la collecte jusqu'à l'analyse , dépendent de chaque source (Table 6.1). À l'hôpital, de nombreux logiciels sont utilisés dans une même structure alors qu'en cabinet de ville, un seul logiciel est utilisé. Les périodes de suivis du patient diffèrent également. Les données du SNDS couvrent la totalité de la vie des patients [27, 225], les EDSH contiennent des données limitées aux séjours hospitaliers (i.e., un ou plusieurs jours). Enfin, les patients, notamment les adolescents et jeunes adultes, postent quotidiennement des informations sur les réseaux sociaux en fonction de leur activité (réseaux sociaux et forums confondus) [226, 227]. En soins premiers, les données sont collectées durant une consultation de quelques minutes et permettent le suivi du patient sur plusieurs années, à condition qu'il ne change pas de médecin traitant.
Concernant le suivi des pathologies, les données hospitalières recensent les affectations aiguës

[57, 215], les blessures et accidents graves [228] et les complications nécessitant une prise en charge complète et rapide [65, 229]. Dans nos travaux, les soins premiers étaient plus souvent représentatifs d'un suivi au long cours, généralement pour le suivi des pathologies aiguës mais bénignes (par exemple, patient souffrant d'un rhume ou d'une angine), des ALD et des pathologies chroniques.
En ville, le médecin généraliste dispose des données de décès. En effet, le médecin traitant peut attester un décès à domicile et remplir le certificat associé [230]. Lors d'un décès à l'hôpital, le médecin traitant reçoit également l'information.

	SNDS	**Hôpital**	**Soins premiers**	**Réseaux sociaux**
Nombre de logiciels en France	-	100	20	20
Nombre de logiciels par structure	-	10 à 100	1	-
Nombre de centres	1	100	2 200	-
Population couverte	10^7	10	10^4	10^9
Période couverte	vie	plusieurs jours	plusieurs années	plusieurs années
Implémentation dans un CDM	(en cours)	+++	++	-
Texte libre	-	+	++	+++
Processus de mutualisation des ETL existants	-	+	+	-
Mesures biométriques	-	+++	+++	+
Diagnostics	+++	+++	+++	+++
Actes	+++	+++	+++	-
Études de pathologies aigües	+	+++	++ (pathologies bénignes)	déclaration des difficultés rencontrées
Études de pathologies chroniques	+	++	+++	déclaration des difficultés rencontrées
Médication disponible	++ (délivrances)	++ (administrations)	++ (prescriptions + posologies)	déclaration des effets perçus
Résultats de biologie	-	++ (réalisée à l'hôpital)	+++ (réalisée en ville)	-
Décès	+++	+	++	-

TABLE 6.1 – Comparaison des informations disponibles sur chaque source de données. *- : données non applicables ; + : peu fréquent ; ++ : assez fréquent ; +++ : très fréquent.*

En France, le grand nombre de structures de soins premiers a nécessité une optimisation du développement des ETL. Pour éviter le développement systématique de nouveaux ETL à chaque intégration de données, les étapes du processus dépendantes du logiciel ont été identifiées, regroupées et placées en amont de l'ETL. Nous avons spécifié un environnement de travail

comprenant un modèle de fichiers pour l'alignement des concepts, une nomenclature d'écriture et d'archivage des scripts, ainsi qu'un ficher de configuration répertoriant toutes les variables de l'environnement et du contexte de l'utilisateur. Cette stratégie a permis de ne développer que les étapes propres au logiciel lors de l'intégration d'un nouveau logiciel, ou d'intégrer des nouvelles données sans modification de l'ETL. La généralisation des étapes indépendantes des logiciels simplifiait également le travail d'alignement sémantique. Cette stratégie n'était pas propre aux soins premiers, et pourrait être utilisée dans d'autres contextes (cf. objectif 2 détaillé dans la section 1.4.2 de l'introduction).

Une fois les données de soins premiers standardisées, des outils facilitant la réutilisation de ces données ont pu être déployés. Nous avons développé un algorithme pour détecter des séquences de traitement. Cet algorithme a permis d'évaluer les prescriptions délivrées aux patients atteints de maladies chroniques. Il calculait la durée entre deux prescriptions d'un même médicament, en intégrant un délai de tolérance afin de considérer les réserves de médicaments au domicile du patient ou les dépannages effectués par la pharmacie. Cet algorithme identifiait le manque de persévérance à un traitement (cf. objectif 3 détaillé dans la section 1.4.2 de l'introduction).
Nous avons implémenté un tableau de bord pour chaque MSP permettant un recul sur l'activité et le suivi des patients. Le tableau de bord interactif intégrait les indicateurs de la ROSP calculés à partir des données actuelles du médecin. Nos indicateurs pouvaient être comparés aux indicateurs fournis par la CNAM en même temps que la rémunération du médecin (cf. objectif 3 détaillé dans la section 1.4.2 de l'introduction).

L'homogénéisation du format des données de soins premiers nous a donné la possibilité de répliquer des analyses sur plusieurs MSP. Dans un premier temps, nous avons croisé les résultats des analyses biologiques avec les traitements prescrits, dans une MSP. Les valeurs de l'hémoglobine glyquée sous différentes familles d'antidiabétiques ont été analysées. Dans un second temps, nous avons étendu cette étude à plusieurs structures afin de les comparer. À partir de notre algorithme de détection de séquences de traitements, nous avons identifié les résultats de biologie propres à chacune de ces séquences. Les valeurs de l'hémoglobine glyquée étaient différentes sur les MSP et identiques selon les familles de traitements d'antidiabétiques. Ces analyses pourraient être partagées avec la communauté de chercheurs utilisant le modèle OMOP. Ces études ont été réalisées sur les données de patients diabétiques mais pourraient également être répliquées sur d'autres pathologies chroniques (cf. objectif 4 détaillé dans la section 1.4.2 de l'introduction).

6.2 Difficultés rencontrées

Lors du développement de chacun des axes présentés dans cette thèse, nous avons rencontré plusieurs difficultés. Tout d'abord, la majorité des informations médicales importantes en soins premiers étaient consignées en texte libre. De plus, certaines données pouvaient être saisies dans différents champs du logiciel. La quantité et la qualité des données textuelles variaient considérablement en fonction des pratiques des professionnels de santé, rendant l'analyse plus complexe que celle des données structurées.
Ensuite, les données extraites des logiciels de soins premiers concernaient un temps précis de la prise en charge du patient (i.e., la consultation), et ne fournissaient pas d'informations sur le suivi en dehors du cabinet. En particulier, les informations de suivi chez un spécialiste et les hospitalisations n'étaient pas collectées. Les comptes rendus envoyés aux MG ne comprenaient pas le détail des prescriptions effectuées. Ainsi, l'analyse des séquences de traitements anti-diabétiques prescrits par le médecin généraliste ne tenait pas compte des prescriptions que le patient pouvait recevoir de son diabétologue. Nous avons calculé les séquences de traitements à partir des données de prescriptions du médecin, sans disposer d'informations sur la délivrance en pharmacie ou la consommation réelle des médicaments par les patients.
Par ailleurs, pour le calcul des indicateurs de la ROSP, nous nous sommes basés également sur les actions des MG. Les prescriptions faites au patient à l'oral n'étaient pas prises en compte, car ces données n'avaient pas été collectées. Ce calcul différait de celui réalisé par la CNAM, qui s'appuie sur les données de facturation des actes réalisés ou des prescriptions délivrées. La CNAM se base sur les réponses du patient, alors que notre calcul portait sur les actions du MG. Cependant, si un patient ne suivait pas les prescriptions malgré les recommandations du médecin, la responsabilité en incombait au médecin par sa sensibilisation insuffisante.

6.3 Perspectives

Tout d'abord, il serait envisageable d'étendre l'intégration des données de soins premiers à d'autres MSP et à d'autres professions, telles que les pharmaciens, les infirmières ou les sages-femmes. Ces nouveaux périmètres de données apporteraient une vue globale de la prise en charge du patient par tous les professionnels de la MSP.
En complément des EDS de soins premiers, l'utilisation du SNDS permettrait de suivre le parcours du patient chez l'ensemble des spécialistes qu'il consulte. Cela offrirait aussi la possibilité d'évaluer la délivrance effective des médicaments prescrits par le MG en pharmacie d'officine.
Les données de soins premiers pourraient être enrichies par les informations extraites des textes libres grâce à des techniques de NLP. Ces techniques seraient employées pour identifier des éléments comme les symptômes, les diagnostics ou les prescriptions d'actes qui seraient intégrés sous forme structurée dans le modèle.

La collaboration avec l'éditeur du logiciel WEDA pourrait faciliter l'intégration de notre tableau de bord, directement dans le logiciel de soins. Une analyse qualitative, réalisée avec les MG, permettrait d'évaluer ce tableau de bord et son utilisation en routine. L'algorithme de détection de séquences de traitement pourrait être intégré à un système d'alerte dans le tableau de bord. En cas d'interruption d'un traitement, une alerte informerait le MG de la durée pendant laquelle le patient ne serait plus couvert par son traitement. De plus, des alertes pourraient être émises en cas de prescriptions inappropriées, grâce à un ensemble de règles appliquées sur les prescriptions.

6.4 Conclusion

Bien que la réutilisation des données soit déjà effective au niveau hospitalier, les données issues des soins premiers restent peu exploitées. Contrairement aux autres sources de données, les soins premiers couvrent un suivi longitudinal du parcours de soins en intégrant les résultats de biologie, les données de consultations chez le médecin généraliste et les prescriptions médicamenteuses. Nous avons traité plusieurs axes pour faciliter la réutilisation de ces données : l'utilisation d'un modèle de données commun pour faciliter le partage des outils et des analyses, l'optimisation des processus d'implémentation des entrepôts de données à grande échelle, le développement d'outils et la réalisation d'études à partir des données de plusieurs MSP.
Dans cette continuité, les données de soins premiers, limitées aux interventions en cabinet, offrent la possibilité de répondre de manière privilégiée aux questions de recherche portant sur l'évolution des valeurs biologiques sur plusieurs années, en lien avec la prise de traitements. Elles constituent également une source permettant le suivi des maladies bénignes.
Des avancées techniques sont également à prévoir pour enrichir ces données. L'implémentation de méthodes NLP, pour extraire les informations du texte, complétera l'entrepôt de données. Une intégration d'un système d'alertes sur la détection de prescriptions inappropriées permettra aux médecins de mieux suivre leurs patients, améliorant ainsi la qualité des soins.

Bibliographie

[1] SCHOEN, C. « A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas » (2012). DOI : 10.1377/hlthaff.2012.0884.

[2] SANTÉ PUBLIQUE, C. de la. *LOI n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé (1)*. In : (2016). URL : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000031912641.

[3] CNIL. *Données de santé : la CNIL rappelle les mesures de sécurité et de confidentialité pour l'accès au dossier patient informatisé (DPI)*. 2024. URL : https://www.cnil.fr/fr/donnees-de-sante-la-cnil-rappelle-les-mesures-de-securite-et-de-confidentialite-pour-lacces-au.

[4] NUMÉRIQUE EN SANTÉ, A. A. du. *Le Dispositif DPI du Couloir Hôpital du Ségur du numérique en santé | Portail Industriels*. URL : https://industriels.esante.gouv.fr/segur-numerique-sante/vague-1/dispositif-dpi-couloir-hopital.

[5] TORKI, A. « Impact du Dossier Patient Informatisé sur la qualité des soins. L'expérience d'un centre hospitalier au Luxembourg » (2022), p. 57-79. DOI : 10.3917/proj.hs03.0057.

[6] CNAM. *Le DMP en pratique*. 2024. URL : https://www.ameli.fr/medecin/sante-prevention/dmp-et-mon-espace-sante/dossier-medical-partage/dmp-en-pratique.

[7] CNIL. *Donnée sensible*. URL : https://www.cnil.fr/fr/definition/donnee-sensible.

[8] NUMÉRIQUE EN SANTÉ, A. A. du. *HDS Certification Hébergeur de Données de Santé*. Agence du Numérique en Santé. URL : https://esante.gouv.fr/produits-services/hds.

[9] CNIL. *Sécurité : Gérer la sous-traitance*. 2024. URL : https://www.cnil.fr/fr/securite-gerer-la-sous-traitance.

[10] CNIL. *Le règlement général sur la protection des données - RGPD*. 2016. URL : https://www.cnil.fr/fr/reglement-europeen-protection-donnees.

[11] CNIL. *RGPD : de quoi parle-t-on ?* 2018. URL : https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on.

[12] MINISTÈRE DE L'ÉCONOMIE, d. f. e. d. l. s. i. e. n. *Le règlement général sur la protection des données (RGPD), mode d'emploi*. 2023. URL : https://www.economie.gouv.fr/entreprises/reglement-general-protection-donnees-rgpd.

[13] CNIL. *Mission 1 - Informer, protéger les droits*. 2024. URL : https://www.cnil.fr/fr/missions/mission-1-informer-proteger-les-droits.

[14] MOSKOLAI, J. et al. « Intégration et interopérabilité des systèmes d'information hétérogènes dans des environnements distribués : vers une approche flexible basée sur l'urbanisation des systèmes d'information ». In : *Conférence de Recherche Internationale en Informatique*. 2013.

[15] SANTÉ, H. H. A. de. *Les logiciels métier des professionnels de santé*. Haute Autorité de Santé. 2016. URL : https://www.has-sante.fr/jcms/r_1506258/fr/les-logiciels-metier-des-professionnels-de-sante.

[16] SANTÉ, H. H. A. de. *Entrepôts de données de santé hospitaliers en France*. Haute Autorité de Santé. 2022. URL : https://www.has-sante.fr/jcms/p_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france.

[17] ZAPLETAL, E. et al. « Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case ». *Stud Health Technol Inform* 160 (2010), p. 193-197.

[18] JANNOT, A.-S. et al. « The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience ». *Int J Med Inform* 102 (2017), p. 21-28. DOI : 10.1016/j.ijmedinf.2017.02.006.

[19] ARTEMOVA, S. et al. « PREDIMED: Clinical Data Warehouse of Grenoble Alpes University Hospital ». *Stud Health Technol Inform* 264 (2019). DOI : 10.3233/SHTI190464.

[20] PALACIOS-FERNANDEZ, S. et al. « Time trends in hospital discharges in patients aged 85years and older in Spain: data from the Spanish National Discharge Database (2000–2015) ». *BMC Geriatrics* 21 (2021), p. 371. DOI : 10.1186/s12877-021-02335-2.

[21] DHALLUIN, T. et al. « Comparison of Unplanned 30-Day Readmission Prediction Models, Based on Hospital Warehouse and Demographic Data ». *Stud Health Technol Inform* 270 (2020), p. 547-551. DOI : 10.3233/SHTI200220.

[22] HUANG, Y. et al. « Machine learning methods to predict 30-day hospital readmission outcome among US adults with pneumonia: analysis of the national readmission database ». *BMC Med Inform Decis Mak* 22 (2022), p. 288. DOI : 10.1186/s12911-022-01995-3.

[23] WASSEL, C. L. et al. « Risk of readmissions, mortality, and hospital-acquired conditions across hospital-acquired pressure injury (HAPI) stages in a US National Hospital Discharge database ». *International Wound Journal* 17.6 (2020). DOI : 10.1111/iwj.13482.

[24] DIRECTION DE LA RECHERCHE des études, d. l. e. d. s. *Qu'est-ce que le SNDS ? | SNDS*. URL : https://www.snds.gouv.fr/SNDS/Qu-est-ce-que-le-SNDS.

[25] J, B. et al. « The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology ». *Pharmacoepidemiology and drug safety* 26.8 (2017). DOI : 10.1002/pds.4233.

[26] DIRECTION DE LA RECHERCHE des études, d. l. e. d. s. *Composantes du SNDS | SNDS*. URL : https://www.snds.gouv.fr/SNDS/Composantes-du-SNDS.

[27] SCAILTEUX, L.-M. et al. « French administrative health care database (SNDS): The value of its enrichment ». *Therapie* 74.2 (2019), p. 215-223. DOI : 10.1016/j.therap.2018.09.072.

[28] YU, Y. et al. « Integrating real-world data to assess cardiac ablation device outcomes in a multicenter study using the OMOP common data model for regulatory decisions: implementation and evaluation ». *JAMIA Open* 6.1 (2023). DOI : 10.1093/jamiaopen/ooac108.

[29] CUGGIA, M. et al. « The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare ». *Yearb Med Inform* 28.1 (2019), p. 195-202. DOI : 10.1055/s-0039-1677917.

[30] HELSEDIREKTORATET. *KUHR-databasen*. Helsedirektoratet. 2006. URL : https://www.helsedirektoratet.no/tema/statistikk-registre-og-rapporter/helsedata-og-helseregistre/kuhr.

[31] SOCIALSTYRELSEN.SE. *Statistikdatabaser - Dödsorsaksstatistik - Val*. 2024. URL : https://sdb.socialstyrelsen.se/if_dor/val_eng.aspx.

[32] NOKLUS. *Norsk diabetesregister for voksne*. 2006. URL : https://www.noklus.no/norsk-diabetesregister-for-voksne/.

[33] SOCIALSTYRELSEN.SE. *Statistikdatabaser - Cancerstatistik - Val*. 2024. URL : https://sdb.socialstyrelsen.se/if_can/val_eng.aspx.

[34] HEALTH {AND} WELFARE, F. I. for. *Finnish National Infectious Diseases Register - THL*. Finnish Institute for Health and Welfare (THL), Finland. 2023. URL : https://thl.fi/en/topics/infectious-diseases-and-vaccinations/surveillance-and-registers/finnish-national-infectious-diseases-register.

[35] MEKARU, S. et al. « One Health in social networks and social media ». *Rev Sci Tech* 33.2 (2014), p. 629-637.

[36] KEMPNY, C. et al. « Web scraping applications in health services research: For web experts only, or a tool for every health services researcher? » *Z Evid Fortbild Qual Gesundhwes* 176 (2023), p. 61-64. DOI : 10.1016/j.zefq.2022.11.010.

[37] FRUCHART, M. et al. « Identification of early symptoms of endometriosis through the analysis of online social networks: A social media study ». *Digit Health* 9 (2023). DOI : 10.1177/20552076231176114.

[38] OSADCHIY, V. et al. « Low Testosterone on Social Media: Application of Natural Language Processing to Understand Patients' Perceptions of Hypogonadism and Its Treatment ». *J Med Internet Res* 22.10 (2020), e21383. DOI : 10.2196/21383.

[39] ZHANG, L. et al. « Utilizing Twitter data for analysis of chemotherapy ». *International Journal of Medical Informatics* 120 (2018), p. 92-100. DOI : 10.1016/j.ijmedinf.2018.10.002.

[40] NGUYEN, A. X.-L. et al. « Determination of Patient Sentiment and Emotion in Ophthalmology: Infoveillance Tutorial on Web-Based Health Forum Discussions ». *J Med Internet Res* 23.5 (2021). DOI : 10.2196/20803.

[41] SARKER, A. et al. « Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource ». *J Am Med Inform Assoc* 27.8 (2020), p. 1310-1315. DOI : 10.1093/jamia/ocaa116.

[42] KRITTANAWONG, C. et al. « Insights from Twitter about novel COVID-19 symptoms ». *European Heart Journal - Digital Health* 1.1 (2020), p. 4-5. DOI : 10.1093/ehjdh/ztaa003.

[43] OBEIDAT, R. et al. « Multi-label multi-class COVID-19 Arabic Twitter dataset with fine-grained misinformation and situational information annotations ». *PeerJ Comput Sci* 8 (2022). DOI : 10.7717/peerj-cs.1151.

[44] MARTY, T. et al. « Patients' testimonies, feelings, complaints and emotional experiences with dermatoses on open social media: The French infodemiologic patient's free speech study ». *J Eur Acad Dermatol Venereol* 38.7 (2024). DOI : 10.1111/jdv.19781.

[45] HENGARTNER, M. P. et al. « Protracted withdrawal syndrome after stopping antidepressants: a descriptive quantitative analysis of consumer narratives from a large internet forum ». *Ther Adv Psychopharmacol* 10 (2020). DOI : 10.1177/2045125320980573.

[46] OYEBODE, O. et al. « Identifying adverse drug reactions from patient reviews on social media using natural language processing ». *Health Informatics J* 29.1 (2023). DOI : 10.1177/14604582221136712.

[47] SCHÜCK, S. et al. « Assessing Patient Perceptions and Experiences of Paracetamol in France: Infodemiology Study Using Social Media Data Mining ». *J Med Internet Res* 23.7 (2021). DOI : 10.2196/25049.

[48] ABD-ALRAZAQ, A. et al. « Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study ». *J Med Internet Res* 22.4 (2020). DOI : 10.2196/19016.

[49] AÏMEUR, E. et al. « Fake news, disinformation and misinformation in social media: a review ». *Soc Netw Anal Min* 13.1 (2023), p. 30. DOI : 10.1007/s13278-023-01028-5.

[50] SARKER, A. et al. « Utilizing social media data for pharmacovigilance: A review ». *Journal of Biomedical Informatics* 54 (2015), p. 202-212. DOI : 10.1016/j.jbi.2015.02.004.

[51] KANG, S. Y. et al. « Comprehensive risk factor evaluation of postoperative delirium following major surgery: clinical data warehouse analysis ». *Neurol Sci* 40.4 (2019), p. 793-800. DOI : 10.1007/s10072-019-3730-1.

[52] BOUZILLÉ, G. et al. « Drug safety and big clinical data: Detection of drug-induced anaphylactic shock events ». *J Eval Clin Pract* 24.3 (2018), p. 536-544. DOI : 10.1111/jep.12908.

[53] KIM, H. S. et al. « Characteristics of Early Pancreatic Cancer: Comparison between Stage 1A and Stage 1B Pancreatic Cancer in Multicenter Clinical Data Warehouse Study ». *Cancers (Basel)* 16.5 (2024), p. 944. DOI : 10.3390/cancers16050944.

[54] LEGGAT-BARR, K. et al. « Early Ascertainment of Breast Cancer Diagnoses Comparing Self-Reported Questionnaires and Electronic Health Record Data Warehouse: The WISDOM Study ». *JCO clinical cancer informatics* 7 (2023). DOI : 10.1200/CCI.23.00019.

[55] SENEVIRATNE, M. G. et al. « Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer ». *EGEMS (Wash DC)* 6.1 (2018), p. 13. DOI : 10.5334/egems.234.

[56] PROFILI, F. et al. « Prevalence and clustering of chronic diseases in Tuscany (Central Italy): evidence from a large administrative data warehouse ». *Epidemiol Prev* 44.5 (2020), p. 385-393. DOI : 10.19191/EP20.5-6.P385.014.

[57] TRIEP, K. et al. « Real-world Health Data and Precision for the Diagnosis of Acute Kidney Injury, Acute-on-Chronic Kidney Disease, and Chronic Kidney Disease: Observational Study ». *JMIR Med Inform* 10.1 (2022). DOI : 10.2196/31356.

[58] BREAULT, J. L. et al. « Data mining a diabetic data warehouse ». *Artif Intell Med* 26.1 (2002), p. 37-54. DOI : 10.1016/s0933-3657(02)00051-9.

[59] JO, K. et al. « Long-term prediction models for vision-threatening diabetic retinopathy using medical features from data warehouse ». *Sci Rep* 12.1 (2022). DOI : 10.1038/s41598-022-12369-0.

[60] WILEY, K. K. et al. « Quantifying Electronic Health Record Data Quality in Telehealth and Office-Based Diabetes Care ». *Appl Clin Inform* 13.5 (2022). DOI : 10.1055/s-0042-1758737.

[61] DELANEROLLE, G. et al. « Methodological Issues in Using a Common Data Model of COVID-19 Vaccine Uptake and Important Adverse Events of Interest: Feasibility Study of Data and Connectivity COVID-19 Vaccines Pharmacovigilance in the United Kingdom ». *JMIR Form Res* 6.8 (2022). DOI : 10.2196/37821.

[62] GROSJEAN, J. et al. « Using Clinical Data Warehouse to Optimize the Vaccination Strategy Against COVID-19: A Use Case in France ». *Stud Health Technol Inform* 290 (2022). DOI : 10.3233/SHTI220050.

[63] MOLTO, A. et al. « Evaluation of the prevalence of new-onset musculoskeletal symptoms in patients hospitalized for severe SARS-CoV-2 infection during the first two COVID waves in France: A descriptive analysis of the clinical data warehouse of 39 hospitals in France ». *Joint Bone Spine* 89.6 (2022). DOI : 10.1016/j.jbspin.2022.105450.

[64] FLEUREN, L. M. et al. « The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients ». *Crit Care* 25.1 (2021), p. 304. DOI : 10.1186/s13054-021-03733-z.

[65] XU, Y. et al. « Associations of Apixaban Dose With Safety and Effectiveness Outcomes in Patients With Atrial Fibrillation and Severe Chronic Kidney Disease ». *Circulation* 148.19 (2023). DOI : 10.1161/CIRCULATIONAHA.123.065614.

[66] DENG, Y. et al. « Comparative effectiveness of second line glucose lowering drug treatments using real world data: emulation of a target trial ». *BMJ Med* 2.1 (2023). DOI : 10.1136/bmjmed-2022-000419.

[67] FEYMAN, Y. et al. « Effect of mental health staffing inputs on suicide-related events ». *Health Serv Res* 58.2 (2023), p. 375-382. DOI : 10.1111/1475-6773.14064.

[68] KITCHEN, C. et al. « Suicide Death Prediction Using the Maryland Suicide Data Warehouse: A Sensitivity Analysis ». *Arch Suicide Res* (2024), p. 1-15. DOI : 10.1080/13811118.2024.2363227.

[69] KWON, E. et al. « A Clinical Data Warehouse Analysis of Risk Factors for Inpatient Falls in a Tertiary Hospital: A Case-Control Study ». *J Patient Saf* 19.8 (2023), p. 501-507. DOI : 10.1097/PTS.0000000000001163.

[70] GEORGIOU, A. et al. « COVID-19: protocol for observational studies utilizing near real-time electronic Australian general practice data to promote effective care and best-practice policy-a design thinking approach ». *Health Res Policy Syst* 19.1 (2021), p. 122. DOI : 10.1186/s12961-021-00772-4.

[71] HOMAYOUNI, H. et al. « Anomaly Detection in COVID-19 Time-Series Data ». *SN Comput Sci* 2.4 (2021). DOI : 10.1007/s42979-021-00658-w.

[72] SUCHARD, M. A. et al. « Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis ». *Lancet* 394 (2019). DOI : 10.1016/S0140-6736(19)32317-7.

[73] YARAHUAN, J. K. et al. « Design, Usability, and Acceptability of a Needs-Based, Automated Dashboard to Provide Individualized Patient-Care Data to Pediatric Residents ». *Applied Clinical Informatics* 13.2 (2022), p. 380-390. DOI : 10.1055/s-0042-1744388.

[74] HUBER, T. C. et al. « Developing an Interactive Data Visualization Tool to Assess the Impact of Decision Support on Clinical Operations ». *J Digit Imaging* 31.5 (2018), p. 640-645. DOI : 10.1007/s10278-018-0065-z.

[75] LAMER, A. et al. « Transforming Anesthesia Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study ». *J Med Internet Res* 23.10 (2021). DOI : 10.2196/29259.

[76] ANSARI, B. et al. « Development of a usability checklist for public health dashboards to identify violations of usability principles ». *Journal of the American Medical Informatics Association: JAMIA* 29.11 (2022), p. 1847-1858. DOI : 10.1093/jamia/ocac140.

[77] DANCIU, I. et al. « Secondary use of clinical data: the Vanderbilt approach ». *J Biomed Inform* 52 (2014), p. 28-35. DOI : 10.1016/j.jbi.2014.02.003.

[78] COOREVITS, P. et al. « Electronic health records: new opportunities for clinical research ». *J Intern Med* 274.6 (2013), p. 547-560. DOI : 10.1111/joim.12119.

[79] WEISKOPF, N. G. et al. « Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research ». *J Am Med Inform Assoc* 20.1 (2013), p. 144-151. DOI : 10.1136/amiajnl-2011-000681.

[80] TERRY, A. L. et al. « A basic model for assessing primary health care electronic medical record data quality ». *BMC Med Inform Decis Mak* 19.1 (2019), p. 30. DOI : 10.1186/s12911-019-0740-0.

[81] SAFRAN, C. « Reuse of clinical data ». *Yearb Med Inform* 9 (2014), p. 52-54. DOI : 10.15265/IY-2014-0013.

[82] PASCOE, S. W. et al. « Identifying patients with a cancer diagnosis using general practice medical records and Cancer Registry data ». *Fam Pract* 25.4 (2008), p. 215-220. DOI : 10.1093/fampra/cmn023.

[83] ALMEIDA, J. R. et al. « A two-stage workflow to extract and harmonize drug mentions from clinical notes into observational databases ». *Journal of Biomedical Informatics* 120 (2021), p. 103849. DOI : 10.1016/j.jbi.2021.103849.

[84] CHENG, K. Y. et al. « ETL Processes for Integrating Healthcare Data - Tools and Architecture Patterns ». *Stud Health Technol Inform* 299 (2022), p. 151-156. DOI : 10.3233/SHTI220974.

[85] PECORARO, F. et al. « Designing ETL Tools to Feed a Data Warehouse Based on Electronic Healthcare Record Infrastructure ». *Stud Health Technol Inform* 210 (2015), p. 929-933.

[86] ONG, T. C. et al. « Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading ». *BMC Med Inform Decis Mak* 17 (2017), p. 134. DOI : 10.1186/s12911-017-0532-3.

[87] LACROIX-HUGUES, V. et al. « Creation of the First French Database in Primary Care Using the ICPC2: Feasibility Study ». *Stud Health Technol Inform* 245 (2017), p. 462-466.

[88] MISHRA, S. et al. « Structured and Unstructured Big Data Analytics ». In : 2017, p. 740-746. DOI : 10.1109/CTCEEC.2017.8454999.

[89] ZHANG, D. et al. « Combining structured and unstructured data for predictive models: a deep learning approach ». *BMC Med Inform Decis Mak* 20.1 (2020), p. 280. DOI : 10.1186/s12911-020-01297-6.

[90] KELOTH, V. K. et al. « Representing and utilizing clinical textual data for real world studies: An OHDSI approach ». *Journal of Biomedical Informatics* 142 (2023), p. 104343. DOI : 10.1016/j.jbi.2023.104343.

[91] BHATTACHARJEE, T. et al. « INSPIRE datahub: a pan-African integrated suite of services for harmonising longitudinal population health data using OHDSI tools ». *Front Digit Health* 6 (2024). DOI : 10.3389/fdgth.2024.1329630.

[92] FRUCHART, M. et al. « Transforming Primary Care Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study ». *JMIR Med Inform* 12 (2024). DOI : 10.2196/49542.

[93] HRIPCSAK, G. et al. « Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers ». *Stud Health Technol Inform* 216 (2015), p. 574-578.

[94] OHDSI. *OHDSI – Observational Health Data Sciences and Informatics*. URL : https://www.ohdsi.org/.

[95] OHDSI. *OMOP CDM v5.4*. URL : https://ohdsi.github.io/CommonDataModel/cdm54.html.

[96] OHDSI. *Athena*. URL : https://athena.ohdsi.org/search-terms/start.

[97] OHDSI. *Rabbit in a Hat*. URL : http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html.

[98] OHDSI. *White Rabbit*. URL : http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html.

[99] OHDSI. *ATLAS*. URL : https://atlas-demo.ohdsi.org/.

[100] OHDSI. *Open-Source Tutorials – OHDSI*. 2023. URL : https://www.ohdsi.org/open-source-tutorials/.

[101] YOO, S. *Transforming Thyroid Cancer Diagnosis and Staging Information from Unstructured Reports to the Observational Medical Outcome Partnership Common Data Model - PubMed*. 2022. URL : https://pubmed.ncbi.nlm.nih.gov/35705182/.

[102] PARIS, N. et al. « Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study ». *JMIR Med Inform* 9.12 (2021). DOI : 10.2196/30970.

[103] KIM, G. L. et al. « The Risk of Osteoporosis and Osteoporotic Fracture Following the Use of Irritable Bowel Syndrome Medical Treatment: An Analysis Using the OMOP CDM Database ». *Journal of Clinical Medicine* 10.9 (2021). DOI : 10.3390/jcm10092044.

[104] BYUN, J. et al. « Analysis of treatment pattern of anti-dementia medications in newly diagnosed Alzheimer's dementia using OMOP CDM ». *Scientific Reports* 12 (2022). DOI : 10.1038/s41598-022-08595-1.

[105] ZHANG, X. et al. « Analysis of treatment pathways for three chronic diseases using OMOP CDM ». *Journal of Medical Systems* 42.12 (2018), p. 260. DOI : 10.1007/s10916-018-1076-5.

[106] CHO, S. et al. « Content Coverage Evaluation of the OMOP Vocabulary on the Transplant Domain Focusing on Concepts Relevant for Kidney Transplant Outcomes Analysis ». *Applied Clinical Informatics* 11.4 (2020), p. 650-658. DOI : 10.1055/s-0040-1716528.

[107] JEON, S. et al. « Proposal and Assessment of a De-Identification Strategy to Enhance Anonymity of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) in a Public Cloud-Computing Environment: Anonymization of Medical Data Using Privacy Models ». *Journal of Medical Internet Research* (2020). DOI : 10.2196/19597.

[108] GLICKSBERG, B. S. et al. « PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model ». *Bioinformatics (Oxford, England)* (2019). DOI : 10.1093/bioinformatics/btz409.

[109] RAVENTÓS, B. et al. « IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model ». *Pharmacoepidemiology and Drug Safety* 33.1 (2024). DOI : 10.1002/pds.5717.

[110] SANTÉ PUBLIQUE, C. de la. *Article L1411-11 - Code de la santé publique*. URL : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000031930722.

[111] BOURGUEIL, Y. et al. « Qu'appelle-t-on « soins primaires » ? » In : *Les soins primaires en question(s)*. Débats Santé Social. 2021, p. 5-13.

[112] AFONSO, M. M. et al. « Les soins primaires : une définition du champ pour développer la recherche ». *Epidemiology and Public Health = Revue d'Epidémiologie et de Santé Publique* 66.2 (2018), p. 157-162. DOI : 10.1016/j.respe.2017.09.004.

[113] ORGANIZATION, W. W. H. *Soins de santé primaires*. 2023. URL : https://www.who.int/fr/news-room/fact-sheets/detail/primary-health-care.

[114] *Déclaration d'Alma-Ata*. URL : https://www.who.int/fr/publications/i/item/WHO-EURO-1978-3938-43697-61471.

[115] COMMUNICATION, D. à l'information et à la. *Stratégie nationale de santé*. Ministère de la santé et de l'accès aux soins. 2019.

[116] ORGANIZATION, W. W. H. *Health Systems Strengthening Glossary*. In : URL : https://cdn.who.int/media/docs/default-source/documents/health-systems-strengthening-glossary.pdf.

[117] MEDECINS, C. N. D. L. D. *Atlas de la démographie médicale en France*. 2022.

[118] LA DREES (DIRECTION DE LA RECHERCHE des études, d. l. e. d. s. *Démographie des professionnels de santé : Qui sont les médecins en 2018 ? Quelle accessibilité aux médecins généralistes ? Combien d'infirmiers en 2040 ? Un outil de projections d'effectifs de médecins*. 2020. URL : https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-08/dossier_presse_demographie.pdf.

[119] GÉNÉRALE, S. E. de médecine. *La définition européenne de la médecine générale – médecine de famille*. In : 2005. URL : https://conseil25.ordre.medecin.fr/sites/default/files/domain-562/1/wonka_-_mg.pdf.

[120] CNAM. *Dispositif du médecin traitant*. 2024. URL : https://www.ameli.fr/medecin/exercice-liberal/facturation-remuneration/dispositif-medecin-traitant/dispositif-medecin-traitant.

[121] ADMINISTRATIVE, D. de l'information légale et. *Médecin traitant et parcours de soins coordonnés*. 2024. URL : https://www.service-public.fr/particuliers/vosdroits/F163.

[122] SANTÉ PUBLIQUE, C. de la. *Article 27 - Arrêté du 20 octobre 2016 portant approbation de la convention nationale organisant les rapports entre les médecins libéraux et l'assurance maladie signée le 25 août 2016 - Légifrance*. 2016. URL : https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037439142?isAdvancedResult=&page=3&pageSize=10&query=ROSP&searchField=ALL&searchProximity=&searchType=ALL&tab_selection=all&typePagination=DEFAULT.

[123] CNAM. *La Rosp*. 2023. URL : https://www.ameli.fr/medecin/exercice-liberal/facturation-remuneration/remuneration-objectifs/nouvelle-rosp.

[124] CNAM. *GUIDE METHODOLOGIQUE Rémunération sur Objectifs de Santé Publique (ROSP) MEDECIN TRAITANT DE L'ADULTE*. 2023. URL : https://www.ameli.fr/sites/default/files/Documents/Guide%20m%C3%A9thodologique-ROSP%202023-M%C3%A9decin%20traitant%20de%20l%27adulte.pdf.

[125] CNAM. *Rosp et forfait structure 2023 : une rémunération moyenne en augmentation de 2,5%*. 2024. URL : https://www.ameli.fr/medecin/actualites/rosp-et-forfait-structure-2023-une-remuneration-moyenne-en-augmentation-de-25.

[126] SANTÉ PUBLIQUE, C. de la. *Article L1111-15 - Code de la santé publique*. URL : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000043908072.

[127] ARS. *Les Maisons de santé pluriprofessionnelles (MSP)*. 2023. URL : https://www.hauts-de-france.ars.sante.fr/les-maisons-de-sante-pluriprofessionnelles-msp.

[128] CNAM. *Constitution d'une maison de santé pluriprofessionnelle (MSP)*. 2023. URL : https://www.ameli.fr/exercice-coordonne/exercice-professionnel/organisations-d-exercice-coordonne/constitution-d-une-maison-de-sante-pluriprofessionnelle-msp.

[129] DGOS et al. *Les maisons de santé*. Ministère du travail, de la santé et des solidarités. 2024. URL : https://sante.gouv.fr/systeme-de-sante/structures-de-soins/article/les-maisons-de-sante-300889.

[130] ANS. *Label e-santé*. Agence du Numérique en Santé. URL : http://esante.gouv.fr/produits-services/label-e-sante.

[131] ANS. *Les solutions labellisées e-santé*. Agence du Numérique en Santé. 2024. URL : https://esante.gouv.fr/offres-services/label-esante/solutions-labellisees.

[132] ARS. *Le Ségur de la santé*. 2023. URL : https://www.auvergne-rhone-alpes.ars.sante.fr/le-segur-de-la-sante.

[133] DGOS. *Ségur Usage Numérique en Établissements de Santé*. Ministère du travail, de la santé et des solidarités. 2024. URL : https://sante.gouv.fr/systeme-de-sante/segur-de-la-sante/sun-es.

[134] ANS. *Le Ségur du numérique en santé*. Agence du Numérique en Santé. 2024. URL : https://esante.gouv.fr/segur.

[135] ANS. *Le Ségur du numérique en santé pour les médecins de ville*. Agence du Numérique en Santé. 2024. URL : https://esante.gouv.fr/segur/medecin-de-ville.

[136] CAREY, I. M. et al. « Prevalence and incidence of neuromuscular conditions in the UK between 2000 and 2019: A retrospective study using primary care data ». *PLoS One* 16.12 (2021). DOI : 10.1371/journal.pone.0261983.

[137] MENÉNDEZ TORRE, E. L. et al. « Prevalence of diabetes mellitus in Spain in 2016 according to the Primary Care Clinical Database (BDCAP) ». *Endocrinol Diabetes Nutr (Engl Ed)* 68.2 (2021), p. 109-115. DOI : 10.1016/j.endinu.2019.12.004.

[138] BOULLENGER, L. et al. « Type 2 diabetics followed up by family physicians: Treatment sequences and changes over time in weight and glycated hemoglobin ». *Prim Care Diabetes* 16.5 (2022), p. 670-676. DOI : 10.1016/j.pcd.2022.07.002.

[139] LOADSMAN, M. E. N. et al. « Impetigo incidence and treatment: a retrospective study of Dutch routine primary care data ». *Fam Pract* 36.4 (2019), p. 410-416. DOI : 10.1093/fampra/cmy104.

[140] MARWAHA, S. et al. « Comanagement of Rashes by Primary Care Providers and Dermatologists: A Retrospective Study ». *Perm J* 25 (2021). DOI : 10.7812/TPP/20.320.

[141] MILEA, D. et al. « Long-Acting Bronchodilator Use in Chronic Obstructive Pulmonary Disease in Primary Care in New Zealand: A Retrospective Study of Treatment Patterns and Evolution Using the HealthStat Database ». *Int J Chron Obstruct Pulmon Dis* 16 (2021). DOI : 10.2147/COPD.S290887.

[142] SOLLIE, A. et al. « Reusability of coded data in the primary care electronic medical record: A dynamic cohort study concerning cancer diagnoses ». *Int J Med Inform* 99 (2017), p. 45-52. DOI : 10.1016/j.ijmedinf.2016.08.004.

[143] KORNFÄLT ISBERG, H. et al. « Increased adherence to treatment guidelines in patients with urinary tract infection in primary care: A retrospective study ». *PLoS One* 14.3 (2019). DOI : 10.1371/journal.pone.0214572.

[144] FRUCHART, M. et al. « Implementation of a Data Warehouse in Primary Care: First Analyses with Elderly Patients ». *Stud Health Technol Inform* 294 (2022), p. 505-509. DOI : 10.3233/SHTI220510.

[145] REGULATORY AGENCY, T. M.
bibinitperiod H. products. *Clinical Practice Research Datalink | CPRD*. URL : https://www.cprd.com/node/120.

[146] S.A.S., G. *Healthcare Data Research | THIN Data*. URL : https://www.the-health-improvement-network.com.

[147] VETERANS AFFAIRS, U. D. of. *Corporate Data Warehouse (CDW)*. 2023. URL : https://www.hsrd.research.va.gov/for_researchers/cdw.cfm.

[148] NETWORK, C. P. C. S. S. *Welcome to CPCSSN*. Canadian Primary Care Sentinel Surveillance Network. URL : https://cpcssn.ca/.

[149] PAPEZ, V. et al. « Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond ». *J Am Med Inform Assoc* 30.1 (2022), p. 103-111. DOI : 10.1093/jamia/ocac203.

[150] WARD, R. et al. « The OMOP common data model in Australian primary care data: Building a quality research ready harmonised dataset ». *PLoS One* 19.4 (2024). DOI : 10.1371/journal.pone.0301557.

[151] PRECIDIAB. *Projet PriCaDa : constitution des premiers Entrepôts de Données de Santé en ambulatoire | Precidiab*. URL : https://www.precidiab.org/actualite/projet-pricada-constitution-des-premiers-entrepots-de-donnees-de-sante-en-ambulatoire/.

[152] WEDA. *2024 - Vidal Expert*. Weda. 2022. URL : https://weda.fr/vidalexpert.

[153] W3SCHOOLS. *XML Introduction*. URL : https://www.w3schools.com/xml/xml_whatis.asp.

[154] OHDSI. *documentation:vocabulary [Observational Health Data Sciences and Informatics]*. URL : https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary.

[155] COHEN, J. « A Coefficient of Agreement for Nominal Scales ». *Educational and Psychological Measurement* 20.1 (1960). Publisher: SAGE Publications Inc, p. 37-46. DOI : 10.1177/001316446002000104.

[156] OHDSI. *Creates Descriptive Statistics Summary for an Entire OMOP CDM Instance*. URL : https://ohdsi.github.io/Achilles/.

[157] KAHN, M. G. et al. « A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data ». *eGEMs* 4.1 (2016). DOI : 10.13063/2327-9214.1244.

[158] HU, C. et al. « Data driven identification of international cutting edge science and technologies using SpaCy ». *PLoS One* 17.10 (2022). DOI : 10.1371/journal.pone.0275872.

[159] BHASURAN, B. et al. « Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases ». *Journal of Biomedical Informatics* 64 (2016), p. 1-9. DOI : 10.1016/j.jbi.2016.09.009.

[160] CORP, X. *Twitter API Documentation*. URL : https://developer.x.com/en/docs/x-api.

[161] *beautifulsoup4: Screen-scraping library*. Version 4.12.3. URL : https://www.crummy.com/software/BeautifulSoup/bs4/.

[162] *selenium: Official Python bindings for Selenium WebDriver*. Version 4.24.0. URL : https://www.selenium.dev.

[163] HUB, H. D. *Transformer le SNDS au format OMOP : Contexte | Documentation du SNDS & SNDS OMOP*. 2023. URL : https://documentation-snds.health-data-hub.fr/omop/introduction/snds_omop.html.

[164] KALIYAR, R. K. et al. « FakeBERT: Fake news detection in social media with a BERT-based deep learning approach ». *Multimed Tools Appl* 80.8 (2021). DOI : 10.1007/s11042-020-10183-2.

[165] WANG, T. « COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model ». *IEEE Access* 8 (2020). DOI : 10.1109/ACCESS.2020.3012595.

[166] BAE, Y. J. et al. « Schizophrenia Detection Using Machine Learning Approach from Social Media Content ». *Sensors (Basel)* 21.17 (2021), p. 5924. DOI : 10.3390/s21175924.

[167] (DICOM), D. à l'information et à la communication. *Les maisons de santé*. Ministère de la Santé et de la Prévention. 2023. URL : https://sante.gouv.fr/systeme-de-sante/structures-de-soins/article/les-maisons-de-sante-300889.

[168] *MSP Corneille*. URL : https://www.mspcorneille.com/.

[169] WEDA. *Weda : Votre Logiciel Médical en Ligne.* Weda. URL : https://weda.fr/.

[170] SANTE, C. *Crossway - Le logiciel de gestion de cabinet pour médecins*. URL : https://www.cegedim-sante.com/solutions-sante-cegedim/solutions-local/crossway/.

[171] FRUCHART, M. *Mathilde Fruchart · GitLab*. GitLab. 2023. URL : https://gitlab.com/mathilde.frchrt.

[172] ONG, T. et al. « A Framework for Classification of Electronic Health Data Extraction-Transformation-Loading Challenges in Data Network Participation ». *EGEMS (Washington, DC)* 5.1 (2017), p. 10. DOI : 10.5334/egems.222.

[173] OHDSI. *The Book of OHDSI*. URL : https://ohdsi.github.io/TheBookOfOhdsi/.

[174] PECORARO, F. et al. « A conceptual framework to design a dimensional model based on the HL7 Clinical Document Architecture ». *Stud Health Technol Inform* 205 (2014), p. 278-282.

[175] MIYOSHI, N. S. B. et al. « Computational framework to support integration of biomolecular and clinical data within a translational approach ». *BMC Bioinformatics* 14 (2013), p. 180. DOI : 10.1186/1471-2105-14-180.

[176] PEDRERA, M. et al. « Making EHRs Reusable: A Common Framework of Data Operations ». *Stud Health Technol Inform* 287 (2021), p. 129-133. DOI : 10.3233/SHTI210831.

[177] QUIROZ, J. C. et al. « Extract, transform, load framework for the conversion of health databases to OMOP ». *PLoS One* 17.4 (2022). DOI : 10.1371/journal.pone.0266911.

[178] ABAD-NAVARRO, F. et al. « A knowledge graph-based data harmonization framework for secondary data reuse ». *Comput Methods Programs Biomed* 243 (2023). DOI : 10.1016/j.cmpb.2023.107918.

[179] LI, K. et al. « Time to lack of persistence with pharmacological treatment among patients with current depressive episodes: a natural study with 1-year follow-up ». *Patient Prefer Adherence* 10 (2016). DOI : 10.2147/PPA.S109941.

[180] GUPTA, P. et al. « How to Screen for Non-Adherence to Antihypertensive Therapy ». *Curr Hypertens Rep* 18.12 (2016), p. 89. DOI : 10.1007/s11906-016-0697-7.

[181] MESSALI, A. J. « Treatment persistence and switching in triptan users: a systematic literature review » (2014). DOI : 10.1111/head.12404.

[182] SOLOMON, M. D. et al. « Primary non-adherence of medications: lifting the veil on prescription-filling behaviors ». *J Gen Intern Med* 25.4 (2010), p. 280-281. DOI : 10.1007/s11606-010-1286-0.

[183] GUERCI, B. et al. « Lack of Treatment Persistence and Treatment Nonadherence as Barriers to Glycaemic Control in Patients with Type 2 Diabetes ». *Diabetes Ther* 10.2 (2019), p. 437-449. DOI : 10.1007/s13300-019-0590-x.

[184] HARBIG, P. et al. « Instantaneous detection of nonadherence: quality, strength, and weakness of an electronic prescription database ». *Pharmacoepidemiol Drug Saf* 21.3 (2012), p. 323-328. DOI : 10.1002/pds.2351.

[185] LAURENT, G. et al. « Development, implementation and preliminary evaluation of clinical dashboards in a department of anesthesia ». *Journal of Clinical Monitoring and Computing* 35.3 (2021), p. 617-626. DOI : 10.1007/s10877-020-00522-x.

[186] LINDER, J. A. et al. « Electronic health record feedback to improve antibiotic prescribing for acute respiratory infections ». *Am J Manag Care* 16.12 (2010).

[187] PAUWELS, K. et al. « Dashboards as a Service: Why, What, How, and What Research Is Needed? » *Journal of Service Research* 12.2 (2009), p. 175-189. DOI : 10.1177/1094670509344213.

[188] RABIEI, R. et al. « Requirements and challenges of hospital dashboards: a systematic literature review ». *BMC Med Inform Decis Mak* 22 (2022), p. 287. DOI : 10.1186/s12911-022-02037-8.

[189] DAGLIATI, A. et al. « A dashboard-based system for supporting diabetes care ». *J Am Med Inform Assoc* 25.5 (2018). DOI : 10.1093/jamia/ocx159.

[190] KARAMI, M. et al. « From Information Management to Information Visualization ». *Appl Clin Inform* 7.2 (2016), p. 308-329. DOI : 10.4338/ACI-2015-08-RA-0104.

[191] MORGAN, M. B. et al. « The radiology digital dashboard: effects on report turnaround time ». *J Digit Imaging* 21.1 (2008), p. 50-58. DOI : 10.1007/s10278-007-9008-9.

[192] KOOPMAN, R. J. et al. « A diabetes dashboard and physician efficiency and accuracy in accessing data needed for high-quality diabetes care ». *Ann Fam Med* 9.5 (2011), p. 398-405. DOI : 10.1370/afm.1286.

[193] AYDIN, C. E. et al. « Beyond Nursing Quality Measurement: The Nation's First Regional Nursing Virtual Dashboard ». In : *Advances in Patient Safety: New Directions and Alternative Approaches (Vol. 1: Assessment)*. Sous la dir. d'HENRIKSEN, K. et al. Advances in Patient Safety. Rockville (MD) : Agency for Healthcare Research et Quality, 2008.

[194] BLAIS, R. et al. « TOCSIN: a proposed dashboard of indicators to control healthcare-associated infections ». *Healthc Q* 12 Spec No Patient (2009), p. 161-167. DOI : 10.12927/hcq.2009.20985.

[195] GHAZISAEIDI, M. et al. « Development of Performance Dashboards in Healthcare Sector: Key Practical Issues ». *Acta Inform Med* 23.5 (2015), p. 317-321. DOI : 10.5455/aim.2015.23.317-321.

[196] KROCH, E. et al. « Hospital Boards and Quality Dashboards ». *Journal of Patient Safety* 2.1 (2006), p. 10.

[197] CIC-IT. *CIC-IT de Lille*. CIC-IT Lille. URL : https://cic-it-lille.com/.

[198] MARCILLY, R. et al. « Usability Checklists for Health Technology: Case Study and Experts' Opinions ». *Studies in Health Technology and Informatics* 316 (2024), p. 1074-1078. DOI : 10.3233/SHTI240596.

[199] LOPUT, C. M. et al. « Evaluation of medication administration timing variance using information from a large health system's clinical data warehouse ». *Am J Health Syst Pharm* 79 (2022). DOI : 10.1093/ajhp/zxab378.

[200] BASTARD, L. et al. « Risk of serious infection associated with different classes of targeted therapies used in psoriatic arthritis: a nationwide cohort study from the French Health Insurance Database (SNDS) ». *RMD Open* 10.1 (2024). DOI : 10.1136/rmdopen-2023-003865.

[201] DONNET, A. et al. « Migraine burden and costs in France: a nationwide claims database analysis of triptan users ». *J Med Econ* 22.7 (2019), p. 616-624. DOI : 10.1080/13696998.2019.1590841.

[202] BURKHARDT, S. et al. « Towards identifying drug side effects from social media using active learning and crowd sourcing ». *Pac Symp Biocomput* 25 (2020), p. 319-330.

[203] LARDON, J. et al. « Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review ». *J Med Internet Res* 17.7 (2015). DOI : 10.2196/jmir.4304.

[204] YANG, Q. et al. « Do we behave differently on Twitter and Facebook: Multi-view social network user personality profiling for content recommendation ». *Front Big Data* 5 (2022). DOI : 10.3389/fdata.2022.931206.

[205] MINISTÈRE DU TRAVAIL, d. l. s. e. d. s. *Diabète*. Ministère du travail, de la santé et des solidarités. 2024. URL : https://sante.gouv.fr/soins-et-maladies/maladies/article/diabete.

[206] SPF. *Bulletin épidémiologique hebdomadaire, 8 novembre 2022*. 2022. URL : https://www.santepubliquefrance.fr/import/bulletin-epidemiologique-hebdomadaire-8-novembre-2022-n-22-journee-mondiale-du-diabete-14-novembre-2022.

[207] ORGANIZATION, W. W. H. *Diabetes*. 2021.

[208] LAGE, M. J. et al. « The relationship between HbA1c reduction and healthcare costs among patients with type 2 diabetes: evidence from a U.S. claims database ». *Curr Med Res Opin* 36.9 (2020). DOI : 10.1080/03007995.2020.1787971.

[209] HAS. *Guide parcours de soins Diabète de type 2 de l'adulte*. Haute Autorité de Santé. 2014. URL : https://www.has-sante.fr/jcms/c_1735060/fr/guide-parcours-de-soins-diabete-de-type-2-de-l-adulte.

[210] FRUCHART, M. et al. « Description of a French Population of Diabetics Treated Followed up by General Practitioners ». *Stud Health Technol Inform* 302 (2023), p. 856-860. DOI : 10.3233/SHTI230289.

[211] ZAFAR, S. et al. « Systemic Adverse Events Among Patients With Diabetes Treated With Intravitreal Anti-Vascular Endothelial Growth Factor Injections ». *JAMA Ophthalmol* 141.7 (2023), p. 658-666. DOI : 10.1001/jamaophthalmol.2023.2098.

[212] SHRIVASTAV, M. et al. « Type 2 Diabetes Management in Primary Care: The Role of Retrospective, Professional Continuous Glucose Monitoring ». *Diabetes Spectrum : A Publication of the American Diabetes Association* 31.3 (2018), p. 279-287. DOI : 10.2337/ds17-0024.

[213] AJROUCHE, S. et al. « HbA1c changes in a deprived population who followed or not a diabetes self-management programme, organised in a multi-professional primary care practice: a historical cohort study on 207 patients between 2017 and 2019 ». *BMC Endocrine Disorders* 24 (2024), p. 72. DOI : 10.1186/s12902-024-01601-9.

[214] GEORGESCU, V. et al. « Primary care visits can reduce the risk of potentially avoidable hospitalizations among persons with diabetes in France ». *European Journal of Public Health* 30.6 (2020), p. 1056-1061. DOI : 10.1093/eurpub/ckaa137.

[215] ROUSSEL, R. et al. « Important Drop in Rate of Acute Diabetes Complications in People With Type 1 or Type 2 Diabetes After Initiation of Flash Glucose Monitoring in France: The RELIEF Study ». *Diabetes Care* 44.6 (2021), p. 1368-1376. DOI : 10.2337/dc20-1690.

[216] FFD. *Norme HbA1c | Hémoglobine Glyquée ou HbA1c | Taux de Bba1c*. 2010. URL : https://www.federationdesdiabetiques.org/information/glycemie/hba1c.

[217] INSERM. *Diabète de type 2 · Inserm, La science pour la santé*. Inserm. 2019. URL : https://www.inserm.fr/dossier/diabete-type-2/.

[218] CNAM. *Diabète de type 2 : la metformine en points clés*. 2024. URL : https://www.ameli.fr/medecin/sante-prevention/pathologies/diabete-type-2/memo-metformine.

[219] *VIDAL : Base de données médicamenteuse pour les prescipteurs libéraux*. VIDAL. URL : https://www.vidal.fr/.

[220] GRYNBERG, M. et al. « Comparative effectiveness of gonadotropins used for ovarian stimulation during assisted reproductive technologies (ART) in France: A real-world observational study from the French nationwide claims database (SNDS) ». *Best Pract Res Clin Obstet Gynaecol* 88 (2023). DOI : 10.1016/j.bpobgyn.2022.102308.

[221] BEZIN, J. et al. « GLP-1 Receptor Agonists and the Risk of Thyroid Cancer ». *Diabetes Care* 46.2 (2023), p. 384-390. DOI : 10.2337/dc22-1148.

[222] CHEN, X. et al. « Patient-Patient Similarity-Based Screening of a Clinical Data Warehouse to Support Ciliopathy Diagnosis ». *Front Pharmacol* 13 (2022). DOI : 10.3389/fphar.2022.786710.

[223] YANG, J.-S. et al. « The Influence of High Blood Pressure on Developing Symptomatic Lumbar Epidural Hematoma after Posterior Lumbar Spinal Fusion Surgery: Clinical Data Warehouse Analysis ». *J Clin Med* 11.15 (2022). DOI : 10.3390/jcm11154522.

[224] KWON, Y.-S. et al. « The Relationship between Perioperative Blood Transfusion and Postoperative Delirium in Patients Undergoing Spinal Fusion Surgery: Clinical Data Warehouse Analysis ». *Medicina (Kaunas)* 58.2 (2022), p. 268. DOI : 10.3390/medicina58020268.

[225] TUPPIN, P. et al. « Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France ». *Revue d'Épidémiologie et de Santé Publique* 65 (2017). DOI : 10.1016/j.respe.2017.05.004.

[226] AIELLO, A. E. et al. « Social media- and internet-based disease surveillance for public health ». *Annual review of public health* 41 (2020), p. 101-118. DOI : 10.1146/annurev-publhealth-040119-094402.

[227] ZAVALA, J. et al. « The Impact of Social Media Use for News on Academic Performance in Underrepresented Undergraduate College Students ». *Cyberpsychology, Behavior and Social Networking* 26.8 (2023), p. 657-661. DOI : 10.1089/cyber.2022.0303.

[228] LOMBARDI, L. R. et al. « Improving identification of crash injuries: Statewide integration of hospital discharge and crash report data ». *Traffic Inj Prev* 23 (sup1 2022). DOI : 10.1080/15389588.2022.2083612.

[229] POURCHER, V. et al. « Outcomes of coronavirus disease 2019-related hospitalization among people with HIV: historical cohort from the Greater Paris area multicenter hospital data warehouse ». *AIDS* 37.12 (2023). DOI : 10.1097/QAD.0000000000003655.

[230] LADEVÈZE, M. et al. « Le médecin généraliste et la mort de ses patients ». 41 (2010), p. 65-72. DOI : 10.3917/pos.411.0065. URL : https://shs.cairn.info/revue-pratiques-et-organisation-des-soins-2010-1-page-65?lang=fr&tab=resume.

Annexes

Chapitre A

Évaluation de la qualité des données dans la base de données

Requêtes	Code SQL	Résultats	Vérifications
Nombre de patients adultes (âgés de 18 ans ou plus) par médecin de famille vus pendant la période du 1er janvier au 7 janvier 2021.	`SELECT count(distinct(vo.person_id)) AS number_patient FROM omop.visit_occurrence vo LEFT JOIN omop.person p ON vo.person_id = p.person_id WHERE vo.visit_start_date >= DATE '01/01/2021' AND vo.visit_end_date <= DATE '07/01/2021' AND (extract(year from vo.visit_start_date) - p.year_of_birth) = 18;`	Dr A : 2 patients Dr B : 2 patients Dr C : 2 patients Dr D : 2 patients	Retrait des patients déclarés auprès du médecin de famille après la date de l'extraction
Nombre de patients adultes (âgés de 18 ans ou plus) vus par médecin de famille en 2020	`SELECT DISTINCT vo.person_id, p.year_of_birth, p2.provider_name FROM omop.visit_occurrence vo LEFT JOIN omop.person p ON vo.person_id = p.person_id LEFT JOIN omop.provider p2 ON p.provider_id = p2.provider_id WHERE (extract(year FROM vo.visit_start_date) - p.year_of_birth) >= 18 AND (extract(year FROM vo.visit_start_date) - p.year_of_birth) <= 20 AND vo.visit_start_date >= DATE '01/01/2020' AND vo.visit_end_date <= DATE '31/12/2020';`	Dr E : 8 patients Dr B : 12 patients Dr F : 10 patients Dr C : 4 patients Dr D : 46 patients Dr A : 34 patients	Les patients n'ayant eu aucune consultation en 2020, les patients dont le dossier a été créé après la date d'extraction des données et les patients déclarés auprès d'un médecin généraliste après la date d'extraction ont été retirés des résultats

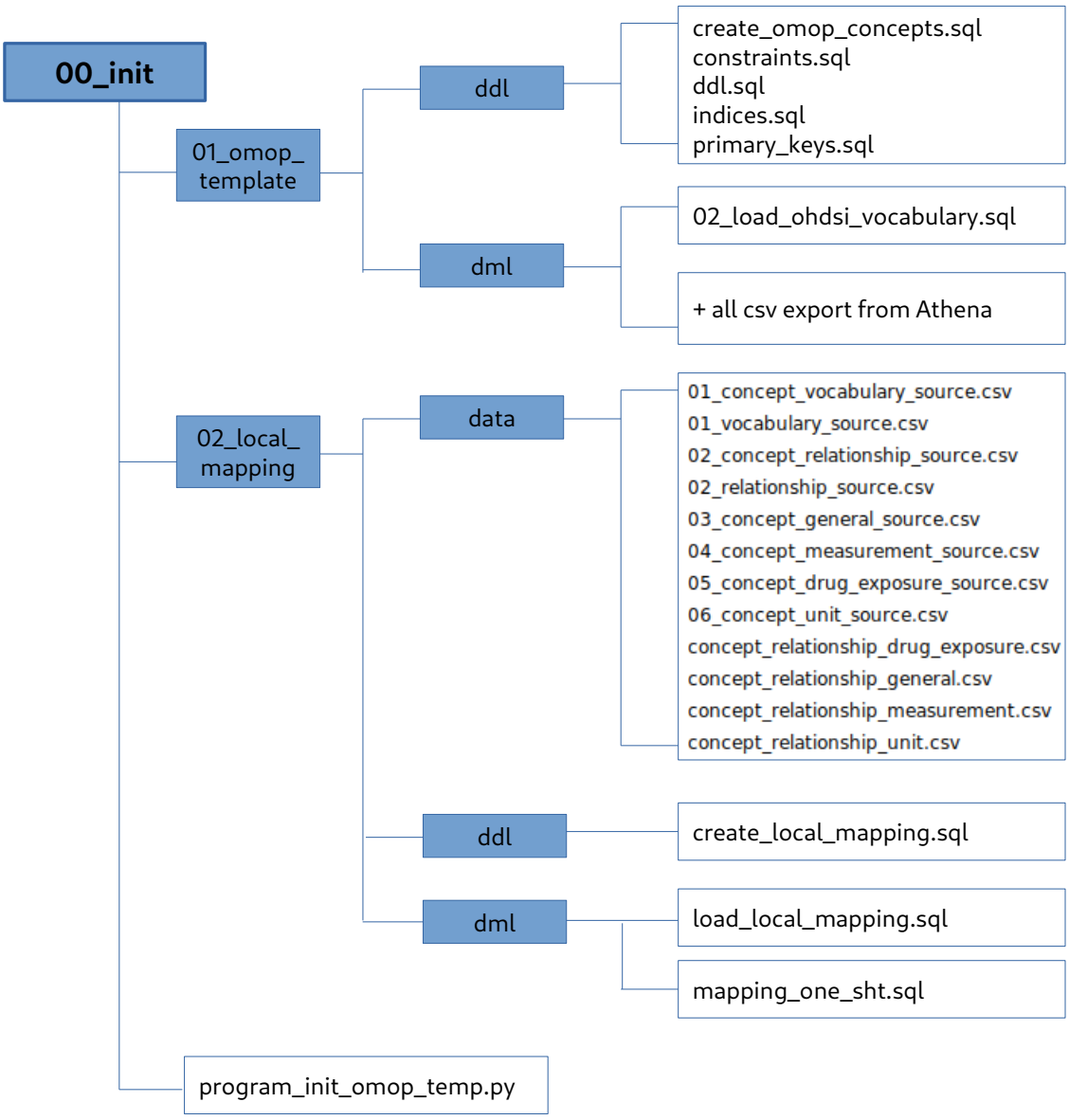
Nombre de patients âgés de 75 ans et plus traités avec Zopiclone (ATC : N05CF01) du 1er janvier au 31 janvier 2021	`SELECT p.gender_concept_id AS sex, count(distinct(de.person_id)) AS number_patient FROM omop.drug_exposure de LEFT JOIN omop.person p ON de.person_id = p.person_id LEFT JOIN omop.concept_relationship r ON de.drug_concept_id = r.concept_id_1 LEFT JOIN omop.concept c ON r.concept_id_2 = c.concept_id WHERE c.concept_code = 'N05CF01' AND c.vocabulary_id = 'ATC' AND de.drug_exposure_start_date >= DATE '01/01/2021' AND de.drug_exposure_start_date <= DATE '31/01/2021' AND (extract(year FROM current_date) - p.year_of_birth) >= 75 GROUP BY p.gender_concept_id;`	9 patientes 4 patients	
Nombre de patients âgés de 75 ans et plus traités avec Zopiclone (ATC : N05CF01) du 1er janvier 2020 au 31 janvier 2021	`SELECT p.gender_concept_id AS sex, count(distinct(de.person_id)) AS number_patient FROM omop.drug_exposure de LEFT JOIN omop.person p ON de.person_id = p.person_id LEFT JOIN omop.concept_relationship r ON de.drug_concept_id = r.concept_id_1 LEFT JOIN omop.concept c ON r.concept_id_2 = c.concept_id WHERE c.concept_code = 'N05CF01' AND c.vocabulary_id = 'ATC' AND de.drug_exposure_start_date >= DATE '01/01/2021' AND de.drug_exposure_start_date <= DATE '31/01/2021' AND (extract(year FROM current_date) - p.year_of_birth) >= 75 GROUP BY p.gender_concept_id;`	24 patientes 9 patients	
Nombre de patients âgés entre 18 et 25 ans ayant eu un résultat de test de laboratoire entre le 1er janvier et le 7 janvier 2021	`SELECT DISTINCT p.person_id FROM omop.measurement m LEFT JOIN omop.person p ON m.person_id = p.person_id WHERE m.measurement_source_concept_id = 2000000005 AND m.measurement_date >= DATE '01/01/2021' AND m.measurement_date <= DATE '07/01/2021' AND (extract(year FROM current_date) - p.year_of_birth) >= 18 AND (extract(year FROM current_date) - p.year_of_birth) <= 25;`	Logiciel : 14 patients Entrepôt de données : 9 patients	Deux résultats de tests de laboratoire n'ont pas été classés au bon endroit dans le logiciel (dans les rapports médicaux) et trois concepts de biologie n'ont pas été alignés dans l'entrepôt de données

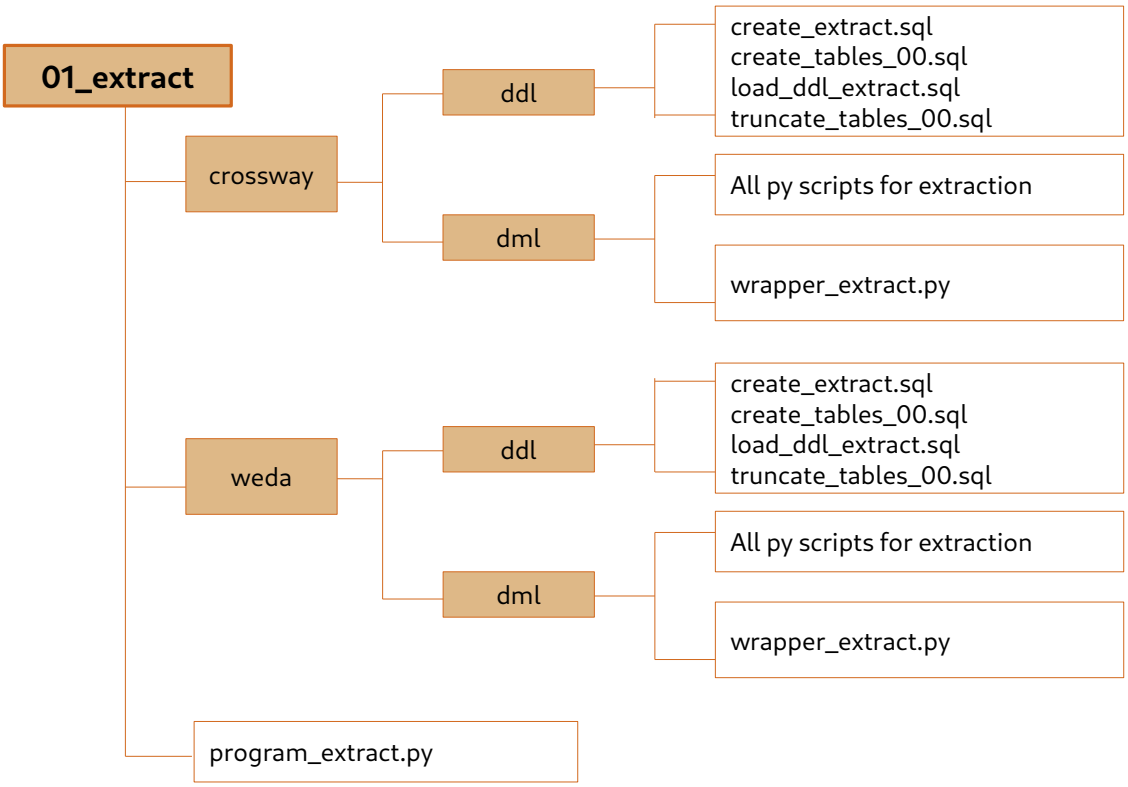
TABLE A.1 – Liste des requêtes pour évaluer la qualité des données dans la base de données.

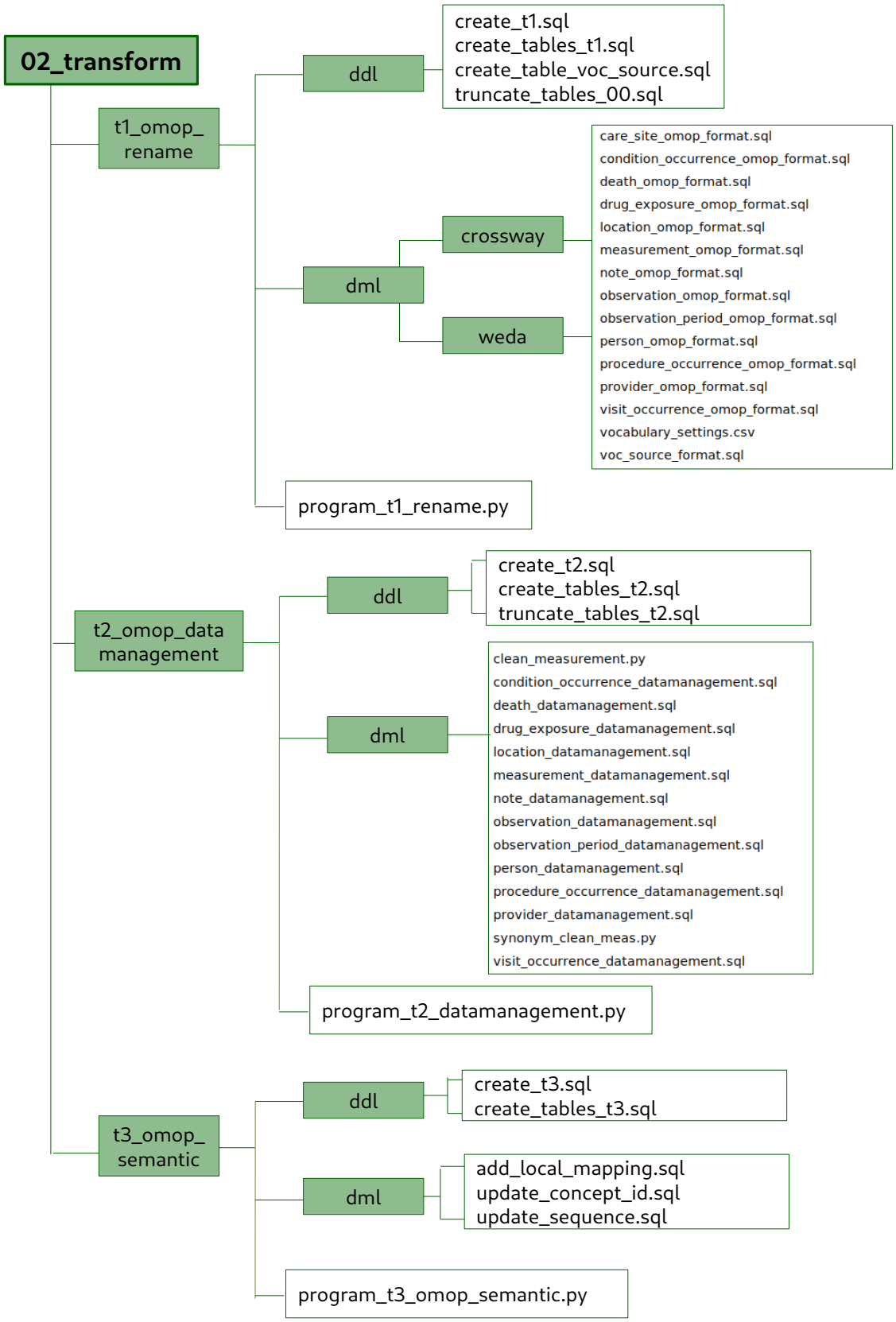
Chapitre B

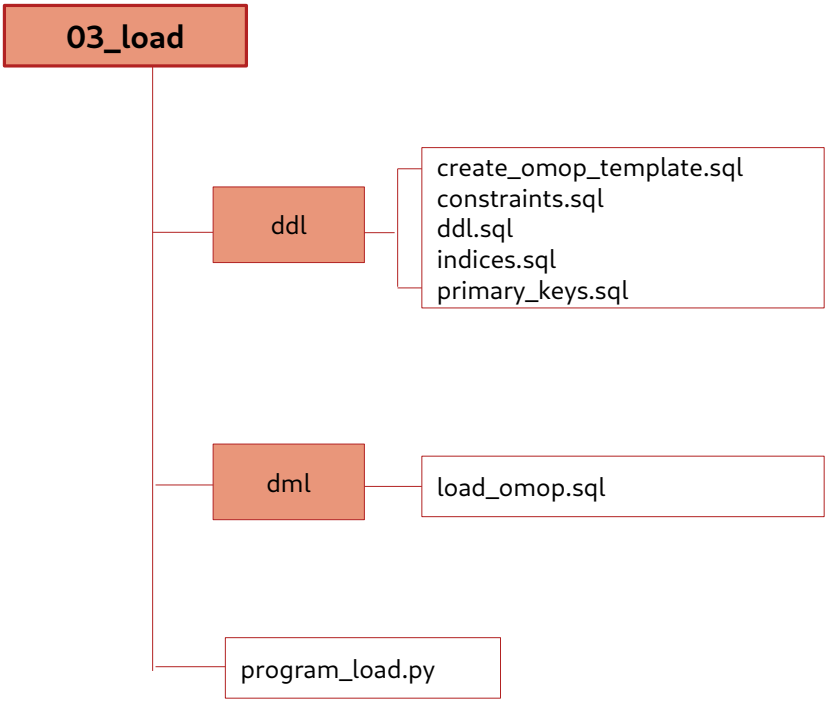
Nomenclature des dossiers et des fichiers de l'ETL optimisé

Chaque étape de l'ETL impliquait la création d'un schéma. Les scripts de chaque étape étaient contenus dans un dossier. Les sous-dossiers étaient divisés en deux : un pour les fichiers DDL (i.e., scripts de création des schémas et structures de tables) et un autre pour les fichiers DML (i.e., scripts de transformation des données).


00_init
01_omop_ template
ddl
create_omop_concepts.sql
constraints.sql
ddl.sql
indices.sql
primary_keys.sql
dml
02_load_ohdsi_vocabulary.sql
+ all csv export from Athena
02_local_ mapping
data
01_concept_vocabulary_source.csv
01_vocabulary_source.csv
02_concept_relationship_source.csv
02_relationship_source.csv
03_concept_general_source.csv
04_concept_measurement_source.csv
05_concept_drug_exposure_source.csv
06_concept_unit_source.csv
concept_relationship_drug_exposure.csv
concept_relationship_general.csv
concept_relationship_measurement.csv
concept_relationship_unit.csv
ddl
create_local_mapping.sql
dml
load_local_mapping.sql
mapping_one_sht.sql
program_init_omop_temp.py


01_extract
crossway
ddl
create_extract.sql
create_tables_00.sql
load_ddl_extract.sql
truncate_tables_00.sql
dml
All py scripts for extraction
wrapper_extract.py
weda
ddl
create_extract.sql
create_tables_00.sql
load_ddl_extract.sql
truncate_tables_00.sql
dml
All py scripts for extraction
wrapper_extract.py
program_extract.py


02_transform
t1_omop_rename
ddl
create_t1.sql
create_tables_t1.sql
create_table_voc_source.sql
truncate_tables_00.sql
dml
crossway
care_site_omop_format.sql
condition_occurrence_omop_format.sql
death_omop_format.sql
drug_exposure_omop_format.sql
location_omop_format.sql
measurement_omop_format.sql
note_omop_format.sql
observation_omop_format.sql
observation_period_omop_format.sql
person_omop_format.sql
procedure_occurrence_omop_format.sql
provider_omop_format.sql
visit_occurrence_omop_format.sql
vocabulary_settings.csv
voc_source_format.sql
weda
program_t1_rename.py
t2_omop_data management
ddl
create_t2.sql
create_tables_t2.sql
truncate_tables_t2.sql
dml
clean_measurement.py
condition_occurrence_datamanagement.sql
death_datamanagement.sql
drug_exposure_datamanagement.sql
location_datamanagement.sql
measurement_datamanagement.sql
note_datamanagement.sql
observation_datamanagement.sql
observation_period_datamanagement.sql
person_datamanagement.sql
procedure_occurrence_datamanagement.sql
provider_datamanagement.sql
synonym_clean_meas.py
visit_occurrence_datamanagement.sql
program_t2_datamanagement.py
t3_omop_semantic
ddl
create_t3.sql
create_tables_t3.sql
dml
add_local_mapping.sql
update_concept_id.sql
update_sequence.sql
program_t3_omop_semantic.py


03_load
ddl
create_omop_template.sql
constraints.sql
ddl.sql
indices.sql
primary_keys.sql
dml
load_omop.sql
program_load.py

Chapitre C

Liste des opérations à suivre pour chaque étape du développement d'ETL

Initiation

(i1) Le modèle ETL "one-shot" est existant et fonctionnel.
(i2) Les étapes sont présentées de manière claire et distincte.
(i3) Les outils et le langage choisis sont décrits.
(i4) Les sources de données, les logiciels intégrés et les types de données sont décrits.
(i5) La documentation sur les outils et les technologies est présentée.
(i6) La nomenclature et la hiérarchie des dossiers et des fichiers sont présentées.
(i7) Un fichier Excel nommé vocabulary_settings.csv est créé et liste les différents domaines de données (concept_id : nom court du domaine ; concept_name : nom complet du domaine ; vocabulary_reference ; vocabulary_version ; id : auto-incrémentation ; concept_vocabulary_id : vide).

Organisation

(o1) Les étapes de l'ETL (*extract*, *T1*,*T2* ,*T3* et *load*) sont respectées sur la base des points suivants. Certains points sont déjà créés si le logiciel source est le même que l'ETL de la phase d'initiation ; il s'agit simplement de vérifier si ces points sont respectés ou de les ajouter si la source est différente. La nomenclature des fichiers et des dossiers ajoutés doit être conforme à celles présentées dans la phase d'initiation (i6) :

- *Extract* (créer un dossier avec le nom du logiciel source s'il est différent des ETL existants et y ajouter des scripts permettant d'exécuter les points suivants) :
  - Sélectionner les variables suivantes à partir des données brutes : (1) données démographiques du patient (i.e., années de naissance, identifiant, date de création du dossier ou période de suivi, origine, sexe) ; (2) données cliniques du patient (i.e., antécédents médicaux, nom du médecin traitant) ; (3) données de consultation (i.e., date de consultation, observations, tests effectués, mesures biométriques prises, nom du médecin, diagnostic, symptômes, correspondance, notes) ; (4) données de visite hors cabinet (i.e., résultats de biologie, laboratoires, dates) ; (5) comptes rendus d'examens hors cabinet.
  - Développer un programme Python pour automatiser l'extraction de ces variables

et les stocker dans une base de données relationnelles PostGres, en respectant la nomenclature des données brutes.

— *T1* (créer un dossier avec le nom du logiciel source s'il est différent des ETL existants et y ajouter des scripts qui permettent l'exécution des points suivants) :

  — Utiliser Rabbit-in-a-hat et WhiteRabbit pour faire correspondre les variables sources aux variables correspondantes du modèle OMOP.

  — Renommer les tables et les colonnes selon la nomenclature OMOP et placer les variables dans les tables correctes.

  — Créer la table "voc_source" en intégrant les données du point (i7) dans la phase d'initiation.

— T2 (ajouter ces points aux scripts existants si la source est différente) :

  — Sauvegarder les identifiants de la source (issus des exports du logiciel) dans une table temporaire et attribuer un nouvel identifiant pour *person_id* et *visit_occurrence _id*. Ces identifiants sont mis à jour dans chaque table. Pour les tables restantes, un identifiant unique est créé.

  — Attribuer le *table_type_concept_id* à chaque domaine (à l'aide de l'outil Athena).

  — Implémenter le *table_source_concept_id* en fonction de la table de vocabulaire créée précédemment.

  — Changer les dates logicielles par défaut (par exemple, '0001-01-01') en NULL. Ajouter une date de fin xx_end_date équivalente à la date de début (xx_start_date) si elle est manquante.

  — Supprimer les lignes sans information (valeur nulle pour observation_source_value, condition_source_value, drug_source_value, address1 + address2 + city + state + zip + country, value_as_number + measurement_source_value, note_text, observation_period_start_date ou observation_period_end_date, person_id, procedure_source_value + procedure_source_date).

  — Supprimer les doublons. Pour la table visit_occurrence, il est nécessaire de regrouper les lignes par visit_occurrence_id, person_id et visit_start_date pour éliminer les données en double.

  — Nettoyer le texte des concepts de tests de laboratoire. Créer une liste de synonymes (i.e., concepts similaires à regrouper sous le même terme). Créer un algorithme basé sur des expressions régulières pour remplacer les concepts similaires, nettoyer le texte (i.e., supprimer les symboles, la ponctuation, les accents).

— Les données sont formatées correctement (i.e., date pour les variables liées à la date, numérique, texte, etc.)

— *T3* : Deux possibilités, soit de nouveaux concepts sont alignés et doivent être intégrés, soit aucun nouveau concept ne doit être ajouté (le processus de mise en relation est donc ignoré). Si un alignement n'existe pas, il est effectué manuellement sur un fichier modèle standardisé, le chemin vers le fichier de correspondance est notifié dans le fichier de configuration et le processus de relations ainsi que plusieurs vérifications doivent être intégrés dans le modèle.

(o2) La documentation des fichiers est présente et claire.
(o3) Le code source est testé avant d'être déployé. Sur GitLab, chaque morceau de code est documenté et testé avant d'être fusionné dans la branche "dev".
(o4) Le fichier de configuration pour l'utilisateur est présent et documenté. Ce fichier permet à l'utilisateur d'entrer des informations sur les données, y compris le chemin d'accès aux données brutes, des informations sur les nouveaux concepts à intégrer (ou non) et le logiciel utilisé pour les données.
(o5) Les wrappers sont fonctionnels. Il existe un wrapper pour chaque phase du processus ETL (*extract, T1, T2, T3, load*).
(o6) Le processus de routine ne prend pas en compte les étapes de correspondance des concepts si aucun concept n'est ajouté.
(o7) Le schéma initial est chargé avec des données provenant d'Athena, en particulier les tables OMOP *VOCABULARY*, *CONCEPT* et *CONCEPT_RELATIONSHIP*, qui sont essentielles pour la correspondance sémantique. Ces données sont essentielles pour la cartographie sémantique, car elles fournissent l'identifiant de concept pour la normalisation de chaque terminologie et sont nécessaires pour le processus de relations.

Développement

(d1) Les interactions avec l'éditeur du logiciel sur la structure et le format des données extraites sont possibles.
(d2) Le format des données brutes et les technologies sont connus et maîtrisés.
(d3) Les données brutes sont intégrées dans un SGBD en conservant la nomenclature source.
(d4) En cas d'ajout d'un nouveau logiciel, un module d'extraction doit être développé et intégré.
(d5) En cas d'ajout de nouveaux logiciels, de nouvelles opérations sont ajoutées dans les scripts existants et de nouveaux concepts sont mis en correspondance.
(d6) La compatibilité entre iOS et les différentes versions de logiciels est gérée.
(d7) Des sauvegardes et des journaux sont fournis en cas d'interruptions ou de problèmes de système.
(d8) La gestion et la préservation des données abérantes dans les valeurs ou les formats de

données sont discutées.
(d9) L'unicité des clés primaires et des relations clé primaire-clé secondaire est vérifiée.

Généralisation

(m1) Le code source de chaque ETL est partagé sur un dépôt GitLab.
(m2) Une personne est désignée comme responsable de chaque phase des ETL : *extract, T2, T2, T3, load*.
(m3) À chaque phase, un wrapper est conçu pour encapsuler les scripts et est nommé "nom_de_la _phase_du_programme".
(m4) Le fichier de configuration permet de définir le nom du logiciel à partir duquel les données sont extraites.
(m5) Les phases d'*extract* et *T1* dépendent du logiciel ; le wrapper exécute les scripts contenus dans le dossier nommé d'après le logiciel choisi.
(m6) Une demande de fusion est initiée sur le dépôt GitLab pendant le développement d'un wrapper.
(m7) La demande de fusion est acceptée une fois que le programme fonctionne avec chaque source de données (données Crossway et Weda).

Qualité

(q1) Vérification de la similitude des données entre l'entrepôt de données et les DSE.
(q2) Les mesures du cadre de Kahn et al. sont calculées et validées (à l'aide de l'outil Atlas).
(q3) Utilisation des outils d'évaluation de la qualité des données développés par l'OHDSI (Atlas, Achilles, DQD).
(q4) Les analyses rétrospectives et les scripts sont basés sur le modèle final et sont reproductibles.
(q5) L'intégration de données provenant d'autres centres à l'aide d'un des logiciels programmés est possible.

Chapitre D

Tableau de suivi du calcul des indicateurs de la ROSP

Ce tableau regroupait les informations des indicateurs de la ROSP calculés sur les données de la MSP de Wattrelos. Les données utilisées en numérateur et en dénominateur ont été inscrites. Les difficultés rencontrées lors des calculs ont aussi été renseignées pour comprendre les différences de taux.

description	resultat_dw	periode	population	numerateur	denominateur	donnees	difficultes	regex
Part des patients MT traités par antidiabétiques Ayant bénéficié d'au moins 2 dosages d'HbA1c dans L'année	**85,50 %**	4e trimestre 2021	patients sous antidiabetiques (au moins 2 prescriptions d'antidiabetique dans les 12 derniers mois) ayant +16ans	Nb de patients sous antidiabetiques ayant eu + = 2 résultats d'HbA1c dans l'année 2021	Nb de patients sous antidiabetiques (au moins 2 prescriptions d'antidiabetique dans les 12 derniers mois). Exclure les patients ayant eu 2 dosages de	antidiabetique : A10 (atc) Hba1c : 3000963, 3034639, 4152671 (measurement_concept_id	mapping hba1c	-
Part des patients MT traités par antidiabétiques ayant bénéficié d'une consultation ou d'un examen du fond d'oeil ou d'une rétinographie dans les deux ans et un trimestre	**15,94 %**	4e trimestre 2021	patients sous antidiabetiques (au moins 2 prescriptions d'antidiabetique dans les 12 derniers mois) ayant +16ans	Nb de patients sous antidiabetiques ayant eu au moins 1 prescription de fond d'oeil dans les 27 derniers mois	Nb de patients sous antidiabetiques (au moins 2 prescriptions d'antidiabetique dans les 12 derniers mois)	antidiabetique : A10 (atc) Mention de fond d'oeil dans les notes CCAM : BGQP002,BGQP007, BGQK001, BGQP140 Actes ophtalmo : C, CS, CA, C2, HS, EXS, SES, V, VS, VA, VU/MU avec	regex fond d'oeil dans les notes	fond doeil|\bfo\b|r
Part des patients MT de moins de 81 ans traités par antidiabétiques ayant bénéficié d'une recherche annuelle de microalbuminurie sur échantillon d'urines et d'un dosage annuel de la créatininémie avec estimation du débit de filtration glomérulaire	**69,35 %**	4e trimestre 2021	patients sous antidiabetiques (au moins 2 prescriptions d'antidiabetiques dans les 12 derniers mois) ayant +81ans	Nb de patients sous antidiabetiques âgés de +81ans ayant eu au moins 1 résultat de microalbumine ET de créatinine dans les 12 derniers mois	Nb de patients sous antidiabetiques + agés de plus 81ans	antidiabetique : A10 (atc) Mention d'albuminerie ou creatinine dans measurement_source_valu e	regex résultats de biologie	microalbuminu| microalbu| albuminu|ualbu| albumine-urin| albuminu
Part des patients MT traités par antidiabétiques ayant bénéficié d'un examen clinique annuel des pieds par le MT ou d'une consultation de podologie dans l'année	**4e trimestre 2021**		patients sous antidiabetiques (au moins 2 prescriptions d'antidiabetique dans les 12 derniers mois)	Nb de patients sous antidiabétiques ayant eu un examen des pieds ou une consultation de podologie dans les 12 derniers mois	Nb de patients sous antidiabetiques			creat|creatinine| clairance|glomeru| dfg|debit-de-filtration-glomerulaire|ckd
Part des patients MT traités par antihypertenseurs ayant bénéficié d'une recherche annuelle de protéinurie ou de microalbuminurie et d'un dosage annuel de la	**45,34 %**	4e trimestre 2021	patients sous antihypertenseurs	Nb de patients sous antihypertenseurs ayant eu au moins 1 résultat de microalbumine OU protéinurie ET de créatinine dans les 12 derniers mois	Nb de patients sous antihypertenseurs dans les 12 derniers mois	antihypertenseurs : C02, C03, C07, C08, C09 (atc), C10BX03 Mention d'albuminerie ou proteinurie et creatinine dans	regex résultats de biologie	microalbuminu| microalbu| albuminu|ualbu| albumine-urin| albuminu| proteinu|protu|

patho_chroniques_page3 | patho_chroniques_page4 | prevention_page5 | prevention_page6 | prevention_page7 | efficience_page8 | efficience_page9 | total_resultats

FIGURE D.1 – Partie du tableau de suivi du calcul des critères de la ROSP

Chapitre E

Recommandations extraites de l'évaluation du tableau de bord par les ergonomes

Un rapport d'évaluation du tableau de bord a été élaboré par des ergonomes pour améliorer l'aspect visuel et didactique. L'évaluation a été faite suivant plusieurs critères. Les critères à améliorer et les recommandations ont été fournis dans un rapport rédigé par des ergonomes.

#	Critère	Description du problème	Recommandation
1	Les informations textuelles et les illustrations sont-elles réparties sur l'ensemble du tableau de bord ?	Sur la page de « prescription », les informations ne sont pas réparties sur l'ensemble de la page.	Dans la partie "prescription", descendre le graphique et l'agrandir
2	Les visualisations sont-elles contrôlées pour vérifier qu'il n'y a pas d'occlusion ou de blocage des informations dans les graphiques, les libellés ou les titres à différents niveaux de zoom ?	Lorsque l'écran est réduit, beaucoup d'éléments se superposent ce qui empêche de lire les informations. Lorsque l'on zoome au sein d'un graphique, on ne peut pas naviguer	Développer en responsive design pour que l'interface soit lisible et organisée même lorsque la taille de l'écran est réduite Au niveau des graphiques, modifier la fonction zoom pour rendre possible la navigation tout en permettant aux utilisateurs de savoir où ils se situent dans le graphique (donc laisser des repères des axes visibles)
3	L'utilisation d'une même couleur pour des concepts différents ou de couleurs différentes pour un même concept est-elle évitée ?	Des thématiques de couleur par page semblent avoir été développées mais elles ne sont pas complétement utilisées (ex. onglets). Par exemple le chapeau/titre de l'onglet n'est pas dans la gamme de couleur de la page qu'il désigne. Il est toujours écrit en blanc sur bleu.	Harmoniser l'utilisation des couleurs dans un même onglet (ex. titre, onglet, données etc.). Lorsqu'un onglet est sélectionné, ne pas séparer son entête du reste de son contenu : utiliser la même couleur de fond, ne pas les séparer visuellement par un trait.
4	Des éléments visuels familiers (par exemple, des icônes et des éléments graphiques) sont-ils utilisés pour tirer parti des connaissances préalables des utilisateurs ?	Certaines icônes sont utilisées pour des significations différentes (ex. le dossier avec une croix signifie à la fois nombre de médicaments, et nombre de consultations par patients)	Une icône ne doit avoir qu'une seule signification et réciproquement chaque signification ne doit être représentée que par une icône (ex. utiliser une icône « agenda » pour le nombre de consultations par patient et une icône « sachet de médicaments » pour le nombre de médicaments).
5	Des titres et des libellés clairs sont-ils utilisés pour guider les utilisateurs dans la lecture des graphiques ?	Les unités ne sont pas toujours indiquées ou parfois de manière incohérente. Dans les graphiques, il n'y a pas d'unité pour les deux axes ce qui empêche de comprendre le contenu. L'usage de la virgule pour les milliers complique la compréhension.	Chaque graphique ou tableau doit contenir un titre précis qui présente son contenu ainsi que des unités pour les axes et des légendes pour les séries de données. Dans le système français, le séparateur de décimales est représenté par une virgule et le séparateur de milliers par une espace.
6	Les graphiques interactifs et statiques sont-ils facilement identifiables par les utilisateurs ?	A l'arrivée de l'utilisateur sur l'onglet « biologie » rien n'indique qu'un graphique peut être présenté (ex. pas de « zone de graphique » avec axes, pas d'espace particulier pour un titre). L'utilisateur non formé pourrait passer sur la page sans rien remarquer.	Proposer par défaut une « zone de graphique » (ex. avec les axes d'abscisses et ordonnées présents par défaut, une zone de donnée avec un fond de couleur légèrement différente). Des instructions dans la zone de données indiqueraient à l'utilisateur qu'un graphique peut être affiché et comment l'afficher : « glisser déposer » dans la zone de données à partir d'une liste de données biologiques disponibles.

FIGURE E.1 – Extrait des critères et recommandations d'amélioration issus du rapport d'évaluation du tableau de bord des ergonomes.

Chapitre F

Concepts standards OMOP associés aux codes ATC d'anti-diabétiques

Concepts standards	Codes ATC
21600761	A10BB12
40251676	A10BD10
21600747	A10BA02
21600790	A10BX02
1123890	A10BK01
43534747	A10AE06
21600758	A10BB09
21600773	A10BD08
21600719	A10AB05
43534751	A10BD15
21600784	A10BH01
21600718	A10AB04
21600772	A10BD07
1123611	A10BJ05
21600739	A10AE04
1588678	A10AE56
1501778	A10BJ06
21600750	A10BB01
21600776	A10BF01
21600715	A10AB01
21600723	A10AC01
21600733	A10AD05
21600740	A10AE05
21600720	A10AB06
1123633	A10BJ01
21600785	A10BH02
21600732	A10AD04
1123630	A10BJ02
21600786	A10BH03

TABLE F.1 – Concepts standards OMOP associés aux codes ATC d'anti-diabétiques.