Mémoire présenté pour obtenir

l'Habilitation à Diriger des Recherches

Discipline : Biostatistiques, informatique médicale et technologies de communication

Antoine Lamer
Maître de Conférences Associé – Data Scientist

préparé à l'ULR 2694 METRICS (Université de Lille, CHU de Lille)

École Doctorale Biologie Santé de Lille

Mise au point de méthodologies de réutilisation des données de santé pour la recherche, le pilotage et l'évaluation de la qualité des soins

---

Présenté publiquement le 21 février 2024 devant le jury composé de

Présidente du jury	Annabelle Deram	Professeur des Universités	Université de Lille
Rapporteur	Fabien Feschet	Professeur des Universités	Université Clermont-Auvergne
Rapporteur	Bastien Rance	Maître de Conférences des Universités - Praticien Hospitalier	Université Paris Cité
Rapporteur	Emilie Olié	Professeur des Universités - Praticien Hospitalier	Université de Lille
Examinateur	Bruno Falissard	Professeur des Universités - Praticien Hospitalier	Université Paris 11
Examinateur	Marc Cuggia	Professeur des Universités - Praticien Hospitalier	Université de Rennes
Garant	Emmanuel Chazard	Professeur des Universités - Praticien Hospitalier	Université de Lille

Jamais dans la tendance, mais toujours dans la bonne direction.

Scred connexion

Remerciements

Je tiens tout d’abord à remercier les Professeurs Émilie Olié et Fabien Feschet, ainsi que le docteur Bastien Rance, pour avoir accepté de faire partie du jury et d’avoir rapporté ce travail d’HDR. Vos remarques pertinentes ont permis d’améliorer ce travail et d’ouvrir de nouvelles perspectives à mes travaux de recherche. Je tiens également à remercier les professeurs Bruno Falissard et Marc Cuggia d’avoir accepté de faire partie de ce jury. Mes remerciements vont également au Professeur Annabelle Deram, pour avoir accepté d’examiner mon travail, ainsi que pour son accueil au sein d’ILIS et la confiance qu’elle me renouvelle chaque année. Enfin, je tiens à remercier le Professeur Emmanuel Chazard pour avoir accepté d’être le garant scientifique de ce mémoire, mais également pour toutes ces années d’échanges concernant la réutilisation des données en santé.

Je souhaite remercier mes collègues de l'ULR METRICS qui ont contribué à créer un environnement de travail propice à la collaboration et à la stimulation de la créativité. En particulier, je tiens à exprimer ma profonde gratitude à Jean-Baptiste Beuscart pour les précieux conseils qu'il m'a prodigués au cours de ces dix dernières années. Je tiens également à le remercier pour les encouragements qu'il partage avec nous chaque jour en tant que directeur d'unité, nous incitant ainsi à aller toujours plus loin. Mes remerciements s'étendent également à Paul Quindroit, Pierre Balayé, Romaric Marcilly, Anaïs Payen, Mathieu Levaillant, Mathilde Fruchart, Grégoire Ficheur, Laurine Robert, Julien Soula, Mélanie Steffe et Renaud Périchon pour leurs précieuses contributions, tant sur le plan amical que professionnel.

Je tiens également à exprimer ma sincère gratitude envers Benjamin Guinhouya, Hervé Hubert, Christian Vilhelm, Florent Occelli, ainsi que tous mes collègues de l’UFR3S ILIS, pour leur précieux accompagnement et leur soutien dans le domaine de l'enseignement et de la pédagogie. Vos conseils, échanges et engagement ont grandement contribué à mon développement en tant qu'enseignant et chercheur, et je vous en suis reconnaissant. Votre collaboration a été essentielle pour l'amélioration de nos pratiques pédagogiques.

La recherche étant le fruit de la collaboration et de l'engagement collectif, je tiens à exprimer ma profonde gratitude envers toutes les personnes que j'ai eu l'honneur de superviser. Vos contributions et réflexions ont joué un rôle essentiel dans les travaux présentés dans ce mémoire. Pour cela, je salue Géry Laurent, Maëlle Baillet, Ikram Ouddarour, Coline Andries, Antoine Teston, ainsi que les étudiants du master Data Science en Santé pour leur précieuse contribution à cette recherche collaborative.

Je souhaite également exprimer ma reconnaissance envers mes collègues de la F2RSM, dont les contributions individuelles ont grandement contribué à mon développement dans le domaine de la recherche en psychiatrie et santé mentale. Je tiens à adresser mes remerciements à Maxime Bubrovszky, Marielle Wathelet, Chloé Saint-Dizier, Delphine Pastureau, Emile Farès, Sophie Loridan, Margot Trimbur, Oumaima El Qaoubii, ainsi qu'à tous les membres de l'équipe de la F2RSM, pour m'avoir chaleureusement accueilli au sein de leur groupe de travail. Votre soutien a été précieux pour mon parcours académique et professionnel, et je vous en suis extrêmement reconnaissant.

Un remerciement spécial à Ali Amad, Thomas Fovet, Charles Edouard-Notredame et Mathilde Horn pour m'avoir initié à la recherche en psychiatrie. J'aimerais également exprimer ma gratitude envers mes collègues anesthésistes-réanimateurs, Benoît Vallet, Benoît Tavernier, Mouhamed

Moussa, ainsi que mes collègues du CIC-IT, Régis Logier, Sylvia Pelayo, Julien de Jonckheere, avec qui j'ai fait mes premiers pas dans le domaine de la recherche.

Je tiens à remercier chaleureusement mes amis d'InterHop, Adrien Parrot, Nicolas Paris, Niels Martignène, Benjamin Popoff, Boris Delange et Chantal Parrot.

Enfin, je tiens à exprimer ma gratitude envers mes parents, ma sœur, et toute ma famille pour le soutien qu'ils apportent à mes projets. Une pensée pour Élise, Camille, William, Repié, Ian, Juliette, Maureen, Briac, Julie, Valentin, Jérèm, Dave, Anna, Quasar, Éli. Votre présence et vos encouragements ont été des piliers essentiels de mon parcours, et je vous en suis profondément reconnaissant.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1 Curriculum Vitae

Sommaire

1.1 Titres et fonctions

1.1.1 Titres et diplômes

2011-2015	**Thèse d'Université, en Recherche Clinique, Innovation Technologique et Santé Publique,** École Doctorale Biologie-Santé, Université de Lille, France. Titre : Contribution à la prévention des risques liés à l'anesthésie par la valorisation des informations hospitalières au sein d'un entrepôt de données. Encadrement : Pr Benoît Vallet (Université de Lille, EA2694) Pr Régis Logier (Université de Lille, EA2694)
2012-2013	**DU de Biostatistiques appliquées à la recherche médicale et à l'épidémiologie**, Université de Lille, France.
2010-2011	**Master Informatique**, Université de Valenciennes, France.
2007-2010	**Diplôme d'Ingénieur**, École Nationale Supérieure des Arts et Industries Textiles, Roubaix, France.

2005-2007	**DUT Mesures physiques**, Institut Universitaire de Technologie, Université de Lille, France

1.1.2 Fonctions

2022 – Présent	**Chercheur, Data Scientist**, Fédération régionale de recherche en psychiatrie et santé mentale des Hauts-de-France.
2020 – Présent	**Maître de conférence associé**, Institut lillois d'ingénierie de la santé, UFR 3S ILIS, Université de Lille.
2011 – 2022	**Data Scientist**, CHU de Lille.

1.1.3 Responsabilités

2019 – Présent	**Membre élu au conseil de laboratoire de l'ULR 2694 METRICS**, Université de Lille - ULR 2694 METRICS | https://metrics.univ-lille.fr/

1.2 Activités de recherche

1.2.1 Contexte

Je suis Maître de Conférences Associé depuis 2019. J'ai été précédemment doctorant de 2011 à 2015, puis ingénieur de recherche de 2015 à 2019. J'effectue actuellement mes activités de recherche au sein de l'Unité Labellisée de Recherche ULR 2694 METRICS : « Évaluation des technologies de santé et des pratiques médicales », dirigée par le Pr Jean-Baptiste Beuscart. Je suis membre de cette équipe depuis octobre 2011, auparavant intitulée EA 2694 « Santé Publique : épidémiologie et qualité des soins » et dirigée par le Pr. Alain Duhamel.

Ma thèse s'est déroulée au sein du laboratoire INSERM CIC-IT 1403 et en collaboration avec le pôle d'Anesthésie-Réanimation et le pôle de Santé Publique, Pharmacie et Pharmacologie du CHU de Lille. Depuis 2022, je suis data scientist à la Fédération régionale de recherche en psychiatrie et santé mentale des Hauts-de-France (F2RSM Psy).

1.2.2 Thématique de recherche : méthodologie et applications de la réutilisation des données de santé

Ma thématique de recherche porte sur la mise au point de méthodologies de réutilisation des données de santé pour répondre à des problématiques de recherche, de pilotage et d'évaluation de la qualité des soins. Cette thématique s'inscrit dans la continuité de mon travail de thèse d'université, soutenue en 2015 et intitulée « Contribution à la prévention des risques liés à l'anesthésie par la valorisation des informations hospitalières au sein d'un entrepôt de données » (1).

La mise au point de méthodologies de réutilisation des données de santé porte principalement sur quatre étapes clés : la collecte des données, l'intégration et la standardisation des données, l'extraction de caractéristiques, et leur exploitation pour retour vers l'utilisateur.

1.2.3 Publications

Type de publication	Quantité
Articles référencés Pubmed	37
Communications orales	30
Communications affichées	4
Organisation de workshops	2

Scores de publication sur la période 2016-2023 : Score SIGAPS : 452 | Score SAMPRA : 642

1.2.3.1 Articles dans des revues internationales à comité de lecture

1.2.3.1.1 Articles en 1er, 2ème, avant-dernier ou dernier auteur (n = 25)

Mastellari T, Saint-Dizier C, Fovet T, Geoffroy PA, Rogers J, **Lamer A**, Amad A. Exploring seasonality in catatonia diagnosis: Evidence from a large-scale population study. Psychiatry Res. 2023 Dec 1;331:115652. doi: 10.1016/j.psychres.2023.115652. Epub ahead of print. PMID: 38071881.

Lamer A, Carette F, Mobi H, Warembourg I, Amariei A, Saint-Dizier C, Bubrovszky M. Organization of French outpatient psychiatric clinics and delay to appointment. Encephale. 2023 Nov 30:S0013-7006(23)00201-4. doi: 10.1016/j.encep.2023.09.005. Epub ahead of print. PMID: 38040509.

Elefterion B, Cirenei C, Kipnis E, Cailliau E, Bruandet A, Tavernier B, **Lamer A**, Lebuffe G. Intraoperative mechanical power and postoperative pulmonary complications in non-cardiothoracic elective surgery patients: a ten-year retrospective cohort-study. Anesthesiology. 2023 Nov 28. doi: 10.1097/ALN.0000000000004848. Epub ahead of print. PMID: 38011027.

Levaillant M, Garabédian C, Legendre G, Soula J, Hamel JF, Vallet B, **Lamer A**. In France, the organization of perinatal care has a direct influence on the outcome of the mother and the newborn: Contribution from a French nationwide study. Int J Gynaecol Obstet. 2023 Jul 24. doi: 10.1002/ijgo.15004. Epub ahead of print. PMID: 37485702.

Doutreligne M, Degremont A, Jachiet PA, **Lamer A**, Tannier X. Good practices for clinical data warehouse implementation: A case study in France. PLOS Digit Health. 2023 Jul 6;2(7):e0000298. doi: 10.1371/journal.pdig.0000298. PMID: 37410797; PMCID: PMC10325086.

Levaillant M, Rony L, Hamel-Broza JF, Soula J, Vallet B, **Lamer A**. In France, distance from hospital and health care structure impact on outcome after arthroplasty of the hip for proximal fractures of the femur. J Orthop Surg Res. 2023 Jun 9;18(1):418. doi: 10.1186/s13018-023-03893-4. PMID: 37296484; PMCID: PMC10257255.

Fovet T, Saint-Dizier C, Wathelet M, Horn M, Thomas P, Guillin O, Coldefy M, D'Hondt F, Amad A, **Lamer A**. Opening the black box of hospitalizations in French high-secure psychiatric forensic units. Encephale. 2023 May 26:S0013-7006(23)00079-9. doi: 10.1016/j.encep.2023.04.008. Epub ahead of print. PMID: 37246100.

Quindroit P, Fruchart M, Degoul S, Périchon R, Soula J, Marcilly R, **Lamer A**. Definition of a practical taxonomy for referencing data quality problems in healthcare databases. Methods Inf Med. 2022 Nov 10. doi: 10.1055/a-1976-2371.

Guardiolle V, Bazoge A, Morin E, Daille B, Toublant D, Bouzillé G, Merel Y, Pierre-Jean M, Filiot A, Cuggia M, Wargny M, **Lamer A**, Gourraud PA. Linking Biomedical Data Warehouse Records With the National Mortality Database in France: Large-scale Matching Algorithm. JMIR Med Inform. 2022 Nov 1;10(11):e36711. doi: 10.2196/36711. PMID: 36318244.

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 Oct 17;10(10):e38936.

Payen A, Godard-Sebillotte C, Sourial N, Soula J, Verloop D, Defebvre MM, Dupont C, Dambre D, **Lamer A**, Beuscart JB. The impact of including a medication review in an integrated care pathway: A pilot study. Br J Clin Pharmacol. 2022 Sep 26.

Lamer A, Moussa MD, Marcilly R, Logier R, Vallet B, Tavernier B. Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project. J Clin Monit Comput. 2022 Aug 6.

Fovet T, Baillet M, Horn M, Chan-Chee C, Cottencin O, Thomas P, Vaiva G, D'Hondt F, Amad A, **Lamer A**. Psychiatric Hospitalizations of People Found Not Criminally Responsible on Account of Mental Disorder in France: A Ten-Year Retrospective Study (2011-2020). Front Psychiatry. 2022 Apr 5;13:812790.

Fovet T, Chan-Chee C, Baillet M, Horn M, Wathelet M, D'Hondt F, Thomas P, Amad A, **Lamer A**. Psychiatric hospitalisations for people who are incarcerated, 2009-2019: An 11-year retrospective longitudinal study in France. EClinicalMedicine. 2022 Apr 8;46:101374.

Oubenali N, Messaoud S, Filiot A, **Lamer A**, Andrey P. Visualization of medical concepts represented using word embeddings: a scoping review. BMC Med Inform Decis Mak. 2022 Mar 29;22(1):83.

Beigné M, **Lamer A**, Eck M, Horn M, Benbouriche M, Thomas P, Amad A, Fovet T. Parcours de soins et expertises psychiatriques pré-sentencielles : une étude descriptive au centre pénitentiaire de Château-Thierry [A descriptive study of psychiatric care and pre-sentencing psychiatric reports in a French high-security prison]. Encephale. 2022 Mar 21:S0013-7006(22)00031-8. French.

Erlich C, **Lamer A**, Moussa MD, Martin J, Rogeau S, Tavernier B. End-tidal Carbon Dioxide for Diagnosing Anaphylaxis in Patients with Severe Postinduction Hypotension. Anesthesiology. 2022 Mar 1;136(3):472-481.

Levaillant M, Marcilly R, Levaillant L, Michel P, Hamel-Broza JF, Vallet B, **Lamer A**. Assessing the hospital volume-outcome relationship in surgery: a scoping review. BMC Med Res Methodol. 2021 Oct 9;21(1):204.

Paris N, **Lamer A**, Parrot A. Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study. JMIR Med Inform. 2021 Dec 14;9(12):e30970.

Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart JB, Tavernier B, Moussa MD. Transforming Anesthesia Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study. J Med Internet Res. 2021 Oct 29;23(10):e29259.

Levaillant M, Marcilly R, Levaillant L, Vallet B, **Lamer A**. Assessing the hospital volume-outcome relationship in surgery: a scoping review protocol. BMJ Open. 2020 Oct 6;10(10):e038201.

Laurent G, Moussa MD, Cirenei C, Tavernier B, Marcilly R, **Lamer A**. Development, implementation and preliminary evaluation of clinical dashboards in a department of anesthesia. J Clin Monit Comput. 2020 May 16:1–10.

Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, Ficheur G, Chazard E, Beuscart JB. Transforming French Electronic Health Records Into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. Appl Clin Inform. 2020 Jan.

Lamer A, Jeanne M, Marcilly R, Kipnis E, Schiro J, Logier R, et al. Methodology to automatically detect abnormal values of vital parameters in anesthesia time-series: Proposal for an adaptable algorithm, Comput. Methods Programs Biomed. 2016 Jun.

Lamer A, De Jonckheere J, Marcilly R, Tavernier B, Vallet B, Jeanne M, et al. A substitution method to improve completeness of events documentation in anesthesia records. J Clin Monit Comput. 2015 Jan 30.

1.2.3.1.2 Articles autres que 1er, 2ème, avant-dernier ou dernier auteur (n = 12)

Jauffret C, Périchon R, **Lamer A**, Cortet B, Chazard E, Paccou J. Association between sarcopenia and risk of major adverse cardiac and cerebrovascular events-UK Biobank database. J Am Geriatr Soc. 2023 Nov 9. doi: 10.1111/jgs.18664. Epub ahead of print. PMID: 37945290.

Jauffret C, Périchon R, **Lamer A**, Cortet B, Chazard E, Paccou J. Association Between Sarcopenia and Fracture Risk in a Population From the UK Biobank Database. J Bone Miner Res. 2023 Jul 17. doi: 10.1002/jbmr.4884. Epub ahead of print. PMID: 37458535.

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 May 21;9:20552076231176114. doi: 10.1177/20552076231176114. PMID: 37228486; PMCID: PMC10204053.

Puigrenier S, Giovannelli J, Lamblin N, De Groote P, Fertin M, Bervar JF, **Lamer A**, Edmé JL, Balquet MH, Sobanski V, Launay D, Hachulla É, Sanges S. Mild pulmonary hemodynamic alterations in patients with systemic sclerosis: relevance of the new 2022 ESC/ERS definition of pulmonary hypertension and impact on mortality. Respir Res. 2022 Oct 15;23(1):284.

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 Sep 2;

Moussa MD, Beyls C, **Lamer A**, Roksic S, Juthier F, Leroy G, Petitgand V, Rousse N, Decoene C, Dupré C, Caus T, Huette P, Guilbart M, Guinot PG, Besserve P, Mahjoub Y, Dupont H, Robin E, Meynier J, Vincentelli A, Abou-Arab O. Early hyperoxia and 28-day mortality in patients on

venoarterial ECMO support for refractory cardiogenic shock: a bicenter retrospective propensity score-weighted analysis. Crit Care. 2022 Aug 26;26(1):257.

Moussa MD, Rousse N, Abou Arab O, **Lamer A**, Gantois G, Soquet J, Liu V, Mugnier A, Duburcq T, Petitgand V, Foulon V, Dumontet J, Deblauwe D, Juthier F, Desbordes J, Loobuyck V, Labreuche J, Robin E, Vincentelli A. Subclavian versus femoral arterial cannulations during extracorporeal membrane oxygenation: A propensity-matched comparison. J Heart Lung Transplant. 2022 Jan 10:S1053-2498(22)00009-2.

Levaillant M, Wathelet M, **Lamer A**, Riquin E, Gohier B, Hamel-Broza JF. Impact of COVID-19 pandemic and lockdowns on the consumption of anxiolytics, hypnotics and antidepressants according to age groups: a French nationwide study. Psychol Med. 2021 Dec 14:1-7.

Perrot J, Hamel JF, **Lamer A**, Levaillant M. The Relationship between the Immigrant Rate and Health Status in the General Population in France. J Pers Med. 2021 Jun 30;11(7):627.

Moussa MD, Soquet J, **Lamer A**, Labreuche J, Gantois G, Dupont A, Abou-Arab O, Rousse N, Liu V, Brandt C, Foulon V, Leroy G, Schurtz G, Jeanpierre E, Duhamel A, Susen S, Vincentelli A, Robin E. Evaluation of Anti-Activated Factor X Activity and Activated Partial Thromboplastin Time Relations and Their Association with Bleeding and Thrombosis during Veno-Arterial ECMO Support: A Retrospective Study. J Clin Med. 2021 May 17;10(10):2158.

Visade F, Babykina G, **Lamer A**, Defebvre MM, Verloop D, Ficheur G, Genin M, Puisieux F, Beuscart JB. Importance of previous hospital stays on the risk of hospital re-admission in older adults: a real-life analysis of the PAERPA study population. Age Ageing. 2020 Jul 20:afaa139.

Moussa, Mouhamed D., Arthur Durand, Guillaume Leroy, Liu Vincent, **A Lamer**, Guillaume Gantois, Olivier Joulin, et al. Central Venous-to-Arterial PCO2 Difference, Arteriovenous Oxygen Content and Outcome after Adult Cardiac Surgery with Cardiopulmonary Bypass: A Prospective Observational Study. Eur J Anaesthesiol. 2019 Jan 16.

1.2.3.2 Articles dans des revues nationales à comité de lecture (n = 0)

Aucune communication

1.2.3.3 Communications orales

1.2.3.3.1 Conférences sur invitation lors de congrès nationaux (n=3)

EMOIS 2023 : Méthodes de représentation du parcours patient

Symposium AIM 2022 : Représentation des données patients dans une vision longitudinale : mise en œuvre du modèle de données OMOP au CHU de Lille

SFIMAR 2013 : Développement d’un entrepôt de données d’anesthésie

1.2.3.3.2 Communications orales à des congrès internationaux référencés Medline (n = 23)

Lamer A, Saint-Dizier C, Fares E, Debien C, Cleva E, Whatelet M, Notredame CE. Automated Monitoring Reports of the Activity of the French National Professional Suicide Prevention Helpline. Stud Health Technol Inform. 2023 May 18;302:474-475. doi: 10.3233/SHTI230177. PMID: 37203721.

Fruchart M, **Lamer A**, Lemaitre M, Beuscart JB, Calafiore M, Quindroit P. Description of a French Population of Diabetics Treated Followed up by General Practitioners. Stud Health Technol Inform. 2023 May 18;302:856-860. doi: 10.3233/SHTI230289. PMID: 37203517.

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 May 18;302:502-503. doi: 10.3233/SHTI230191. PMID: 37203735.

Saint-Dizier C, **Lamer A**, Zaanouar M, Amariei A, Quindroit P. OpenDataPsy: An Open-Data Repository with Standardized Storage and Description for Research in Psychiatry. Stud Health Technol Inform. 2023 May 18;302:851-855. doi: 10.3233/SHTI230288. PMID: 37203516.

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 Aug 31;298:82-86.

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 Aug 31;298:51-55.

Kerisit E, Legrand B, Calafiore M, Rochoy M, Chazard E, Marcilly R, **Lamer A**. Awareness and Perception of Google® Reviews Among French GPs. Stud Health Technol Inform. 2022 Jun 6;290:1118-1119.

Chazard E, Balaye P, Balcaen T, Genin M, Cuggia M, Bouzille G, **Lamer A**. "Book Music" Representation for Temporal Data, as a Part of the Feature Extraction Process: A Novel Approach to Improve the Handling of Time-Dependent Data in Secondary Use of Healthcare Structured Data. Stud Health Technol Inform. 2022 Jun 6;290:567-571.

Fruchart M, Guinhouya B, Pelayo S, Vilhelm C, **Lamer A**. Jupyter Notebooks for Introducing Data Science to Novice Users. Stud Health Technol Inform. 2022 May 25;294:823-824.

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 May 25;294:705-706.

Martignene N, Amad A, Bellet J, Tabareau J, D'Hondt F, Fovet T, **Lamer A**. Goupile: A New Paradigm for the Development and Implementation of Clinical Report Forms. Stud Health Technol Inform. 2022 May 25;294:540-544.

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 May 25;294:505-509.

Boudis F, Clement G, Bruandet A, **Lamer A**. Automated Generation of Individual and Population Clinical Pathways with the OMOP Common Data Model. Stud Health Technol Inform. 2021 May 27;281:218-222. doi: 10.3233/SHTI210152. PMID: 34042737.

Lamer A, Filiot A, Bouillard Y, Mangold P, Andrey P, Schiro J. Specifications for the Routine Implementation of Federated Learning in Hospitals Networks. Stud Health Technol Inform. 2021 May 27;281:128-132. doi: 10.3233/SHTI210134. PMID: 34042719.

Martignene N, Balcaen T, Bouzille G, Calafiore M, Beuscart JB, **Lamer A**, Legrand B, Ficheur G, Chazard E. Heimdall, a Computer Program for Electronic Health Records Data Visualization. Stud Health Technol Inform. 2020 Jun 16;270:247-251.

Laurent G, Guinhouya B, Whatelet M, **Lamer A**. Automatic Exploitation of YouTube Data: A Study of Videos Published by a French YouTuber During COVID-19 Quarantine in France. Stud Health Technol Inform. 2020 Nov 23;275:112-116.

Mangold P, Filiot A, Moussa M, Sobanski V, Ficheur G, Andrey P, **Lamer A**. A Decentralized Framework for Biostatistics and Privacy Concerns. Stud Health Technol Inform. 2020 Nov 23;275:137-141.

Laurent G, Guinhouya B, Whatelet M, **Lamer A**. Automatic Exploitation of YouTube Data: A Study of Videos Published by a French YouTuber During COVID-19 Quarantine in France. Stud Health Technol Inform. 2020 Nov 23;275:112-116.

Lamer A, Laurent G, Pelayo S, El Amrani M, Chazard E, Marcilly R. Exploring Patient Path Through Sankey Diagram: A Proof of Concept. Stud Health Technol Inform. 2020 Jun 16;270:218-222.

Chazard E, Ficheur G, Caron A, **Lamer A**, Labreuche J, Cuggia M, Genin M, Bouzille G, Duhamel A. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. Stud Health Technol Inform. 2018;255:15-19.

Lamer A, Demay A, Marcilly R. Data Reuse Through Anesthesia Data Warehouse: Searching for New Use Contexts. Stud Health Technol Inform. 2018;255:102-106.

Lamer A, Ficheur G, Rousselet L, van Berleere M, Chazard E, Caron A. From Data Extraction to Analysis: Proposal of a Methodology to Optimize Hospital Data Reuse Process. Stud Health Technol Inform. 2018;247:41-45.

Lamer A, Jeanne M, Ficheur G, Marcilly R. Automated Data Aggregation for Time-Series Analysis: Study Case on Anaesthesia Data Warehouse. Stud Health Technol Inform. 2016;221:102-6.

1.2.3.3.3 Communications orales à des congrès nationaux (n = 4)

Lamer A, Quindroit P, Ismail M, Boudis F, Beuscart JB, Chazard E. Méthodes et outils de représentation du parcours patient. EMOIS, 2023.

Lamer A, Fruchart M, Paris N, Popoff B, Payen A, Balcaen T, Gacquer W, Cuggia M, Doutreligne M, Chazard E. Description standardisée du processus d'extraction de caractéristiques afin d'améliorer la réutilisation des données. EMOIS, 2023.

Lamer A, Moussa M, Tavernier B. Entrepôt de données et réutilisation des données d’anesthésie. JLAR, 2021.

Lamer A. Représentation des données patients dans une vision longitudinale : mise en œuvre du modèle de données OMOP au CHU de Lille. Séminaire AIM, 2021.

1.2.3.3.4 Communications affichées à des congrès nationaux ou internationaux (n = 4)

Saint-Dizier C, **Lamer A**, Bubrovszky M. Comparaison des patients hospitalisés en France en 2022 dans des hôpitaux psychiatriques publics, privés à but non lucratif et privés à but lucratif : une étude transversale nationale. Congrès Français de Psychiatrie, 2023.

Lamer A, Carette F, Mobi H, Warembourg I, Amariei A, Saint-Dizier C, Bubrovszky M. Organisation et délais de rendez-vous dans les centres médico-psychologiques adultes des Hauts-de-France. Congrès Français de Psychiatrie, 2023.

Saint-Dizier C, Kfoury P, Fovet T, Amad A, Wathelet M, **Lamer A**. Description des hospitalisations avant et après un passage aux urgences pour tentatives de suicide. Congrès Français de Psychiatrie, 2022.

Fovet T, Saint-Dizier C, Wathelet M, Horn M, Thomas P, Guillin O, Coldefy M, D'Hondt F, Amad A, **Lamer A**. Les hospitalisations en unités pour malades difficiles de 2012 à 2021. Congrès Français de Psychiatrie, 2022.

1.2.3.3.5 Workshops (n=2)

Conférence Medical Informatics in Europe 2022 à Nice : «Participatory Definition of Feature Extraction» (organisateur principal, animateur).

Conférence Medical Informatics in Europe 2023 à Göteborg : « Lessons Learned from the Implementation of Health Data Warehouses : Participatory Development of Recommendations » (organisateur principal, animateur).

1.2.3.4 Logiciels (n=1)

J'ai encadré la conception et le développement de Goupile, un outil de conception d'eCRF libre et gratuit qui s'efforce de rendre la création de formulaires et la saisie de données à la fois puissantes et faciles. Goupile est placé sous licence AGPL v3. Site web : www.goupile.fr

Communication associée :

Martignene N, Amad A, Bellet J, Tabareau J, D'Hondt F, Fovet T, **Lamer A**. Goupile: A New Paradigm for the Development and Implementation of Clinical Report Forms. Stud Health Technol Inform. 2022 May 25;294:540-544.

1.2.4 Encadrement de thèses d’université et de masters recherche

1.2.4.1 Encadrement de thèses d’université

2023-Présent	Chloé Saint-Dizier, Thèse d’Université, École Doctorale Biologie-Santé, Université de Lille Développement d’un pipeline d'analyses spatio-temporelles pour aider au pilotage territorialisé en psychiatrie Directeur : Dr. Michaël Génin. Rôle : co-encadrant.
2022-Présent	Mathilde Fruchart, Thèse d’Université, École Doctorale Biologie-Santé, Université de Lille. Optimisation de la réutilisation des données de soins primaires pour le suivi de l’activité et la recherche Directeur : Pr. Benjamin Guinhouya. Rôle : co-encadrant.
2019-2022	Mathieu Levaillant, Thèse d’Université, École Doctorale Biologie-Santé, Université de Lille. Planification territoriale des soins en France pour assurer la qualité et la sécurité des soins : apport des études sur les bases de données administratives françaises. Directeur : Pr. Benoît Vallet. Rôle : co-encadrant.

1.2.4.2 Encadrement de Masters recherche (7 M1 et 9 M2)

2023	Colines Andries, M1 Data Science pour la Santé, ILIS, Université de Lille. Développement d’un package R pour la représentation des données de psychiatrie.
	Ikram Ouddarour, M1 Data Science pour la Santé, ILIS, Université de Lille. Développement d'une cartographie de la santé mentale des territoires des HDF et d'Occitanie
	Avave Itir, M1 Data Science pour la Santé, ILIS, Université de Lille. Génération de jeux de données synthétiques en santé à partir de méthodes de perturbation.
	Habiba El-bali, M1 Data Science pour la Santé, ILIS, Université de Lille, 2023. Réutilisation des données de visites médicales à l’Institut Pasteur.
	Chloé Saint-Dizier, M2 Data Science pour la Santé, ILIS, Université de Lille, 2022-2023. Études rétrospectives en psychiatrie et santé mentale à partir des bases de données médico-administratives nationales.
2022	Chloé Saint-Dizier, M2 Data Science pour la Santé, ILIS, Université de Lille, 2022-2023. Développement de tableaux de bords pour le suivi de l’activité psychiatrique dans les Hauts-de-France.

	Claire Butaye, M2 Data Science pour la Santé, ILIS, Université de Lille, 2022. Développement d'un package pour faciliter l'extraction de caractéristiques à partir du modèle de données OMOP.
	Christophe Huz, M2 Biologie Santé, Faculté de médecine, Université de Lille. Association entre hypotension artérielle peropératoire, bas débit cardiaque estimé à partir du CO2 expiré et morbi-mortalité postopératoire en chirurgie non-cardiaque.
	Maelle Baillet, Master 2 Data Science pour la santé, Université de Lille. Lien COVID et tentatives de suicide par joint disease mapping. Durée : 5 mois. Encadrement : 40% (Co-encadré avec Michael Génin et Marielle Wathelet).
2021	Fabio Boudis, M2 Data Science pour la Santé, ILIS, Université de Lille. Proposition d'une méthode automatiser la représentation du parcours patient.
	Sabrina Messaoud, M2 Data Science pour la Santé, ILIS, Université de Lille. Visualisation de concepts médicaux par la méthode de « word embeddings ».
	Naima Oubenali, M2 Data Science pour la santé, ILIS, Université de Lille. Visualisation de concepts médicaux par la méthode de « word embeddings ».
	Alexandre Bourgeois, M2 Biologie Santé, Faculté de médecine, Université de Lille. Hypotension artérielle peropératoire et mortalité postopératoire en chirurgie non-cardiaque : le traitement par vasoconstricteurs améliore-t-il le pronostic ? Étude d'une cohorte de 250 000 patients.
	Haris Patel, M1 Data Science pour la Santé, ILIS, Université de Lille. Réutilisation des données d'un cabinet de médecine générale pour la recherche.
2020	Mathilde Fruchart, Master 1 Data Science pour la Santé, ILIS, Université de Lille. Développement de fonctions pour faciliter l'extraction de caractéristiques à partir de données d'anesthésie-réanimation.
2019	Géry Laurent, Master 1 Data Science pour la Santé. Implémentation et test d'une plateforme de tableaux de bord pour le suivi de l'activité en anesthésie-réanimation.

1.2.5 Activités d'animation et de rayonnement

1.2.5.1 Prix et distinctions scientifiques

Prix du meilleur poster au congrès Medical Informatics in Europe 2016 :

Marcilly R, **Lamer A.** Evidence-Based Usability Database for Medication Alerting Systems. EFMI; 2016. https://hal.science/hal-04146995

1.2.5.2 Appartenance à des comités éditoriaux

Je suis membre du comité éditorial de journaux internationaux :

Computer Methods and Programs in Biomedicine Update

BMC Research Notes.

1.2.5.3 Reviewer pour des journaux internationaux à comité de lecture

Je réalise régulièrement des reviews pour les journaux internationaux suivants :

- Journal of Medical Internet Research
- JMIR mHealth and uHealth
- JMIR Medical informatics
- JMIR Public Health and Surveillance
- JMIR Mental Health
- JMIR Perioperative Medicine
- JMIR Infodemiology

1.2.5.4 Sociétés savantes

J'adhère aux sociétés suivants :

- European Federation of Medical Informatics
- Association d'Informatique Médicale

1.2.6 Collaborations

Nous collaborons avec le CERN (Conseil européen pour la recherche nucléaire, Suisse) pour établir une plate-forme de données ouvertes et pour analyser les données temporelles. Ce projet s'intègre aussi localement dans le CPER Tec'santé, avec les équipes membres de la SFR-TSM (directeur Nicolas Blanchemain), notamment pour la mise en place de la plateforme e-santé (Jean-Baptiste Beuscart et Emmanuel Chazard).

Nous investiguons le calcul fédéré avec l'équipe Inria MAGNET (Marc Tommasi, Aurélien Bellet, Paul Mangold) et la représentation du parcours patient avec l'équipe Inria MODAL (Pr. Christophe Biernacki, Pr. Cristian Preda, Pr Sophie Dabo).

Nos études liées à la santé mentale et la psychiatrie sont menées avec l'équipe LilNcog (Lille Neuroscience & CognitionServices, Ali Amad, Thomas Fovet, Mathilde Horn, Fabien D'Hondt).

Mon activité à la Fédération régionale de recherche en psychiatrie et santé mentale (F2RSM) s'appuie sur les collaborations avec les établissements de psychiatrie et de santé mentale des Hauts-de-France, et en particulier le CHU de Lille, le CH de Calais, l'EPSM de l'agglomération lilloise et l'établissement de santé mentale de Lille de la MGEN. Nous collaborons activement avec

l’ARS Hauts-de-France pour la mise en place d’indicateurs de suivi de la santé mentale, et pour l’évaluation des politiques de santé.

Nos travaux autour de la standardisation des données, l’utilisation du modèle OMOP, l’interopérabilité et l’extraction des caractéristiques sont menées conjointement avec les pôles d’anesthésie-réanimations du CHU de Lille, du CHU d’Amiens, du CHU de Rouen, du CH Saint-Malo et de l’Hôpital Foch.

Enfin, nous collaborons avec l’équipe Inserm UMR 1085 (Ester - Equipe d'épidémiologie en santé au travail et ergonomie) pour la conduite d’études rétrospectives à partir des bases de données nationales de l’assurance maladie.

1.3 Activités d’enseignement liées à la recherche

1.3.1 Contexte

Mes activités d’enseignements se déroulent principalement au sein de l’UFR3S, dans le département Ingénierie et de Management de la Santé de l’Université de Lille (doyen : Pr. Annabelle Deram). Les sections suivantes ne présentent que les activités d’enseignement liées à la recherche et ne reflètent pas la totalité de mes enseignements.

1.3.2 Masters recherche

Master Data Science en Santé, Faculté d'Ingénierie et Management de la Santé, Université de Lille:

- Programmation R (15h)
- Bases de données relationnelles (12h)
- Bases de données NoSQL (6h)
- Technologies Big Data (12h)
- Initiation à la recherche (3h)
- Communication (3h)

Master Ingénierie de la Santé - Parcours Healthcare business et recherche clinique, Faculté d'Ingénierie et Management de la Santé, Lille:

- Initiation aux Big Data (8h)

Chapitre 2 Introduction générale à la thématique de recherche

Sommaire

Ce chapitre mettra en lumière la distinction entre les approches classiques de collecte de données en recherche et le nouvel horizon qu'offre la réutilisation des données. Nous explorerons comment les entrepôts de données et les modèles de données communs simplifient la réutilisation des données, favorisant ainsi les collaborations entre institutions. Nous examinerons également l'importance de l'extraction de caractéristiques pour transformer les données brutes en informations exploitables. La visualisation des données sera abordée dans ce contexte. Enfin, nous détaillerons la structure de ce mémoire ainsi que les différentes contributions collaboratives.

2.1 Réutilisation des données

En recherche en santé, qu'elles soient observationnelles ou interventionnelles, prospectives ou rétrospectives, les méthodologies traditionnelles de recherche impliquent la collecte de données spécifiquement pour la recherche. Ce recueil de données est souvent réalisé manuellement, conformément au protocole de recherche, qui définit comment chaque variable doit être collectée, et à l'aide d'un formulaire papier ou informatique (également connu sous le nom de *clinical report form*, CRF en anglais) (2). Les données englobent des éléments tels que les critères d'inclusion, les variables d'évaluation (comme la durée de séjour à l'hôpital ou la survie), les expositions (comme la prise de médicaments ou des procédures chirurgicales), et les variables de contrôle (par exemple, l'âge, le sexe et les antécédents du patient), qui serviront à des fins d'analyse statistique. Ces variables ont généralement été définies au regard de la littérature scientifique du domaine étudié. Ces variables sont prédéfinies et collectées manuellement au fil du temps, avec l'expertise humaine, une valeur à la fois, tout en tenant compte du contexte clinique. Si nécessaire, il est possible de recourir à des sources de données tierces ou à l'expertise de professionnels de la santé. Les informations ainsi recueillies sont intellectuellement traitées par la personne responsable du recueil, en respectant les normes indiquées dans le protocole. L'ensemble final de données est composé d'informations explicites qui ne nécessitent pas de calculs supplémentaires et peuvent être directement utilisées lors de l'analyse statistique. Cependant, cette méthode s'avère coûteuse et demande beaucoup de temps. De plus, comme il s'agit de recherche « sur les personnes », ce processus requiert au préalable l'obtention d'une autorisation du comité de protection des personnes (CPP). Cette autorisation inclut la limitation du recueil au nombre de

sujets nécessaire pour répondre à l'objectif énoncé dans le protocole. Par conséquent, généralement, elle ne permet de rassembler qu'un échantillon de taille limitée, utilisé pour une seule étude (3).

Une autre approche, la *réutilisation des données*, consiste à s'appuyer sur des sources de données déjà produites et disponibles à moindre coût (4,5). Dans le cadre de notre travail, nous adopterons la définition de Safran et al. pour définir la réutilisation des données comme « l'utilisation non directe des informations personnelles de santé, y compris, sans s'y limiter, l'analyse, la recherche, la mesure de la qualité/sécurité, la santé publique, le paiement, la certification ou l'accréditation des prestataires de soins, ainsi que le marketing et d'autres activités commerciales, y compris les activités strictement commerciales » (4).

La réutilisation des données entre dans le cadre de la recherche sur les données, par opposition à la recherche sur les personnes. Elle tire parti de l'informatisation croissante des dossiers médicaux au cours des dernières décennies, ce qui a produit une quantité importante de données cliniques structurées disponibles sous format électronique (6–8). Initialement, ces logiciels, et les bases de données qui y sont associées, étaient développés dans le but de collecter et gérer des données pour des besoins liés à la prestation de soins, à la gestion des séjours médicaux et à la facturation des services de santé. En plus de ces objectifs d'origine, ces grands volumes de données ouvrent aussi la voie à des opportunités de réutilisation, dans des domaines tels que la recherche, l'amélioration de la qualité des soins, la gestion des opérations médicales et la santé publique (4,9).

2.2 Entrepôt de données et modèles de données commun

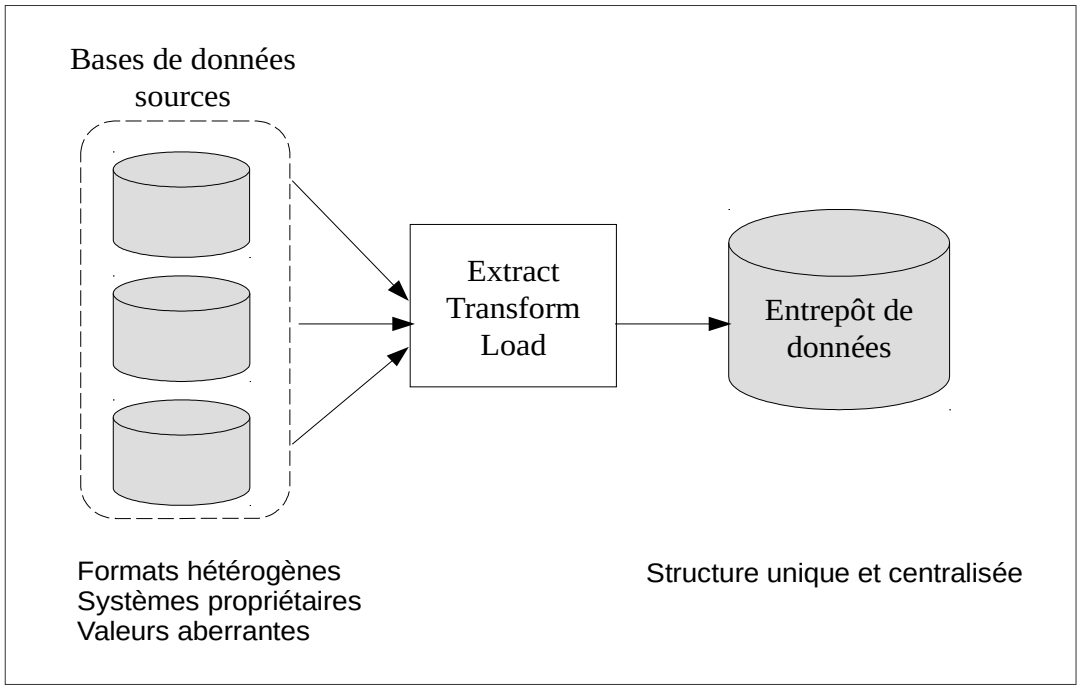


Figure 1: Des bases de données sources à l'entrepôt de données

Malgré les opportunités qu'elle offre, la réutilisation des données rencontre de nombreuses difficultés avant de pouvoir être mise en œuvre efficacement. Tout d'abord, les données brutes sont stockées dans des fichiers ou des bases de données aux formats hétérogènes, et propriétés des éditeurs de logiciels. Les données sont affectées par de nombreux problèmes de qualité qui découlent de la manière dont elles ont été saisies ou collectées (10). A ce stade, il est compliqué, voir impossible de croiser des sources de données différentes.

Pour lever ces barrières, plusieurs opérations sont réalisées pour implémenter une nouvelle structure, unique et centralisée, l'entrepôt de données. La première opération consiste à sélectionner les données pertinentes dans chacune des sources de données. Les données sont ensuite transformées et nettoyées afin de (i) filtrer ou corriger les valeurs aberrantes, (ii) convertir, normaliser ou unifier les formats et les unités, (iii) aligner les identifiants entre les sources de données, et enfin (iv) les regrouper et les charger dans l'entrepôt de données, en vue de faciliter l'analyse et la prise de décision (11,12) (Figure 1). Ce processus est appelé ETL, pour *extract-transform-load* (13). Il a vocation a être automatisé pour alimenter régulièrement l'entrepôt de données avec un minimum d'interventions humaines.

Les entrepôts de données de santé (EDS) ont principalement été développés à partir des logiciels hospitaliers et contiennent les informations relatives aux différentes étapes du parcours du patient, incluant les passages dans les unités de soins, le statut vital, les actes médicaux effectués, les diagnostics posés, les résultats d'analyses biologiques, ainsi que les administrations de médicaments. Des sources plus spécialisées peuvent également enrichir ces EDS avec des mesures de signaux physiologiques provenant des services de réanimation et des salles d'opération, des données d'imagerie médicale, ainsi que des comptes-rendus médicaux (14). En dehors de l'hôpital, les bases de données médico-administratives contiennent les éléments nécessaires au remboursement des prestations de soins comme les consultations des professionnel de santé, les délivrances de médicaments en pharmacie d'officine, les hospitalisations (15,16).

Même si de nombreuses études ont pu être réalisée à partir des EDS, l'hétérogénéité des structures de données et des vocabulaires locaux complique la mise en commun des données, le partage d'outils et de méthodes, ainsi que la comparaison des résultats entre plusieurs entités (i.e. établissements de soins, institutions, pays) (17,18). Des initiatives ont vu le jour pour promouvoir la réutilisation des données grâce au partage et à la fédération des données cliniques à grande échelle et à la mise en œuvre de modèles de données communs (19–23). Le modèle le plus récent et sans doute le plus répandu, le modèle OMOP (Observational Medical Outcomes Partnership), a été développé par le consortium international OHDSI (Observational Health Data Sciences and Informatics) (24–27). Ce modèle est conçu pour les études observationnelles, la pharmaco-épidémiologie et la modélisation prédictive au niveau du patient. Le modèle est structuré de manière standardisée, et le vocabulaire est uniformisé afin de surmonter les disparités liées aux modèles de données et aux vocabulaires qui varient entre les éditeurs de logiciels, les établissements médicaux et les pays. La communauté OHDSI partage des méthodes et des outils pour l'utilisation du modèle de données. Plus de 3000 collaborateurs de 80 pays étaient impliqués dans la communauté OHDSI fin 2023 (28). Au delà des développements méthodologiques, la communauté OHDSI a également pu produire des études observationnelles multicentriques internationales (29,30). La Figure 2 illustre le processus de standardisation des données, depuis les bases de données hétérogènes, jusqu'au partage de méthodes et de résultats. La Figure 3 présente le modèle de données OMOP.

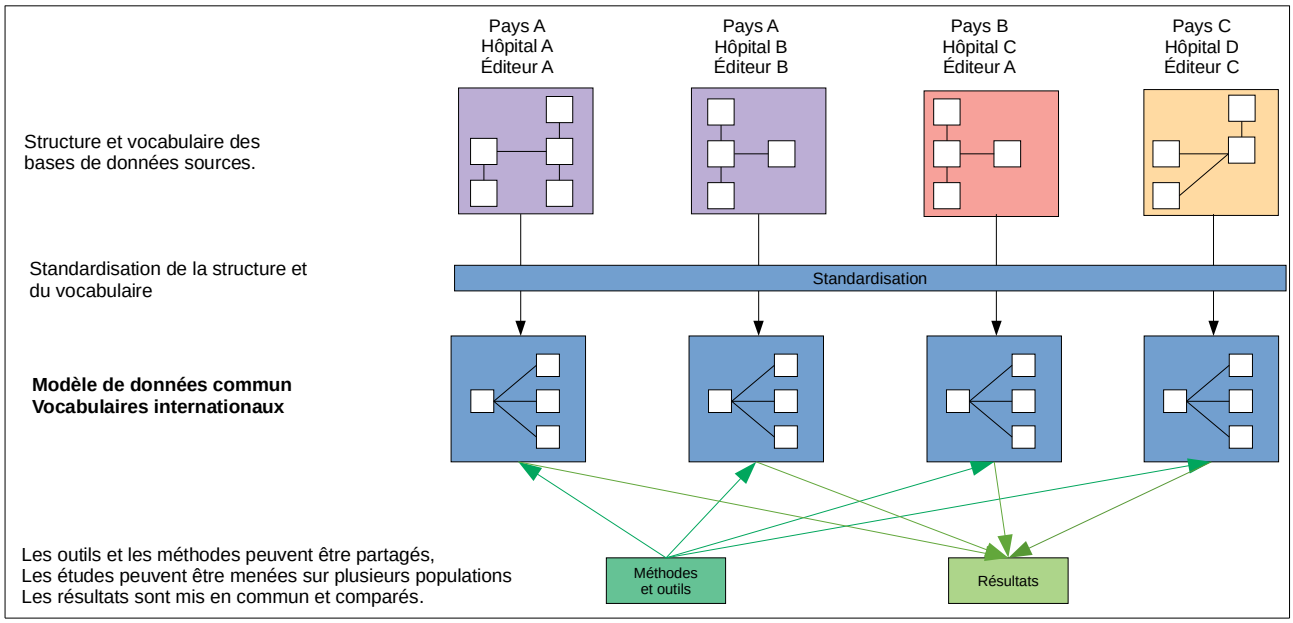


Figure 2: Standardisation des données, depuis les bases de données hétérogènes, jusqu'au partage de méthodes et de résultats

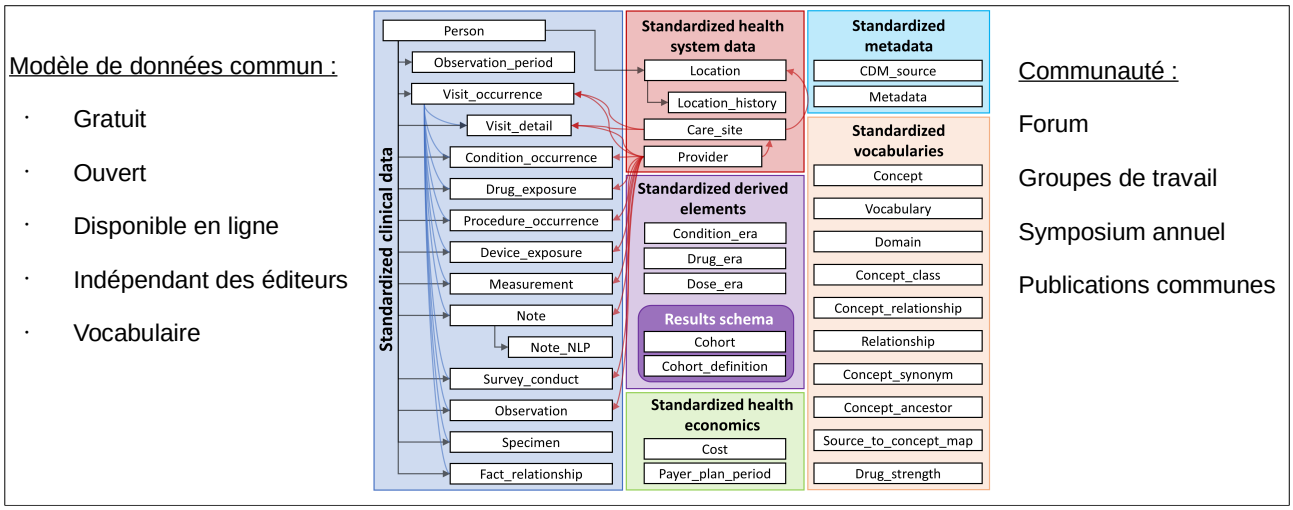


Figure 3: Modèle de données commun OMOP

2.3 Extraction de caractéristiques

Après intégration et standardisation des données brutes au sein de l'EDS, les informations nécessaires pour répondre à la question de recherche ne sont pas toujours disponibles directement, et doivent être calculées dans un second temps à partir des données brutes.

Dans la recherche clinique traditionnelle basée sur des formulaires, cette transformation des données brutes en information analysables par le statisticien est réalisée par la personne en charge du recueil de données. Elle s'appuie donc sur une intelligence humaine, capable de calculs, d'intégration de données hétérogènes multi-sources, et éventuellement d'interprétation. En

réutilisation de données, dès lors qu'on ne dispose que des données brutes telles qu'elles sont disponibles, par exemple, dans le système d'information hospitalier, un des enjeux est de reproduire algorithmiquement ce processus.

Nous pouvons par exemple disposer de toutes les valeurs des mesures pressions artérielles enregistrées au bloc opératoire, pour chacun des patients, alors que ce dont nous aurons besoin est le nombre d'épisodes d'hypotension et la durée totale qui y est associée (Figure 4).

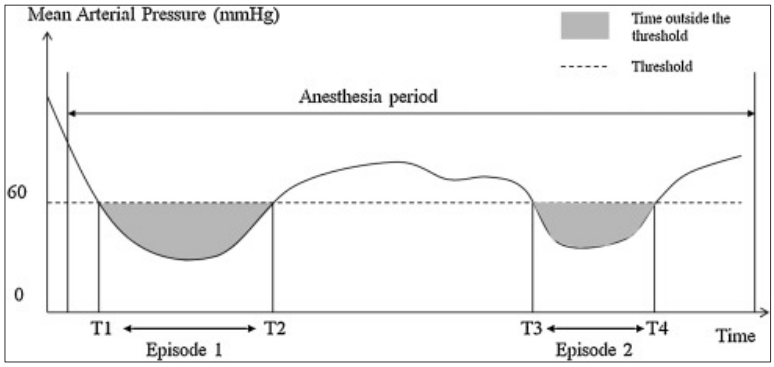
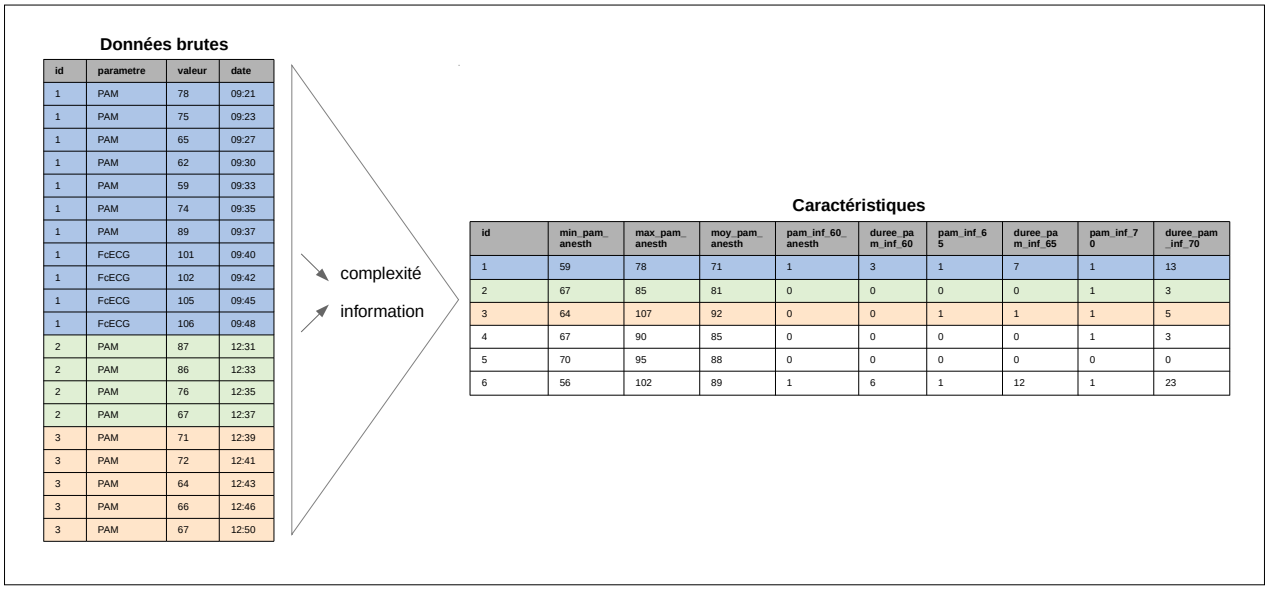


Figure 4: Mesures de pression artérielle et hypotension inférieure à 60 mmHg.

Cette opération est généralement appelée "transformation des données", "agrégation des données" ou encore "extraction de caractéristiques" (31). Ce processus n'est pas trivial et pose des problèmes méthodologiques. En effet, les caractéristiques sont extraites d'une base de données statique (déjà enregistrée et clôturée), pour un grand nombre d'enregistrements, et pour des patients dont l'événement de soins a déjà été réalisé, souvent des années auparavant. Tous les scénarios doivent être pris en considération afin d'éviter la modification individuelle et manuelle des enregistrements extraits, avant l'analyse.



Données brutes

id	parametre	valeur	date
1	PAM	78	09:21
1	PAM	75	09:23
1	PAM	65	09:27
1	PAM	62	09:30
1	PAM	59	09:33
1	PAM	74	09:35
1	PAM	89	09:37
1	FcECG	101	09:40
1	FcECG	102	09:42
1	FcECG	105	09:45
1	FcECG	106	09:48
2	PAM	87	12:31
2	PAM	86	12:33
2	PAM	76	12:35
2	PAM	67	12:37
3	PAM	71	12:39
3	PAM	72	12:41
3	PAM	64	12:43
3	PAM	66	12:46
3	PAM	67	12:50

Caractéristiques

id	min_pam_anesth	max_pam_anesth	moy_pam_anesth	pam_inf_60_anesth	duree_pam_inf_60	pam_inf_65	duree_pam_inf_65	pam_inf_70	duree_pam_inf_70
1	59	78	71	1	3	1	7	1	13
2	67	85	81	0	0	0	0	1	3
3	64	107	92	0	0	1	1	1	5
4	67	90	85	0	0	0	0	1	3
5	70	95	88	0	0	0	0	0	0
6	56	102	89	1	6	1	12	1	23

Figure 5: Transformation des données brutes en caractéristiques

Dans la Figure 5, nous présentons une table typique de mesures enregistrées au bloc opératoire. Chaque patient a plusieurs mesures de pression artérielle moyenne et de fréquence cardiaque. Pour l'analyse statistique, nous aurons besoin, pour chaque patient, de caractéristiques telles que les valeurs minimales, maximales et moyennes de ces deux paramètres. Nous devrons également identifier les patients pour lesquels la pression artérielle moyenne est passée sous 60, 65, 70 mmHg ainsi que les durées cumulées pendant lesquels ils sont restées en dessous de ces seuils. L'extraction de caractéristiques conduit à une réduction de la complexité des données initiales et à une augmentation du nombre d'informations, puisque que nous passons d'un tableau avec plusieurs lignes de données par patient à un tableau avec une ligne par unité statistique de l'étude (ici, le patient) et une caractéristique par colonne.

2.4 Visualisation des données

La visualisation des données est un processus de représentation graphique des informations pour aider à comprendre les tendances, les modèles, et les relations au sein des données (32). La visualisation des données repose sur la production de graphiques qui s'intégreront dans des rapports, des tableaux de bord, des infographies ou d'autres outils d'analyse visuelle (33). Dans le cadre de la réutilisation des données, la visualisation répond à plusieurs objectifs essentiels, aussi bien pour les méthodologistes que pour les utilisateurs finaux. Le Tableau 1 énumère plusieurs de ces objectifs.

Tableau 1: Apports de la visualisation des données

Objectif	Description	Acteurs
Compréhension des données	Comprendre la structure et la distribution des données de santé, facilitant ainsi l'identification de tendances, de schémas ou d'anomalies.	Informaticiens, data engineers, data scientists
Détection d'anomalies	Mettre en évidence des anomalies ou des points aberrants dans les données, contribuant à l'identification d'erreurs de saisie, de valeurs aberrantes ou d'autres problèmes de qualité des données.	Informaticiens, data engineers, data scientists
Validation des modèles	Visualiser les performances des modèles prédictifs ou des algorithmes pour évaluer la qualité des prédictions et identifier les domaines nécessitant des ajustements.	Data scientists, statisticiens, chercheurs
Communication des résultats	Rendre les résultats de l'analyse de données complexes plus accessibles aux parties prenantes non techniques, facilitant ainsi la communication des découvertes et des perspectives.	Chercheurs, professionnels de santé
Interprétation Clinique	Interpréter les résultats des analyses de données d'une manière cliniquement significative, favorisant ainsi une intégration plus directe des insights dans la pratique médicale.	Professionnels de santé
Prise de Décision	Aider les professionnels de la santé à prendre des décisions éclairées en mettant en évidence les informations pertinentes et en permettant une comparaison rapide entre différentes variables.	Décideurs
Surveillance et suivi	Surveiller en continue des indicateurs de santé, afin de détecter rapidement les tendances, les fluctuations ou les ruptures.	Décideurs

2.5 Structuration du mémoire et principaux collaborateurs

Ce mémoire présente mes contributions à la réutilisation des données en santé depuis ma thèse de doctorat soutenue en 2015. Ces travaux de recherche ont été menés au sein de l'Unité Labellisée de Recherche 2694 METRICS au sein de laquelle j'ai préparé mon doctorat. Depuis 2020, j'occupe la fonction de Maître de Conférences Associé. Ces travaux de recherche s'inscrivent dans l'axe 3 de l'ULR 2964 qui a pour thème l'évaluation des technologies de santé et des pratiques médicales en population réelle.

Le Chapitre 3 de ce mémoire s'attache à décrire mes **travaux de recherche méthodologiques** dans le champ de la réutilisation de données.

- La section 3.1 de ce chapitre présente de nouvelles méthodes de **collecte de données** liées, en particulier, aux réseaux sociaux. Ce travail a été réalisé en collaboration avec les étudiants du Master Data Science en Santé à ILIS (Université de Lille), ainsi que mes collègues Mathilde Fruchart, Djamel Zitouni et Benjamin Guinhouya (ULR2694, Université de Lille). Cette partie traite également de la normalisation et de la conservation des données ouvertes destinées à la recherche en psychiatrie, en collaboration avec la F2RSM Psy. Nous présentons aussi le développement et l'évaluation d'un algorithme pour l'appariement d'enregistrements entre les fichiers décès INSEE et les patients des EDS, en collaboration avec les CHUs de Lille, Nantes et Rennes (Vianney Guardiolle, Adrien Bazogue, Morgane Pierre-Jean, Marc Cuggia).
- La section 3.2 concerne l'**intégration des données**, et en particulier la standardisation des données avec le modèle commun OMOP. Ces travaux ont été menés avec les data scientists de l'association InterHop (Adrien Parrot, Nicolas Paris) et la DRESS (Matthieu Doutreligne), les anesthésistes-réanimateurs des CHUs de Lille (Mouhamed Moussa, Benoît Tavernier), de Rouen (Benjamin Popoff), de Saint-Malo (Adrien Parrot) et d'Amiens (Osama Abou-Arab).
- La section 3.3 décrit les travaux qui mettent en œuvre les méthodes d'**extraction de caractéristiques**. Pour cette thématique, nous avons collaboré avec les équipes des EDS du CHU de Rennes (Marc Cuggia, Guillaume Bouzillé), d'Amiens (Thibaut Balcaen, William Gacquer), et de Rouen Benjamin Popoff.
- La section 3.4 porte sur les méthodes de **visualisation de données** pour la diffusion des résultats. Ces projets ont été réalisés en collaboration avec le pôle d'anesthésie-réanimation du CHU de Lille, et l'équipe de coordination du 3114 (Charles-Édouard Notredame).
- La section 3.5 traite des travaux relatifs à l'**ensemble du processus de réutilisation des données**. Nous y abordons la mise en place d'une méthodologie d'exploitation de l'EDS d'anesthésie du CHU de Lille, en collaboration avec la Plateforme d'Aide Méthodologique du CHU de Lille. Nous présentons également les travaux exploratoires que nous avons mené avec l'équipe Inria Magnet sur le calcul décentralisé. Enfin, en partenariat avec la Haute Autorité de Santé, nous dressons un panorama des mises en œuvre d'EDS en France. En capitalisant sur 10 années d'expérience dans l'exploitation des données d'anesthésie du CHU de Lille, nous proposons des recommandations pour la mise en place de ce type de projet.

Le Chapitre 4 présente mes **travaux de recherche appliquée** dans le domaine de la psychiatrie et de la santé mentale, de l'anesthésie-réanimation, de la médecine générale, et d'autres spécialités.

- La section 4.1 présente mes travaux portant sur la **psychiatrie et la santé mentale**, à partir de données issues des bases de données médico-administratives françaises. Ces travaux sont issus de nombreuses collaborations avec des chercheurs cliniciens qui se sont développées au fur et à mesure des rencontres au CHU de Lille, puis à la F2RSM Psy. A titre d'exemple, nous citerons les professeurs Ali Amad et Mathilde Horn, ainsi que les docteurs Thomas Fovet, Maxime Bubrovszky et Marielle Wathelet.
- La section 4.2 décrit nos travaux concernant la ré-utilisation de bases de données médico-administratives pour la **santé publique**. Cette section s'inscrit dans la collaboration avec la cours des comptes et la thèse d'Université de Mathieu Levaillant, co-encadrée avec le professeur Benoît Vallet.
- La section 4.3 résume nos travaux d'autres champs cliniques, comme l'**anesthésie-réanimation** avec le Pôle d'Anesthésie-réanimation du CHU de Lille, les **soins premiers** avec le Département de Médecine Générale de la Faculté de Médecine de Lille et les maisons de santé pluridisciplinaires de Wattrelos, Tourcoing, Lille-Moulins et Guesnain.

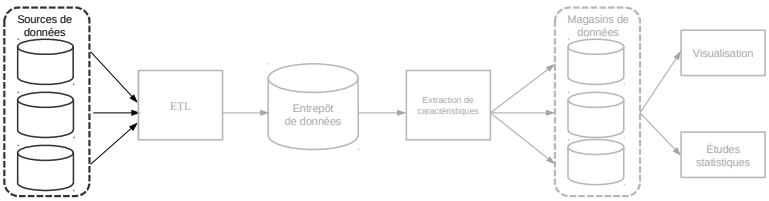
Le Chapitre 5 vient conclure ce mémoire et présente mes perspectives de recherche, à la fois dans le domaine de la recherche méthodologique que dans le cadre de la recherche appliquée.

Chapitre 3 Travaux méthodologiques

Sommaire

Ce chapitre récapitule mes contributions méthodologiques dans le domaine de la réutilisation des données. La section 3.2 expose mes travaux relatifs à la collecte de données. La section 3.3 traite de l'intégration des données brutes au sein d'EDS, ainsi que de la standardisation structurelle et sémantique visant à faciliter leur réutilisation. La section 3.4 se concentre sur l'extraction de caractéristiques. La section 3.5 présente mes travaux de visualisation de données, visant à faciliter la présentation des résultats. Enfin, la section 3.6 expose mes contributions au processus global de la réutilisation des données.

3.1 Collecte des données

Étape :	
Publications :	Laurent G, Guinhouya B, Whatelet M, **Lamer A**. Automatic Exploitation of YouTube Data: A Study of Videos Published by a French YouTuber During COVID-19 Quarantine in France. Stud Health Technol Inform. 2020 Nov 23;275:112–6.
	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 May 25;294:705–6.
	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 May 21;9:20552076231176114.
	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 May 18;302:502–3.
	Saint-Dizier C, **Lamer A**, Zaanouar M, Amariei A, Quindroit P. OpenDataPsy: An Open-Data Repository with Standardized Storage and Description for Research in Psychiatry. Stud Health Technol Inform. 2023 May 18;302:851–5.
	Martignene N, Amad A, Bellet J, Tabareau J, D'Hondt F, Fovet T, **et al**. Goupile: A New Paradigm for the Development and Implementation of Clinical Report Forms. Stud Health Technol Inform. 2022 May 25;294:540–4.
	Guardiolle V, Bazoge A, Morin E, Daille B, Toublant D, Bouzillé G, **et al**. Linking Biomedical Data Warehouse Records With the National Mortality Database in France: Large-scale Matching Algorithm. JMIR Med Inform. 2022 Nov 1;10(11):e36711.

La première étape du processus de réutilisation des données consiste à identifier les sources de données et à en extraire les données pertinentes. Les données peuvent provenir de sources très diverses, telles que : (i) des logiciels développés par des entreprises ou des logiciels en open source, (ii) des applications de bureau, web ou mobile, (iii) des logiciels propriétaires, open source ou des données ouvertes, entre autres. Lorsqu'elles sont structurées sous forme de tableaux, leur collecte présente généralement peu de difficultés, à l'exception de la gestion des autorisations d'accès aux logiciels propriétaires. Ces données sont habituellement intégrées dans les entrepôts de données. Cependant, les obstacles rencontrés lors de leur intégration et de leur exploitation, seront abordés dans les sections 3.2, 3.3 et 3.4.

Les réseaux sociaux ont ouvert la voie à de nouveaux types de données non structurées, offrant des perspectives riches et diverses pour la compréhension des comportements humains et des tendances. Ces données incluent des éléments tels que les publications textuelles, les commentaires, les images, les vidéos, ainsi que des interactions sous forme de *likes*, partages ou commentaires. Leur nature non structurée rend leur traitement et leur analyse plus complexes, nécessitant souvent des approches avancées de traitement du langage naturel, de l'apprentissage automatique et de l'analyse de données multimédias. Nous aborderons ces nouvelles sources de données dans la section 3.1.1.

3.1.1 Réseaux sociaux

Cette dernière décennie, l'utilisation des réseaux sociaux a explosé. Ces nouveaux médias permettent aux utilisateurs de partager rapidement leurs émotions, avis ou intérêts pour des situations, de discuter de sujets d'actualité ou de leur état de santé. Les données issues des réseaux sociaux pourraient compléter les données de santé collectées habituellement dans les logiciels hospitaliers du fait qu'elles sont recueillies en dehors du contexte de soins, souvent axées sur le quotidien, et partagées de manière plus directe et spontanée. Cependant, de par leur volumétrie, leur vélocité et leur variété, elles présentent des difficultés pour leur réutilisation.

Tableau 2: Catégories des réseaux sociaux

Catégorie	**Réseaux sociaux**
Généralistes	Facebook, Twitter
Professionnels	LinkedIn, Viadeo
Visuels	Instagram, Pinterest
Partage de vidéos	YouTube, TikTok, Snapchat
Messagerie	WhatsApp, Messenger, Telegram, Signal
Thématiques	Reddit, Goodreads, Babelio
Partage d'informations	Twitter

Les réseaux sociaux produisent des données, telles que des messages sous format texte, image ou vidéo. Ces données sont complétées par un ensemble de méta-données qui vont renseigner sur le contexte du message, avec sa date de publication et l'utilisateur qui l'a publié, mais également les réactions qu'ont généré la publication avec un nombre de commentaires, de réactions positives ou négatives, et des repartages. Bien que lisibles pour l'humain grâce au réseau social lui-même, ces données et métadonnées ne sont pas directement récoltables et exploitables pour le chercheur. Il faudra pour cela employer des méthodes d'extraction et d'exploitation automatiques que nous verrons dans les paragraphes suivants. Le Tableau 2 liste les réseaux par catégories d'usage. L'exploitation des données des réseaux sociaux suit plusieurs étapes, décrites dans la Figure 6 et dans les paragraphes suivants.

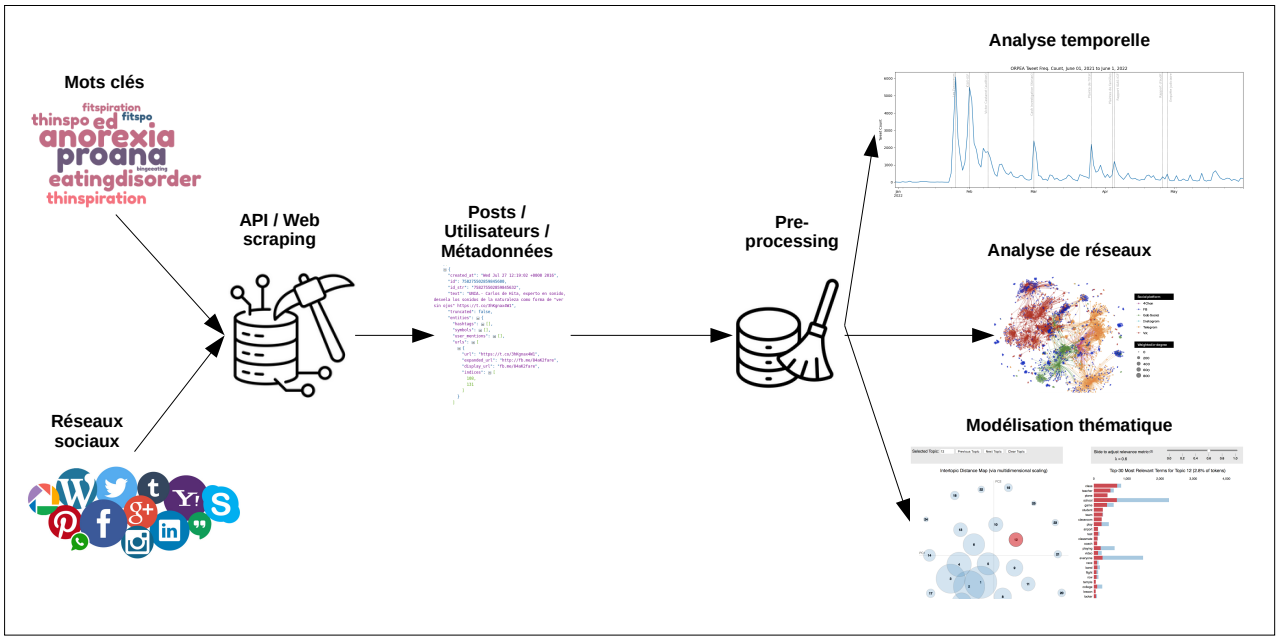


Figure 6: Processus d'exploitation des données issues des réseaux sociaux

3.1.1.1 Extraction

Les deux principales méthodes pour collecter des données sur les médias sociaux sont l'Interface de Programmation d'Application (API) et le web scraping.

Une API est une interface informatique mise en place par le réseau social, permettant un accès aux données non pas par visualisation des pages, mais par interrogation directe du site. Les API, après inscription et obtention des droits d'accès pour certaines d'entre elles, permettent d'interroger et de collecter des données à partir de requêtes prédéfinies, comme par exemple extraire des publications liées à un mot-clé, obtenir les commentaires et les indicateurs d'une publication, identifier les relations d'un profil. La plupart des API proposent des options gratuites et payantes. Les éditions gratuites n'offrent qu'un nombre limité de requêtes et un historique de recherche limité (par exemple, 7 jours pour TwitterAPI), tandis que les éditions payantes facturent des volumes supplémentaires, un historique complet et des fonctionnalités plus avancées.

Pour surmonter les limitations ou l'absence d'API officielles, il est possible d'automatiser une extraction d'informations, à partir du code HTML des pages qu'un humain pourrait visiter. Ceci suppose d'une part de charger des pages entières en masse, éventuellement de suivre des liens hypertexte au sein de ces pages, et ensuite d'identifier les informations pertinentes dans ces pages et les ranger dans des structures de données. L'ensemble de ce processus s'appelle le *web scrapping*. Cette technique adopte une approche différente des API, en simulant la navigation sur un site web et en extrayant le contenu des pages web en utilisant les balises HTML (34). Certains webscrappers spécifiques ont également été développés : Tweepy et snscrape pour Twitter, facebook-scraper pour Facebook, et instascrape pour Instagram. Avec cette méthode, il est possible de récupérer de la page web tout ce qui est visible pour l'utilisateur.

Que ce soit à travers des API ou le web scraping, les données sont renvoyées au format JSON, dans des formats différents selon le réseau social.

3.1.1.2 Pre-processing

Après extraction, les messages sont nettoyés en suivant plusieurs méthodes de traitement du langage naturel (NLP) : les accents, caractères spéciaux, majuscules et *stop words* (e.g., le, la, et, ou) sont supprimés, puis les mots sont ramenés à leur forme la plus simple, quels que soient leurs accords et déclinaisons. Le Tableau 3 présente le résultat du pre-processing de deux messages issus de Twitter.

Tableau 3: Pre-processing du texte brut issu d'un réseau social

Message original	**Message nettoyé et préparé pour analyse**
Ma mère comprend pas d'où viennent mes TCA ...	mere / comprendre / pas / ou / venir / TCA
Je viens de me goinfrer de tartes au sucre, je suis en pleine montée de glycémie. Putain de boulimie 😈	je / venir / se goinfrer / tarte / sucre / je / etre / plein / monte / glycemie / putain / boulimie

3.1.1.3 Analyses

Plusieurs types d'analyses peuvent ensuite être conduites, à la fois sur les données et les métadonnées des publications : (i) une description simple des messages et comptes (e.g., longueur des tweets, nombre de réactions, nombre de partage, nombre d'abonnées, etc.), (ii) une description de la dynamique de publication (e.g., nombre de tweets par jour, par heure), (iii) une analyse du réseau avec de la détection de communauté entre utilisateurs, et (iv) une modélisation thématique pour identifier et catégoriser automatiquement les thèmes prépondérants dans un ensemble de messages. Dans ces approches courantes, les contenus non-textuels ne sont pas encore analysés (e.g., images, contenu des vidéos).

Nous présenterons ci-dessous des exemples de collecte et d'exploitation de données issues réseaux sociaux.

3.1.1.4 YouTube - HugoDecrypt

Nous avons testé la faisabilité de l'extraction et de l'exploitation automatiques des données de YouTube (35). Pour cela, nous avons identifié les principaux thèmes dans les vidéos du youtubeur français « HugoDécrypte », et les réactions du public à ces sujets. Nous avons analysé les vidéos de la playlist "actualités quotidiennes" entre mars et mai 2020 pour suivre l'évolution des actualités sur la COVID-19 en France.

Sur cette période, 49 vidéos liées à la flambée de COVID-19 ont été publiées dans cette playlist. Ces vidéos ont reçu 38 725 commentaires dans les 24 heures suivant leur publication. Pour chaque vidéo, le nombre médian [Q1;Q3] de commentaires était de 771 [682 ; 865]. 135 sujets uniques ont été identifiés. Les 10 sujets les plus discutés sont présentés dans la Figure 7. Les deux sujets avec le pourcentage le plus élevé de commentaires positifs ou neutres étaient "merci" et "Hugo", avec respectivement 30 % et 27 %. Le travail du youtubeur a été bien accueilli par les abonnés, tandis que les sujets abordés ont une polarité négative.

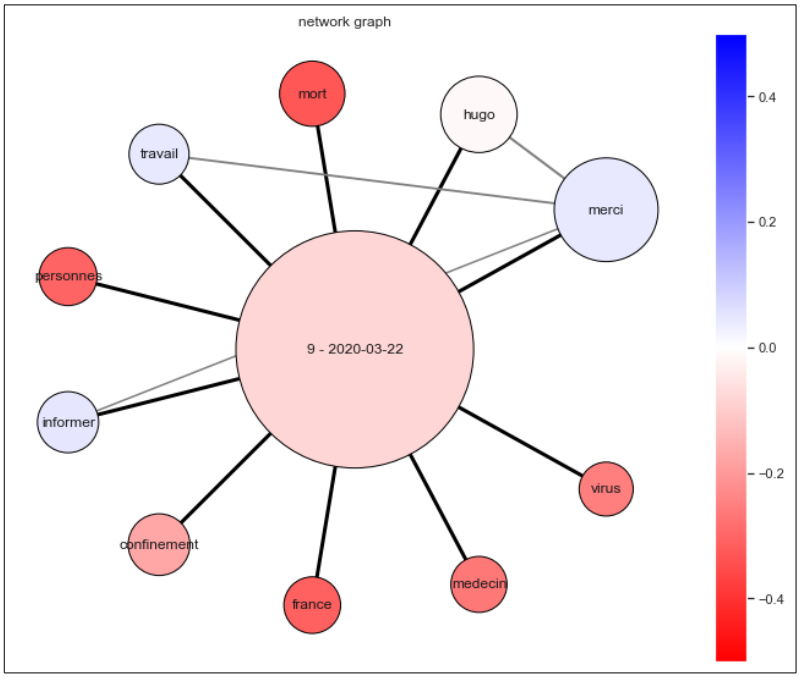


Figure 7: Polarité des thèmes abordés dans les commentaires des vidéos d'HugoDécrypt (35)

3.1.1.5 Twitter – Tableau de bord automatisé

Lors d'un second travail, nous avons développé un tableau de bord pour permettre aux utilisateurs novices d'extraire et d'analyser facilement et de manière autonome les données de Twitter (36). Nous avons évalué leur capacité à l'utiliser et recueilli leurs commentaires sur la valeur d'un tel outil pour leur pratique quotidienne. Le tableau de bord permet à l'utilisateur de créer un jeu de données, sans avoir à coder, en fonction de quatre critères : la requête textuelle, la plage de temps souhaitée, le nombre de tweets à récupérer et la langue des tweets.

Le tableau de bord propose trois onglets avec (i) une navigation à travers le jeu de données et son exportation au format CSV, (ii) une analyse temporelle avec le décompte des tweets pour chaque jour et pour chaque heure de la journée, et (iii) une analyse textuelle avec un comptage des mots et un nuage de mots. Nous avons proposé cet outil à des personnes qui ne maîtrisent pas la programmation (médecins, pharmaciens et *community manager*), en leur donnant pour instruction de mener des recherches avec un mot-clé lié à leur domaine d'intérêt.

Figure 8: Tableau de bord automatisé pour l'analyse des données de Twitter (36)

Les figures 8 et 9 illustrent un cas d'usage sur les salles de shoot, avec le nombre de tweets par jour, le nombre de tweets par heure, et un nuage de mots avec la fréquence des termes contenus dans les tweets. Il met en évidence les villes où les sites d'injection supervisée sont ouverts ou en cours d'ouverture.

Figure 9: Résumé des thèmes de discussion présents dans les tweets traitant des salles de shoot (36)

L’outil que nous proposons permet d’extraire automatiquement des tweets et fournit des statistiques descriptives. Il peut être utilisé pour suivre les tendances des sujets actuels discutés sur Twitter. Avec son module de recherche, il n'est pas nécessaire de programmer.

3.1.1.6 Twitter - Scandale Orpea

Le scandale Orpea a débuté le 24 janvier 2022 avec la publication d'un livre-enquête écrit par le journaliste français Victor Castanet. Le livre, intitulé "Les Fossoyeurs", a révélé des mauvais traitements infligés aux personnes âgées résidant dans des maisons de retraite. Ce scandale a provoqué un mouvement politique qui a dénoncé d'autres groupes de maisons de retraite (comme Korian). Les objectifs de cette étude étaient d'étudier les tendances temporelles et la dynamique des publications pendant le scandale d'Orpea, ainsi que d'identifier les principaux sujets de discussion sur Twitter.

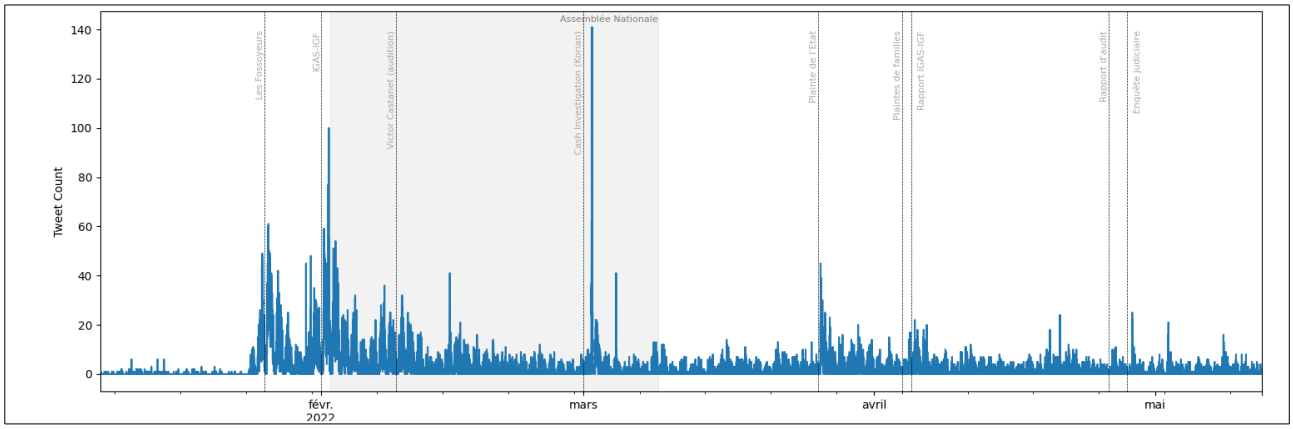


Figure 10: Distribution des publications liées au scandale Orpéa sur Twitter (78)

Nous avons extrait des tweets contenant les mots clés liés au scandale Orpea (« Orpea », « Cash Investigation », etc.) du 1er juin 2021 au 1er juin 2022. Nous avons extrait 85 342 tweets uniques liés au scandale. Les principaux pics concernent la publication de "Les Fossoyeurs", puis l'enquête menée par les autorités françaises (IGAS-IGF) et enfin l’émission Cash Investigation (Figure 10). La plupart des tweets ont été publiés entre 6h et 12h (n = 22 662, 26%). Nous avons identifié 6 clusters de tweets à partir du modèle de l'allocation latente de Dirichlet non supervisé (37) (Tableau 4). Nous avons identifié 389 tweets publiés par le compte @Orpea_ entre septembre 2019 et novembre 2022. La plus forte activité du compte Orpea a eu lieu en juin 2022 avec 15 tweets le 01/06/2022. En revanche, il y a une absence d'activité aux dates du scandale.

Tableau 4: Sujets identifié par allocation de Dirichlet latente

Cluster	Pourcentage de Tweets	Label attribué	Exemples de mots contenus dans le sujet
1	21,0 %	Influence politique et volonté d'action	Faire, avoir, pouvoir, Olivier Veran, Brigitte Bourguignon
2	18,2 %	Réactions sur le fonctionnement d'Orpea	Personne, âge, public, résident, personnel, actionnaire, argent, payer
3	17,0 %	Avis des proches et soignants	Aller, bien, place, éthique, retraire
4	15,6 %	Médiatisation du scandale Orpea	Orpea, scandale, enquête ,affaire, état, gouvernement, rapport, directeur
5	15,2 %	Dénonciation des maltraitance	Orpea, groupe, livre, maltraitance, action, enquête, dénoncer, plainte
6	13,0 %	Impact du scandale Orpea sur le groupe Korian	Korian, Cash Investigation, question, soir, Elise Lucet

3.1.2 Forums

Après YouTube et Twitter, nous nous sommes intéressés aux forums en ligne. Alors que Twitter et Youtube proposent du contenu limités en taille ou au format vidéo, les forums offrent un espace pour des discussions plus approfondies et des échanges spécifiques sur des sujets particuliers. Les discussions peuvent être manière anonyme, ce qui encourage le partage d'expériences personnelles plus détaillées et authentiques.

Notre étude visait à explorer les discussions portant sur l'endométriose afin d'identifier les signes précoces associés à cette pathologie (38). Nous avons extrait automatiquement les posts de 10 forums français abordant des discussions liées à l'endométriose. Après nettoyage des données, nous avons identifié les symptômes évoqués dans les posts et les avons connectés au dictionnaire de termes médicaux MedDRA. Nous avons également relevé les marqueurs temporels évoqués à proximité des symptômes pour identifier les symptômes précoces.

Au total, nous avons collecté 7148 fils de discussion et 78 905 messages (Figure 11).

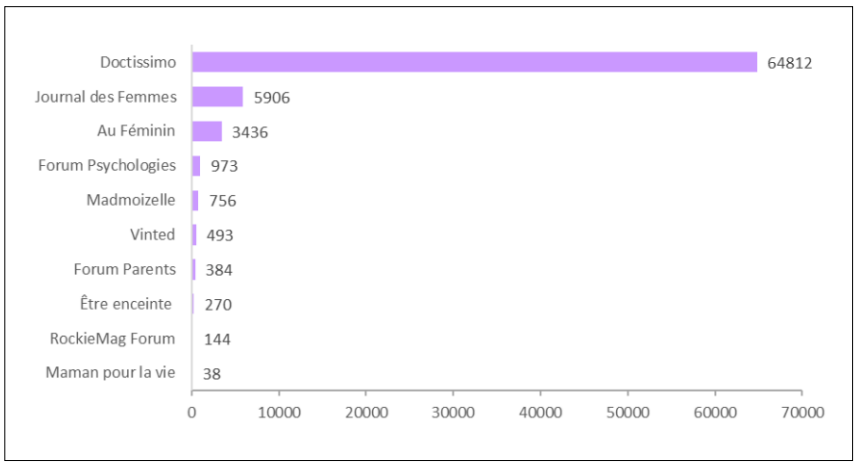


Figure 11: Nombre de messages extraits par forum(38)

Nous avons extrait 41 groupes de symptômes contextualisés, dont 20 groupes de symptômes précoces associés à l'endométriose (Figure 12). Parmi ces groupes de symptômes précoces, 13 se sont avérés représenter des signes déjà connus d'endométriose. Les 7 groupes de symptômes précoces restants sont les suivants : œdème des membres, douleurs musculaires, névralgies, hématurie, démangeaisons vaginales, altération de l'état général (i.e. vertiges, fatigue, nausées) et bouffées de chaleur. Les présents résultats offrent la possibilité d'explorer plus avant les processus biologiques précoces qui déclenchent cette maladie.

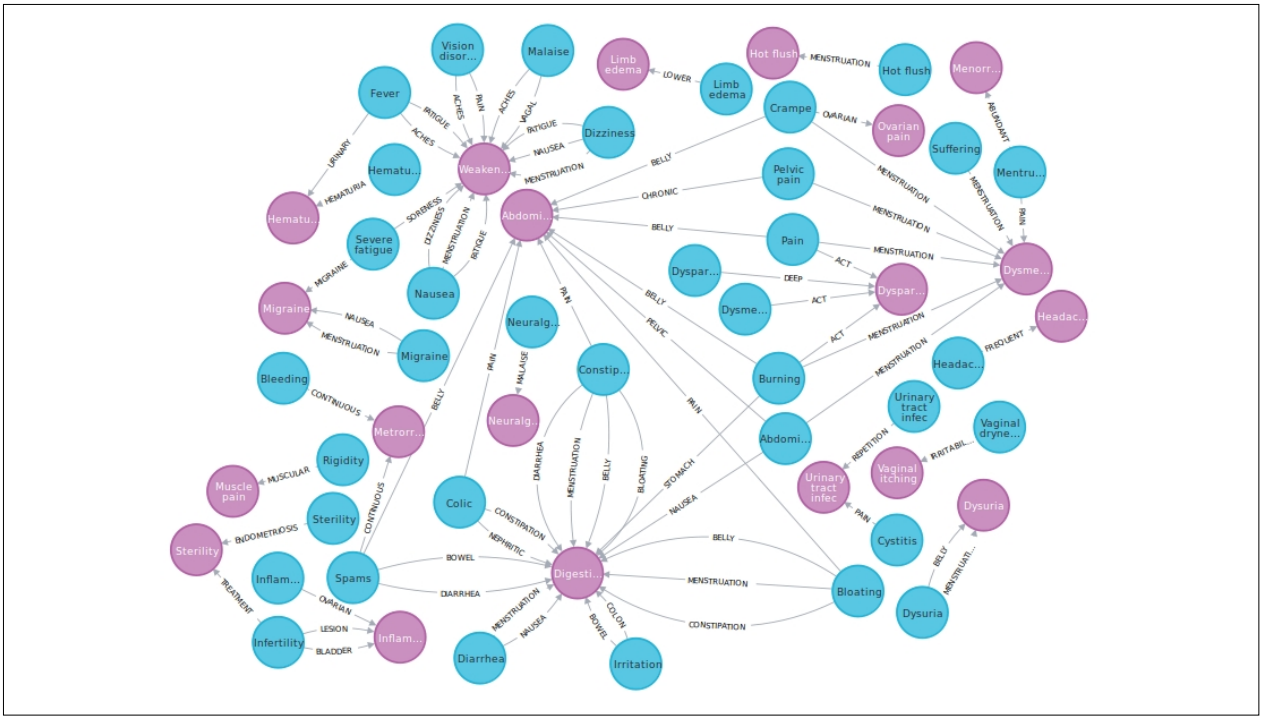


Figure 12: Symptômes précocement détectés associés aux symptômes annotés avec les mots de contextualisation (38). Nœuds bleus : symptômes précocement détectés. Arêtes : mots de contextualisation associés au symptôme. Nœuds violets : classes de symptôme annoté.

3.1.3 Données ouvertes

Depuis quelques années, nous observons une tendance mondiale vers l'ouverture des données, un mouvement "open data", motivé par le désir d'accroître la transparence et de faciliter la réutilisation des informations. En France, de nombreuses institutions ont adopté cette approche en mettant à disposition leurs données, notamment l'Insee.

Le partage de données liées à la santé offre la possibilité d'éviter la duplication des efforts lors de la collecte de données, de réduire les coûts inutiles pour les futures études, et de promouvoir la collaboration ainsi que la diffusion des données au sein de la communauté scientifique. Plusieurs organismes nationaux et équipes de recherche ont créé des dépôts de données accessibles au public (e.g., http://data.gouv.fr, http://opendata.cern.ch). Ces dépôts sont souvent organisés en fonction de critères spatiaux ou temporels, ou sont spécifiquement consacrés à des domaines particuliers. Ils partagent des caractéristiques et des fonctionnalités communes, comme un moteur de recherche, des métadonnées décrivant les jeux de données, le suivi des versions, ainsi que la

génération d'un Identifiant de Document Numérique (DOI). Cependant, il n'existe pas actuellement de norme universelle pour le stockage et la documentation des jeux de données.

L'objectif de cette étude était de mettre en place un système normalisé de stockage et de description pour des jeux de données ouvertes destinés à la recherche (39). Pour ce faire, nous avons sélectionné huit jeux de données, couvrant la démographie, l'emploi, l'éducation et la psychiatrie. Nous avons analysé les formats, les nomenclatures (i.e., noms de fichiers et de variables, modalités des variables qualitatives récurrentes), ainsi que les descriptions de ces jeux de données. Nous avons ensuite proposé un format et une description normalisés et uniformes pour ces jeux de données (Tableau 5).

Nous avons mis ces jeux de données à disposition dans un dépôt GitLab public. Chaque jeu de données est accompagné des éléments suivants : le fichier de données brutes dans son format d'origine, le fichier de données nettoyées au format CSV, une description détaillée des variables, un script de data management, ainsi que des statistiques descriptives. Les noms de ces fichiers sont standardisés et présentés dans le Tableau 6. Les statistiques descriptives sont générées automatiquement en fonction des types de variables documentées précédemment (Tableau 7).

Tableau 5: Documentation du fichier README

Section	Description
Description	Fichier source, sans modification
Aperçu	Script de data management pour nettoyer et standardiser le fichier brut
Licence et condition d'évaluation	Fichier nettoyé et standardisé
Fichier source	Lien vers la page de téléchargement du fichier source
Références	Liens vers des exemples d'utilisation du jeu de données (articles scientifiques, rapports, sites web, etc.)

Tableau 6: Nom et description normalisés des fichiers disponibles pour chaque ensemble de données

Fichiers	Description
dm_XXX_annee.R	Script de data management pour nettoyer et standardiser le fichier brut
head_XXX_annee.png	Aperçu des premières lignes
raw_XXX_annee.xlsx	Fichier source, sans modification
cleaned_XXX_year.csv	Fichier nettoyé et standardisé
desc_XXX_year.html	Statistiques descriptives
variables_XXX_year.csv	Cahier de variables (nom technique, libellé lisible, type de variable)

Tableau 7: Indicateurs statistiques et graphiques appropriés à chaque type de variables

Type de variable	Indicateurs statistiques	Graphiques
Binaire	Nombre, pourcentage, pourcentage de valeurs manquantes	Doughnut
Qualitative	Nombre, pourcentage, pourcentage de valeurs manquantes	Diagramme en barres
Quantitative continue	Quartiles, min, max, pourcentage de valeurs manquantes	Histogramme et courbe de densité
Quantitative discrète	Quartiles, min, max, pourcentage de	Diagramme en barres

	valeurs manquantes	
Date	Min, max, nombre d'événements par date avec quartiles	-

3.1.4 Données décès

Depuis 2019, l'Insee met à disposition les fichiers des personnes décédées (40). Ils contiennent les identités des personnes décédées depuis 2010, avec les dates et communes de naissance et de décès, ainsi que les nom, prénoms et nom marital. Ces données sont extrêmement précieuses pour compléter les informations des EDS hospitaliers, limitées à la durée du séjour et aux réadmissions au sein du même établissement hospitalier. Les données de l'Insee permettent d'obtenir des informations sur le statut vital des patients en dehors du cadre hospitalier, même lorsqu'ils sont pris en charge dans d'autres structures médicales.

Nous avons développé un algorithme d'appariement déterministe basé sur un préparation avancée des données à partir d'une connaissance du système de dénomination et de la distance de Damerau-Levenshtein (DLD) (41). Ainsi, nous avons relevé que dans le cas des personnes avec un prénom composé ou un deuxième prénom, il pouvait y avoir eu des combinaisons différentes de ces prénoms dans les EDS hospitaliers. Après un nettoyage élémentaire (i.e., passage en minuscules, éliminations des caractères spéciaux et accents), nous avons créé plusieurs prénoms candidats avec la première partie du prénom composé (1er prénom 1), le prénom composé en entier (1er prénom 1-2) et la concaténation des 1er et 2ème prénoms (1er et 2ème prénoms) (Tableau 8).

Tableau 8: Transformation des prénoms avant appariement

Données initiales		**Données**		
1er prénom	**2ème prénom**	**1er prénom 1**	**1er prénom 1-2**	**1er et 2ème prénoms**
Jean	N/A	jean	jean	jean
Marie	Claire	marie	marie	marieclaire
Pierre-Olivier	Christian	pierre	pierreolivier	pierreolivierchristian
Elon-Louis	N/A	elon	elonlouis	elonlouis

La performance de l'algorithme a été évaluée de manière indépendante en utilisant les données des EDS de 3 hôpitaux universitaires : Lille, Nantes et Rennes. La spécificité a été évaluée avec les patients vivants au 1er janvier 2016 (c'est-à-dire les patients ayant au moins une rencontre hospitalière avant et après cette date). La sensibilité a été évaluée avec les patients enregistrés comme décédés entre le 1er janvier 2001 et le 31 décembre 2020. L'algorithme basé sur la DLD a été comparé à un algorithme d'appariement direct avec un nettoyage minimal des données comme référence. En combinant l'ensemble des centres, la sensibilité était de 11 % plus élevée pour l'algorithme basé sur la DLD (93,3 %, IC à 95 % 92,8-93,9) que pour l'algorithme direct (82,7 %, IC à 95 % 81,8-83,6 ; $p < 0,001$). Cela montre l'importance du nettoyage des données et de la connaissance d'un système de dénomination. L'utilisation d'un algorithme d'appariement déterministe tel que l'algorithme basé sur la DLD est un exemple de combinaison de données externes en open source pour améliorer la valeur des EDSs.

3.1.5 Goupile

Malgré l'informatisation croissante des soins de santé, certains segments du parcours patient sont encore documentés au format papier. Ainsi, il arrive que, dans une recherche sur les données (et non sur les personnes), certaines données complémentaires nécessitent toutefois d’être collectées par des formulaires et non par réutilisation directe de données structurées.

Trois outils principaux permettent de collecter manuellement ces données pour compléter les bases de données informatisées : les tableurs, les eCRF (Case Report Form), et les applications de saisie développées au cas par cas. Les tableurs, comme les logiciels de bureautique classiques, sont largement utilisés pour leur accessibilité, mais ils peuvent être modifiés accidentellement et ne génèrent pas de données normalisées. Les eCRF offrent une saisie et une validation conviviales mais ont des fonctionnalités limitées et les coûts de licence pourraient poser problème pour les projets scientifiques. Enfin, les applications personnalisées offrent des fonctionnalités illimitées mais exigent des compétences avancées en programmation et prennent du temps à développer.

Nous avons développé un outil de conception d'eCRF open source (libre et gratuit, placé sous licence AGPL v3) qui s'efforce de rendre la création de formulaires et la saisie de données à la fois puissantes et faciles (42). Cet outil se base sur un nouveau paradigme pour concevoir les eCRF. Nous proposons, dans un langage de programmation usuel (Javascript), des fonctions prédéfinies permettant de créer le formulaire instantanément, et en le déployant immédiatement pour test ou utilisation (Figure 13). Il dispose également d’un panneau de suivi des saisies (Figure 14).

Comme tout édtieur de formulaires, Goupile peut également être utilisé pour des recherches cliniques plus traditionnelles, qui ne s’appuient pas nécessairement sur la réutilisation de données.

À ce jour, Goupile a été utilisé pour 30 projets, dont 3 projets multi-centres. La conception de Goupile était suffisamment souple pour nous permettre de mettre en place les 17 modules de l'évaluation MINI pour le DSM-V en ligne, malgré la complexité du flux des questions. Nous avons réussi à conduire une étude interactive d'envergure nationale, incluant plusieurs évaluations neuropsychologiques où les utilisateurs devaient visionner plusieurs vidéos et photos, puis répondre à des questions qui leur étaient associées.

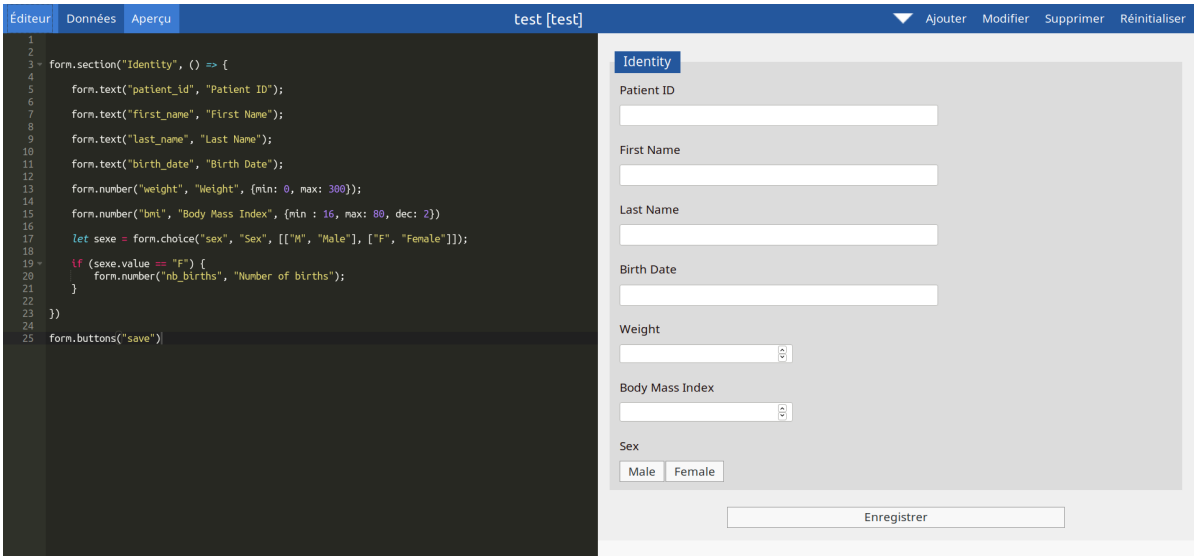


Figure 13: Éditeur de formulaire (à gauche) et interface de saisie (à droite) de Goupile.

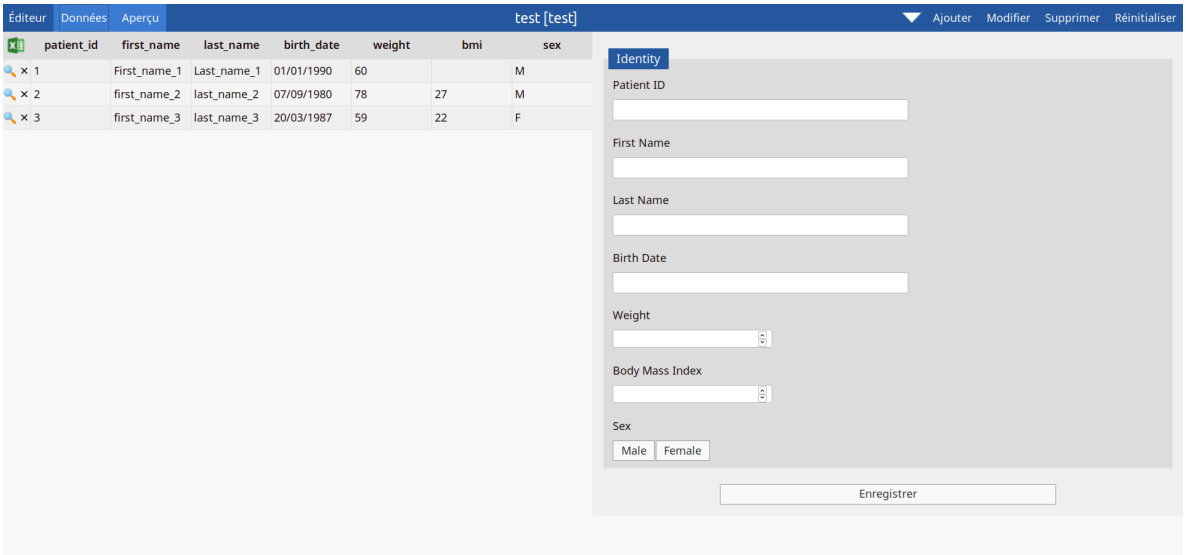
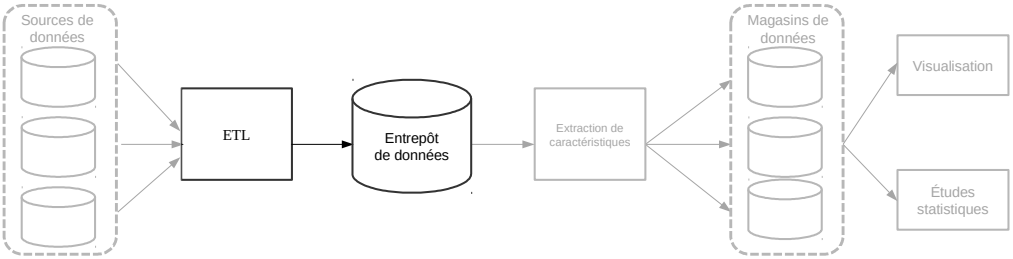


Figure 14: Suivi des enregistrements (à gauche) et interface de saisie (à droite) de Goupile

3.2 Intégration des données

Étape :	
Publications :	Quindroit P, Fruchart M, Degoul S, Périchon R, Soula J, Marcilly R, **et al**. Definition of a practical taxonomy for referencing data quality problems in healthcare databases. Methods Inf Med. 2022 Nov 10
	Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. Appl Clin Inform. 2020 Jan;11(1):13–22.
	Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart JB, et al. Transforming Anesthesia Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study. J Med Internet Res. 2021 Oct 29;23(10):e29259.
	Paris N, **Lamer A**, Parrot A. Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study. JMIR Med Inform. 2021 Dec 14;9(12):e30970.

L'intégration des données consiste à rassembler des ensembles de données disparates et souvent hétérogènes provenant de différentes sources. Cette étape vise à créer un environnement où ces données variées peuvent être réunies, organisées et rendues compatibles. Elle implique plusieurs actions telles que l'identification, la collecte, la fusion, la normalisation et l'organisation des données provenant de sources multiples et parfois incompatibles. L'objectif principal est de créer une base de données unifiée et cohérente, prête à être analysée et exploitée pour répondre à des besoins de recherche spécifiques ou à des objectifs d'analyse de données.

3.2.1 Qualité des données

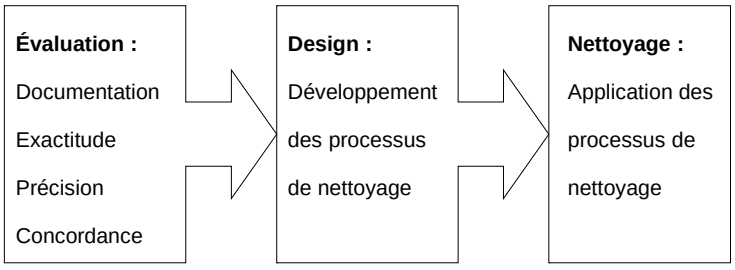


Figure 15: Processus de nettoyage des données

Les bases de données transactionnelles contiennent des incohérences dues à des erreurs de saisie ou à une mauvaise documentation de la part des utilisateurs, à des artefacts de surveillance provenant des moniteurs, ou à l'hétérogénéité de la structure des bases de données. Par

conséquent, l'exploitation de ces données peut conduire à des résultats erronés. La qualité des données peut être gérée par le processus de nettoyage des données. Ce type de processus se compose de trois étapes principales décrites dans la figure 15.

Même si la problématique de la qualité des données est fréquemment abordée en informatique médicale et dans d'autres domaines, il n'existait pas, à notre connaissance, de taxonomie des problèmes de qualité des données qui abordait tous les types possibles de problèmes techniques de l'enregistrement unique aux multiples sources de données, y compris les problèmes liés à la structure des données et aux valeurs elles-mêmes.

Le problème de la qualité des données est complexe, et va des problèmes formels identifiables dans les données (ex : type de valeur illégal), aux incohérences métier individuelles (ex : cancer de la prostate chez une femme) ou statistiques (ex : proportion d'hommes de 75 % en gériatrie). Dans ce premier travail, nous avons choisi de nous concentrer sur les problèmes formels, qui sont un obstacle aux opérations suivantes.

Nous avons recherché dans la littérature scientifique les publications une liste ou une taxonomie des problèmes de qualité des données. Le principal critère d'inclusion était de proposer une définition pratique et/ou une illustration permettant d'identifier les problèmes de qualité des données. Nous nous sommes concentrés sur les problèmes techniques, à l'exclusion des problèmes liés à la confidentialité, à l'autorisation d'accès, à la sécurité, etc. Au final, 286 cas de problèmes de qualité des données ont été extraits de 12 publications. En comparant et synthétisant les travaux existants, nous avons pu proposer une taxonomie opérationnelle des problèmes de qualité des données, illustrée par des exemples concrets tirés de bases de données cliniques (43). Cette taxonomie est organisée suivant les 7 niveaux de granularité illustrés dans la figure 16, et contient 53 problèmes de qualité de données.

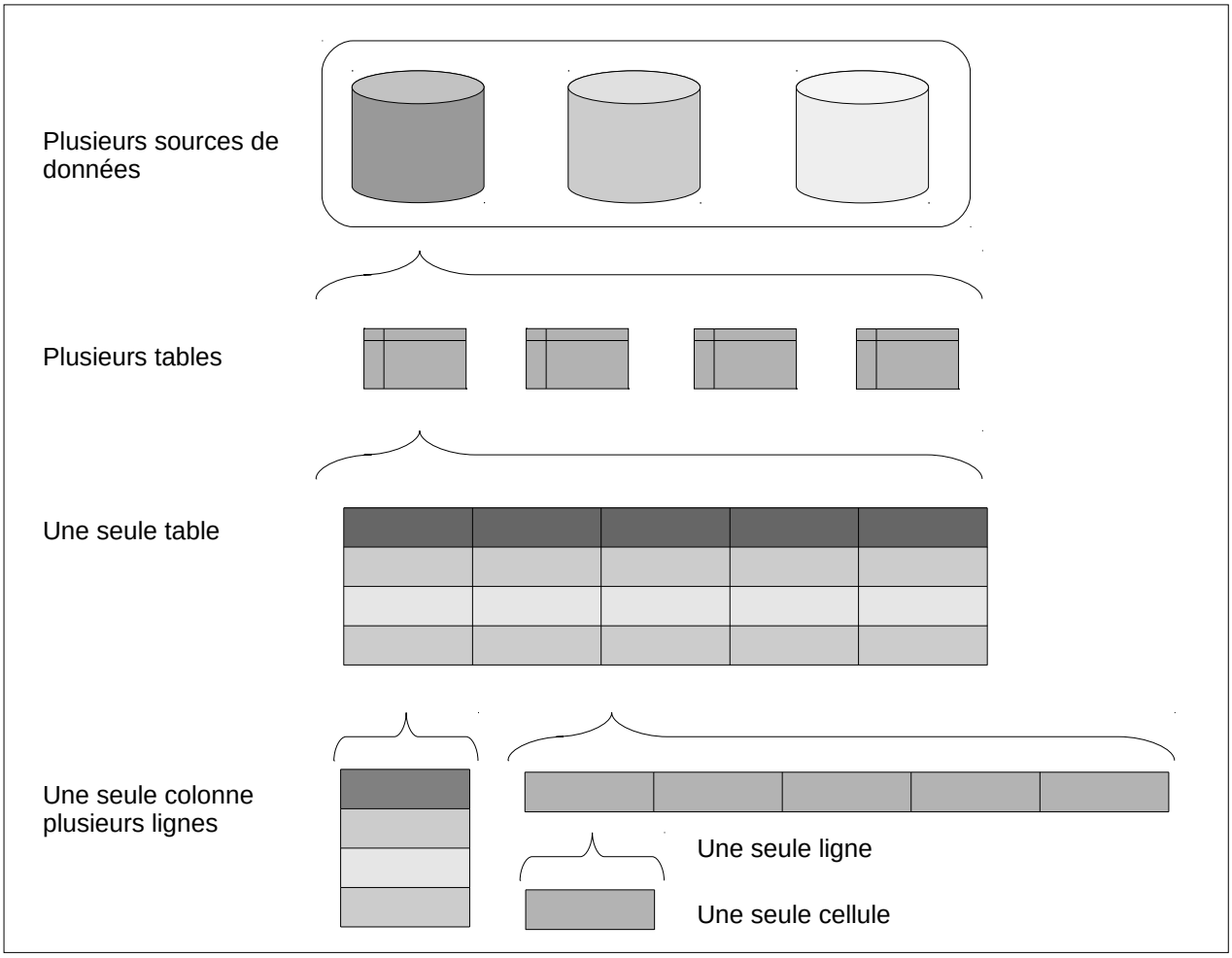


Figure 16: Niveau de granularité de la taxonomie

Tableau 9: Éléments de la taxonomie

Niveau de granularité	Problème de qualité	Définition	Exemple
Une seule cellule	Valeur manquante	La valeur de la cellule est nulle.	Dans la table PATIENT, la colonne DATE_NAISSANCE d'un enregistrement est nulle. PATIENT (ID_PATIENT = 444908, PRENOM = "JOSIANE", NOM = "DEWALLE", DATE_NAISSANCE = ""
Une seule ligne	Valeur dérivée erronée	La valeur d'une colonne calculée à partir d'autres colonnes est fausse.	Dans la table MEDICAMENT, la colonne TOTAL ne correspond pas au produit de la dose, de la concentration et de la durée d'administration : MEDICAMENT (…, ID_MEDICAMENT : 118, DATE_DEBUT : "14/12/2012 08:20:05", DATE_FIN = "14/12/2012 10:45:50", DOSE = 2, UNITE = "mL/h", CONCENTRATION = 1, UNITE_CONCENTRATION = "mg/ml", TOTAL = 7, UNITE_TOTAL = "mg")
Une seule colonne plusieurs lignes	Violation de la contrainte d'unicité	Une colonne a la même valeur dans plusieurs enregistrée, alors qu'elle doit être unique à un enregistrement	Dans table PATIENT, les lignes suivantes ont le même identifiant patient : PATIENT (PATIENT_ID = 102310, INPATIENT_IDENTIFIER = "1002392301", ...) PATIENT (PATIENT_ID = 104913, INPATIENT_IDENTIFIER = "1002392301", ...)
Une seule table	Violation d'une contrainte métier.	Un enregistrement ne respecte pas une contrainte métier.	Dans la table EVENEMENT, la date d'un enregistrement de « Fin de chirurgie » est antérieure à la date d'un enregistrement de « Début de chirurgie » : EVENT (ID_INTERVENTION = 250931, ID_EVENEMENT = 158 ["Début de chirurgie"], DATE = "02/01/2012 13:21:04", ...) EVENT (ID_INTERVENTION = 250931, ID_EVENEMENT = 159 ["Fin de chirurgie"], DATE = "02/01/2012 08:40:14", ...)
Plusieurs tables	Violation d'une contrainte d'intégrité référentielle	La valeur d'une clé étrangère ne correspond à aucun enregistrement dans la table référencée.	La valeur de clé étrangère ID_UNITE d'un enregistrement de la table INTERVENTION ne correspond à aucun enregistrement dans la table UNITE : INTERVENTION (ID_INTERVENTION = 230291, …, ID_UNITE = 214, …) UNITE (ID_UNITE = 213, …) UNITE (ID_UNITE = 215, ...)
Plusieurs sources de données	Synonymes entre plusieurs schémas	Des noms différents sont utilisés pour nommer le même objet dans deux bases de données.	Dans les bases de données de biologie et de séjour administratif, l'entité patient est nommée PAT et PATIENT, respectivement.

3.2.2 Standardisation des données

La diversité des modèles de données locaux et des vocabulaires complique la mise en commun des données ainsi que le partage d'algorithmes, d'outils et de résultats (17,18). Nous avons décidé de travailler à partir du modèle commun OMOP (25,26), qui en 2015, avait déjà intégré plus de 200

millions de patients grâce à l'adoption précoce du projet par plusieurs pays, dont les États-Unis, le Royaume-Uni, les Pays-Bas, la Suède, l'Italie, la Corée et Taïwan (26).

Malgré la diffusion internationale de ce modèle de données, à notre connaissance, en 2019, les données du Système National des Données de Santé (SNDS) n'avaient pas encore été converties au format OMOP. De même, des données en dehors du champ médico-administratif, telles que celles du bloc opératoire et des soins primaires, n'avaient pas non plus été intégrées vers ce modèle.

3.2.2.1 Mise au format OMOP du Système National des Données de Santé

Le modèle OMOP a été mis en avant en premier lieu par les équipes de l'EDS de l'Assistance Publique - Hôpitaux de Paris (APHP). Nous avons entamé un travail collaboratif avec les data scientists de l'AP-HP et de la DREES pour implémenter au format OMOP une extraction du SNDS (44). Les données ont été extraites du SNDS dans le cadre du projet national français "Personnes Agées en Risque de Perte d'Autonomie" (PAERPA), déployé par le Ministère des Affaires sociales et de la Santé de d'octobre 2014 à décembre 2019. Ce programme expérimental est mis en œuvre dans 16 zones administratives et se concentre sur les adultes fragiles âgés de 75 ans et plus. Dans cette population, le parcours de santé du patient est coordonné par un médecin de famille et implique au moins un autre professionnel de la santé, le plus souvent une infirmière et/ou un pharmacien de quartier.

Les principaux critères d'extraction de données étaient les suivants : résider dans la région du Valenciennois-Quercitain dans le nord de la France et avoir 75 ans ou plus au 1er janvier 2015. Les données ont été extraites pour la période allant du 1er janvier 2014 au 31 décembre 2017. Cette extraction couvrait à la fois des données hospitalières et des donnés ambulatoires (i.e., consultations de médecines générales, soins infirmiers).

Au total, les données de 38 730 individus ont été intégrées. Dix-sept vocabulaires correspondant au contexte français ont été ajoutés aux concepts du modèle OMOP. Trois terminologies françaises (UCD, CIP13 et CIM10) ont été automatiquement alignées vers des vocabulaires standardisés. Pour la CIM10, il a été nécessaire de tronquer certains codes, pour trouver une correspondance dans la version internationale équivalente, l'*International Statistical Classification of Diseases and Related Health Problems intervention* (ICD10). En effet, certains codes français étaient plus précis que ceux disponibles dans l'ICD10. Par exemple, le code CIM10 W11.38 (Chute sur ou d'une échelle, lieu de sport, en participant à d'autres activités précisées) a été associé au code ICD10 W11 (fall on and from a ladder). En revanche, il n'a pas été possible d'aligner automatiquement la CCAM à une terminologie internationale et standard d'actes médicaux. Les codes devront être alignés manuellement, au besoin, lors de la réalisation d'études.

Nos résultats ont démontré que les données du système de santé national français pouvaient être intégrées dans le modèle OMOP CDM. L'un des principaux défis résidait dans l'utilisation de concepts OMOP internationaux pour annoter des données enregistrées dans un contexte français. L'utilisation de terminologies locales constituait un obstacle. A l'exception d'une adaptation de la Classification internationale des maladies, 10e révision, le système de santé français n'utilise pas de terminologies internationales.

3.2.2.2 Mise au format OMOP de données d'anesthésie-réanimation

Le modèle OMOP a été conçu initialement pour mener des études pharmaco-épidémiologiques à partir de données issus des remboursements, telles que celles du SNDS par exemple. Nous avons

été les premiers à implémenter vers ce modèle les données du bloc opératoire et de la réanimation, plus détaillées et précises (45,46). Le challenge de ces champs d'activité résidait dans la vélocité d'enregistrements des données (i.e., avec plusieurs mesures à la minute). Il concernait également les plages temporelles spécifiques dédiées à la salle d'opération et à la salle de réveil, pendant lesquelles se déroulaient de nombreux autres événements tels que l'administration de médicaments, les moments clés de la chirurgie, ainsi que les mesures de la pression sanguine et de la respiration.

Concernant la réanimation, nous avons transformé la base de données ouvertes MIMIC-III au format OMOP après un alignement sémantique et structurel (46,47). Au final, 64 % des éléments des 26 tables de MIMIC ont été alignés vers la structure du modèle OMOP et 78 % des éléments de vocabulaires sources ont été alignés avec des terminologies de référence. Le modèle a prouvé sa capacité à soutenir les besoins de la communauté et a été bien accueilli lors d'un datathon rassemblant 160 participants et 15 000 requêtes exécutées en une minute au maximum. Le jeu de données MIMIC-OMOP résultant a été le premier jeu de données MIMIC-OMOP disponible gratuitement avec de vraies données désidentifiées prêtes pour la recherche reproductible en réanimation.

En partenariat avec des anesthésistes-réanimateurs de Lille, Rouen, Amiens et Paris, nous avons recensé 522 éléments de vocabulaire associés aux soins anesthésiques. Ces éléments englobent des données démographiques, des unités de mesure, des étapes en salle d'opération, des médicaments, des périodes pertinentes et des caractéristiques spécifiques. Après alignement sémantique, 353 (67,7 %) de ces concepts anesthésiques ont été associés aux concepts standards proposés par la communauté OHDSI. Les éléments manquants concernaient les concepts liés aux périodes spécifiques à la prise en charge au bloc opératoire et aux caractéristiques calculés secondairement (e.g., hypotension, tachycardie). Ensuite, 10 tables du modèle OMOP ont été alimentées et 2 nouvelles tables (EPISODE et FEATURE) ont du être ajoutées pour stocker les caractéristiques calculées secondairement (Figure 17).

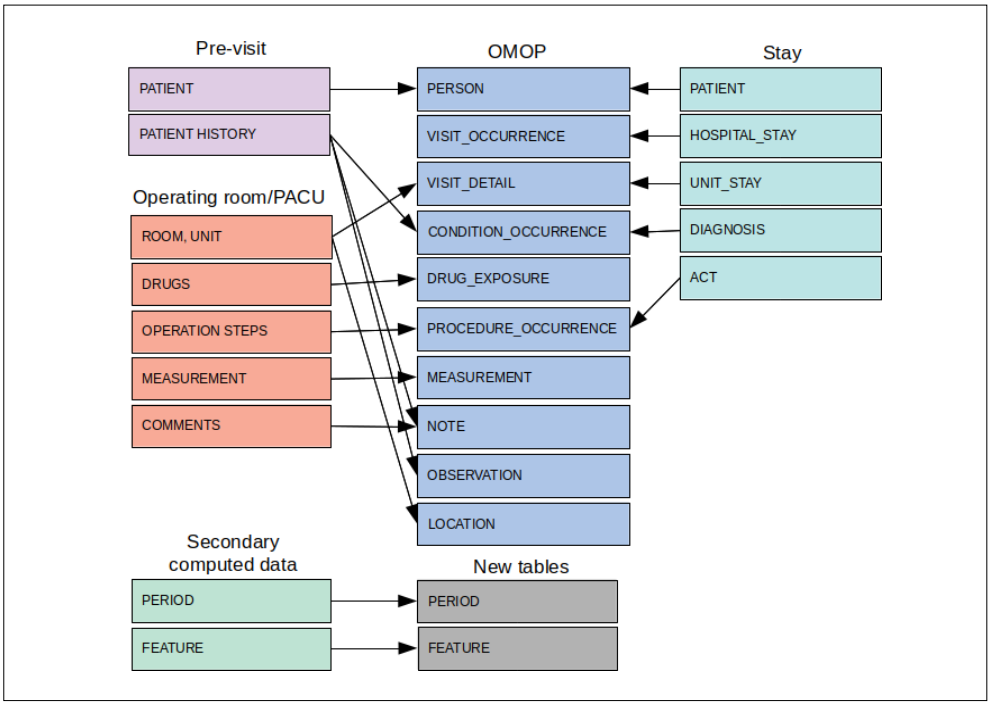


Figure 17: Transformation structurelle des données d'anesthésie vers le modèle OMOP (45)

Au final, nous avons intégré les données de 572 609 interventions chirurgicales. Afin de partager notre travail avec la communauté, nous avons rendu disponible le code de 8 requêtes et 4 tableaux de bord liés aux soins anesthésiques (Figure 18) (48). Malgré quelques adaptations nécessaires, ce travail nous a permis de valider l'utilisation du modèle OMOP pour stocker des données plus précises que celles habituellement utilisées pour des études de pharmaco-épidémiologie.

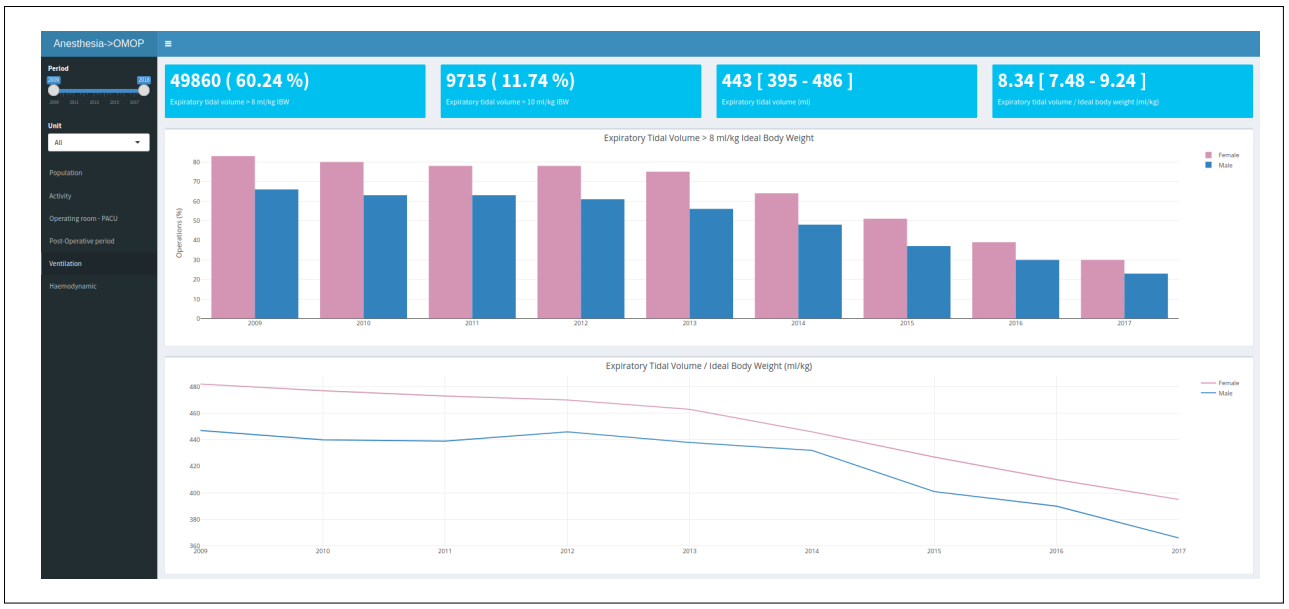


Figure 18: Tableau de bord pour le suivi des recommandations en anesthésie (45)

Depuis ce travail, nous animons un réseau avec les établissements de Rouen, Amiens, Foch, Toulouse, Lille et Saint-Malo. L'objectif est de mutualiser les ETLs, de développer collaborativement des outils et de conduire des études multicentriques à partir des données d'anesthésie-réanimation.

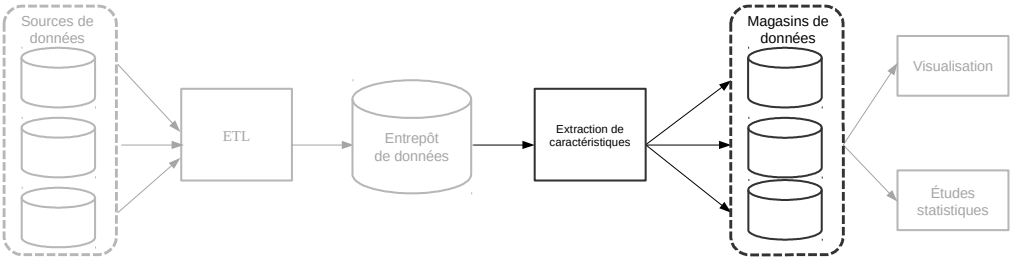
3.2.2.3 Mise au format OMOP de données de soins primaires

Le projet PriCaDa (pour Primary Care Data warehouse) est un projet ayant pour objectif de réutiliser les données de soins primaires afin d'améliorer les pratiques des professionnels de santé (médecins généralistes, infirmiers, pharmaciens). Ce projet est mené en collaboration avec l'éditeur de logiciels WEDA (WEDA, Montpellier, France) et les quatre maisons de santé pluridisciplinaires universitaires (MSPU) de Wattrelos, Lille, Tourcoing et Guesnin, dans le nord de la France.

Les données collectées lors des soins primaires apportent des informations complémentaires à celles disponibles dans les bases de données de facturation, ou dans les bases de données cliniques hospitalières. Ainsi, lors d'une consultation de médecine générale, le praticien recueille les symptômes, mesures les constantes vitales, pose un diagnostic et prescrit des soins. Le praticien reçoit également les résultats d'analyses ou les compte-rendus de consultation de la part d'autres professionnels de santé.

Nous avons été les premiers à intégrer au format OMOP les données de soins primaires d'une maison de santé pluridisciplinaire et universitaire (49). Pour ce travail, nous avons utilisé les données du CSM de Wattrelos et avons travaillé avec l'éditeur de leur logiciel, WEDA (WEDA, Montpellier, France).

3.3 Extraction de caractéristiques

Étape :	
Publications :	**Lamer A**, Jeanne M, Marcilly R, Kipnis E, Schiro J, Logier R, et al. Methodology to automatically detect abnormal values of vital parameters in anesthesia time-series: Proposal for an adaptable algorithm. Comput Methods Programs Biomed. 2016 Jun;129:160–71.
	Lamer A, Jeanne M, Ficheur G, Marcilly R. Automated Data Aggregation for Time-Series Analysis: Study Case on Anaesthesia Data Warehouse. Stud Health Technol Inform. 2016;221:102–6.
	Chazard E, Ficheur G, Caron A, **Lamer A**, Labreuche J, Cuggia M, et al. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. Stud Health Technol Inform. 2018;255:15–9.
	Chazard E, Balaye P, Balcaen T, Genin M, Cuggia M, Bouzille G, **et al.** "Book Music" Representation for Temporal Data, as a Part of the Feature Extraction Process: A Novel Approach to Improve the Handling of Time-Dependent Data in Secondary Use of Healthcare

	Structured Data. Stud Health Technol Inform. 2022 Jun 6;290:567–71.
	Lamer A, Fruchart M, Paris N, Popoff B, Payen A, Balcaen T, et al. Standardized Description of the Feature Extraction Process to Transform Raw Data Into Meaningful Information for Enhancing Data Reuse: Consensus Study. JMIR Med Inform. 2022 Oct 17;10(10):e38936.

Dans l’entrepôt de données, malgré les transformations réalisées lors de l’ETL, et en particulier lors de la standardisation, les données restent complexes car hétérogènes, multidimensionnelles, déséquilibrées (i.e., la fréquence des modalités varient fortement entre elles) et dépendantes du temps. L'hétérogénéité de ces données provient de la diversité des variables impliquées pour extraire des caractéristiques : mesures des signes vitaux (e.g., pression artérielle et fréquence cardiaque) et des signaux ventilatoires (e.g., pression partielle d'oxygène et volume courant), administrations de médicaments, analyses biologiques. En plus de leur hétérogénéité, les bases de données sont multidimensionnelles, ce qui implique que les tables qui les composent ont des unités statistiques différentes. Ainsi, chaque patient aura un nombre différent d'enregistrements dans les autres tables (procédures, diagnostics, mesures, médicaments, etc.), en fonction de la durée du séjour à l'hôpital, des soins reçus et de la durée du suivi (Figure 19). Ensuite, les modalités des variables sont nombreuses et déséquilibrées, c'est-à-dire que chaque terminologie comporte des milliers de codes, dont certains sont largement utilisés, tandis que d'autres ne sont presque jamais nécessaires.

Le format de données requis pour faciliter les analyses est généralement un tableau à structure plate, offrant une perspective unidimensionnelle, où chaque ligne représente un individu statistique distinct et chaque colonne correspond à une variable spécifique. Pour optimiser leur utilisation, ces données doivent être conçues de manière à être indépendantes du temps, ce qui garantit une certaine stabilité dans les analyses longitudinales. Idéalement, les modalités associées à chaque variable devraient être équilibrées et maintenues dans un nombre limité, simplifiant ainsi la manipulation et permettant une interprétation plus directe des résultats.

La figure 19 représente un échantillon de base de données multidimensionnelles d’un service de chirurgie viscérale. Les données sont ventilées sur plusieurs tables, avec des unités statistiques différentes. A l’inverse, la Figure 20 représente le tableau plat utilisé pour l’analyse statistique, avec une ligne par individu et une colonne par variable.

Patient

patient_id	birth_date	sex	zip_code
1	1994-02-02	M	59000
2	1941-03-11	F	59100
3	1965-08-22	M	59200
4	1954-04-08	M	59100
5	2000-01-12	F	59000

Hospital stay

stay_id	patient_id	admission	discharge	discharge_status
1	1	2019-09-23	2019-09-30	home
2	2	2019-09-20	2019-10-10	home
3	1	2019-11-12	2019-11-20	transfer
4	3	2019-10-20	2019-10-25	home
5	4	2019-10-23	2019-10-27	home
6	5	2019-10-24	2019-10-25	death

Procedure

stay_id	procedure_date	procedure_code	procedure_label
1	2019-09-24	HNFA016	Appendectomy
2	2019-09-20	HNFA016	Appendectomy
2	2019-09-20	ZCQK002	Abdominal radiography
3	2019-11-12	LMMA012	Unilateral hernia repair
3	2019-11-13	DEPQ003	Electrocardiography
4	2019-10-20	LMMA012	Unilateral hernia repair
5	2019-10-23	ZCQK002	Abdominal radiography
5	2019-10-23	LMMA012	Unilateral hernia repair
6	2019-10-24	DEPQ003	Electrocardiography
6	2019-10-24	HNFA016	Appendectomy

Unit stay

unit_stay_id	stay_id	admission	discharge	unit_type
1	1	2019-09-23	2019-09-24	emergency
2	1	2019-09-24	2019-09-30	conventional
3	2	2019-09-23	2019-09-25	Intensive care
4	2	2019-09-25	2019-09-27	conventional
5	2	2019-09-27	2019-10-10	conventional
6	3	2019-11-12	2019-11-13	Intensive care
7	3	2019-11-13	2019-11-20	conventional
8	4	2019-10-20	2019-10-25	conventional
9	5	2019-10-23	2019-10-27	conventional
10	6	2019-10-24	2019-10-25	emergency
11	6	2019-10-25	2019-10-27	Intensive care

Figure 19: Données brutes multidimensionnelles et dépendantes du temps

patient_id	stay_id	age	emergency	intensive_care	appendectomy	hernia treatment	length_of_stay	death
1	1	25	1	0	1	0	7	0
2	2	78	0	1	1	0	20	0
1	3	25	0	1	0	1	8	0
3	4	54	0	0	0	1	5	0
4	5	65	0	0	0	1	4	0
5	6	19	1	1	1	0	1	1

Figure 20: Tableau plat utilisé pour l'analyse statistique

La transition des données brutes, multidimensionnelles, vers un format de tableau plat représente une opération complexe et souvent non documentée, créant ainsi une sorte de "boîte noire". Cette absence de documentation claire et de procédures standardisées complique la reproduction des études. Cette problématique essentielle souligne un défi majeur dans le domaine de la recherche liée à la réutilisation des données de santé : comment établir des méthodologies cohérentes et reproductibles pour cette transformation fondamentale des données.

Dans les sections suivantes, nous allons rapporter notre expérience dans l’extraction de caractéristiques, et comment nous envisageons de standardiser cette étape.

3.3.1 Retour d’expérience et première normalisation

Dans un premier temps, et en se basant sur notre expérience, nous avons récapitulé les opérations à réaliser lors du processus d’extraction de caractéristiques (31).

Certaines variables pouvaient être directement analysées, telles que l'âge et le genre. D'autres caractéristiques devaient être extraites. Une méthode courante pour extraire des caractéristiques est de considérer des seuils de normalité définis par des experts. Cela a permis de définir des événements temporels. Parfois, des données manquantes peuvent également être imputées : il est généralement admis que si un patient présentait un symptôme lié à l'hypoglycémie ou à l'hyperglycémie, la glycémie aurait immédiatement été mesurée (Mr Smith sur la Figure 21). Dans cette étude de cas, l'absence de valeur a été déduite comme l'absence d'hypoglycémie et d'hyperglycémie.

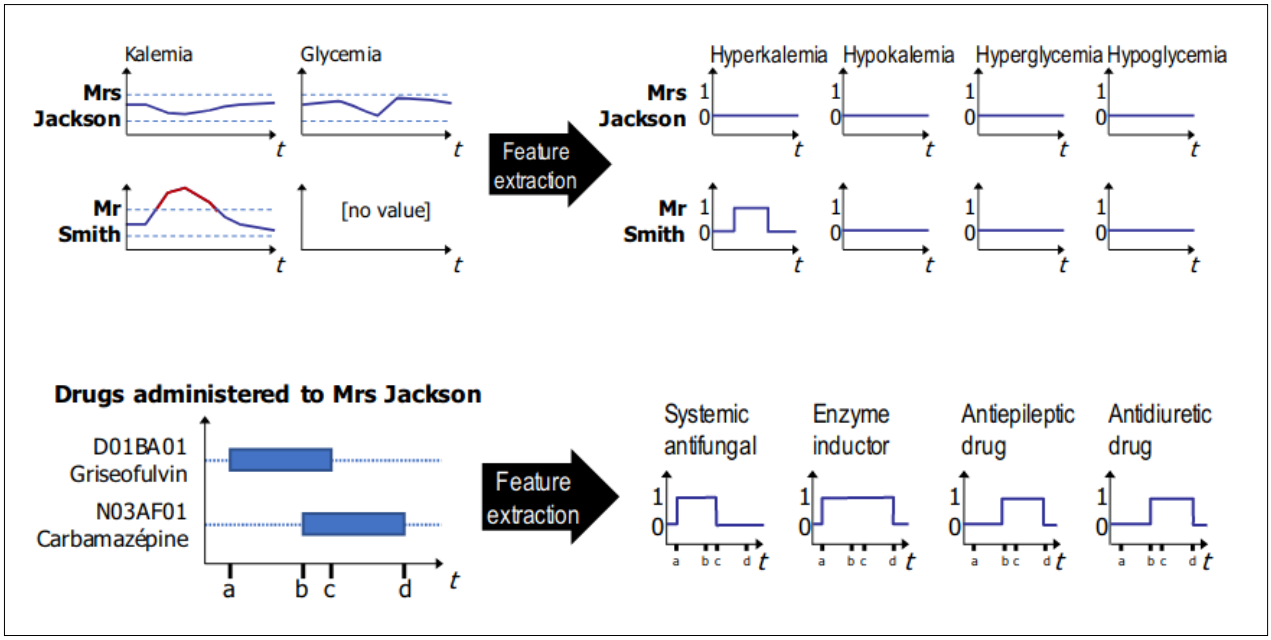


Figure 21: Extraction de caractéristiques à partir de données de biologie et d'administrations de médicaments (31)

Le Tableau 10 liste des opérations usuelles de data management pour extraire les caractéristiques. Dans le cas de données temporelles, il est nécessaire d'identifier une période pour filtrer le signal et une méthode d'agrégation pour arriver à l'unité statistique désirée. Ces deux opérations sont présentées avec plusieurs cas d'usage sur la Figure 22. Tout d'abord, les périodes de bloc opératoire, d'anesthésie et de chirurgie sont identifiées à partir des événements documentés au bloc opératoire. Dans un second temps, il est possible de filtrer les données de monitorage (e.g., fréquence cardiaque, pression artérielle) et les administrations de médicaments et les agréger pour calculer (i) des valeurs moyennes, minimales, maximales, (ii) des périodes passés au-delà de valeurs seuils prédédinies, et (iii) des doses totales administrées.

Tableau 10: Opérations usuelles de data management

Opérations	Exemple
Filtrer les enregistrements sur une période d'intérêt	Conserver les mesures de pression artérielle moyenne entre l'induction anesthésique et l'extubation pour s'affranchir des mesures lors de l'installation du patient.
Réduire la complexité des données à une seule table de données (avec une ligne par individu statistique et une colonne par variable)	Fusionner au sein d'un même tableau les administrations de médicaments, les diagnostics, et les actes médicaux de séjours hospitaliers.
Introduire des seuils spécifiques au domaine pour des variables quantitatives	Identifier un épisode d'hypotension à partir de mesures de pression artérielle supérieures à 65 mmHg
Réduire le déséquilibre des données en permettant un regroupement cohérent basé sur les connaissances	Regrouper les codes de médicaments ou de diagnostics en fonction d'une catégorie. Exemple : les troubles de l'humeur à partir du code CIM10 F3.
Gérer les données hétérogènes sous forme d'événements génériques dépendant du temps.	Calculer la valeur maximale de créatinine dans les deux jours qui suivent une intervention chirurgicale.

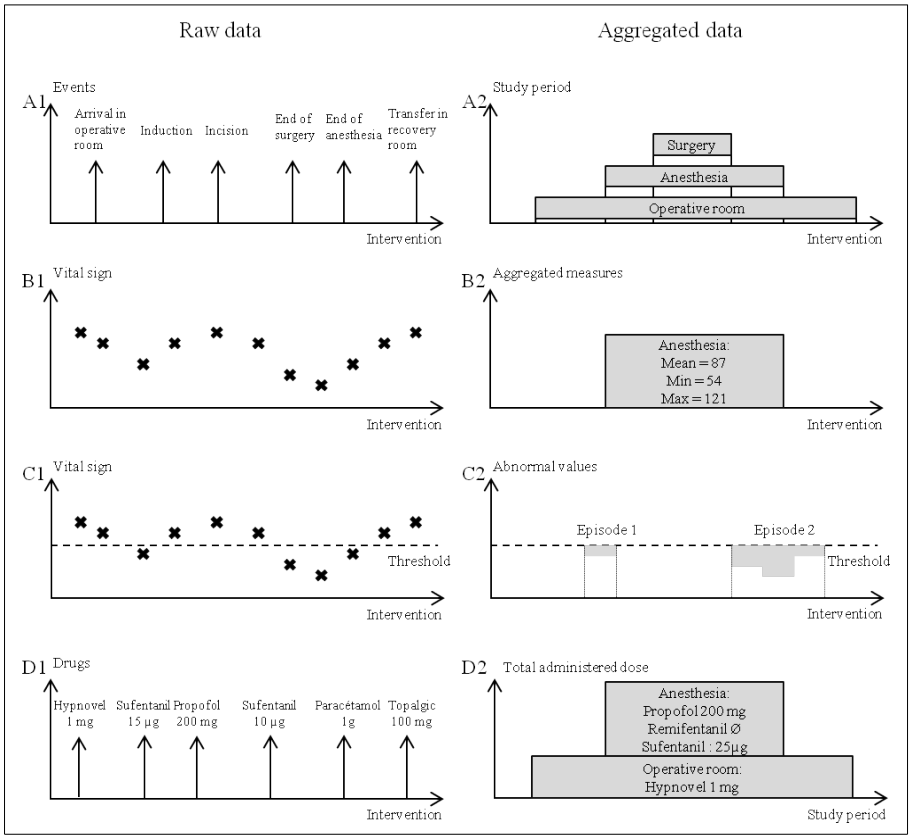


Figure 22: Filtre et agrégation de données (50)

Ces cas d'usage nous ont conduit à résumer l'extraction de caractéristiques par 4 éléments : l'unité statistique, le signal de départ, la période d'intérêt, la fonction d'agrégation pour réduire les données multiples en une valeur par individu statistique (50).

3.3.2 Implémentation

Les variations anormales des paramètres vitaux, telles que l'hypotension ou la tachycardie, surviennent fréquemment pendant les procédures d'anesthésie et peuvent entraîner un dysfonctionnement des organes, entraînant ainsi une prolongation de la durée de séjour à l'hôpital, une augmentation de la morbidité et, dans certains cas, la mort (51). Ces variations sont liées aux effets secondaires des médicaments anesthésiques tels que la vasoplégie ou la dépression myocardique, aux complications liées à la procédure telles que les saignements ou la perte de liquide, ainsi qu'aux facteurs de risque et aux comorbidités liés au patient. Les occurrences de ces variations anormales peuvent être détectées grâce aux paramètres mesurés au bloc opératoire et enregistrés par la feuille informatisée d'anesthésie. Divers seuils sont souvent appliqués sans qu'une méthode spécifique ne soit consensuellement établie dans la littérature pour détecter ces variations.

L'objectif de notre travail était de proposer une méthode robuste et explicite pour identifier ces variations anormales, en contribuant à une plus grande transparence et reproductibilité dans le processus d'extraction de caractéristiques à partir des données du bloc opératoire (52).

Pour cela, nous avons identifié et défini un ensemble de paramètres qui doivent être obligatoirement explicités. Ils permettront de préciser à partir de quand il faudra commencer à comptabiliser un épisode et quand le clôturer, comment interpoler les valeurs entre deux mesures successives, comment tenir compte de la fréquence d'échantillonnage du signal et des valeurs manquantes, quel indicateur qualité utiliser pour qualifier l'enregistrement et le conserver ou non pour l'analyse statistique (Figure 23). Ces paramètres sont détaillés dans le Tableau 11.

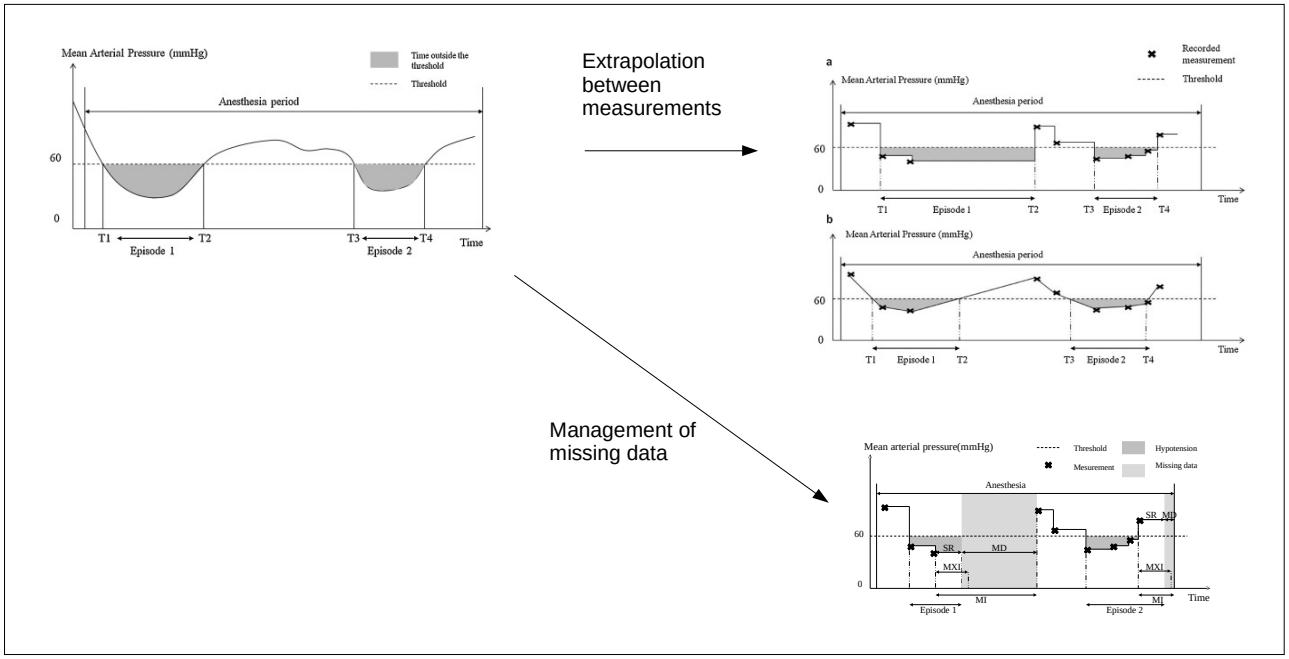


Figure 23: Détection de l'hypotension à partir des mesures de pression artérielle (52)

Cette méthode permet d'extraire plusieurs caractéristiques: le nombre d'épisodes au-delà du seuil, la durée totale passée au-delà du seuil, la durée et la proportion de temps pendant laquelle aucune mesure n'était disponible.

Tableau 11: Paramètres nécessaires lors de l'extraction de caractéristiques à partir de mesures

Paramètres	Description	Exemple
Période d'étude	Sélection des mesures sur la période d'intérêt pour l'étude	Induction-Extubation Incision-Fin de chirurgie
Signal	Paramètre mesuré	Fréquence cardiaque, pression artérielle moyenne, désaturation en oxygène
Valeur seuil	Valeur seuil absolue ou relative (par rapport à une valeur de référence) au-delà de laquelle un épisode est comptabilisé	60 mmHg 80 % de la valeur initiale à l'arrivée au bloc
Méthode d'interpolation	Méthode utilisée pour estimer les valeurs intermédiaires entre deux points de données consécutifs	Interpolation linaire Dernière observation reportée (LOCF, Last Observation Carried Forward)
Intervalle maximal entre deux mesures	Intervalle de temps maximal entre deux mesures pour ne pas considérer qu'il y ait des données manquantes	360 secondes pour un signal mesuré normalement toutes les 5 minutes
Proportion de temps maximale sans mesure	Rapport entre la durée total sans mesure, et la durée de la période d'intérêt pour l'étude. Au delà de cette valeur, l'intervention sera exclue de l'analyse	25 %

Dans ce travail, nous avons proposé un algorithme, et 4 enregistrements fictifs pour tester son implémentation et un modèle de données pour stocker les résultats.

3.3.3 Des données brutes, aux tracks, puis aux caractéristiques

Le principal objectif de ce travail est de présenter une description normalisée des étapes et des transformations nécessaires lors du processus d'extraction des caractéristiques lors de la réalisation d'études observationnelles rétrospectives (53). Un objectif secondaire est d'identifier comment les caractéristiques pourraient être stockées dans le schéma d'un entrepôt de données.

Cette étude impliquait trois étapes principales : (i) la collecte de cas d'étude pertinents liés à l'extraction de caractéristiques et basés sur l'utilisation automatique et secondaire des données ; (ii) la description normalisée des données brutes, des étapes et des transformations communes à l'ensemble des cas d'étude ; et (iii) l'identification d'une table appropriée pour stocker les caractéristiques dans le modèle de données commun OMOP) Nous avons interviewé 10 chercheurs provenant de 3 hôpitaux universitaires français et d'une institution nationale, impliqués dans 8 études rétrospectives et observationnelles.

À partir de ces études, 2 états (piste et caractéristique) et 2 transformations (définition de la piste et agrégation de la piste) ont émergé. Une *piste* est un signal dépendant du temps ou une période d'intérêt, définie par une unité statistique, une valeur et 2 repères temporels (un événement de début et un événement de fin). Une *caractéristique* est une information de haut niveau indépendante du temps avec une dimensionnalité identique à l'unité statistique de l'étude, définie par un libellé et une valeur. La dimension temporelle est devenue implicite dans la valeur ou le nom de la variable.

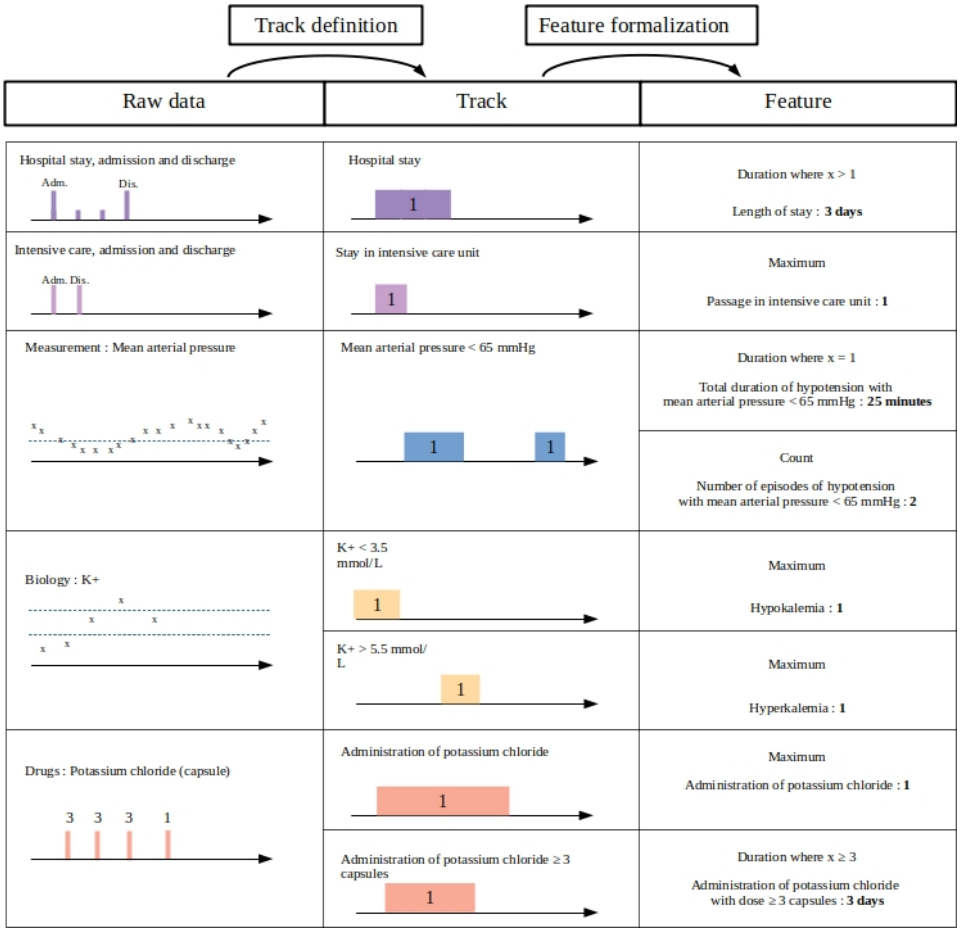
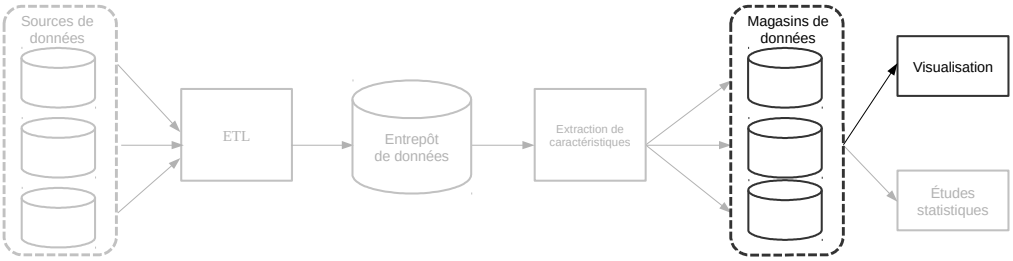


Figure 24: Transformation des données brutes en tracks (pistes) puis en caractéristiques (features) (53)

Nous proposons les tables TRACK et FEATURE pour stocker les variables obtenues lors de l'extraction des caractéristiques et étendre le modèle de données commun CDM OMOP.

Au final, la normalisation des données brutes en pistes exige une grande expertise en ce qui concerne les données, mais permet l'application d'un nombre infini de transformations complexes. En revanche, l'agrégation des pistes pour obtenir les caractéristiques est une opération très simple avec un nombre fini de possibilités (e.g., valeurs moyenne, minimale, maximale, compte, etc.). Une description complète de ces étapes pourrait améliorer la reproductibilité des études rétrospectives en évitant les « boîtes noires » que sont habituellement les étapes de data management.

3.4 Visualisation des données

Étape :	
Publications :	**Lamer A**, Laurent G, Pelayo S, El Amrani M, Chazard E, Marcilly R. Exploring Patient Path Through Sankey Diagram: A Proof of Concept. Stud Health Technol Inform. 2020 Jun 16;270:218–22.
	Boudis F, Clement G, Bruandet A, **Lamer A**. Automated Generation of Individual and Population Clinical Pathways with the OMOP Common Data Model. Stud Health Technol Inform. 2021 May 27;281:218–22.
	Laurent G, Moussa MD, Cirenei C, Tavernier B, Marcilly R, **Lamer A**. Development, implementation and preliminary evaluation of clinical dashboards in a department of anesthesia. J Clin Monit Comput. 2020 May 16;
	Lamer A, Saint-Dizier C, Fares E, Debien C, Cleva E, Whatelet M, et al. Automated Monitoring Reports of the Activity of the French National Professional Suicide Prevention Helpline. Stud Health Technol Inform. 2023 May 18;302:474–5.

La visualisation de données constitue un élément essentiel de la chaîne de réutilisation des données. Elle s'articule autour de deux objectifs majeurs : explorer les données, souvent de manière semi-automatisée, et communiquer les résultats afin d'éclairer les prises de décision. Un enjeu crucial réside dans la sélection précise des indicateurs, qui devront résumer l'information contenue dans les données complexes.

Dans cette partie, nous présenterons plusieurs réalisations de visualisation de données complexes à des fins décisionnelles.

3.4.1 Représentation graphique du parcours patient

3.4.1.1 Aperçu du parcours patient avec le diagramme de Sankey

La représentation d'informations complexes comme le parcours du patient est crucial en santé afin d'évaluer la qualité des soins, gérer les ressources des établissements, et générer des hypothèses de recherche. Cependant, la complexité des données brutes et la diversité des variables rendent difficile l'extraction d'informations pertinentes. De plus, la variabilité des séjours des patients complique la représentation graphique des flux.

Plusieurs types de diagrammes sont couramment utilisés pour représenter des flux. D'un point de vue pratique, un flux est une succession d'étapes, chacune étant définie par un état initial, un état final et une quantité. D'un point de vue graphique, le flux est représenté par des liens représentant une quantité ou un volume, des nœuds de départ aux nœuds d'arrivée. En santé, les nœuds correspondent généralement à des étapes ou des événements de la prise en charge, comme le passage dans des unités de soins, des administrations de médicaments, des procédures

médicales, ou des diagnostics. Les liens représentent un ensemble de patients passant d'une étape à une autre. La largeur des nœuds et des liens est proportionnelle au nombre de patients.

Nous avons comparé les différents types de diagrammes de flux : le diagramme en réseau, le diagramme de Chord, le dendogramme, le *sunburst*, le diagramme de Sankey, l'agenda, et la matrice de connexion (54). La plupart de ces représentations ne permettent de visualiser qu'une seule étape d'un parcours. En revanche, le diagramme de Sankey se distingue en permettant la représentation de plusieurs étapes avec des informations sur les volumes de flux, ce qui le rend adapté à la complexité des parcours des patients.

Nous avons évalué l'utilité du diagramme de Sankey pour représenter les parcours des patients de manière lisible par les cliniciens. Deux cas d'étude ont été examinés : (i) les parcours des patients avec une pancréaticoduodénectomie après la chirurgie, et (ii) la fréquence des nouvelles interventions et la mortalité après une première chirurgie au sein d'un même séjour hospitalier.

Les opérations de pancréaticoduodénectomie ont été sélectionnées avec le code CCAM HNFA007 à partir de l'entrepôt de données d'anesthésie du CHU de Lille (55). Entre 2010 et 2018, 551 patients ont été inclus. Deux étapes clés du séjour ont été étudiés : l'unité post-opératoire et la dernière unité avant la sortie, toutes deux caractérisées par leur niveau de criticité (i.e., conventionnelle, soins continus, soins intensifs/réanimation) et le statut de sortie (mortalité).

Deux gestionnaires et deux médecins ont examiné le diagramme, décrivant les flux de patients dans chaque unité après l'opération (Figure 25). Les médecins ont noté des parcours atypiques, soulignant la pertinence du diagramme pour la compréhension. Bien qu'ils le trouvent plus lisible qu'un tableau, ils ont exprimé le besoin d'informations contextuelles lors de la première consultation. Les médecins prévoient d'investiguer pourquoi certains patients n'ont pas été dirigés vers une unité de soins appropriée plus tôt.

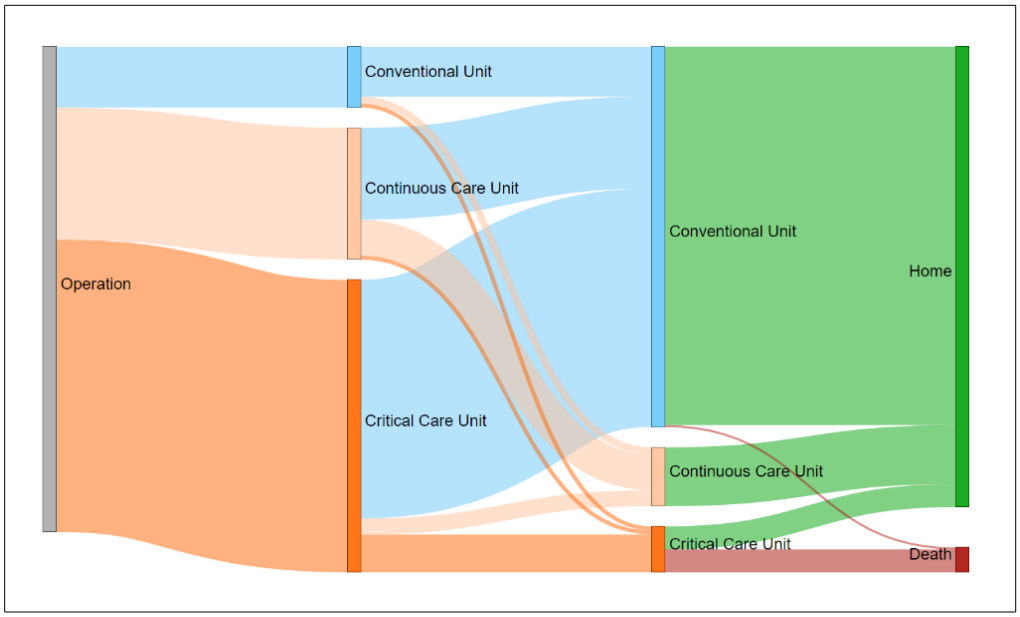


Figure 25: Représentation du parcours patient après une pancréaticoduodénectomie à partir du diagramme de Sankey (54)

Concernant le second cas d'étude, entre 2010 et 2018, 473 953 patients ont subi une opération chirurgicale. Pour chaque patient, nous avons calculé le nombre d'opérations et le statut de sortie final (vivant ou décès). Pour tirer le meilleur parti du diagramme, les résultats devaient être affichés

en pourcentage pour pouvoir comparer les modalités des étapes avec un faible effectif (3ème et 4ème interventions chirurgicales) avec celles ayant un plus grand effectif (1ère et 2ème interventions chirurgicales). Nous pouvons observer que les taux de mortalité et de réintervention augmentent avec le nombre d'opérations réalisées lors du même séjour hospitalier. Ce diagramme a été présenté à deux gestionnaires et deux médecins, et tous ont convenu de l'intuitivité des informations affichées. Ils ont simplement demandé des informations contextuelles sur la source des données. Selon eux, ce résultat constituera le point de départ d'un programme de recherche sur l'impact du passage par plusieurs opérations sur les taux de mortalité.

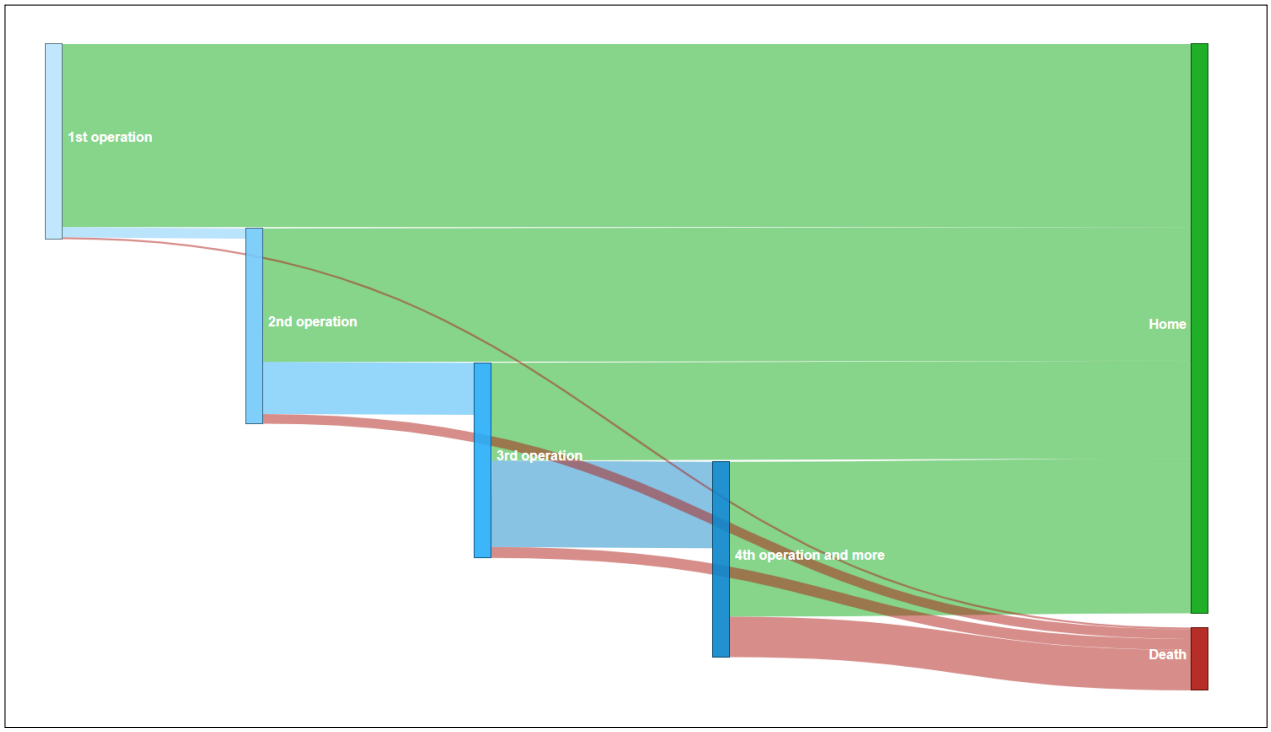


Figure 26: Ré-intervention et décès en fonction du nombre de chirurgies (54)

Il en ressort que le diagramme de Sankey peut être utilisé pour représenter plusieurs étapes successives du parcours patient. Cependant, pour éviter une surcharge d'informations, il est nécessaire de sélectionner les étapes pertinentes, regrouper des modalités, et filtrer les parcours singuliers. En conclusion, il n'y a pas de représentation qui permettent de répondre à toutes les questions, mais une représentation graphique adaptée à chaque question.

3.4.1.2 Production automatisée du parcours patient

En cherchant à automatiser la production de cette représentation du parcours patient, nous avons rencontré plusieurs challenges. Tout d'abord, les grands nombres de patients et l'hétérogénéité des variables augmentent la complexité des ensembles de données et entraînent une surcharge d'informations. Ensuite, les étapes de la prise en charge ne surviennent pas forcément au même moment d'un patient à l'autre. Ainsi, un premier patient pourrait avoir une intervention chirurgical dès son entrée à l'hôpital, alors qu'un second patient pour passer au bloc opératoire après plusieurs jours.

Les données ont été extraites de la base de données du Centre Hospitalier Universitaire de Lille, et couvraient les séjours hospitaliers, les procédures médicales et les diagnostics pour l'année

2019. Les données locales ont été mises au format OMOP avant traitement, afin de pouvoir partager nos développements plus facilement (56). Nous avons choisi trois cas d'étude en lien avec les problématiques que nous traitons régulièrement : la chirurgie totale de la hanche (cas 1), le pontage coronarien (cas 2) et l'implantation transcathéter d'une valve aortique (cas 3). Ces trois procédures bénéficient de différents parcours de soins, et nous souhaitions visualiser les étapes précédant la chirurgie dans le cas 1, étudier le parcours des patients décédés dans le cas 2 et visualiser l'ensemble du séjour hospitalier dans le cas 3.

Afin de tenir compte de l'hétérogénéité des séjours, nous proposons de définir une séquence générique des étapes. Pour cela, quatre listes d'indices sont développées : la première en partant de la première étape de la séquence (liste A), la seconde en partant de la dernière étape de la séquence (liste B), la troisième en partant de l'événement d'inclusion en se déplaçant vers les points finaux (liste C) et et la dernière en partant des points finaux vers l'événement d'inclusion (liste D). Par la suite, les étapes d'intérêt sont sélectionnées en fonction de leurs indices dans la séquence d'étapes. Un exemple des listes d'indices est présenté dans la Figure 27. Le diagramme de Sankey est indexé (ou centré) sur l'événement d'inclusion, et seules les étapes préalablement sélectionnées sont représentées. Nous avons utilisé ces 4 listes d'indices pour sélectionner les étapes pertinentes aux trois cas d'usage (Figure 28).

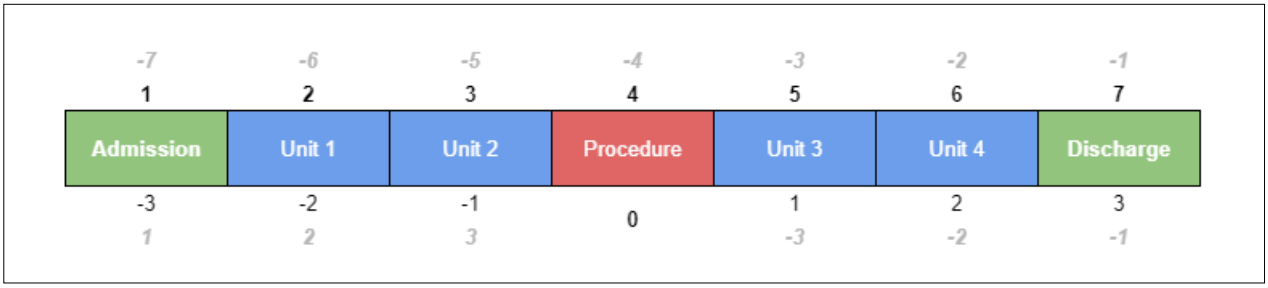


Figure 27: Séquence d'indices

Dans le cas A, il y avait deux flux principaux avant la chirurgie : (i) les patients venaient de leur domicile et étaient admis dans des unités conventionnelles, ou (ii) les patients étaient admis en unités d'urgence pour une admission non planifiée (Figure 28 - A).

Pour le cas B, tous les patients étaient directement dirigés vers les unités de soins intensifs. Le diagramme a montré que certains des patients étaient revenus aux soins continus avant que leur situation ne se détériore et qu'ils décèdent, ou étaient réadmis aux soins intensifs avant de décéder (flux orange, Figure 28 - B).

En ce qui concerne le cas C, le diagramme décrit le parcours global avant et après la chirurgie (Figure 28 - C). Les patients passaient par plusieurs types d'unités, en proportions équivalentes. Cela a mis en évidence l'hétérogénéité des soins. Environ un quart des patients étaient transférés vers un autre hôpital à la fin du séjour.

En proposant la sélection des étapes à travers différentes listes d'indices, et en simplifiant les libellés hétérogènes, notre méthode contribue à réduire la complexité du parcours du patient.

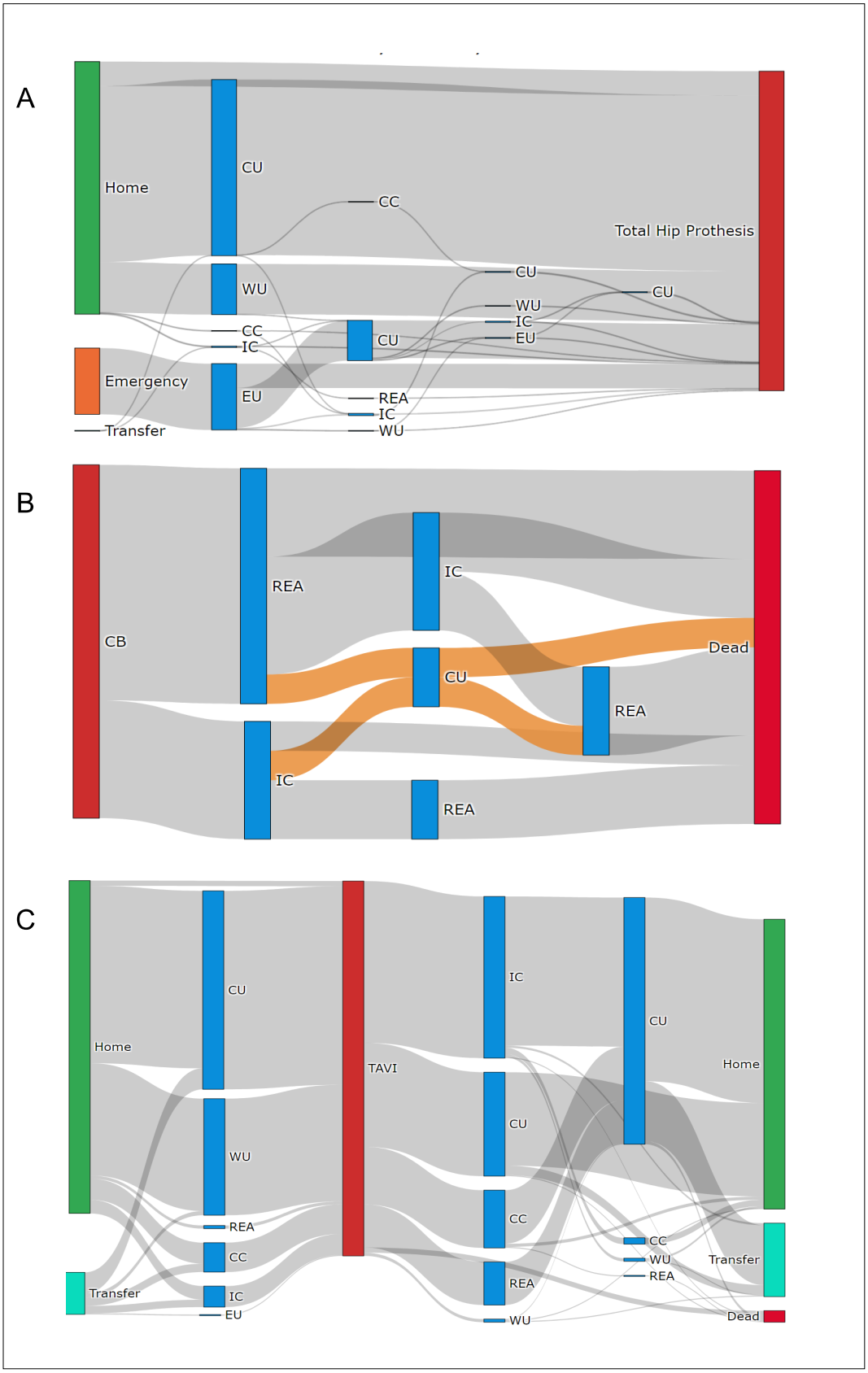


Figure 28: Parcours des patients opérés pour une chirurgie totale de la hanche (A), un pontage coronarien (B) et une implantation transcathéter d'une valve aortique (C) (56)

3.4.2 Tableaux de bord et rapports automatisés

En complément des opportunités pour la recherche, les données à disposition dans les EDS permettent aussi de fournir des indicateurs de suivi de l'activité des services et de la prise en charge des patients. A partir de l'entrepôt d'anesthésie du CHU de Lille, nous avons développé des tableaux de bord concernant la gestion de l'unité d'anesthésie et l'évaluation de la qualité des soins (57).

Les besoins des utilisateurs et les exigences techniques ont été identifiés lors d'entretiens semi-dirigés, puis synthétisés. Plusieurs représentations ont ensuite été développées (selon les bonnes pratiques en matière de visualisation) et soumises aux utilisateurs finaux pour évaluation. Enfin, les tableaux de bord ont été implémentés et rendus accessibles pour une utilisation quotidienne via le réseau du centre médical. Après une période d'utilisation, les retours des utilisateurs finaux sur la plateforme du tableau de bord ont été collectés sous forme d'un score de convivialité du système (échelle de 0 à 100). Dix-sept thèmes (correspondant à 29 questions et 42 indicateurs) ont été identifiés. Après priorisation et évaluation de la faisabilité, 10 tableaux de bord ont finalement été mis en œuvre et déployés. Les tableaux de bord abordaient divers aspects tels que l'activité globale de l'unité, la conformité aux directives sur l'hémodynamique peropératoire, la ventilation (Figure 29), ainsi que la documentation de la procédure d'anesthésie. Le score moyen (écart-type) de convivialité du système était de 82,6 (11,5), ce qui correspond à une excellente convivialité.

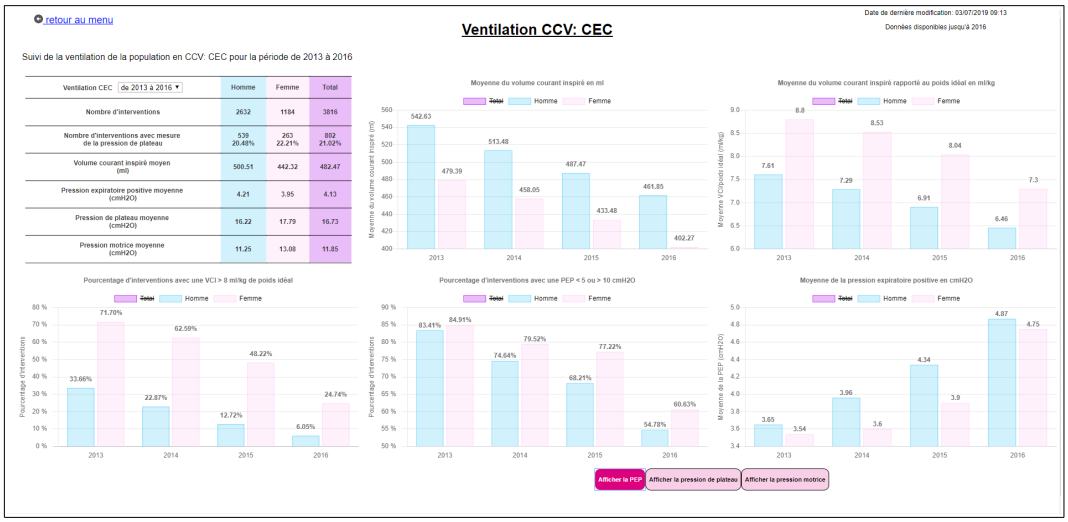


Figure 29: Suivi des recommandations sur la ventilation protectrice (57)

Le numéro national de prévention du suicide, le 3114, a été lancée le 1er octobre 2021. Le 3114 est une ligne professionnelle, gratuite et confidentielle. Des professionnels de la santé de quinze centres régionaux différents gèrent les appels quotidiens. Des indicateurs sur le nombre d'appels reçus, les taux d'appels répondus, leur durée et leur répartition par centre étaient nécessaires pour évaluer l'activité de la ligne d'assistance. Nous avons élaboré des rapports et des présentations automatisés à l'aide de Rmarkdown. Deux formats de rapports ont été élaborés : un rapport national destiné à être présenté à la Direction Générale de la Santé (DGS), et un rapport régional pour chaque centre d'appels. Ils comportent des indicateurs communs tels que le nombre d'appels reçus, d'appels répondus et les taux de réponse (Figure 30). L'utilisation de Rmarkdown a permis d'automatiser la production de 16 rapports hebdomadaires (un par centre, et un national) et 1

rapport national mensuel pour la DGS. En complément des rapports, nous avons également développé une application web interactive permettant de visualiser les transferts d'appels entre les centres (Figure 31). Cette application utilise les diagrammes de Sankey présentaient dans la section 3.5.1.

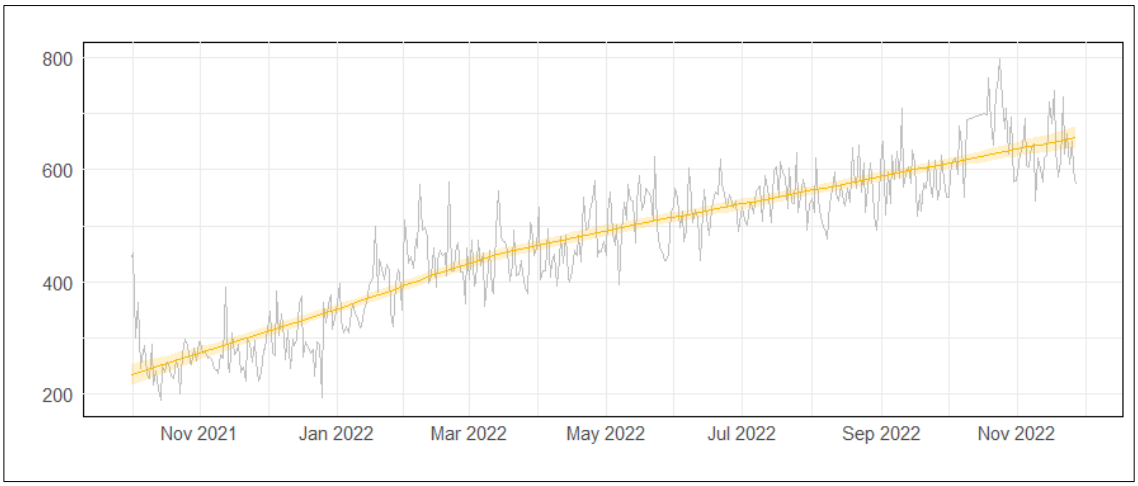


Figure 30: Nombre d'appels reçus par le 3114 depuis son lancement

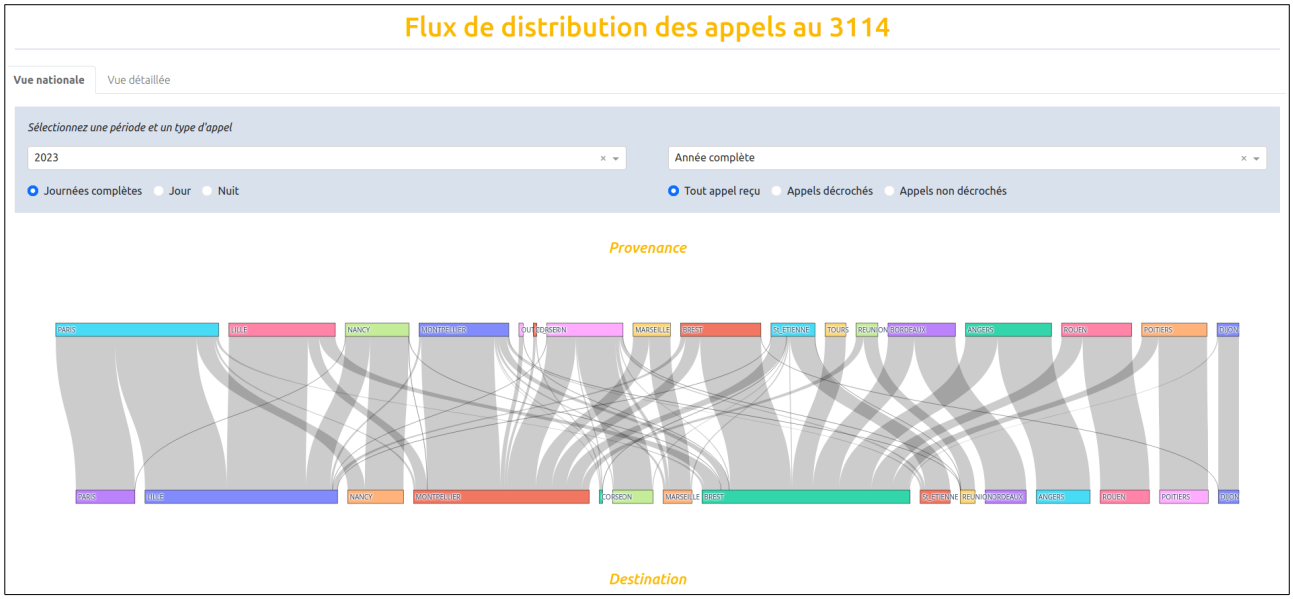
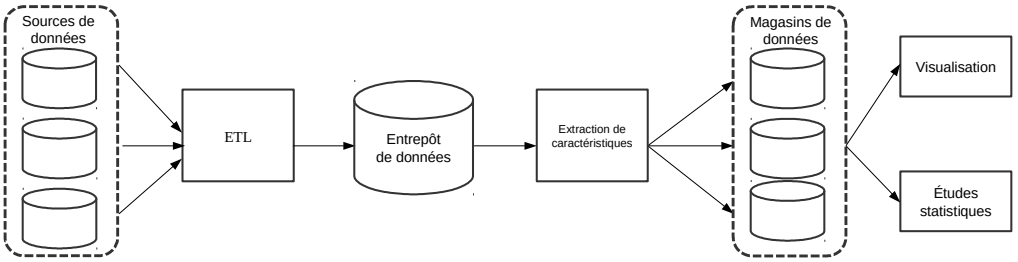


Figure 31: Transferts d'appels entre les centres répondants du 3114

3.5 Mise en place de la réutilisation des données et retour d'expériences

Études :	
Publications :	**Lamer A**, Ficheur G, Rousselet L, van Berleere M, Chazard E, Caron A. From Data Extraction to Analysis: Proposal of a Methodology to Optimize Hospital Data Reuse Process. Stud Health Technol Inform. 2018;247:41–5.
	Mangold P, Filiot A, Moussa M, Sobanski V, Ficheur G, Andrey P, **et al**. A Decentralized Framework for Biostatistics and Privacy Concerns. Stud Health Technol Inform. 2020 Nov 23;275:137–41.
	Lamer A, Filiot A, Bouillard Y, Mangold P, Andrey P, Schiro J. Specifications for the Routine Implementation of Federated Learning in Hospitals Networks. Stud Health Technol Inform. 2021 May 27;281:128–32.
	Lamer A, Moussa MD, Marcilly R, Logier R, Vallet B, Tavernier B. Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project. J Clin Monit Comput. 2023 Apr;37(2):461–72.
	Doutreligne M, Degremont A, Jachiet PA, **Lamer A**, Tannier X. Good practices for clinical data warehouse implementation: A case study in France. PLOS Digit Health. 2023 Jul;2(7):e0000298.

3.5.1 Processus standardisé d'utilisation d'un entrepôt de données

La mise en place d'études prospectives pour les thèses et mémoires d'internes est coûteuse en temps et présente des risques. Les premiers résultats de l'entrepôt de données en anesthésie ont conduit à une augmentation des demandes d'études. Cependant, le manque de connaissance du processus de recherche par les cliniciens et les internes entrave la réalisation d'études de qualité et perturbe le service.

De plus, au démarrage de l'entrepôt de données, les étapes étaient réalisées de manière séquentielle avec peu d'interaction entre les acteurs, ce qui entraînait des demandes d'extraction de données non justifiées sur le plan médical et incompatibles avec l'analyse statistique ultérieure. Le processus était également entravé par des boucles de rétroaction après l'analyse statistique, provoquant une perte de temps significative et empêchant la reproductibilité de la recherche.

Pour remédier à ces problèmes, nous avons mis en place un cadre pour structurer l'utilisation de l'entrepôt de données en anesthésie à des fins de recherche clinique observationnelle (58).

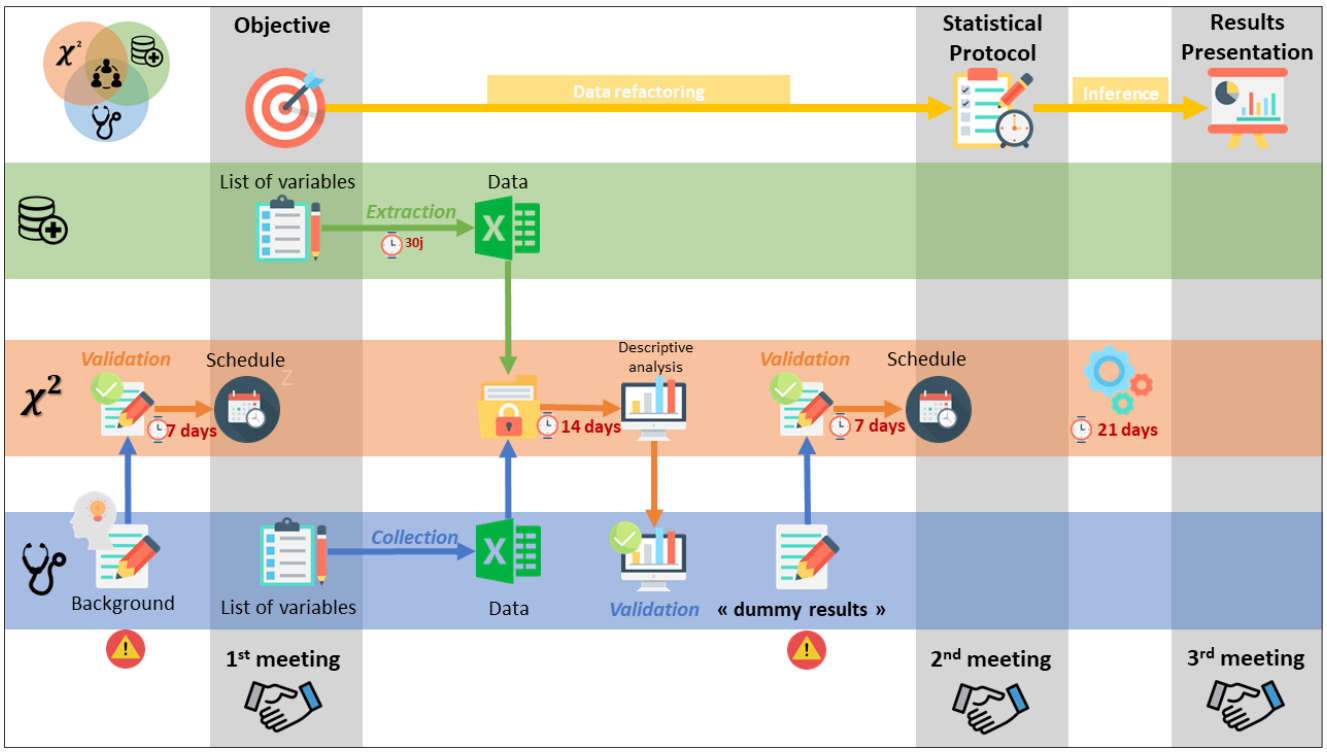


Figure 32: Processus standardisé d'utilisation d'un entrepôt de données (58)

Le cadre proposé est structuré autour de trois réunions entre les cliniciens, les informaticiens et les statisticiens (Figure 32). Le data scientist agit en tant que coordinateur, anime les réunions et vérifie chaque jalon.

Pour la première réunion, les cliniciens doivent fournir un contexte détaillé et répertorier toutes les variables pertinentes (critères finaux, expositions, facteurs de confusion, etc.) grâce à une revue de la littérature. De plus, un aperçu de la méthodologie de chaque étude (y compris la taille de l'échantillon) est également requis. Un modèle basé sur les recommandations STROBE est utilisé pour structurer la revue (59). L'objectif de la première réunion est de décider des objectifs principaux et secondaires de l'étude. La faisabilité est évaluée par le data scientist qui compare les objectifs aux données disponibles dans l'entrepôt de données. La possibilité de calculer de nouvelles variables à partir des données existantes peut être envisagée. Lorsque les données ne sont disponibles que sur papier, la faisabilité de la collecte manuelle et de la fusion ultérieure est discutée entre le clinicien et le data scientist.

À la suite de cette première réunion, une liste de variables à extraire et/ou collecter est établie. Une analyse statistique descriptive complète est effectuée, ce qui permet au clinicien de contrôler la qualité des données.

Avant la deuxième réunion, les cliniciens sont priés de fournir un « modèle de résultats », c'est-à-dire des tableaux, du texte et des graphiques vides, montrant le type de résultats qu'ils aimeraient obtenir (Figure 33). L'objectif de la deuxième réunion est de valider le protocole statistique proposé par le statisticien en fonction de ce « modèle de résultats » et de la qualité des données disponibles.

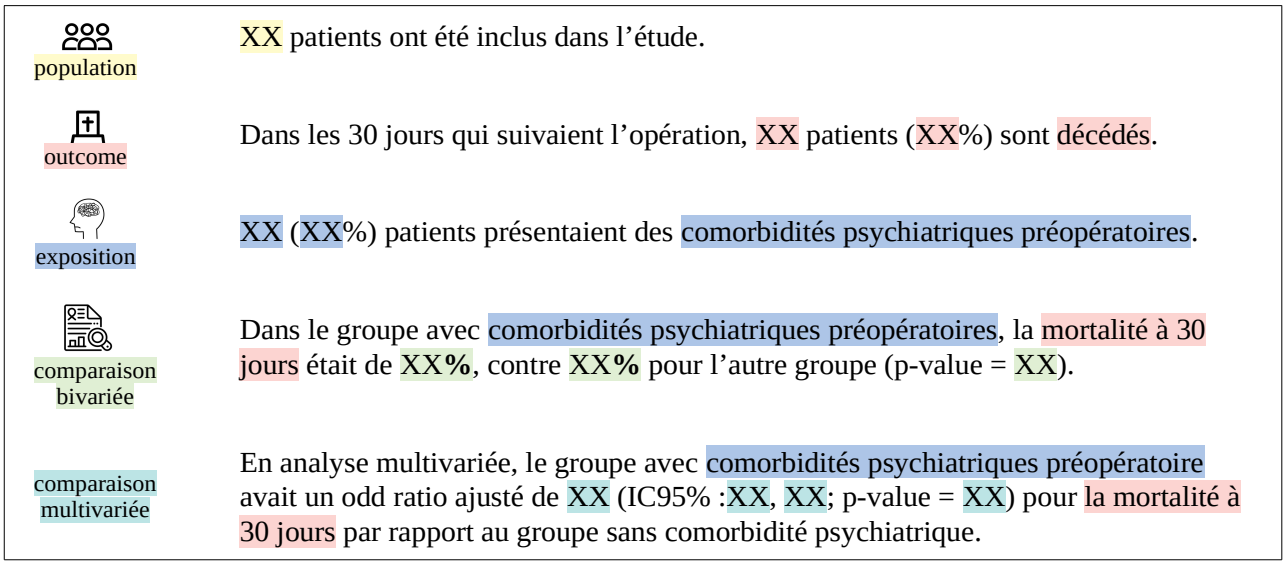


Figure 33: Modèle de résultats attendus

L'analyse statistique est ensuite réalisée, et les résultats sont présentés et expliqués au début de la troisième réunion. Les résultats peuvent ensuite être discutés entre les différents collaborateurs. Un paragraphe clé de l'analyse statistique est rédigé par le statisticien afin d'être inséré dans la future publication.

3.5.2 Calcul décentralisé

La constitution de bases de données avec des effectifs représentatifs de la population est souvent indispensable pour disposer d'une puissance statistique suffisante et pour améliorer la généralisation des modèles. Une approche courante consiste donc à centraliser les données provenant de plusieurs centres sur un site principal et à mener l'étude à partir de là (Figure 34). Dans le domaine des données médicales, cette centralisation constitue souvent un défi pratique, car les données sont sensibles et doivent être gérées dans un environnement contrôlé respectant des contraintes légales et éthiques strictes.

Une approche alternative, connue sous le nom d'apprentissage fédéré, consiste à former des modèles statistiques de manière décentralisée, en laissant les données sur chaque site, en effectuant des calculs localement et en communiquant des informations agrégées entre les centres (Figure 35). Une telle approche a déjà été appliquée à la médecine dans quelques études, dans le but de préserver la confidentialité des données sensibles (60), ainsi que la souveraineté des producteurs de données.

Notre travail constituait une première étape vers la définition et la mise en œuvre d'un cadre d'apprentissage décentralisé pour la médecine, se distinguant des travaux précédents en ce sens qu'il permet une décentralisation totale, c'est-à-dire qu'aucun centre n'était tenu de jouer un rôle central dans les calculs (bien que cela reste une option). Notre objectif était de démontrer que ce cadre pouvait produire des résultats virtuellement identiques à ceux obtenus dans un cadre centralisé. Pour ce faire, nous avons utilisé deux ensembles de données distincts sur lesquels nous avons appliqués des régressions logistiques. Nous retrouvons les mêmes coefficients de régression logistique qu'avec le modèle monocentrique centralisé (61).

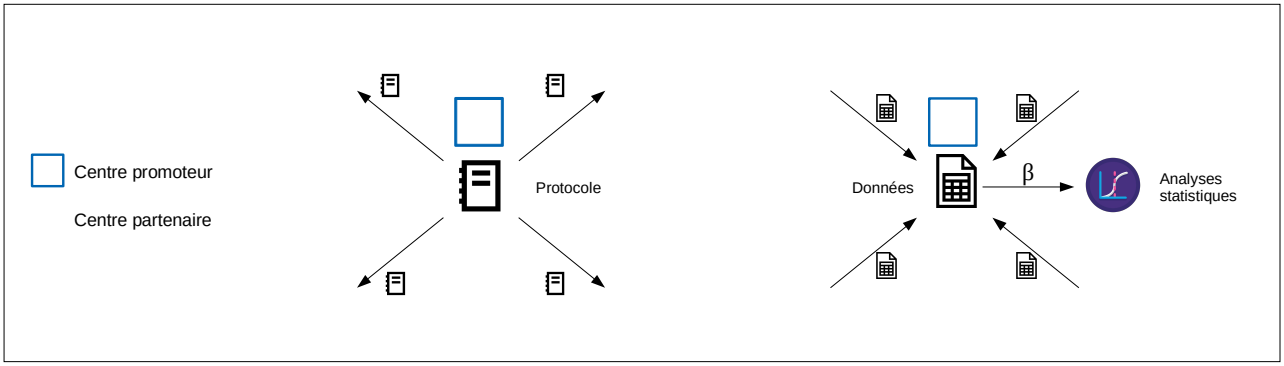


Figure 34: Étude multicentrique centralisée

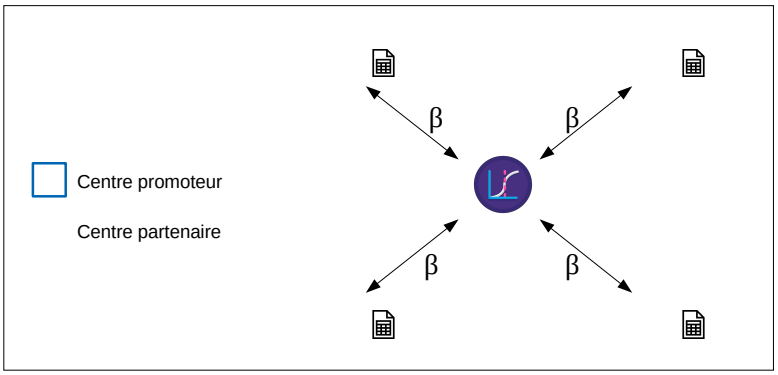


Figure 35: Étude multicentrique avec calcul décentralisé

Malgré la diffusion significative de l'apprentissage fédéré, les recommandations sur la manière de le déployer pour une utilisation courante restent, à notre connaissance, rares. Nous avons collecté les besoins des acteurs hospitaliers afin de proposer des spécifications pour concevoir et mettre en œuvre efficacement l'apprentissage fédéré au sein d'un réseau de partenaires hospitaliers (62).

Des entretiens semi-structurés ont été réalisés par téléphone dans 7 hôpitaux universitaires français (Amiens, Caen, Dijon, Lille, Marseille, Rouen, Toulouse). Les personnes interviewées étaient des décideurs, des chercheurs, des médecins, des informations et des data scientists. Ces entretiens avaient pour objectif de comprendre comment mettre en œuvre l'apprentissage fédéré dans la pratique courante, d'identifier les acteurs impliqués à chaque étape, leurs rôles et préoccupations, ainsi que leurs idées et propositions pour améliorer cette pratique.

Nous avons identifiés sept étapes pour déployer un cadre d'apprentissage fédéré (Figure 36) : la constitution du consortium, la mise en place de l'architecture, la définition du protocole, la collecte des données, l'initialisation du calcul, la modélisation et le partage des résultats.

Un avantage du processus proposé est de rester proche du processus bien établi des études multicentriques centralisées. En effet, le processus d'apprentissage fédéré partage la plupart des étapes classiques en place dans les centres de recherche clinique : la mise en place d'un consortium, la proposition et l'acceptation de l'étude, ainsi que la diffusion des résultats.

À notre avis et selon notre expérience personnelle, ces spécifications peuvent être utiles à toute équipe intéressée par l'apprentissage fédéré d'un point de vue statistique et/ou clinique, et souhaitant concevoir de nouvelles études décentralisées à partir de zéro.

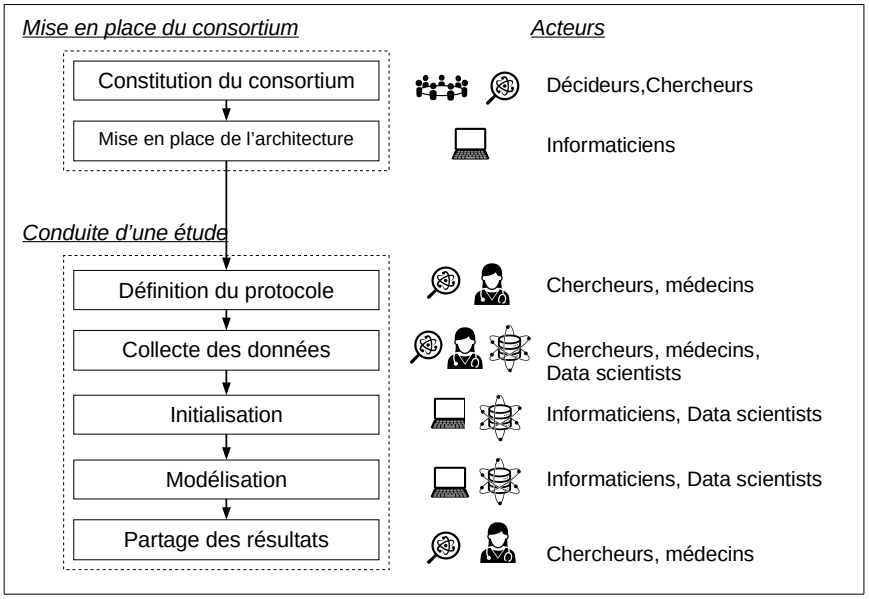


Figure 36: Spécifications pour la mise en place et l'utilisation en routine du calcul décentralisé (62)

3.5.3 Retours d'expériences, barrières et recommandations EDS

3.5.3.1 Panorama des EDS en France

Le paysage des entrepôts de données cliniques en France remonte à 2011 et s'est accéléré à la fin des années 2020. Nous avons conduit une première études exhaustives des EDS dans les hôpitaux régionaux et universitaires (CHU) en France afin d'obtenir une vision globale des éléments clés de leur mise en oeuvre : gouvernance, types de données intégrées, objectifs principaux de réutilisation des données, outils techniques, documentation et processus de contrôle de qualité des données (63).

Sur les 32 CHU, 14 disposaient d'un EDS en production, 5 en phase expérimentale, 5 avaient un projet d'EDS en cours de développement, et 8 n'avaient aucun projet d'EDS au moment des échanges (Figure 37).

La pérennité des EDS était accompagnée par la construction d'un environnement coopératif entre différents acteurs : le Département d'Information Médicale (DIM), le Direction des Ressources Numériques (DRN), la Direction de la Recherche Clinique et de l'Innovation (DRCI), les professionnels de santé, et avec le soutien de la direction ou de la Commission Médicale d'Établissement (CME). Elle était également accompagnée par la création d'une équipe ou d'une entité dédiée à la maintenance et à la mise en œuvre de l'entrepôt de données cliniques. Bien qu'une équipe opérationnelle pour l'entrepôt de données cliniques soit toujours présente, les ressources humaines qui lui sont allouées varient considérablement : d'un demi-équivalent temps plein à 80 personnes pour l'AP-HP, avec une médiane de 6 personnes. Avant de démarrer, les projets sont systématiquement analysés par un comité scientifique et éthique.

Historiquement, les premiers entrepôts de données cliniques reposaient sur le développement de solutions internes. Plus récemment, des acteurs privés proposaient leurs services pour la mise en place et l'implémentation des entrepôts de données cliniques.

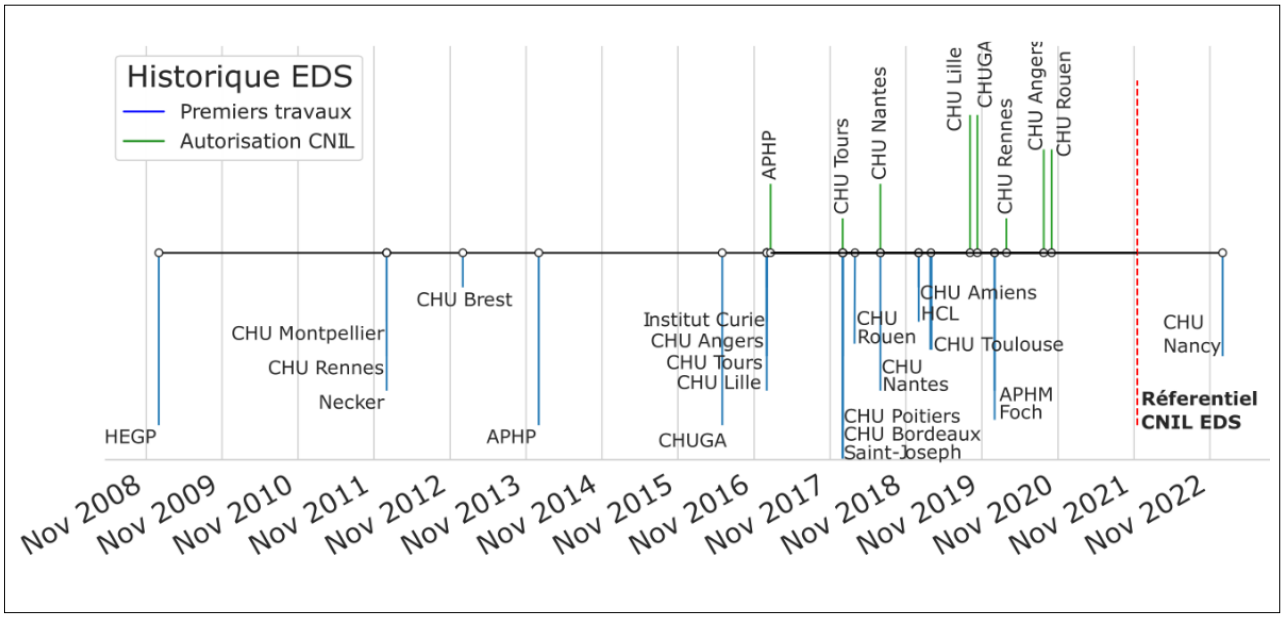


Figure 37: Répartition des EDS en France (63)

La base commune de tous les entrepôts de données cliniques était constituée par les données du logiciel de Gestion Administrative des Patients (identification des patients, mouvements hospitaliers) et les codes de facturation (Tableau 12). Dans un second temps, les flux de données étaient progressivement développés à partir des différents logiciels qui composent le SI.

Tableau 12: Sources de données intégrées dans les EDS

Source de données	Nombre d'EDS (%)
Gestion administrative des malades	21 (100%)
Facturation (autorisation des unités, mouvements dans les unités, diagnostics, actes médicaux)	20 (95%)
Biologie	20 (95%)
Compte-rendus médicaux	20 (95%)
Médicaments	16 (76%)
Imagerie	4 (19%)
Compte-rendus infirmiers	4 (19%)
Anatomopathologie	3 (14%)
Réanimation	2 (10%)
Dispositifs médicaux	2 (10%)

Aujourd'hui, l'utilisation principale des EDS est la recherche scientifique. Les études sont principalement observationnelles (non interventionnelles). Les études se concentrent principalement sur la caractérisation de la population (25%), suivie du développement de

processus d’aide à la décision (24%), de l'étude des facteurs de risque (18%) et de l’évaluation des effets des traitements (16%).

Pour la plupart des institutions interrogées, il subsiste toujours un manque de ressources et une maturité insuffisante en termes de méthodes et d'outils pour mener des recherches interinstitutionnelles (comme dans la région du Grand-Ouest de la France) ou via des appels à projets européens (EHDEN). Ces deux réseaux de recherche sont rendus possibles par une gouvernance supra-locale et un schéma de données commun, respectivement, eHop (64) et OMOP (26). Les hôpitaux de Paris, grâce à leur couverture régionale et au choix d'OMOP, sont également bien avancés dans la recherche multicentrique. Parallèlement, la région Grand-Est est en train de construire un réseau d'entrepôts de données cliniques basé sur le modèle de la région du Grand-Ouest, utilisant également eHop.

Cette étude met en avant une homogénéisation progressive mais encore incomplète des EDS. Des projets nationaux et européens émergent, soutenant les initiatives locales en matière de normalisation, de travaux méthodologiques et d'outils. Une attention particulière doit être portée à la durabilité des équipes de l'entrepôt de données et à la gouvernance multi-niveaux. La transparence des outils de transformation des données et des études doit s'améliorer pour permettre une réutilisation réussie des données multicentriques ainsi que des innovations pour les patients.

3.5.3.2 EDS d’Anesthésie-Réanimation du CHU de Lille

De 2011 à 2021, nous avons développé, maintenu et exploité un entrepôt de données d’anesthésie au CHU de Lille (55). L’entrepôt de données était principalement alimenté par des données collectées par le logiciel d'anesthésie et le logiciel de PMSI. Il pouvait également héberger des données de biologie ou de dispositifs médicaux, en fonction des études. Entre 2010 et 2021, 636 784 dossiers d'anesthésie ont été intégrés pour 353 152 patients. Nous avons rapporté les principales barrières et obstacles rencontrés lors du développement de ce projet et avons fourni 8 conseils pour les surmonter.

3.5.3.3 Recommandations pour l’implémentation d’entrepôt de données de santé

Au cours d’un atelier organisé lors de la Conférence européenne d'informatique médicale MIE2023 à Göteborg, nous avons pu compléter ce premier travail. L’atelier était intitulé « Lessons Learned from the Implementation of Health Data Warehouses : Participatory Development of Recommendations ». Nous avons invité des experts dans la mise en œuvre de systèmes de gestion de données en santé. Ces experts ont décrit leurs systèmes et les difficultés qu'ils ont rencontrés à chaque étape : (i) le lancement du projet d’entrepôt de données, (ii) la mise en place de l’entrepôt de données et (iii) l'utilisation de l’entrepôt de données en routine. Ils ont également été invités à proposer des solutions qu'ils ont réussi à mettre en œuvre pour surmonter les obstacles précédemment signalés. Afin de catégoriser les obstacles rencontrés lors de la mise en place du système de gestion de données en santé, nous avons choisi le cadre du Système d'Ingénierie pour la Sécurité des Patients 2.0 (SEIPS 2.0) (63). Le SEIPS 2.0 est un cadre largement reconnu et bien établi, initialement développé dans le contexte des systèmes de santé. Le choix du SEIPS 2.0 a été motivé par son approche globale visant à comprendre et optimiser les systèmes. Il englobe divers facteurs tels que les personnes, la technologie, les tâches, l'environnement et l'organisation, qui sont tous des composantes essentielles pour le déploiement réussi d'un système de gestion de données en santé.

Après synthèse et consensus, un total de 26 obstacles ont été identifiés, dont 10 étaient liés aux tâches, 5 aux outils et technologies, 4 aux personnes, 4 à l'organisation et 3 à l'environnement externe. La figure 38 présente ces barrières en fonction des étapes du projets et de leur catégorie. Pour faire face à ces défis, un ensemble de 15 recommandations pratiques est proposé, couvrant des aspects essentiels tels que la gouvernance, l'engagement des parties prenantes, la collaboration interdisciplinaire et l'utilisation de l'expertise externe.

Ces recommandations constituent une ressource précieuse pour les établissements de santé qui cherchent à établir et optimiser les systèmes de gestion de données en santé, offrant un plan d'action pour exploiter les données cliniques à des fins de recherche, d'amélioration de la qualité et de soins aux patients améliorés.

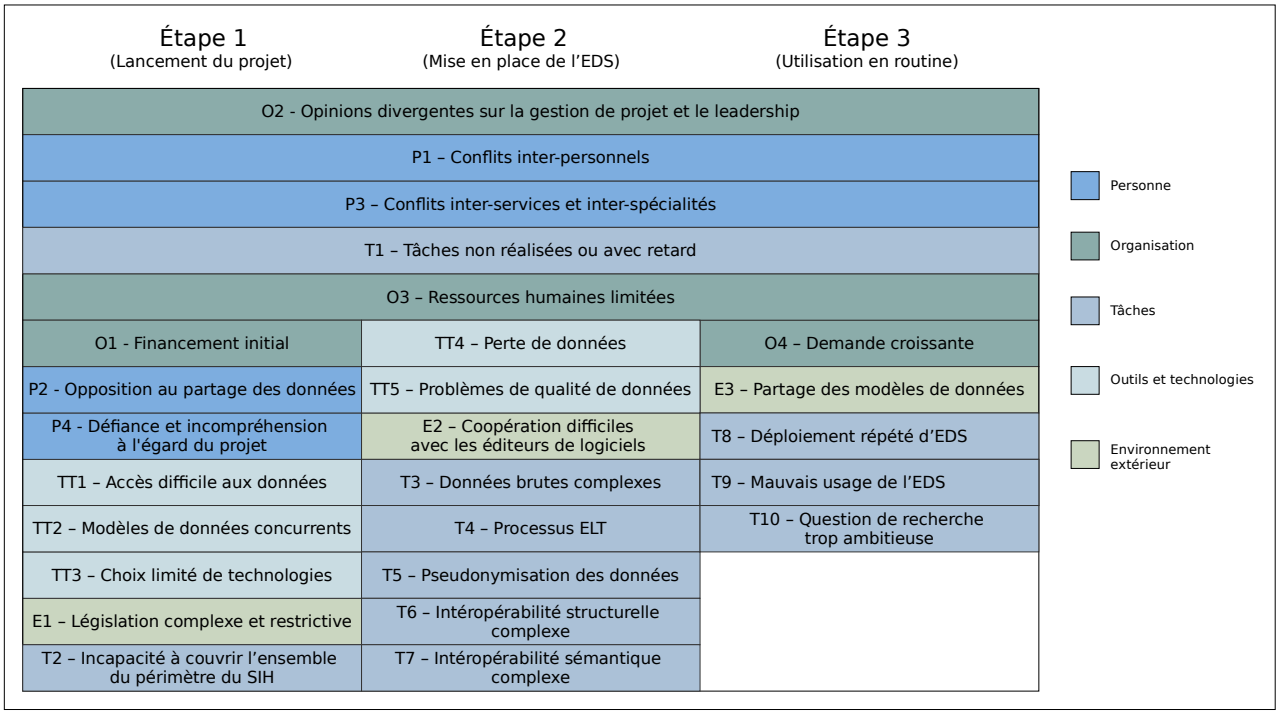


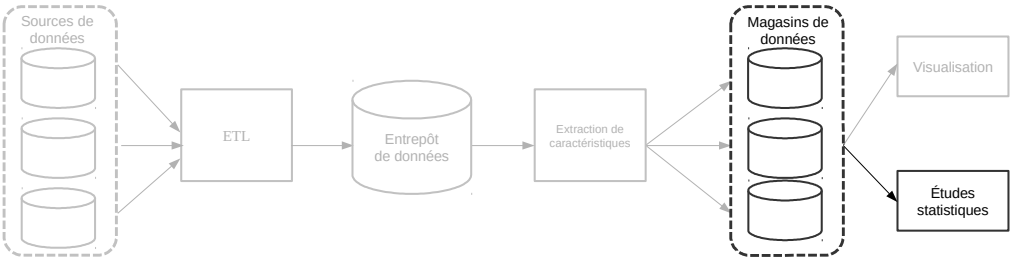
Figure 38: Barrières rencontrées lors d'un projet d'EDS

Chapitre 4 Travaux appliqués

Sommaire

Les travaux appliqués que nous présenterons ont été réalisés à partir des bases de données du PMSI, du SNDS, de l'entrepôt d'anesthésie du CHU de Lille, de l'entrepôt de données de soins primaires, ou des bases de données ouvertes.

4.1 Psychiatrie

Étape :	
Publications :	Levaillant M, Wathelet M, **Lamer A**, Riquin E, Gohier B, Hamel-Broza JF. Impact of COVID-19 pandemic and lockdowns on the consumption of anxiolytics, hypnotics and antidepressants according to age groups: a French nationwide study. Psychol Med. 2023 May;53(7):2861–7.
	Fovet T, Baillet M, Horn M, Chan-Chee C, Cottencin O, Thomas P, **et al**. Psychiatric Hospitalizations of People Found Not Criminally Responsible on Account of Mental Disorder

	in France: A Ten-Year Retrospective Study (2011-2020). Front Psychiatry. 2022;13:812790.
	Fovet T, Chan-Chee C, Baillet M, Horn M, Wathelet M, D'Hondt F, **et al**. Psychiatric hospitalisations for people who are incarcerated, 2009-2019: An 11-year retrospective longitudinal study in France. EClinicalMedicine. 2022 Apr;46:101374.
	Beigné M, **Lamer A**, Eck M, Horn M, Benbouriche M, Thomas P, et al. [A descriptive study of psychiatric care and pre-sentencing psychiatric reports in a French high-security prison]. Encephale. 2022 Mar 21;S0013-7006(22)00031-8.
	Fovet T, Saint-Dizier C, Wathelet M, Horn M, Thomas P, Guillin O, **et al**. Opening the black box of hospitalizations in French high-secure psychiatric forensic units. Encephale. 2023 May 26;S0013-7006(23)00079-9.

4.1.1 Consommation d'anxiolytiques, antidépresseurs et hypnotiques après le début de l'épidémie de la COVID-19.

Au début de la pandémie du COVID-19, beaucoup d'études ont rapporté l'impact négatif de la pandémie et de ses mesures sanitaires associées sur la santé mentale, en particulier chez les adolescents et les jeunes adultes (65). Nous avons entrepris une étude de cohorte historique afin d'examiner la consommation d'anxiolytiques, d'antidépresseurs, et d'hypnotiques au cours de la première année de la pandémie de COVID-19 par rapport aux cinq années précédentes (66).

À cette fin, nous avons extrait les données de délivrance des anxiolytiques, des antidépresseurs et des hypnotiques du SNDS, entre le 1er janvier 2015 et le 28 février 2021. Les personnes ayant reçu un remboursement pour l'un de ces traitements au cours de la semaine sélectionnée ont été identifiées en tant que consommateurs d'antidépresseurs, d'anxiolytiques et d'hypnotiques. Le taux de nouveaux consommateurs a été défini comme la différence entre le nombre de nouveaux consommateurs de médicaments par semaine, avant et après la période de COVID-19, divisée par le nombre total de consommateurs de médicaments. Les individus ont été répartis en cinq groupes d'âge : 12-18, 19-25, 26-50, 51-75 et plus de 75 ans.

L'évolution du nombre hebdomadaire de consommateurs au fil du temps a été analysée à l'aide de modèles de régression linéaire. Des modèles distincts ont été réalisés pour chaque médicament étudié et pour chaque classe d'âge considérée. Pour tous ces modèles, la variable dépendante était le nombre de consommateurs par semaine, et les variables explicatives étaient le temps et la période de COVID-19 (à partir du 17 mars 2020, c'est-à-dire la date du premier confinement sanitaire en France). Une interaction entre ces deux covariables a également été prise en compte pour évaluer l'effet de la pandémie de COVID-19 sur la consommation de médicaments.

Une interaction significative entre le temps et la période de la COVID-19 a été mise en évidence pour l'ensemble des médicaments et des groupes d'âge considérés, à l'exception de la consommation spécifique d'antidépresseurs chez les patients de plus de 75 ans (Figure 39).

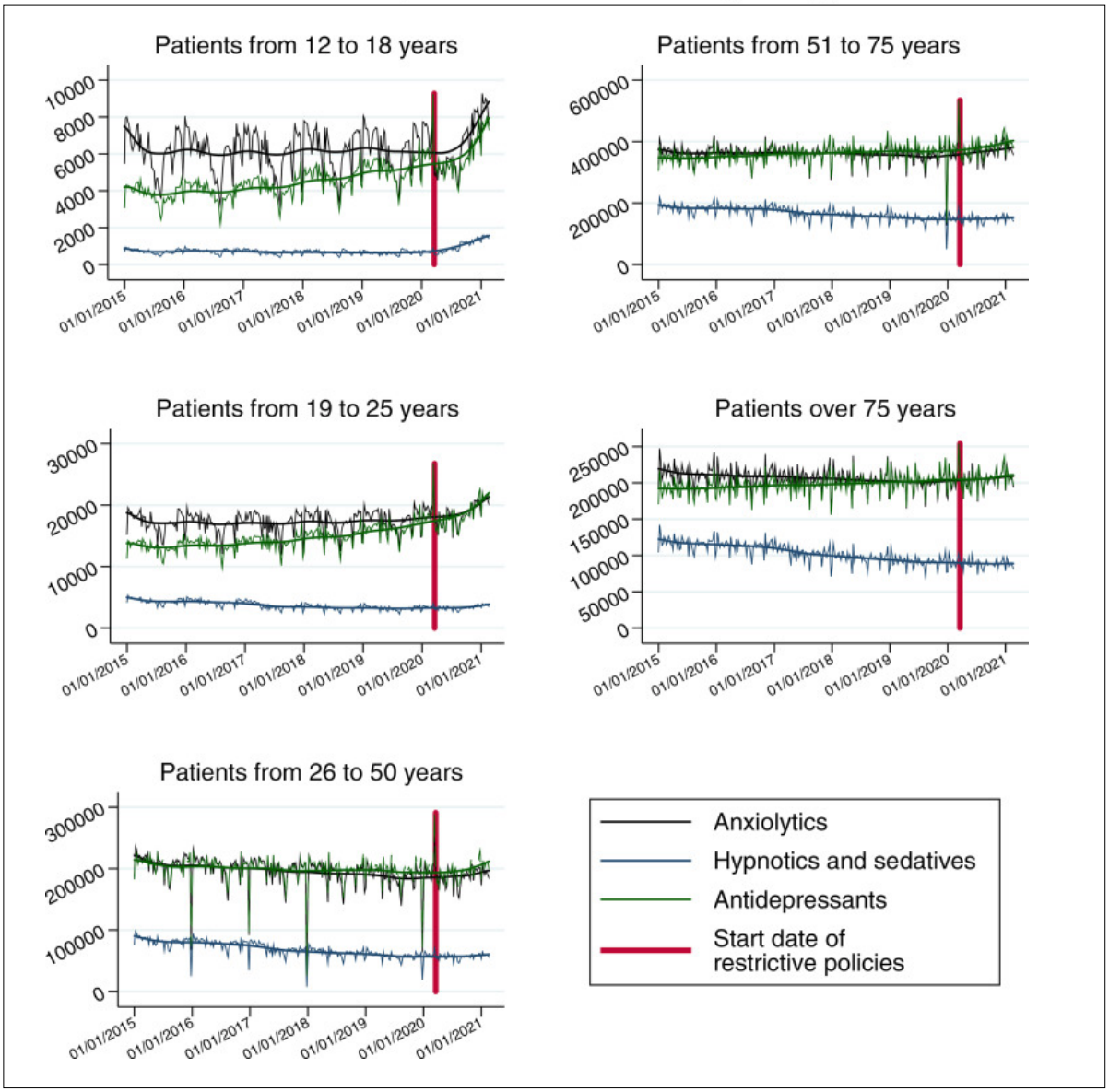


Figure 39: Nombre de patients consommant des psychotropes par classe d'âge

Plus les patients étaient jeunes, plus l'ampleur de ces interactions était prononcée. Ainsi, une augmentation très marquée de la consommation d'antidépresseurs, d'hypnotiques et d'anxiolytiques a été observée chez les personnes âgées de 12 à 18 ans pendant la période de la pandémie. Pour ce groupe d'âge, la consommation d'hypnotiques et d'anxiolytiques est passée d'une légère diminution à une forte augmentation, tandis que les antidépresseurs, qui étaient déjà en légère augmentation, ont connu une augmentation considérable.

Le taux de nouveaux consommateurs de médicaments en raison de la période de COVID-19 dépendait des médicaments considérés et des groupes d'âge. Pour les antidépresseurs, l'augmentation du nombre de consommateurs était respectivement cinq et quatre fois plus élevée parmi les 12-18 ans et les 19-25 ans par rapport aux 26-50 ans. Pour les anxiolytiques, cette augmentation était respectivement quatre et deux fois plus élevée parmi les 12-18 ans et les 19-25 ans par rapport aux 26-50 ans. Enfin, l'augmentation des utilisateurs d'hypnotiques était respectivement neuf et 1,6 fois plus élevée parmi les 12-18 ans et les 19-25 ans par rapport aux 26-50 ans.

La surveillance de la consommation de médicaments psychiatriques pourrait être d'un grand intérêt, car des indicateurs fiables sont essentiels pour la planification des stratégies de santé publique. Une politique post-crise comprenant une surveillance fiable de la santé mentale doit être anticipée.

4.1.2 Hospitalisations psychiatriques des personnes jugées irresponsables en France

La responsabilité pénale est un concept clé dans la sanction pénale des personnes diagnostiquées avec des troubles de santé mentale ayant commis des crimes. En France, sur la base des recommandations d'un ou de plusieurs psychiatres experts, un juge peut déclarer qu'une personne n'est pas pénalement responsable en raison d'un trouble mental si, au moment de l'infraction, la personne présentait un trouble psychiatrique qui abolissait ou altérait sa capacité de discernement et/ou sa capacité à contrôler ses actions. Dans un tel cas, le juge ordonne généralement une hospitalisation psychiatrique sans consentement. Les objectifs de cette étude étaient de (1) décrire les hospitalisations psychiatriques des irresponsables en France, (2) identifier l'âge, le sexe et les diagnostics principaux de ces individus, et (3) caractériser les trajectoires de leurs soins psychiatriques avant et après l'hospitalisation psychiatrique (67).

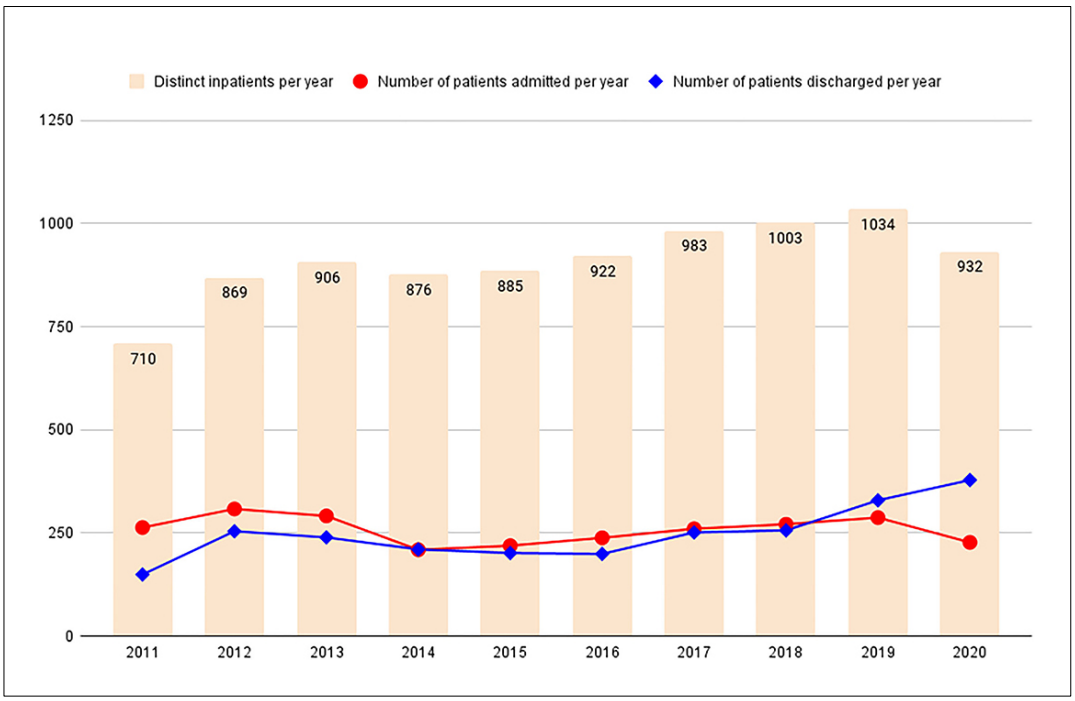


Figure 40: Nombre de personnes jugées irresponsables et hospitalisées en France (2011–2020)

Le nombre d'admissions sur ces motifs est resté stable au cours de cette période, allant de 263 en 2011 à 227 en 2021. Il s'agissait principalement de jeunes hommes diagnostiqués avec un trouble psychotique (62%). La majorité (87%) a été hospitalisée dans des hôpitaux psychiatriques généraux, et seuls 13% ont été admis dans des unités de haute sécurité (Unités pour malades difficiles, UMD). La durée médiane d'hospitalisation pour ces patients était de 13 mois. Nos résultats montrent que 73 % des patients avaient déjà été hospitalisés avant leur hospitalisation

pour trouble mental non criminel ayant mené à une décision. Le taux de ré-hospitalisations dans les 5 ans suivant la sortie de l'hospitalisation psychiatrique pour trouble mental non criminel était de 62 %.

4.1.3 Hospitalisations psychiatriques des personnes détenues en France

Bien que l'état de santé mentale précaire des personnes incarcérées soit évident, peu d'études ont examiné le nombre d'hospitalisations psychiatriques dans cette population. Depuis 2010, la France a progressivement ouvert neuf services psychiatriques hospitaliers à temps plein exclusivement dédiés aux personnes incarcérées, appelés "unités hospitalières spécialement aménagées" (UHSA, 440 lits). Cette étude visait à présenter les taux annuels d'hospitalisations psychiatriques et les diagnostics psychiatriques principaux chez les personnes incarcérées en France de 2009 à 2019, à partir du SNDS (68).

Entre le 1er janvier 2009 et le 31 décembre 2019, 32 228 personnes incarcérées (92,2 % d'hommes, n = 29 721 ; 7,8 % de femmes, n = 2 507) ont été hospitalisées pour des soins psychiatriques (64 481 séjours). Les principaux diagnostics étaient les troubles psychotiques (27,4 %), les troubles de la personnalité (23,2 %) et les troubles liés au stress (20,2 %). Le nombre annuel de personnes incarcérées hospitalisées en soins psychiatriques est passé de 3 263 en 2009 à 4 914 en 2019. La Figure 41 représente l’augmentation annuelle en pourcentage (comparée au nombre mesuré en 2009) du nombre de (1) personnes incarcérées hospitalisées en établissements psychiatriques (en bleu), (2) personnes non incarcérées hospitalisées en soins psychiatriques (en jaune), et (3) personnes incarcérées (en vert) en France (2009-2019). L'augmentation progressive de l'activité des unités hospitalières spécialement aménagées (UHSA) (300 hospitalisations en 2010 contre 3 252 en 2019) n'a pas été associée à une réduction du taux d'hospitalisation des personnes incarcérées dans les hôpitaux psychiatriques locaux (Figure 42).

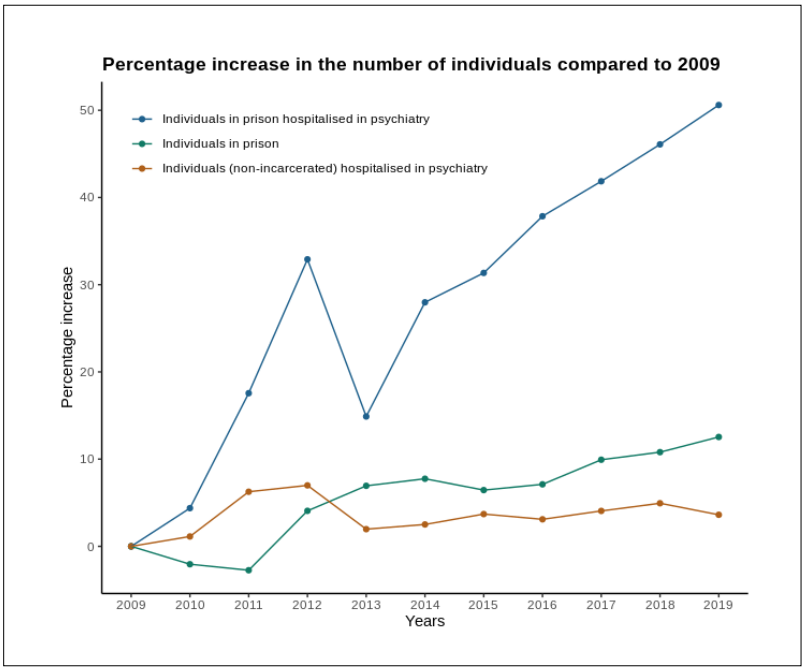


Figure 41: Augmentation annuelle des personnes détenues hospitalisées en psychiatrie, des personnes détenues, et des patients hospitalisées en psychiatrie

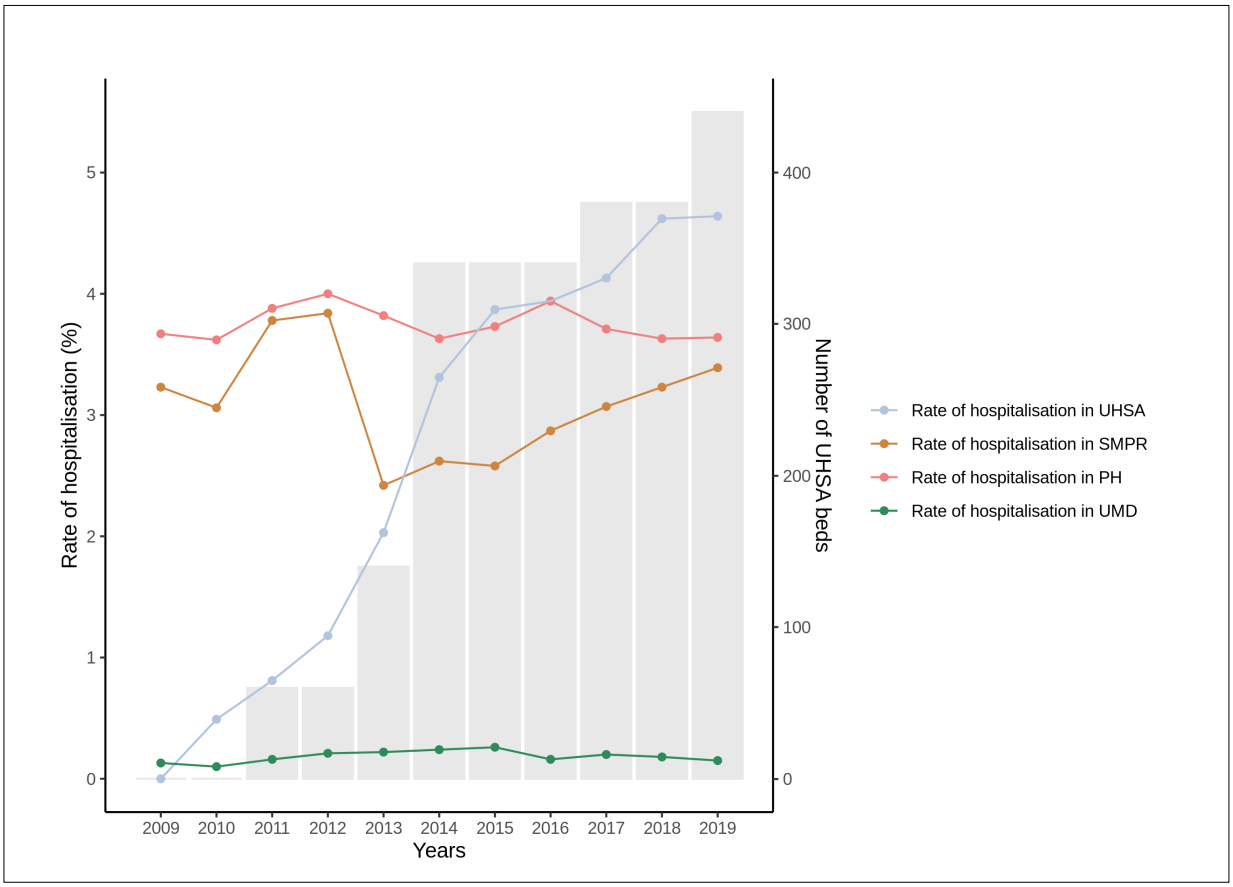


Figure 42: Taux annuels d'hospitalisation psychiatrique pour les personnes incarcérées par type d'établissement et nombre de lits dans les UHSA (barres grises) entre 2009 et 2019 en France.

4.1.4 Description des soins psychiatriques et des rapports psychiatriques pré-sentenciels dans une prison française de haute sécurité

Le centre pénitentiaire de Château-Thierry accueille des détenus dont l'adaptation à un milieu carcéral standard est considérée difficile en raison de problèmes comportementaux. Ce travail est la première étude à décrire le parcours judiciaire et médical de ces détenus (69). Tous les détenus hébergés dans la section "quartier maison centrale" entre mai et septembre 2019 ont été inclus dans cette étude transversale. Sur la période étudiée, 68 des 70 détenus (97 %) ont été pris en compte, avec l'analyse de 92 évaluations psychiatriques pré-sentencielles. Il s'agissait exclusivement d'hommes d'environ 40 ans, avec un statut socio-économique bas et des antécédents judiciaires fréquents (79 % avaient déjà été confrontés à la justice avant leur incarcération). Près de la moitié avait été hospitalisée en psychiatrie avant leur détention (46 %), et pendant leur incarcération, plus de trois quarts ont eu recours à une hospitalisation psychiatrique (79 %). Leur expérience carcérale était marquée par des sanctions disciplinaires fréquentes (72 % ayant séjourné au moins une fois en quartier disciplinaire) et des condamnations pour des infractions commises en détention (57 %). Lorsque des évaluations psychiatriques ont été réalisées (29 personnes ont eu une seule évaluation, 27 en ont eu plusieurs), dans près de la moitié des cas (44 %), au moins un psychiatre a conclu à une "altération du discernement". La figure 43 présente les résultats des expertises psychiatriques pré-sentencielles des personnes détenues au centre pénitentiaire de Château-Thierry. Chaque cercle représente une expertise (1er cercle = première expertise, 2^{e} cercle = deuxième expertise, etc.). Les couleurs représentent le degré de discernement retenu par l’expert psychiatre pour l’expertise concernée. Au total, 56 personnes détenues ont bénéficié d’au moins une expertise psychiatrique pré-sentencielle. Nous pouvons lire cette figure comme suit. Pour 19 d’entre elles, le premier expert sollicité a conclu à un discernement altéré. Parmi ces 19 personnes, 8 ont bénéficié d’une deuxième expertise (3 expertises ont conclu à un discernement altéré, 2 expertises à un discernement aboli et 3 expertises à un discernement ni altéré ni aboli). Cinq des 8 personnes concernées ont ensuite bénéficié d’une 3^{e} et dernière expertise.

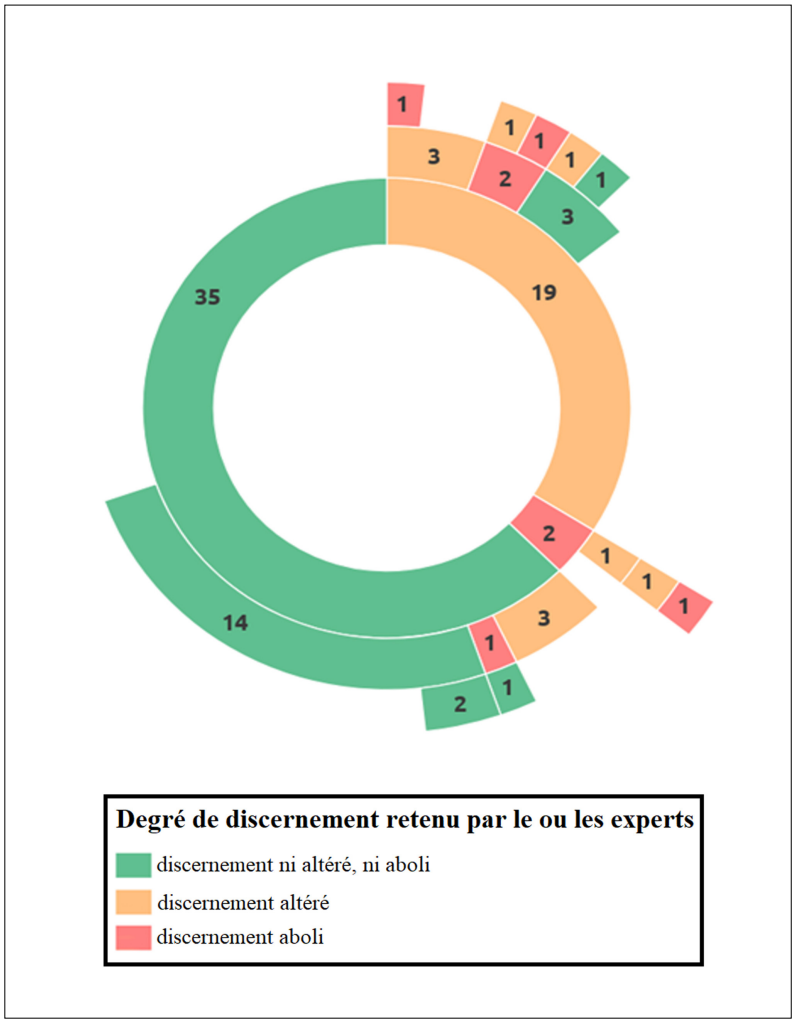


Figure 43: Expertises psychiatriques pré-sentencielles des personnes détenues au centre pénitentiaire de Château-Thierry

Cette étude souligne la prévalence des troubles psychiatriques chez les détenus de Château-Thierry, remettant en question l'absence d'alternatives à l'incarcération en France pour les personnes considérées pénalement responsables de manière atténuée.

4.1.5 Hospitalisations psychiatriques en unités pour malades difficiles en France

Les données épidémiologiques de base sont rares concernant l'activité des établissements spécialisés en psychiatrie légale en France. A partir de la base de données du PMSI, nous avons étudié l'activité des dix unités pour malades difficiles (UMD) françaises (640 lits), en décrivant les caractéristiques et l'évolution des hospitalisations psychiatriques dans ces unités entre 2012 et 2021 (70). Lors de cette période, 4857 patients ont été hospitalisés en UMD (6082 séjours). Parmi eux, 897 (18,5%) ont eu plus d'un séjour. Le nombre d'admissions variait entre 434 et 632 par an, et le nombre de sorties variait entre 473 et 609 par an. La durée de séjour médiane (Q1-Q3) était de 7,3 (4,0-14,4) mois. Parmi les 6082 séjours, 5721 (94,1%) concernaient des patients de sexe masculin. L'âge médian (Q1-Q3) était de 33 (26-41) ans. Les diagnostics psychiatriques principaux

les plus fréquents étaient les troubles psychotiques et les troubles de la personnalité. Le nombre de personnes hospitalisées dans des structures spécialisées en psychiatrie légale est stable depuis 10 ans en France et reste inférieur à celui de la plupart des pays européens.

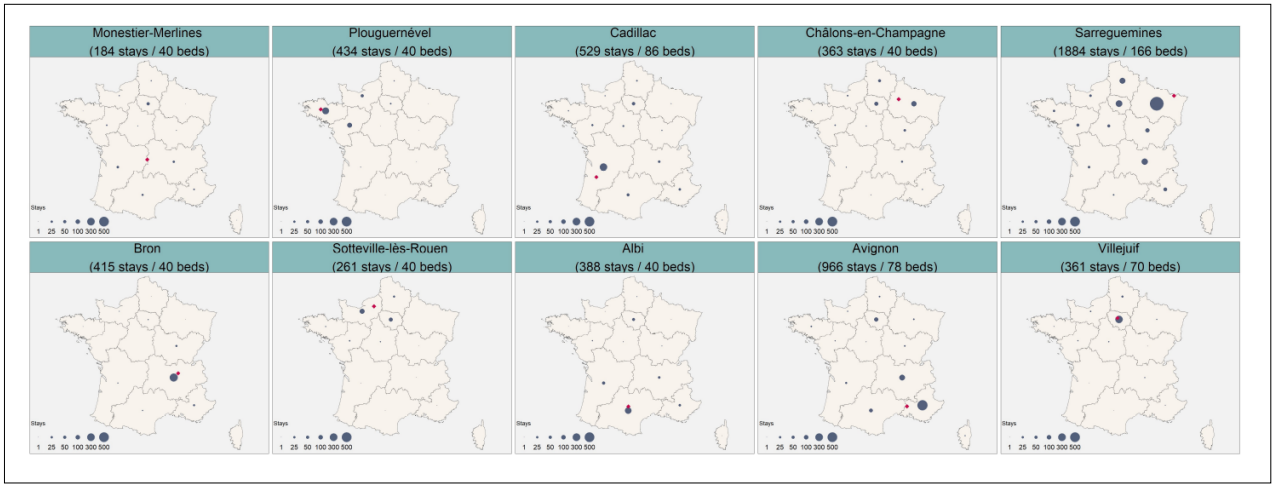
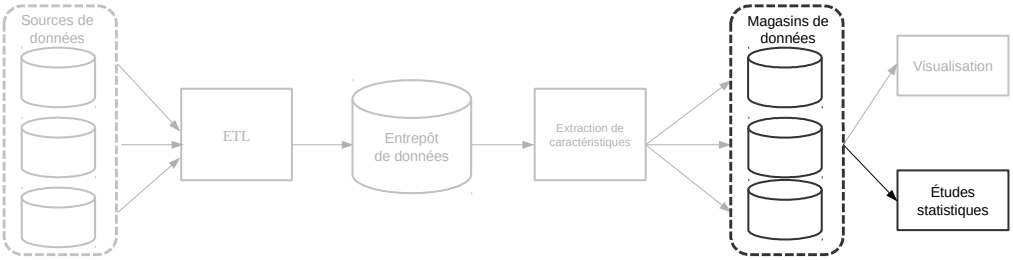


Figure 44: Répartition des régions de domiciliation des patients pour les séjours de chaque UMD entre 2012 et 2021

Bien que les UMD n'aient pas de vocation régionale, nous avons constaté que ces établissements accueillaient principalement des personnes résidant dans la région où ils se trouvaient. Sur la figure 44, pour chaque UMD (représenté par des points rouges sur les cartes), le nombre total de séjours sur la période 2012-2021 est présenté entre parenthèses, ainsi que le nombre de lits dans l'établissement. La répartition des séjours selon la région de résidence des patients admis est représentée sur la carte, pour chaque UMD, par des points bleus (la taille du point étant proportionnelle au nombre absolu de patients résidant dans une région donnée, admis dans l'UMD concerné). La proportion moyenne de patients résidant dans la même région que l'UMD dans laquelle ils ont été hospitalisés était de 55 (14) %, variant de 40 % au minimum à 81 % au maximum.

4.2 Santé publique

Étape :	
Publications	Perrot J, Hamel JF, **Lamer A**, Levaillant M. The Relationship between the Immigrant Rate and Health Status in the General Population in France. J Pers Med. 2021 Jun 30;11(7):627.
	Levaillant M, Rony L, Hamel-Broza JF, Soula J, Vallet B, **Lamer A**. In France, distance from hospital and health care structure impact on outcome after arthroplasty of the hip for proximal fractures of the femur. J Orthop Surg Res. 2023 Jun 9;18(1):418.
	Levaillant M, Garabédian C, Legendre G, Soula J, Hamel JF, Vallet B, **et al**. In France, the organization of perinatal care has a direct influence on the outcome of the mother and the newborn: Contribution from a French nationwide study. Int J Gynaecol Obstet. 2023 Jul 24;

4.2.1 Relation entre le taux d'immigration et l'état de santé dans la population générale

Principalement étudiée au niveau individuel, l'analyse de l'état de santé des immigrants à un niveau populationnel peut offrir une perspective différente pour investiguer, intégrant les déterminants sociaux dans l'explication de leur relation avec l'état de santé en France. Nous avons analysé des bases de données librement accessibles, gérées par des organismes publics français.

Les variables dépendantes étaient le taux de mortalité et l'âge moyen au décès. Le taux d'immigration et les covariables associées à l'un ou l'autre de ces résultats ont été explorés dans des modèles univariés et multivariés. Des modèles linéaires ont été utilisés pour expliquer l'âge moyen au décès, tandis que des modèles Tobit ont été utilisés pour expliquer le taux de mortalité. Le taux d'immigration variait considérablement d'un département à l'autre, tout comme l'accessibilité aux soins de santé, le profil d'âge de la population et les covariables économiques. En considérant les modèles univariés, presque toutes les covariables étudiées étaient significativement associées aux résultats. Le taux d'immigration était associé à un taux de mortalité plus faible et à un âge de décès plus bas. Dans les modèles multivariés, le taux d'immigration n'était plus associé à l'âge au décès mais restait négativement associé au taux de mortalité. La figure 45 représente la distribution des variables dépendantes et des variables co-variables qui y étaient significativement associées dans les les analyses multivariées.

En France, les départements avec une proportion plus élevée d'immigrants étaient ceux avec un taux de mortalité plus faible, possiblement parce que les immigrants sont attirés par les zones économiquement prospères.

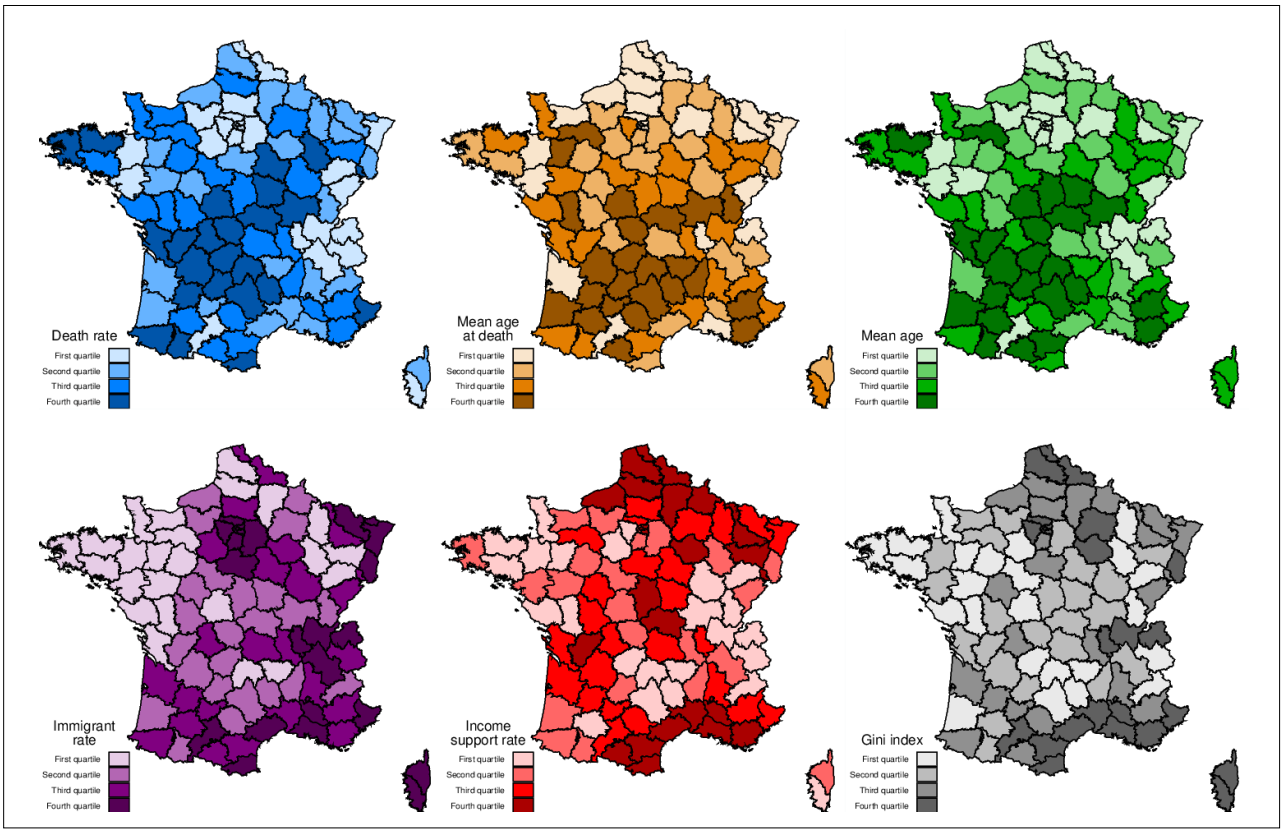


Figure 45: Co-variables associées significativement avec le taux de mortalité et l'âge au décès dans les analyses mutlivariées.

4.2.2 Arthroplastie de la hanche pour les fractures proximales du fémur

Ces dernières années, de nombreuses études ont étudié la relation entre le volume d'activité d'un établissement et le devenir du patient. Même si ces résultats semblent être associés plusieurs types de chirurgies, les données fournies ne sont pas suffisantes pour établir des seuils chirurgicaux ni pour fermer les centres à faible volume.

La prothèse de hanche est une procédure fréquemment réalisée en chirurgie orthopédique, réalisée dans presque toutes les hôpitaux pour la fracture et la coxarthrose. Avec cette étude, nous avons souhaité identifier les facteurs chirurgicaux, liés aux soins de santé et territoriaux influençant la mortalité et la réadmission des patients après une prothèse de hanche pour une fracture du col du fémur en 2018 en France (71). Les données ont été du SNDS, pour tous les patients ayant subi une arthroplastie de hanche pour une fracture du col du fémur en 2018. Le critère de jugement des patients était la mortalité à 90 jours et le taux de réadmission à 90 jours après la chirurgie.

Sur les 36 252 patients ayant subi une prothèse de hanche pour fracture en France en 2018, 0,7 % sont décédés dans les 90 jours suivant l'année et 1,2 % ont été réadmis. Le sexe masculin et l'indice de comorbidité de Charlson étaient associés à un taux de mortalité à 90 jours et de réadmission plus élevé dans l'analyse multivariée. Un volume élevé était associé à un taux de mortalité plus faible. Ni le temps de trajet ni la distance jusqu'à l'établissement de santé n'étaient associés à la mortalité ni au taux de réadmission dans l'analyse.

Même si le volume semble être associé à un taux de mortalité plus faible même pour des distances et des temps de trajet plus longs, la persistance de facteurs exogènes non documentés dans les bases de données françaises suggère que la régionalisation de la prothèse de hanche devrait être organisée avec prudence.

4.2.3 Soins périnatals et influence directe sur l'issue pour la mère et le nouveau-né

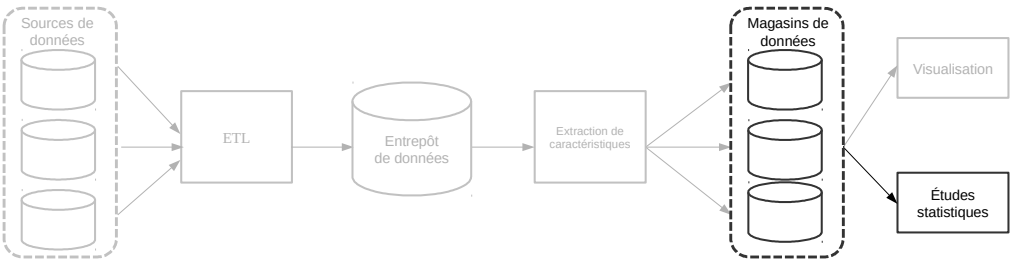
Dans le suite de cette première étude, nous avons étudié les résultats maternels et néonatals après un accouchement en France en 2019, en fonction des caractéristiques hospitalières, ainsi que l'impact de la distance et du temps de trajet sur la mère et le nouveau-né (72).

Toutes les parturientes de plus de 18 ans ayant accouché en 2019 et identifiées dans la base de données de l'assurance maladie française, avec leurs nouveau-nés, ont été incluses dans cette étude de cohorte rétrospective. Les principaux critères de jugement étaient le score de Morbidité Maternelle Sévère et l'Indicateur d'Événements Indésirables Néonatals.

Parmi les 733 052 grossesses incluses, 10 829 ont présenté une morbidité maternelle sévère (1,48 %) et 77 237 ont eu un événement indésirable néonatal (10,4 %). Les facteurs associés à un résultat maternel ou néonatal défavorable étaient l'Indice de Comorbidité Obstétricale, la primiparité et l'accouchement par césarienne ou instrumental. La prématurité était associée à une morbidité maternelle moins sévère mais à davantage d'événements indésirables néonatals. Un temps de trajet supérieur à 30 minutes était associé à une valeur plus élevée du score de morbidité maternelle.

Les résultats suggèrent l'efficacité de la régionalisation des soins périnatals en France, bien qu'une différence persiste dans les résultats en fonction du volume des unités, ce qui évoque la nécessité d'une étape supplémentaire dans la concentration des soins périnatals. L'organisation des soins périnatals devrait se concentrer sur la cartographie du territoire avec des maternités de haut niveau et à haut volume sur l'ensemble du territoire ; cela suggère la fermeture des unités à haut volume et l'amélioration de celles à faible volume pour maintenir une cartographie cohérente.

4.3 Autres disciplines

Étape :	
Publications :	Moussa MD, Beyls C, **Lamer A**, Roksic S, Juthier F, Leroy G, et al. Early hyperoxia and 28-day mortality in patients on venoarterial ECMO support for refractory cardiogenic shock: a bicenter retrospective propensity score-weighted analysis. Crit Care. 2022 Aug 26;26(1):257.
	Erlich C, **Lamer A**, Moussa MD, Martin J, Rogeau S, Tavernier B. End-tidal Carbon Dioxide for Diagnosing Anaphylaxis in Patients with Severe Postinduction Hypotension.

	Anesthesiology. 2022 Jan 18;
	Moussa MD, Rousse N, Abou Arab O, **Lamer A**, Gantois G, Soquet J, et al. Subclavian versus femoral arterial cannulations during extracorporeal membrane oxygenation: A propensity-matched comparison. J Heart Lung Transplant. 2022 May;41(5):608–18.
	Moussa MD, Soquet J, **Lamer A**, Labreuche J, Gantois G, Dupont A, et al. Evaluation of Anti-Activated Factor X Activity and Activated Partial Thromboplastin Time Relations and Their Association with Bleeding and Thrombosis during Veno-Arterial ECMO Support: A Retrospective Study. J Clin Med. 2021 May 17;10(10):2158.
	Rozencwajg S, Blet A, **Lamer A**, Boisson M, Clavier T, Abou-Arab O, et al. SARS-CoV-2 vaccination efficacy on hospitalisation and variants. Anaesth Crit Care Pain Med. 2021 May 1;40(3):100874.
	Degoul S, Chazard E, **Lamer A**, Lebuffe G, Duhamel A, Tavernier B. Intraoperative administration of 6% hydroxyethyl starch 130/0.4 is not associated with acute kidney injury in elective non-cardiac surgery: A sequential and propensity-matched analysis. Anaesthesia Critical Care & Pain Medicine. 2020 Apr 1;39(2):199–206.
	Moussa MD, Durand A, Leroy G, Vincent L, **Lamer A**, Gantois G, et al. Central venous-to-arterial PCO2 difference, arteriovenous oxygen content and outcome after adult cardiac surgery with cardiopulmonary bypass: A prospective observational study. Eur J Anaesthesiol. 2019 Jan 16;
	Fruchart M, Quindroit P, Patel H, Beuscart J, Calafiore M, Lamer A. Implementation of a Data Warehouse in Primary Care: First Analyses with Elderly Patients. MIE. 2022;
	Fruchart M, Lamer A, Lemaitre M, Beuscart JB, Calafiore M, Quindroit P. Description of a French Population of Diabetics Treated Followed up by General Practitioners. Stud Health Technol Inform. 2023 May 18;302:856–60.

4.3.1 Anesthésie-Réanimation

4.3.1.1 Dioxyde de carbone expiré comme outil de diagnostic de l'anaphylaxie chez les patients présentant une hypotension sévère post-induction

Les réactions d'hypersensibilité périopératoires peuvent être difficiles à diagnostiquer pendant une anesthésie générale. L'hypotension post-induction est le signe le plus courant mais n'est pas spécifique. Il a récemment été suggéré que de faibles niveaux de dioxyde de carbone expiré (ETco2) pourraient être un marqueur d'anaphylaxie (réactions d'hypersensibilité immédiate de grade III à IV selon les grades de Ring et Messmer) chez les patients hypotendus sous ventilation mécanique. Pour tester cette hypothèse, nous avons comparé l'ETco2 chez des patients diagnostiqués avec une anaphylaxie et chez des patients présentant une hypotension sévère due à une autre cause après l'induction de l'anesthésie (73).

Il s'agissait d'une étude de cas-témoins rétrospective à centre unique dans laquelle deux groupes ont été formés à partir de l'EDS d'anesthésie du CHU de Lille. Le groupe anaphylaxie a été formé sur la base de données d'analyse de tryptase/histamine et de données d'enquête allergique enregistrées sur la période 2010-2018. Le groupe témoin (hypotension) était constitué de tous les patients ayant présenté une hypotension sévère (pression artérielle moyenne inférieure à 50 mmHg pendant 5 minutes ou plus) pour une cause autre que l'anaphylaxie après l'induction de l'anesthésie en 2017.

Les groupes anaphylaxie et hypotension comprenaient respectivement 49 patients (grade III : n = 38 ; grade IV : n = 11) et 555 patients. La valeur minimale d'ETco2 était significativement plus basse dans le groupe anaphylaxie (médiane [intervalle interquartile] : 17 [12 à 23] mmHg) que

dans le groupe hypotension (32 [29 à 34] mmHg ; P < 0,001) (Figure 46). L'aire sous la courbe ROC (IC 95 %) pour l'ETco2 était de 0,95 (0,91 à 0,99). La sensibilité et la spécificité (IC 95 %) pour la valeur seuil optimale étaient respectivement de 0,92 (0,82 à 0,98) et 0,94 (0,92 à 0,99) (Figure 47). Dans une analyse multivariée, la valeur minimale d'ETco2 était associée à l'anaphylaxie après ajustement pour les facteurs confusionnels et les prédicteurs concurrents, notamment la pression artérielle, la fréquence cardiaque et la pression maximale des voies respiratoires (rapport de cotes [IC 95 %] pour l'ETco2 : 0,51 [0,38 à 0,68] ; P < 0,001).

En cas d'hypotension sévère après l'induction de l'anesthésie, un faible ETco2 contribue au diagnostic de l'anaphylaxie, en plus des signes classiques d'hypersensibilité immédiate périopératoire.

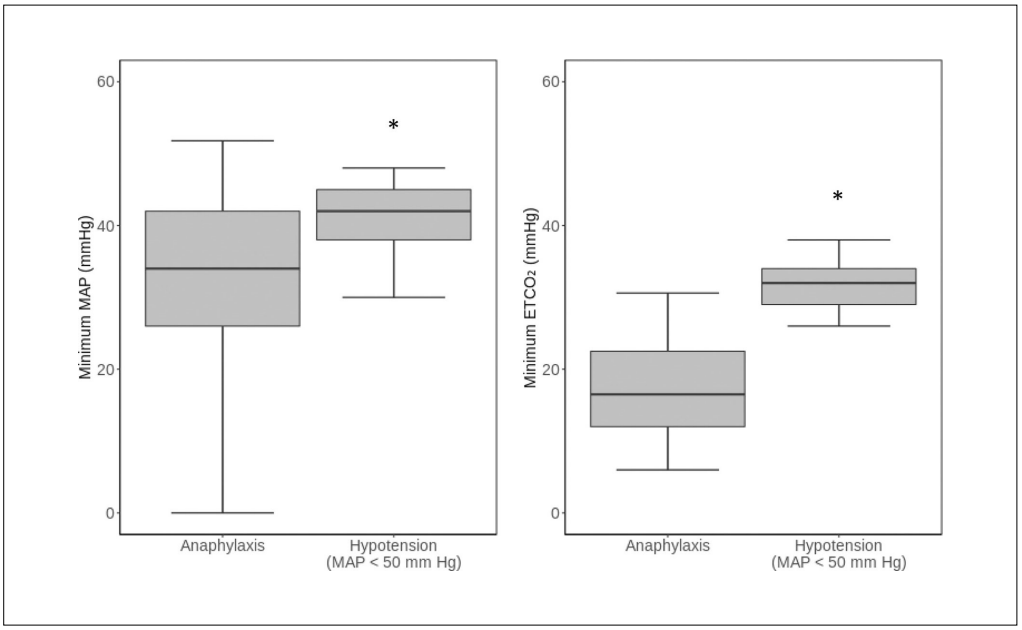


Figure 46: Pression artérielle moyenne minimale et dioxyde de carbone expiré minimum chez les patients souffrant d'anaphylaxie par rapport aux patients présentant une hypotension post-induction

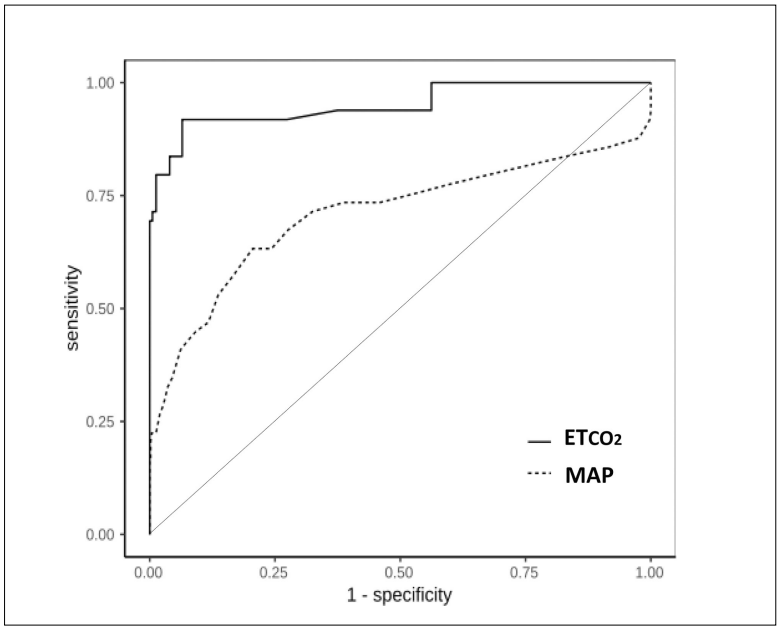


Figure 47: Courbe ROC de la capacité du dioxyde de carbone expiré minimum (ETco2) et de la pression artérielle moyenne minimum (MAP) à différencier l'anaphylaxie de l'hypotension post-induction

4.3.1.2 Administration peropératoire d'hydroxyéthylamidon et risque d'insuffisance rénale aigüe

La prise en charge hémodynamique peropératoire nécessite généralement la perfusion de solutions cristalloïdes ou colloïdales. Cependant, les solutions colloïdales présentent plusieurs effets indésirables, notamment sur la fonction rénale. Plusieurs essais contrôlés randomisés ont spécifiquement démontré que l'hydroxyéthylamidon (HEA), une forme de solution colloïdale, accroissait le risque d'insuffisance rénale aiguë (IRA). Ce risque était particulièrement notable chez les patients en état de choc septique par rapport à l'utilisation de gélatines et de cristalloïdes (74,75).

Les données sur de grandes populations font défaut pour évaluer si l'utilisation peropératoire d'HEA pouvait augmenter le risque d'IRA postopératoire. A partir de l'EDS d'anesthésie du CHU de Lille, nous avons cherché à évaluer si l'administration intraopératoire de 6 % d'HEA 130/0,4 était associée à une IRA dans le cas de chirurgies non cardiaques (76). Dans cette étude rétrospective, nous avons inclus les chirurgies abdominales, urologiques, thoraciques et vasculaires périphériques électives de 2010 à 2015. Les patients traités avec et sans HEA ont été appariéss et comparées à l'aide d'un score de propension. L' IRA postopératoire, définie par le stade 3 du score KDIGO (Kidney Disease Improving Global Outcomes), était le critère principal. Comme l'utilisation d'HEA a considérablement diminué en 2013, des analyses supplémentaires, limitées à la période 2010-2012, ont également été réalisées (Figure Figure 48).

Au total, 23 045 et 11 691 patients ont été inclus sur les périodes 2010-2015 et 2010-2012, respectivement. La réduction de l'utilisation d'HEA n'a été accompagnée d'aucun changement dans l'incidence de l'IRA. L'association non ajustée entre HEA et l'IRA de stade 3 selon le KDIGO était significative (OR [IC à 95 %] de 2,13 [1,67, 2,71]). Pour la période entière, 6460 patients ont été appariés. Les odds-ratios pour l'IRA de stade 3 et pour l'IRA de tous les stades lors de

l'utilisation d'HEA (10,3 ± 4,7 mL.kg-1) étaient respectivement de 1,20 (IC à 95 % [0,74, 1,95]) et de 1,21 (IC à 95 % [0,95, 1,54]). Aucune association n'a été observée avec la transplantation rénale ou la mortalité hospitalière. Des résultats similaires ont été obtenus pour la période restreinte.

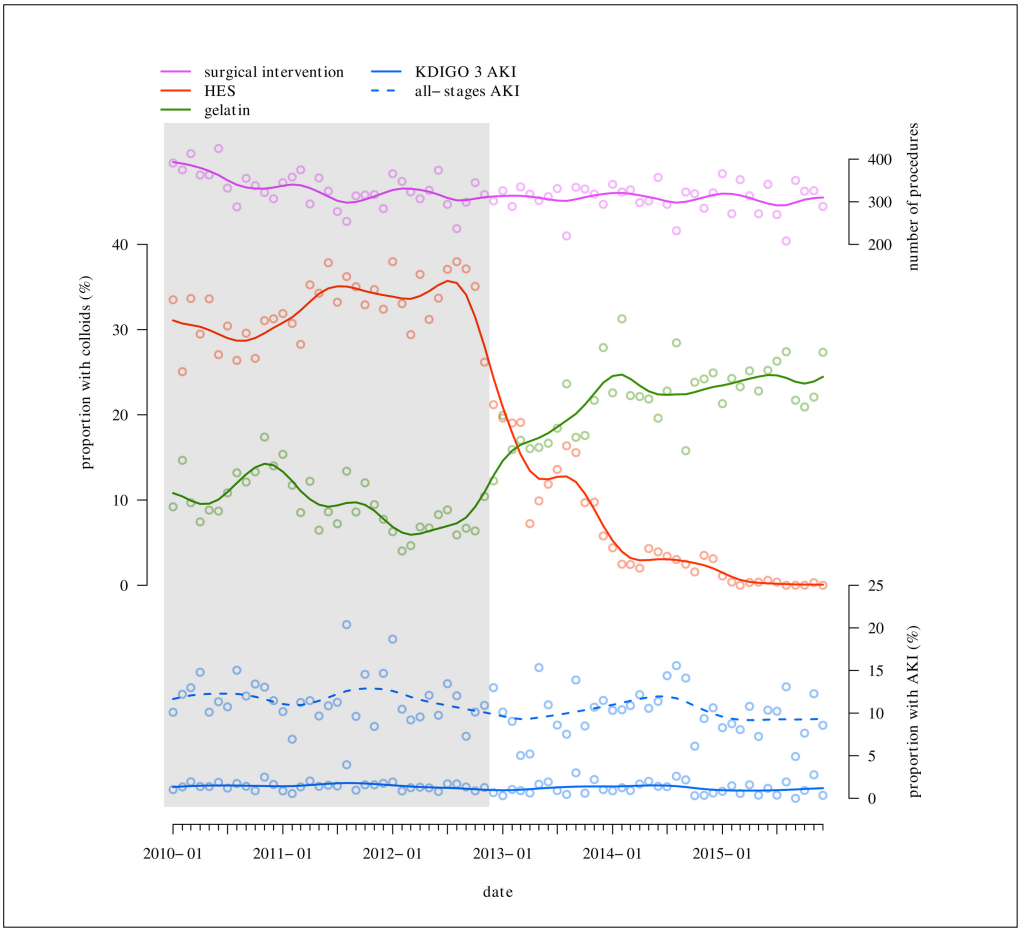
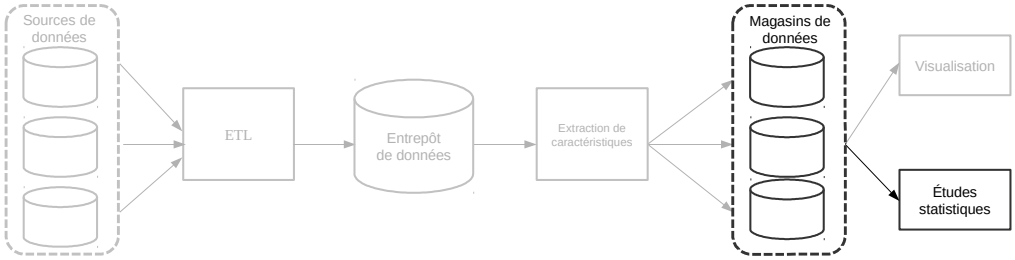


Figure 48: Séries temporelles mensuelles avec ajustement par spline cubique confrontant l'utilisation de colloïdes et l'insuffisance rénale aigüe

L'utilisation d'HEA 130/0,4 n'était pas associée à un risque accru d'IRA. Aucune conclusion ne peut être tirée pour des doses plus élevées d'HEA.

4.3.2 Soins premiers

Étape :	
Publications :	Fruchart M, Quindroit P, Patel H, Beuscart J, Calafiore M, Lamer A. Implementation of a Data Warehouse in Primary Care: First Analyses with Elderly Patients. MIE. 2022;
	Fruchart M, Lamer A, Lemaitre M, Beuscart JB, Calafiore M, Quindroit P. Description of a French Population of Diabetics Treated Followed up by General Practitioners. Stud Health Technol Inform. 2023 May 18;302:856–60.

Les travaux d'intégration de données et de standardisation que nous avons présenté en partie 3.2.2 nous ont permis d'exploiter les données de soins primaires de la MSPU de Wattrelos. L'objectif d'une première étude était d'évaluer la faisabilité de réutilisation des données de soins primaires, et en particulier d'étudier les consultations et les prescriptions du patient âgé contenues dans notre entrepôt de données de soins primaires (49).

La population âgée (de plus de 75 ans) représente 5,5 % de la population du cabinet, soit 900 patients pour 13 867 consultations (7,6 % des consultations) (Figure 49). Cette population a bénéficié en moyenne de 4,5 consultations par an avec 5,97 (3,72) prescriptions de médicaments par consultation.

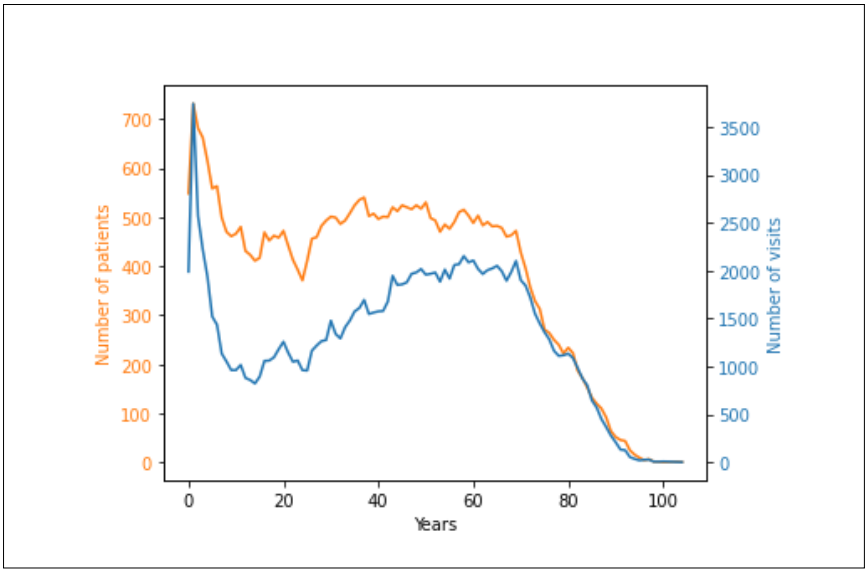


Figure 49: Nombre de patients et consultations par âge (49)

Les médicaments les plus prescrits traitaient la douleur (Paracétamol) pour 10 % des prescriptions, suivis des anti-inflammatoires non-stéroïdiens (Acide acétylsalicylique) pour 2,5 % et de l'Ésoméprazole pour 2,3 % des prescriptions.

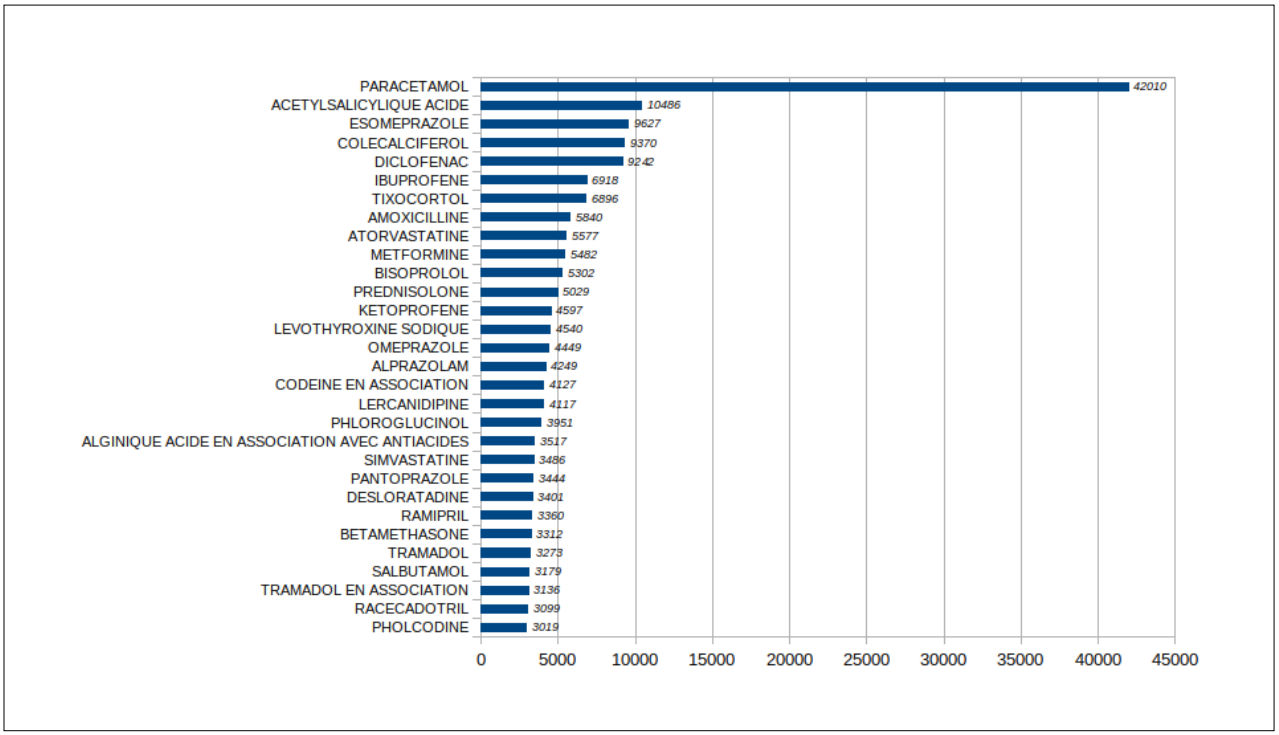


Figure 50: Médicaments les plus prescrits pour la population de plus de 75 ans (49)

La réutilisation des données de soins primaires permet d'étudier des données cliniques ambulatoires telles que les résultats de laboratoire et les prescriptions médicamenteuses, qui ne sont pas documentées dans les bases de données des réclamations et des hôpitaux. Dans cette étude, nous avons sélectionné la population des diabétiques traités dans l'entrepôt de données de soins primaires Wattrelos, dans le nord de la France. Un deuxième travail portait sur la description des patients diabétiques (77). Dans un premier temps, nous avons étudié les résultats de laboratoire des diabétiques en identifiant si les recommandations de la Haute Autorité de Santé (HAS) étaient respectées. Dans un second temps, nous avons étudié les prescriptions des diabétiques en identifiant les traitements par hypoglycémiants oraux et par insulines.

La population diabétique représente 690 patients de la MSPU. Les recommandations sur le laboratoire sont respectées pour 84% des diabétiques. La majorité des diabétiques sont traités par hypoglycémiants oraux 68,6%. Comme le recommande la HAS, la metformine est le traitement de première intention dans la population diabétique.

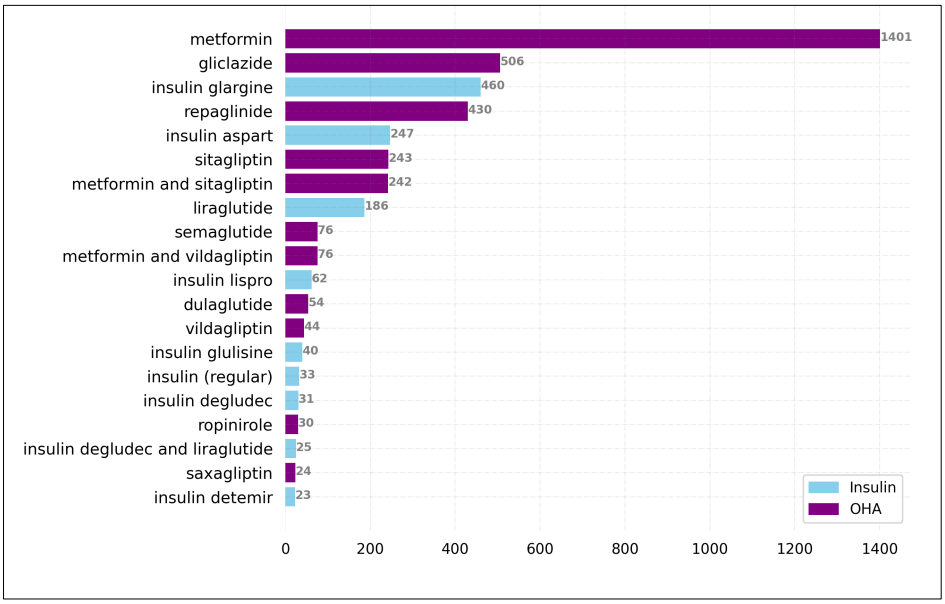


Figure 51: Prescription d'antidiabétiques (77)

Sur la période de 2015 à 2020, 9 435 patients adultes ont consulté au centre médical de Wattrelos. La population adulte diabétique représente 690 patients (soit 7,3 % des patients adultes) traités avec des médicaments antidiabétiques. L'âge moyen (écart type) de ces patients est de 65,7 (13,2) ans lors de la consultation, et la population comprend 46,8 % de femmes. La population de diabétiques représente en moyenne (écart type) 3 152 (785,4) consultations par an (soit 12,8 % de la population adulte), 15 596 (2 387,9) prescriptions de médicaments (23,5 % de la population adulte) et 21 031 (2 699,8) résultats de laboratoire par an (18,5 % de la population adulte). En ce qui concerne les recommandations de la HAS, en moyenne, 84,4 % des patients diabétiques ont effectué plus de deux résultats de laboratoire HbA1c, et 84,3 % ont réalisé au moins un résultat de laboratoire pour la créatinine. La majorité des patients (68,6 %) se sont vu prescrire uniquement des OHAs, tandis que 22,2 % ont reçu des prescriptions pour à la fois de l'insuline et des OHAs au cours de l'année (Figure 2). En ce qui concerne les prescriptions générales (sans prescriptions d'antidiabétiques), le paracétamol (7,0 %), l'acide acétylsalicylique (4,4 %) et l'atorvastatine (3,5 %) sont les plus prescrits (Figure 3). Pour les deux catégories d'antidiabétiques, la metformine (43,8 % des OHAs) et l'insuline glargine (41,5 % des traitements à l'insuline) étaient les plus prescrits (Figure 3).

Chapitre 5 Conclusion et perspectives

5.1 Conclusion des travaux réalisés

Dans ce mémoire d'habilitation à diriger des recherches, j'ai exposé mes contributions dans le domaine de la réutilisation des données en santé. Mes travaux s'adressent à toute la chaîne de la réutilisation des données, à savoir la collecte des données depuis des sources variées, leur intégration au sein d'un entrepôt de données, leur standardisation vers un modèle de données commun, l'extraction de caractéristiques, et la visualisation des résultats. Ces développements méthodologiques ont été concrétisés à travers diverses applications cliniques, notamment en psychiatrie, en santé publique, en anesthésie-réanimation et en soins primaires.

Il existe encore cependant de nombreuses pistes d'exploration et d'amélioration.

1. Malgré une utilisation croissante, les **données ouvertes demeurent largement méconnues et insuffisamment exploitées**. De plus, la **communauté scientifique n'investigue que très peu les réseaux sociaux**, des plateformes encore mal appréhendées, souvent fréquentées par une population très jeune. Cette sous-exploration laisse échapper des informations cruciales concernant les opinions exprimées librement sur ces plateformes, qui pourraient enrichir les connaissances acquises lors des interactions avec les professionnels de santé. Des événements significatifs comme par exemple l'augmentation de publications et de réactions concernant un nouveau « challenge » lié à l'anorexie ou des effets secondaires liés à la prise d'un traitement, pourraient également servir de signaux d'alerte importants, d'indicateurs épidémiologiques en ligne, et permettre de mieux appréhender la représentation des soins et des maladies.

2. **L'extraction de caractéristiques reste encore peu abordée dans la littérature scientifique** et nécessiterait une normalisation approfondie afin de rendre les études reproductibles et d'améliorer l'accès à ces données.

3. En comparaison avec la dynamique internationale, **la communauté scientifique française présente un retard significatif dans l'adoption du modèle OMOP**, dans le développement des outils et méthodes qui y sont adossés, et dans la participation à des études internationales de grande ampleur à partir de ce modèle de données. De même, malgré l'importance fondamentale des technologies open-source en data science, le partage d'outils reste rare.

4. Les domaines de la data science et de l'informatique sont dynamiques, avec l'émergence régulière de nouvelles méthodes et technologies. Cependant, il existe un **écart considérable entre l'activité des professionnels qui manipulent les données au quotidien et les thématiques de recherche en science des données et en informatique médicale orientées par les chercheurs**. Ce décalage est également observé entre les professionnels qui soignent les patients et ceux qui sont engagés dans la recherche.

5. **Les professionnels de santé font face à un manque de formation en recherche**, ce qui se traduit par des demandes de réutilisation des données souvent inadaptées aux bases de données disponibles et aux designs d'études envisageables. Cette situation engendre une perte de temps considérable pour les équipes méthodologiques et aboutit rarement à des résultats valorisables. De plus, il existe un **manque de confiance parmi les professionnels de santé concernant la remontée de propositions novatrices et l'engagement dans des projets de recherche**, alors que le pratique quotidienne est source d'idées. Enfin, les professionnels de santé et les data

scientists ne comprennent pas assez le potentiel et les limites respectives de leur domaine. Cette lacune entrave l'interaction entre ces métiers et complique également l'utilisation optimale des données pour la recherche.

Dans la section 5.2, nous présenterons les pistes qui nous animeront ces prochaines années pour lever ces barrières.

5.2 Perspectives thématiques

Dans les années à venir, nos efforts se concentreront sur la résolution des barrières et l'amélioration des points de blocage identifiés dans la section 5.1.

Ce projet s'articulera à partir des points suivants : (i) automatiser la collecte des données, l'intégration et l'évaluation de la qualité des données, en particulier à partir des réseaux sociaux ou des nouvelles sources de données qui apparaîtront; (ii) formaliser l'extraction de caractéristiques et proposer des outils et méthodes pour améliorer la reproductibilité de ce processus ; (iii) proposer un pipeline d'analyse épidémiologique afin d'orienter vers les stratégies optimales d'analyses statistique et faciliter l'accès et la compréhension des résultats ; (iv) favoriser l'implémentation de modèles de données commun et développement et le partage des méthodes et outils open-source dédiés à la recherche ; (v) former les professionnels de santé et les data scientists à la recherche, et favoriser les échanges entre les différentes spécialités. Voici les axes que nous aborderons pour lever ces obstacles :

5.2.1 Collecte et intégration des données

Une approche anticipative consisterait à **surveiller la diffusion de nouveaux ensembles de données ouvertes** émis par des institutions, ainsi que l'émergence de nouveaux producteurs de données, notamment au sein des réseaux sociaux. Pour les sources de données pertinentes, nous entreprendrons la **caractérisation de leur format et contenu**. Dans le cas des données spatiales et temporelles, il sera essentiel de **normaliser les dimensions des bases de données** afin d'assurer une uniformité dans les granularités spatiales et temporelles pour favoriser l'interopérabilité et la capacité à croiser les données provenant de différentes sources. Nous veillerons tenir à jour et/ou garder un historique des modifications dans les découpages territoriaux qui évolueraient.

Nous aborderons la mise en place d'une collecte automatique de données sur des plateformes comme Reddit, Twitter, Youtube, Instagram, Snapchat et TikTok. L'objectif sera de créer un modèle standard de stockage des publications pour générer des rapports synthétiques et repérer rapidement des contenus potentiellement alarmants.

Nous continuerons nos efforts pour évaluer la qualité des données en mettant en œuvre la taxonomie exposée dans ce mémoire. Nous proposerons une **évaluation automatisée des problèmes de qualité et une synthèse des résultats via un tableau de bord**, accompagnés de scores associés aux différentes dimensions de la qualité des données. Cet outil pourrait servir à deux fins : (i) évaluer et comparer de manière systématique toute nouvelle source de données, (ii) surveiller les alimentations d'EDS et détecter des variations inhabituelles dans leur qualité.

5.2.2 Extraction de caractéristiques

L'extraction de caractéristiques nécessite une standardisation approfondie pour éviter la « boîte noire » et améliorer la reproductibilité des analyses. Il est crucial de guider et formaliser les échanges avec les professionnels de santé pour recueillir leurs définitions de ces caractéristiques. Nous proposons d'investiguer une **méthode mixte intégrant approches qualitative et quantitative, exploitant l'expertise des professionnels pour créer les caractéristiques**. Cette méthodologie inclurait (i) une sélection aléatoire des dossiers de patients, (ii) une analyse cas par cas par les professionnels de santé pour identifier les éléments clés qui composeraient les caractéristiques, (iii) le développement et l'exécution de l'algorithme pour calculer les caractéristiques, et (iv) la validation manuelle par les professionnels de santé des transformations des données brutes en caractéristiques. La mise en œuvre de cette méthodologie pourrait être intégrée à la *checklist* Strobe et décrite systématiquement.

Nous estimons aussi essentiel d'améliorer la **stockage pérenne et optimal des caractéristiques et des méta-données** qui caractérisent leur calcul. Pour cela, nous investiguerons le *feature store* (magasin de caractéristiques) et les technologies big data. Nous pourrons nous inspirer des concepts et outils déjà mis en œuvre dans des domaines tels que le e-commerce, un secteur accoutumé au stockage de catalogues volumineux.

5.2.3 Pipeline d'analyse épidémiologique

Nous souhaitons proposer un **pipeline méthodologique** qui intégrerait plusieurs outils pour aborder de manière efficace des questions de recherche ou faciliter le processus de décision, comme prévoir à court-terme et moyen-terme les besoins de santé sur des territoires. Cette méthode aborderait deux aspects clés : (i) orienter vers les stratégies optimales d'analyses statistique et (ii) faciliter l'accès et la compréhension des résultats.

Ce processus reposera sur une proposition exhaustive des méthodes statistiques et épidémiologiques utilisables. Il sera appuyé par un arbre décisionnel pour sélectionner et appliquer la méthode statistique la plus appropriée. Chaque méthode sera préalablement testée et documentée.

Pour faciliter l'accès aux résultats et leur compréhension, nous créerons des outils interactifs destinés à la communauté scientifique, aux autorités publiques et aux professionnels sur le terrain. En guidant les décisions à travers des données scientifiques, cette approche permettra à la communauté de s'approprier cet outil comme un support pour les prises de décision.

5.2.4 Développement et implémentation d'outils et technologies open-source pour la recherche

En tirant parti de notre expérience dans l'intégration des données d'anesthésie-réanimation, de soins primaires et du SNDS, nous visons à **encourager davantage l'adoption de modèles de données standardisés** et à **développer l'écosystème d'outils et méthodes open-sources**.

Nous souhaitons proposer un **processus générique d'ETL adapté au modèle OMOP** pour déployer plus rapidement des entrepôts de données dans ce format. Ce processus reposerait sur une architecture commune et un paramétrage indépendants de la source de données.

5.2.5 Accompagnement et formation des professionnels de santé pour la recherche et l'amélioration des pratiques

Nous aspirons à **établir des liens interdisciplinaires solides** qui encouragent des échanges fructueux à toutes les étapes des projets de recherche. Cette collaboration implique un rapprochement entre les data scientists et les ingénieurs, spécialistes de la manipulation quotidienne des données, et les professionnels de santé qui interagissent directement avec les patients. Cette coopération vise également à i**dentifier et à mettre en lumière les besoins et les idées**, réduisant ainsi l'écart entre la recherche académique et les réalités du terrain.

Dans les **ouvertures aux autres professionnels de santé**, nous aborderons les fonctions du champ médico-social (i.e., infirmiers, psychologues, orthophonistes, ergothérapeutes.) et proposeront des études qui dépassent le parcours de soins pour aller vers le parcours de vie.

5.3 Organisation et structuration du projet de recherche à venir

Au cours des prochaines années, ces thématiques de recherche seront menées dans quatre cadres principaux, l'ULR2694 METRICS, l'UFR3S ILIS de l'Université de Lille, la Fédération régionale de recherche en psychiatrie et santé mentale des Hauts-de-France, et l'association InterHop.

En collaboration entre l'ULR2694, l'UFRS ILIS et la F2RSM, nous mettrons l'accent sur la **formation à la recherche des étudiants** en formation initiale et sur la création d'un vivier de chercheurs émergents. Cela se concrétisera à travers les cours d'initiation à la recherche et de data science en Licence 3, Master 1 et Master 2 à ILIS. Les stages et projets intégrés à ces enseignements offriront aux étudiants désireux de poursuivre vers une thèse universitaire un accompagnement adapté. À l'heure actuelle, ce dispositif a donné lieu à deux thèses en cours menées par Mathilde Fruchart et Chloé Saint-Dizier. Nous sommes en préparation pour les prochaines thèses prévues en 2024 et 2025.

Nous continuerons également à **dynamiser l'activité de recherche à ILIS**, en la valorisant comme une vitrine pour les étudiants et les évaluateurs (HCERES). Cela incluera la diffusion des résultats de recherche des enseignants-chercheurs, le soutien à la présentation des travaux réalisés par les étudiants lors de congrès scientifiques, ainsi que l'organisation de séminaires de recherche.

Dans notre démarche visant à développer les compétences des professionnels de santé en matière de recherche, nous implémenterons des **programmes de formation à la recherche** ainsi que des **ateliers pratiques**. Ces formations et ateliers pourront conduire à la **réalisation conjointe de projets de recherche** impliquant des chercheurs expérimentés de la F2RSM et des professionnels de santé des établissements de santé mentale et psychiatrie.

Au sein d'InterHop et de la F2RSM, nous aurons également pour objectif de **développer et de partager des outils open-source et libres**. Nous envisageons de mettre en place une plateforme de data science pour l'échange de scripts d'analyse, des tutoriels de statistiques et de graphiques, pour favoriser ainsi la diffusion et l'apprentissage dans le domaine de la recherche en santé.

Nous continuerons à **animer la communauté scientifique OMOP en France** en collaboration avec l'association InterHop, à travers la mise en place de plusieurs groupes de travail. Ces

groupes auront pour objectif d'introduire les nouveaux collaborateurs à l'utilisation du modèle OMOP, de les former à la mise en œuvre des processus ETL, et de les accompagner dans la réalisation d'études multicentriques. Ils seront également des espaces dédiés pour recueillir les retours d'expérience sur la réutilisation des données, l'adoption d'outils et de technologies open-source, l'exploitation de données ouvertes, et pour mener des projets collaboratifs visant à co-développer des outils et des méthodes.

Nous jouerons un rôle actif dans l'**obtention de nouveaux financements**, notamment en répondant à des appels à projets, en mettant l'accent sur les **projets centrés sur les réseaux sociaux**, des outils encore largement sous-exploités dans le domaine de la recherche.

A la suite des deux workshops organisés aux congrés MIE2022 et MIE2023, nous échangeons avec la société européenne d'informatique médicale (European Federation of Medical Informatics, EFMI) quant à la **création d'un groupe de travail européen sur la réutilisation des données**. Ce groupe de travail aurait vocation à animer la communauté européenne investie dans la thématique de réutilisation des données. Ce groupe de travail pourra bénéficier du réseau européen d'EFMI pour organiser des workshops et séminaires.

Nous envisageons de **créer un journal scientifique dédiée à la valorisation de la recherche en science des données**, et en particulier en réutilisation des données avec les problématiques liées à la collecte des données, l'intégration et le nettoyage des données, l'extraction de caractéristiques, les développements de méthodes statistiques adaptés à la réutilisation des données et aux études prospectives, et enfin la visualisation des résultats.

Bibliographie

1. Lamer A. Contribution à la prévention des risques liés à l'anesthésie par la valorisation des informations hospitalières au sein d'un entrepôt de données [Internet]. Lille 2; 2015 [cited 2016 Mar 17]. Available from: http://www.theses.fr/2015LIL2S021
2. Nahm M, Shepherd J, Buzenberg A, Rostami R, Corcoran A, McCall J, et al. Design and implementation of an institutional case report form library. Clin Trials. 2011 Feb;8(1):94–102.
3. Krishnankutty B, Bellary S, Kumar NBR, Moodahadu LS. Data management in clinical research: An overview. Indian J Pharmacol. 2012;44(2):168–72.
4. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association. 2007 Jan 1;14(1):1–9.
5. Safran C. Reuse of Clinical Data. Yearb Med Inform. 2014 Aug 15;9(1):52–4.
6. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. Health Aff (Millwood). 2015 Dec;34(12):2174–80.
7. Kim YG, Jung K, Park YT, Shin D, Cho SY, Yoon D, et al. Rate of electronic health record adoption in South Korea: A nation-wide survey. International Journal of Medical Informatics. 2017 May 1;101:100–7.
8. Esdar M, Hüsers J, Weiß JP, Rauch J, Hübner U. Diffusion dynamics of electronic health records: A longitudinal observational study comparing data from hospitals in Germany and the United States. International Journal of Medical Informatics. 2019 Nov 1;131:103952.
9. William R. Hersh MD. Adding Value to the Electronic Health Record Through Secondary Use of Data for Quality Assurance, Research, and Surveillance. American Journal of Managed Care [Internet]. 2007 Jun 1 [cited 2015 Jan 30];13(June 2007-Part 1 6-Pt 1). Available from: http://www.ajmc.com/publications/issue/2007/2007-06-vol13-n6-Pt1/Jun07-2487p277-278/
10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144–51.
11. Kimball R. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons; 1998. 801 p.
12. Inmon WH. Building the Data Warehouse. Wiley; 1992. 320 p.
13. Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. International Journal of Medical Informatics. 2016 Oct 1;94:271–4.
14. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inform. 2017 Jun;102:21–8.
15. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc JL, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne. 2015 Jun;36(6):411–7.
16. Scailteux LM, Droitcourt C, Balusson F, Nowak E, Kerbrat S, Dupuy A, et al. French administrative health care database (SNDS): The value of its enrichment. Therapie. 2019 Apr;74(2):215–23.
17. Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies--Investigating partial mappings between HPO and SNOMED CT. J Biomed Semantics. 2016;7:3.
18. Krumm R, Semjonow A, Tio J, Duhme H, Bürkle T, Haier J, et al. The need for harmonized structured documentation and chances of secondary use – Results of a systematic

analysis with automated form comparison for prostate and breast cancer. Journal of Biomedical Informatics. 2014 Oct;51:86–99.
19. Xu Y, Zhou X, Suehs BT, Hartzema AG, Kahn MG, Moride Y, et al. A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance. Drug Saf. 2015 Aug;38(8):749–65.
20. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus M. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016 Oct 28;
21. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc. 2016;23(5):909–15.
22. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. J Biomed Inform. 2016 Apr;60:352–62.
23. Liyanage H, Liaw ST, Jonnagaddala J, Hinton W, de Lusignan S. Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs. Stud Health Technol Inform. 2018;255:60–4.
24. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010 Nov 2;153(9):600.
25. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Feb;19(1):54–60.
26. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.
27. Informatics OHDS and. The Book of OHDSI [Internet]. [cited 2023 Jan 10]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/
28. Collaborators – OHDSI [Internet]. [cited 2021 Apr 5]. Available from: https://www.ohdsi.org/who-we-are/collaborators/
29. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet. 2019 16;394(10211):1816–26.
30. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MTF, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun [Internet]. 2020 Oct 6 [cited 2020 Nov 5];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7538555/
31. Chazard E, Ficheur G, Caron A, Lamer A, Labreuche J, Cuggia M, et al. Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features. Stud Health Technol Inform. 2018;255:15–9.
32. Thomas JJ, Cook KA. Illuminating the Path: The Research and Development Agenda for Visual Analytics. 2005 Jan 1 [cited 2020 Mar 17]; Available from: https://www.hsdl.org/?abstract&did=
33. Nelson O, Sturgis B, Gilbert K, Henry E, Clegg K, Tan JM, et al. A Visual Analytics Dashboard to Summarize Serial Anesthesia Records in Pediatric Radiation Treatment. Appl Clin Inform. 2019 Aug;10(4):563–9.
34. Glez-Peña D, Lourenço A, López-Fernández H, Reboiro-Jato M, Fdez-Riverola F. Web scraping technologies in an API world. Brief Bioinformatics. 2014 Sep;15(5):788–97.
35. Laurent G, Guinhouya B, Whatelet M, Lamer A. Automatic Exploitation of YouTube Data: A Study of Videos Published by a French YouTuber During COVID-19 Quarantine in France. Stud Health Technol Inform. 2020 Nov 23;275:112–6.
36. H P, R P, D Z, B G, M F, A L. Automated Twitter Extraction and Visual Analytics with Dashboards: Development and First Experimentations. Studies in health technology and informatics [Internet]. 2022 May 25 [cited 2023 May 12];294. Available from: https://pubmed.ncbi.nlm.nih.gov/35612183/

37. Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLOS ONE. 2020 Sep 25;15(9):e0239441.
38. Fruchart M, El Idrissi F, Lamer A, Belarbi K, Lemdani M, Zitouni D, et al. Identification of early symptoms of endometriosis through the analysis of online social networks: A social media study. Digit Health. 2023;9:20552076231176114.
39. Saint-Dizier C, Lamer A, Zaanouar M, Amariei A, Quindroit P. OpenDataPsy: An Open-Data Repository with Standardized Storage and Description for Research in Psychiatry. Stud Health Technol Inform. 2023 May 18;302:851–5.
40. Fichiers des personnes décédées depuis 1970 | Insee [Internet]. [cited 2023 Nov 12]. Available from: https://www.insee.fr/fr/information/4190491
41. Guardiolle V, Bazoge A, Morin E, Daille B, Toublant D, Bouzillé G, et al. Linking Biomedical Data Warehouse Records With the National Mortality Database in France: Large-scale Matching Algorithm. JMIR Med Inform. 2022 Nov 1;10(11):e36711.
42. Martignene N, Balcaen T, Bouzille G, Calafiore M, Beuscart JB, Lamer A, et al. Heimdall, a Computer Program for Electronic Health Records Data Visualization. Stud Health Technol Inform. 2020 Jun 16;270:247–51.
43. Quindroit P, Fruchart M, Degoul S, Périchon R, Soula J, Marcilly R, et al. Definition of a practical taxonomy for referencing data quality problems in healthcare databases. Methods Inf Med. 2022 Nov 10;
44. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. Appl Clin Inform. 2020 Jan;11(1):13–22.
45. Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart JB, et al. Transforming Anesthesia Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study. J Med Internet Res. 2021 Oct 29;23(10):e29259.
46. Paris N, Lamer A, Parrot A. Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study. JMIR Med Inform. 2021 Dec 14;9(12):e30970.
47. Johnson AEW, Pollard TJ, Shen L, Lehman L wei H, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3(1):1–9.
48. GitLab [Internet]. 2021 [cited 2023 Nov 15]. Antoine Lamer / omop_anesthesia · GitLab. Available from: https://gitlab.com/antoinelamer/omop_anesthesia
49. Fruchart M, Quindroit P, Patel H, Beuscart J, Calafiore M, Lamer A. Implementation of a Data Warehouse in Primary Care: First Analyses with Elderly Patients. MIE. 2022;
50. Lamer A, Jeanne M, Ficheur G, Marcilly R. Automated Data Aggregation for Time-Series Analysis: Study Case on Anaesthesia Data Warehouse. Stud Health Technol Inform. 2016;221:102–6.
51. Sessler DI, Sigl JC, Kelley SD, Chamoun NG, Manberg PJ, Saager L, et al. Hospital stay and mortality are increased in patients having a "triple low" of low blood pressure, low bispectral index, and low minimum alveolar concentration of volatile anesthesia. Anesthesiology. 2012 Jun;116(6):1195–203.
52. Lamer A, Jeanne M, Marcilly R, Kipnis E, Schiro J, Logier R, et al. Methodology to automatically detect abnormal values of vital parameters in anesthesia time-series: Proposal for an adaptable algorithm. Comput Methods Programs Biomed. 2016 Jun;129:160–71.
53. Lamer A, Fruchart M, Paris N, Popoff B, Payen A, Balcaen T, et al. Standardized Description of the Feature Extraction Process to Transform Raw Data Into Meaningful Information for Enhancing Data Reuse: Consensus Study. JMIR Med Inform. 2022 Oct 17;10(10):e38936.
54. Lamer A, Laurent G, Pelayo S, El Amrani M, Chazard E, Marcilly R. Exploring Patient Path Through Sankey Diagram: A Proof of Concept. Stud Health Technol Inform. 2020 Jun 16;270:218–22.

55. Lamer A, Moussa MD, Marcilly R, Logier R, Vallet B, Tavernier B. Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project. J Clin Monit Comput. 2022 Aug 6;
56. Boudis F, Clement G, Bruandet A, Lamer A. Automated Generation of Individual and Population Clinical Pathways with the OMOP Common Data Model. Stud Health Technol Inform. 2021 May 27;281:218–22.
57. Laurent G, Moussa MD, Cirenei C, Tavernier B, Marcilly R, Lamer A. Development, implementation and preliminary evaluation of clinical dashboards in a department of anesthesia. J Clin Monit Comput. 2020 May 16;
58. Lamer A, Ficheur G, Rousselet L, van Berleere M, Chazard E, Caron A. From Data Extraction to Analysis: Proposal of a Methodology to Optimize Hospital Data Reuse Process. Stud Health Technol Inform. 2018;247:41–5.
59. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007 Oct 20;370(9596):1453–7.
60. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. Int J Radiat Oncol Biol Phys. 2017 Oct 1;99(2):344–52.
61. Mangold P, Filiot A, Moussa M, Sobanski V, Ficheur G, Andrey P, et al. A Decentralized Framework for Biostatistics and Privacy Concerns. Stud Health Technol Inform. 2020 Nov 23;275:137–41.
62. Lamer A, Filiot A, Bouillard Y, Mangold P, Andrey P, Schiro J. Specifications for the Routine Implementation of Federated Learning in Hospitals Networks. Stud Health Technol Inform. 2021 May 27;281:128–32.
63. Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: A case study in France. PLOS Digit Health. 2023 Jul;2(7):e0000298.
64. Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny ML, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019 Aug 21;264:1536–7.
65. Wathelet M, Duhem S, Vaiva G, Baubet T, Habran E, Veerapa E, et al. Factors Associated With Mental Health Disorders Among University Students in France Confined During the COVID-19 Pandemic. JAMA Netw Open. 2020 Oct 23;3(10):e2025591.
66. Levaillant M, Wathelet M, Lamer A, Riquin E, Gohier B, Hamel-Broza JF. Impact of COVID-19 pandemic and lockdowns on the consumption of anxiolytics, hypnotics and antidepressants according to age groups: a French nationwide study. Psychol Med. 2021 Dec 14;1–7.
67. Fovet T, Baillet M, Horn M, Chan-Chee C, Cottencin O, Thomas P, et al. Psychiatric Hospitalizations of People Found Not Criminally Responsible on Account of Mental Disorder in France: A Ten-Year Retrospective Study (2011-2020). Front Psychiatry. 2022;13:812790.
68. Fovet T, Chan-Chee C, Baillet M, Horn M, Wathelet M, D'Hondt F, et al. Psychiatric hospitalisations for people who are incarcerated, 2009-2019: An 11-year retrospective longitudinal study in France. EClinicalMedicine. 2022 Apr;46:101374.
69. Beigné M, Lamer A, Eck M, Horn M, Benbouriche M, Thomas P, et al. [A descriptive study of psychiatric care and pre-sentencing psychiatric reports in a French high-security prison]. Encephale. 2022 Mar 21;S0013-7006(22)00031-8.
70. Fovet T, Saint-Dizier C, Wathelet M, Horn M, Thomas P, Guillin O, et al. Opening the black box of hospitalizations in French high-secure psychiatric forensic units. Encephale. 2023 May 26;S0013-7006(23)00079-9.
71. Levaillant M, Rony L, Hamel-Broza JF, Soula J, Vallet B, Lamer A. In France, distance from hospital and health care structure impact on outcome after arthroplasty of the hip for proximal fractures of the femur. J Orthop Surg Res. 2023 Jun 9;18(1):418.

72. Levaillant M, Garabédian C, Legendre G, Soula J, Hamel JF, Vallet B, et al. In France, the organization of perinatal care has a direct influence on the outcome of the mother and the newborn: Contribution from a French nationwide study. Int J Gynaecol Obstet. 2023 Jul 24;
73. Erlich C, Lamer A, Moussa MD, Martin J, Rogeau S, Tavernier B. End-tidal Carbon Dioxide for Diagnosing Anaphylaxis in Patients with Severe Postinduction Hypotension. Anesthesiology. 2022 Jan 18;
74. Schortgen F, Lacherade JC, Bruneel F, Cattaneo I, Hemery F, Lemaire F, et al. Effects of hydroxyethylstarch and gelatin on renal function in severe sepsis: a multicentre randomised study. Lancet. 2001 Mar 24;357(9260):911–6.
75. Brunkhorst FM, Engel C, Bloos F, Meier-Hellmann A, Ragaller M, Weiler N, et al. Intensive insulin therapy and pentastarch resuscitation in severe sepsis. N Engl J Med. 2008 Jan 10;358(2):125–39.
76. Degoul S, Chazard E, Lamer A, Lebuffe G, Duhamel A, Tavernier B. Intraoperative administration of 6% hydroxyethyl starch 130/0.4 is not associated with acute kidney injury in elective non-cardiac surgery: A sequential and propensity-matched analysis. Anaesthesia Critical Care & Pain Medicine. 2020 Apr 1;39(2):199–206.
77. Fruchart M, Lamer A, Lemaitre M, Beuscart JB, Calafiore M, Quindroit P. Description of a French Population of Diabetics Treated Followed up by General Practitioners. Stud Health Technol Inform. 2023 May 18;302:856–60.
78. 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 May 18;302:502–3.