Méthodes de clustering fondées sur les réseaux pour stratifier des patients à partir de données longitudinales et hiérarchiques

Judith Lambert

Université Paris Cité
École doctorale Pierre Louis de santé publique : Épidémiologie et Sciences de l'Information Biomédicale - ED393
Centre de Recherche des Cordeliers (UMR-S 1138)
Équipe HeKA

Méthodes de clustering fondées sur les réseaux pour stratifier des patients à partir de données longitudinales et hiérarchiques

Thèse de doctorat de Biostatistique et Biomathématiques

Par Judith LAMBERT

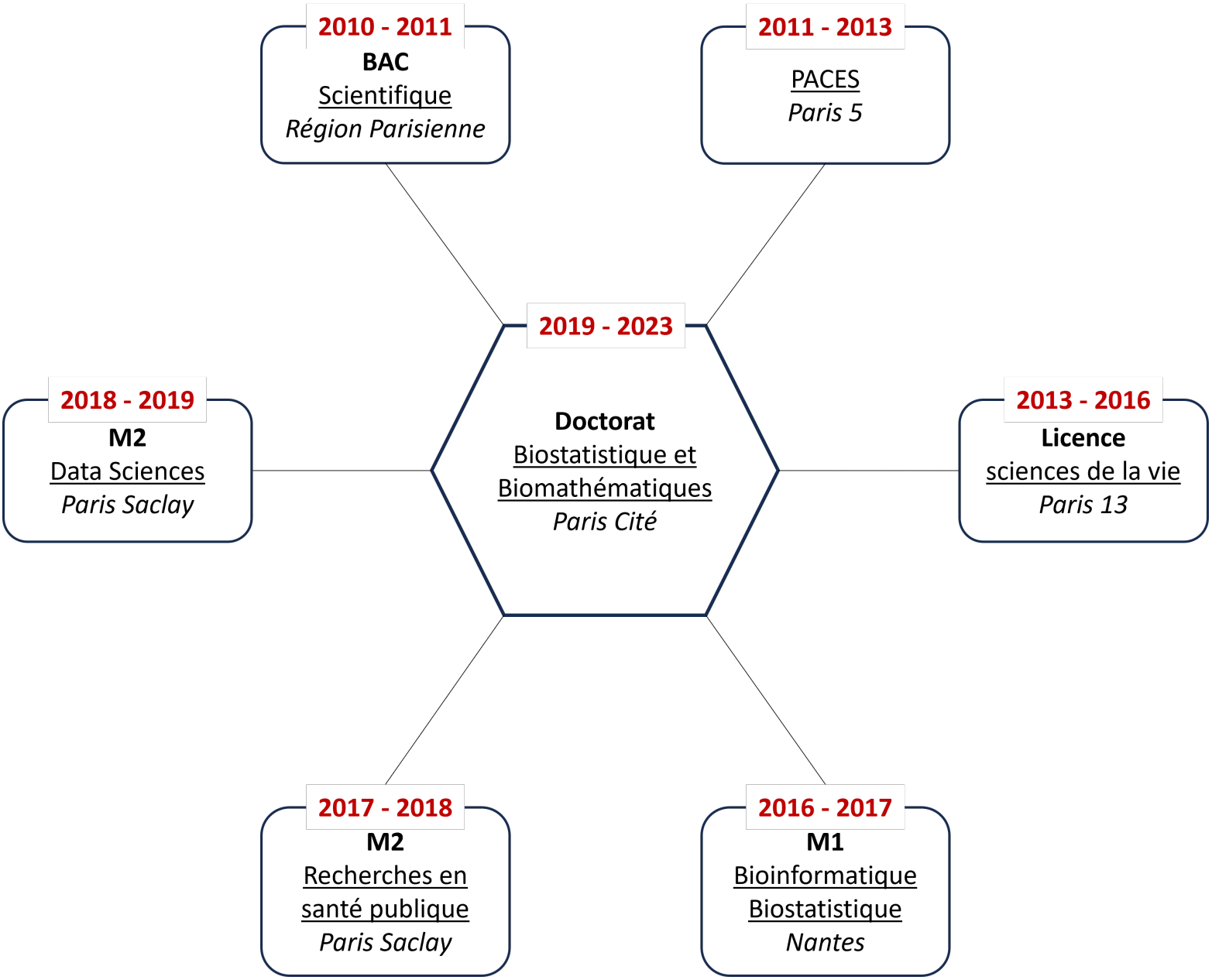
Sous la direction de Anne-Sophie JANNOT
et la co-direction de Anaïs BAUDOT

Présentée et soutenue publiquement le 21 décembre 2023

Devant un jury composé de :

Jean-Baptiste BEUSCART	PU-PH, Université de Lille	Rapporteur
Nicolas JAY	PU-PH, Université de Lorraine	Rapporteur et Président du jury
Jean-Daniel FEKETE	DR, Université Paris-Saclay	Examinateur
Emmanuelle GÉNIN	DR, Université de Brest	Examinatrice
Hervé PERDRY	MCU-HDR, Université Paris-Saclay	Examinateur
Anne-Louise LEUTENEGGER	CR, Université Paris Cité	Membre invité
Anne-Sophie JANNOT	MCU-PH, Université Paris Cité	Directrice de thèse
Anaïs BAUDOT	DR, Université Aix-Marseille	Co-directrice de thèse

Parcours


2010 - 2011
BAC
Scientifique
Région Parisienne
2011 - 2013
PACES
Paris 5
2019 - 2023
Doctorat
Biostatistique et
Biomathématiques
Paris Cité
2018 - 2019
M2
Data Sciences
Paris Saclay
2013 - 2016
Licence
sciences de la vie
Paris 13
2017 - 2018
M2
Recherches en
santé publique
Paris Saclay
2016 - 2017
M1
Bioinformatique
Biostatistique
Nantes

Remerciements

Je tiens à remercier Jean-Baptiste Beuscart et Nicolas Jay d'avoir accepté d'être rapporteurs de ma thèse et d'évaluer ce travail. Je remercie également Jean-Daniel Fekete, Emmanuelle Genin et Hervé Perdry pour avoir accepté d'examiner ce travail.

C'est tout naturellement que je remercie mes deux directrices de thèse, Anaïs Baudot et Anne-Sophie Jannot pour avoir choisi de me faire confiance pour ce travail. Et je tiens également à remercier Anne-Louise Leutenegger qui a consacré son temps à l'encadrement de ma thèse. Elles ont toutes les trois toujours été présentes pour me guider et me conseiller, que ce soit pour les expérimentations, la préparation aux séminaires ou l'écriture d'articles combien fastidieux pour moi. Elles ont toujours trouvé les mots pour m'encourager et me rassurer quand je ne comprenais pas ou dans les moments de découragements.

Je remercie tous ceux qui ont contribué de près ou de loin à ce travail : Pierre Sabatier qui m'a permis d'avoir accès aux données et qui a pris le temps de me les expliquer ; Anthony qui m'a si souvent aidé en apportant son point de vue et de nouvelles idées ; Louise Labatie avec qui j'ai eu la chance de travailler lors de son stage en lien avec ma thèse ; les membres du projet GOLD et toutes les personnes que j'ai rencontré lors des séminaires et qui m'ont posé des questions ou qui ont partagé leurs expériences de travail afin que j'améliore le mien.

Un grand merci à tous les membres de la team Heka, croisés au CRC ou à PSC, avec qui j'ai partagé des moments précieux : Lillian et Safa grâce à qui mon niveau d'anglais est bien meilleur et avec qui j'ai partagé de beaux moments culinaires ; Juliette, un ange tombée du ciel, qui prend le temps d'aider et d'écouter les autres ; Sandrine qui a toujours les mots pour faire rire ; Alice qui se bat de toute son âme pour ce qui est juste ; Olivier pour sa gentillesse et Camila pour sa bonne humeur ; tous les doctorants Linus, Agathe, Louis, Solange, Fleur, Pierre et Enora et les post-doc Jong Ho et Nadim pour leur bienveillance, sans oublier Aurélie et Maïsen qui ont su apporter une bouffée de fraîcheur.

Je remercie également tous les membres de la team MMG croisés à la Timone ou à Luminy, avec qui j'ai eu des discussions enrichissantes. Mes remerciements vont à Ozan, Céline, David, Nadine ainsi qu'Elisabeth, Laurent et Brigitte pour leur accueil si chaleureux.

Je n'oublie pas les mousquetaires, Morgane et Elva, avec qui j'ai partagé de beaux moments et qui m'ont supporté quand je venais à Marseille, même sans prévenir.

Un merci spécial aux "jeunes cadres" pour leur joie de vivre et leur soutien.

Je remercie Ursule dont j'ai eu le privilège de croiser la route un jour, et ce sans aucun regret.

Je tiens à remercier Sarah, dont le coeur est en or.

Je remercie Mélissa pour son sens du partage et dont la folie est contagieuse.

Je suis reconnaissante envers Fafa pour son soutien inconditionnel.

Je remercie tout spécialement ma soeur Viputha de toujours croire en moi, de m'encourager et de m'aimer.

Je remercie la love family pour la joie et le bonheur qu'ils m'apportent et je remercie tout le reste de la famille.

Un merci tout particulier à la dynamique, Farah et Béthanyel, pour avoir toujours su trouver les mots et les gestes pour me rassurer, me soutenir et m'encourager.

Je remercie plus que tout ma Mounette, mon Pouni et ma Marraine qui ont su me donner les moyens de réussir par leur sacrifice, leur soutien et surtout leur amour.

Je remercie mes frères, Judicaël et Judison, pour tout leur conseil, leur amour et leur force, car à trois, on est plus fort.

Et pour finir, je tiens à remercier J.C. dont l'amour a toujours su me guider et me montrer la voie à suivre.

Résumé

Méthodes de clustering fondées sur les réseaux pour stratifier des patients à partir de données longitudinales et hiérarchiques

La stratification des patients est importante pour mieux comprendre l'hétérogénéité des maladies, évaluer l'efficacité d'un traitement et faciliter l'appariement des patients. Cette stratification repose sur des méthodes de clustering utilisant des données de santé provenant, par exemple, de bases médico-administratives. Ces données sont nombreuses et variées, qualitatives ou quantitatives, avec ou sans labels organisés en nomenclatures. De plus, ce sont des données longitudinales complexes et parfois tronquées. Ces spécificités limitent les approches de clustering.

L'objectif de ma thèse a été de développer de nouvelles approches de clustering pour identifier des sous-groupes homogènes (clusters) de patients en tenant compte de la complexité des données médico-administratives. Nous avons développé deux approches. La première, nommée "cluster-tracking", identifie des clusters dans des réseaux de patients construits à chaque période de temps. Nous avons identifié des trajectoires de clusters cliniquement significatives. De manière importante, notre approche ne nécessite pas l'imputation de données tronquées ni l'exclusion de patients. La seconde approche intègre les relations entre labels au sein des nomenclatures dans les mesures de similarités. En comparaisons avec des mesures qui ne tiennent pas compte de ces relations, nos mesures pondérées permettent d'identifier des clusters plus pertinents cliniquement.

En considérant les spécificités des données médico-administratives, nos deux nouvelles approches ont permis d'améliorer la stratification des patients en fonction de leur état de santé.

Mots-clés : clustering longitudinal de patients, *cluster tracking*, bases de données médico-administratives, réseaux de patients, connaissance préalable d'experts, labels de variables hiérarchiques, stratification de patients, mesures de similarité.

Abstract

Network-based clustering methods for patient stratification from longitudinal and hierarchical data

Patient stratification is important to better understand disease heterogeneity, assess treatment efficacy, and facilitate patient matching. This stratification relies on clustering approaches using health data from sources such as medico-administrative databases. These data are numerous and diverse, qualitative or quantitative, with or without labels organized in nomenclatures. Furthermore, these are complex and sometimes truncated longitudinal data. These specificities limit clustering approaches.

The aim of my thesis was to develop new clustering approaches to identify homogeneous subgroups (clusters) of patients while considering the complexity of medico-administrative data. We developed two approaches. The first one, called "cluster-tracking," identifies clusters in patient networks constructed at each time period. We identified clinically significant cluster trajectories. Importantly, our approach does not require imputation of truncated data or patient exclusion. The second approach incorporates label relationships within nomenclatures into similarity measures. Compared to measures that do not consider these relationships, our weighted measures allow for the identification of more clinically relevant clusters.

By considering the specificities of medico-administrative data, our two new approaches have improved patient stratification regarding their health status.

Keywords : longitudinal patient clustering, cluster tracking, medico-administrative databases, patient networks, prior expert knowledge, hierarchical variable labels, patient stratification, similarity measures.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

Productions scientifiques

Articles publiés :

- **Lambert Judith**, Leutenegger Anne-Louise, Jannot Anne-Sophie, Baudot Anaïs. Tracking clusters of patients over time enables extracting information from medico-administrative databases. Journal of Biomedical Informatics, 2023, vol. 139, p. 104309. https://doi.org/10.1016/j.jbi.2023.104309
- **Lambert Judith**, Leutenegger Anne-Louise, Baudot Anaïs, Jannot Anne-Sophie. Improving patient clustering by incorporating structured label relationships in similarity measures. medRxiv, 2023, p. 2023.06. 06.23291031. https://doi.org/10.1101/2023.06.06.23291031

Communications orales

- 2021
  - European Mathematical Genetics Meeting (EMGM), Visioconférence, "Identifying association between genotypes and patient care trajectories using patients' network"
- 2022
  - Journées de Biostatistique, Rennes (France), "Tracking clusters of patients over time enables extracting information from medico-administrative databases"
- 2023
  - Congrès EMOIS, Nancy (France), "Suivi temporel de clusters de patients pour extraire les informations à partir des bases de données médico-administratives"

Posters

- 2022
  - Congrès EMOIS, Dijon (France), "Suivi temporel de clusters identifiés à partir de réseaux de patients"
  - Medical Informatics Europe (MIE), Nice (France), "Tracking temporal clusters from patient network"
- 2023
  - Colloque Données de Santé en vie réelle, Paris (France), "Stratification de patients à l'aide de réseaux construits à partir de différentes mesures de similarité : application au SNDS"

Liste des abréviations

ATC *Anatomical Therapeutic Chemical classification system.*

B01 Agents antithrombotiques.

BIRCH *Balanced Iterative Reducing and Clustering using Hierarchies.*

CIM-10 Classification Internationale des Maladies, 10e révision.

CNAM Caisse Nationale de l'Assurance Maladie.

CURE *Clustering Using Representatives.*

CépiDC Centre d'épidémiologie sur les causes médicales de décès.

DBSCAN *Density-Based Spatial Clustering of Applications with Noise.*

DPMM *Dirichlet Process Mixture Models.*

DTW *Dynamic Time Warping*.

EGB Echantillon Généraliste des Bénéficiaires.

ELFE Étude Longitudinale Française depuis l'Enfance.

ESND Échantillon du Système National des Données de santé.

FHS *Framingham Heart Study*.

GMM *Growth Mixture Modeling*.

HMM *Hidden Markov Models.*

HPO *Human Phenotype Ontology*.

INSERM Institut National de la Santé Et de la Recherche Médicale.

KNN *K Nearest Neighbors.*

LCGA *Latent Class Growth Analysis.*

LCS *Longest Common Subsequence.*

LOINC *Logical Observation Identifiers Names and Codes.*

MCL *Markov Cluster Algorithm.*

MeSH *Medical Subject Headings.*

MIMIC *Medical Information Mart for Intensive Care.*

NHS *Nurses' Health Study.*

NIS *National Inpatient Sample.*

NLM *National Library of Medicine.*

OMS Organisation mondiale de la santé.

OPTICS *Ordering Points To Identify the Clustering Structure.*

PEPR SN Programme et Equipement Prioritaire de Recherche Santé Numérique.

PMSI Programme de Médicalisation des Systèmes d'Information.

SNDS Système National des Données de Santé.

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

SNOMED-CT *Systematized Nomenclature Of Medicine Clinical Terms.*

STS *Short Time Series distance.*

TGCC *Temporal Global Clustering Consensus.*

UMLS *Unified Medical Language System.*

Chapitre 1

Introduction

1.1 Les données de santé

1.1.1 Présentation générale

Définitions

Les données de santé sont des données qui font référence à l'état de santé physique ou mentale d'une personne ou d'une population [1]. Elles peuvent être générées principalement de deux manières : soit dans le cadre du soin, soit dans le cadre de la recherche. Les données de

soin sont collectées et utilisées dans le but de diagnostiquer, traiter et surveiller l'état de santé d'un patient tandis que les données de recherche sont destinées à des études médicales, épidémiologiques ou cliniques. Les données de santé sont fondamentales pour l'étude des maladies à l'échelle des individus et de la population. En effet, elles peuvent englober diverses informations : des données relatives à une personne telles que l'âge et le sexe, des données collectées lors d'un test ou d'un examen médical telles que les résultats de laboratoire et les données génétiques, ou des données liées à une maladie telles que les symptômes et les traitements. Toutes ces informations sont sensibles et confidentielles. Elles sont, pour cela, soumises à un régime juridique particulier.

Complexité des données de santé

Les données de santé se caractérisent par leur complexité (voir *Figure 1.1*). Tout d'abord, elles présentent un volume important. Cette abondance s'explique notamment par la multitude de sources de données disponibles. Ces sources incluent, par exemple, les données provenant du Dossier Patient Informatisé (*Electronic Health Record* en anglais) qui stocke et centralise l'ensemble des données lié au parcours de soins des patients et les données de remboursement associées aux données issues du soin [2]. Concernant les données collectées spécifiquement pour une recherche, elles peuvent provenir d'enquêtes de santé via des questionnaires par exemple, de registres, de cohortes ou encore d'essais cliniques. La numérisation croissante des données de santé facilite ainsi leur conservation et leur archivage mais génère de la complexité liée au volume des données disponibles.

Par ailleurs, les données de santé sont hétérogènes. On distingue principalement deux catégories : les données de santé structurées et les données de santé non structurées. Les données de santé structurées contiennent des informations organisées dans des formats prédéfinis avec une syntaxe précise. Cela inclut par exemple les données démographiques, les données de facturation ou encore les données codées telles que les diagnostics ou les médicaments. Grâce à cette organisation structurée, l'exploitation de ces données est facilitée. Les données de santé non structurées, au contraire, contiennent des informations dont l'organisation ne présente pas de formats prédéfinis. Elles sont souvent sous forme de texte libre que l'on retrouve, par exemple, dans les rapports médicaux, mais également sous forme d'images médicales comme les radiographies. L'exploitation de ces données non structurées est plus complexe, car elle nécessite, au préalable, une étape de pré-traitement. Ainsi, les données de santé peuvent prendre diverses formes. Cette diversité est une caractéristique importante des données de santé.

Un autre aspect essentiel des données de santé est leur nature longitudinale. En effet, les données sont recueillies à plusieurs moments dans le temps sur un même individu, ce qui permet de suivre l'évolution de ses caractéristiques de santé sur une période donnée. Ce suivi peut se faire de deux manières : soit à temps fixe, soit à temps variable. Le suivi à temps fixe consiste à recueillir les données à des intervalles de temps réguliers et prédéfinis. Dans ce cas, on a un suivi identique pour chaque individu et à des moments équidistants dans le temps. C'est le cas en général des données collectées de manière prospective pour une étude spécifique. Le suivi à temps variable consiste à recueillir les données à des intervalles de temps différents en fonction de l'apparition d'événements ou de conditions spécifiques. Il peut s'agir, par exemple, de la mesure de la glycémie chez un patient diabétique lors d'une hospitalisation liée à son diabète. C'est le cas des données issues du soin.

La complexité des données de santé se caractérise également par la présence de données manquantes. Au cours du suivi d'un patient, les données manquantes peuvent survenir en raison, par exemple, du manquement d'une visite médicale ou de l'oubli de renseigner une mesure par le praticien. Un cas particulier de données manquantes est celui des données tronquées. En général, les données disponibles couvrent une période de la vie du patient plutôt que la totalité de sa vie. Par conséquent, il y a une absence de données à la fois avant et après la recherche ou l'extraction des données de soin. Cela se traduit par une troncature à gauche et à droite. Dans la suite, je me focaliserai particulièrement sur les données tronquées.

La complexité des données de santé peut rendre leur exploitation difficile. Toutefois, cette exploitation demeure essentielle pour la prise de décisions médicales, la recherche clinique et l'amélioration des soins de santé.

1.1.2 Sources de données de santé pour la recherche

Les données de santé utilisées pour une recherche biomédicale peuvent provenir de deux sources différentes, se distinguant par la manière dont ces données ont été initialement collectées. D'une part, les données peuvent être collectées spécifiquement pour une étude particulière. Cette collecte est planifiée par les chercheurs dans le but de répondre à des questions de recherche spécifiques. Ces données sont donc analysées de manière prospective. D'autre part, elles peuvent être issues de la réutilisation des données initialement collectées dans le cadre du soin. La collecte de ces données est faite dans le cadre de la prestation des soins médicaux sans intention initiale de recherche. Ces données sont donc analysées de manière rétrospective.

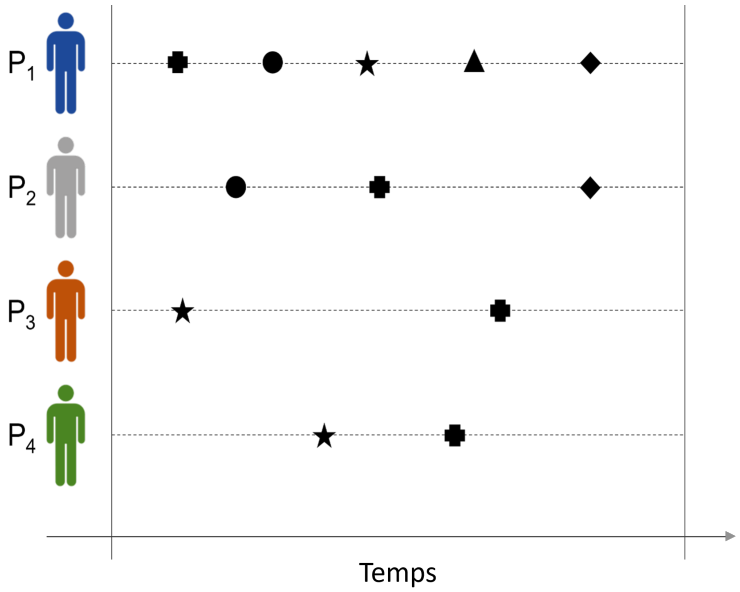


Figure 1.1 – Complexité des données de santé
Chaque patient est caractérisé par un grand nombre de données de divers formats et recueillies à des moments différents dans le temps

Données de santé collectées pour une étude spécifique

La collecte des données de santé dans le cadre d'une étude spécifique est planifiée en suivant un protocole de recherche préalablement établi. Ce protocole détaille notamment les critères d'inclusion et d'exclusion des participants et les analyses médicales nécessaires. Les données collectées permettent d'obtenir des informations précises sur un groupe d'individus particulier dans le but de répondre à des questions de recherche spécifiques. Cette collecte de données implique souvent la création de cohortes. Les cohortes regroupent des individus qui partagent des caractéristiques ou des antécédents de santé similaires et qui sont suivis pendant une période de temps définie afin d'étudier la survenue d'évènements de santé. De nombreuses cohortes existent. Par exemple, UK Biobank est une cohorte anglaise créée en 2006 dans le but d'étudier les déterminants génétiques des maladies à l'âge adulte [3]. Elle est composée de plus de 500 000 participants âgés de 49 à 69 ans. Les données phénotypiques et génétiques collectées dans cette cohorte sont issues de questionnaires, de mesures physiques, d'analyses d'échantillons biologiques, d'imagerie et de génotypage. ELFE (Étude Longitudinale Française depuis l'Enfance) est une cohorte française lancée en 2011 qui suit le développement et la santé de 20 000 enfants depuis leur naissance jusqu'à l'âge adulte [4]. Elle a pour objectif de mieux comprendre les facteurs affectant le développement, la santé et la socialisation des enfants. Les données collectées dans cette cohorte comprennent des informations médi-

cales, développementales, familiales et biologiques. Un autre exemple de cohorte française est Constances. Lancée en 2012, Constances constitue un échantillon aléatoire de 200 000 adultes âgés de 18 à 69 ans représentatif de la population adulte française [5]. Elle a pour but d'étudier les déterminants de santé au sein de la population française. Les données de cette cohorte proviennent d'examens de santé réalisés tous les cinq ans, de questionnaires auxquels les volontaires répondent annuellement, ainsi que des bases de données médico-administratives de la Caisse Nationale de l'Assurance Maladie (CNAM, voir SNDS dans la section suivante). NHS (*Nurses' Health Study*) est une cohorte américaine créée en 1976 qui suit 120 000 infirmières âgées de 30 à 55 ans [6]. Elle permet d'étudier les facteurs de risque des maladies chroniques chez les femmes. Les données sont collectées régulièrement par le biais d'entretiens et d'examens médicaux. Enfin, FHS (*Framingham Heart Study*) est une cohorte américaine lancée en 1948 qui suit plus de 5 000 participants âgés de 30 à 62 ans. Elle a pour but d'identifier et de mieux comprendre les facteurs de risque des maladies cardiaques. Les données sont collectées régulièrement lors de visites médicales et via des questionnaires.

En dehors des cohortes, la collecte de données de santé pour une étude spécifique inclut aussi la création de registres de patients ou la mise en place d'essais cliniques. Les registres de patients sont des systèmes organisés qui visent à regrouper des informations concernant tous les patients présentant des caractéristiques communes, comme une même maladie par exemple, sur un territoire donné. Les essais cliniques sont des études au cours desquelles des individus sont recrutés dans le but d'évaluer l'efficacité et la tolérance de plusieurs traitements.

Réutilisation des données de santé initialement collectées dans le cadre du soin

Les données de santé collectées dans le cadre du soin sont qualifiées de données de vie réelle car elles sont recueillies lors de la pratique médicale courante, en dehors de tout cadre expérimental. Initialement, ces données sont destinées à la prise en charge du patient, en vue d'élaborer un traitement ou d'établir un diagnostic. Cela correspond à leur usage primaire. Cependant, elles peuvent également être réutilisées à des fins de recherche, ce qui constitue un usage secondaire.

Afin de faciliter leur analyse secondaire, ces données peuvent être regroupées au sein d'entrepôts de données de santé. Ces entrepôts sont des bases de données qui permettent de centraliser et homogénéiser les informations médicales provenant de différentes sources [7]. Ils garantissent un accès rapide et sécurisé aux données pour les chercheurs et les professionnels de santé.

Les entrepôts de données de santé se déclinent en plusieurs catégories. Une première catégorie comprend les entrepôts de données hospitaliers. Ces entrepôts contiennent des données collectées auprès des établissements de santé tels que les hôpitaux, les cliniques ou les centres médicaux. La base de données MIMIC (*Medical Information Mart for Intensive Care*) est un exemple d'entrepôts de données hospitaliers. MIMIC regroupe les données médicales de patients admis dans les unités de soins intensifs de l'hôpital général du Massachusetts [8]. Cette base de données contient notamment des signaux et des mesures enregistrés périodiquement à l'aide de moniteurs ainsi que des données cliniques telles que les diagnostics extraits des dossiers médicaux des patients. Un autre exemple d'entrepôts de données hospitaliers est la base de données NIS (*National Inpatient Sample*). NIS est la plus grande base de données publique regroupant des informations sur les hospitalisations aux État-Unis [9]. Cette base recueille les données de près de sept millions de séjours hospitaliers provenant de plus de 1 000 hôpitaux différents. Ces données concernent les diagnostics, la durée d'hospitalisation ou encore les caractéristiques des hôpitaux. En France, le PMSI (Programme de Médicalisation des Systèmes d'Information) est un exemple d'entrepôts de données hospitaliers. Cet entrepôt contient des données concernant notamment les diagnostics et les actes médicaux.

Une deuxième catégorie d'entrepôts de données de santé est celle regroupant les données médico-administratives [10]. Aux États-Unis, il existe un entrepôt de données médico-administratives contenant les données d'environ 60 millions d'individus bénéficiant du programme *Medicare* [11]. Ce programme s'adresse aux personnes âgées de plus de 65 ans, aux personnes de moins de 65 ans souffrant d'une affection chronique et aux patients en insuffisance rénale terminale. Cet entrepôt de données est accessible pour mener des études épidémiologiques et des études sur les services de santé. En Finlande, il existe également un entrepôt de données regroupant les données médico-administratives de l'ensemble des citoyens ainsi que les données génétiques d'un sous-ensemble de la population [12]. Cet entrepôt propose un accès sécurisé à distance pour les personnes désirant utiliser ces données à des fins de recherche. En France, la base principale de données du Système National des Données de Santé (SNDS) englobe plusieurs composantes : le PMSI (détaillé précédemment), le Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM) qui rassemble les remboursements médicamenteux de la CNAM et le Centre d'épidémiologie sur les causes médicales de décès (CépiDC) qui centralise les causes médicales de décès [13] (voir *Figure 1.2*). Ces données sont mises à disposition des chercheurs dans des espaces sécurisés. Elles couvrent l'ensemble de la population résidant en France, soit près de 70 millions d'individus. Afin de faciliter la recherche, ces données ont été échantillonnées, à la fois pour simplifier les analyses et garantir

l'anonymat des individus, permettant d'étudier les parcours de soins. Un de ces échantillons est l'Echantillon Généraliste des Bénéficiaires (EGB) qui représente 1/97ème des individus inclus dans le SNDS [14]. Il a été conçu pour être représentatif en âge et en genre de la population française. En juin 2022, la CNAM a créé un nouvel échantillon, l'Échantillon du Système National des Données de santé (ESND), destiné à remplacer l'EGB. L'ESND rassemble les parcours de soins en ville (SNIIRAM) et à l'hôpital (PMSI) pour 2% de la population présente dans le SNDS. Il est construit sur la même architecture que le SNDS afin de faciliter la réutilisation des programmes informatiques d'analyse de données entre l'échantillon et la base de données complète.

L'usage secondaire des données de santé présente de nombreux avantages [15]. Comme les données ont déjà été collectées, elles sont facilement et rapidement accessibles, ce qui permet un gain de temps significatif et une réduction des coûts d'accès par rapport aux données collectées pour une étude spécifique. De plus, ces données sont largement disponibles grâce à l'existence d'entrepôts de données qui permettent de les regrouper. Ces avantages d'accès et de disponibilité font de l'usage secondaire des données de santé un outil essentiel pour améliorer la qualité et la gestion des soins, réduire les coûts et faciliter la recherche clinique [16]. Cependant, l'usage secondaire des données de santé est également limité par leur qualité parfois incompatible avec les exigences d'une recherche et la complexité de ces données. En effet, comme énoncé dans la section 1.1.1, les données de santé sont des données longitudinales nombreuses et variées, ce qui rend leur exploitation difficile.

Pendant ma thèse, mon travail s'est porté spécifiquement sur le développement de méthodes d'analyse visant à faciliter l'usage secondaire de données issues de base de données médico-administratives.

1.1.3 Format des variables de santé

Une donnée de santé est définie comme une mesure spécifique d'une variable chez un individu. Les variables peuvent être de différentes formes. Pour introduire ce point, je vais distinguer, dans la suite, la valeur et le label associés aux variables. La valeur représente l'observation ou l'attribut spécifique que peut prendre une variable et le label représente le nom ou la description attribuée à la variable afin de l'identifier. Les variables de données structurées sont des variables dont les valeurs peuvent être quantitatives ou catégorielles, tandis que les variables de données non structurées sont principalement des variables à valeurs textuelles. Ces variables peuvent être définies par des labels organisés sous forme de nomenclature. Cette

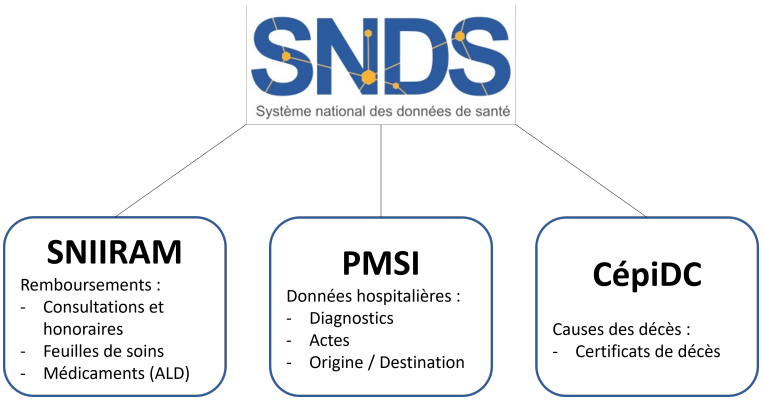


Figure 1.2 – Composantes du Système National des Données de Santé
ALD : Affections de Longue Durée (maladies chroniques pour lesquelles les médecins réalisent une déclaration spécifique)

diversité reflète la complexité des informations que les données de santé contiennent.

Les variables de santé quantitatives

Les variables quantitatives sont des variables dont les valeurs sont exprimées sous forme de nombres. Elles peuvent être de deux types : continues ou discrètes. Les variables quantitatives continues contiennent des mesures qui peuvent prendre n'importe quelle valeur dans un certain intervalle. Elles sont souvent obtenues à partir de mesures précises et peuvent inclure des valeurs décimales. Un exemple de variables quantitatives continues est la pression artérielle. Les variables quantitatives discrètes contiennent des valeurs entières qui comptent ou dénombrent des éléments et ne peuvent prendre que des valeurs spécifiques. Ce sont généralement des nombres entiers. Un exemple de variables quantitatives discrètes est le nombre annuel de visites médicales.

Les variables de santé catégorielles

Les variables catégorielles sont des variables qualitatives regroupées en catégories ou en classes dont les valeurs ne sont pas quantitatives. Elles peuvent être de deux types : binaires ou non binaires. Les variables catégorielles binaires sont des variables qui ne peuvent prendre que deux valeurs distinctes. Ces valeurs peuvent être codées par des libellés ou des nombres, et elles n'ont pas d'ordre particulier. Les variables catégorielles binaires sont souvent utilisées pour représenter des situations où la valeur se résume à deux options exclusives (par exemple, Oui/Non ou présence/absence). Les variables catégorielles non binaires sont des va-

riables qualitatives qui peuvent prendre plus de deux valeurs distinctes. Ces valeurs peuvent être nominales (sans ordre) comme le groupe sanguin d'un patient ou ordinales (avec un ordre) comme le niveau de gravité d'une maladie. Les variables catégorielles non binaires sont couramment utilisées pour représenter des situations où la valeur peut prendre plusieurs options (par exemple, la couleur des yeux).

Les variables de santé textuelles

Les variables textuelles sont des variables de données non structurées dont les valeurs sont exprimées sous forme de texte brut, à la différence des variables quantitatives et catégorielles. En santé, les variables textuelles sont souvent utilisées pour enregistrer des symptômes, des diagnostics ou d'autres observations cliniques liés à la santé du patient. Les variables textuelles peuvent être des comptes rendus médicaux. Ces comptes rendus ont généralement des formats standardisés pour faciliter l'enregistrement systématique et la communication des données médicales. Les variables textuelles peuvent également faire référence à du texte libre. Contrairement aux comptes rendus, le texte libre n'est pas standardisé. Il permet une expression plus libre et naturelle des informations. Ce peut être, par exemple, un commentaire laissé par un médecin dans le dossier médical d'un patient.

Les variables de santé avec des labels nomenclaturals

Pour certaines variables quantitatives, catégorielles ou textuelles, les labels peuvent être des nomenclatures. Ces nomenclatures peuvent être des ontologies, des hiérarchies ou des classifications systématiques.

Il existe de nombreuses nomenclatures utilisées dans les labels des variables de santé. HPO (*Human Phenotype Ontology*) est une nomenclature utilisée pour décrire et classifier les anomalies phénotypiques des pathologies humaines [17]. C'est une ontologie, ce qui signifie qu'elle représente un ensemble de termes et de relations hiérarchiques entre ces termes.

L'ATC (*Anatomical Therapeutic Chemical classification system*) est une classification internationale utilisée pour regrouper les médicaments en fonction de leur composition et de leur usage thérapeutique [18]. L'ATC attribue un code alphanumérique unique à chaque médicament. Chaque élément du code permet de distinguer les critères anatomiques, thérapeutiques, pharmacologiques et chimiques spécifiques du médicament (voir *Figure 1.3*).

La CIM-10 (Classification Internationale des Maladies, 10e révision) est une classification internationale développée par l'Organisation mondiale de la santé (OMS) [19]. Elle permet de regrouper les maladies et les problèmes de santé en leur attribuant un code alphanumérique unique.

Le SNOMED-CT (*Systematized Nomenclature Of Medicine Clinical Terms*) est une hiérarchie médicale multinationale multilingue [20]. Il contient un vaste ensemble de concepts cliniques utilisés pour décrire différents aspects de la santé comme les diagnostics. La structure hiérarchique de cette nomenclature offre une représentation détaillée des liens entre les concepts cliniques, ce qui permet de représenter les connaissances médicales de manière plus précise.

Le MeSH (*Medical Subject Headings*) est une hiérarchie développée par la *National Library of Medicine* (NLM) des États-Unis [21]. Il est utilisé pour indexer et classifier les articles de recherche biomédicale dans les bases de données médicales. Chaque terme MeSH représente un concept médical spécifique, tel qu'une maladie ou un symptôme. Sa structure hiérarchique permet de regrouper les articles communs et faciliter ainsi la recherche d'informations médicales.

Le LOINC (*Logical Observation Identifiers Names and Codes*) est une nomenclature utilisée pour identifier et codifier les observations de laboratoire et les résultats d'examens médicaux [22]. Chaque type d'observations médicales est codé numériquement.

L'UMLS (*Unified Medical Language System*) est un système d'organisation des terminologies et des classifications médicales [23]. Il a été développé par la NLM des États-Unis dans le but de créer un vocabulaire médical unifié et normalisé. L'UMLS regroupe une grande variété de termes et de codes issus de nomenclatures telles que l'ATC, la CIM-10, SNOMED-CT, ou le MeSH. Cela permet donc de mettre en relation différentes terminologies médicales provenant de différents systèmes de nomenclatures médicales.

1.1.4 Analyse des données de santé

L'analyse des données de santé est un outil majeur de la prise en charge des patients. Elle contribue à l'amélioration des soins, à la prévention des maladies et à la prise de décisions médicales. Cette analyse présente de nombreux objectifs et enjeux.

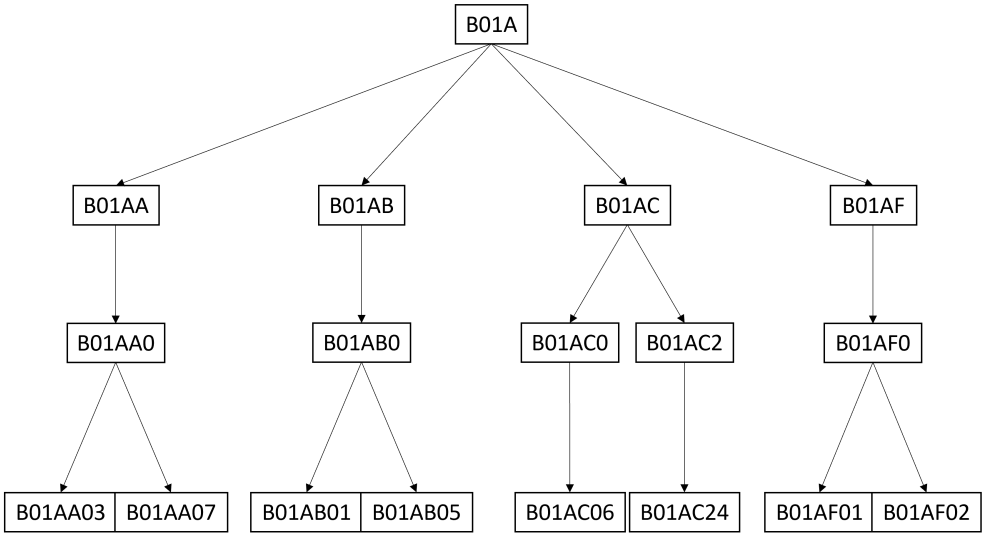


Figure 1.3 – Extrait de la nomenclature *Anatomical Therapeutic Chemical classification system*
Il s'agit de l'extrait de la classification des médicaments appartenant à la classe des agents antithrombotiques (B01)

Les objectifs de l'analyse des données de santé

L'analyse des données de santé comporte de multiples objectifs [24]. Tout d'abord, elle vise à identifier les facteurs de risque, les causes et les déterminants des maladies afin de mieux comprendre leurs hétérogénéités et prévenir les risques pour la santé. Un autre objectif essentiel consiste à établir le diagnostic d'une maladie, permettant d'identifier sa présence chez un patient et de formuler des recommandations médicales appropriées. L'analyse des données de santé permet également de prédire la survenue d'une maladie, offrant ainsi la possibilité d'anticiper son évolution et de mettre en place des interventions médicales précoces. Par ailleurs, elle contribue à étudier l'effet et l'efficacité d'un traitement, permettant d'établir ses bénéfices par rapport à d'autres traitements et d'évaluer les éventuels effets indésirables garantissant la sécurité des patients.

De nombreuses méthodes d'analyse des données de santé existent pour atteindre ces multiples objectifs [25]. Parmi ces méthodes, les tests statistiques de base comme le test de Student, le test de Fisher ou l'ANOVA peuvent être utilisés pour comparer l'efficacité de différents traitements entre deux ou plusieurs groupes de patients. Par exemple, l'utilisation du test de Fisher a permis à Fenaux et al. de démontrer que l'azacitidine, en comparaison avec les traitements conventionnels, permettait d'augmenter la durée de vie et de réduire le risque de progression vers une leucémie chez les patients atteints du syndrome myélodysplasique [26]. L'ANOVA a été employée par Mohanty et al. pour mettre en évidence que la combinaison de deux

traitements spécifiques était prometteuse pour prolonger la durée de vie des patients atteints d'ostéosarcome [27].

Les régressions, de type linéaire ou logistique par exemple, peuvent servir à identifier les facteurs de risques associés à une maladie ou prédire le pronostic d'une maladie. Par exemple, Bui, Edwards et Finlayson ont utilisé la régression logistique pour identifier les facteurs de risques associés à une infection chez les patients atteints d'ulcères de jambe [28]. D'autre part, Cassidy et al. ont employé la régression logistique pour prédire le risque de survenue du cancer du poumon chez un patient sur une période de cinq ans [29].

Les algorithmes d'intelligence artificielle sont également utilisés pour établir et améliorer le diagnostic de maladies, notamment en analysant des images médicales. Par exemple, Lakhani et Sundaram ont eu recours à des réseaux de neurones convolutifs pour détecter la tuberculose à partir de radiographies pulmonaires [30].

Un autre exemple de méthodes d'analyse est le clustering qui est une approche visant à regrouper les patients présentant des caractéristiques médicales similaires telles que les symptômes, les diagnostics ou les réponses à un traitement. Cette méthode est particulièrement pertinente dans l'analyse des données de santé [31]. En effet, le clustering permet de segmenter les patients afin d'obtenir des groupes homogènes de malades qui serviront à mieux comprendre l'hétérogénéité des maladies et évaluer des traitements ou des parcours de soin. Dans la section 1.2, je présenterai plus en détail cette méthode.

Les enjeux de l'analyse des données de santé

L'analyse des données de santé implique de nombreux enjeux dus à la complexité de ces données [32]. En raison de la numérisation presque systématique des données de santé et de leurs nombreuses sources, un volume important d'informations est généré. Cela entraîne des difficultés dans le stockage, le temps de traitement et la gestion des données. En effet, des infrastructures de stockage robustes sont nécessaires pour conserver les données afin de faciliter leur disponibilité lors des analyses. De plus, le temps de traitement peut être augmenté par rapport à un nombre réduit de données. Cela nécessite alors des méthodes d'analyses puissantes pour gérer les données de manière rapide et efficace tout en maintenant la performance et la fiabilité.

En plus d'être nombreuses, les données de santé présentent une grande hétérogénéité en termes de formats. Cette diversité nécessite souvent une étape préalable d'harmonisation des

données car les méthodes d'analyse des données de santé existantes ont tendance à préférer les données uniformes. Chaque format de données possède ses propres caractéristiques, ce qui complique l'analyse simultanée de données différentes. Chacun de ces formats requiert ainsi l'application de méthodes spécifiques, ce qui peut accroître la difficulté et la durée de l'analyse des données.

Un des enjeux majeurs dans l'analyse des données de santé est la présence de données tronquées. Comme énoncé dans la section 1.1.1, les données peuvent présenter une troncature à gauche ou à droite. Cette troncature provient du fait que les données de vie réelle sont collectées à des moments différents de la vie du patient, entraînant une disponibilité de ces données qui n'est pas homogène entre les patients. Ainsi, quand on se focalise sur une période spécifique de la vie (par exemple l'âge) ou d'une maladie, des mesures peuvent être absentes pour certains patients juste après le début ou avant la fin du suivi. Les données tronquées impliquent une perte d'informations. Lors d'une analyse, cela peut induire une réduction de la puissance statistique et conduire ainsi à une perte de précision. Il est donc important de pré-traiter ce genre de données afin de minimiser leurs impacts dans les analyses.

Ainsi, l'analyse des données de santé est confrontée à une série d'enjeux importants en raison du volume considérable de ces données, de leur diversité de formats ainsi que de la présence fréquente de données tronquées. Ces enjeux exigent le développement de méthodes innovantes pour exploiter pleinement le potentiel de ces données. Ces innovations doivent notamment offrir des solutions de stockage des données robustes, une flexibilité permettant de gérer la variété des formats et une gestion adéquate des données tronquées. Comme énoncé précédemment, parmi l'une des méthodes d'analyse des données de santé, le clustering se révèle particulièrement pertinent. Cette méthode permet de mieux appréhender l'hétérogénéité des patients en permettant d'identifier des groupes homogènes. C'est pour cette raison que j'ai choisi d'axer mon travail de thèse sur le développement de méthodes basées sur le clustering de patients.

1.2 Le clustering

1.2.1 Définition

Le clustering est une méthode d'analyse couramment utilisée pour traiter les données de santé [33]. Cette méthode consiste à former des groupes homogènes de patients, appelés clusters, à partir de leurs caractéristiques cliniques. L'objectif est d'avoir des patients similaires

au sein d'un même cluster et des patients différents entre des clusters distincts. Il existe deux catégories principales de clustering : le clustering supervisé et le clustering non supervisé. Le clustering supervisé a pour but de classer les patients dans des groupes prédéfinis et connus à l'avance. A l'inverse, le clustering non supervisé vise à classer les patients dans des groupes qui ne sont pas définis à l'avance. Dans le cadre de ma thèse, mon intérêt s'est porté sur les méthodes de clustering non supervisé.

Le clustering est important dans l'analyse des données de santé car il permet de stratifier les patients en sous-groupes homogènes. Cette stratification est essentielle car il existe une grande hétérogénéité entre les patients concernant leurs caractéristiques cliniques, leurs maladies ou encore leurs réponses aux traitements [31]. Plusieurs études ont utilisé l'approche du clustering pour stratifier des patients. Par exemple, Ahlqvist et al. ont utilisé deux approches de clustering pour identifier cinq sous-groupes de patients diabétiques [34]. De même, Jia et al. ont identifié 22 sous-groupes de patients asthmatiques en utilisant une approche de clustering basée sur leurs comorbidités [35]. Au sein d'une population atteinte de pathologies complexes, Grant et al. ont réussi à identifier 7 sous-groupes de patients présentant des profils cliniques distincts [36]. Ces exemples soulignent l'importance de la stratification pour mieux décrire et comprendre les différents profils de santé présents dans une population de patients.

Cette stratification des patients est particulièrement pertinente dans le contexte de la médecine stratifiée. La médecine stratifiée est une stratégie qui vise à adapter les soins de santé de manière spécifique aux patients d'un même sous-groupe partageant des caractéristiques cliniques communes [37]. L'analyse des spécificités de chaque sous-groupe de patients identifié par le clustering permet donc de mieux cibler les traitements et les interventions, facilitant ainsi une approche plus personnalisée de la prise en charge médicale.

1.2.2 Les mesures de similarités

L'identification de sous-groupes de patients homogènes par le clustering repose sur la capacité à évaluer la similarité entre les patients. En effet, cette notion de similarité permet de déterminer quels patients doivent être regroupés dans un même cluster. Elle est essentielle pour garantir que les patients au sein d'un même sous-groupe soient plus similaires que ceux appartenant à des sous-groupes différents. Afin de calculer cette similarité entre patients, diverses mesures existent. Le choix de la mesure dépend principalement du format des variables utilisées pour calculer la similarité entre les patients. Les mesures de similarité les plus couramment utilisées sont la distance Euclidienne, la similarité Cosinus et la distance de Jaccard

[38].

La distance Euclidienne

La distance euclidienne est la mesure de similarité la plus standard. C'est une distance géométrique qui exprime la plus courte distance entre deux points dans un espace multidimensionnel. Elle se calcule en prenant la racine carrée de la somme des carrés des différences entre les coordonnées des deux points. Soit deux ensembles de données A et B de longueur n appartenant à deux patients distincts. La distance euclidienne entre ces deux patients est définie par l'équation suivante :

$$Euc(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{1.1}$$

où A_i et B_i représente respectivement la ième observation (ou coordonnée) des ensembles de données A et B.
Plus la distance euclidienne est petite, plus les patients sont similaires, et inversement.

La similarité Cosinus

La similarité cosinus est une mesure qui est utilisée pour calculer la similarité entre deux ensembles dans un espace vectoriel [39]. Elle calcule le cosinus de l'angle (θ) entre ces deux ensembles, ce qui permet de quantifier leur orientation. Cette mesure est le plus souvent utilisée dans l'analyse de texte. Lorsqu'elle est appliquée entre deux ensembles de données A et B appartenant à deux patients distincts, la similarité cosinus est définie par l'équation suivante :

$$\cos_\theta(A, B) = \frac{A.B}{||A||.||B||} \tag{1.2}$$

Les valeurs prises par la similarité cosinus sont comprises entre -1 et 1. Une valeur proche de -1 indique que les ensembles sont opposés. Une valeur proche de 0 indique que les ensembles sont différents (orthogonaux). Une valeur proche de 1 indique que les ensembles sont identiques (alignés). Par conséquent, plus l'angle θ entre les deux ensembles de données est petit, plus les patients sont considérés comme similaires.

La distance de Jaccard

La distance de Jaccard permet de calculer la similarité entre deux ensembles [40]. Elle se définit comme le rapport de l'intersection des deux ensembles sur leur union. Cette mesure est particulièrement adaptée pour les variables binaires, mais elle peut également être utilisée pour calculer la similarité entre deux ensembles de variables textuelles en comparant le nombre de mots en commun. Lorsqu'elle est appliquée à deux ensembles de données A et B appartenant à deux patients distincts, la distance de Jaccard est calculée selon l'équation suivante :

$$J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{1.3}$$

où $|A \cap B|$ représente le nombre d'éléments en communs aux ensembles A et B (leur intersection) et $|A \cup B|$ représente le nombre total d'éléments uniques dans les deux ensembles (leur union).
La distance de Jaccard varie entre 0 lorsque les ensembles sont différents et 1 lorsque les ensembles sont similaires.

Autres mesures

En plus de la distance Euclidienne, de la similarité Cosinus et de la distance de Jaccard, de nombreuses autres mesures existent pour calculer la similarité entre patients. Pour les variables quantitatives, on peut utiliser les distances de Manhattan, Minkowski, Chebyshev et Mahalanobis [41, 42]. La distance de Manhattan est une mesure qui calcule la somme des différences absolues entre les coordonnées (c'est-à-dire les valeurs) de deux ensembles de données. La distance de Minkowski est la généralisation des distances Euclidienne et de Manhattan. La distance de Chebyshev calcule la plus grande différence absolue entre les coordonnées de deux ensembles de données. La distance de Mahalanobis calcule la similarité entre deux ensembles de données en tenant compte de la covariance entre les variables. Pour les variables binaires, on peut utiliser la distance de Hamming [43]. Cette distance calcule le nombre d'éléments qui diffèrent entre deux ensembles de données.

L'ensemble des mesures présentées ici permet de calculer une similarité entre chaque paire de patients au sein d'une population d'étude. Les similarités obtenues sont stockées dans une matrice de similarité $M = [m_{p_1,p_2}]^N$ où N représente le nombre total de patients présents dans la population d'étude et m_{p_1,p_2} représente la similarité calculée entre les patients p_1 et p_2 en utilisant l'une des mesures de similarité. C'est à partir de cette matrice de similarité que

les méthodes de clustering vont former les clusters de patients. La diversité des mesures de similarité offre donc une grande flexibilité aux approches de clustering. Cependant, le choix de ces mesures doit se faire avec discernement car il exerce un impact significatif sur les résultats du clustering [44].

1.2.3 Les différentes méthodes de clustering

Il existe plusieurs catégories de méthodes de clustering [45]. Les cinq catégories de clustering les plus couramment utilisées sont : le clustering fondé sur les centroïdes, le clustering hiérarchique, le clustering de densité, le clustering flou et le clustering de réseaux (voir *Table 1.1*). Dans cette section, je détaillerai ces méthodes dans le contexte de l'analyse des données de santé non longitudinales. Toutefois, il existe des catégories spécifiques de clustering pour les données longitudinales que j'explorerai dans le chapitre 2.

Catégories	**Algorithmes**
Clustering fondé sur les centroïdes	K-means K-medoids
Clustering hiérarchique	BIRCH CURE
Clustering de densité	DBSCAN OPTICS
Clustering flou	Fuzzy C-means Possibilistic C-means
Clustering de réseaux	Louvain MCL

Table 1.1 – Les différentes catégories de clustering
MCL : *Markov Cluster Algorithm*. Pour chaque catégorie de méthode de clustering est présenté quelques exemples d'algorithmes les plus couramment utilisés.

Clustering fondé sur les centroïdes

Le clustering fondé sur les centroïdes consiste à diviser un ensemble de données en un nombre prédéfini et non superposé de clusters en se basant sur les centroïdes qui représentent les centres des clusters. Chaque cluster doit contenir au moins un point de données et chaque point de données doit appartenir à un cluster unique. Parmi les méthodes de clustering fondé sur les centroïdes les plus utilisées, on trouve K-means [46] et K-medoids [47]. A partir de la

matrice de similarité, ces méthodes cherchent à regrouper les données en K clusters en minimisant la distance entre les points de données et les centroïdes correspondants. Le processus commence par le choix aléatoire des centroïdes initiaux pour chaque cluster. Ensuite, chaque point de données est attribué au centroïde le plus proche en termes de distance. De nouveaux centroïdes sont alors recalculés en fonction des points attribués. Ces étapes sont répétées jusqu'à ce que les centroïdes convergent vers une position stable, indiquant la formation des clusters finaux. La différence principale entre K-means et K-medoids réside dans la manière dont les centroïdes des clusters sont établis. Pour l'algorithme du K-means, les centroïdes sont calculés comme étant la moyenne des points du cluster, ce qui implique qu'ils peuvent ne pas correspondre à des points réels des données. En revanche, pour l'algorithme du K-medoids, les centroïdes correspondent à des points réels des données, ce qui le rend plus robuste aux données aberrantes par rapport au K-means.

Par exemple, Fuente-Tomas et al. ont appliqué la méthode du K-means sur des variables démographiques et cliniques quantitatives pour identifier cinq clusters de patients atteints de troubles bipolaires [48]. Les clusters obtenus permettaient de distinguer des groupes de patients de gravité différente. Une autre étude, menée par Abbas et al., a comparé les méthodes du K-means et du K-medoids pour classer des femmes en fonction de leurs antécédents d'accouchement [49]. Ces deux méthodes ont été appliquées sur des données issues de questionnaires concernant la grossesse, les antécédents de maternité, l'état de santé, la gestation et la vie sociale de ces femmes. Cette comparaison a révélé que la méthode du K-medoids a généré de meilleurs clusters que K-means.

Clustering hiérarchique

Le clustering hiérarchique consiste à créer une structure arborescente de clusters à partir de la matrice de similarité. Au sein de cette arborescence, chaque cluster est lié à au moins deux clusters successeurs, formant ainsi une hiérarchie. Il existe deux grands types de clustering hiérarchique en fonction de la manière dont l'arborescence est construite : soit de manière ascendante, soit de manière descendante.

L'approche ascendante, également appelée clustering agglomératif, part du principe que chaque point de données forme un cluster individuel. Les clusters sont ensuite fusionnés en fonction de leur similarité jusqu'à ce que tous les points soient rassemblés dans un seul grand cluster. Dans l'approche descendante, également appelée clustering divisif, le procédé est inversé. Au départ, tous les points de données sont contenus dans un seul grand cluster. Ensuite,

ce cluster est progressivement divisé jusqu'à obtenir autant de clusters que de points.

L'arborescence obtenue par le clustering hiérarchique est généralement représentée sous forme de dendrogramme. Ce graphique est composé de branches qui illustrent les fusions et divisions des clusters à différentes échelles de granularité. Selon la profondeur à laquelle une branche est coupée dans le dendrogramme, un nombre différent de clusters est obtenu.

Deux exemples d'algorithmes de clustering hiérarchiques sont BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) [50] et CURE (*Clustering Using Representatives*) [51]. BIRCH commence par résumer les données initiales en les stockant dans un arbre nommé *Clustering Feature* (CF). L'arbre CF est réduit en fusionnant les parties de l'arbre les plus proches et en éliminant les points les plus éloignés selon la matrice de similarité. Cela permet d'obtenir une représentation plus compacte des données, réduisant ainsi l'espace mémoire requis. Une approche ascendante est ensuite appliquée en utilisant les informations stockées dans l'arbre CF réduit et les clusters sont identifiés en coupant le dendrogramme construit à partir de l'arborescence obtenue. De manière similaire, CURE réduit l'espace mémoire requis en sélectionnant aléatoirement des points représentatifs parmi les données initiales. A partir de ces points, l'arborescence est construite en utilisant une approche ascendante. Ces deux algorithmes sont particulièrement adaptés pour clusteriser des ensembles de données volumineux.

Par exemple, Mamatha Bai, Nalini et Majumdar ont appliqué l'algorithme BIRCH à partir de huit variables quantitatives relatives au diabète pour identifier trois clusters de patients diabétiques [52].

Clustering de densité

Le clustering de densité consiste à identifier les clusters en se basant sur la densité des points de données. Les clusters sont définis comme étant des régions contiguës de haute densité, séparées des autres clusters par des régions moins denses. Le processus de formation des clusters commence par le calcul de la densité locale de chaque point de donnée à partir de la matrice de similarité. La densité locale d'un point correspond au nombre de points voisins se trouvant dans un rayon de distance maximale noté ϵ. Ensuite, si à partir d'un point de données, une région contenant au moins un nombre de points $MinPts$ dans le rayon ϵ est identifiée, alors cette région est considérée comme dense et forme ainsi un cluster.

La méthode de clustering de densité la plus couramment utilisée est DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) [53]. Les deux paramètres ϵ et $MinPts$

qui définissent la taille des clusters doivent être spécifiés par l'utilisateur. Si la taille d'une région satisfait ces paramètres, DBSCAN la désigne comme formant un cluster. En revanche, les points ne satisfaisant pas ces paramètres (c'est-à-dire les points isolés) sont considérés comme du bruit. Cette caractéristique confère à l'algorithme une robustesse face aux données aberrantes.

Une autre méthode largement utilisée est OPTICS (*Ordering Points To Identify the Clustering Structure*) [54]. Contrairement à DBSCAN, OPTICS ne nécessite pas la spécification du paramètre de distance ϵ. Au lieu de cela, l'algorithme propose des techniques pour estimer automatiquement ce paramètre. OPTICS génère un ordonnancement des points de données en fonction de leur densité locale. Dans cet ordonnancement, les points denses seront les plus proches les uns des autres. L'identification des clusters repose sur la détection des régions de l'ordonnancement qui satisfont le paramètre MinPts.

A titre d'exemple, Shukla et al. ont utilisé DBSCAN pour identifier neuf clusters composés de patientes atteintes d'un cancer du sein et présentant des caractéristiques communes [55]. Les données analysées comprenaient des informations sur les caractéristiques de la tumeur, le stade de la maladie, les caractéristiques sociodémographiques des patientes ainsi que sur la mortalité. Yan et al. ont appliqué OPTICS pour identifier dix clusters au sein d'une population de patients consommant des soins coûteux en utilisant des variables cliniques et administratives. Cette étude a également révélée que OPTICS a généré des résultats supérieurs à ceux obtenus avec le clustering hiérarchique ascendant et l'algorithme du K-medoids.

Clustering flou

Le clustering flou (ou *fuzzy clustering*) permet à chaque point de données d'appartenir à plusieurs clusters plutôt que d'être assigné à un cluster unique. Les points de données se voient attribuer une probabilité d'appartenance à chaque cluster, reflétant l'incertitude dans leur regroupement. Cette approche est particulièrement utile quand les patients ne sont pas nettement séparables et qu'il existe des chevauchements entre les clusters. Parmi les méthodes de clustering flou, fuzzy C-means est la plus couramment utilisée [57]. Cet algorithme commence par attribuer aléatoirement des probabilités d'appartenance à chaque point de données, ainsi que des centroïdes pour les clusters. Les probabilités d'appartenance sont ensuite mises à jour en fonction de la similarité entre les points et les centroïdes, et de nouveau centroïdes sont calculés en les pondérant par ces probabilités. Ce processus itératif se poursuit jusqu'à ce que les probabilités d'appartenance convergent vers une position stable.

Un exemple d'application de fuzzy C-means est celui de Sanakal et Jayakumari qui ont utilisé cet algorithme pour identifier des clusters de patients en se basant sur les résultats de leur test de diagnostic du diabète [58]. Cette étude avait pour objectif d'améliorer la capacité de pronostic de la maladie.

Une extension de l'algorithme Fuzzy C-means, appelée possibilistic C-means, a également été développée [59]. Cette extension permet d'éliminer la contrainte probabiliste imposée par fuzzy C-means selon laquelle la somme des probabilités d'appartenance doit être égale à un. Grâce à cette modification, les points de données relativement éloignés de tous les clusters, correspondant à des données aberrantes, peuvent avoir une probabilité d'appartenance négligeable à tous les clusters.

Clustering de réseaux

Le clustering de réseaux consiste à créer des réseaux à partir des données et à identifier les clusters au sein de ces réseaux. Contrairement aux méthodes de clustering flou, de densité, hiérarchique et fondé sur les centroïdes, qui s'appliquent directement sur les variables brutes, le clustering de réseaux s'applique sur les réseaux dérivés de la transformation des variables brutes. La première étape de cette approche implique la construction d'un réseau. Un réseau est un graphe composé de noeuds et de liens qui connectent les noeuds entre eux. Les réseaux peuvent être dirigés, pondérés ou encore cycliques. Les réseaux dirigés possèdent des liens avec une direction spécifique, indiquant une relation unidirectionnelle entre deux noeuds. Par exemple, un lien partant d'un noeud A vers un noeud B n'implique pas nécessairement un retour de B vers A, contrairement aux réseaux non dirigés. Dans les réseaux pondérés, chaque lien est associé à un poids, c'est-à-dire une valeur permettant de quantifier l'intensité de la relation entre les noeuds. En revanche, ces poids sont absents des liens dans les réseaux non pondérés. Dans les réseaux cycliques, il existe au moins un chemin direct permettant de revenir au noeud de départ, tandis que les réseaux acycliques ne le permettent pas. Les noeuds peuvent représenter diverses entités telles que des patients, des maladies, ou encore des protéines. Dans notre cas, nous nous intéressons aux réseaux pondérés non orientés dans lesquels les noeuds sont les patients, comme illustré dans la *Figure 1.4*. Les réseaux de patients présentent l'avantage de mieux préserver la vie privée car ce sont les relations entre les patients qui sont considérées plutôt que les données absolues [60]. Ces réseaux peuvent donc être partagés plus facilement.

Dans les réseaux, les liens représentent les interactions entre ces noeuds et sont établis à

l'aide de mesures de similarité. La construction du réseau s'effectue en établissant un seuil dans la matrice de similarité. Ce seuil est généralement choisi de manière à connecter uniquement les paires de noeuds présentant une forte similarité. Par conséquent, dans le réseau résultant, toutes les paires de noeuds dont la similarité dépasse le seuil choisi seront connectées par un lien dont la pondération correspond à cette similarité. Il est important de noter que d'autres méthodes peuvent être utilisées pour construire un réseau en dehors de la création d'une matrice de similarité. Par exemple, les interactions entre les noeuds peuvent être identifiées au moyen de corrélations, de régressions, d'analyses bayésiennes ou de la théorie de l'information [61].

Une fois le réseau construit, la deuxième étape consiste à identifier les clusters dans ce réseau. Dans un réseau, un cluster est défini comme un groupe de noeuds étroitement liés entre eux. Il existe un grand nombre d'algorithmes issu de la théorie des graphes permettant de réaliser du clustering à partir des réseaux. Parmi ces algorithmes, on retrouve le clustering spectral, les méthodes fondées sur la modularité ou encore les méthodes fondées sur les marches aléatoires [62]. Cette diversité d'algorithmes constitue un avantage de la construction de réseaux.

Le clustering spectral utilise les propriétés spectrales de la matrice d'adjacence d'un réseau pour identifier les clusters [63]. La matrice d'adjacence est une matrice binaire où la valeur de 1 indique la présence d'un lien entre deux noeuds du réseau associé, tandis que 0 indique l'absence de lien. Dans le clustering spectral, l'identification des clusters commence par le calcul de la matrice Laplacienne à partir de cette matrice d'adjacence. La matrice Laplacienne L est définie par $L = D - A$, où D est la matrice des degrés dont la diagonale indique le nombre de liens qui connecte chaque noeud, et A est la matrice d'adjacence. Ensuite, les valeurs et vecteurs propres sont extraits à partir de la matrice Laplacienne pour réduire la dimension des données. Enfin, un algorithme de clustering, tel que K-means, est appliqué dans cet espace de dimension réduit afin d'identifier les clusters.

Les méthodes fondées sur la modularité consistent à identifier les clusters en optimisant la mesure de modularité. La modularité permet d'évaluer la structure d'un réseau en quantifiant la force de division d'un réseau en clusters [64]. Elle compare le nombre de liens à l'intérieur d'un cluster avec le nombre de liens qui seraient attendus au hasard. Un exemple courant d'algorithme est Louvain [65]. Cet algorithme identifie les clusters en déplaçant les noeuds d'une communauté à une autre de manière itérative jusqu'à ce que la modularité maximale soit atteinte.

Les méthodes fondées sur les marches aléatoires permettent d'identifier les clusters en exploitant les propriétés des marches aléatoires. Les marches aléatoires sont des processus stochastiques permettant de simuler des flux au sein du réseau. Ces flux parcourent les noeuds et ceux qui sont fréquemment visités ensemble sont plus susceptibles de faire partie du même cluster. MCL (*Markov Cluster Algorithm*) est un exemple d'algorithme exploitant les marches aléatoires [66]. Cet algorithme détecte les clusters en alternant entre deux paramètres : l'inflation et l'expansion. L'inflation favorise les flux au sein des clusters, permettant d'identifier tous les noeuds appartenant au même cluster, tandis que l'expansion favorise les flux entre les clusters, permettant ainsi de déterminer leur séparation.

Plusieurs études ont utilisé les réseaux pour effectuer du clustering de patients. Par exemple, Wang et al. ont construit des réseaux de patients atteints de divers cancers à partir de plusieurs types de variables omiques qui ont permis d'identifier des clusters de pronostic différent [67]. Li et al. ont construit un réseau de patients à partir de leurs similarités cliniques pour identifier différents sous-types de diabète de type 2 [68]. Sánchez-Valle et al. ont construit un réseau de patients à partir de leurs variables transcriptomiques dans le but de stratifier les patients en fonction de leurs comorbidités [69].

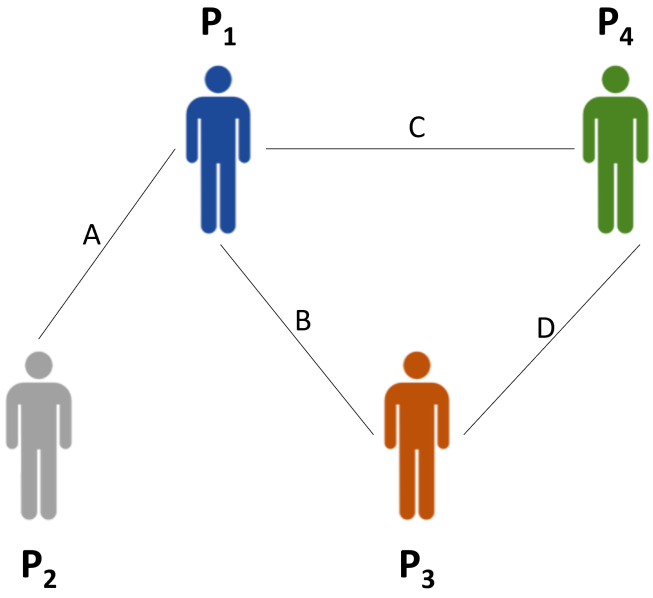


Figure 1.4 – Exemple d'un réseau de patients
Les noeuds représentent les patients et les liens représentent les interactions entre ces noeuds. Ce réseau est pondéré, non orienté et cyclique. Chaque lien est associé à un poids (A, B, C ou D). Les patients P_1, P_3 et P_4 forment un cycle.

1.2.4 Les challenges du clustering

Les méthodes de clustering sont diverses et variées, mais elles sont confrontées à plusieurs challenges qui doivent être pris en compte à la fois lors de leur développement et lors de leur exécution [70]. Nous pouvons distinguer deux catégories de challenges : ceux liés aux algorithmes et ceux liés aux données de santé.

Les challenges liés aux algorithmes

Un des challenges majeurs associés aux algorithmes de clustering réside dans le choix des paramètres adéquats. De nombreux algorithmes nécessitent que l'utilisateur fournisse certains paramètres. Par exemple, le choix du nombre de clusters constitue un paramètre essentiel. Il doit souvent être prédéfini dans les méthodes de clustering fondé sur les centroïdes, de clustering hiérarchique ou de clustering flou. Cependant, la sélection du nombre de clusters exige une connaissance préalable de la structure des données, ce qui n'est pas toujours évident. Bien qu'il existe divers critères pour tenter de déterminer le nombre optimal de clusters [71], il arrive fréquemment que ces critères aboutissent à des paramètres différents (voir section 1.2.5 dans la sous-partie des mesures de qualité internes). Un autre exemple de paramètre à spécifier est la mesure de similarité. En fonction de la mesure choisie, la similarité entre les données varie ce qui influence directement les résultats du clustering [44]. Ainsi, les performances des algorithmes sont étroitement liées aux paramètres choisis.

Un autre challenge réside dans les problèmes de convergence. La convergence fait référence au moment où l'algorithme atteint une solution de clustering stable et cohérente. Cependant, il arrive parfois que les algorithmes aient des difficultés à converger. Un exemple en est lorsque l'algorithme converge vers un minimum local plutôt que vers la solution globale optimale. Cette situation peut conduire à la formation de clusters incorrects et est particulièrement fréquente dans les méthodes de clustering fondé sur les centroïdes et de clustering flou. Les problèmes de convergence peuvent être étroitement liés aux conditions initiales des algorithmes. Pour atténuer ce problème, différentes initialisations peuvent être testées, mais cela nécessite d'exécuter l'algorithme plusieurs fois, ce qui augmente le coût computationnel.

L'interprétation des résultats de clustering constitue également un challenge important. En effet, les clusters obtenus ne sont généralement pas directement interprétables car les caractéristiques spécifiques de ces clusters ne sont pas explicitement détaillées par les algorithmes. Il est donc nécessaire de recourir à des outils de validation et de visualisation des clusters

afin de faciliter leur interprétation. Deux catégories de mesures de qualité du clustering sont régulièrement utilisées pour évaluer la qualité des clusters. Ces deux catégories varient en fonction des données que les mesures exploitent : les mesures de qualité externes exploitent des données non impliquées dans la formation des clusters et les mesures de qualité internes exploitent les données impliquées dans la formation des clusters [72]. Une description détaillée de ces mesures est donnée dans la section 1.2.5. Concernant les outils de visualisation des clusters, les représentations graphiques telles que les nuages de points, les dendrogrammes ou les *heatmaps* permettent d'offrir une vue plus claire et intuitive des clusters identifiés. En fonction de l'objectif de l'analyse, le choix des outils de validation et de visualisation est crucial pour parvenir à donner un sens clinique cohérent aux clusters identifiés.

Les challenges liés aux données de santé

Comme évoqué précédemment, les données de santé sont complexes. Cette complexité peut largement influer sur la performance des algorithmes de clustering, entraînant plusieurs challenges.

Le premier challenge est dû au volume important des données de santé. Lorsque le nombre de variables et de patients est élevé, cela peut entraîner un ralentissement significatif de l'exécution des algorithmes de clustering. En particulier, le calcul de la similarité entre toutes les paires de patients peut être chronophage. Par exemple, dans le clustering hiérarchique et le clustering des réseaux, la création de l'arbre ou du réseau peut devenir une opération longue et complexe. Par ailleurs, la gestion de ce volume massif de données peut exiger des ressources informatiques plus importantes, notamment en termes de capacité de stockage. En conséquence, les algorithmes doivent être capables de gérer avec efficacité ces contraintes de mémoire pour assurer leur bon fonctionnement.

L'hétérogénéité des variables de santé constitue également un challenge important dans les méthodes de clustering. De nombreux algorithmes sont initialement conçus pour traiter des variables quantitatives. Cependant, les variables de santé peuvent présenter une variété de formats, tels que les formats quantitatifs, catégoriels ou textuels. De plus, les labels de ces variables sont très souvent organisés sous forme de nomenclature. Ainsi, certains labels ont des relations plus étroites que d'autres au sein d'une nomenclature donnée, qu'il serait intéressant de prendre en compte. Cependant, les méthodes de clustering se fondent uniquement sur les valeurs des variables, sans considérer les relations entre labels. Dans ce contexte, nous avons élaboré une approche permettant de considérer à la fois les valeurs et les relations entre les

labels des variables que je présenterai dans le chapitre 3.

Un dernier challenge important lié aux données de santé dans le contexte du clustering est la présence de données tronquées. L'option de simplement ignorer ou de supprimer les variables lorsque les patients présentent des données tronquées entraîne une perte d'information car l'algorithme n'utilise alors pas la totalité des patients disponibles. Cela introduit des biais affectant ainsi la formation des clusters. Par conséquent, il est essentiel de traiter les données tronquées pour réaliser le clustering. L'approche la plus couramment utilisée pour traiter les données tronquées est l'imputation. Elle consiste à remplacer les données tronquées par des données estimées, permettant ainsi de considérer la totalité des patients dans le clustering. Plusieurs techniques d'imputation existent [73]. Par exemple, les données tronquées peuvent être remplacées par une valeur unique comme la moyenne, la médiane ou le mode de la distribution. La technique du KNN (*K Nearest Neighbors*) peut être employée pour remplacer les données tronquées par les données non tronquées les plus proches en termes de similarité. Dans certains cas, les données tronquées peuvent être prédites au moyen de modèles de régression. En présence de variables longitudinales, les données tronquées peuvent être remplacées par les données non tronquées qui les précèdent ou les suivent dans le temps. Cependant, malgré la diversité des techniques d'imputation, l'estimation des données tronquées reste un processus incertain. Par conséquent, il est essentiel de développer des algorithmes de clustering capables de gérer les données tronquées sans avoir recours à leur suppression ou à leur imputation. C'est l'objectif que nous nous sommes fixé en développant des approches réseaux que je présenterai dans le chapitre 2 de cette thèse.

1.2.5 Les outils d'évaluation du clustering

Une fois que les clusters sont formés, évaluer leur qualité est important pour s'assurer de leur cohérence, de leur pertinence et de la performance globale du clustering. Cette évaluation a pour but de vérifier si les clusters contiennent bien des patients similaires tout en étant distincts les uns des autres. Pour ce faire, il existe diverses mesures de qualité du clustering classées en deux grandes catégories : les mesures de qualité externes et les mesures de qualités internes [72]. Ces mesures de qualité diffèrent par les variables exploitées pour l'évaluation. Les mesures de qualité externes exploitent des variables externes. Ces variables sont dites externes car elles n'ont pas été utilisées pour la formation des clusters. En revanche, les mesures de qualité internes exploitent les variables qui ont été directement impliquées dans la formation des clusters. Par exemple, dans le cadre où le clustering de patients a été réalisé à partir de

leurs remboursements médicamenteux, les mesures de qualité externes pourraient exploiter les variables concernant les maladies de ces patients alors que les mesures de qualité internes exploiteraient directement les remboursements.

Les mesures de qualité externes

De nombreuses mesures de qualité externes existent pour évaluer la performance du clustering en utilisant des variables externes. L'entropie, par exemple, est utilisée pour mesurer la pureté des clusters [74]. Pour un cluster donné c, son entropie $H(c)$ est définie comme suit :

$$H(c) = -\sum_{l \in L} P(c_l) \log_2 P(c_l), \tag{1.4}$$

où l est une donnée externe appartenant à l'ensemble des données externes L et $P(c_l)$ est la probabilité qu'un patient classé dans le cluster c possède la donnée externe $l \in L$. Cette probabilité peut être calculée empiriquement à partir de l'équation suivante :

$$P(c_l) = \frac{n_{c_l}}{|c|} \tag{1.5}$$

où n_{c_l} est le nombre de patients du cluster c possédant la donnée externe l et $|c|$ est le nombre de patients contenus dans le cluster c.
L'entropie global $H(C)$ du clustering s'obtient ainsi :

$$H(C) = \sum_{c \in C} H(c) \frac{|c|}{N}, \tag{1.6}$$

où C est l'ensemble des clusters identifiés, $H(c)$ est l'entropie du cluster c, $|c|$ est le nombre de patients contenus dans le cluster c et N est le nombre total de patients.
Si un cluster contient des patients possédant les mêmes données externes, son entropie est de 0 indiquant que ce cluster est pur. A l'inverse, plus un cluster contient des patients possédant des données externes différentes, plus son entropie sera élevée et moins il sera pur.

L'indice de Rand est une mesure de qualité externe qui permet d'évaluer la consistance des clusters obtenus par rapport aux données externes [75]. Il se base sur un tableau de contingence qui répertorie les paires de patients regroupées dans les mêmes clusters et ayant soit des données externes identiques, soit différentes. Le tableau de contingence est défini comme suit :

	Données externes identiques	Données externes différentes
Cluster identique	A'	B'
Clusters différents	C'	D'

Table 1.2 – Tableau de contingence utilisé pour le calcul de l'indice de Rand

Dans ce tableau, A' représente le nombre de paires de patients regroupés dans un même cluster et possédant les mêmes données externes, B' représente le nombre de paires de patients groupés dans un même cluster mais possédant des données externes différentes, C' représente le nombre de paires de patients situés dans des clusters différents mais possédant les mêmes données externes et D' représente le nombre de patients situés dans des clusters différents et possédant des données externes différents.
A partir de ce tableau de contingence, l'indice de Rand (RI) est calculé à l'aide de la formule suivante :

$$RI = \frac{A' + D'}{A' + B' + C' + D'}. \tag{1.7}$$

L'indice de Rand varie entre 0 et 1, où 0 indique que les clusters regroupent des patients possédant des données externes différentes et 1 indique que les clusters regroupent des patients possédant les mêmes données externes.
L'inconvénient de l'indice de Rand est qu'il ne tient pas compte que des similarités entre les données pourraient survenir par hasard. L'indice de Rand ajusté a été conçu pour remédier à cette limitation. Il permet en effet de prendre en considération la possibilité que des clusters puissent être trouvés par hasard.

L'information mutuelle est une mesure de qualité externe qui permet d'évaluer la dépendance entre les clusters obtenus et les données externes des patients [76]. Elle est définie comme suit :

$$MI(C, L) = \sum_{c \in C} \sum_{l \in L} P(c, l) \log \frac{P(c, l)}{P(c)P(l)}, \tag{1.8}$$

où $P(c, l)$ est la probabilité conjointe du cluster c et de la donnée externe l et $P(c)$ et $P(l)$ sont les probabilités marginales du cluster c et de la donnée externe l respectivement.
Une information mutuelle élevée indique que les clusters dépendent fortement des données externes, ce qui signifie qu'ils regroupent des patients possédant les mêmes données externes. En revanche, une information mutuelle faible indique que les clusters sont indépendants des données externes, ce qui signifie qu'ils regroupent des patients possédants des données externes différentes.

A titre d'exemple, Akbarpour, Akbari et Motameni ont développé trois mesures de similarité qu'ils ont utilisées avec l'algorithme K-means en testant plusieurs valeurs pour le nombre de clusters et pour l'initialisation [77]. Ils ont analysé dix jeux de données différents pour lesquels les clusters réels étaient connus et dont trois concernaient le clustering de patients. Afin de comparer les performances de leurs mesures de similarité, ils ont utilisé plusieurs mesures de qualité externes, notamment l'indice de Rand et l'information mutuelle. Les données externes utilisées étaient les labels des clusters réels. L'indice de Rand a démontré une sensibilité accrue au nombre de clusters. En effet, il tendait à converger vers 1 lorsque le nombre de clusters augmentait. Pour tenir compte de cette sensibilité, l'indice de Rand ajusté est utilisé car il permet de tenir compte de la possibilité que des clusters puissent être trouvés par hasard. En revanche, l'information mutuelle s'est révélée moins sensible au nombre de clusters.

Outre la sensibilité au nombre de clusters, l'indice de Rand et l'information mutuelle sont également sensibles à la taille des clusters [78]. Si les clusters identifiés ont des tailles très variables, ces mesures de qualité vont accorder plus d'importance à la concordance entre les clusters les plus grands qu'entre les clusters les plus petits. Cette limitation s'applique également à l'entropie [79]. De plus, en ce qui concerne l'information mutuelle, comme il faut calculer des probabilités conjointes, cela peut entraîner un coût computationnel plus élevé.

Il est important de noter que toutes ces mesures de qualité du clustering exploitent les données qui n'ont pas été impliquées dans la formation des clusters. Cependant, ces données externes ne sont pas toujours disponibles. Dans de tels cas, les mesures de qualité internes deviennent plus pertinentes pour l'évaluation du clustering car elles se basent sur les données qui ont été directement impliquées dans la formation des clusters et qui sont donc toujours disponibles.

Les mesures de qualité internes

Il existe de nombreuses mesures de qualité internes. Le coefficient silhouette est la mesure de qualité interne la plus couramment utilisée. Il permet d'évaluer la qualité de séparation des clusters [80].
Le coefficient silhouette $s(p)$ d'un patient p est défini de la manière suivante :

$$s(p) = \frac{b(p) - a(p)}{\max(a(p), b(p))}, \tag{1.9}$$

où $a(p)$ est la distance moyenne entre le patient p et les autres patients de son propre cluster et $b(p)$ est la distance moyenne entre le patient p et les patients du cluster voisin le plus proche.

Le coefficient silhouette global S du clustering est calculé ainsi :

$$S = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|c|} \sum_{p \in c} s(p), \tag{1.10}$$

où $|C|$ est le nombre total de clusters identifiés, $|c|$ le nombre de patients contenus dans le cluster c et $s(p)$ le coefficient silhouette du patient p appartenant au cluster c.
Le coefficient silhouette varie entre -1 et 1. Les valeurs proches de 1 indiquent que les clusters sont bien séparés. Les valeurs proches de 0 indiquent des chevauchements entre les clusters. Les valeurs négatives indiquent que les clusters sont moins distincts que ce qui serait attendu par hasard. Dans une étude menée par Ogbuabor et Ugwoke, deux méthodes de clustering, à savoir K-means et DBSCAN, ont été comparées. Elles ont été utilisées dans le but d'identifier des clusters de patients en se basant sur des données d'activité physique enregistrées au moyen de capteurs [81]. L'évaluation du clustering à l'aide du coefficient silhouette, a démontré que K-means obtenait de meilleures performances que DBSCAN.

L'indice de Calinski-Harabasz est une autre mesure de qualité interne qui évalue la séparation des clusters en analysant les variances inter-cluster et intra-cluster [82]. Il est défini par l'équation suivante :

$$CH = \frac{tr(X)}{tr(W)} . \frac{N - |C|}{|C| - 1}, \tag{1.11}$$

où $tr(X)$ est la trace de la matrice de covariance inter-cluster, $tr(W)$ est la trace de la matrice de covariance intra-cluster, N est le nombre total de patients et $|C|$ est le nombre total de clusters identifiés.
Plus l'indice de Calinski-Harabasz est élevé, meilleure est la séparation entre les clusters. Par exemple, Ma et al. ont utilisé l'algorithme du K-means pour stratifier des patients atteints de la maladie de Parkinson à partir de leurs données cliniques et démographiques [83]. Ils ont calculé l'indice de Calinski-Harabasz pour déterminer le nombre optimal de clusters dans l'algorithme, conduisant à un choix de quatre clusters. Les quatre clusters ainsi identifiés ont permis de caractériser différents sous-types cohérents de la maladie de Parkinson.

L'indice de Davies-Bouldin est une mesure de qualité interne qui permet d'évaluer la séparation des clusters en calculant la distance de chaque cluster à son cluster le plus proche [84]. Pour chaque paire de clusters identifié, cet indice commence par calculer le ratio suivant :

$$R_{cc'} = \frac{Y_c + Y_{c'}}{M_{cc'}}, \tag{1.12}$$

où Y_c est la distance moyenne entre chaque patient du cluster c et du centroïde de ce cluster et $M_{cc'}$ est la distance entre les centroïdes des clusters c et c'.

L'indice de Davies-Bouldin est ensuite calculé à l'aide de l'équation suivante :

$$DB = \frac{1}{|C|} \sum_{c \in C} \max_{c \neq c'} R_{cc'}, \tag{1.13}$$

où $|C|$ est le nombre total de clusters identifiés.
Par exemple, l'étude menée par Urban et al. a utilisé l'algorithme du K-medoids pour stratifier des patients atteints d'insuffisance cardiaque aiguë en se basant sur leurs données cliniques et biochimiques [85]. L'indice de Davies-Bouldin a été calculé pour déterminer le nombre optimal de clusters ainsi que la mesure de similarité appropriée pour ces données. Les résultats ont montré que la meilleure valeur de cet indice a été obtenue avec six clusters, identifiés en utilisant le coefficient de corrélation.

Il existe plusieurs inconvénients associés au coefficient silhouette, à l'indice de Calinski-Harabasz et à l'indice de Davies-Bouldin. Ces inconvénients ont été mis en évidence dans une étude menée par Van Craenendonck et Blockeel. Au cours de cette étude, les auteurs ont appliqué six algorithmes de clustering à 27 jeux de données différents et les trois mesures de qualité internes mentionnées précédemment ont été calculées pour comparer les performances de ces algorithmes [86]. Ces mesures se sont révélées sensibles à la taille des clusters. Par exemple, le coefficient silhouette a obtenu de meilleur score quand les clusters étaient de taille variable. De plus, elles ont montré une sensibilité au bruit, favorisant les algorithmes qui sont plus robustes au bruit, comme DBSCAN. Ces mesures se sont également révélées sensibles à la structure des données analysées. Par exemple, dans l'un des jeux de données où les clusters à identifier avaient des formes ellipsoïdales, ces mesures n'ont pas réussi à identifier correctement les solutions de clusters. Enfin, cette étude a comparé les trois mesures de qualité internes avec l'indice de Rand, une mesure de qualité externe, pour déterminer le nombre optimal de clusters. Les résultats ont montré que ces deux catégories de mesures conduisent à des nombres de clusters différents.

L'ensemble des challenges existant dans l'analyse des données de santé fait que les méthodes de clustering utilisées habituellement sont limitées. Ces limitations concernent la gestion du volume important de variables, de l'hétérogénéité des formats, de la nature longitudinale des variables et de la présence de données tronquées. Par conséquent, il devient essentiel de développer des algorithmes de clustering capables de répondre à ces limitations. C'est précisément cet objectif que j'ai poursuivi au cours de ma thèse.

1.3 Objectif et plan de la thèse

L'objectif de ma thèse a été de mettre au point des méthodes de clustering de patients à partir de données médico-administratives tout en tenant compte des caractéristiques et complexités spécifiques de ces données de santé. J'ai choisi pour cela de développer de nouvelles approches de clustering. Le développement de nouvelles méthodes de clustering est essentiel dans le contexte de l'usage secondaire des données de santé qui consiste à réutiliser les données collectées dans le cadre du soin à des fins de recherche. La complexité qui réside au sein de ces données constitue plusieurs défis pour les méthodes de clustering les plus couramment utilisées. Dans le cadre de ma thèse, je me suis particulièrement intéressée à deux challenges et j'ai développé deux méthodes de clustering novatrices pour répondre à ces challenges.

Le premier challenge a été de prendre en compte la nature longitudinale des données médico-administratives et plus particulièrement la présence des données tronquées. Les méthodes couramment utilisées, présentées dans la section 1.2.3, ont tendance à exclure les patients ou les variables présentant des données tronquées, ou à effectuer des imputations. Cependant, comme évoqué précédemment, de telles procédures peuvent biaiser la formation des clusters. Il est donc essentiel de développer des approches de clustering capables de prendre en compte efficacement les données tronquées, tout en préservant la qualité des regroupements. Pour répondre à ce challenge, j'ai développé une première méthode, appelée "cluster-tracking", qui a permis de prendre en compte la nature longitudinale des données de santé. Le "cluster-tracking", détaillé dans le chapitre 2, vise à identifier les clusters de patients à chaque instant temporel et à suivre leur évolution dans le temps. Cette méthode n'a nécessité ni exclusion, ni imputation, car les données sont exploitées uniquement aux temps où elles sont disponibles. Ainsi, la totalité des patients est utilisée dans l'analyse.

Le deuxième challenge a été de prendre en compte la nomenclature des labels des variables. Certaines variables de santé sont en effet caractérisées par des labels associés à des nomenclatures, au sein desquelles certains labels présentent des relations plus étroites que d'autres. Toutefois, les méthodes de clustering couramment utilisées se fondent uniquement sur les valeurs des variables, sans considérer les relations entre labels. En conséquence, le regroupement des données se limite à l'identification des valeurs similaires des variables. Il est donc nécessaire de développer des approches de clustering qui intègrent efficacement ces relations pour obtenir des regroupements plus significatifs et pertinents. La deuxième approche de clustering que j'ai développé a permis d'intégrer la notion de relations entre labels dans les mesures de similarité couramment utilisées. Cette démarche, détaillée dans le chapitre 3, permet de tenir

compte à la fois des valeurs des labels ainsi que de leurs relations lors du clustering. De cette manière, deux patients ayant des variables présentant à la fois des valeurs similaires et une relation étroite entre leurs labels sont considérées plus proches que deux patients ayant des variables partageant des valeurs similaires mais présentant une relation plus distante entre leurs labels.

Ces deux nouvelles méthodes ont également permis de répondre au challenge algorithmique lié au choix du nombre de clusters. Ce paramètre qui doit être prédéfini dans la majorité des méthodes de clustering, est souvent difficile à déterminer. Nous avons donc décidé d'identifier les clusters de patients en utilisant les réseaux. En effet, les algorithmes de clustering appliqués aux réseaux ne nécessitent pas de spécifier le nombre de clusters au préalable.

Pour l'application de ces méthodes de clustering novatrices, j'ai utilisé les données issues de l'EGB. Ces données fournissent des informations concernant les caractéristiques médicales et socio-démographiques des patients ainsi que sur leurs prestations remboursées. Mon attention s'est focalisée particulièrement sur les remboursements de médicaments. Dans un premier temps, j'ai appliqué le "cluster-tracking" pour effectuer le clustering des patients en prenant en considération la nature longitudinale des remboursements perçus sur la période allant de 2008 à 2018. Par la suite, j'ai exploré la relation entre les labels des médicaments dans le clustering. Ces labels sont structurés selon la nomenclature internationale ATC. L'intégration des relations entre les labels des médicaments a permis de prendre en compte les similarités en termes de composition et d'usage thérapeutique des médicaments.

Afin d'évaluer la qualité et d'interpréter les résultats de ces deux nouvelles méthodes de clustering, j'ai utilisé plusieurs mesures de qualité du clustering et outils de visualisation. Cela nous a permis de mettre en évidence des clusters cohérents et qui faisait sens au niveau clinique dans la méthode du cluster tracking. Dans notre deuxième méthode, nous avons pu mettre en évidence une amélioration de la qualité des clusters identifiés en intégrant les relations entre labels dans les mesures de similarité entre patients.

Dans la suite, le chapitre 2 introduira en détail les méthodes usuellement employées pour effectuer un clustering longitudinal, ainsi que l'application de la méthode du "cluster-tracking". Puis, le chapitre 3 abordera les métriques conçues pour incorporer les relations entre les labels dans le clustering. Enfin, dans le chapitre 4, les deux nouvelles approches de clustering développées seront discutées et des perspectives futures seront évoquées.

Chapitre 2

Clustering de patients à partir de variables longitudinales

Dans le chapitre précédent, nous avons exploré la complexité des données de santé et comment cette complexité limite l'application des méthodes de clustering usuelles. Une des principales complexités des données de santé que nous avons cherché à aborder au cours de ma thèse est leur nature longitudinale, qui engendre notamment la présence fréquente de données tronquées. Afin de tenir compte de cette complexité, nous avons développé une méthode appelée "cluster-tracking" pour identifier les clusters de patients à partir de données médico-administratives. De nombreuses méthodes de clustering ont été spécifiquement développées pour l'analyse de variables longitudinales. Ces méthodes se répartissent dans différentes catégories, chaque catégorie étant adaptée à des situations spécifiques. En introduction, j'ai présenté les méthodes de clustering d'un point de vue général, et je vais ici détailler plus particulièrement les méthodes de clustering de variables longitudinales.

2.1 Les principales approches de clustering longitudinal

Le clustering longitudinal des données de santé a pour objectif d'identifier des sous-groupes homogènes de patients en fonction de l'évolution de leurs caractéristiques au cours du temps. Cette analyse prend en compte les trajectoires individuelles de chaque patient et regroupe les patients ayant des trajectoires similaires. Cela permet donc de stratifier les patients et de mieux comprendre l'hétérogénéité et l'évolution des maladies. Il existe trois grandes catégories d'approches pour le clustering de données longitudinales [87] : les approches utilisant comme variables les données brutes, les approches utilisant comme variables les caractéristiques des variations longitudinales et les approches modélisant la variation longitudinale (voir *Table 2.1*). La principale différence entre ces trois catégories repose sur la manière dont sont pré-traitées les données initiales.

2.1.1 Approches à partir des variables brutes

Les approches de clustering longitudinal à partir des variables brutes sont directement appliquées aux variables longitudinales dans leur forme initiale, en considérant simultanément l'ensemble des instants temporels. Ces approches utilisent les méthodes usuelles de clustering (voir section 1.2.3), mais en ajustant les mesures de similarité pour qu'elles tiennent compte de la nature longitudinale des données. La mesure de similarité la plus basique est la distance Euclidienne appliquée dans un espace multidimensionnel où chaque dimension représente un instant temporel. Par exemple, les deux algorithmes Kml et Kml3d implémentent le K-means en utilisant, parmi d'autres, cette distance Euclidienne adaptée [88]. Une étude menée par Mullin et al. a appliqué Kml à des données de prescriptions d'opioïdes, aboutissant à l'identification de trois clusters de patients distincts [89]. En dehors de la distance Euclidienne, d'autre mesures spécifiques sont couramment utilisées pour quantifier la similarité entre les variables longitudinales, notamment DTW (*Dynamic Time Warping*), LCS (*Longest Common Subsequence*) ou STS (*Short Time Series distance*) [90].

DTW (*Dynamic Time Warping*)

DTW est une mesure qui calcule la similarité entre deux ensembles de variables longitudinales en minimisant la somme des distances entre les points temporels de ces ensembles [91]. Cette mesure présente l'avantage de tenir compte des décalages temporels entre les deux

ensembles de données. Ainsi, elle peut calculer aisément la similarité entre deux patients qui reçoivent des traitements à des moments temporels différents, par exemple. En revanche, la distance Euclidienne ne serait pas adéquate pour calculer cette similarité car elle exige que les points temporels correspondent exactement entre les deux patients. Là où la distance euclidienne requiert un alignement parfait des données, DTW offre une flexibilité dans cet alignement. Par exemple, Hebbrecht et al. ont appliqué le clustering hiérarchique en utilisant DTW pour identifier cinq clusters de patients souffrant de dépression et présentant des symptômes différents [92].

LCS (*Longest Common Subsequence*)

LCS est une mesure qui calcule la similarité entre deux ensembles de variables longitudinales en identifiant les séquences d'éléments en communs entre ces ensembles [93]. Cette mesure commence par analyser les séquences similaires et identifie parmi celles-ci la séquence la plus longue. Tout comme DTW, LCS autorise les décalages temporels. Elle est particulièrement adaptée pour les données longitudinales qui présentent des nombres de points temporels différents et qui comportent du bruit. Par exemple, Vogt, Scholz et Sundmacher ont appliqué l'algorithme du K-medoids avec LCS pour identifier des clusters de patients atteints d'insuffisance cardiaque selon leur traitement [94].

STS (*Short Time Series distance*)

STS est une mesure de similarité spécifiquement conçue pour la comparaison de données longitudinales de courte durée, c'est-à-dire les données pour lesquelles on ne dispose pas de beaucoup d'instants temporels. Cette mesure repose sur l'hypothèse que même avec une longueur réduite, de telles données peuvent contenir des informations importantes. STS calcule la similarité entre deux ensembles de données en analysant à la fois leur forme et leur amplitude. L'idée sous-jacente est que deux ensembles de données partageant des formes et amplitudes communes ont plus de probabilité d'être similaires que deux ensembles aux formes et amplitudes différentes. A titre d'exemple, Möller-Levet et al. ont développé une approche de clustering flou en intégrant STS dans le but d'analyser les données longitudinales d'expression de gènes [95].

Les diverses mesures de similarité mentionnées dans la section précédente permettent de prendre en compte divers aspects des données longitudinales, tels que leur longueur et les

décalages temporels. Cela permet ainsi d'adapter les approches du clustering longitudinal à partir des données brutes aux différentes données longitudinales. Cependant, il est important de noter que ces données brutes sont souvent de grande dimension en raison des nombreux instants temporels par variable et des nombreuses variables, rendant leur exploitation difficile. Dans de tels cas, une solution consisterait à recourir aux approches de clustering longitudinal à partir de l'extraction de caractéristiques qui permettent de réduire la dimension des données.

2.1.2 Approches à partir de l'extraction de caractéristiques

Les approches de clustering longitudinal à partir de l'extraction de caractéristiques impliquent un pré-traitement des données longitudinales brutes. A partir de ces données brutes, un ensemble de caractéristiques des variations longitudinales est extrait afin de synthétiser au mieux l'information temporelle contenue dans ces données. Cela permet donc d'éliminer la dimension longitudinale de ces données. Les caractéristiques extraites peuvent ensuite être utilisées dans les méthodes de clustering traditionnellement appliquées (voir section 1.2.3) pour identifier les clusters. Un grand nombre de caractéristiques peut être extrait à partir des données longitudinales. Les caractéristiques les plus fréquemment extraites sont la moyenne, l'écart type, le kurtosis et l'asymétrie [96]. D'autres caractéristiques courantes incluent la médiane, la variance, la pente ainsi que les mesures de corrélations.

Le kurtosis

Le kurtosis est une mesure statistique utilisée pour évaluer la forme de la distribution des données longitudinales par rapport à une distribution normale (gaussienne). En pratique, il fournit des indications sur la concentration des points autour de la moyenne ainsi que dans les extrémités de la distribution. Un kurtosis élevé suggère que la distribution présente des extrémités plus épaisses que celles d'une distribution normale. Par conséquent, cette mesure peut renseigner sur la présence éventuelle de données aberrantes.

L'asymétrie

Comme son nom l'indique, cette mesure statistique donne des informations sur le degré d'asymétrie de la distribution des données par rapport à une distribution normale. Elle permet de déterminer s'il existe un déséquilibre entre les extrémités de la distribution par rapport à la

moyenne. Une asymétrie positive indique que l'extrémité droite est plus étendue, tandis qu'une asymétrie négative indique que l'extrémité gauche est plus étendue. Une asymétrie proche de zéro indique que la distribution est parfaitement symétrique.

Un package Python nommé tsfresh a été spécialement conçu pour extraire automatiquement jusqu'à 794 caractéristiques à partir des données longitudinales [97]. Tiano, Bonifati et Ng ont développé une méthode utilisant l'algorithme du K-medoids avec des caractéristiques extraites via tsfresh et l'ont appliquée pour identifier des clusters de patients souffrant de maladies rénales [98].

Il est donc possible d'extraire un grand nombre de caractéristiques à partir des données longitudinales. Cependant, la sélection des caractéristiques les plus pertinentes pour synthétiser au mieux l'information temporelle des données peut s'avérer une tâche complexe. Pour contourner cette difficulté, les approches de clustering longitudinal basées sur les modèles offrent une alternative. Elles permettent de représenter les données longitudinales sous forme de modèles, éliminant ainsi le besoin d'extraire des caractéristiques ou de manipuler directement les données brutes.

2.1.3 Approches à partir des modèles

Les approches de clustering longitudinal à partir de modèles supposent que la trajectoire des différentes variables longitudinales brutes est générée à partir d'un modèle statistique sous-jacent. Dans ce contexte, les clusters contiennent des données dont les modèles génératifs sont similaires. Ces approches visent à estimer les paramètres optimaux des modèles, tels que la pente ou l'ordonnée à l'origine, pour chaque cluster. Ensuite, chaque variable est affectée au cluster dont les paramètres de modèle correspondent le mieux à ceux de la variable respective. Dans le domaine de la santé, deux approches de clustering longitudinal à partir des modèles sont couramment utilisées : LCGA (*Latent Class Growth Analysis*) et GMM (*Growth Mixture Modeling*) [99].

LCGA (*Latent Class Growth Analysis*)

LCGA est une approche de clustering qui caractérise chaque variable longitudinale par un modèle de croissance spécifique [100]. Cette approche impose la contrainte qu'il n'y ait aucune variance ni covariance entre les paramètres de croissance estimés au sein d'un même cluster. Par conséquent, ces paramètres sont constants à l'intérieur de chaque cluster. Cela implique

que les modèles qui caractérisent les données contenues dans un même cluster ne présentent aucune variation, ce qui ne reflète pas toujours la complexité des données réelles. Un exemple d'application de LCGA est l'étude menée par Downie et al. où des clusters de patients souffrant de douleurs lombaires aiguës ont été identifiés en utilisant des scores de douleurs collectés sur une période de 12 semaines [101].

GMM (*Growth Mixture Modeling*)

Tout comme LCGA, GMM caractérise chaque variable longitudinale par un modèle de croissance [100]. Cependant, GMM autorise une faible variation entre les paramètres de croissance estimés au sein d'un même cluster. Cette flexibilité permet d'estimer une plus grande diversité de modèles par rapport à LCGA, ce qui conduit à un meilleur ajustement aux données. Cependant, cette flexibilité rend également l'algorithme plus complexe car le nombre de paramètres à estimer augmente en conséquence. Un exemple d'application de GMM a été menée par Colder et al. pour identifier des clusters d'adolescents dont les habitudes de tabagisme ont été relevées sur une période de quatre ans [102].

Autres approches de clustering à partir de modèles

En plus de LCGA et GMM, d'autres approches sont utilisées pour réaliser du clustering longitudinal à partir de modèles. Par exemple, *gaussian mixture model* est une approche qui caractérise les données longitudinales par des modèles gaussiens [103]. Les paramètres de ces modèles sont estimés à l'aide de l'algorithme EM (Expectation-Maximization) qui maximise la vraisemblance des données. Un autre exemple d'approche est HMM (*Hidden Markov Models*) qui considère que les variables sont générées par un processus stochastique qui n'est pas directement observable [104]. Ce processus est composé d'une série d'états cachés qui évoluent selon des probabilités de transition. Les paramètres sont également estimés à l'aide de l'algorithme EM. De plus, DPMM (*Dirichlet Process Mixture Models*) est une approche bayésienne qui suppose que les variables sont générées à partir d'un mélange infini de modèles [105]. Dans cette approche, le nombre de clusters n'est pas fixé à l'avance, mais est considéré comme une variable aléatoire. Cela permet d'ajuster automatiquement le nombre de clusters en fonction des données, offrant une plus grande flexibilité dans la modélisation. Ces différents exemples illustrent la variété d'approches disponibles pour le clustering longitudinal à partir de modèles.

Approches	Principes	Limites
à partir des variables brutes	algorithmes appliqués directement sur les données brutes	- choix du nombre de clusters - imputation des données tronquées
à partir de l'extraction de caractéristiques	algorithmes appliqués sur des caractéristiques extraites à partir des données brutes	- choix du nombre de clusters - choix des caractéristiques pertinentes à extraire - temps d'exécution croissant avec le nombre de caractéristiques
à partir de modèles	algorithmes estimant les paramètres optimaux des modèles sous-jacents aux données brutes	- choix du nombre de clusters - imputation des données tronquées - analyse simultanée de plus de 3 variables difficile - temps d'exécution long

Table 2.1 – Les principales approches de clustering longitudinal

2.1.4 Les limites des approches de clustering longitudinal

Les approches de clustering longitudinal présentent plusieurs limites, comme indiqué dans la *table 2.1*). Les approches de clustering longitudinal "à partir des variables brutes" et "à partir de l'extraction des caractéristiques" se basent sur les méthodes usuelles de clustering (non longitudinales) pour identifier les clusters. Par conséquent, les challenges énoncés dans la section 1.2.4 limitent également ces approches. Parmi ces limites, on retrouve les difficultés liées au choix du nombre de clusters et de la mesure de similarité, les problèmes de convergence des algorithmes ainsi que la complexité d'interprétation des résultats. De plus, en raison de la complexité des données longitudinales, ces approches peuvent rencontrer des contraintes en termes de temps d'exécution et de capacités de stockage, ainsi que des difficultés à capturer les formes irrégulières des données et à gérer les données tronquées.

En plus du choix du nombre de clusters et de la mesure de similarité, les approches de clustering longitudinal "à partir de l'extraction de caractéristiques" nécessitent de choisir les caractéristiques des variations longitudinales à extraire des données. En effet, une diversité de caractéristiques peut être utilisée pour décrire les données longitudinales. Selon les caractéristiques choisies, le résultat du clustering peut varier.

En ce qui concerne les approches de clustering longitudinal "à partir des modèles", il est souvent nécessaire de prédéfinir le nombre de clusters. De plus, ces approches exigent fréquemment l'imputation des données tronquées pour assurer leur fonctionnement. Par ailleurs, la plupart de ces approches se basent sur l'utilisation d'une seule variable pour réaliser le clus-

tering car l'analyse simultanée de plus de trois variables devient généralement compliquée d'un point de vue computationnel.

Face à ces diverses limites inhérentes aux approches de clustering longitudinal, j'ai développé une nouvelle méthode intitulée "cluster-tracking". Cette méthode vise à surmonter plusieurs de ces limitations en permettant l'identification des clusters de patients à chaque instant temporel et en suivant ensuite l'évolution des clusters au cours du temps, comme illustré dans la *figure 2.1*. Ainsi, elle offre la possibilité de capturer les trajectoires de clusters regroupant des patients partageant des évolutions similaires dans leurs caractéristiques de santé. Cette nouvelle méthode ambitionne de répondre aux défis liés au choix du nombre de clusters, à l'analyse simultanée de multiples variables longitudinales et à la gestion des données tronquées. Ce travail a fait l'objet d'une publication dans le "*Journal of Biomedical Informatics*" et est présenté dans la section 2.2. Le matériel supplémentaire lié à cette publication est accessible dans l'annexe A.

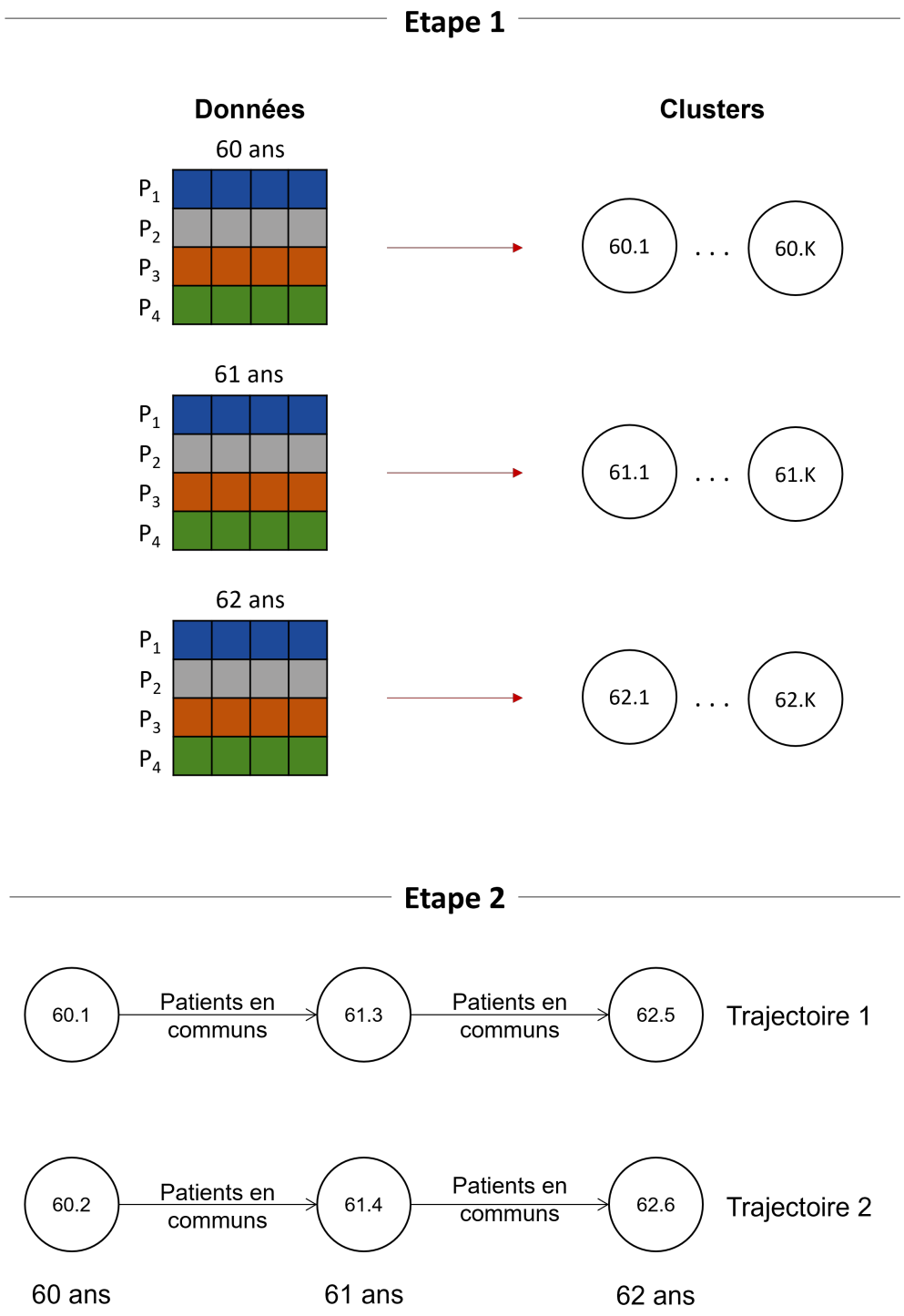


Figure 2.1 – Principe du "cluster-tracking"
L'étape 1 consiste à identifier les clusters par temps (ici par âge). L'étape 2 consiste à suivre l'évolution des clusters au cours du temps à partir du nombre de patients en communs

2.2 Publication numéro 1 : Tracking clusters of patients over time enables extracting information from medico-administrative databases

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

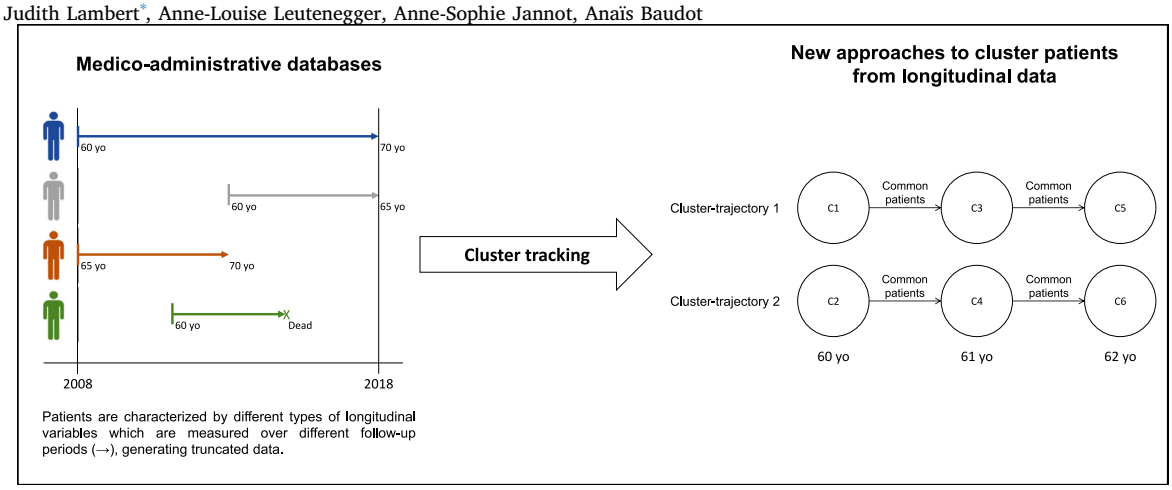
journal homepage: www.elsevier.com/locate/yjbin

Graphical Abstract

Tracking clusters of patients over time enables extracting information from medico-administrative databases

Judith Lambert*, Anne-Louise Leutenegger, Anne-Sophie Jannot, Anaïs Baudot

Journal of Biomedical Informatics xxx (xxxx) xxx



Patients are characterized by different types of longitudinal variables which are measured over different follow-up periods (→), generating truncated data.

Graphical abstract and Research highlights will be displayed in online search result lists, the online contents list and the online article, but will not appear in the article PDF file or print unless it is mentioned in the journal specific style requirement. They are displayed in the proof pdf for review purpose only.

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Tracking clusters of patients over time enables extracting information from medico-administrative databases

Judith Lambert [a,b,c,*], Anne-Louise Leutenegger [d], Anne-Sophie Jannot [b,e,f,1], Anaïs Baudot [c,g,h,1]

[a] Sorbonne Université, Université Paris Cité, INSERM, Centre de Recherche des Cordeliers, F-75006 Paris, France
[b] HeKA, Inria Paris, F-75015 Paris, France
[c] Aix Marseille Univ, INSERM, MMG, UMR1251, Marseille, France
[d] Université Paris Cité, INSERM, NeuroDiderot, UMR1141, 75019 Paris, France
[e] Université Paris Cité, Sorbonne Université, INSERM, Centre de Recherche des Cordeliers, F-75006 Paris, France
[f] French National Rare Disease Registry (BNDMR), Greater Paris University Hospitals (AP-HP), Paris, France
[g] CNRS, Marseille, France
[h] Barcelona Supercomputing Center, Barcelona, Spain

ARTICLE INFO

Keywords:
Longitudinal clustering
Cluster tracking
Medico-administrative databases
Patient networks

ABSTRACT

Context: Identifying clusters (i.e., subgroups) of patients from the analysis of medico-administrative databases is particularly important to better understand disease heterogeneity. However, these databases contain different types of longitudinal variables which are measured over different follow-up periods, generating truncated data. It is therefore fundamental to develop clustering approaches that can handle this type of data.
Objective: We propose here cluster-tracking approaches to identify clusters of patients from truncated longitudinal data contained in medico-administrative databases.
Material and Methods: We first cluster patients at each age. We then track the identified clusters over ages to construct cluster-trajectories. We compared our novel approaches with three classical longitudinal clustering approaches by calculating the silhouette score. As a use-case, we analyzed antithrombotic drugs used from 2008 to 2018 contained in the Échantillon Généraliste des Bénéficiaires (EGB), a French national cohort.
Results: Our cluster-tracking approaches allow us to identify several cluster-trajectories with clinical significance without any imputation of data. The comparison of the silhouette scores obtained with the different approaches highlights the better performances of the cluster-tracking approaches.
Conclusion: The cluster-tracking approaches are a novel and efficient alternative to identify patient clusters from medico-administrative databases by taking into account their specificities.

1. Introduction

The reuse of medico-administrative databases is nowadays extremely popular. Such databases are indeed increasingly available for epidemiological, clinical and healthcare research to study a large range of health-related issues [1]. However, medico-administrative databases are complex and appropriate analysis methods are required [2]. First, each patient is described through a large number of variables. Analysis methods able to deal with high dimensional data are hence needed. Second, these variables are of a different nature (e.g., drug reimbursements, diagnoses, hospitalizations), and the methods need to consider heterogeneity. Finally, the variables vary over time and are measured over different follow-up periods, thereby generating truncated data when focusing on a given stage of life or disease. This time dimension is very difficult to apprehend and, overall, only few methods can deal with high dimensional truncated longitudinal data.

Among the various objectives targeted by the reuse of medico-administrative databases, the identification of clusters (i.e., subgroups) of patients is particularly significant. Indeed, given the complexity and the heterogeneity of human diseases, we have to move from a "one size fits all" paradigm towards a more personalized care and a better understanding of disease heterogeneity [3,4]. In general, clusters of patients related to a given disease are identified using the coded diagnoses. However, in medico-administrative databases, the diagnoses are often missing due to truncated patient history. For example, if a patient had an infarction twenty years ago, the hospital stay related to this event will not be available in the database but the patient will still

* Correspondence to: PariSanté Campus, 10 rue d'Oradour-sur-Glane, 75015 Paris, France.
E-mail address: judith.lambert@inserm.fr (J. Lambert).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jbi.2023.104309
Received 3 October 2022; Received in revised form 22 December 2022; Accepted 6 February 2023

have treatments for secondary prevention of cardiovascular diseases. Patient history could hence be inferred from their current treatments.

To the best of our knowledge, three categories of approaches are available to cluster patients using longitudinal data. These longitudinal clustering approaches are raw-data-based, feature-based and model-based [5]. In raw-data-based approaches, classical (non-longitudinal) clustering algorithms, such as Kmeans, adapt their similarity measure to be applied to the raw longitudinal data. For instance, Kmeans adapted to raw longitudinal data has been used to identify clusters of children based on inattention and hyperactivity during elementary school [6], or to assess the relationships between fibrosis and bio-clinical parameters [7]. In feature-based approaches, features are first extracted from the raw longitudinal data. These extracted features are then used as input for classical (non-longitudinal) clustering algorithms. For instance, Wang et al. extracted several features from longitudinal data in three (non-clinical) benchmark datasets [8]. They then used the extracted features as input in hierarchical clustering and in an unsupervised neural network algorithm. Although only a small number of features are used for the clustering, the identified clusters are similar to the clusters identified using all the data. Finally, model-based approaches assume that the raw longitudinal data are generated by a mixture of models and intend to extract the parameters of these models. Model-based approaches are, to the best of our knowledge, the most frequently used in biomedical research. The two prevailing model-based approaches are Growth Mixture Modeling (GMM) and Latent Class Growth Analysis (LCGA) [9]. These methods identify clusters of patients based on the common evolution of their longitudinal variables over time. GMM allows small variations around this common evolution between patients within cluster whereas LCGA assumes no variation [10]. Mora et al. applied GMM to identify clusters of women according to the magnitude and timing of depressive symptomatology from pregnancy to two years postpartum [11]. Colder et al. also used GMM to identify clusters of adolescents based on their smoking behavior over four years [12]. LGCA was used by Downie et al. to identify clusters of patients with acute low back pain from pain scores over twelve weeks [13] and by Landa et al. to identify clusters of babies at high risk for autism based on their language, motor and nonverbal cognitive functioning from 6 to 36 months [14].

However, raw-data-based, feature-based and model-based longitudinal clustering approaches have some limitations. For instance, truncated data are not handled. Patients with truncated data must be removed or their data must be imputed. In the context of medico-administrative databases, truncated data are an inescapable issue, as patients are followed-up over a fixed period. In addition, the number of clusters must be specified *a priori*. To determine the optimal number of clusters, criteria are usually used to assess the quality of the clustering [15]. These criteria include for instance the silhouette score [16,17] or the Davies–Bouldin criterion [18,19]. However, the optimal number of clusters might differ depending on the criterion chosen [20]. Another limitation specific to the model-based approaches is that the majority of the studies focus on only one longitudinal variable. The joint analysis of two or three longitudinal variables is possible [21–24], but becomes computationally challenging for more than three variables. Finally, in all three categories of approaches, each patient is assigned to only one cluster over the entire time period.

An alternative strategy for clustering patients from longitudinal data could be cluster tracking. Cluster tracking is an approach mainly used in the field of social network analysis [25]. It is a two-step strategy. In the first step, the clusters are identified at each time point. In the second step, the clusters are matched between the different time points to allow their tracking along the timeline. Clusters are identified at each time point using non-longitudinal clustering algorithms [26,27].

Different methods can be used to identify clusters including methods such as Kmeans or network clustering algorithms. For instance, Li et al. constructed a patient network based on clinical similarity and performed a clustering approach in order to identify subtypes of type 2 diabetes [28]. Wang et al. constructed patient networks from omics data and identified clusters of cancer patients with different survival profiles [29]. Patient networks have the advantage of preserving privacy because the interactions between patients are considered rather than absolute data [30]. In addition, a large number of algorithms exist for clustering networks [31]. However, to our knowledge, current network-based approaches to identify patient clusters do not consider longitudinal data.

We propose here novel cluster-tracking approaches to identify patient clusters and trajectories from longitudinal data contained in medico-administrative databases. Our approaches starts by identifying clusters of patients at each time step. Patient clusters are identified using two clustering strategies: Kmeans directly applied to the raw data or the Markov Cluster algorithm (MCL) applied to patient networks constructed from raw data. We then track the clusters identified at the different time steps based on their sharing of patients. As a use-case, we analyzed drug reimbursements contained in the national cohort managed by the French health insurance, called the Échantillon Généraliste des Bénéficiaires (EGB). Our aim was to identify clusters of patients that could be related to given diseases using only drug reimbursements and in the absence of any coded diagnoses. We identified different trajectories of patient clusters with clinical interest. Finally, we compared these cluster-tracking approaches with three existing types of longitudinal clustering approaches, by calculating a modified silhouette score. The best modified silhouette scores were obtained with the two cluster-tracking approaches.

2. Material and methods

2.1. Cluster-tracking approach

We propose novel approaches for clustering patients from longitudinal data extracted from medico-administrative databases. These approaches start by identifying clusters of patients at each time step. To this goal, we used two different clustering strategies: the Markov Cluster algorithm (MCL) applied to patient networks built from raw data and Kmeans applied directly on raw data. Clusters are then tracked over time steps to define cluster-trajectories.

2.1.1. Identifying clusters of patients from patient networks

The first clustering strategy used to identify clusters of patients relies on the construction of patient networks. We started by constructing a patient network for each time step. We then applied the MCL clustering algorithm on each network.

Constructing patient networks. A patient network is a graph $G = (V, E)$ with V patient nodes and E edges representing interactions between patient nodes. We built a network for each time step. Each network is constructed using a similarity matrix $M_i = [m_{p_1,p_2}]^n$ where n is the number of patients, i is the time step and m_{p_1,p_2} is the similarity between patients p_1 and p_2 at the time step i. This similarity matrix is symmetrical, with $m_{p_1,p_2} = m_{p_2,p_1}$.

The similarity between patients at time step i can be computed using different similarity measures. We tested four different similarity measures: the Cosine similarity, the opposite of the normalized Euclidean distance, the Jaccard index and the generalized Jaccard index (*Supplementary section S1*).

The similarity matrices built for each time step are then filtered according to a threshold t. The goal of the filtering step is to obtain networks with a reduced number of edges [32]. The filtered matrices are next used to build patient networks. We tested different thresholds. For each threshold t, the filtered matrix M_i^t is obtained as follows:

$$M_i^t = \begin{cases} m_{p_1,p_2} & \text{if } m_{p_1,p_2} \geq t \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where a null value indicates that patients p_1 and p_2 have a similarity value below the threshold t and will thereby not be connected in the patient network. From each similarity matrix M_i^t, the associated patient network can be constructed. An edge between patients P_1 and P_2 is weighted by the value m_{p_1,p_2} of the matrix.

Reducing the number of edges may lead to disconnected nodes. Therefore, we selected the threshold t in the similarity matrices which allowed us to obtain the minimum number of isolated patient nodes in any network (*Supplementary section S2*).

Clustering patient networks. We applied the Markov Cluster algorithm (MCL) [33] on the largest connected component of the patient networks. The MCL algorithm uses random walks to simulate flows on the network. The flows allow to distinguish network areas where nodes are strongly connected, which correspond to the clusters. We used the version 0.0.6.dev0 of the "markov-clustering" Python package with the default parameters.

2.1.2. Identifying clusters of patients from raw data

We described in the previous section a clustering strategy based on patient networks. We also used Kmeans as a second clustering strategy [34]. Kmeans is applied directly on raw data, for each time step. In Kmeans, the number of clusters must be specified *a priori*. We determined the optimal number of clusters per time step by calculating the silhouette score [35]. The silhouette score assesses the clustering quality by computing the separation distance between the obtained clusters.

Let us define

$$a^i(p) = \frac{1}{|C_p^i| - 1} \sum_{j \in C_p^i, j \neq p} d(p, j), \tag{2}$$

the mean distance of patient p to their cluster C_p^i at time step i, with $|C_p^i|$ the number of patients in C_p^i and $d(p, j)$ the Euclidean distance between patients p and j belonging to C_p^i, and let

$$b^i(p) = \min_{C_z^i \neq C_p^i} \frac{1}{|C_z^i|} \sum_{z \in C_z^i} d(p, z), \tag{3}$$

be the mean distance of a patient p to their neighboring cluster C_z^i at time step i, with $|C_z^i|$ the number of patients in C_z^i and $d(p, z)$ the Euclidean distance between the patient p belonging to C_p^i and the patient z belonging to C_z^i.

We start by calculating the silhouette score for each patient at time step i as follows:

$$s^i(p) = \frac{b^i(p) - a^i(p)}{\max(a^i(p), b^i(p))}, \tag{4}$$

The silhouette score at a given time step i over all the patients is obtained as follows:

$$S^i = \frac{1}{K^i} \sum_{k=1}^{K^i} \frac{1}{|C_k|} \sum_{p \in C_k} s^i(p), \tag{5}$$

with K^i the number of clusters at time step i, $|C_k|$ the number of patients in the cluster C_k.

The silhouette score varies between −1 and 1. Values close to 1 indicate that the clusters are well-separated. Values close to 0 indicate overlapping clusters. Negative values indicate that the clusters are worse than random.

2.1.3. Tracking the clusters over time steps

In the previous step, we identified sets of clusters per time step either from patient networks with MCL or from raw data with Kmeans. We then intend to follow the clusters over the different time steps. Let C^i and C^{i+1} be two sets of clusters identified at 2 consecutive time steps, i and $i + 1$. We computed the intersection (i.e., the number of common patients) between every pair of clusters (c, c') obtained at 2 consecutive time steps:

$$Q^i(c, c') = |c \cap c'| \ \forall i, \tag{6}$$

with $c \in C^i$ and $c' \in C^{i+1}$.

Next, for each cluster $c \in C^i$, we identified the cluster from the set of clusters C^{i+1} having the greatest number of common patients as follows:

$$T_c^i = \underset{c'}{\operatorname{argmax}}\, Q^i(c, c'). \tag{7}$$

Please note that if, for the cluster c, there is more than one cluster match in T_c^i (i.e., if there is more than one cluster with the same maximum number of common patients), all the clusters are included in T_c^i.

We visualized the tracking of clusters with an alluvial plot, in which the blocks represent the clusters and the stream fields between the blocks represent the number of common patients. The height of the blocks and the thickness of the stream fields are proportional to the number of patients.

2.1.4. Identifying cluster-trajectories

We identified in the previous section sets of successive clusters. We called the sets of successive clusters cluster-trajectories. Patients in the same cluster-trajectory are considered to follow the same evolution over time for the longitudinal variables of interest.

The cluster-trajectories are visualized using a flowchart composed of blocks representing the clusters. The arrow thickness between the blocks represents the number of common patients. All clusters identified are described using the meta-information available for the patients.

2.2. Longitudinal clustering approaches

We compared the performance of the cluster-tracking approaches proposed in this work to existing state-of-the-art approaches dedicated to clustering patients using longitudinal data. The three categories of state-of-the-art longitudinal clustering approaches are raw-data-based, feature-based and model-based approaches [5,36]. We selected three specific methods, each representative of a category of approach. All longitudinal clusters identified with these methods are described using the meta-information available for the patients.

2.2.1. Raw-data-based approach

Raw-data-based approaches work directly with longitudinal raw data [5,36]. We selected Kml3d, an R package providing an implementation of Kmeans specifically designed for longitudinal data [37]. This package takes as input a 3-dimensional matrix $M(n, i, y)$ with n the patients, i the time step and y the set of variables characterizing the patients. The algorithm calculates the Euclidean distance between all patients (in n-dimensional space). Patients with the smallest distance are grouped in the same cluster. Importantly, the number of cluster needs to be defined *a priori*.

Kml3d cannot handle truncated data but allows imputation using different methods. We used the copy mean method (default), which imputes data using a linear interpolation and adds a variation to adapt the shape of the interpolation to the shape of the mean of the other values [38]. Patients are removed from the analysis when their number of truncated data are greater than $|I| - 2$, with I the set of time steps.

2.2.2. Feature-based approach

Raw data usually have a high dimension. The goal of the feature-based approaches is to reduce the dimensions by extracting several features characterizing the longitudinal data [5,36]. These features can then be used as input in classic (non-longitudinal) clustering algorithms, such as Kmeans or hierarchical clustering. We extracted the most common features: mean, standard deviation, kurtosis and skewness [39]. The kurtosis and the skewness describe the shape of the distribution of longitudinal data. We therefore obtained four features per patient and per longitudinal variable. These features were used as input in Kmeans.

2.2.3. Model-based approach

In model-based approaches, each longitudinal variable is characterized by a model or a mixture of models [5,36]. We applied Growth Mixture Modeling (GMM), which assumes that a model with a given mean and shape is associated with each cluster [10]. Let y_p be a longitudinal variable of the patient p composed of j repeated observations and K the number of clusters, distributed with probabilities π_k with $k = 1, \dots, K$, $\pi_k \in [0, 1]$ and $\sum_k \pi_k = 1$. A growth mixture modeling is defined as follows:

$$y_{p,j|k} = \beta_{0p}^k + \beta_{1p}^k \cdot i_j + \epsilon_{pj}^k, \tag{8}$$

with i_j the time step at the jth observation of the variable y, ϵ_{pj}^k the time-specific residual errors, and $(\beta_{0p}^k, \beta_{1p})$ the patient-specific coefficients.

In GMM, analyzing several variables simultaneously is computationally challenging. GMM can be applied separately for each variable, but this assumes that all longitudinal variables are independent from each other. We hence decided to use an aggregated variable $Y_p = [\sum_{v^i \in V_p^i} v^i \ \forall i \in I_p]$, with I_p the set of time steps of the patient p and V_p^i the set of longitudinal variables of the patient p at time step i. This aggregated variable allows us to apply a single GMM.

GMM calculates for every patient their posterior probability of belonging to each cluster using this aggregated variable as input. The cluster assigned to each patient is the one with the greatest posterior probability.

2.2.4. Determining the optimal number of clusters

In the raw-data-based, the feature-based and the model-based approaches, the number of clusters must be specified as a parameter *a priori*. In order to determine the optimal number of clusters, we calculated several classic clustering quality criteria (*Supplementary section S3)*). In the raw-data-based and the feature-based approaches, we calculated the Calinski–Harabasz criterion [40], the Kryszczuk variant of Calinski–Harabasz criterion [41], the Genolini variant of Calinski–Harabasz criterion [37], the opposite of Ray–Turi criterion [42] and the opposite of Davies–Bouldin criterion [43]. In the model-based approach, we calculated the Akaike Information Criterion (AIC) [44] and the Bayesian Information Criterion (BIC) [45]. Furthermore, for all the approaches, we calculated a modified silhouette score as follows:

$$S = \frac{1}{|I|} \sum_{i \in I} S^i, \tag{9}$$

with S^i the silhouette score at the time step i (Eq. (5)) and I the set of time steps. In this modified silhouette score, we calculated the silhouette score S^i at each time step rather than over the entire period. This avoids imputing truncated data.

2.3. Choice of the metric to compare the performances of the different approaches

In the cluster-tracking approaches, we used two clustering strategies: one based on network (Section 2.1.1) and one based on raw data (Section 2.1.2). In order to compare the clustering quality of these two clustering strategies, we calculated the modified silhouette score (Eq. (9)). We also calculated this modified silhouette score in the three longitudinal-clustering approaches. This allowed us to compare the clustering quality of the different approaches.

We estimated the 95% confidence interval of the modified silhouette score using the percentile bootstrap method [46]. We generated 100 bootstrap samples by resampling with replacement patients present in the population of interest. In each bootstrap sample, we applied the different approaches and we calculated the modified silhouette score. We obtained the confidence interval by taking the 2.5th and the 97.5th percentile of the distribution of the modified silhouette scores.

Table 1
Example of drug reimbursements contained in the EGB.

Patient ID	Reimbursement date	ATC class	Drug name
P_1	01/04/2008	M01	Ibuprofen
P_1	01/12/2015	B01	Aspirin
P_2	01/02/2010	N02	Tramadol
P_3	01/05/2016	B01	Clopidogrel

M01: Anti-inflammatory and antirheumatic products, B01: antithrombotic agents, N02: Analgesics.

2.4. Use-case: the Echantillon Généraliste des Bénéficiaires

We used longitudinal health data from the Echantillon Généraliste des Bénéficiaires (EGB), a French medico-administrative database. The EGB is a random sample from the French health insurance database [47]. It is representative of the French population and contains approximately 660,000 individuals followed over a period of 11 years. This study has been declared to INSERM (Institut National de la Santé et de la Recherche Médicale, https://www.inserm.fr/). The information provided to individuals in EGB on the possible reuse of their data and the procedures for exercising their rights comply with the legislative and regulatory provisions applicable to the processing of personal data in the SNDS (Système National des Données de Santé, https://www.snds.gouv.fr/SNDS/Accueil). According to French regulation, individuals in SNDS database are informed of the reuse of their data for research and can opposed to this reuse as defined by Articles 92 to 95 of Decree No. 2005-1309 of 20 October 2005 (https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037300884/). As required from French regulation, EGB data can be reuse for research projects from authorized persons once the research project is declared to their institution (INSERM).

Among others, EGB contains drug reimbursements, which are longitudinal high dimensional data that can be used to identify subgroups of patients (Fig. 1). We extracted data on drugs reimbursements between 2008 and 2018. For each patient, the date of reimbursement, the Anatomical Therapeutic Chemical (ATC) class and the name of the reimbursed drugs are indicated (see example *Table* 1). The ATC class is an international classification of drugs established by the World Health Organization (WHO) [48]. We only considered reimbursement of drugs belonging to the ATC class of antithrombotic agents (i.e., B01). We obtained 164,942 patients with such reimbursements. We further selected patients aged 60 to 70 and having had at least one drug reimbursement for two or more consecutive months. Our goal was to focus only on patients with sustained reimbursements. Our final dataset is composed of 30,111 different patients and 19 different drugs. There is a majority of men in this population, with a sex ratio (men/women) of 0.61. This is consistent with the fact that cardiovascular diseases, which accounts for the majority of antithrombotic use, is more common in men.

We also extracted data on long-term illnesses (i.e., illnesses that last at least 6 months) from the EGB. 23,063 patients out of the 30,111 patients studied experienced at least 1 long-term illness between 60 and 70 years old. These long-term illnesses represent 865 distinct diseases. Each disease is coded with the 10th revision of the international statistical classification of diseases and related health problems (ICD-10 code).

We decided to choose the age of the patient as time steps. Indeed, we did not have information about patients' thrombotic events nor about the initial intake of antithrombotic drugs. Choosing age as time steps is also consistent with the fact that the antithrombotic use strongly hinges on age. We hence calculated, for each patient, the number reimbursements for each drug at a given age (see example Table 2). We therefore obtained a table per patient age. Focusing on patients aged 60 to 70 years old, we obtained a total of 11 tables.

Importantly, we observed three types of truncated data (Fig. 2).

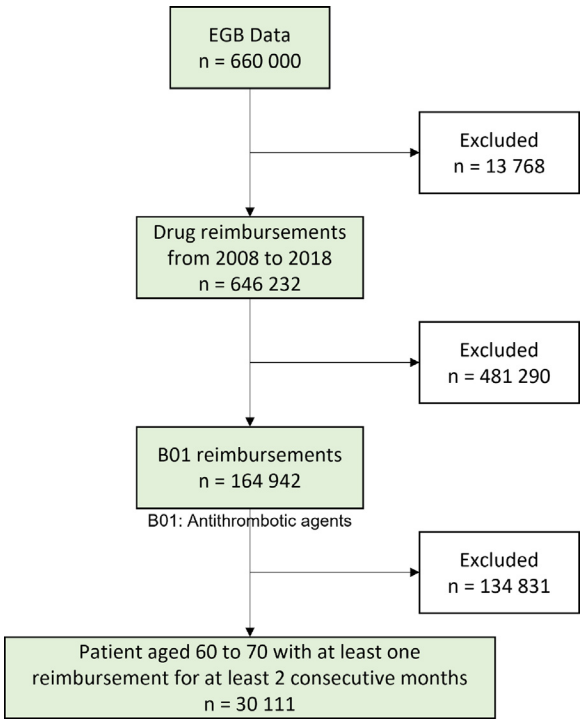


Fig. 1. Extraction of longitudinal data from the EGB, considered as a use-case in this study.
From the EGB medico-administrative database, we extracted antithrombotic drugs reimbursed for at least two consecutive months from 2008 to 2018 in patients ages 60 to 70. We therefore keep here only patients with sustained reimbursements. n: number of patients, B01: antithrombotic agents.

Table 2
Example of total number of reimbursements that three patients aged 60 years received for three drugs.

Patient ID	B01_1	B01_2	B01_3
P_1	0	10	5
P_2	1	8	4
P_3	2	6	3

B01_1, B01_2 and B01_3 are three different drugs belonging to the ATC class B01, the antithrombotic agents.

In France, no more than one month's treatment can be dispensed. We therefore considered that the number of reimbursement for a drug is a good proxy for annual drug use. In the following, we suppose that when patients have reimbursements for a drug, they are exposed to that drug.

We also applied our approaches to cluster patients with primary sclerosing cholangitis contain in pbcseq, a public database (*Supplementary section S8*). This database is a clinical trial including, among another, laboratory measurements.

3. Results

3.1. Cluster-tracking approaches allow identifying and tracking patient clusters over ages to identify cluster-trajectories

We first apply two different clustering strategies to identify clusters of patients at each age. The first clustering strategy is applied to patient networks (Material and methods 2.1.1). The second clustering strategy is directly applied to raw data (Material and methods 2.1.2). The clusters are then tracked over ages to define cluster-trajectories.

3.1.1. Identifying cluster-trajectories with the cluster-tracking approach based on networks

The first clustering strategy used in the cluster-tracking approach relies on the construction of patient networks (Material and methods 2.1.1). Patient networks are constructed using similarity matrices. Different measures can be computed to calculate similarities between patients and construct the similarity matrices (*Supplementary section S1*). We selected the Cosine similarity because it has the greatest variance. Using this Cosine similarity, we constructed 11 similarity matrices. In each matrix, the similarities are computed between all patients of a given age (from 60 to 70 years old). For example, the 60-year-old matrix is constructed by computing the similarities between all patients aged 60 between 2008 and 2018. Patient networks are then constructed by applying a threshold on the similarity matrices. Patients associated with a similarity higher than the threshold will be linked by an edge in the patient network. We tested different Cosine similarity thresholds and selected a threshold of 0.8. This threshold was chosen as the best trade-off to minimize the number of isolated patients while reducing the number of edges (*Supplementary section S2*). We obtained 11 patient networks (one by age, see Table 3 and Fig. 3 for the network of patients aged 60 years old).

We then applied the Markov Cluster algorithm (MCL) to identify clusters of patients (Material and methods 2.1.1). The MCL algorithm is applied systematically on all the 11 patient networks, revealing different numbers of clusters per network (Table 3). For example, in the patient network constructed at 60 years old, 127 clusters are identified (Fig. 3).

We next computed the number of common patients between clusters identified at consecutive ages (Material and methods 2.1.3). This allows tracking the evolution of the clusters over consecutive ages (Fig. 4) and identifying cluster-trajectories. We identified 12 cluster-trajectories composed of clusters with at least 100 patients (*Supplementary section S4*). We described the clusters that compose these trajectories with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses. Most of the 12 identified trajectories are composed of clusters with a majority of men. This is explained by the presence of a majority of men in our study population (i.e., 30,111 patients). Indeed, the sex ratio of this population is 0.61.

We next focused on the 3 cluster-trajectories (A,B and C) with the largest number of patients (Fig. 5 and *Supplementary section S4*). The trajectory A is the one with the largest number of patients. By analyzing clusters at all ages of this trajectory, we observed that all patients used aspirin. Furthermore, more than half of the patients present in any cluster of the trajectory A are also present in the following cluster. For instance, among the 4238 patients of the cluster 60.1 identified at age 60, 3 209 (i.e., 76 %) are present in the cluster 61.1 of age 61. Thus, for the majority of the patients, aspirin is used for at least two consecutive years. In addition, at 63 and 64 years old, two clusters are observed in the trajectory A. The first cluster (63.1 and 64.1) is associated with aspirin use only and the second cluster (63.14 and 64.11) is associated with enoxaparin use in addition to aspirin. These two clusters merge into the same cluster at the following age (64.1 and 65.1) in which only aspirin is used. This implies that, when enoxaparin is used in addition to aspirin, the majority of the patients switch to aspirin-only use the following year. The most frequent long-term illnesses observed in clusters that compose this trajectory is diabetes (ICD-10 code E11). This diagnosis is also observed in all the 12 trajectories identified. The other long-term illness observed in the trajectory A is chronic ischemic heart disease (ICD-10 code I25).

The trajectory B is composed of clusters in which clopidogrel, an antiplatelet drug, is used by all patients. Two clusters are systematically observed at each age. For example at age 60, in the first cluster 60.2, clopidogrel is the only drug used. In the second cluster 60.8, aspirin is used in addition to clopidogrel. These two clusters merge into the same cluster 61.2 at the following age in which clopidogrel is the only drug

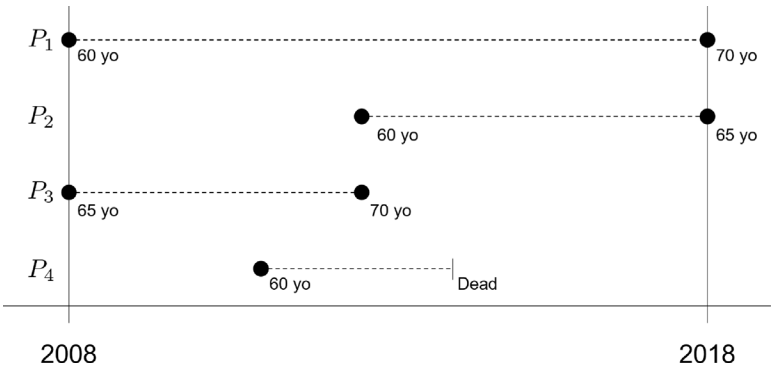


Fig. 2. Example of patient follow-up in the EGB.
P_1 has no truncated data because they were 60 years old in 2008 and therefore they have data for the entire period. P_2 has truncated data because they were 60 years old after 2008 and therefore they have no data before then. P_3 has truncated data because they were 70 years old before 2018 and therefore they have no data after that. P_4 has two types of truncated data because they were 60 years old after 2008 and died before 2018.

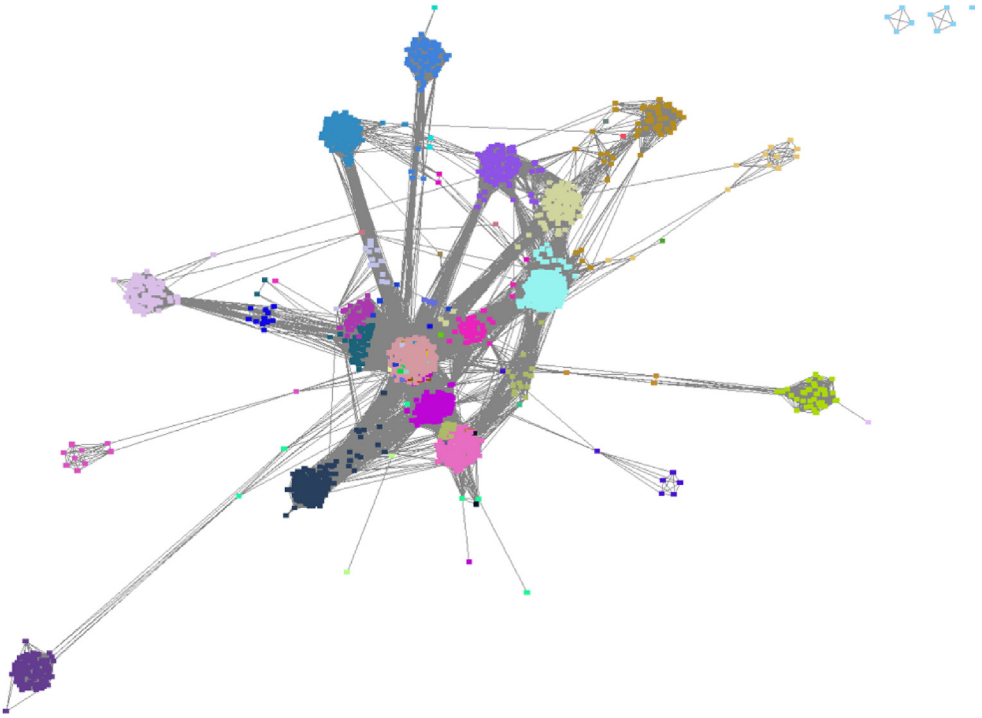


Fig. 3. 60-year-old patients network.
In this network, nodes represent all patients aged 60 between 2008 and 2018 and edges represent the interactions between those patients having a Cosine similarity of at least 0.8. The length of edges is inversely proportional to the Cosine similarity. Nodes of the same color belong to one of the 127 clusters identified with the Markov Cluster algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

used. Hence, we can observe that when aspirin is used in addition to clopidogrel, the majority of the patients switch to clopidogrel-only use the following year. The most frequent long-term illness in addition to diabetes is peripheral arterial disease (ICD-10 code I702).

The trajectory C is composed of clusters of patients who use fluindione; about 12 % of the patients also use enoxaparin. More than half of the patients present in any cluster of the trajectory C are also present in a cluster of the following year. For instance, among the 679 patients present in the cluster 60.3 identified at age 60, 503 (i.e., 74 %) are present in the cluster 61.3 identified at the age 61. Thus, we can conclude that, for the majority of the patients, fluindione is used for at least two consecutive years. The most frequent long-term illness in this trajectory is atrial fibrillation (ICD-10 code I48).

The same interpretations were carried out for the 9 remaining cluster-trajectories (*Supplementary section S4*). In each cluster that compose these trajectories, we always observe a drug used by all patients (i.e., predominant drug). Most of the time, more than half of the patients present in the clusters of these trajectories are also present in the following-age clusters. Thus, the predominant drugs are usually used for at least two consecutive years. However, this is not the case in the cluster-trajectory D. In this trajectory, two types of clusters are usually observed at each age. The first cluster contains patients who all used enoxaparin and the second cluster contains patients who all used tinzaparin. These two clusters systematically merge into the cluster 0 at the following age (e.g., cluster 61.0 at age 60). The cluster 0 is composed of patients with no antithrombotic use. Thus, the majority of patients with enoxaparin or tinzaparin use in this trajectory no longer use antithrombotics at the following year. This cluster-trajectory D is also the only one with clusters composed of a majority of women (i.e., sex ratio about 0.40). Associated comorbidities are scarce, with

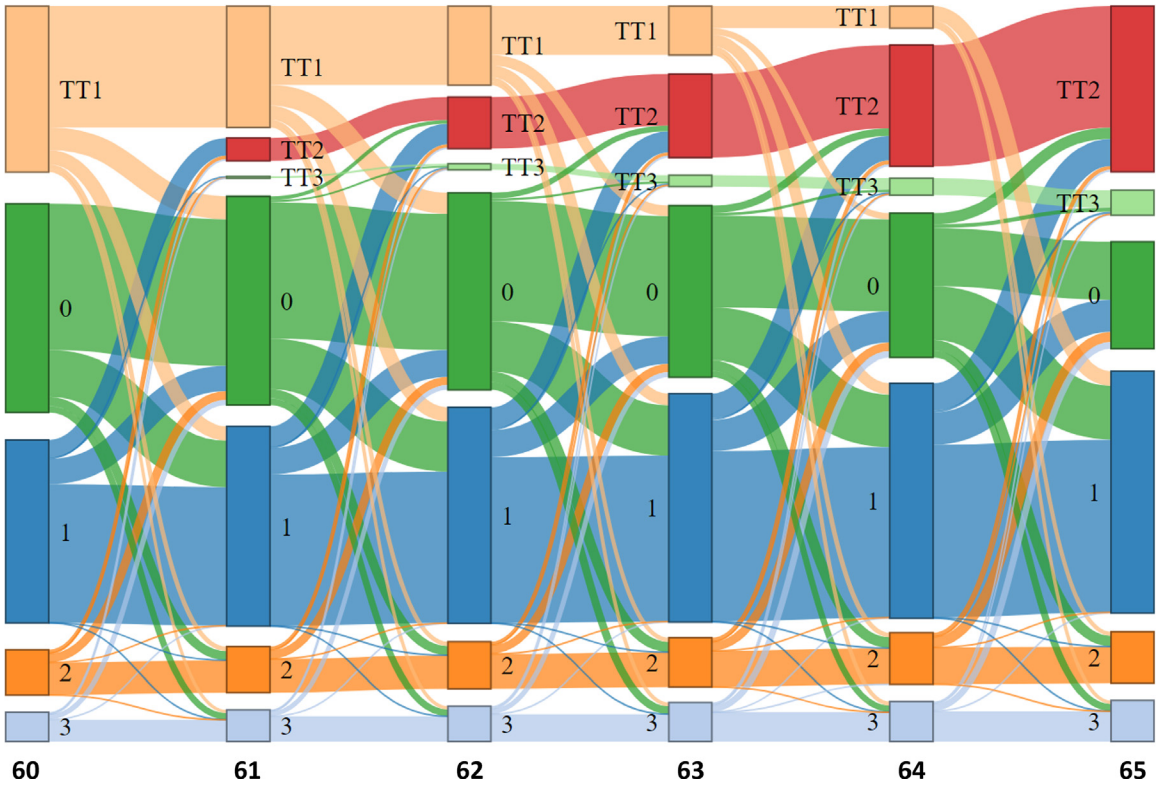


Fig. 4. Tracking of clusters identified from patient networks.
The alluvial plot represents the tracking of the clusters identified from 60 to 65 years old. The clusters were identified at each age based on patient networks with the MCL algorithm. At each age, the color blocks represent the different clusters of patients. Stream fields between blocks represent the number of common patients between clusters of consecutive ages. The height of the blocks and the thickness of the stream fields are proportional to the number of patients. At each age, only the clusters containing more than 500 patients are represented (corresponding to blocks 1 to 3). The blocks 0 correspond to the clusters of patients with no antithrombotic use. The three blocks TT1 (Truncated Type 1), TT2 (Truncated Type 2) and TT3 (Truncated Type 3) are the clusters of patients with truncated data. TT1 contains patients aged 70 before 2018; TT2 contains patients aged 60 after 2008 and TT3 contains patients who have died before 2018 (Fig. 2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Number of nodes, edges and clusters in 60 to 70 years old patient networks.

Age	Number of nodes	Number of edges (10^7)	Number of clusters
60	8268	1.25	127
61	8884	1.46	144
62	9555	1.70	162
63	10042	1.87	149
64	10466	2.03	168
65	10761	2.15	165
66	11097	2.27	150
67	11392	2.37	168
68	11492	2.43	207
69	11664	2.45	205
70	11687	2.48	220

the most frequent long-term illnesses being cancers (ICD-10 codes C50, C34, C18).

3.1.2. Identifying cluster-trajectories with the cluster-tracking approach based on raw data

In the previous Section 3.1.1, we identified cluster-trajectories using a network-based cluster-tracking approach. We also implemented a cluster-tracking approach using Kmeans applied to raw data (Material and methods 2.1.2). In this second strategy, we applied a Kmeans per patient age, from 60 to 70 years old.

In Kmeans, the number of clusters must be specified *a priori*. We calculated the silhouette score and identified an optimal number of clusters at each patient age (*Supplementary section S5*). The optimal number of clusters was between 6 and 8. We then tracked the clusters identified by Kmeans over ages (Material and methods 2.1.3). We identified 9 cluster-trajectories composed of clusters with at least 100 patients (*Supplementary section S6*). We described these trajectories with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses. We observed that all trajectories are composed of a majority of men. This is explained by the presence of a majority of men in our study population (i.e., 30,111 patients).

For the sake of simplicity, we next focused on three cluster-trajectories (A,B and C). We represented them from 60 to 65 years old (Fig. 6). The trajectory A is the one with the largest number of patients. Aspirin is used by all patients in the clusters that compose this trajectory. In all the clusters of the trajectory B, clopidogrel is used by all patients. In all the clusters of the trajectory C, fluindione is used by all patients and enoxaparin is used by about 12 % of patients. In addition, more than half of the patients present in any cluster of these three trajectories are also present in the following-age clusters. Thus, we can conclude that, for the majority of the patients, aspirin, clopidogrel and fluindione are used for at least two consecutive years in the trajectories A, B, and C, respectively. As in the network-based cluster-tracking approach, diabetes (ICD-10 code E11) is one of the most frequent long-term illnesses observed in clusters of all identified trajectories. The other long-term illness observed in the trajectory A is chronic ischemic heart disease (ICD-10 code I25). In trajectory B, the most frequent long-term illness in addition to diabetes in the clusters of age 60 (60.4) and 61 (61.3) is peripheral arterial disease (ICD-10 code I702). In the clusters identified from 62 years old, the most frequent long-term illness is chronic ischemic heart disease (ICD-10 code I25) In trajectory C, the most frequent long-term illness is atrial fibrillation (ICD-10 code I48).

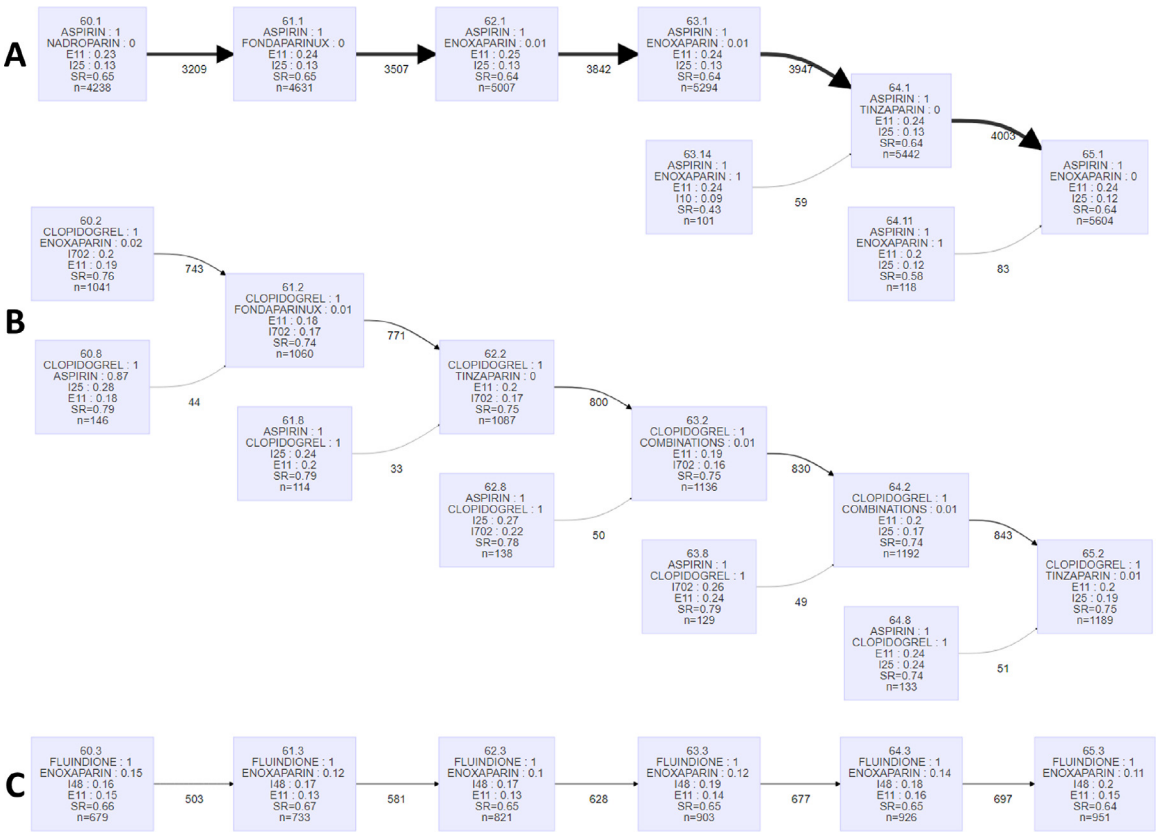


Fig. 5. Subset of patient cluster-trajectories identified with the cluster-tracking approach based on network.
We represented 3 cluster-trajectories (A,B and C) out of the 12 identified. We represented them from 60 to 65 years old. In these 3 cluster-trajectories, each block represents a cluster. Each cluster is named as follows: "x.y", with x the age at which it was identified and y its cluster number in the alluvial plot (Fig. 4). The clusters are characterized by the two most frequently reimbursed drugs (name, percentage of patients receiving the drug), the two most frequent long-term illnesses (ICD-10 code, percentage of patients), the sex ratio (SR) and the total number of patients (n). The number under arrows is the number of common patients between the two blocks. The arrow thickness is proportional to this number. Combinations: combination of two platelet aggregation inhibitors. ICD-10 code E11: type 2 diabetes mellitus, I25: chronic ischemic heart disease, I10: essential primary hypertension, I702: atherosclerosis of arteries of extremities, I48: atrial fibrillation.

These same descriptions were carried out for the 6 other cluster-trajectories (*Supplementary section S6*). In each cluster that compose these trajectories, we always observe a predominant drug used by all patients. Hence, we can conclude that the predominant drug is used for at least two consecutive years. This is not the case in the cluster-trajectories D and F. In the trajectory D, several clusters merge into the cluster 0 (e.g., cluster 61.0 at age 60), which is composed of patients with no antithrombotic use. Thus, most of the patients in this trajectory no longer use antithrombotics at the following year. Contrarily to what we previously observed in the network-based cluster-tracking approach, this trajectory D is not composed of a majority of women (i.e., sex ratio about 0.53). In the trajectory F, combinations (i.e., combination of two platelet aggregation inhibitors) are used to all patients in clusters identified from 61 to 67 years old. Then aspirin is used by about 60% of patients in clusters identified from 68 years old.

3.2. Comparing the two clustering strategies used in the cluster-tracking approaches

We identified the cluster-trajectories with the cluster-tracking approaches using two different clustering strategies: one based on the construction of patient networks by applying the MCL algorithm and one based on raw data by applying Kmeans. We aimed to compare the performances of these two clustering strategies.

We observed that the trajectories A in the two cluster-tracking approaches are composed of clusters having a similar description (*Supplementary sections S4 and S6*). Indeed, aspirin is used by all the patients and the two most frequent long-term illnesses are the same in all the clusters. We also observed a similar description between the clusters of the trajectories C and E of the two cluster-tracking approaches. The clusters of the two trajectories G also have a similar description, but the two trajectories do not begin at the same age. The first cluster is identified at 60 years old with the network-based cluster-tracking approach and at 64 years old with the raw-data-based cluster-tracking approach. The two trajectories H also begin at different ages. In both cases, the cluster-trajectories identified with the network-based cluster-tracking approach start at earlier ages than the cluster-trajectories identified with the raw-data-based cluster-tracking approach.

We calculated the modified silhouette score (S) and its 95% confidence interval to assess clustering quality in the two cluster-tracking approaches (Material and methods 2.3). We obtained $S = 0.50$ ([0.46 ; 0.55]) with the network-based cluster-tracking approach and $S = 0.57$ ([0.53 ; 0.58]) with the raw-data-based cluster-tracking approach (Table 4B.). *A priori*, the cluster-tracking approach seems to be more efficient using a raw-data-based than a network-based strategy. But we can observe that the confidence intervals of the modified silhouette scores obtained with the network-based and raw-based clustering approaches overlap.

3.3. Comparing the cluster-tracking approach with the longitudinal-clustering approaches

We compared the performance of the cluster-tracking approaches based on network and raw-data with three methods representative

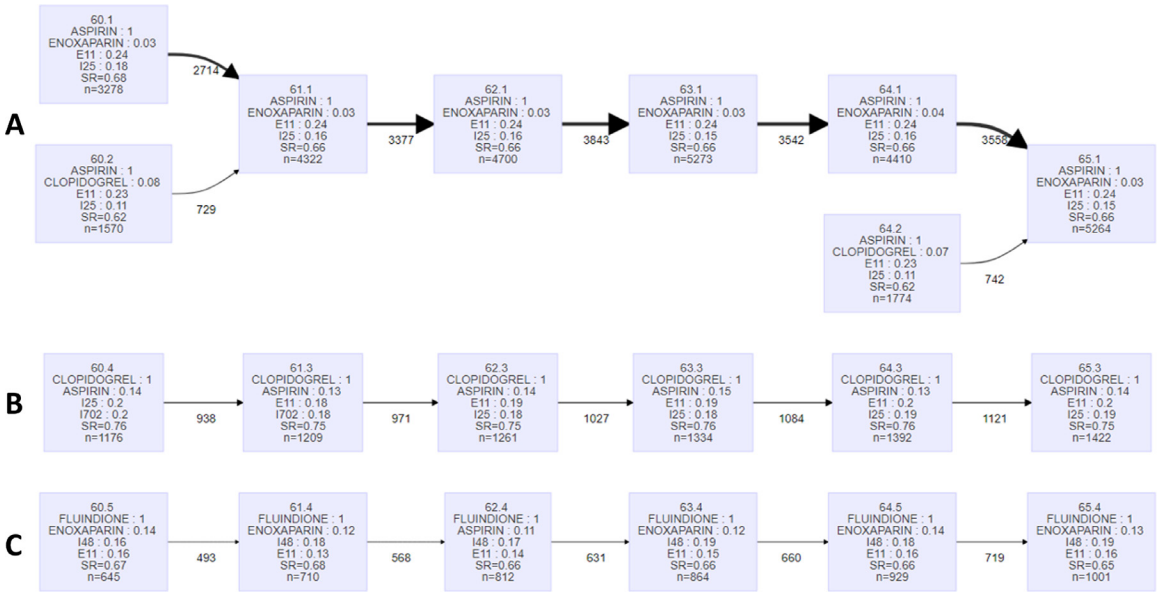


Fig. 6. Subset of patient cluster-trajectories identified with the raw-data-based cluster-tracking approach.
We represented 3 cluster-trajectories out (A,B and C) of the 9 identified. We represented them from 60 to 65 years old. In these 3 trajectories, each block represents a cluster. Each cluster is named as follows: "x.y", with x the age at which it was identified and y the number of the cluster. The clusters are characterized by the two most frequently reimbursed drugs (name, percentage of patients), the two most frequent long-term illnesses (ICD-10 code, percentage of patients), the sex ratio (SR) and the number of patients (n). The number under arrows is the number of patients in common between the two blocks. The arrow thickness is proportional to this number. ICD-10 code E11: type 2 diabetes mellitus, I25: chronic ischemic heart disease, I702: atherosclerosis of arteries of extremities, I48: atrial fibrillation.

of the three types of longitudinal clustering approaches, namely raw-data-based, feature-based and model-based approaches (Material and methods 2.2). We used the same longitudinal data extracted from EGB in patients aged from 60 to 70 years old in all the approaches.

3.3.1. *Choosing the optimal number of clusters*

In the three longitudinal-clustering approaches, the number of clusters need to be specified *a priori*. In order to select an optimal number of clusters, we calculated several classic clustering quality criteria (Material and methods 2.2.4). These criteria however do not point to clear optimums (*Supplementary section S3*). Hence, we next tried to use the modified silhouette score. We also failed to find a clear optimum with this approach. Indeed, the greatest silhouette scores (i.e., global maximum) was obtained for the smallest number of clusters (*Supplementary section S3*). We therefore decided to specify the number of clusters as 12 clusters. This number corresponds to the number of cluster-trajectories identified with the network-based cluster-tracking approach.

3.3.2. *Identifying clusters with the raw-data-based longitudinal-clustering approach*

We applied Kml3d [37], the selected raw-data-based longitudinal clustering approach (Material and methods 2.2.1) to the longitudinal data extracted from the EGB. First, 1737 patients are removed by the Kml3d algorithm because they have more than 9 truncated data (which is the limit with 11 different ages). We applied the Kml3d algorithm with 12 clusters as parameter and we described all the identified longitudinal-clusters with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses.

Among the 12 longitudinal-clusters identified by Kml3d, 10 are composed of at least 100 patients (Table 4A.). At least one of the two most frequently reimbursed drugs is used by more than 60 % of patients. For instance, aspirin, clopidogrel, combinations, warfarin and ticlopidine are used by all patients in longitudinal-clusters B, C, F, H and L, respectively. Each longitudinal-cluster identified is therefore characterized by a drug that is predominantly used by patients. More than 20 % of patients have diabetes (ICD-10 code E11) in all the longitudinal-clusters except in the longitudinal-cluster G. Atrial fibrillation (ICD-10 code I48) is one of the two most frequent long-term illnesses in longitudinal-clusters D, E, H, I and K. In these clusters, at least 70 % of patients use vitamin K antagonist (such as fluindione, warfarin or acenocoumarol) or non-vitamin K antagonist oral anticoagulants (such as rivaroxaban or apixaban). Chronic ischemic heart disease (ICD-10 code I25) is always observed in the longitudinal-clusters when aspirin is one of the two most frequently reimbursed drugs.

Our goal is to compare the 12 longitudinal-clusters obtained with the raw-data-based longitudinal clustering approach with the cluster-trajectories identified with the cluster-tracking approaches. At least one of the two most frequently reimbursed drugs is used by more than 60 % of patients in all the clusters that compose the cluster-trajectories (*Supplementary sections S4 and S6*) and in all the longitudinal-clusters (Table 4A.). This is not the case in the raw-data-based-cluster-trajectory D where aspirin is used by about 38 % of patients and enoxaparin is used by about 16 % of patients. Therefore, the majority of cluster-trajectories and longitudinal-clusters are characterized by a predominantly used drug. These trajectories and longitudinal-clusters are composed of a majority of men except in the network-based-cluster-trajectory D where the sex ratio is about 0.40. Breast cancer (ICD-10 code C50) is usually one of the two most frequent long-term illnesses in the clusters that compose the network-based-cluster-trajectory D. Several cluster-trajectories and longitudinal-clusters have a common drug description. For instance, aspirin and enoxaparin are both used in the longitudinal-cluster B and in the two cluster-trajectories A of the cluster-tracking approaches. The two most frequent long-term illnesses are also the same. Conversely, the raw-data-based longitudinal clustering approach is the only one to have identified three longitudinal-clusters characterized by use of ticagrelor-aspirin, prasugrel-aspirin and ticlopidine-aspirin (G, J and L respectively in Table 4A.). Similarly, the network-based cluster-tracking approach is the only one to have identified cluster-trajectories characterized by use of enoxaparin-tinzaparin, aspirin-fluindione and dabigatran-enoxaparin (D, J and L respectively in *Supplementary section S4*). Therefore, additional information are given with the raw-data-based longitudinal clustering

Table 4
Longitudinal-clusters identified with the three longitudinal clustering approaches and comparison with the cluster-tracking approaches.

A.

	Raw-data-based longitudinal-clustering				Feature-based longitudinal-clustering				Model-based longitudinal-clustering			
	n	SR	Top 2 drugs (%)	Top 2 diseases (%)	n	SR	Top 2 drugs (%)	Top 2 diseases (%)	n	SR	Top 2 drugs (%)	Top 2 diseases (%)
A	12550	0.53	Aspirin (65) Enoxaparin (22)	E11 (23) I25 (7)	11510	0.62	Aspirin (100) Enoxaparin (11)	E11 (32) I25 (17)	14461	0.65	Aspirin (76) Clopidogrel (27)	E11 (28) I25 (18)
B	8665	0.64	Aspirin (100) Enoxaparin (15)	E11 (32) I25 (21)	7484	0.51	Aspirin (41) Enoxaparin (28)	E11 (17) I48 (6)	6822	0.50	Aspirin (52) Enoxaparin (23)	E11 (19) I10 (6)
C	2937	0.75	Clopidogrel (100) Aspirin (60)	E11 (29) I25 (28)	2827	0.73	Clopidogrel (100) Aspirin (44)	E11 (27) I25 (24)	2481	0.77	Aspirin (92) Clopidogrel (66)	I25 (44) E11 (29)
D	1794	0.65	Fluindione (99) Enoxaparin (43)	I48 (24) E11 (22)	2460	0.62	Aspirin (100) Enoxaparin (20)	E11 (31) I25 (15)	2198	0.59	Aspirin (74) Enoxaparin (18)	E11 (29) I10 (10)
E	1013	0.65	Rivaroxaban (70) Aspirin (41)	I48 (37) E11 (23)	2050	0.60	Fluindione (100) Enoxaparin (38)	I48 (24) E11 (19)	2033	0.55	Aspirin (70) Fluindione (17)	E11 (25) I10 (8)
F	402	0.81	Combinations (100) Aspirin (79)	I25 (45) E11 (28)	1114	0.74	Clopidogrel (100) Aspirin (89)	I25 (41) E11 (30)	1296	0.61	Aspirin (82) Enoxaparin (22)	E11 (30) I25 (13)
G	345	0.77	Ticagrelor (98) Aspirin (97)	I25 (46) I21 (33)	615	0.77	Aspirin (100) Clopidogrel (100)	I25 (37) E11 (31)	803	0.58	Aspirin (78) Fluindione (18)	E11 (31) I10 (12)
H	243	0.62	Warfarin (100) Enoxaparin (46)	E11 (23) I48 (19)	576	0.62	Rivaroxaban (100) Aspirin (16)	I48 (32) E11 (16)	15	0.80	Aspirin (100) Clopidogrel (87)	I25 (67) E11 (13)
I	233	0.64	Acenocoumarol (99) Enoxaparin (42)	E11 (23) I48 (18)	509	0.80	Combinations (100) Aspirin (97)	I25 (51) E11 (33)	2	1.00	Aspirin (100) Fondaparinux (50)	C61 (50) K74 (50)
J	106	0.82	Prasugrel (96) Aspirin (93)	I25 (62) E11 (28)	454	0.70	Fluindione (100) Aspirin (96)	E11 (28) I48 (22)				
K	73	0.66	Apixaban (82) Aspirin (33)	I48 (30) E11 (18)	410	0.63	Rivaroxaban (100) Aspirin (60)	I48 (34) E11 (23)				
L	13	0.85	Ticlopidine (100) Aspirin (54)	E11 (23) C34 (15)	102	0.62	Warfarin (100) Aspirin (61)	E11 (26) I25 (17)				

B.

Network-based cluster-tracking	Raw-data-based cluster-tracking	Raw-data-based longitudinal-clustering	Feature-based longitudinal-clustering	Model-based longitudinal-clustering
0.50 [0.46 ; 0.55]	0.57 [0.53 ; 0.58]	0.27	0.20 [−0.02 ; 0.20]	−0.33 [−0.26 ; 0.01]

A. n: number of patients, SR: sex ratio (percentage of men), Top 2 drugs: the two most frequently reimbursed drugs with the percentage of patients, Top 2 diseases: the two most frequent long-term illnesses (ICD-10 code) with the percentage of patients. In all approaches, the identified longitudinal-clusters are ranked from the largest to the smallest. Combinations: combination of two platelet aggregation inhibitors. ICD-10 codes C34: malignant neoplasm of bronchus and lung, C50: malignant neoplasm of breast, C61: malignant neoplasm of prostate, E11: type 2 diabetes mellitus, I10: essential primary hypertension, I21: acute myocardial infarction, I25: chronic ischemic heart disease, I702: atherosclerosis of arteries of extremities, I48: atrial fibrillation, K74: fibrosis and cirrhosis of liver.
B. silhouette scores calculated in the different approaches and their 95% confidence intervals.

approach and the network-based cluster-tracking approach compared to the raw-data-based cluster-tracking approach.

Furthermore, we calculated the modified silhouette score (S) in the raw-data-based longitudinal clustering approach and in the cluster-tracking approaches to compare the clustering quality (Material and methods 2.3). We obtained $S = 0.27$ for the raw-data-based longitudinal clustering approach, $S = 0.50$ for the network-based cluster-tracking approach and $S = 0.57$ for the raw-data-based cluster-tracking approach (Table 4B.). The 95% confidence intervals of the two strategies of cluster-tracking approach overlap (Table 4B.). Overall, we obtained a better clustering quality with the cluster-tracking approaches compared to the raw-data-based longitudinal clustering approach.

3.3.3. *Identifying clusters with the feature-based longitudinal-clustering approach*

We extracted 4 standard features from the antithrombotic drug use contained in the EGB: the mean, the standard deviation, the kurtosis and the skewness (Material and methods 2.2.2). We therefore obtained a total of 76 features per patient (i.e., 4 features extracted over the 19 antithrombotic drugs). We then used these features as input in Kmeans. Here, the Kmeans clustering is applied over all the ages jointly. As for the raw-data-based longitudinal clustering approach, we applied the Kmeans clustering selecting 12 clusters as parameter. We described the identified longitudinal-clusters with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses.

The 12 longitudinal-clusters identified with the feature-based longitudinal clustering approach are all composed of at least 100 patients (Table 4A.). One of the two most used drugs is always used by all patients except in the cluster B. In this cluster, aspirin is used by 41 % of the patients and enoxaparin is used by 28 % of the patients. The majority of the identified longitudinal-clusters is therefore characterized by a predominantly used drug. At least 15 % of patients have diabetes (ICD-10 code E11) in all the clusters. Chronic ischemic heart disease (ICD-10 code I25) is always observed in the clusters where aspirin is one of the two most frequently reimbursed drugs.

We compared the 12 longitudinal-clusters obtained in the feature-based longitudinal clustering approach with the cluster-trajectories identified in the cluster-tracking approaches (*Supplementary sections S4 and S6*). We observe that the longitudinal-clusters A and D have a common drug and long-term illness description (Table 4A.). Indeed, aspirin and enoxaparin are both used by a similar proportion of patients and the two most frequent long-term illnesses are the same (i.e., ICD-10 codes E11 and I25). This type of redundant information is not observed in the cluster-trajectories identified with the two cluster-tracking approaches.

We then calculated the modified silhouette score (S) in the feature-based longitudinal clustering approach to compare the clustering quality with the other clustering approaches (Material and methods 2.3). We obtained $S = 0.20$ for the feature-based longitudinal clustering approach (Table 4B.). This score indicates that patients are less well assigned in clusters with the feature-based longitudinal clustering approach than with the cluster-tracking approach and with the raw-data-based longitudinal clustering approach. The clustering quality is therefore better with the cluster-tracking approaches.

3.3.4. *Identifying clusters with the model-based longitudinal-clustering approach*

The model-based approach that we applied to the antithrombotic drug use is GMM (Material and methods 2.2.3). We used an aggregated variable with this algorithm because the simultaneous analysis of several variables is computationally challenging [49]. This aggregated variable is calculated, for each patient, as the total number of drugs

used at a given age. As before, we applied GMM selecting 12 clusters as parameter. We described the identified longitudinal-clusters with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses.

The GMM algorithm assigns patients to the cluster for which they have the greatest posterior probability of belonging. Although we chose 12 clusters as parameter, none of the patients had a greatest posterior probability of belonging to three out of the 12 selected clusters. Therefore, only 9 longitudinal-clusters were identified.

The longitudinal-clusters A to G are composed of more than 100 patients (Table 4A.). The two remaining clusters are composed of less than 20 patients. In the 9 longitudinal-clusters, we observed that aspirin is used by more than 50 % of patients. All these longitudinal-clusters are therefore characterized by the same predominantly used drug. Diabetes (ICD-10 code E11) is always one of the two most frequent long-term illnesses except in longitudinal-cluster I. The longitudinal-cluster I is very small with only two patients. One of the patients has prostate cancer (ICD-10 code C61) and the other has fibrosis and cirrhosis of liver (ICD-10 code K74).

We compared the 9 longitudinal-clusters with the cluster-trajectories identified with the cluster-tracking approaches (*Supplementary ections S4 and S6)*). The longitudinal-clusters are highly different compared to the cluster-trajectories. Indeed, aspirin is used by a majority of patients in all these longitudinal-clusters, which is not the case in the cluster-trajectories. Furthermore, the diversity of the two most frequently reimbursed drugs is lower in the longitudinal-clusters since only aspirin, clopidogrel, enoxaparin, fluindione or fondaparinux are observed. In the cluster-trajectories, other drugs such as warfarin, combinations or rivaroxaban are additionally observed. The model-based longitudinal clustering approach therefore identified longitudinal-clusters where patients are more heterogeneous compared to the cluster-tracking approach.

We then calculated the modified silhouette score (S) in the model-based longitudinal clustering approach to compare the clustering quality with the other clustering approaches (Material and methods 2.3). We obtained $S = -0.33$ for the model-based longitudinal clustering approach (Table 4B.). This negative score indicates that the clusters are worse than random. The model-based longitudinal clustering approach therefore fails to identify patient clusters. Among all the analyzed approaches, the best clustering quality is obtained with the cluster-tracking approaches.

4. Discussion

We proposed here novel approaches based on cluster-tracking, with the objective of clustering patients using longitudinal data extracted from medico-administrative databases. We applied these new approaches to the analysis of antithrombotic drugs extracted from the Echantillon Généraliste des Bénéficiaires (EGB). We extracted the data from 2008 to 2018 and focused on patients aged from 60 to 70 years old. We aimed to identify clusters of patients that could be related to given diseases using only drug reimbursements and in the absence of any coded diagnoses. We showed that cluster-tracking approaches are efficient to identify patient trajectories from medico-administrative databases. They are able to consider the longitudinal, multidimensional and truncated nature of data. We were able to identify clusters of patients related to given diseases based only on drug reimbursements. We compared these new cluster-tracking approaches with three classical longitudinal clustering approaches using a modified silhouette score. We showed that the cluster-tracking approaches had a higher performance than the classical approaches.

We here applied all the approaches using age as time steps. However, it is to note that different data types can be used as input of our approaches. Depending on those input data, different time steps can be chosen. For example, we applied our two cluster-tracking approaches to the pbcseq database [50] using patient visits as time steps. This allowed us to cluster patients with primary sclerosing cholangitis based on their laboratory measurements (*Supplementary section S8*).

We identified 12 and 9 cluster-trajectories with the network-based and raw-data-based cluster-tracking approaches, respectively. We described the clusters that compose the cluster-trajectories with their number of patients, sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses. Of note, for both approaches, the top three largest cluster-trajectories had similar characteristics. The trajectories with the highest number of patients identified with the two cluster-tracking approaches (trajectories A) are composed of patients with aspirin use and chronic ischemic heart disease. Antithrombotic therapy is a key part of secondary prevention in patients with chronic ischemic heart disease and patients with this illness are considered for long-term aspirin treatment [51]. The trajectories B identified with the two cluster-tracking approaches are composed of patients with clopidogrel use and coded arteriopathies as long-term illnesses (i.e., peripheral arterial disease and chronic ischemic heart disease). This is in accordance with clopidogrel being the preferred antiplatelet drug indicated in patients with arteriopathies that are symptomatic or have undergone revascularization [52]. The trajectory B identified with the network-based cluster-tracking approach also shows patients using aspirin with clopidogrel and switching to the use of clopidogrel-only the following year. This is in accordance with the fact that after myocardial infarction and percutaneous coronary intervention, a switch to mono-therapy is recommended after one year of dual antiplatelet [53]. The two trajectories C, the third largest trajectories identified with the two cluster-tracking approaches, are composed of patients with fluindione use and coded atrial fibrillation. Fluindione, which is a vitamin K antagonist, has been shown to strongly reduce stroke in patients with atrial fibrillation [54]. Furthermore, in the trajectory C identified by the network-based cluster-tracking approach, we observed a switch of drugs from age 67. Recently, non-vitamin K antagonist oral anticoagulants (e.g., apixaban and rivaroxaban) have been recommended in replacement of vitamin K antagonists [55]. Because non-vitamin K antagonist oral anticoagulants are more convenient to use, the switch of drugs observed from age 67 with the network-based cluster-tracking approach is consistent. The two identified trajectories D are composed of patients using low molecular weight heparin (i.e., enoxaparin or tinzaparin) over a short period of time. Indeed, these patients do not use antithrombotics the following year. We hypothesize that these trajectories captured patients having an acute venous thromboembolism event. However, the trajectory D identified with the network-based cluster-tracking approach was the only one composed of patients who were mostly women with cancers. There is a known significant increase of thromboembolism event requiring low molecular weight heparin in these patients [56]. Moreover, it is well-known that women have a higher risk of thromboembolism event than men [57]. The trajectory F identified with the raw-data-based cluster-tracking approach was the only one with patients first using combination of two antithrombotic drugs and then aspirin at older age. As hemorrhage risk increases with age, patients at older age switch to only one platelet aggregation inhibitor [53]. As a side note, regarding long-term illnesses, diabetes was among the two most frequent long-term illnesses in all the cluster-trajectories. No specific antithrombotic drugs are recommended for patients suffering from diabetes. However, diabetes increases cardiovascular risk and therefore many patients with antithrombotic drugs have diabetes [58].

We compared these new cluster-tracking approaches with three classical longitudinal clustering approaches. The better modified silhouette score was obtained with the cluster-tracking approaches. This higher performance might arise from a better usage of the available information. Indeed, clustering per age allows us to take into account a maximum number of patients: as the clustering is performed by age, patient follow-ups over the entire period are not required and missing data can be handled. Contrarily, classical longitudinal clustering approaches require patient follow-ups over the entire period. Longitudinal

clustering approaches hence either impute data or exclude patients with truncated data. Our new cluster-tracking approaches are therefore less sensitive to small sample sizes. However, large sample sizes increase computation time for all the approaches (supplementary Table S1). Another interesting feature of the cluster-tracking approaches is that patients can switch clusters as their age progresses. A patient can therefore belong to several cluster-trajectories. This allows considering some uncertainty in patient clustering compared to the longitudinal-clustering approaches where a patient belongs to a single longitudinal-cluster.

The modified silhouette score also showed comparable performances between the two cluster-tracking approaches. However, it is to note that the network-based cluster-tracking approach does not require the number of clusters to be defined *a priori*. This is an advantage as the number of clusters might be a parameter difficult to set-up. In addition, the network-based cluster-tracking approach has also the advantage of preserving privacy because the interactions between patients are considered rather than absolute data. Another advantage is the flexibility of this approach, as many different measures can be used to compute the similarity between patients. These similarity measures can then be tuned depending on the data and question at hand. Moreover, a large number of algorithms exist for clustering networks.

Code availability

The code for our two cluster-tracking approaches is available on GitHub (https://github.com/JudithLamb/Cluster-tracking). For privacy reasons, antithrombotic drug reimbursements extracted from the EGB cannot be shared publicly. We hence generated a simulated dataset of 5594 patients with their drug use from these extracted data. The results obtained from this simulated sample dataset can be visualized in an R Shiny app also available from the GitHub repository.

Funding

This work was supported by the Inserm cross-cutting program Genomic variability 2018 GOLD.

Acknowledgements

The authors would like to acknowledge Pierre Sabatier for extracting and formatting the data. The authors would also like to thank Anthony Baptista for his contribution in the Methods section. We would like to thank David Hirst, Céline Chevalier, Morgane Terezol and Ozan Ozisik for their many comments after proofreading the article. And finally, we would like to thank all the members of MMG and Heka teams for their feedback.

CRediT authorship contribution statement

Judith Lambert: Writing – original draft, Visualization, Conceptualization, Methodology. **Anne-Louise Leutenegger:** Supervision, Conceptualization, Methodology, Writing – review & editing. **Anne-Sophie Jannot:** Supervision, Conceptualization, Methodology, Writing – review & editing. **Anaïs Baudot:** Supervision, Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104309.

It contains the method to choose the similarity measure and the threshold for the construction of patient networks; the assessment of the optimal number of clusters in the raw-data-based cluster-tracking approach and the longitudinal clustering approaches; the complete figures of all the cluster-trajectories identified with the cluster-tracking approaches; the computation time and the time complexity of the different approaches; the application of the cluster-tracking approaches to pbcseq database.

References

[1] Cristina Mazzali, Piergiorgio Duca, Use of administrative data in healthcare research, Intern. Emerg. Med. 10 (4) (2015) 517–524.
[2] Ivo D. Dinov, Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data, Gigascience 5 (1) (2016).
[3] Sula Windgassen, et al., The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome, J. Ment. Health 27 (2) (2018) 94–96.
[4] Anna Okula Basile, Marylyn DeRiggi Ritchie, Informatics and machine learning to define the phenotype, Expert Rev. Mol. Diagn. 18 (3) (2018) 219–226.
[5] T. Warren Liao, Clustering of time series data—a survey, Pattern Recognit. 38 (11) (2005) 1857–1874.
[6] Jean-Baptiste Pingault, et al., Childhood trajectories of inattention and hyperactivity and prediction of educational attainment in early adulthood: a 16-year longitudinal population-based study, Am. J. Psychiatry 168 (11) (2011) 1164–1170.
[7] Adeline Divoux, et al., Fibrosis in human adipose tissue: composition, distribution, and link with lipid metabolism and fat mass loss, Diabetes 59 (11) (2010) 2817–2825.
[8] Xiaozhe Wang, Kate Smith, Rob Hyndman, Characteristic-based clustering for time series data, Data Min. Knowl. Discov. 13 (3) (2006) 335–364.
[9] Daniel S. Nagin, Candice L. Odgers, Group-based trajectory modeling in clinical research, Annu. Rev. Clin. Psychol. 6 (2010) 109–138.
[10] Moritz Herle, et al., Identifying typical trajectories in longitudinal data: modelling strategies and interpretations, Eur. J. Epidemiol. 35 (3) (2020) 205–222.
[11] Pablo A Mora, et al., Distinct trajectories of perinatal depressive symptomatology: evidence from growth mixture modeling, Am. J. Epidemiol. 169 (1) (2009) 24–32.
[12] Craig R Colder, et al., Identifying trajectories of adolescent smoking: an application of latent growth mixture modeling., Health Psychol. 20 (2) (2001) 127.
[13] Aron S Downie, et al., Trajectories of acute low back pain: a latent class growth analysis, Pain 157 (1) (2016) 225–234.
[14] Rebecca J Landa, et al., Latent class analysis of early developmental trajectory in baby siblings of children with autism, J. Child Psychol. Psychiatry 53 (9) (2012) 986–996.
[15] Lucas Vendramin, Ricardo J.G.B. Campello, Eduardo R. Hruschka, Relative clustering validity criteria: A comparative overview, Stat. Anal. Data Min.: ASA Data Sci. J. 3 (4) (2010) 209–235.
[16] Steven J Van Laere, et al., Uncovering the molecular secrets of inflammatory breast cancer biology: an integrated analysis of three distinct affymetrix gene expression datasets, Clin. Cancer Res. 19 (17) (2013) 4685–4696.
[17] Lovisa Lovmar, Annika Ahlford, Mats Jonsson, Ann-Christine Syvänen, Silhouette scores for assessment of SNP genotype clusters, BMC Genomics 6 (1) (2005) 1–6.
[18] Victor M Vergara, et al., Determining the number of states in dynamic functional connectivity using cluster validity indexes, J. Neurosci. Methods 337 (2020) 108651.
[19] Jordi A Matias-Guiu, et al., Clustering analysis of FDG-PET imaging in primary progressive aphasia, Front. Aging Neurosci. 10 (2018) 230.
[20] Yanchi Liu, et al., Understanding and enhancement of internal clustering validation measures, IEEE trans. cybern. 43 (3) (2013) 982–994.
[21] Zuyun Liu, et al., Joint trajectories of cognition and frailty and associated burden of patient-reported outcomes, J. Am. Med. Dir. Assoc. 19 (4) (2018) 304–309.
[22] Tracy Vaillancourt, John D. Haltigan, Joint trajectories of depression and perfectionism across adolescence and childhood risk factors, Dev. Psychopathol. 30 (2) (2018) 461–477.
[23] Mitzi M Gonzales, et al., Joint trajectories of cognition and gait speed in Mexican American and European American older adults: The San Antonio longitudinal study of aging, Int. J. Geriatr. Psychiatry 35 (8) (2020) 897–906.
[24] William Fung, et al., Joint trajectories of disease activity, and physical and mental health-related quality of life in an inception lupus cohort, Rheumatology 59 (10) (2020) 3032–3041.

[25] Narimene Dakiche, et al., Tracking community evolution in social networks: A survey, Inf. Process. Manage. 56 (3) (2019) 1084–1102.
[26] Derek Greene, Donal Doyle, Padraig Cunningham, Tracking the evolution of communities in dynamic social networks, in: 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2010, pp. 176–183.
[27] Yang Sun, et al., Matrix based community evolution events detection in online social networks, in: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom, SmartCity, IEEE, 2015, pp. 465–470.
[28] Li Li, et al., Identification of type 2 diabetes subgroups through topological analysis of patient similarity, Sci. Transl. Med. 7 (311) (2015) 311ra174.
[29] Bo Wang, et al., Similarity network fusion for aggregating data types on a genomic scale, Nature Methods 11 (3) (2014) 333.
[30] Shraddha Pai, Gary D. Bader, Patient similarity networks for precision medicine, J. Mol. Biol. 430 (18) (2018) 2924–2938.
[31] Sarvenaz Choobdar, et al., Assessment of network module identification across complex diseases, Nature Methods 16 (9) (2019) 843–852.
[32] Santo Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.
[33] Stijn vanDongen, A cluster algorithm for graphs, Inf. Syst. [INS] (R 0010) (2000).
[34] J. MacQueen, Classification and analysis of multivariate observations, in: 5th Berkeley Symp. Math. Statist. Probability, 1967, pp. 281–297.
[35] Peter J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. (1987) 53–65.
[36] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, Teh Ying Wah, Time-series clustering–a decade review, Inf. Syst. 53 (2015) 16–38.
[37] Christophe Genolini, et al., kml and kml3d: R packages to cluster longitudinal data, J. Stat. Softw. 65 (4) (2015) 1–34.
[38] Christophe Genolini, Hélène Jacqmin-Gadda, et al., Copy mean: a new method to impute intermittent missing values in longitudinal studies, Open J. Stat. 3 (04) (2013) 26.
[39] Alex Nanopoulos, Rob Alcock, Yannis Manolopoulos, Feature-based classification of time-series data, Int. J. Comput. Res. 10 (3) (2001) 49–61.
[40] Tadeusz Caliński, Jerzy Harabasz, A dendrite method for cluster analysis, Comm. Statist. Theory Methods 3 (1) (1974) 1–27.
[41] Krzysztof Kryszczuk, Paul Hurley, Estimation of the number of clusters using multiple clustering validity indices, in: International Workshop on Multiple Classifier Systems, Springer, 2010, pp. 114–123.
[42] Siddheswar Ray, Rose H. Turi, Determination of number of clusters in k-means clustering and application in colour image segmentation, in: Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, Citeseer, 1999, pp. 137–143.
[43] David L. Davies, Donald W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. (2) (1979) 224–227.
[44] Hirotogu Akaike, Information theory and an extension of the maximum likelihood principle, in: Selected Papers of Hirotugu Akaike, Springer, 1998, pp. 199–213.
[45] Gideon Schwarz, Estimating the dimension of a model, Ann. Statist. (1978) 461–464.
[46] Ron Wehrens, Hein Putter, Lutgarde M.C. Buydens, The bootstrap: a tutorial, Chemometr. Intell. Lab. Syst. 54 (1) (2000) 35–52.
[47] P1 Tuppin, et al., French national health insurance information system and the permanent beneficiaries sample, Rev. Epidemiol. Sante Publique 58 (4) (2010) 286–290.
[48] Armin Skrbo, Begler Begović, Selma Skrbo, Classification of drugs using the ATC system (anatomic, therapeutic, chemical classification) and the latest changes, Med. Arh. 58 (1 Suppl 2) (2004) 138—141.
[49] Jin Liu, Robert A. Perera, Extending growth mixture model to assess heterogeneity in joint development with piecewise linear trajectories in the framework of individual measurement occasions, 2020, arXiv preprint arXiv:2010.13325.
[50] Thomas R. Fleming, David P. Harrington, Counting Processes and Survival Analysis, John Wiley & Sons, 1991.
[51] Juhani Knuuti, et al., 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes: The task force for the diagnosis and management of chronic coronary syndromes of the European society of cardiology (ESC), Eur. Heart J. 41 (3) (2020) 407–477.
[52] Victor Aboyans, et al., ESC scientific document group. 2017 ESC guidelines on the diagnosis and treatment of peripheral arterial diseases, in collaboration with the European society for vascular surgery (ESVS): Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteriesendorsed by: the European stroke organization (ESO) the task force for the diagnosis and treatment of peripheral arterial diseases of the European society of cardiology (ESC) and of the European society for vascular surgery (ESVS), Eur. Heart J. 39 (9) (2018) 763–816.
[53] Marco Valgimigli, et al., 2017 ESC focused update on dual antiplatelet therapy in coronary artery disease developed in collaboration with EACTS: The task force for dual antiplatelet therapy in coronary artery disease of the European society of cardiology (ESC) and of the European association for Cardio-Thoracic surgery (EACTS), Eur. Heart J. 39 (3) (2018) 213–260.
[54] Robert G. Hart, Lesly A. Pearce, Maria I. Aguilar, Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation, Ann. Intern. Med. 146 (12) (2007) 857–867.
[55] Gerhard Hindricks, et al., 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European association for cardio-thoracic surgery (EACTS) the task force for the diagnosis and management of atrial fibrillation of the European society of cardiology (ESC) developed with the special contribution of the European heart rhythm association (EHRA) of the ESC, Eur. Heart J. 42 (5) (2021) 373–498.
[56] Deirdre P Cronin-Fenton, et al., Hospitalisation for venous thromboembolism in cancer patients and the general population: a population-based cohort study in Denmark, 1997–2006, Br. J. Cancer 103 (7) (2010) 947–953.
[57] Emmanuel Oger, EPI-GET.B.O. study group, et al., Incidence of venous thromboembolism: a community-based study in Western France, Thromb. Haemost. 83 (05) (2000) 657–660.
[58] Karine Chevreul, Karen Berg Brigham, Clara Bouché, The burden and treatment of diabetes in France, Glob. Health 10 (1) (2014) 1–9.

2.3 Conclusion et discussion

Le "cluster-tracking" est une approche novatrice qui vise à identifier des sous-groupes homogènes de patients à partir de variables longitudinales. Cette approche repose sur deux étapes essentielles : l'identification des clusters à chaque instant temporel, puis leur suivi au cours du temps en vue de déterminer des trajectoires de clusters. Nous avons opté pour deux stratégies distinctes pour l'identification des clusters. La première stratégie consiste à appliquer directement un K-means sur les données brutes à chaque instant temporel. Nous avons appliqué le K-means dans cette première stratégie car c'est la méthode de clustering la plus communément utilisée. La deuxième stratégie consiste à construire au préalable des réseaux de patients. Cette seconde stratégie est ainsi une méthode moins classique reposant sur l'utilisation des réseaux. L'utilisation de réseaux permet de remédier à certaines des limitations du K-means, telles que le choix du nombre de clusters.

Nous avons testé cette approche du "cluster-tracking" dans le but de stratifier des patients âgés de 60 à 70 ans ayant reçus des remboursements de médicaments antithrombotiques entre 2008 et 2018. Ces données ont été extraites de la base de données médico-administrative EGB. Nous avons pu identifier plusieurs trajectoires de clusters regroupant des patients similaires et cliniquement significatifs. Cette interprétation a été facilitée grâce à la visualisation des trajectoires de clusters au moyen d'un diagramme composé de blocs. Chaque bloc a servi à représenter un cluster de la trajectoire en fournissant une description de ses caractéristiques, notamment les deux médicaments les plus fréquemment remboursés (évaluation interne) et les deux maladies les plus observées chez les patients (évaluation externe). Cette représentation a révélé que chaque trajectoire de cluster était associée à un médicament majoritairement prescrit et que la recommandation initiale de ce médicament correspondait à la maladie majoritairement identifiée chez les patients. En complément des trajectoires, l'évolution des clusters au cours du temps a été visualisée à l'aide d'un *alluvial plot*. Dans ce graphique, les clusters sont représentés par des blocs. Chaque cluster identifié à un âge donné est relié aux clusters suivant par des branches représentant le nombre de patients en commun. Cette visualisation permet d'avoir un aperçu de la répartition des patients d'un âge à l'autre dans le clustering. En fin de compte, ces approches visuelles offrent une représentation instantanée et claire des résultats, simplifiant ainsi leur interprétation.

Dans le but d'évaluer la performance de cette approche de "cluster-tracking", nous l'avons comparée à trois méthodes classiques de clustering longitudinal, chacune appartenant à l'une des catégories majeures abordées précédemment dans la section 2.1. En utilisant le coefficient

silhouette, une mesure de qualité du clustering interne (voir section 1.2.5), nous avons pu démontrer que le "cluster-tracking" obtenait de meilleures performances que les trois autres méthodes.

Le "cluster-tracking" représente une solution efficace pour aborder les défis auxquels sont confrontées les méthodes usuelles de clustering longitudinal. Dans un premier temps, cette approche permet de gérer les données tronquées sans nécessiter d'imputations ou d'exclusions de patients. Cette gestion est rendue possible par la segmentation par temps de l'identification des sous-groupes de patients. Contrairement aux méthodes classiques, cette segmentation ne requiert plus un suivi continu des patients, mais permet plutôt d'utiliser les données des patients uniquement aux moments où elles sont disponibles. De plus, cette segmentation permet de réduire la quantité de données à traiter en même temps, ce qui se traduit par des gains de temps et de stockage, et facilite l'analyse simultanée de variables longitudinales. Une autre caractéristique notable du "cluster-tracking", et spécifique à la deuxième stratégie utilisant les réseaux, réside dans le fait qu'il n'est pas nécessaire de spécifier le nombre de clusters au préalable. Ceci représente un avantage important compte tenu de la difficulté de déterminer ce nombre de manière optimale. En effet, dans la première stratégie qui consistait à utiliser le K-means, nous avons dû recourir au coefficient silhouette pour identifier le nombre de clusters optimal à chaque instant temporel. Cela a impliqué de tester plusieurs valeurs de clusters, allant de 2 à 200. Cette procédure a contribué à augmenter le temps et la complexité de l'analyse. De plus, en fonction de la mesure utilisée pour déterminer le nombre de clusters, les résultats obtenus peuvent varier.

Tous ces avantages offerts par le "cluster-tracking" le positionne comme une alternative intéressante pour aborder les problématiques complexes du clustering longitudinal de données de santé. Cependant, plusieurs limites sont à considérer dans cette approche. Tout d'abord, lors du processus de construction des réseaux de patients, plusieurs paramètres requièrent une attention particulière. Le premier de ces paramètres est la mesure de similarité entre patients qui doit être choisie en fonction du type de données analysé. En effet, dans notre étude, nous avons dû évaluer plusieurs mesures afin de sélectionner celle qui convenait le mieux à nos données. Ces mesures comprenaient la similarité Cosinus, la distance Euclidienne normalisée, la distance de Jaccard et la distance de Jaccard généralisée. En calculant la variance de chaque distribution, nous avons mis en évidence que la similarité Cosinus était la mesure qui distinguait le mieux les patients similaires des patients dissimilaires. Un autre paramètre à prendre en considération dans la construction des réseaux de patients est le seuil appliqué dans la matrice de similarité. Toutes les paires de patients dont la similarité dépasse le seuil

choisi seront connectés dans le réseau. Ce seuil est généralement choisi pour conserver dans le réseau uniquement les patients présentant une forte similarité, ce qui permet de réduire le bruit au sein des données. La règle de décision que nous avons adoptée pour choisir ce seuil n'est pas universelle et doit être ajustée en fonction des caractéristiques spécifiques des données, ce qui peut impacter la structure finale des réseaux. En ce qui concerne l'identification des clusters, les paramètres inhérents des algorithmes de clustering utilisés, à savoir K-means et MCL, sont à prendre en compte. Dans notre étude, nous avons utilisé les valeurs par défaut des paramètres, mais il est important de noter que leur modification peut influencer les résultats du clustering. Pour la construction des trajectoires des clusters, le paramètre crucial est la règle de décision relative au regroupement des clusters au sein de ces trajectoires. Pour notre étude, la règle que nous avons décidé de choisir pour la construction des trajectoires repose sur l'identification des clusters qui partagent le plus de patients avec les clusters suivants. Cependant, cette règle de décision pourrait être plus flexible en utilisant un seuil au lieu du nombre maximal de patients ou totalement différente, en se basant par exemple sur le nombre de remboursements de médicaments en commun. Changer la règle de décision pourrait induire des modifications sur la composition des trajectoires. Il en résulte que certains éléments de notre méthode doivent être adaptés pour l'utiliser sur d'autres types de variables.

Une extension de ce travail pourrait être de réaliser le clustering de manière globale. En effet, un réseau temporel constitué de couches successives, chacune représentant un réseau à un instant temporel donné, pourrait être envisagé. Ces couches seraient interconnectées en reliant les mêmes noeuds et ne seraient donc pas traitées comme des réseaux individuels, contrairement au "cluster-tracking". A partir de ces réseaux temporels, l'objectif serait d'identifier les clusters en utilisant un algorithme de clustering qui prendrait en compte à la fois les liens inter-couches et les liens intra-couches. Cependant, il n'existe pour le moment pas d'algorithme qui réalise cette tâche. La réalisation de cet algorithme a été initiée en 2021 par Louise Labatie qui a effectué un stage de six mois au sein de notre équipe. Elle a pour cela initié la création d'une fonction Python appelée TGCC (*Temporal Global Clustering Consensus*), conçue pour identifier des clusters à travers toutes les couches du réseau temporel. Elle a développé cette fonction en modifiant l'approche de clustering implémentée dans le package Python Teneto qui contient un ensemble d'outils destinés à l'analyse des données de réseaux temporels [106]. Afin de tester sa méthode, elle a entrepris un travail approfondi consistant à simuler plusieurs réseaux pour identifier les paramètres optimaux nécessaires à son fonctionnement. Cependant, les réseaux simulés étaient de petites tailles, ce qui signifie que l'application de cette méthode à des réseaux plus vastes nécessiterait l'identification de nouveaux paramètres

optimaux. Une fois cette étape réalisée, cette approche de clustering pourrait ainsi représenter une alternative au cluster-tracking.

En conclusion, dans cette première étude, nous avons développé une approche de clustering de patients à partir des remboursements de médicaments. Cette approche permet à partir des données brutes d'utiliser l'ensemble des patients disponibles quelle que soit la période pendant laquelle ils ont été suivis, ce qui en fait une méthode particulièrement adaptée aux données médico-administratives. Cette approche offre l'avantage d'être particulièrement interprétable. L'analyse fine des résultats obtenus sur notre cas d'usage nous a conduit à identifier un autre challenge dans le traitement de ces données, à savoir que deux patients ayant des remboursements de médicaments avec des caractéristiques communes (par exemple, principe actif proche) seront considérés de la même manière que deux patients dont les médicaments remboursés différent. Cela conduit à une perte d'information dans la construction des clusters à chaque pas de temps. C'est ce challenge que nous allons aborder dans le chapitre 3.

Chapitre 3

Amélioration de la qualité du clustering en considérant les relations entre les labels des variables dans le calcul de la similarité entre patients

Dans le chapitre précédent, j'ai présenté une nouvelle méthode, appelée "cluster-tracking", qui a permis de prendre en compte la nature longitudinale des données contenues dans les bases médico-administratives pour identifier des clusters de patients. Nous avons pu mettre en évidence des trajectoires de clusters pertinentes et cliniquement significatives. Ce travail s'est inscrit dans le premier objectif de ma thèse. Le deuxième objectif a été de prendre en compte la nature nomenclaturale des labels associés aux données de santé médico-administratives. En effet, certaines données de santé ont des labels organisés en nomenclature. Les labels sont ainsi reliés de manière plus ou moins étroite entre eux dans la nomenclature. Afin de prendre en

compte cette notion de relation entre labels, nous l'avons intégrée dans les mesures de similarités entre patients. Nous avons cherché à évaluer si les méthodes de clustering de patients présentées dans le chapitre précédent pouvaient être améliorées en intégrant ces relations entre labels.

3.1 Mesures prenant en compte les relations entre labels

3.1.1 Les valeurs et labels des variables de santé

Les variables sont en général caractérisées par des valeurs et des labels. La valeur d'une variable fait référence à la mesure spécifique que prend cette variable. Cette mesure peut être quantitative, catégorielle ou textuelle, comme nous l'avons mentionné en introduction. Le label d'une variable, quant à lui, fait référence au nom ou à la description attribuée à cette variable afin de l'identifier. Par exemple, dans le contexte des remboursements de médicaments, la valeur peut représenter le nombre de remboursements et le label le nom du médicament en question. Certaines variables de santé contiennent des labels organisés en nomenclature. Les nomenclatures sont des systèmes d'organisation spécifiques dans lesquels les labels sont interconnectés par des relations plus ou moins étroites dans les nomenclatures. Par exemple, au sein de la nomenclature de l'ATC utilisée pour classifier les médicaments, ceux qui partagent des compositions ou des usages thérapeutiques similaires sont regroupés dans la même classe. Ainsi, ces médicaments sont plus proches au sein de cette nomenclature. La CIM-10, quant à elle, permet de regrouper dans une même classe les maladies partageant des caractéristiques communes. D'autres exemples de nomenclatures sont également abordés dans la section 1.1.3. Les relations entre labels au sein de ces nomenclatures pourraient être une information capitale à exploiter dans les approches de clustering. En effet, deux patients qui prennent des médicaments anti-diabétiques devraient être considérés comme plus similaires que deux patients prenant des médicaments de types différents. De même, deux patients atteints chacun d'une maladie endocrinienne seraient plus similaires que deux patients atteints chacun d'une maladie de classe différente. Cependant, la plupart des approches de clustering se basent uniquement sur les valeurs des variables pour calculer la similarité entre les patients. Intégrer la notion de relations entre labels dans le clustering apporterait une perspective clinique plus approfondie et pertinente. Cela permettrait de mieux saisir les liens entre les caractéristiques de santé des patients et ainsi de définir des groupes plus cohérents et cliniquement significatifs.

3.1.2 Les mesures existantes pour analyser les relations entre labels

De nombreuses mesures ont été développées pour analyser les relations entre les labels des variables [107]. Parmi ces mesures, celles de Wu et Palmer, de Resnik et de Lin sont les plus couramment employées. Parallèlement, d'autres mesures telles que la mesure de Jiang et Conrath ou la mesure de Leacock-Chodorow existent. Afin d'introduire ces différentes mesures, nous considérons les labels d'un ensemble de variables I organisés selon une nomenclature dont la structure est un arbre. Le sommet de l'arbre est composé d'un label unique représentant la racine, tandis que les labels se trouvant au niveau le plus bas de l'arbre sont les feuilles. Entre la racine et les feuilles de l'arbre, les labels sont connectées à des labels parents (qui les précèdent) et des labels enfants (qui les suivent) par des liens qui illustrent les relations existantes entre les différents labels. Ces séquences de liens définissent des chemins au sein de l'arbre.

La mesure de Wu et Palmer

La mesure de Wu et Palmer analyse les relations entre les labels des variables en prenant en compte la structure de l'arbre de nomenclature des labels [108]. Cette mesure est calculée à l'aide de la formule suivante :

$$WP(x,y) = \frac{2 \times Profondeur(LCA(x,y))}{Profondeur(x) + Profondeur(y)}, \tag{3.1}$$

où $Profondeur(x) = D(x)/D$ avec $D(x)$ le nombre de liens entre le plus court chemin de la racine de l'arbre au label de la variable x et D la profondeur totale de l'arbre, c'est-à-dire le nombre total de liens entre le plus court chemin de la racine aux feuilles de l'arbre ; $LCA(x,y)$ représente l'ancêtre commun des labels des variables x et y le plus bas dans l'arbre, c'est-à-dire le label de la variable de l'ensemble I le plus bas ayant à la fois x et y comme descendants.
La mesure de Wu et Palmer varie entre 0 lorsque les labels des variables ne présentent aucune relation, et 1 lorsque les labels des variables présentent une relation étroite dans l'arbre.

La mesure de Resnik

La mesure de Resnik analyse les relations entre les labels des variables en se basant sur la structure de l'arbre de nomenclature et l'*information content* (IC) [109]. L'*information content* est utilisée pour mesurer la quantité d'information présente dans une variable. Plus une variable est rare dans l'arbre, plus elle sera considérée comme spécifique et informative. Par

conséquent, contrairement à la mesure de Wu et Palmer, la mesure de Resnik permet de considérer la fréquence des variables au sein de l'arbre. La formule de Resnik s'exprime comme suit :

$$R(x, y) = IC(LCA(x, y)), \tag{3.2}$$

où $IC(z) = -\log P(z)$ avec $P(z)$ la probabilité d'occurrence d'une variable z estimée par sa fréquence.
Plus la valeur de la mesure de Resnik est élevée, plus les labels des variables partagent d'informations et plus leur relation dans l'arbre est étroite.

la mesure de Lin

Tout comme la mesure de Resnik, la mesure de Lin analyse les relations entre les labels des variables en se basant sur la structure de l'arbre de nomenclature et l'*information content*. Cependant, la différence entre ces deux mesures réside dans le fait que, là où la mesure de Resnik ne tient compte que de l'*information content* de l'ancêtre commun des labels, la mesure de Lin considère également l'*information content* individuel de chacun des deux labels comparés. Elle est définie par la formule suivante :

$$Lin(x, y) = \frac{2 \times IC(LCA(x, y))}{IC(x) + IC(y)}. \tag{3.3}$$

La mesure de Lin varie entre 0 lorsque les labels des variables ne présentent aucune relation, et 1 lorsque les labels des variables présentent une relation étroite dans l'arbre.
En tenant compte de l'*information content* individuel de chacun des deux labels comparés, la mesure de Lin est capable de faire des distinctions plus précises entre les labels (par rapport à la mesure de Resnik). En effet, au sein d'une nomenclature, de nombreux labels peuvent partager le même ancêtre commun. Par conséquent, si seule l'*information content* de l'ancêtre commun est prise en compte, comme c'est le cas pour la mesure de Resnik, ces labels ayant le même ancêtre commun auraient des valeurs identiques [110].

La mesure de Jiang et Conrath

La mesure de Jiang et Conrath, tout comme la mesure de Lin, analyse les relations entre les labels des variables en se basant sur la structure de l'arbre et en combinant l'*information content* individuel de chacun des deux labels et de celui de l'ancêtre commun de ces deux

labels [111]. Cette combinaison lui confère également l'avantage de faire des distinctions plus précises entre les labels par rapport à la mesure de Resnik. Sa formule est définie comme suit :

$$JC(x,y) = \frac{1}{IC(x) + IC(y) - 2 \times IC(LCA(x,y))}. \tag{3.4}$$

Plus la mesure de Jiang et Conrath est élevée, plus la relation entre les labels des variables est étroite dans l'arbre.

La combinaison de l'*information content* individuel de chacun des deux labels et de celui de leur ancêtre commun s'opère différemment entre les mesures de Lin et de Jiang et Conrath. Dans la mesure de Lin, cette combinaison se fait au moyen d'un ratio, tandis que dans la mesure de Jiang et Conrath, elle est faite au moyen d'une différence.

Pedersen et al. ont évalué les relations entre 30 paires de labels appartenant à la nomenclature du SNOMED-CT. Ils ont pour cela utilisé plusieurs mesures, notamment celles de Resnik, de Lin et de Jiang et Conrath. Ces relations ont également été soumises à l'évaluation d'experts médicaux. La qualité des relations calculées par les différentes mesures a été évaluée en comparant leurs corrélations avec les relations établies par les experts. Les résultats ont montré que les corrélations des relations obtenues avec la mesure de Lin était plus proche de l'avis des experts que la mesure de Jiang et Conrath. D'autre part, les corrélations des relations obtenues avec la mesure de Resnik étaient similaires à celles obtenues avec la mesure de Jiang et Conrath.

La mesure de Leacock-Chodorow

De la même manière que la mesure de Wu et Palmer, la mesure de Leacock-Chodorow analyse les relations entre les labels des variables en prenant en compte la structure de l'arbre [112]. Cependant, il ne considère pas la profondeur individuelle des deux labels comparés dans l'arbre, mais la profondeur total de l'arbre comme suit :

$$LC(x,y) = -\log \frac{D(LCA(x,y))}{2 \times D}, \tag{3.5}$$

où $D(LCA(x,y))$ représente le nombre de liens entre le plus court chemin de la racine de l'arbre à l'ancêtre commun le plus bas des labels des variables x et y et D représente la profondeur totale de l'arbre, c'est-à-dire le nombre total de liens entre le plus court chemin de la racine aux feuilles de l'arbre.

Plus la valeur de la mesure de Leacock-Chodorow est élevée, plus la relation entre les variables est étroite dans l'arbre.

Selon le même protocole utilisé dans [110], Alonso et Contreras ont évalué les relations entre plusieurs paires de labels provenant de la nomenclature du SNOMED-CT en utilisant plusieurs mesures dont celles de Wu et Palmer et Leacock-Chodorow. La qualité des relations calculées par ces mesures a été évaluée en les comparant avec les relations établies par des experts médicaux. Les résultats ont révélé que les corrélations des relations obtenues avec la mesure de Wu et Palmer étaient plus proches de l'avis des experts que celles obtenues avec la mesure de Leacock-Chodorow. De manière similaire, McInnes, Pedersen et Pakhomov ont évalué les relations entre 25 paires de labels mais issus de la nomenclature du MeSH. Dans cette étude, c'est la mesure de Leacock-Chodorow qui a obtenu de meilleures corrélations avec l'avis des experts que la mesure de Wu et Palmer. Par conséquent, la structure de l'arbre de nomenclature peut influer sur la qualité des mesures de relation entre labels. De plus, dans l'étude de Pedersen et al., précédemment décrite, outre les mesures de Resnik, de Lin et de Jiang et Conrath, la mesure de Leacock-Chodorow a également été calculée. La mesure de Leacock-Chodorow a obtenu de meilleures corrélations avec l'avis des experts que les trois autres mesures. De manière générale, les mesures qui tiennent compte à la fois de la structure de l'arbre de nomenclature et de la fréquence des variables grâce à l'*information content* permettent d'évaluer les relations entre labels de manière plus précise que les mesures qui se basent uniquement sur la structure de l'arbre [115]. Cependant, il est essentiel de noter que la qualité des mesures de relation entre labels peut varier en fonction des données analysées. Le choix de la mesure doit donc être fait en fonction du contexte spécifique de l'analyse.

3.1.3 Les enjeux de la prise en compte des relations entre labels dans les mesures de similarité

Diverses études ont pris en compte la notion de relation entre labels dans le calcul de la similarité entre patients. Par exemple, Ni et al. ont calculé la similarité entre patients à partir des labels et valeurs associés à leurs codes diagnostiques. Les relations entre les labels de ces diagnostics, issus de la CIM-10, ont été calculées en se basant sur la structure de cette nomenclature[116]. Les valeurs associées à ces labels indiquaient la présence ou l'absence du diagnostic chez le patient. De même, Girardi et al. ont pris en compte les relations entre les labels des diagnostics dans la distance de Jaccard en s'appuyant sur la structure de la CIM-10 [117]. Les valeurs associées à ces labels indiquaient également la présence ou l'absence du diagnostic chez le patient. Par ailleurs, Jia et al. ont pris en compte les relations entre labels des diagnostics dans le calcul de la similarité entre patients en utilisant la structure de la CIM-10 et l'*information content* [118]. Dans cette étude aussi, les valeurs associées à ces labels indiquaient

la présence ou l'absence du diagnostic chez le patient. Toutes ces mesures de similarité ont permis de considérer à la fois les valeurs et les relations entre labels des variables. Cependant, elles ont été conçues pour être appliquées spécifiquement aux variables ayant des valeurs binaires, ce qui les rend inadaptées pour l'analyse de variables quantitatives pourtant fréquentes dans le domaine de la santé. Pour remédier à cette limitation, j'ai développé plusieurs mesures de similarité prenant en compte à la fois les relations entre labels et les valeurs quantitatives des variables de santé. Ces mesures ont ensuite été utilisées dans des approches de clustering pour le calcul de la similarité entre patients. L'objectif de l'utilisation de ces mesures est de parvenir à obtenir des clusters plus cohérents et cliniquement pertinents en comparaison à ceux obtenus avec les mesures de similarité qui négligent les relations entre labels. Ce travail est actuellement soumis pour publication dans "*BMC Medical Research Methodology*" et est présenté dans la section 3.2.

3.2 Publication numéro 2 : Improving patient clustering by incorporating structured variable label relationships in similarity measures

Improving patient clustering by incorporating structured variable label relationships in similarity measures

Judith Lambert[a,b,c], **Anne-Louise Leutenegger**[d], **Anaïs Baudot**[c,e,f,1] **and Anne-Sophie Jannot**[b,g,h,1]

[a]Sorbonne Université, Université Paris Cité, INSERM, Centre de Recherche des Cordeliers, F-75006 Paris, France

[b]HeKA, Inria Paris, F-75015 Paris, France

[c]Aix Marseille Univ, INSERM, MMG, UMR1251, Marseille, France

[d]Université Paris Cité, INSERM, NeuroDiderot , UMR1141, 75019 Paris, France

[e]CNRS, Marseille, France

[f]Barcelona Supercomputing Center, Barcelona, Spain

[g]Université Paris Cité, Sorbonne Université, INSERM, Centre de Recherche des Cordeliers, F-75006 Paris, France

[h]French National Rare Disease Registry (BNDMR), Greater Paris University Hospitals (AP-HP), Paris, France

Corresponding author: Judith Lambert, PariSanté Campus, 10 rue d'Oradour-sur-Glane, 75015 Paris, France, judith.lambert@inserm.fr

Abstract

Background Patient stratification is the cornerstone of numerous health investigations, serving to enhance the estimation of treatment efficacy and facilitating patient matching. To stratify patients, similarity measures between patients can be computed from clinical variables contained in medical health records. These variables have both values and labels structured in ontologies or other classification systems. The relevance of considering variable label relationships in the computation of patient similarity measures has been poorly studied.

Objective We propose and evaluate several weighted versions of the Cosine similarity that consider structured label relationships to compute patient similarities from a medico-administrative database.

Material and Methods As a use case, we clustered patients aged 60 years from their annual medicine reimbursements contained in the *Échantillon Généraliste des Bénéficiaires*, a random sample of a French medico-administrative database. We used four patient similarity measures: the standard Cosine similarity, a weighted Cosine similarity measure that includes variable frequencies and two weighted Cosine similarity measures that consider variable label relationships. We construct patient networks from each similarity measure and identify clusters of patients using the Markov Cluster algorithm. We evaluate the performance of the different similarity measures with enrichment tests based on patient diagnoses.

[1] These authors contributed equally to this work.

Results The weighted similarity measures that include structured variable label relationships perform better to identify similar patients. Indeed, using these weighted measures, we identify more clusters associated with different diagnose enrichment. Importantly, the enrichment tests provide clinically interpretable insights into these patient clusters.

Conclusion Considering label relationships when computing patient similarities improves stratification of patients regarding their health status.

Keywords: prior expert knowledge; structured variable labels; patient stratification; patient clustering; patient networks; similarity measures

1 BACKGROUND

Identifying similar patients can serve multiple purposes in healthcare. In routine practice, finding patients similar to a given patient can help elucidate undiagnosed cases, particularly in the case of rare diseases [1]. Similar patients can also provide prognostic guidance. In research, identifying groups of similar patients is useful to stratify the population. This enables, for instance, a more precise estimation of medicine efficacy for a given patient profile or matching similar patients in case-control studies [2,3].

The similarity between patients can be computed from medical health records such as medico-administrative databases. Medico-administrative databases contain data used for health care reimbursement purposes, including information about hospitalization, medicine and medical device consumption. They therefore provide a comprehensive perspective on the entire healthcare pathway of a given patient.

The variables contained in these databases are labeled by terms that can be related to each other. For instance, medicines are labeled by codes organized into classification trees such as the Anatomical Therapeutic Chemical (ATC) classification. Within this classification, all anti-diabetic medicines belong to the same class. Thus, the labels of two medicines used to treat diabetes are related according to the classification, and patients treated with these medicines are expected to be more similar than patients treated with medicines from different classes. Other classification systems are available for various types of medical data. For instance, the International Classification of Diseases (ICD) is used for diagnoses [4] and SNOMED-CT is used for clinical information [5].

The objective of similarity measures is to identify patients sharing characteristics, such as taking the same medicine or having the same diagnoses at a specific age. The most commonly used

measures to compute similarities between patients are the Euclidean distance, the Jaccard index, and the Cosine similarity [6]. Weighted measures can also be incorporated to these similarity measures to account for the frequency of the variables. For instance, the Inverse Document Frequency is a weighted measure that assigns greater importance to rare variables [7,8]. However, these classical similarity measures rely only on the variable values and do not consider other information associated with the variables, such as their labels. For instance, when considering medications, two patients are similar if they take similar dosage of similar medicine. In this case, similarity is defined at both the dosage level (i.e., variable value) and at the medicine level (i.e., variable label relationships). Hence, analyzing the relationships between variable labels, by leveraging a label classification system, is expected to provide pertinent information for computing patient similarity from a clinical perspective.

Several measures have been proposed to analyze variable label relationships. The Wu and Palmer measure examines the relationships between two variable labels by considering their depth in the classification tree [9] while the Lin measure considers both their depth in the classification tree and their frequency [10]. These measures have been incorporated into the computation of patient similarities in various studies. For instance, Ni et al. compute patient similarities based on their ICD-10 diagnosis, using a weighted similarity measure that considers the classification depth of ICD-10 codes [11]. Girardi et al. adapted the Jaccard distance to include diagnose relationships in patient similarity computation based on the depth of their ICD-10 codes [12]. However, these weighted patient similarity measures are limited to variables associated with binary values (i.e., Boolean variables) and cannot be applied to quantitative variables. To the best of our knowledge, similarity measures able to simultaneously consider variable label relationships from classifications and quantitative values of variables are lacking in the field.

The efficiencies of the similarity measures are usually estimated by assessing the quality of the clusters obtained using the measures. The clustering performance is evaluated with metrics such as silhouette score or accuracy. However, interpreting these performance metrics from a clinical perspective can be challenging. An alternative method to assess the performance of clustering involves using external variables that were not used initially to compute patient similarities. These external variables can be related to prognosis [13] or tumor characteristics [14], for instance.

In this paper, we propose to weight the Cosine similarity to include variable label relationships in order to identify similar patients. We further aim to assess the added value of incorporating this information thanks to an evaluation protocol that can be interpreted clinically. Our study focuses on a specific use case related to medicine reimbursement in a national French medico-administrative database. We first compute several weighted similarity measures and employ them to cluster patients, thereby revealing groups of similar patients. We then assess the performance of the different similarity measures in identifying clusters of patients thanks to

enrichment tests based on external variables (i.e., diagnoses). We observed that taking into account the relationships between the variable labels in the computation of patient similarities improves the quality of the identified patient clusters.

2 MATERIAL AND METHODS

Let *I* be the set of variables (i.e., medicines). Let *X* and *Y* be the vectors of variables from *I* for two patients. We compute the similarity between patients using four different measures. The first two measures (Cosine similarity and Cosine similarity weighted by the Inverse Document Frequency) rely on quantitative variables, while the remaining two (Cosine similarity weighted by the Wu and Palmer measure and Cosine similarity weighted by the Lin measure) rely on both quantitative variables and label relationships of the variables.

2.1 Defining patient similarity measures

2.1.1 Cosine similarity

The Cosine similarity between two patient vectors *X* and *Y* is defined as the cosine of the angle (θ) between the two vectors [15]:

$$\cos_\theta(X,Y) = \frac{X.Y}{||X||.||Y||} \quad (1)$$

The Cosine similarity values range from -1 to 1 , with value equal to -1 when vectors are opposite, 0 when vectors are different (i.e., orthogonal) and 1 when they are identical.

2.1.2 Cosine similarity weighted by the Inverse Document Frequency

The Cosine similarity weighted by the Inverse Document Frequency (IDF) is defined as follows for two patient vectors *X* and *Y* [7]:

$$\cos_{\theta,IDF}(X,Y) = \frac{\sum_{i\in I} IDF(i)X_iY_i}{\sqrt{\sum_{i\in I} IDF(i)X_i^2}\sqrt{\sum_{i\in I} IDF(i)Y_i^2}} \quad (2)$$

with $IDF(i) = \log \frac{N_I}{N_i}$, where N_I is the total number of observations of all the variables in the set *I* and N_i is the total number of observations of the variable *i*.
As for the standard Cosine similarity, the values of this weighted version range from -1 to 1.

2.1.3 Cosine similarity weighted by the Wu and Palmer measure

The Cosine similarity weighted by the Wu and Palmer measure is defined as follows for two patient vectors *X* and *Y* [9]:

$$\cos_{\theta,WP}(X,Y) = \frac{\sum_{i,j\in I} WP(ij)X_iY_j}{\sqrt{\sum_{i,j\in I} WP(ij)X_iX_j}\sqrt{\sum_{i,j\in I} WP(ij)Y_iY_j}} \quad (3)$$

with *WP*(*i,j*) being the Wu and Palmer measure between the labels of the two variables *i* and *j* of the set *I*.

The labels of the variables of the set *I* are organized into a classification tree consisting of successive levels (*Figure 1*). The label composing the top level is the root and the labels composing the lowest level are the leaves. Each level in the classification is connected to the next level through edges representing the relationship between variable labels in the classification. A sequence of edges represents a path in the classification.

The Wu and Palmer measure is computed from the variable labels as follows:

$$WP(i,j) = \frac{2 \times depth(LCA(i,j))}{depth(i) + depth(j)} \quad (4)$$

with *depth*(*z*) = E_z/E where E_z is the number of edges between the root and the variable label *z* in the classification tree and *E* is the total depth in the classification tree (i.e., the number of edges in the shortest path from the root to the leaves); *LCA*(*i,j*) is the Lowest Common Ancestor of the labels of the variables *i* and *j* in the classification (i.e., the lowest label of the variable of set *I* that has both *i* and *j* as descendants). For example, the Wu and Palmer measure between the medicine labels B1 and B22 from the classification of the *Figure 1* is computed as follows: $WP(B1,B22) = \frac{2\times depth(B)}{depth(B1)+depth(B22)} = \frac{2\times(1/3)}{(2/3)+(3/3)} = 0.40.$

As for the standard Cosine similarity, the values of this weighted version range from -1 to 1.

2.1.4 Cosine similarity weighted by the Lin measure

The Cosine similarity weighted by the Lin measure is defined as follows for two patient vectors *X* and *Y*:

$$\cos_{\theta,Lin}(X,Y) = \frac{\sum_{i,j\in I} Lin(ij)X_iY_j}{\sqrt{\sum_{i,j\in I} Lin(ij)X_iX_j}\sqrt{\sum_{i,j\in I} Lin(ij)Y_iY_j}} \quad (5)$$

With *Lin*(*i,j*) the Lin measure between the labels of the two variables *i* and *j* of the set *I*.

The Lin measure analyzes the relationship between two variables *i* and *j* by considering the information content (IC) of the labels of the two variables and the information content of their lowest common ancestor [10]:

$$Lin(i,j) = \frac{2 \times IC(LCA(i,j))}{IC(i) + IC(j)} \quad (6)$$

with *IC*(*z*) = −log*P*(*z*) where *P*(*z*) is the probability of occurrence of the variable *z* estimated by its frequency. For example, the Lin measure between the medicine labels B1 and B22 from the

classification of the *Figure 1* is computed as follows: $Lin(B1, B22) = \frac{2 \times IC(B)}{IC(B1)+IC(B2)} = \frac{2 \times \log(0.4)}{\log(0.1)+\log(0.18)} = 0.46$

As for the standard Cosine similarity, the values of this weighted version range from -1 to 1.

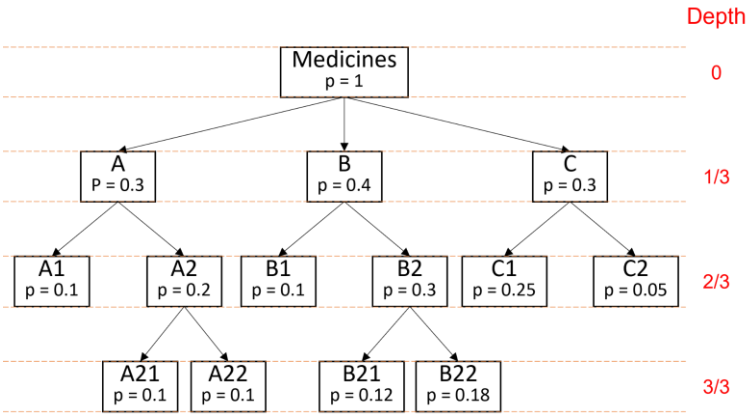


Figure 1: Example of a classification tree of medicine data

The classification is composed of several medicine labels organized in successive levels interconnected by edges. The depth of a given variable label is the number of edges between the root (i.e., label "Medicines") and that given variable label, divided by the number of edges in the shortest path from the root to the leaves. *p* is the medicine label frequency

2.2 Identifying clusters of patients from patient networks

Various clustering methods can be used to identify clusters of patients from their similarity measures. Some examples include K-means, hierarchical clustering, or the Markov cluster algorithm applied to patient networks [16,17]. In a previous work, we showed that building patient networks using Cosine similarity on medicine data and clustering the networks was a pertinent approach to identify patient clusters and trajectories [18]. Therefore, here, we build patient networks using Cosine similarity or its weighted versions on medicine data, and cluster the networks with the Markov Cluster algorithm to identify clusters of patients.

2.2.1 Constructing patient networks

A patient network is a graph $G = (V,E)$ with V patient nodes and E edges representing interactions between patient nodes. The network is constructed using a similarity matrix. Let $M = [m_{X,Y}]^n$ be the similarity matrix where n is the number of patients and $m_{X,Y}$ is the similarity between patient vectors X and Y. This similarity matrix is symmetrical, with $m_{X,Y} = m_{Y,X}$. We compute four similarity matrices, each corresponding to a specific similarity measure. We then apply a threshold t to the similarity matrices to construct the patient networks. Two patients are connected in the network (i.e., an edge between the patients is present) if their similarity is above the threshold t. The connection between patients X and Y is weighted by the value $m_{X,Y}$ of the matrix.

To ensure comparable networks for the different similarity measures, we select a distinct threshold *t* for each measure. These thresholds are chosen to obtain approximately 5000 patient nodes in the largest connected component of each network (*supplementary Table S1*).

2.2.2 Clustering patient networks

We apply the Markov Cluster algorithm (MCL) [19] on the largest connected component of each patient network. The MCL algorithm uses random walks to simulate flows on the network. The flows allow to distinguish network areas where nodes are strongly connected, which correspond to the clusters. We use the version 0.0.6.dev0 of the "markov-clustering" Python package with the default parameters.

2.3 Cluster enrichment analysis

Let external variables be binary variables that are not used to compute the similarities between patients. The aim of the enrichment analysis is to assess if each external variable has a frequency higher than expected in a cluster. For each external variable and each cluster of patients, we compare patients inside and outside the cluster using Fisher's exact test [20]. This procedure involves performing a number of tests equal to the product of the number of clusters times the number of external variables. We adjust this multiple testing with the Benjamini-Hochberg procedure. We consider that a variable is enriched in a given cluster if its adjusted p-value is lower than 0.05.

2.4 Use-case: the Échantillon Généraliste des Bénéficiaires

We use health data from the Échantillon Généraliste des Bénéficiaires (EGB), a French medico-administrative database. The EGB is a random sample of the French health insurance database [21]. It is representative of the French population and contains approximately 660,000 individuals followed over a period of 11 years.

We extract from the EGB data on medicine reimbursements between 2008 and 2018 (*Figure 2*), including the date of reimbursement and the medicine classification in the Anatomical Therapeutic Chemical (ATC) class (see example *Table 1*). The ATC class is an international classification of medicines established by the World Health Organization (WHO) [22]. We exploit this classification in the patient similarity measures. We then select patients aged 60 and who had received reimbursement for at least one medicine for two or more consecutive months. We therefore keep only patients with sustained reimbursements. We also extract chronic disease diagnoses declared by the patient to the French health insurance. These diagnoses are coded with the 10*th* revision of the international statistical classification of diseases and related health problems (i.e., using ICD-10 code). We thus exclude from our analysis the patients with no

declared chronic diseases. Importantly, diabetes appears as the most frequent chronic disease observed within the population. We analyze female and male patient datasets separately. In each dataset, we calculate for each patient, the number of reimbursements they had for each medicine at age 60 (see example *Table 2*).

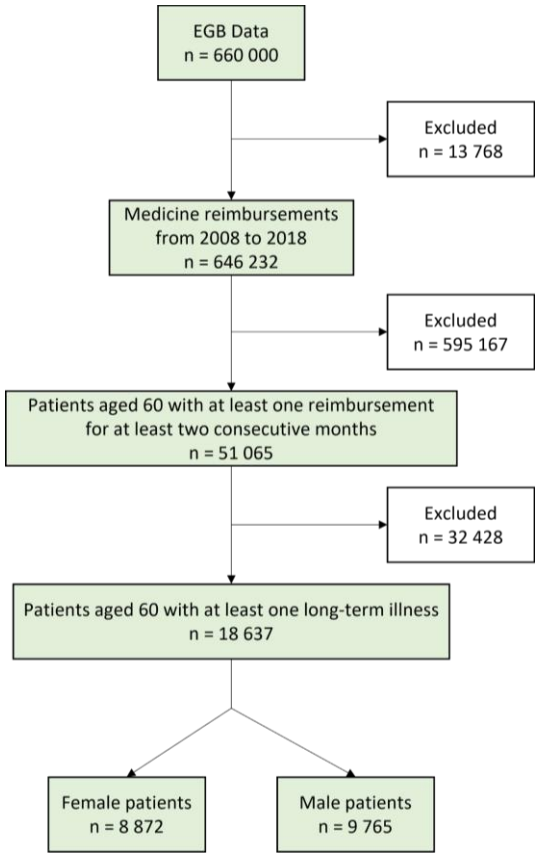


Figure 2: Flowchart of the medicine data extraction process from the Échantillon Généraliste des Bénéficiaires (EGB)

Patient ID	Reimbursement date	ATC class	Medicine name
P_1	01/04/2008	M01AE01	Ibuprofen
P_1	01/12/2015	B01AC06	Aspirin
P_2	01/02/2010	N02AX02	Tramadol
P_3	01/05/2016	B01AC04	Clopidogrel

Table 1: Example of medicine reimbursements contained in the Échantillon Généraliste des Bénéficiaires (EGB)
ATC: Anatomical Therapeutic Chemical

Patient ID	Tramadol	Aspirin	Ibuprofen
P_1	0	10	5
P_2	1	8	4
P_3	2	6	3

Table 2: Example of total number of reimbursements that three patients aged 60 years received for three different medicines

3 RESULTS

Our two use-case datasets are composed of 8,872 female and 9,765 male patients. For each dataset, we compute the similarity between patients, build networks, and identify clusters. We assess the performance of the different patient similarity measures with enrichment tests on the patient clusters using declared chronic diseases.

3.1 Similarity measures including variable label relationships have higher similarity values in the use-case populations

We first compare patient similarities computed from medicine reimbursements using four similarity measures, i.e., the standard Cosine similarity and its weighted versions (*Material and methods 2.1*).

In the dataset of female patients, the Cosine similarity weighted by the Wu and Palmer measure and the Cosine similarity weighted by the Lin measure identify more patient pairs with similarities with non-zero values ($n_0 = 3.89\times10^7$ for these two measures) as compared to the Cosine similarity and the Cosine similarity weighted by IDF ($n_0 = 3.02\times10^7$ for the two other measures) (*Figure 3*). Additionally, the Cosine similarity weighted by the Wu and Palmer measure and the Cosine similarity weighted by the Lin measure show a higher variability. Thus, the weighted Cosine similarity measures that include variable label relationships information have higher similarity values. Similar results are observed in the dataset of male patients (*Figure S1*).

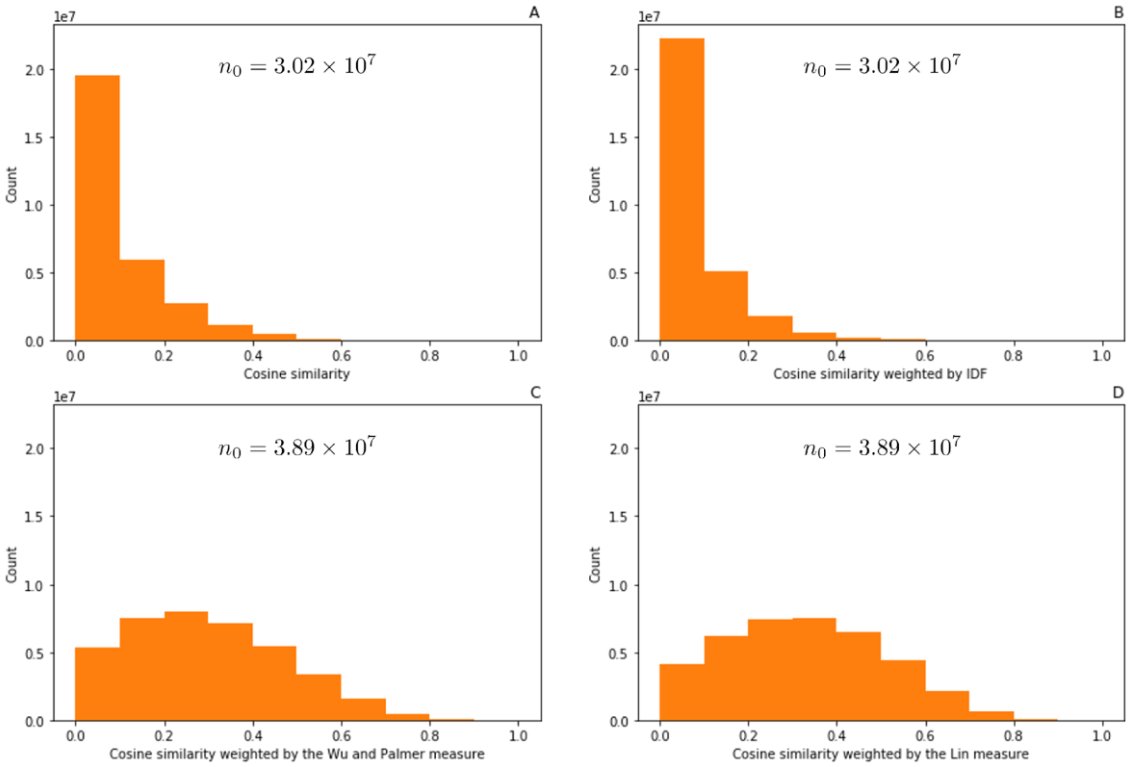


Figure 3: Similarity distributions in the female patient dataset

A: Distribution of the Cosine similarity, B: Distribution of the Cosine similarity weighted by the Inverse Document Frequency (IDF), C: Distribution of the Cosine similarity weighted by the Wu and Palmer measure, D: Distribution of the Cosine similarity weighted by the Lin measure. n_0: Total number of patient pairwise similarities with non-zero values.

3.2 Similarity measures including variable label relationships improve patient cluster quality

A patient network is constructed for each of the four similarity measures, in both male and female datasets, leading to 8 different patient networks (*Material and methods 2.1*). The *Figure 4* shows the two networks constructed for the dataset of female patients using the Cosine similarity and the Cosine similarity weighted by the Wu and Palmer measure. The network constructed with the Cosine similarity (*Figure 4*A) displays a highly connected structure. Conversely, the network constructed with the Cosine similarity weighted by the Wu and Palmer measure (*Figure 4*B) reveals distinct subnetworks.

In the network of female patients built with the Cosine similarity, we identify 12 clusters composed of at least 50 patients. We carry out an enrichment analysis to identify, in each cluster, potential enrichments in chronic diseases (Figure 5A). The enrichment analysis reveals several

significant enrichments. For instance, we observe significant enrichments of patients with thyroid and breast cancers in cluster 1, cerebrovascular diseases in cluster 5 and depressive episodes in cluster 2. Similarly, in the dataset of male patients, we identify 11 clusters (*Figures 6A*). The enrichment analysis reveals significant enrichments of patients with cerebrovascular diseases in cluster 5, atherosclerosis in cluster 6, prostate cancer in cluster 8 and thyroid cancer in cluster 9. Of note, we identify several clusters with the same enriched chronic diseases. For instance, female clusters 2, 3, 9 and 11, and male clusters 2, 4, 7, 10 and 11 are all enriched in type 2 diabetes patients. Female clusters 6 and 7 are enriched in breast cancer patients, female clusters 10 and 12 in autoimmune disorder patients and male clusters 1 and 3 in coronary diseases patients. Overall, the use of Cosine similarity allows to identify clusters of similar patients. However, several clusters are redundant regarding their chronic disease enrichments. Similar results are obtained with the Cosine similarity weighted by IDF (*Figures 5 B and 6 B*).

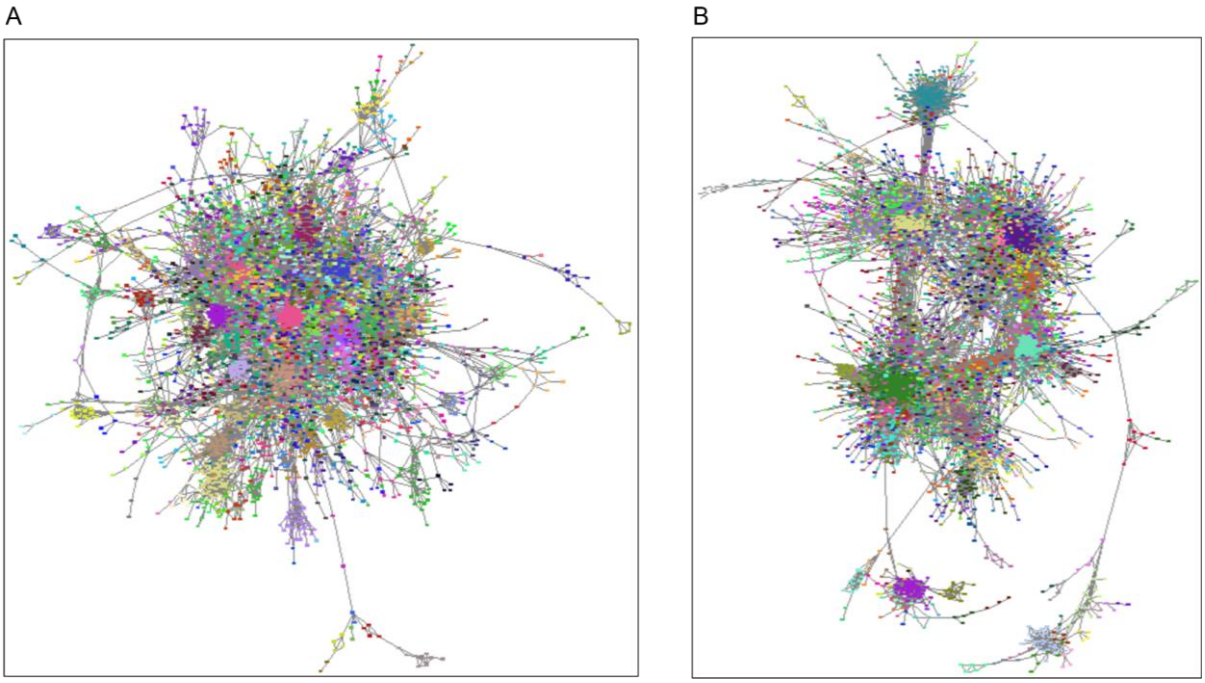


Figure 4: Patient networks built from Cosine similarity and Cosine similarity weighted by the Wu and Palmer measure

Networks are built from Cosine similarity (A) and Cosine similarity weighted by the Wu and Palmer measure (B), on the female patient dataset. Nodes represent patients aged 60 and edges represent the interactions between those patients. The length of edges is inversely proportional to the Cosine similarity or the Cosine similarity weighted by the Wu and Palmer measure. Node colors represent the clusters identified with the Markov Clustering algorithm. For the sake of visualization, we only represent the largest connected component of each network.

Patient networks constructed with the Cosine similarity weighted by the Wu and Palmer measure and the Lin measure result in a higher number of clusters significantly enriched in chronic diseases and less redundant clusters (*Figures 5C, 5D, 6C and 6D*). Using the Cosine similarity weighted by the Wu and Palmer measure (*Figures 5C for the female dataset and 6C for the male dataset*), the enrichment analysis reveals clusters significantly enriched in respiratory disease patients (female cluster 4 and male cluster 7), psychiatric disorders patients with psychotic side (female cluster 1 and male cluster 6), coronary disease patients (female cluster 5 and male cluster 1) and patients with type 2 diabetes associated with its comorbidities (female and male clusters 2). In the dataset of female patients, we find clusters significantly enriched in thyroid and breast cancer patients (clusters 3 and 6). In the dataset of male patients, we find clusters significantly enriched in atherosclerosis patients (cluster 3), depressive disorders patients (cluster 4) and heart failure patients (cluster 9). Similar results are obtained using the Cosine similarity weighted by the Lin measure (*Figures 5D and 6D*). Notably, all clusters are significantly enriched in diabetes patients, in both datasets. This is explained by the fact that diabetes is the most frequent chronic disease in our population (see overall columns in *Figures 5 and 6*).

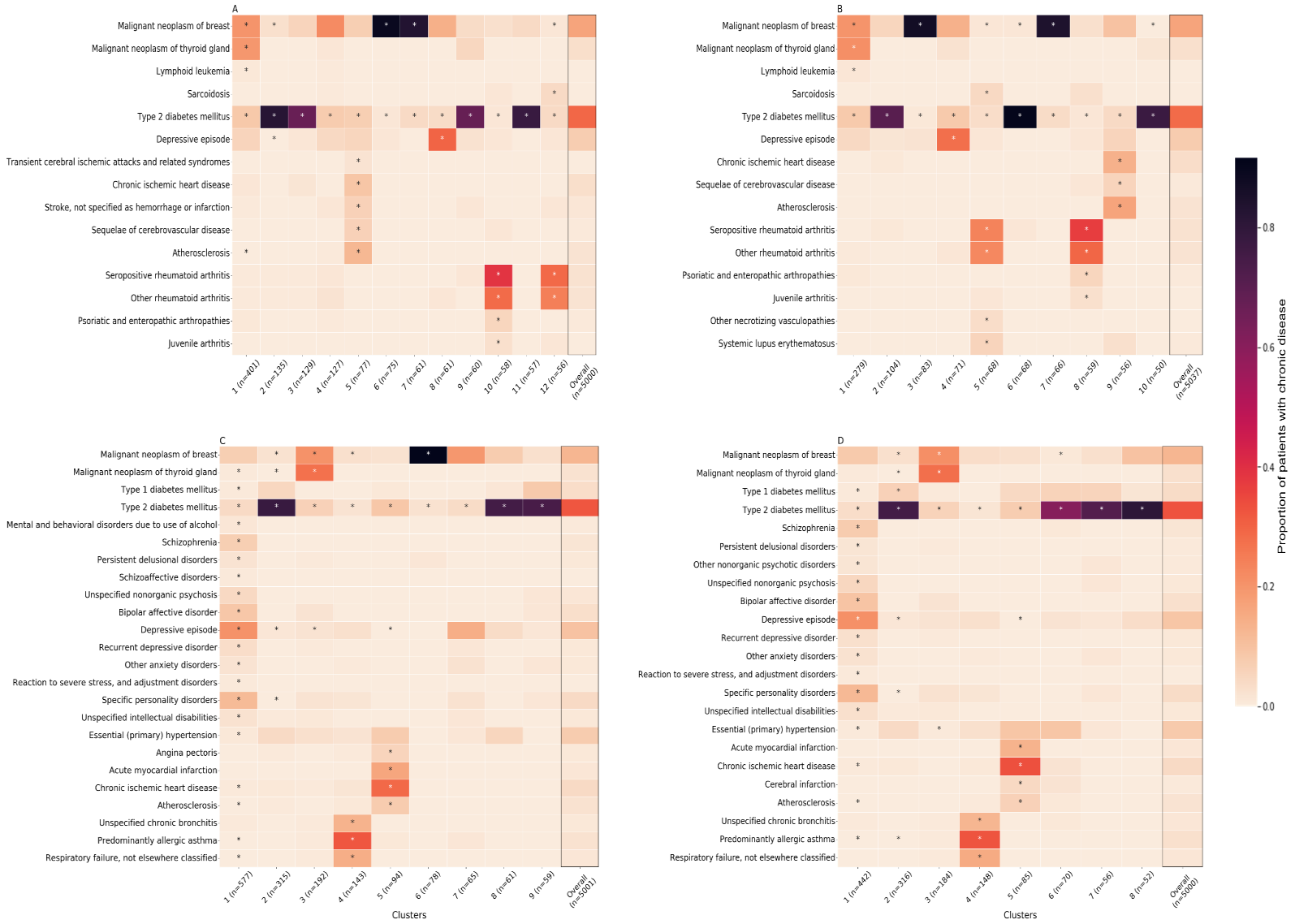


Figure 5: Chronic disease enrichments in patient clusters obtained from the female patient dataset

Clusters are identified in networks built from Cosine similarity (A), Cosine similarity weighted by the Inverse Document Frequency (B), Cosine similarity weighted by the Wu and Palmer measure (C), and Cosine similarity weighted by the Lin measure (D), on the female patient dataset. The numbered columns represent the clusters composed of at least 50 patients, ranked from the largest to the smallest. The last column, named overall, represents all the patients found in the network's largest connected component. n: number of patients identified in each cluster or in the network largest connected component. The rows correspond to the chronic diseases. Box colors represent the proportion of patients with a given chronic disease. Stars represent significant enrichments (p-value lower than 0.05 after Benjamini-Hochberg correction). For the sake of visualization, we only represent chronic diseases that are significant in at least one cluster.

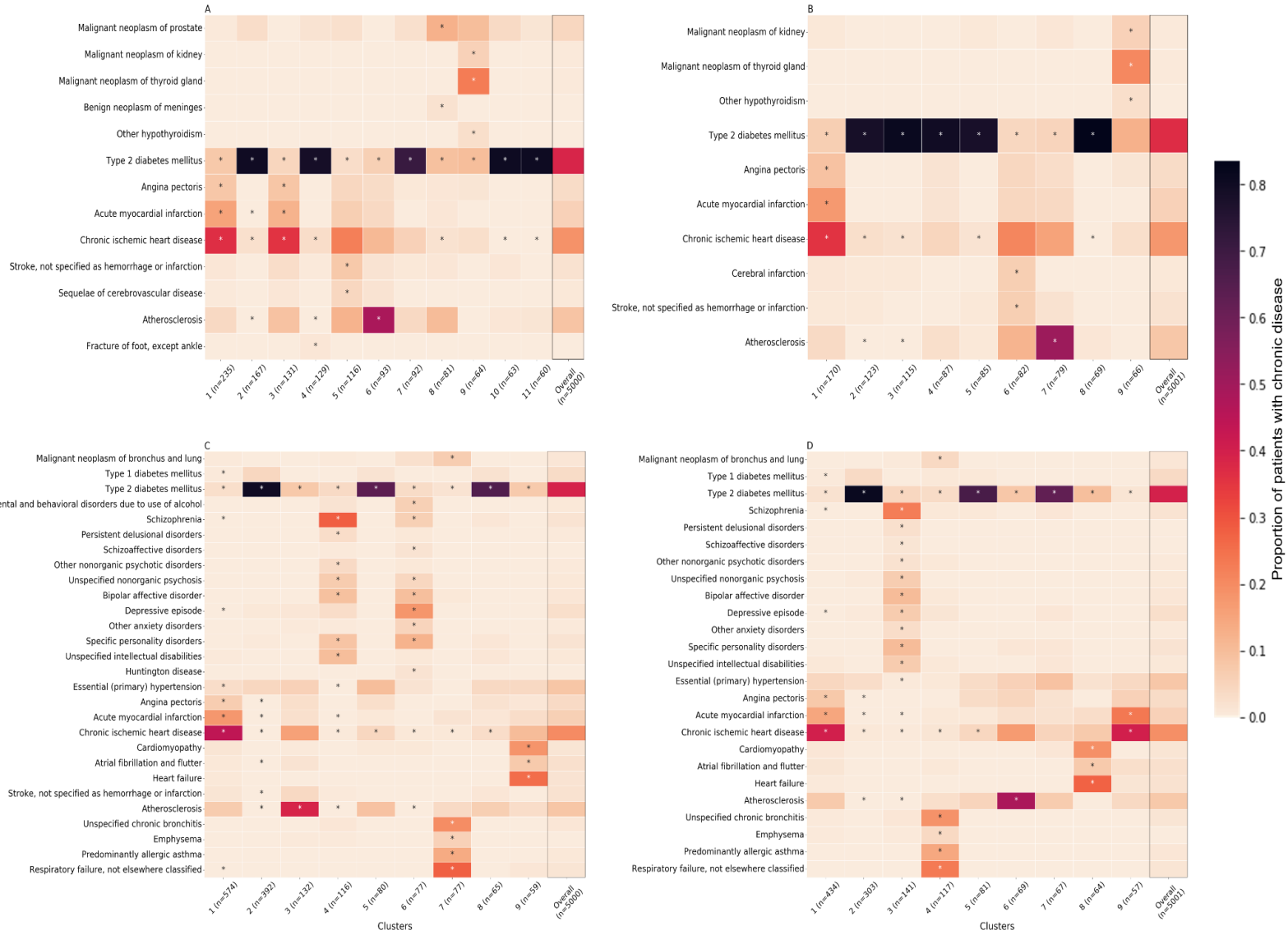


Figure 6: Chronic diseases enrichments in patient clusters obtained from the male patient dataset

Clusters are identified in networks built from Cosine similarity (A), Cosine similarity weighted by the Inverse Document Frequency (B), Cosine similarity weighted by the Wu and Palmer measure (C), and Cosine similarity weighted by the Lin measure (D), on the male patient dataset. The numbered columns represent the clusters composed of at least 50 patients, ranked from the largest to the smallest. The last column, named overall, represents all the patients found in the network's largest connected component. n: number of patients identified in each cluster or in the network largest connected component. The rows correspond to the chronic diseases. Box colors represent the proportion of patients with a given chronic disease. Stars represent significant enrichments (p-value lower than 0.05 after Benjamini-Hochberg correction). For the sake of visualization, we only represent chronic diseases that are significant in at least one cluster.

4 DISCUSSION

In this paper, we proposed two novel weighted similarity measures for health quantitative variables. These similarity measures were weighted by considering the variable labels relationships. They performed better in identifying clusters of patients suffering from different diseases as compared to unweighted similarity measures that did not consider label relationships. Overall, our analysis highlighted the interest of considering variable label relationships when calculating patient similarities to improve patient stratification.

In recent years, there has been a growing interest in computing patient similarities using Electronic Health Records (EHR). However, most papers focused on computing similarities using variables extracted from medical texts through natural language processing methods. These papers also focused on the development of methods to automatically learn variable label relationships from data [23]. However, using variable label relationships from existing medical classifications has been poorly addressed, despite the ready availability of this expert information. In this study, we underline the value of integrating such an expert knowledge into patient similarity measures to increase clustering performance. This is particularly relevant to analyze records obtained from administrative claim databases. Indeed, these databases gather medical variables with labels that are always organized into classifications such as SNOMED-CT or ICD-10.

While our study only considered variable label relationships organized according to a classification, other types of variable label organization exist. For instance, the Human Phenotype Ontology (HPO) is a directed acyclic graph (DAG). Previous studies have already proposed variable label relationship measures for this type of label organization. For example, Köhler et al. developed a variable label relationship measure that exploits the structure of HPO to improve clinical diagnostics [24]. Xue, Peng, and Shang derived another measure that exploits both the DAG structure and the phenotype term definition of HPO in order to improve the prediction of disease-related phenotypes [25]. However, these measures were originally designed for binary variables and would need to be adapted for quantitative variables commonly found in medical health records as well as in many biological and omics datasets. A recent work demonstrated the interest of considering prior knowledge representation in the context of omics data [13].

In this study, we used the Cosine similarity because we have previously shown that this measure was more effective than others to deal with our specific use-cases [18]. However, depending on the data, other similarity measures could be used, and weighted, to compute similarities between patients. We also explored the interest of incorporating variable frequencies in the computation of patient similarities. Indeed, we weighted the Cosine similarity by the Inverse Document Frequency (IDF) to take into account the frequency of the usage of the medicines. Our hypothesis was that it would better capture similarity information as two patients taking the same uncommon medicine would be considered more similar than two patients taking the same

common medicine. However, we observed that exploiting the medicine frequency did not enhance the performance of the similarity measures. This may be attributed to the fact that a single medicine can be used to treat multiple conditions, making it challenging to associate a medicine with a specific pathology. However, considering the frequency of medicines may be more effective in the context of rare diseases [26]. In such cases, these diseases are typically treated with orphan medicines that have specific indications.

In this paper, we assessed the performance of the different similarity measures to cluster patients using external binary variables (i.e., chronic diseases). We employed these external variables in cluster enrichment analyses. This novel approach deviates from the typical reliance on internal criteria such as silhouette score, which do not offer clinically interpretable insights. Using these enrichment analyses, we were able to interpret the clusters clinically and to compare the different similarity measures from an expert point of view. Although previous works have already used enrichment analyses to study enrichments of HPO terms in the literature [27], to the best of our knowledge, it was never used on patient medical data.

As a conclusion, we recommend considering variable label relationships when computing patient similarities to improve stratification of patients regarding their health status.

FUNDING

This work was supported by the Inserm cross-cutting program Genomic variability 2018 GOLD.

ACKNOWLEDGMENTS

The authors would like to acknowledge Pierre Sabatier for extracting and formatting the data. The authors would also like to thank Morgane Terezol and Ozan Ozisik for their comments after proofreading the article. And finally, we would like to thank all the members of MMG and Heka teams for their feedback.

DECLARATIONS

Ethics approval and consent to participate

All methods were carried out in accordance with the RECORD (REporting of studies Conducted using Observational Routinely-collected Data) statement. This study has been declared to INSERM (Institut National de la Santé et de la Recherche Médicale, https://www.inserm.fr/). The data from this study are extracted from the EGB (Échantillon Généraliste des Bénéficiaires), a permanent 1/97 representative sample of the National Health Data System (Système National de Données de Santé, SNDS). The information provided to individuals in EGB on the possible re-use of their data

and the procedures for exercising their rights comply with the legislative and regulatory provisions applicable to the processing of personal data in the SNDS. According to French regulations, informed consent is not required for secondary data reuse, but patient information is mandatory. Therefore, individuals in SNDS database are informed of the reuse of their data for research and can opposed to this reuse as defined by Articles 92 to 95 of Decree No. 2005-1309 of 20 October 2005 (https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037300884/). EGB data, which is part of the SNDS, can be reused for research projects from authorized persons once the research project is declared to their institution (INSERM). Institutional review board approval was not required for this study because the research was based on the secondary use of the EGB (Échantillon Généraliste des Bénéficiaires) medico-administrative database, which is a random sample of 1/97 of the French National medico-administrative database.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are not publicly available due to the access policy of the Échantillon Généraliste des Bénéficiaires database.

Competing interests

The authors have declared no competing interest.

Author's contributions

JL contributed to the writing-original draft preparation, the visualization, the conceptualization and the methodology of the study. AL, AB and AJ contributed to the supervision, the conceptualization, the methodology, the writing-reviewing and the editing of the study.

References

1. Garcelon N, Neuraz A, Salomon R, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. *Orphanet journal of rare diseases*. 2018;13(1):1–11.

2. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med*. 2015;7(311):311ra174.

3. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*. 2018May;6(5):361–9.

4. Hong Y, Zeng ML. International classification of diseases (ICD). *KO KNOWLEDGE ORGANIZATION*. 2023;49(7):496–528.

5. Donnelly K, et al. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*. 2006;121:279.

6. Irani J, Pise N, Phatak M. Clustering techniques and the similarity measures used in clustering: A survey. *International journal of computer applications*. 2016;134(7):9–14.

7. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*. 1972;28(1):11–21.

8. Conroy B, Xu-Wilson M, Rahman A. Patient similarity using population statistics and multiple kernel learning. *In: Machine Learning for Healthcare Conference*. PMLR; 2017.p. 191–203.

9. Wu Z, Palmer M. Verb semantics and lexical selection. arXiv preprint cmp-lg/9406033. 1994.

10. Lin D, et al. An information-theoretic definition of similarity. *In: Icml*. 1998.p. 296–304.

11. Ni J, Liu J, Zhang C, et al. Fine-grained patient similarity measuring using deep metric learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.p.1189–98.

12. Girardi D, Wartner S, Halmerbauer G, et al. Using concept hierarchies to improve calculation of patient similarity. *Journal of biomedical informatics*. 2016;63:66–73.

13. Kańduła MM, Aldoshin AD, Singh S, et al. ViLoN-a multi-layer network approach to data integration demonstrated for patient stratification. *Nucleic Acids Res*. 2023Jan11;51(1):e6.

14. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001Sep11;98(19):10869–74.

15. Singhal A et al.. Modern information retrieval: A brief overview. *IEEE Data Eng Bull*. 2001;24(4):35–43.

16. Xu R, WunschII D. Survey of Clustering Algorithms. *IEEE Trans Neural Netw*. 2005 May;16(3):645–78.

17. Schaeffer SE. Graph clustering. *Computer Science Review*. 2007Aug;1(1):27–64.

18. Lambert J, Leutenegger AL, Jannot AS, Baudot A. Tracking clusters of patients over time enables extracting information from medico-administrative databases. *Journal of Biomedical Informatics*. 2023;139:104309.

19.vanDongen S. A cluster algorithm for graphs. *Information Systems [INS]*. 2000;(R 0010).

20.Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the royal statistical society*. 1922;85(1):87–94.

21.Tuppin P, De Roquefeuil L, Weill A, et al. French national health insurance information system and the permanent beneficiaries sample. *Revue d'epidemiologie et de sante publique.* 2010;58(4):286–90.

22.Skrbo A, Begović B, Skrbo S. Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes. *Medicinski arhiv*. 2004;58(1Suppl 2):138—141.

23.Choi E, Bahadori MT, Searles E, et al. Multi-layer representation learning for medical concepts. In: proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.p.1495–504.

24.Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*. 2009;85(4):457–64.

25.Xue H, Peng J, Shang X. Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. *BMC systems biology*. 2019;13(2):1–12.

26.Chen X, Garcelon N, Neuraz A, et al. Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping. *Journal of Biomedical Informatics*. 2019;100:103308.

27.Deng Y, Gao L, Wang B, et al. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PloS one*. 2015;10(2):e0115692.

3.3 Conclusion et discussion

Nous avons proposé une approche novatrice visant à améliorer le clustering des patients en intégrant les relations entre les labels des variables dans le calcul de similarité entre ces variables. A notre connaissance, notre approche est la première permettant d'utiliser des variables quantitatives. Pour ce faire, nous avons pondéré la similarité Cosinus en utilisant deux mesures distinctes : la mesure de Wu et Palmer qui tient compte de la structure de la nomenclature et la mesure de Lin qui tient compte à la fois de la structure de la nomenclature et de la fréquence des variables. Nous avons utilisé ces mesures pondérées pour calculer la similarité entre les patients lors de la construction des réseaux de patients au sein desquels les clusters ont été identifiés. Nous avons mis en oeuvre cette approche dans l'analyse des remboursements annuels de médicaments d'un échantillon de patients âgés de 60 ans extrait de la base de données médico-administrative EGB. Nous avons comparé les performances de ces mesures pondérées à la similarité Cosinus standard en évaluant la qualité du clustering. A cette fin, nous avons utilisé une mesure de qualité externe qui consiste à réaliser des tests d'enrichissement pour chaque cluster identifié en utilisant comme variables externes les affections de longue durée des patients (maladies chroniques pour lesquelles les médecins réalisent une déclaration spécifique). Ces tests d'enrichissement sont régulièrement utilisés dans le contexte des données génomiques pour identifier les enrichissements d'un ensemble de gènes donné à partir de leurs annotations [119]. Nous avons décidé d'adapter ces tests d'enrichissement dans le contexte de l'évaluation du clustering dans le but de faciliter l'interprétabilité des clusters obtenus. Les résultats de cette évaluation ont révélé que l'emploi des mesures de similarité pondérées considérant les relations entre labels permettait d'identifier des clusters de patients souffrant des mêmes maladies de manière plus précise que la similarité Cosinus seule qui ne tient pas compte de ces relations.

Cette intégration des relations entre labels des médicaments dans la stratification de patients offre ainsi un moyen de regrouper les variables de manière plus informative et pertinente. De plus, l'emploi des tests d'enrichissement pour évaluer les clusters combiné avec une visualisation des résultats sous forme de *heatmaps* a permis de fournir une interprétation clinique simplifiée par rapport aux outils usuels d'évaluation du clustering présenté dans la section 1.2.5. Cette approche a facilité la compréhension de la stratification de patients en mettant en évidence les liens entre les médicaments et les maladies chroniques, ce qui constitue un atout significatif pour l'analyse des données de santé.

Dans cette étude, nous avons exclusivement considéré la similarité Cosinus. Cette décision

découle du fait que dans l'approche du "cluster-tracking" appliquée à l'analyse des remboursements de médicaments, nous avions identifié cette mesure de similarité comme la plus adaptée parmi celles testées. Toutefois, comme énoncé dans la discussion du chapitre précédent, le choix de la mesure de similarité dépend largement du type de données analysé. Par conséquent, la pondération d'autres mesures de similarité pour tenir compte des relations entre labels reste un domaine à explorer. De plus, contrairement à l'étude précédente du "cluster-tracking", nous n'avons pas pris en compte la dimension temporelle des données dans cette étude. Nous nous sommes limités à considérer un seul âge de patients, alors que l'étude précédente s'est étendue sur une période de 10 ans. Il serait donc intéressant d'explorer la valeur ajoutée de l'intégration des relations entre labels dans la similarité des patients pour le "cluster tracking".

Chapitre 4

Discussion et perspectives

Au cours de ma thèse, j'ai exploré deux aspects des caractéristiques des données médico-administratives dans le développement de méthodes de clustering de patients. La première étude a porté sur l'aspect longitudinal des données, aboutissant au développement de l'approche du "cluster-tracking". Grâce à cette approche, nous avons été en mesure de répondre à plusieurs défis liés au clustering de données longitudinales, notamment le choix du nombre optimal de clusters, l'analyse simultanée de multiples données longitudinales, la gestion des données tronquées, ainsi que l'interprétation et la visualisation des résultats qui sont des limitations importantes dans les méthodes usuelles de clustering longitudinal. Nous avons obtenu des clusters regroupant des patients présentant une similarité clinique plus pertinents que ceux obtenus par les méthodes classiques. La deuxième étude a exploré un autre aspect des caractéristiques des données médico-administratives, à savoir les relations entre les labels des variables au sein des nomenclatures. Dans cette perspective, nous avons pondéré la similarité Cosinus avec deux mesures distinctes permettant de considérer la structure des nomenclatures, soit seule, soit en combinaison avec la fréquence des variables. Cette pondération a permis d'améliorer la stratification des patients en générant des clusters plus cohérents et présentant une signification clinique. Nous avons aussi proposé une méthode inspirée des tests d'enrichissement pour évaluer le clustering. Cette évaluation offre l'avantage de rendre les résultats plus interprétables.

Les deux approches développées ont été appliquées à l'analyse des remboursements de médicaments contenus dans la base de données médico-administrative EGB. Plus spécifiquement, dans le "cluster-tracking", nous avons analysé les remboursements de médicaments antithrombotiques pour chaque âge de patients allant de 60 à 70 ans. Nous avons sélectionné cette

tranche d'âge car une augmentation significative du nombre de remboursements est observé à partir de 60 ans, comme illustré dans la *figure 4.1*. En revanche, dans la deuxième étude, nous avons analysé les remboursements des médicaments de toutes les classes ATC pour uniquement un seul âge. Une perspective intéressante consisterait à combiner ces deux approches pour prendre en compte à la fois l'aspect temporel et les relations entre labels dans l'identification des sous-groupes homogènes de patients. Cela impliquerait de calculer la similarité entre patients à chaque instant temporel en utilisant la similarité Cosinus pondérée à partir des remboursements des médicaments de toutes classes. De plus, il serait envisageable d'élargir l'intervalle au-delà des 10 ans initialement choisis.

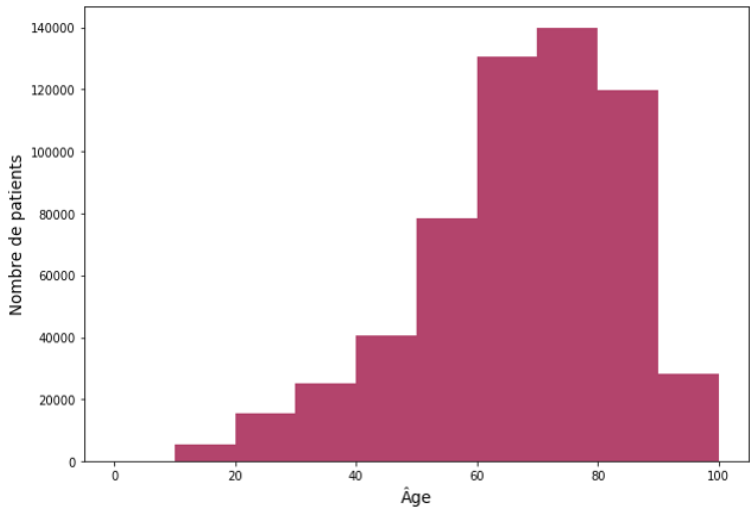


Figure 4.1 – Distribution des patients ayant reçu des remboursements de médicaments antithrombotiques par âge

A chaque âge est représenté le nombre de patients ayant reçu au moins un remboursement de médicaments antithrombotiques entre 2008 et 2018

Au-delà de l'analyse des remboursements de médicaments, ces approches pourraient être étendues aux autres types de données présentes dans les bases de données médico-administratives. En effet, ces bases contiennent également des données sur les hospitalisations, les consultations, les examens complémentaires et d'autres consommations de soins en ville. Compte tenu de la diversité de ces données, il sera nécessaire d'adapter nos travaux en envisageant un large éventail de mesures de similarité. Ces mesures devraient être conçues pour être applicables à des données ponctuelles et convenir à des données de très grande dimension. Nos approches pourraient également être étendues à d'autres types de bases de données de santé. Par exemple, nous avions envisagé d'appliquer la méthode du "cluster-tracking" à partir des données médicamenteuses contenues dans la cohorte UK Biobank. Cependant, contrairement à l'EGB, les

données de UK Biobank comportent un nombre limité d'instants temporels, souvent seulement deux ou trois, correspondant au nombre de visites médicales. Il aurait donc fallu adapter le "cluster-tracking" dans le but d'identifier des trajectoires de clusters pertinentes à partir de données longitudinales de courte durée. Cependant, cette analyse n'a pas pu être réalisée, car l'accès aux données de UK Biobank, en tant que cohorte, nécessitait une demande longue et fastidieuse, prenant environ un an pour être approuvée.

Ma thèse s'inscrit dans le cadre du projet du programme transversal variabilité génomique (projet GOLD) [120], initié par l'Institut National de la Santé Et de la Recherche Médicale (INSERM) en 2018. Ce projet a pour objectif de mieux comprendre les liens complexes entre la variabilité génomique et phénotypique en génotypant et séquençant un échantillon des patients de la cohorte Constances, présentée dans la section 1.1.2 du chapitre 1. Les données génétiques seront disponibles pour 10 000 volontaires de cette cohorte, parmi lesquels 4 000 auront bénéficié d'un séquençage complet de leur génome. Le projet GOLD s'articule autour de deux défis majeurs : le développement d'outils pour synthétiser l'information riche et complexe contenue dans les phénotypes et l'identification des variants génétiques expliquant au mieux ces phénotypes.

Grâce aux outils développés au cours de ma thèse qui permettent de synthétiser l'information phénotypique en établissant des clusters d'individus, il sera désormais possible de mener l'étude sur les associations entre génotype et phénotype. Cette étude vise notamment à identifier si certains génotypes sont plus fréquents au sein de certains clusters de patients. Ce travail sera poursuivi notamment dans le cadre du Programme et Equipement Prioritaire de Recherche Santé Numérique (PEPR SN), contribuant ainsi à mieux comprendre la génétique des phénotypes complexes.

Cependant, la perspective majeure de ce travail réside dans l'application des méthodes développées aux données du SNDS. En effet, ces données sont amenées à être de plus en plus utilisées à des fins d'évaluation des produits de santé et de vigilance mais également d'analyse des parcours de soins. Dans ce contexte, disposer de méthodes adaptées pour définir des sous-groupes de patients homogènes est crucial. Par exemple, ces méthodes peuvent être utilisées pour identifier des comparateurs pertinents, définir des parcours de soins type, ou repérer des populations présentant des risques spécifiques. Bien que ce travail de thèse ne constitue qu'une première étape pour atteindre cet objectif, les perspectives offertes pour les méthodes développées sont très importantes.

Annexe A

Matériel supplémentaire de la publication numéro 1

Tracking clusters of patients over time enables extracting information from medico-administrative databases

Judith Lambert[1,2], Anne-Louise Leutenegger[3], Anne-Sophie Jannot[1,4,*] and Anaïs Baudot[2,5,6,*]

[1]INSERM, Université Paris Cité, CRC, HeKA team, UMR1138, Paris, France
[2]Aix Marseille Univ, INSERM, MMG, UMR1251, Marseille, France
[3]INSERM, Université Paris Cité, NeuroDiderot , UMR1141, Paris, France
[4]Department of Statistics, Medical Informatic and Public Health, HEGP, AP-HP
[5]CNRS, Marseille, France
[6]Barcelona Supercomputing Center, Barcelona, Spain

S1 Similarity measures used for constructing patient networks per age

We used a similarity matrix per patient age to construct each patient network. The similarity between patients is calculated with a similarity measure. In this study, we tested four similarity measures: the Cosine similarity, the opposite of the normalized Euclidean distance, the Jaccard index and the generalized Jaccard index. In the following, we consider two sample sets A and B of length X representing the sum of the reimbursements per drug that two different patients had at a given age.

S1.1 Cosine similarity

The Cosine similarity between two sample sets A and B calculates the cosine of the angle (θ) between them. It is most commonly used in information retrieval or text mining [1]. This measure is defined as follows:

$$\cos_\theta(A, B) = \frac{A.B}{||A||.||B||} \tag{1}$$

The values range from -1 to 1 , with -1 when the samples are opposite, 0 when samples are different (i.e., orthogonal) and 1 when they are identical.

S1.2 Normalized Euclidean distance

In a n-dimensional space, the normalized Euclidean distance between two sample sets A and B is defined as follows [2]:

$$\begin{aligned} NED(A, B) &= \frac{1}{2}.\frac{||(A - \mathbb{E}[A]) - (B - \mathbb{E}[B])||^2}{||(A - \mathbb{E}[A])||^2 + ||(B - \mathbb{E}[B])||^2} \\ &= \frac{1}{2}.\frac{Var(A - B)}{Var(A) + Var(B)}, \end{aligned} \tag{2}$$

with $\mathbb{E}[A]$ and $\mathbb{E}[B]$, the expectation of A and B.
By calculating the opposite of the normalized Euclidean distance, values range from 0 when the samples are different to 1 when they are identical.

*These authors contributed equally to this work.

S1.3 Jaccard index

The Jaccard index calculates the similarity between two sample sets A and B by the ratio of their intersection over their union [3]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{3}$$

When A and B are numeric, we transform each of their element x_j by $x_j^{new} = \mathbb{1}_{x_j \geq 1}$. We then obtain two new sample sets $A' \in \{0,1\}$ and $B' \in \{0,1\}$. The Jaccard index between them is defined as follows:

$$J(A',B') = \frac{M_{11}}{X - M_{00}}, \tag{4}$$

with X the length of the two sample sets, $M_{11} = \sum_{i=1}^{X}(\mathbb{1}_{A'_i=1} \times \mathbb{1}_{B'_i=1})$ and $M_{00} = \sum_{i=1}^{X}(\mathbb{1}_{A'_i=0} \times \mathbb{1}_{B'_i=0})$. This index gives a value between 0 when the two samples are different and 1 when they are identical.

S1.4 Generalized Jaccard index

The generalized version of the Jaccard index allows to calculates the similarity between two numeric vectors, without transforming their elements. The generalized Jaccard index between two sample sets A and B is defined as follows [4]:

$$Jg(A,B) = \frac{\sum_{i=1}^{X} \min(A_i, B_i)}{\sum_{i=1}^{X} \max(A_i, B_i)} \tag{5}$$

with X the length of sample sets A and B. The obtained values are ranges from 0 when the samples are different to 1 when they are identical.

We calculated the Cosine similarity, the Jaccard index, the opposite of the normalized Euclidean distance and the generalized Jaccard index between all patients in our use-case, for each age from 60 to 70 years old. We selected the Cosine similarity for constructing networks because this is the similarity measure having the greatest variance (*Figure S1*). This similarity measure is the one that best distinguishes similar patients from dissimilar patients.

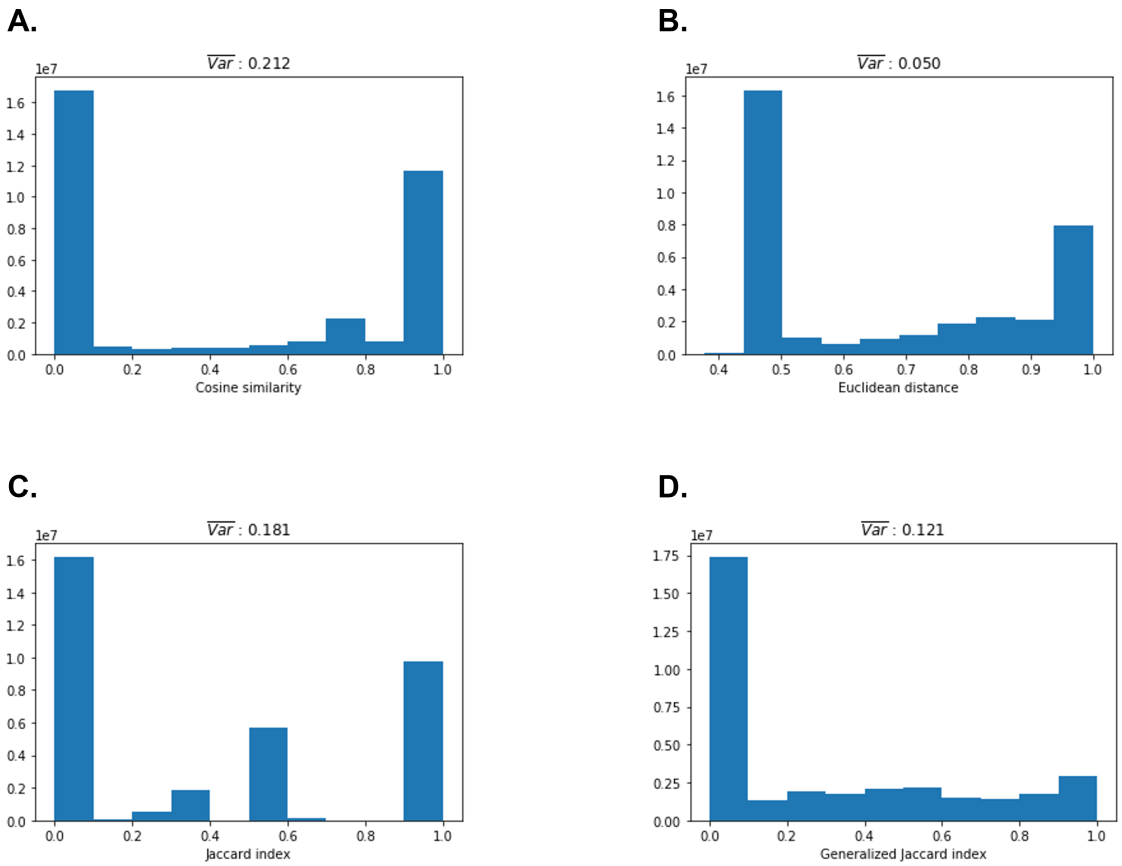


Figure S1: Similarity measure distributions and the mean variance

A: Distribution of the Cosine similarity, B: Distribution of the opposite of the normalized Euclidean distance, C: Distribution of the Jaccard index, D: Distribution of the generalized Jaccard index. The different distributions relate patients aged 60.

$\overline{Var}$: Mean variance from 60 to 70 years old. For each similarity measure, we calculated the distribution variance at each age and we took the mean.

S2 Choice of the threshold applied in similarity matrix

We varied the Cosine similarity from 0 to 1 with a step of 0.1 to choose the threshold. For each threshold tested, we calculated the number of edges and the number of isolated patients (i.e., patient connected to none of the other patients) obtained in the associated network (*Figure S2*). The number of isolated patients is under 1% from 0 to 0.9. We selected 0.8 as threshold because this is where we observe the fastest decrease in the number of edges.

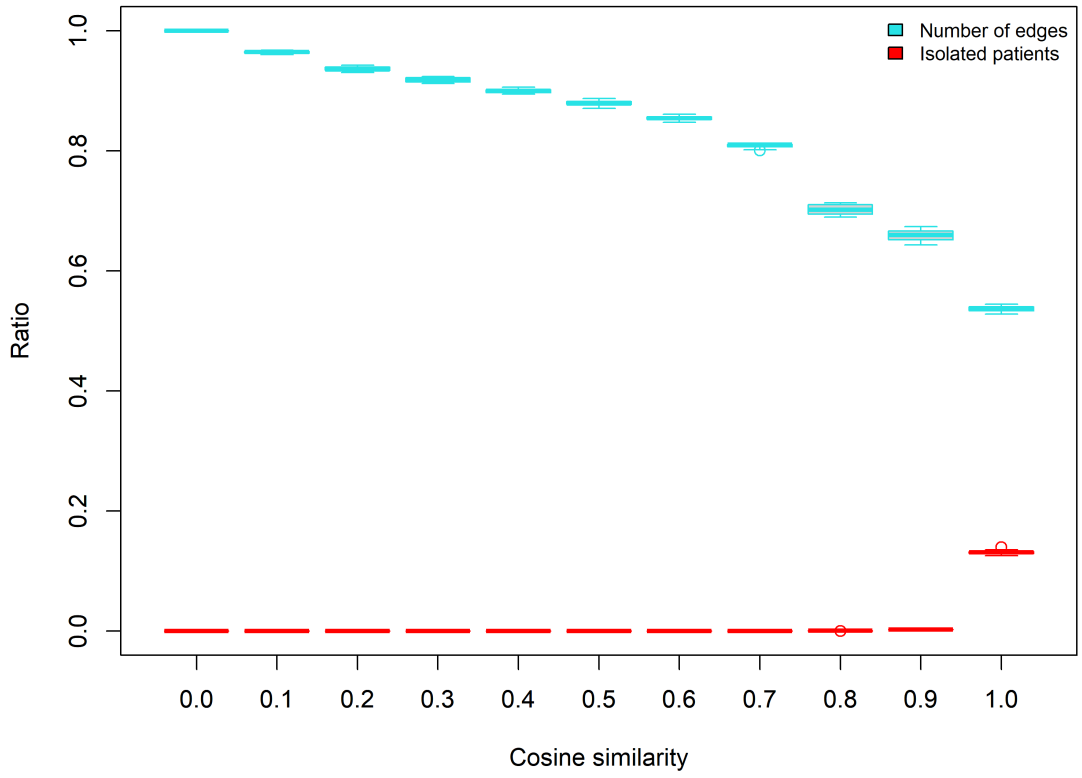


Figure S2: Choice of the Cosine similarity threshold
For each value of Cosine similarity, the blue box represents the number of edges in each patient network from 60 to 70 years old, and the red box represents the isolated patients (i.e., patient connected to none of the other patients)

S3 Assessing the optimal number of clusters using several quality criteria in the three longitudinal clustering approaches

The number of clusters must be specified in the three longitudinal clustering approaches. We used several quality criteria to find the optimal number of clusters. Kml3d, the selected raw-data-based longitudinal-clustering approach, computes five different quality criteria to help selecting the optimal number of clusters: Calinski-Harabasz criterion [5], Kryszczuk variant of Calinski-Harabasz criterion [6], Genolini variant of Calinski-Harabasz criterion [7], the opposite of Ray-Turi criterion [8] and the opposite of Davies-Bouldin criterion [9] (*Figure S3 A*). We also used these five quality criteria in the the feature-based longitudinal-clustering approach to find the optimal number of clusters (*Figure S3 B*). We used the Akaike Information Criterion (AIC) [10] and the Bayesian Information Criterion (BIC) [11] in GMM, the selected model-based longitudinal-clustering approach, as they are calculated by the algorithm (*Figure S3 C*). We also calculated in all the longitudinal approaches, the modified silhouette score (*Figure S4*).

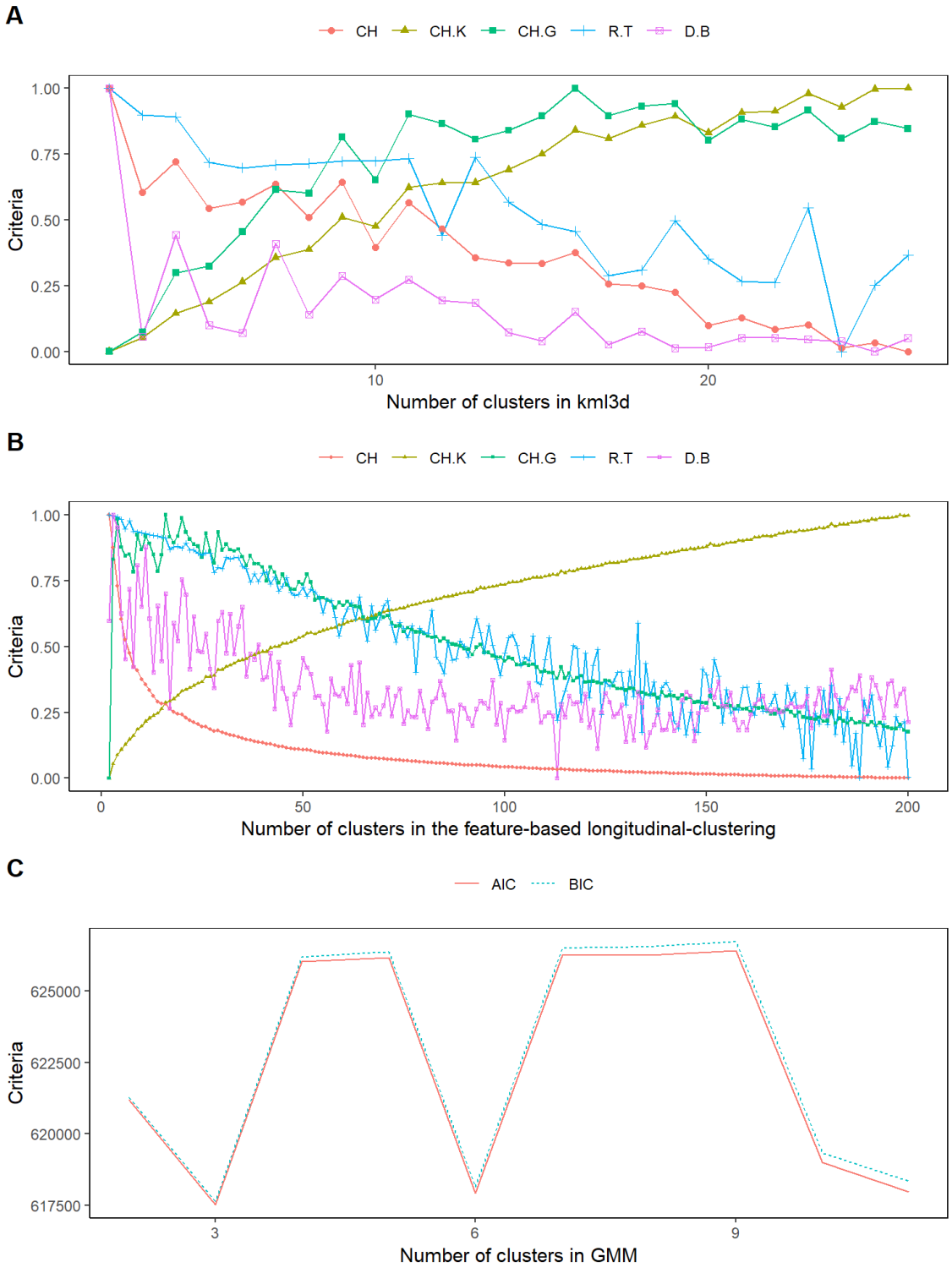


Figure S3: Choice of the optimal number of clusters with several quality criteria computed by the different longitudinal clustering approaches

A: Kml3d, the selected raw-data-based longitudinal-clustering approach, allowed to vary the number of clusters from 2 to 26, B: We varied the number of clusters from 2 to 200 in the feature-based longitudinal-clustering approach, C: We varied the number of clusters only from 2 to 10 in GMM, the selected model-based longitudinal-clustering approach, because of the complexity in the computational time.

CH: Calinski-Harabasz criterion, CH.K: Calinski-Harabasz Kryszczuk variant criterion, CH.G: Calinski-Harabasz Genolini variant criterion, R&T: Ray-Turi criterion, D&B: Davies-Bouldin criterion, AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion

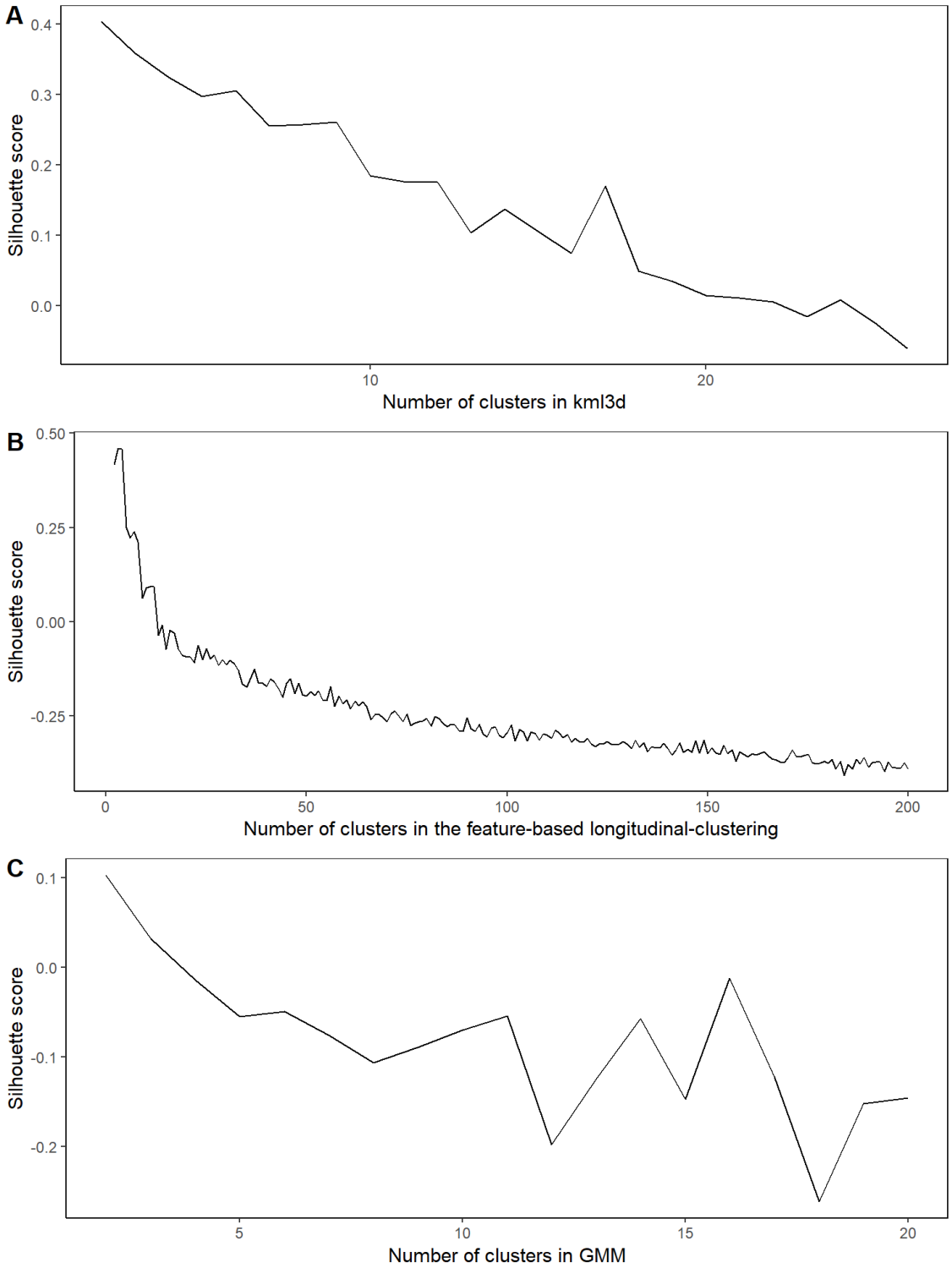


Figure S4: Choice of the optimal number of clusters with the modified silhouette score in the three longitudinal clustering approaches

A: Kml3d, the selected raw-data-based longitudinal-clustering approach, allowed to vary the number of clusters from 2 to 26, B: We varied the number of clusters from 2 to 200 in the feature-based longitudinal-clustering approach, C: We varied the number of clusters only from 2 to 20 in GMM, the selected model-based longitudinal-clustering approach because of the complexity in the computational time (11 days).

S4 Cluster-trajectories identified with the network-based cluster-tracking approach

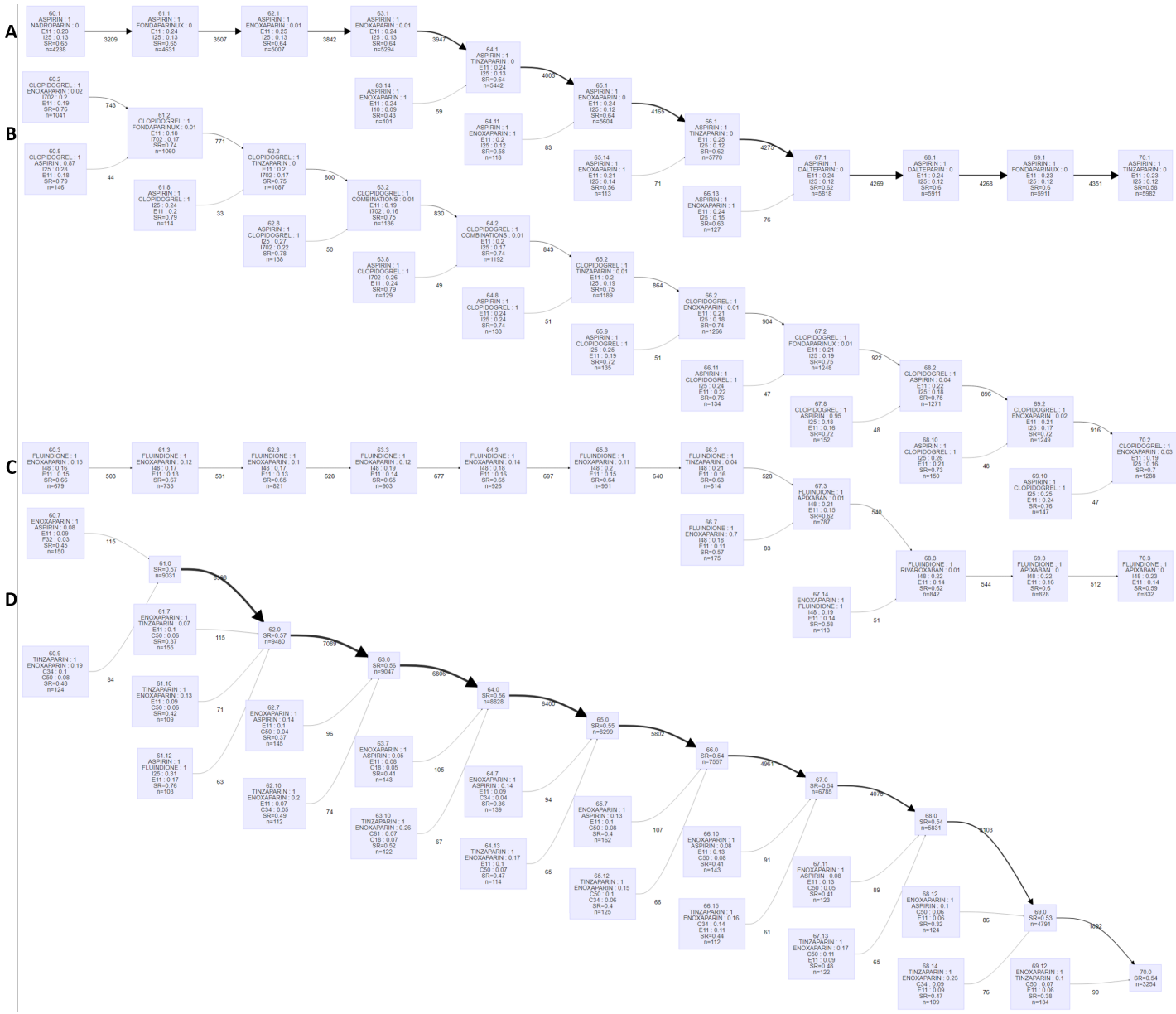


Figure S5: Network-based Cluster-trajectories part 1

Each block represents a cluster. Each cluster is named as follows: "x.y", with x the age time at which it was identified and y the number of the cluster. The clusters are characterized by the two most frequently reimbursed drugs (name, percentage of patients), the two most frequent long-term illnesses (ICD-10 code, percentage of patients), the sex ratio (SR) and the number of patients (n). The number under arrows is the number of patients in common between the two blocks. The arrow thickness is proportional to this number. Combinations: combinations of two platelet aggregation inhibitors.

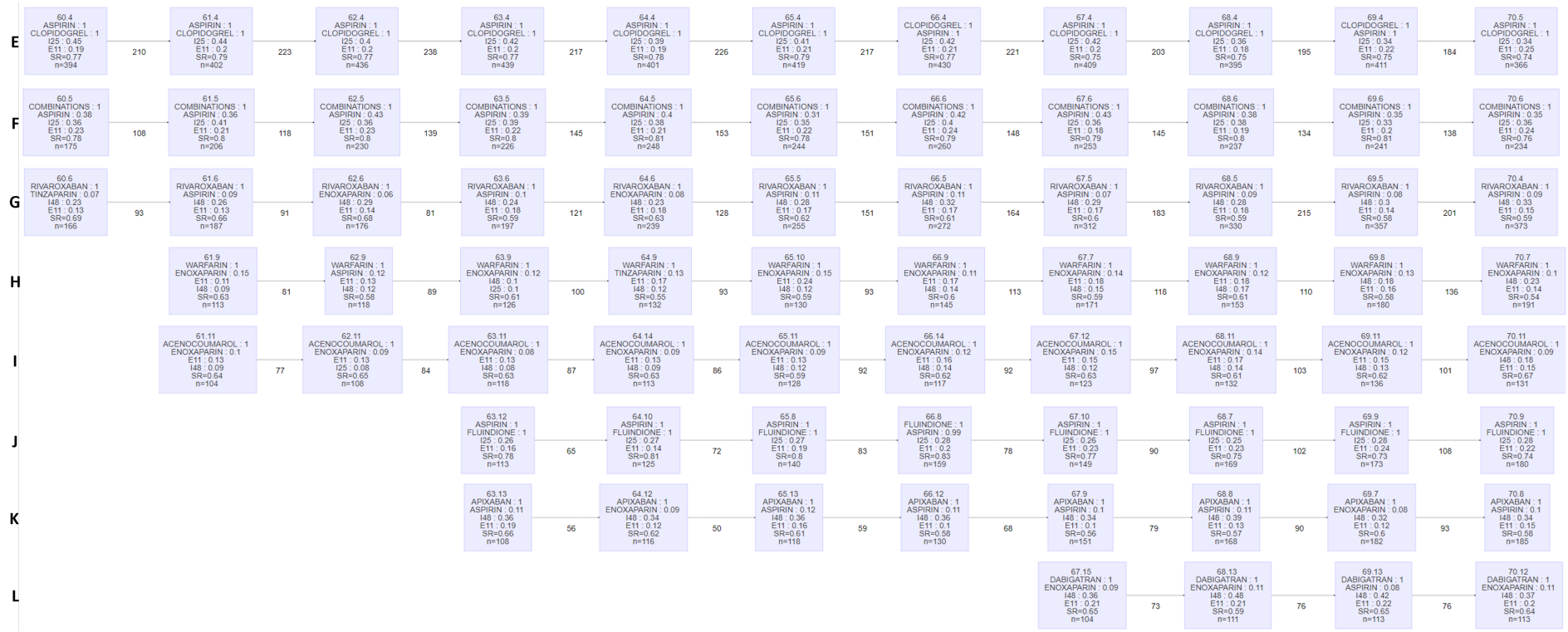


Figure S6: Network-based Cluster-trajectories part 2

Each block represents a cluster. Each cluster is named as follows: "x.y", with x the age time at which it was identified and y the number of the cluster. The clusters are characterized by the two most frequently reimbursed drugs (name, percentage of patients), the two most frequent long-term illnesses (ICD-10 code, percentage of patients), the sex ratio (SR) and the number of patients (n). The number under arrows is the number of patients in common between the two blocks. The arrow thickness is proportional to this number. Combinations: combinations of two platelet aggregation inhibitors.

S5 Assessing the optimal number of clusters using the silhouette score in the raw-data-based cluster-tracking approach

In the raw-data-based cluster-tracking approach, we applied a Kmeans to raw data, for each age considered. In Kmeans, the number of clusters must be specified *a priori*. We determined the optimal number of clusters per age by calculating the silhouette score (*Figure S5*).

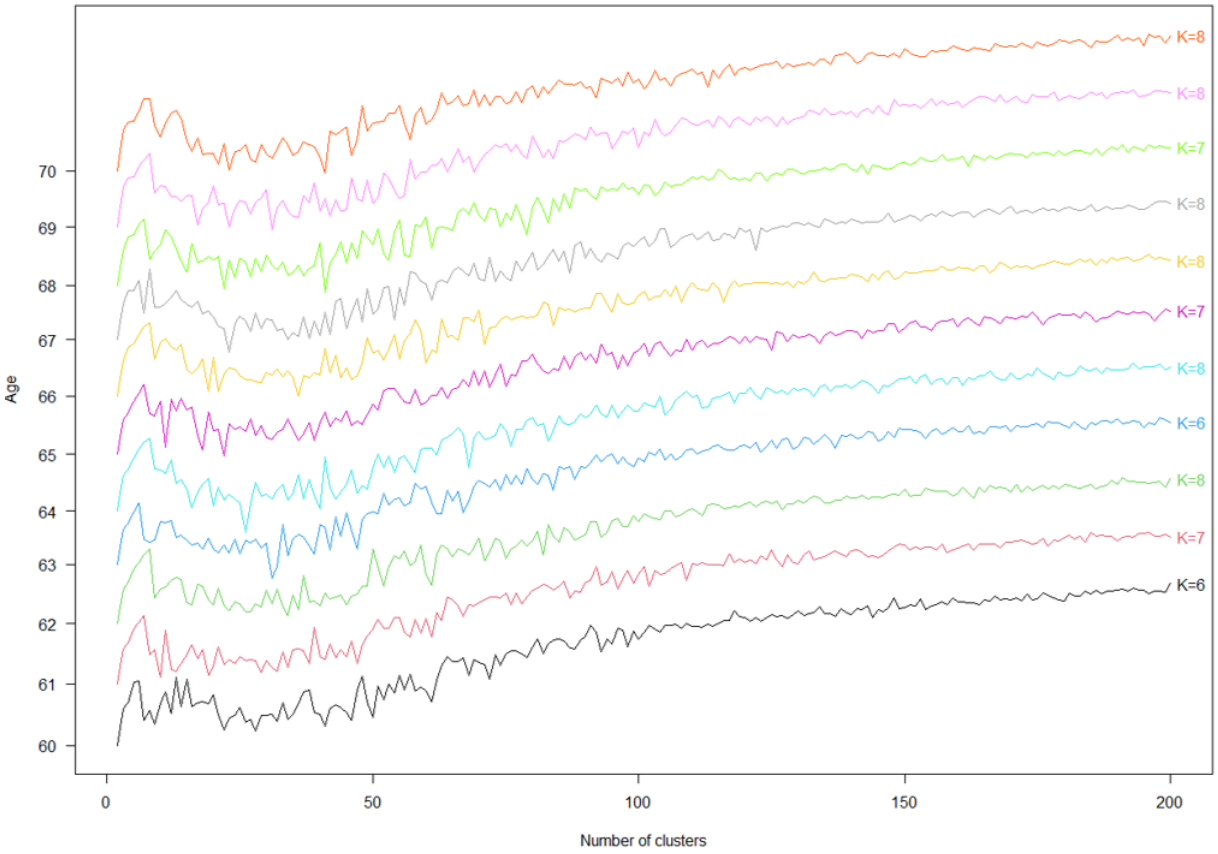


Figure S7: Silhouette score used to determine the optimal number of clusters at each age in the raw-data-based cluster-tracking approach

We calculated the silhouette score at each age, from 60 to 70 years old. We varied the number of clusters from 2 to 200. A specific optimal number of clusters K is identified at each age.

S6 Cluster-trajectories identified with the raw-data-based cluster-tracking approach

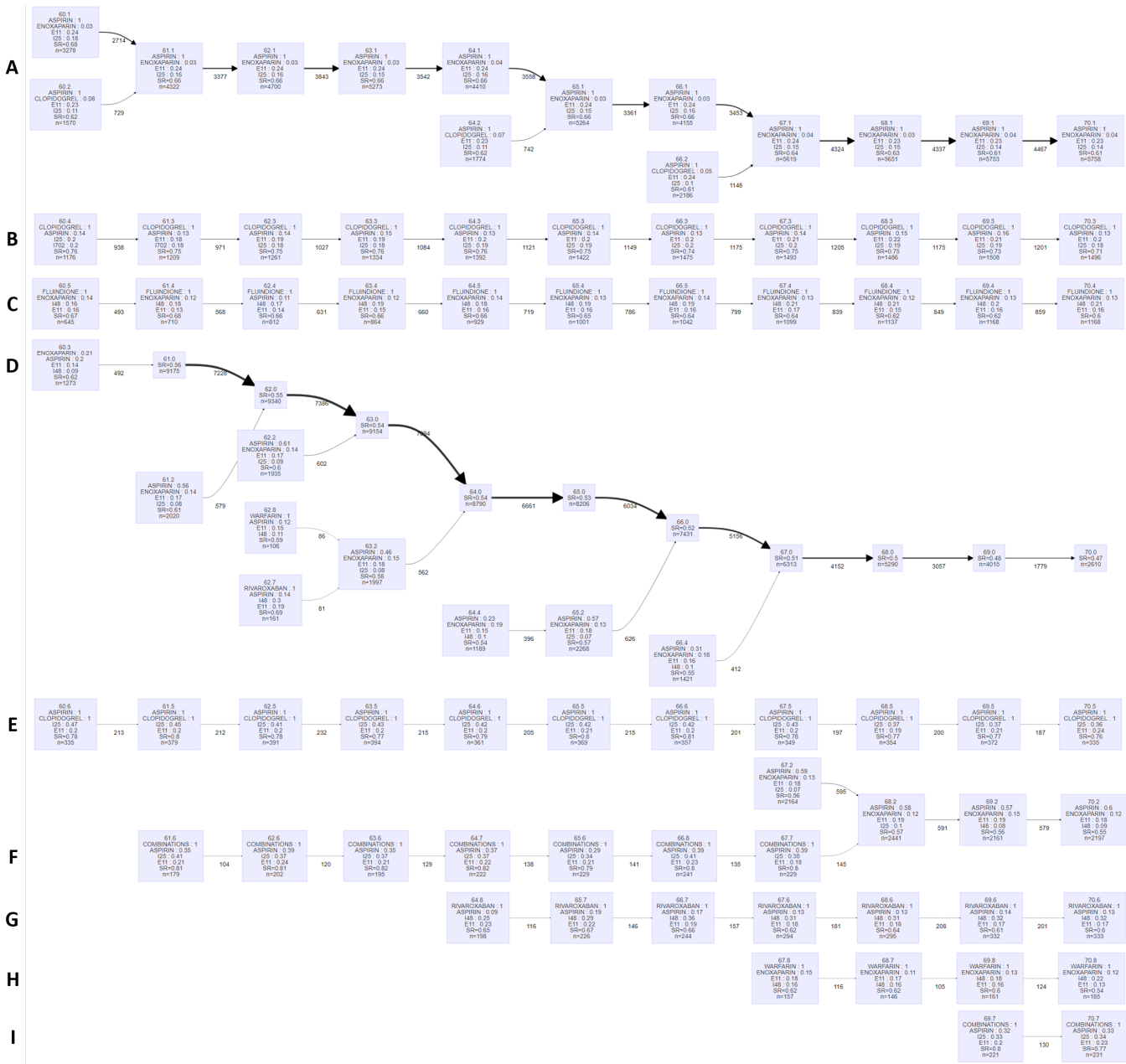


Figure S8: Cluster-trajectories in the raw-data-based cluster-tracking approach

Each block represents a cluster. Each cluster is named as follows: "x.y", with x the age time at which it was identified and y the number of the cluster. The clusters are characterized by the two most frequently reimbursed drugs (name, percentage of patients), the two most frequent long-term illnesses (ICD-10 code, percentage of patients), the sex ratio (SR) and the number of patients (n). The number under arrows is the number of patients in common between the two blocks. The arrow thickness is proportional to this number. Combinations: combinations of two platelet aggregation inhibitors.

S7 Computation time and time complexity of the different approaches

The five clustering approaches were applied in a desk computer with intel Xeon W-1270P at 3.80 GHz, 62 GB of RAM and 18 CPUs. We used parallel computing for the cluster-tracking approaches. The time complexities are those of the algorithms used, namely MCL for the network-based cluster-tracking approach, Kmeans for the raw-data-based cluster-tracking approach and the feature-based longitudinal-clustering approach, kml3d for the raw-data-based longitudinal-clustering approach and GMM for the model-based longitudinal-clustering approach. The computation time of the network-based cluster-tracking approach includes the construction of the similarity matrices. The computation time of the feature-based longitudinal-clustering approach includes the feature extraction.

	Computation time	**Time complexity**
Network-based cluster-tracking	1h50	$O(v_i^3)$
Raw-data-based cluster-tracking	20s	$O(t_i k_i n_i d_i)$
Raw-data-based longitudinal-clustering	2h37	$O(tkndI)$
Feature-based longitudinal-clustering	1min43	$O(tkndf)$
Model-based longitudinal-clustering	2 days	$O(knd^3I)$

Table S1: Computation time and time complexity of the 5 clustering approaches
i is the age, v is the number of nodes in the network, t is the number of iterations, k is the number of clusters, n is the number of patients, d is the number of drugs, I is the number of ages considered and f is the number of features extracted

S8 Cluster-tracking approaches applied to pbcseq database

S8.1 Another use-case: pbcseq

We applied our two cluster-tracking approaches to pbcseq [12], a clinical trial database. pbcseq is publicly available and contains, among other, laboratory measurements on patients with primary sclerosing cholangitis, an autoimmune disease. The clinical trial was conducted between 1974 and 1984. The laboratory measurements were collected during patient visits at 6 months, 1 year, and annually thereafter. We applied our cluster-tracking approaches to cluster patients based on the laboratory measurements using patient visits as time steps. We focused on the first six visits (i.e, 0, 182, 365, 730, 1095 and 1460 days $\pm$ 90 days) and used four different laboratory measurement variables: serum bilirubin, serum albumin, aspartate aminotransferase and standardized blood clotting time (*Table S2*). We centered and scaled these four variables to mean 0 and standard deviation 1 before applying our approaches. The dataset is composed of 312 different patients.

S8.2 Identifying cluster-trajectories with the cluster-tracking approach based on networks

We first constructed six similarity matrices, one for each visit from the first to the sixth visit, using the Cosine similarity. The similarities are computed between all patients for a given visit. Patient networks are then constructed by applying a threshold on the similarity matrices. We tested different Cosine similarity thresholds and selected a threshold of 0.6. This threshold was chosen as the best trade-off to

Patient ID	Day	Bilirubin	Albumin	Aspartate	Protime
1	0	14.5	2.60	138.0	12.2
1	192	21.3	2.94	6.20	11.2
2	0	1.1	4.14	113.5	10.6
2	182	0.8	3.60	139.5	11.0
2	365	1.0	3.55	144.2	11.6

Table S2: Extract of laboratory variables from pbcseq databases used to apply our cluster-tracking approaches

Day is the number of days between enrollment and this visit date, Bilirubin is the serum bilirubin (mg/dl), Albumin is the serum albumin (mg/dl), Aspartate is the aspartate aminotransferase (U/ml) and Protime is the standardized blood clotting time.

minimize the number of isolated patients while reducing the number of network edges (*Figure S9*). We obtained 6 patient networks, one per visit. We then applied the Markov Cluster algorithm (MCL) to identify clusters of patients. The MCL algorithm reveals 3 to 6 clusters per network. We next computed the number of common patients between clusters identified at consecutive visits, in order to track the clusters. We identified 2 cluster-trajectories composed of clusters with at least 10 patients (*Figure S10*). We then estimated the mortality rate for each trajectory. To this goal, we first computed the probability for each patient to belong to each trajectory. Then, for each trajectory, we estimated the mortality rate by the sum over all patients of the product of patient death status by the probability for the patient to belong to this trajectory. Trajectory A has an estimated mortality rate of 0.23 ± 0.05 and trajectory B has an estimated mortality rate of 0.64 ± 0.05 (*Table S3*). We then estimated, using the same approach, the values of the laboratory variables in each trajectory. The estimated value of serum bilirubin and aspartate aminotransferase were different between the two trajectories. Indeed, the estimated value of serum bilirubin and aspartate aminotransferase were, respectively, 1.71 ± 0.22 and 106.02 ± 5.32 in the trajectory A, and 6.95 ± 0.66 and 157.63 ± 6.74 in the trajectory B. Patients in trajectory B were therefore characterized by a higher risk of mortality and higher values of serum bilirubin and aspartate aminotransferase than patients of the trajectory A.

	Mortality rate	Sex ratio	Bilirubin	Albumin	Aspartate	Protime
A	0.23 ±0.05	0.09	1.71 ±0.22	3.61 ±0.03	106.02 ±5.32	10.48 ±0.07
B	0.64 ±0.05	0.15	6.95 ±0.66	3.20 ±0.04	157.63 ±6.74	11.49 ±0.13

Table S3: Characteristics of the two cluster-trajectories identified with the network-based cluster-tracking approaches

For each trajectory, we estimated the death rate, the sex ratio, the serum bilirubin (mg/dl), the serum albumin (mg/dl), the aspartate aminotransferase (U/ml) and the standardized blood clotting time. We also represented, for each value, the 95% confidence interval.

S8.3 Identifying cluster-trajectories with the cluster-tracking approach based on raw data

We applied a Kmeans per patient visit, one for each visit from the first to the sixth visit. For each Kmeans, we calculated the silhouette score and identified an optimal number of clusters (*Figure S11*). The same optimal number of 2 clusters was identified for each visit. We then tracked the identified clusters over the visits. We identified 2 cluster-trajectories composed of clusters with at least 10 patients

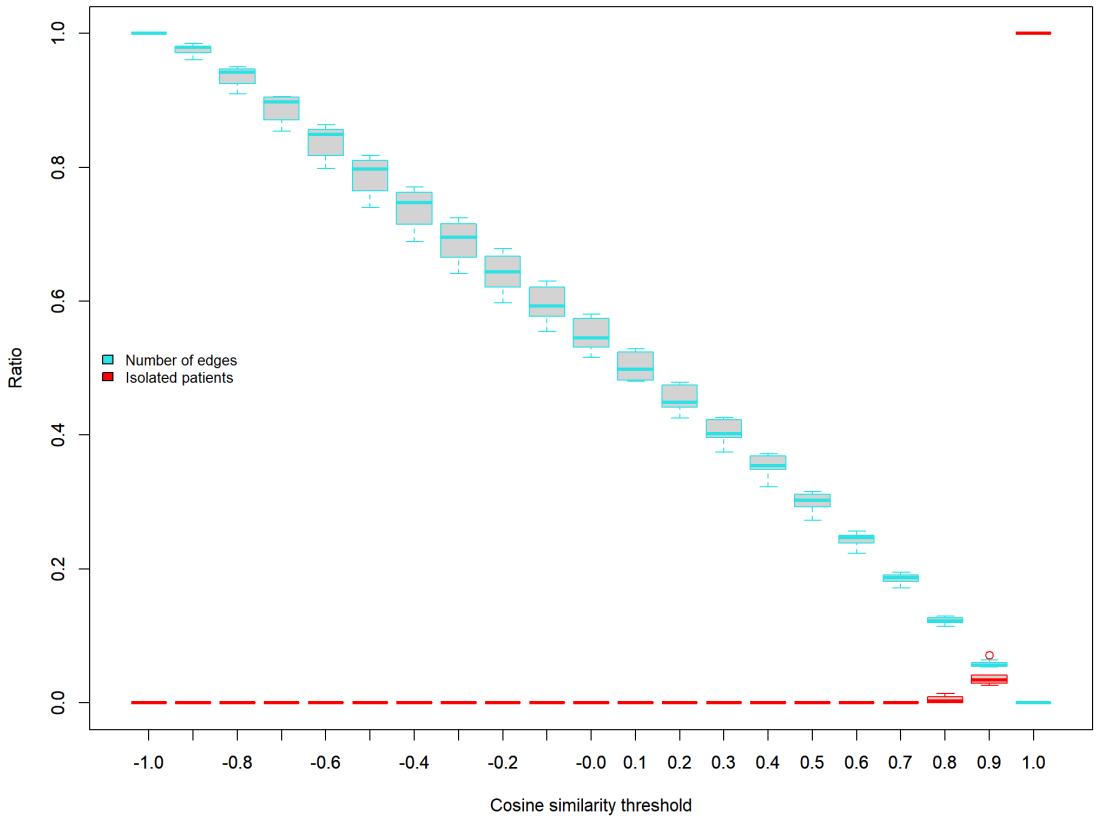


Figure S9: Choice of the Cosine similarity threshold from pbcseq database
For each value of Cosine similarity, the blue box represents the number of edges in each patient network from the first to the sixth visit, and the red box represents the isolated patients (i.e., patient connected to none of the other patients)

(*Figure S12*). The clusters of each trajectory were characterized by the estimated patient mortality rate. We estimated the mortality rate by the sum over all patients of the product of patient death status by the probability for the patient to belong to this trajectory. Trajectory A has a mortality rate of 0.30 ± 0.05 and trajectory B has a mortality rate of 0.79 ± 0.05 (*Table S4*). Regarding the laboratory variables, serum bilirubin and aspartate aminotransferase were different between the two trajectories. Indeed, serum bilirubin and aspartate aminotransferase have, respectively, values of 2.10 ± 0.24 and 112.26 ± 5.52 in the trajectory A and 10.42 ± 0.69 and 182.67 ± 6.55 in the trajectory B. Patients in trajectory B are therefore characterized by a higher risk of mortality and higher values of serum bilirubin and aspartate aminotransferase than patients of trajectory A. The same differences were observed between trajectories identified with the network-based cluster-tracking approach. However, the differences are higher between the two trajectories identified here with the raw-data-based cluster-tracking approach.

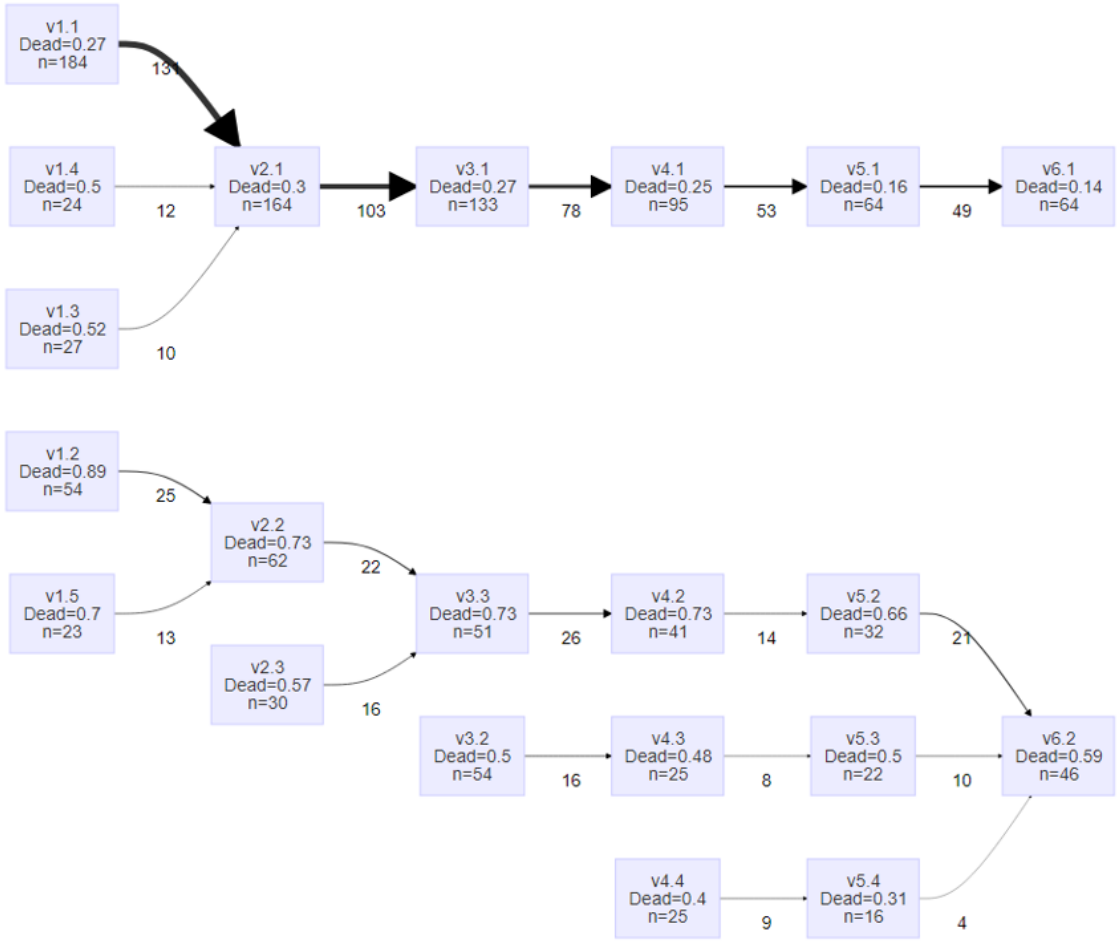


Figure S10: Cluster-trajectories identified with the network-based cluster-tracking approach from pbcseq database

Each block represents a cluster. Each cluster is named as follows: "vx.y", with vx the visit at which it was identified and y the number of the cluster. The clusters of each trajectory are characterized by the patient mortality (Dead) and the number of patients (n). The number under arrows is the number of patients in common between the two blocks. The arrow thickness is proportional to this number.

	Mortality rate	**Sex ratio**	**Bilirubin**	**Albumin**	**Aspartate**	**Protime**
A	0.30 ±0.05	0.12	2.10 ±0.24	3.53 ±0.04	112.26 ±5.52	10.62 ±0.08
B	0.79 ±0.05	0.11	10.42 ±0.69	3.08 ±0.04	182.67 ±6.55	11.92 ±0.15

Table S4: Characteristics of the two cluster-trajectories identified with the raw-data-based cluster-tracking approaches

For each trajectory, we estimated the death rate, the sex ratio, the serum bilirubin (mg/dl), the serum albumin (mg/dl), the aspartate aminotransferase (U/ml) and the standardized blood clotting time. We also represented, for each value, the 95% confidence interval.

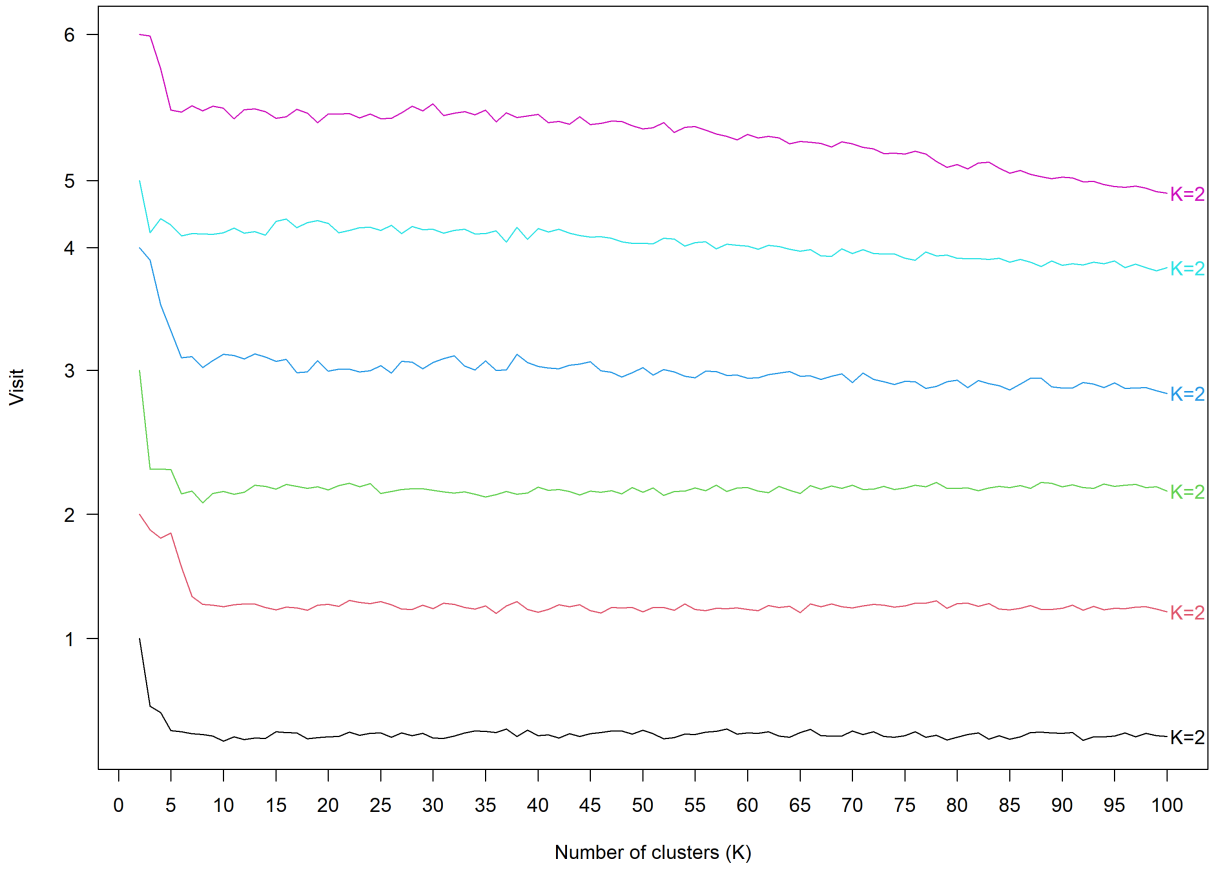


Figure S11: Silhouette score used to determine the optimal number of clusters for each patient visit in the raw-data-based cluster-tracking approach applied to pbcseq database

We calculated the silhouette score for each patient visit, from the first to the sixth visit. We varied the number of clusters from 2 to 100. An optimal number of clusters K is identified at each visit.

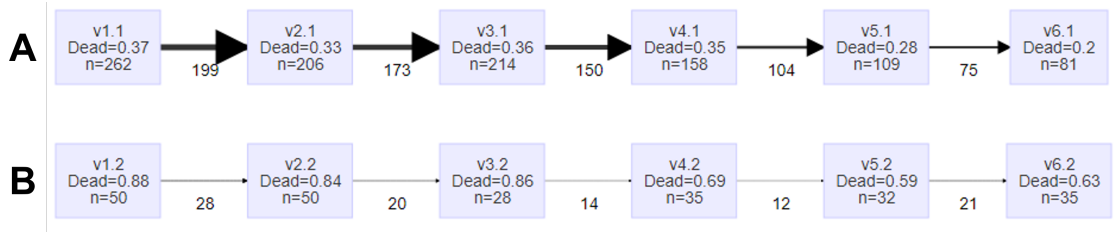


Figure S12: Cluster-trajectories identified with the raw-data-based cluster-tracking approach from pbcseq database

Each block represents a cluster. Each cluster is named as follows: "vx.y", with vx the visit patient at which it was identified and y the number of the cluster. The clusters of each trajectory are characterized by the patient mortality (Dead) and the number of patients (n). The number under arrows is the number of patients in common between the two blocks. The arrow thickness is proportional to this number.

References

[1] Amit Singhal et al. "Modern information retrieval: A brief overview". In: *IEEE Data Eng. Bull.* 24.4 (2001), pp. 35–43.

[2] Wolfram Research. *NormalizedSquaredEuclideanDistance*. 2010. URL: `https://reference.wolfram.com/language/ref/NormalizedSquaredEuclideanDistance.html`.

[3] Paul Jaccard. "The distribution of the flora in the alpine zone. 1". In: *New phytologist* 11.2 (1912), pp. 37–50.

[4] Wayne D. Blizard et al. "Multiset theory". In: *Notre Dame Journal of formal logic* 30.1 (1989), pp. 36–66.

[5] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[6] Krzysztof Kryszczuk and Paul Hurley. "Estimation of the number of clusters using multiple clustering validity indices". In: *International workshop on multiple classifier systems*. Springer. 2010, pp. 114–123.

[7] Christophe Genolini et al. "kml and kml3d: R packages to cluster longitudinal data". In: *Journal of Statistical Software* 65.4 (2015), pp. 1–34.

[8] Siddheswar Ray and Rose H Turi. "Determination of number of clusters in k-means clustering and application in colour image segmentation". In: *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*. Citeseer. 1999, pp. 137–143.

[9] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227.

[10] Hirotogu Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[11] Gideon Schwarz. "Estimating the dimension of a model". In: *The annals of statistics* (1978), pp. 461–464.

[12] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 1991.

Bibliographie

[1] URL : https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante (visité le 05/07/2023).

[2] Martin R Cowie et al. "Electronic health records to facilitate clinical research". In : *Clinical Research in Cardiology* 106 (2017), p. 1-9.

[3] Cathie Sudlow et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In : *PLoS medicine* 12.3 (2015), e1001779.

[4] Marie Aline Charles et al. "Cohort profile: the French national cohort of children (ELFE): birth to 5 years". In : *International journal of epidemiology* 49.2 (2020), 368-369j.

[5] MARIE Zins et MARCEL Goldberg. *The CONSTANCES cohort*. 2015.

[6] Graham A Colditz, Joann E Manson et Susan E Hankinson. "The Nurses' Health Study: 20-year contribution to the understanding of health among women". In : *Journal of women's health* 6.1 (1997), p. 49-62.

[7] Alaa Hamoud, Ali Salah Hashim et Wid Akeel Awadh. "Clinical data warehouse: a review". In : *Iraqi Journal for Computers and Informatics* 44.2 (2018).

[8] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In : *Scientific data* 3.1 (2016), p. 1-9.

[9] URL : https://hcup-us.ahrq.gov/nisoverview.jsp (visité le 05/07/2023).

[10] Suzanne M Cadarette et Lindsay Wong. "An introduction to health care administrative data". In : *The Canadian journal of hospital pharmacy* 68.3 (2015), p. 232.

[11] Katherine E Mues et al. "Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US". In : *Clinical epidemiology* (2017), p. 267-277.

[12] URL : https://findata.fi/en/ (visité le 16/10/2023).

[13] Yvon Merlière. "L'histoire des systèmes d'information de santé en France: connaître, comprendre pour mieux réguler la dynamique des dépenses d'Assurance maladie au pilotage par la donnée pour améliorer la santé de la population au moindre coût". In : *Journal de Droit de la Santé et de l'Assurance Maladie* 31 (2022).

[14] P1 Tuppin et al. "French national health insurance information system and the permanent beneficiaries sample". In : *Revue d'epidemiologie et de sante publique* 58.4 (2010), p. 286-290.

[15] Martin Jungkunz et al. "Secondary use of clinical data in data-gathering, non-interventional research or learning activities: definition, types, and a framework for risk assessment". In : *Journal of Medical Internet Research* 23.6 (2021), e26631.

[16] Stephane M Meystre et al. “Clinical data reuse or secondary use: current status and potential future progress”. In : *Yearbook of medical informatics* 26.01 (2017), p. 38-52.

[17] Sebastian Köhler et al. “The human phenotype ontology in 2017”. In : *Nucleic acids research* 45.D1 (2017), p. D865-D876.

[18] Armin Skrbo, Begler Begović et Selma Skrbo. “Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes”. In : *Medicinski arhiv* 58.1 Suppl 2 (2004), p. 138-141.

[19] Yi Hong et Marcia Lei Zeng. “International classification of diseases (ICD)”. In : *KO KNOWLEDGE ORGANIZATION* 49.7 (2023), p. 496-528.

[20] Kevin Donnelly et al. “SNOMED-CT: The advanced terminology and coding system for eHealth”. In : *Studies in health technology and informatics* 121 (2006), p. 279.

[21] Margaret H Coletti et Howard L Bleich. “Medical subject headings used to search the biomedical literature”. In : *Journal of the American Medical Informatics Association* 8.4 (2001), p. 317-323.

[22] Clement J McDonald et al. “LOINC, a universal standard for identifying laboratory observations: a 5-year update”. In : *Clinical chemistry* 49.4 (2003), p. 624-633.

[23] Olivier Bodenreider. “The unified medical language system (UMLS): integrating biomedical terminology”. In : *Nucleic acids research* 32.suppl_1 (2004), p. D267-D270.

[24] Mohamed Khalifa. “Health analytics types, functions and levels: a review of literature”. In : *Data, Informatics and Technology: An Inspiration for Improved Healthcare* (2018), p. 137-140.

[25] Neesha Jothi, Wahidah Husain et al. “Data mining in healthcare–a review”. In : *Procedia computer science* 72 (2015), p. 306-313.

[26] Pierre Fenaux et al. “Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study”. In : *The lancet oncology* 10.3 (2009), p. 223-232.

[27] Suchismita Mohanty et al. “Improving the efficacy of osteosarcoma therapy: combining drugs that turn cancer cell ‘don’t eat me’signals off and ‘eat me’signals on”. In : *Molecular Oncology* 13.10 (2019), p. 2049-2061.

[28] Ut T Bui, Helen Edwards et Kathleen Finlayson. “Identifying risk factors associated with infection in patients with chronic leg ulcers”. In : *International wound journal* 15.2 (2018), p. 283-290.

[29] Adrian Cassidy et al. “The LLP risk model: an individual risk prediction model for lung cancer”. In : *British journal of cancer* 98.2 (2008), p. 270-276.

[30] Paras Lakhani et Baskaran Sundaram. “Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks”. In : *Radiology* 284.2 (2017), p. 574-582.

[31] Sula Windgassen et al. *The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome*. 2018.

[32] Hosagrahar V Jagadish et al. “Big data and its technical challenges”. In : *Communications of the ACM* 57.7 (2014), p. 86-94.

[33] Tyler J Loftus et al. “Phenotype clustering in health care: a narrative review for clinicians”. In : *Frontiers in Artificial Intelligence* 5 (2022), p. 842306.

[34] Emma Ahlqvist et al. “Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables”. In : *The lancet Diabetes & endocrinology* 6.5 (2018), p. 361-369.

[35] Gengjie Jia et al. “Discerning asthma endotypes through comorbidity mapping”. In : *Nature communications* 13.1 (2022), p. 6712.

[36] Richard W Grant et al. “Use of latent class analysis and k-means clustering to identify complex patient profiles”. In : *JAMA network open* 3.12 (2020), e2029068-e2029068.

[37] Mark R Trusheim, Ernst R Berndt et Frank L Douglas. “Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers”. In : *Nature reviews Drug discovery* 6.4 (2007), p. 287-293.

[38] Jasmine Irani, Nitin Pise et Madhura Phatak. “Clustering techniques and the similarity measures used in clustering: A survey”. In : *International journal of computer applications* 134.7 (2016), p. 9-14.

[39] Amit Singhal et al. “Modern information retrieval: A brief overview”. In : *IEEE Data Eng. Bull.* 24.4 (2001), p. 35-43.

[40] Paul Jaccard. “The distribution of the flora in the alpine zone. 1”. In : *New phytologist* 11.2 (1912), p. 37-50.

[41] Archana Singh, Avantika Yadav et Ajay Rana. “K-means with Three different Distance Metrics”. In : *International Journal of Computer Applications* 67.10 (2013).

[42] Igor Melnykov et Volodymyr Melnykov. “On K-means algorithm with the use of Mahalanobis distances”. In : *Statistics & Probability Letters* 84 (2014), p. 88-95.

[43] Mohammad Norouzi, David J Fleet et Russ R Salakhutdinov. “Hamming distance metric learning”. In : *Advances in neural information processing systems* 25 (2012).

[44] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi et Teh Ying Wah. “A comparison study on similarity and dissimilarity measures in clustering continuous data”. In : *PloS one* 10.12 (2015), e0144059.

[45] Dongkuan Xu et Yingjie Tian. “A comprehensive survey of clustering algorithms”. In : *Annals of Data Science* 2 (2015), p. 165-193.

[46] J MacQueen. “Classification and analysis of multivariate observations”. In : *5th Berkeley Symp. Math. Statist. Probability*. 1967, p. 281-297.

[47] Leonard Kaufman. “Partitioning around medoids (program pam)”. In : *Finding groups in data* 344 (1990), p. 68-125.

[48] Lorena de la Fuente-Tomas et al. “Classification of patients with bipolar disorder using k-means clustering”. In : *PloS one* 14.1 (2019), e0210314.

[49] Syed Ali Abbas et al. “K-means and k-medoids: Cluster analysis on birth data collected in city Muzaffarabad, Kashmir”. In : *IEEE Access* 8 (2020), p. 151847-151855.

[50] Tian Zhang, Raghu Ramakrishnan et Miron Livny. “BIRCH: an efficient data clustering method for very large databases”. In : *ACM sigmod record* 25.2 (1996), p. 103-114.

[51] Sudipto Guha, Rajeev Rastogi et Kyuseok Shim. “CURE: An efficient clustering algorithm for large databases”. In : *ACM Sigmod record* 27.2 (1998), p. 73-84.

[52] BG Mamatha Bai, BM Nalini et Jharna Majumdar. “Analysis and detection of diabetes using data mining techniques—a big data application in health care”. In : *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2018, Volume 1*. Springer. 2019, p. 443-455.

[53] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In : *kdd*. T. 96. 34. 1996, p. 226-231.

[54] Mihael Ankerst et al. “OPTICS: Ordering points to identify the clustering structure”. In : *ACM Sigmod record* 28.2 (1999), p. 49-60.

[55] Nagesh Shukla et al. “Breast cancer data analysis for survivability studies and prediction”. In : *Computer methods and programs in biomedicine* 155 (2018), p. 199-208.

[56] Jiali Yan et al. “Applying machine learning algorithms to segment high-cost patient populations”. In : *Journal of general internal medicine* 34 (2019), p. 211-217.

[57] Joseph C Dunn. “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters”. In : (1973).

[58] Ravi Sanakal et T Jayakumari. “Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine”. In : *International Journal of Computer Trends and Technology* 11.2 (2014), p. 94-98.

[59] Raghuram Krishnapuram et James M Keller. “A possibilistic approach to clustering”. In : *IEEE transactions on fuzzy systems* 1.2 (1993), p. 98-110.

[60] Shraddha Pai et Gary D Bader. “Patient similarity networks for precision medicine”. In : *Journal of molecular biology* 430.18 (2018), p. 2924-2938.

[61] Michael M Saint-Antoine et Abhyudai Singh. “Network inference in systems biology: recent developments, challenges, and applications”. In : *Current opinion in biotechnology* 63 (2020), p. 89-98.

[62] Sarvenaz Choobdar et al. “Assessment of network module identification across complex diseases”. In : *Nature methods* 16.9 (2019), p. 843-852.

[63] Ulrike Von Luxburg. “A tutorial on spectral clustering”. In : *Statistics and computing* 17 (2007), p. 395-416.

[64] Mark EJ Newman. “Modularity and community structure in networks”. In : *Proceedings of the national academy of sciences* 103.23 (2006), p. 8577-8582.

[65] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In : *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[66] Stijn vanDongen. “A cluster algorithm for graphs”. In : *Information Systems [INS]* R 0010 (2000).

[67] Bo Wang et al. “Similarity network fusion for aggregating data types on a genomic scale”. In : *Nature methods* 11.3 (2014), p. 333.

[68] Li Li et al. “Identification of type 2 diabetes subgroups through topological analysis of patient similarity”. In : *Science translational medicine* 7.311 (2015), 311ra174-311ra174.

[69] Jon Sánchez-Valle et al. “Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships”. In : *Nature Communications* 11.1 (2020), p. 2854.

[70] Parul Agarwal, M Afshar Alam et Ranjit Biswas. “Issues, challenges and tools of clustering algorithms”. In : *arXiv preprint arXiv:1110.2610* (2011).

[71] Glenn W Milligan et Martha C Cooper. “An examination of procedures for determining the number of clusters in a data set”. In : *Psychometrika* 50 (1985), p. 159-179.

[72] Hui Xiong et Zhongmou Li. “Clustering validation measures”. In : *Data Clustering*. Chapman et Hall/CRC, 2018, p. 571-606.

[73] Tlamelo Emmanuel et al. “A survey on missing data in machine learning”. In : *Journal of Big Data* 8.1 (2021), p. 1-37.

[74] Claude Elwood Shannon. “A mathematical theory of communication”. In : *The Bell system technical journal* 27.3 (1948), p. 379-423.

[75] William M Rand. “Objective criteria for the evaluation of clustering methods”. In : *Journal of the American Statistical association* 66.336 (1971), p. 846-850.

[76] Richard E Blahut. *Principles and practice of information theory*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[77] Najmeh Akbarpour, Ebrahim Akbari et Homayun Motameni. “External clustering validity index based on extended similarity measures”. In : *Journal of Computational Science* 72 (2023), p. 102116.

[78] Matthijs J Warrens et Hanneke van der Hoef. “Understanding the adjusted rand index and other partition comparison indices based on counting object pairs”. In : *Journal of Classification* 39.3 (2022), p. 487-509.

[79] Hanneke van der Hoef et Matthijs J Warrens. “Understanding information theoretic measures for comparing clusterings”. In : *Behaviormetrika* 46.2 (2019), p. 353-370.

[80] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In : *Journal of computational and applied mathematics* (1987), p. 53-65.

[81] Godwin Ogbuabor et FN Ugwoke. “Clustering algorithm for a healthcare dataset using silhouette score value”. In : *Int. J. Comput. Sci. Inf. Technol* 10.2 (2018), p. 27-37.

[82] Tadeusz Caliński et Jerzy Harabasz. “A dendrite method for cluster analysis”. In : *Communications in Statistics-theory and Methods* 3.1 (1974), p. 1-27.

[83] Ling-Yan Ma et al. “Heterogeneity among patients with Parkinson’s disease: cluster analysis and genetic association”. In : *Journal of the neurological sciences* 351.1-2 (2015), p. 41-45.

[84] David L Davies et Donald W Bouldin. “A cluster separation measure”. In : *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), p. 224-227.

[85] Szymon Urban et al. “Novel Phenotyping for Acute Heart Failure—Unsupervised Machine Learning-Based Approach”. In : *Biomedicines* 10.7 (2022), p. 1514.

[86] Toon Van Craenendonck et Hendrik Blockeel. “Using internal validity measures to compare clustering algorithms”. In : *Benelearn 2015 Poster presentations (online)* (2015), p. 1-8.

[87] T Warren Liao. “Clustering of time series data—a survey”. In : *Pattern recognition* 38.11 (2005), p. 1857-1874.

[88] Christophe Genolini et al. “kml and kml3d: R packages to cluster longitudinal data”. In : *Journal of Statistical Software* 65.4 (2015), p. 1-34.

[89] Sarah Mullin et al. “Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes”. In : *Journal of biomedical informatics* 122 (2021), p. 103889.

[90] Tomasz Górecki et Paweł Piasecki. “An Experimental Evaluation of Time Series Classification Using Various Distance Measures”. In : *Archives of Data Science, Series A (Online First)* 5.1 (2018), p. 07.

[91] Donald J Berndt et James Clifford. "Using dynamic time warping to find patterns in time series". In : *Proceedings of the 3rd international conference on knowledge discovery and data mining*. 1994, p. 359-370.

[92] Kaat Hebbrecht et al. "Understanding personalized dynamics to inform precision medicine: a dynamic time warp analysis of 255 depressed inpatients". In : *BMC medicine* 18.1 (2020), p. 1-15.

[93] Michail Vlachos, George Kollios et Dimitrios Gunopulos. "Discovering similar multidimensional trajectories". In : *Proceedings 18th international conference on data engineering*. IEEE. 2002, p. 673-684.

[94] Verena Vogt, Stefan M Scholz et Leonie Sundmacher. "Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data". In : *The European Journal of Public Health* 28.2 (2018), p. 214-219.

[95] Carla S Möller-Levet et al. "Fuzzy clustering of short time-series and unevenly distributed sampling points". In : *International symposium on intelligent data analysis*. Springer. 2003, p. 330-340.

[96] Alex Nanopoulos, Rob Alcock et Yannis Manolopoulos. "Feature-based classification of time-series data". In : *International Journal of Computer Research* 10.3 (2001), p. 49-61.

[97] Maximilian Christ et al. "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)". In : *Neurocomputing* 307 (2018), p. 72-77.

[98] Donato Tiano, Angela Bonifati et Raymond Ng. "FeatTS: Feature-based time series clustering". In : *Proceedings of the 2021 International Conference on Management of Data*. 2021, p. 2784-2788.

[99] Daniel S Nagin et Candice L Odgers. "Group-based trajectory modeling in clinical research". In : *Annual review of clinical psychology* 6 (2010), p. 109-138.

[100] Tony Jung et Kandauda AS Wickrama. "An introduction to latent class growth analysis and growth mixture modeling". In : *Social and personality psychology compass* 2.1 (2008), p. 302-317.

[101] Aron S Downie et al. "Trajectories of acute low back pain: a latent class growth analysis". In : *Pain* 157.1 (2016), p. 225-234.

[102] Craig R Colder et al. "Identifying trajectories of adolescent smoking: an application of latent growth mixture modeling." In : *Health Psychology* 20.2 (2001), p. 127.

[103] Martin HC Law, Mario AT Figueiredo et Anil K Jain. "Simultaneous feature selection and clustering using mixture models". In : *IEEE transactions on pattern analysis and machine intelligence* 26.9 (2004), p. 1154-1166.

[104] Padhraic Smyth. “Clustering sequences with hidden Markov models”. In : *Advances in neural information processing systems* 9 (1996).

[105] Yuelin Li, Elizabeth Schofield et Mithat Gönen. “A tutorial on Dirichlet process mixture modeling”. In : *Journal of mathematical psychology* 91 (2019), p. 128-144.

[106] William Hedley Thompson, Per Brantefors et Peter Fransson. “From static to temporal network theory: Applications to functional brain connectivity”. In : *Network Neuroscience* 1.2 (2017), p. 69-99.

[107] Mingxin Gan, Xue Dou, Rui Jiang et al. “From ontology to semantic similarity: calculation of ontology-based semantic similarity”. In : *The Scientific World Journal* 2013 (2013).

[108] Zhibiao Wu et Martha Palmer. “Verb semantics and lexical selection”. In : *arXiv preprint cmp-lg/9406033* (1994).

[109] Philip Resnik. “Using information content to evaluate semantic similarity in a taxonomy”. In : *arXiv preprint cmp-lg/9511007* (1995).

[110] Ted Pedersen et al. “Measures of semantic similarity and relatedness in the biomedical domain”. In : *Journal of biomedical informatics* 40.3 (2007), p. 288-299.

[111] Jay J Jiang et David W Conrath. “Semantic similarity based on corpus statistics and lexical taxonomy”. In : *arXiv preprint cmp-lg/9709008* (1997).

[112] Claudia Leacock. “Combining local context and WordNet similarity for word sense identification”. In : *WordNet: A Lexical Reference System and its Application* (1998), p. 265-283.

[113] Israel Alonso et David Contreras. “Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach”. In : *Expert Systems with Applications* 44 (2016), p. 386-399.

[114] Bridget T McInnes, Ted Pedersen et Serguei VS Pakhomov. “UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity”. In : *AMIA annual symposium proceedings*. T. 2009. American Medical Informatics Association. 2009, p. 431.

[115] Vijay N Garla et Cynthia Brandt. “Semantic similarity in the biomedical domain: an evaluation across knowledge sources”. In : *BMC bioinformatics* 13.1 (2012), p. 1-13.

[116] Jiazhi Ni et al. “Fine-grained patient similarity measuring using deep metric learning”. In : *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, p. 1189-1198.

[117] Dominic Girardi et al. “Using concept hierarchies to improve calculation of patient similarity”. In : *Journal of biomedical informatics* 63 (2016), p. 66-73.

[118] Zheng JIA et al. “Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity”. In : *BMC medical informatics and decision making* 19 (2019), p. 1-11.

[119] Jui-Hung HUNG et al. “Gene set enrichment analysis: performance evaluation and usage guidelines”. In : *Briefings in bioinformatics* 13.3 (2012), p. 281-291.

[120] URL : https://www.inserm.fr/actualite/programme-transversal-variabilite-genomique-2018-ouverture-appel-projets/ (visité le 05/07/2023).