Analyse de données longitudinales, causalité et parcours de soin : une application aux bases médico-administratives françaises

Camille Nevoret

université PARIS-SACLAY

Analyse de données longitudinales, causalité et parcours de soin : une application aux bases médico-administratives françaises

Longitudinal data analysis, causality and care pathways : application to French nationwide claims database

Thèse de doctorat de l'université Paris-Saclay

École doctorale (n°574) de Mathématique Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques, Référent : Université d'Évry Val d'Essonne

Thèse préparée dans le laboratoire de Mathématiques et Modélisation d'Évry (Université Paris-Saclay, CNRS, Univ Evry) et le Centre de Recherche des Cordeliers, sous la direction d'**Agathe GUILLOUX**, professeure des universités, la co-direction de **Sandrine KATSAHIAN**, professeure des universités-praticienne hospitalier et la co-supervision de **Stéphane BOUÉE**, docteur

Thèse soutenue à Paris-Saclay, le 25 Avril 2023, par

Camille NEVORET

Composition du jury

Membres du jury avec voix délibérative

Christophe Ambroise Professeur, Université d'Evry Val d'Essonne	Président
Olivier Bouaziz Maître de conférence, Université de Paris	Rapporteur
Vivan Viallon Maître de conférence, Université Lyon 1, IARC	Rapporteur
Cécile Proust-Lima Directrice de recherche, INSERM Université de Bordeaux	Examinatrice

THESE DE DOCTORAT
NNT : 2023UPASM006

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Analyse de données longitudinale, causalité et parcours de soin : une application aux bases médico-administratives françaises

Mots clés : analyse causale, ATT, parcours de soin, données longitudinales, SNDS

Résumé : Le domaine de la santé connaît depuis plusieurs années une augmentation importante des données disponibles. En France, de nouvelles plateformes ont vu le jour pour centraliser les données de santé, en commençant par les données des bases médico-administratives. Initialement recueillies pour des besoins tarifaires, ces données sont une grande opportunité pour l'études en vie réelle des consommations de soin et leur évolution dans le temps.

L'efficacité des traitements est une problématique récurrente dans les études sur données de santé. Notre première contribution porte sur un modèle d'estimation de l'effet moyen du traitement chez les patients traités (ATT) en présence de facteurs de confusion dépendants du temps. Nous avons proposé une estimation débiaisée de l'ATT basée sur une généralisation du modèle de Gran en présence de plusieurs facteurs de confusion continus et dépendants du temps et d'un résultat qui peut être répétés dans le temps, comme des réhospitalisations. Dans un second temps, une extension multivariée du modèle INGARCH a été proposée pour prendre en compte des facteurs de confusion discrets dans l'estimation de l'ATT.

L'étude des parcours de soin est une autre thématique très souvent étudiée sur les données de vie réelle. Nous proposons dans nos travaux une procédure d'analyse des parcours de soin hospitalier afin d'évaluer leur association avec la mortalité. Des méthodes d'analyse de séquences prenant en compte l'ordre de la survenue des hospitalisations ont été combinées. Cette procédure d'analyse a été appliquée à l'étude de l'association des hospitalisations cardiovasculaires avec le décès chez des patients atteints d'insuffisance cardiaque.

Title : Longitudinal data analysis, causality and care pathways : application to French nationwide claims database

Keywords : causal analysis, ATT, care pathway, longitudinal data, SNDS

Abstract : For several years, the health field has seen a significant increase in the amount of data available. In France, new platforms have been created to centralize health data, starting with data from claims database. Initially collected for pricing purposes, these data are a great opportunity for real-life studies of healthcare consumption and its evolution over time.

Treatment effectiveness is a recurrent issue in health data studies. Our first contribution concerns a model for estimating the average treatment effect in treated patients (ATT) in the presence of time-dependent confounders. We proposed an unbiased estimate of ATT based on a generalization of Gran's model in the presence of several continuous, time-dependent confounders and a repeatable outcome, such as rehospitalizations. A simulation study demonstrated the value of the correction and the model was applied to MIMIC-III data. In a second step, a multivariate extension of the INGARCH model was proposed to take into account discrete confounders in Gran's model.

The study of care pathways is another issue often studied on real-life data. We propose a process to analysed hospital care pathways in order to assess their association with mortality. Sequence analysis methods taking into account the order of occurrence of hospitalizations have been combined. This analysis process was applied to the study of the association of cardiovascular hospitalisations with death in patients with heart failure.

Remerciements

J'aimerais tout d'abord remercier mes directrices de thèse, Agathe Guilloux et Sandrine Katsahian, pour m'avoir permis de mener à bien ce travail de thèse. Merci de m'avoir accompagnée dans ce projet, merci pour vos conseils, vos partages de connaissances et vos précieuses relectures.

Je tiens à remercier Monsieur Vivan Viallon et Monsieur Olivier Bouaziz pour avoir accepté de juger mon travail de thèse en tant que rapporteurs. Je remercie également Madame Cécile Proust-Lima et Monsieur Christophe Ambroise pour leur participation au jury de cette thèse en tant qu'examinateur.

Je remercie la direction de CEMKA et en particulier Stéphane Bouée et Corinne Emery pour m'avoir accueillie et pour la confiance qu'ils m'ont accordé tout au long de cette thèse. Je tiens également à remercier tous mes collègues et tout particulier l'équipe Statistique pour leur accueil, leur bienveillance et leur soutien.

Je ne veux pas oublier mes anciens collègues de l'équipe de l'unité de recherche clinique de l'hôpital européen Georges-Pompidou pour nos différents échanges et collaborations très enrichissantes.

Je remercie enfin ma famille et mes amis pour leur encouragement. Merci à mes parents, mon frère et ma sœur pour leur soutien sans faille et pour leurs précieux conseils. Alexandre, je te remercie tout particulièrement pour ta patience tout au long de ces années de thèse. Merci d'avoir été là dans les bons et les moins bons moments et de m'avoir encouragée lorsque j'étais prise de doutes.

Valorisation scientifique

Articles en lien avec la thèse

— Article en révision dans le journal Statistical Methods in Medical Research :
Nevoret, C., Katsahian, S. & Guilloux, A. (2022). Debiasing the estimate of treatment effect on the treated with time-varying counfounder. *Statistical methods in medical research* (Chapitre 3)
— Article en révision dans le European journal of heart failure :
Nevoret, C., Tran, Y., Guendouz, S., Lavenu, A., Katsahian, S., Damy, T. & Tropeano, AI. (2022). Cardiovascular healthcare trajectories and mortality in heart failure. *European heart journal* (Chapitre 5)

Articles en collaboration

— **Nevoret, C.**, Gervaise, N., Delemer, B., Bekka, S., Detournay, B., Benkhelil, A., Bahloul, A., d'Orsay, G., Penfornis, A. (2023) The Effectiveness of an App (Insulia) in Recommending Basal Insulin Doses for French Patients With Type 2 Diabetes Mellitus : Longitudinal Observational Study. *JMIR Diabetes*.
— Renard, E.,**Nevoret, C.**, Borot, S., Delemer, B., Mohammedi, K., Sultan, A., ... & Penfornis, A. (2022). Prise en charge du diabète de type 1 chez les adultes en France : l'étude SAGE. *Médecine des Maladies Métaboliques*.
— Escudier, B., de Zélicourt, M., Bourouina, R., **Nevoret, C.**, & Thiery-Vuillemin, A. (2022). Management and Health Resource Use of Patients With Metastatic Renal Cell Carcinoma treated With Systemic Therapy Over 2014-2017 in France : A National Real-World Study. *Clinical Genitourinary Cancer*.
— Harrow, B., Fagnani, F., **Nevoret, C.**, Truong-Thanh, X. M., de Zélicourt, M., & de Mestier,

L. (2022). Patterns of Use and Clinical Outcomes with Long-Acting Somatostatin Analogues for Neuroendocrine Tumors : A Nationwide French Retrospective Cohort Study in the Real-Life Setting. *Advances in Therapy*, 39(4), 1754-1771.

— Benhamouda, N., Sam, I., Epaillard, N., Gey, A., Phan, L., Pham, H. P., ..., **Nevoret, C.**, ... & Tartour, E. (2022). Plasma CD27, a surrogate of the intratumoral CD27-CD70 interaction, correlates with immunotherapy resistance in renal cell carcinoma. *Clinical Cancer Research*.

— Lampros, A., Montardi, C., Journeau, L., Georgin-Lavialle, S., Hanslik, T., Dhote, R., ..., **Nevoret, C.**, ... & Steichen, O. (2020). Association des comorbidités psychiatriques avec la durée de séjour des patients en médecine interne d'aval des urgences. *La Revue de Médecine Interne*, 41(6), 360-367.

— Pieragostini, R., Perrin, G., **Nevoret, C.**, Amar, L., Jannot, A. S., Sabatier, P., ... & Sabatier, B. (2020). Conditional prescriptions of oral antihypertensive drugs for the management of hypertension urgencies in the inpatient setting : An observational study. *Journal of Clinical Pharmacy and Therapeutics*, 45(2), 282-289.

— Foult, J. M., Katsahian, S., **Nevoret, C.**, Hoffman, O., Sabouret, P., Attal, B., & Friedlander, G. (2019). P1529 Coronary and aortic atheroma are not identical diseases : A calcium score comparative study in 1010 patients with a normal SPECT. *European Heart Journal*, 40(Supplement 1), ehz748-0291.

— Commereuc, M., **Nevoret, C.**, Radermacher, P., Katsahian, S., Asfar, P., & Schortgen, F. (2019). Hyperchloremia is not associated with AKI or death in septic shock patients : results of a post hoc analysis of the "HYPER2S" trial. *Annals of intensive care*, 9(1), 1-9.

— Krane-Gartiser, K., Scott, J., **Nevoret, C.**, Benard, V., Benizri, C., Brochard, H., ... & Etain, B. (2019). Which actigraphic variables optimally characterize the sleep-wake cycle of individuals with bipolar disorders ?. *Acta Psychiatrica Scandinavica*, 139(3), 269-279.

— Prunas, C., Krane-Gartiser, K.,**Nevoret, C.**, Benard, V., Benizri, C., Brochard, H., ... & Etain, B. (2019). Does childhood experience of attention-deficit hyperactivity disorder symptoms increase sleep/wake cycle disturbances as measured with actigraphy in adult patients with bipolar disorder ?. *Chronobiology International*, 36(8), 1124-1130.

— **Nevoret, C.**, Jannot, A. S., & Pallet, N. (2019). Clinical and Pharmacological Aspects of Hospital-Acquired Acute Kidney Injuries Outside the Intensive Care Unit : A Phenome-Wide Association Study. *Kidney Diseases*, 5(4), 272-280.

— Clement, O., Dewachter, P., Mouton-Faivre, C., **Nevoret, C.**, Guilloux, L., Morot, E. B., ... & Zins, M. (2018). Immediate hypersensitivity to contrast agents : the French 5-year CIRTACI study. *EClinicalMedicine*, 1, 51-61.

Communications orales

— EMOI 2022, **Nevoret, C.**, Tessier, C., Laurendeau, C., Voinot, C., Kab, S., & Goldberg, M. (2022). Apports et limites du «machine learning» dans la prédiction du changement du stade de sévérité de l'asthme en France : une analyse du Système national des données de santé (SNDS). *Revue d'Épidémiologie et de Santé Publique*, 70, S6.

— ISCB 2020, **Nevoret, C.**, Katsahian, S. & Guilloux, A. (2022). Estimating causal effects from the large observational, with an application to multiple sclerosis.

Poster en collaboration

— Vataire, A. L., **Nevoret, C.**, Bouée, S., Duvivier, A., & Coppo, P. (2022). POSB4 Effectiveness of Caplacizumab on Reducing Acute Mortality in Acquired Thrombotic Thrombocytopenic Purpura : Results from a French National Registry (CNR-MAT). *Value in Health*, 25(1), S25.

— Leo, M., Buleux, E., Le-Tutour, A., Duflos, A. S., Macabiau, C., **Nevoret, C.**, & Duburcq, A. (2021). L'impact des traitements immunosuppresseurs sur la vie des patients transplantés : une étude construite par et pour les patients. *Revue d'Épidémiologie et de Santé Publique*, 69, S97.

— Molins, M., Duburcq, A., **Nevoret, C.**, de Durat, G., & Molinier, G. (2021). Sclérose en plaques et Covid19 : une étude construite avec les patients (es) pour comprendre l'impact de la crise sanitaire. *Revue d'Épidémiologie et de Santé Publique*, 69, S95.

— **Nevoret, C.**, Benhamou, P. Y., Cariou, B., Fontaine, P., Patel, S., Petite, J. Z., & Detournay, B. (2020). PDB20 Cost of Cardiovascular Disease in TYPE2 Diabetes in France. a Nationwide Claims DATA Analysis. *Value in Health*, 23, S508-S509.

Table des matières

Table des figures

Liste des tableaux

Acronymes

ALD Affections de Longue Durée

ATE Effet total moyen du traitement ou Average Treatment Effect

ATIH Agence Technique de l'Information sur l'Hospitalisation

ATT Effet moyen du traitement chez les patients traités ou Average Treatment effect on Treated

CCAM Classification Commune des Actes Médicaux

CHU Centre Hospitalier Universitaire

CIM-10 Classification Internationale des Maladies version 10

CNIL Commission Nationale de l'Informatique et des Libertés

CépiDC Centre d'épidémiologie sur les causes médicales de Décès

DAG Graphe orienté acyclique ou Directed Acyclic Graph

DAS Diagnostics Associés

DP Diagnostic Principal

EDS Entrepôts de Données de Santé

EGB Échantillon du Système National des Données de Santé

EGB Échantillon Généraliste des Bénéficiaires

GHM Groupes Homogènes de Malades

HDH Health Data Hub

IC Insuffisance Cardiaque

IPCW Pondération par la probabilité inverse d'être censuré ou Inverse Probability of Censoring Weighting

IPTW Pondération par la probabilité inverse d'être traité ou Inverse Probability Treatment Weighting

MCO Médecine-Chirurgie-Obstétrique

NIR Numéro d'Inscription au Répertoire ou numéro de sécurité sociale

PMSI Programme de Médicalisation des Systèmes d'Information

RSS Résumés de Séjour Standardisé

RUM Résumé d'Unité Médicale

SEP Sclérose En Plaques

SNDS Système National des Données de Santé

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

SNM Modèles structuraux emboîtés ou Structural Nested Model

T2A Tarification A l'Activité

Chapitre 1

Introduction

Ce manuscrit présente le développement d'algorithmes pour l'analyse des données des bases médico-administratives qui, en France, ont la particularité de couvrir presque toute la population et qui, après autorisation, sont librement accessibles à des fins de recherche. Ces deux aspects les distinguent d'autres sources de données comme celles des essais cliniques et présentent une alternative importante pour la recherche dans le domaine de la santé. Ces données seront présentées dans la section 1.1. Les parties 1.2 et 1.3 présenteront rapidement les deux problématiques étudiées dans ces travaux.

1.1 - Les données de santé

Le domaine de la santé connaît depuis plusieurs années une augmentation importante des données disponibles. Ce volume de données a été multiplié par plus de 30 entre 2010 et 2020, en passant de 2 à 64 zettaoctets [85]. Des modèles de prévision estiment qu'en 2025, celui-ci devrait dépasser les 180 zettaoctets [85]. Cette augmentation est notamment rendue possible par le développement des outils techniques de génération de données comme les systèmes de séquençage à haut débit conduisant aux données « omiques » (i.e (méta)génomique, (méta)transcriptomique...) et de stockage. Le volume de données d'e-santé double tous les 73 jours dans le monde [172].

L'augmentation du volume de données de santé et la volonté de pouvoir les interroger constituent un enjeu important tant sur les aspects techniques (stockage et interopérabilité des données) que réglementaires. En France, les autorités s'intéressent de plus en plus à cette problématique. En 2018, suite au rapport Villani, a été créé le Health Data Hub (HDH) ayant pour objectif de centraliser les données de santé françaises dans un seul et unique entrepôt de données, voir [193]. En fin d'année 2021, la Commission nationale de l'informatique et des libertés a édité un référentiel sur les Entrepôts de Données de Santé (EDS) spécifiant ainsi le cadre juridique et technique de leur constitution. Enfin, un appel à projet de 50 millions d'euros est prévu par le plan France 2030 pour la constitution d'EDS au sein des différents centres hospitaliers.

Parmi les données de santé, les données recueillies dans le cadre d'essais cliniques sont à différencier

de celles recueillies dans les registres ou en routine à l'hôpital. Ces dernières sont souvent accessibles pour des travaux de recherche moyennant des démarches réglementaires contrairement aux données d'essais cliniques, souvent propriétés des laboratoires. Elles permettent notamment d'étudier des populations plus larges puisqu'elles sont sélectionnées sur des critères moins stricts, et d'avoir une vision des pratiques de soin dans des conditions de vie réelle. En France, les premiers registres épidémiologiques ont vu le jour au milieu des années 1970. Plus récemment, différents centres hospitaliers ont eu l'autorisation de constituer un EDS permettant de collecter leurs données pour les réutiliser dans des projets de recherche ou à des visées d'évaluation et de suivi d'activité. C'est par exemple le cas de l'Assistance Publique des Hôpitaux de Paris en 2017 [53], du Centre Hospitalier Universitaire (CHU) de Nantes en 2018 [54], des CHU de Lilles [55] et Grenoble [56] en 2019 ou du CHU de Rennes en 2020 [57]. Ces bases de données peuvent souffrir des spécificités géographiques ou d'un manque de diversité des pathologies représentées, liées aux priorités nationales ou de santé publique favorisant le développement de certains registres spécifiques. Il existe un autre type de données largement disponibles en France via le Health Data Hub que nous décrirons dans le paragraphe suivant.

1.1.1 . Les bases de données médico-administratives françaises

La France possède depuis plusieurs années un ensemble de bases médico-administratives alimentées par les remboursements des consommations de soin par l'Assurance Maladie. Initialement constituées dans un but de suivi d'activité et de comptabilité, elles sont de plus en plus utilisées à des visées de recherche.

Aujourd'hui l'ensemble de ces bases sont accessibles via le Système National des Données de Santé (SNDS) créer par la loi de modernisation du système de santé [124]. Ce dernier regroupe les données du Système National d'Information Inter-Régime de l'Assurance Maladie (SNIIRAM), du Programme de Médicalisation des Systèmes d'Information (PMSI), du Centre d'épidémiologie sur les causes médicales de Décès (CépiDC) et sera alimenté prochainement par les données médico-sociales liées au handicap, fournies par les maisons départementales des personnes handicapées et un échantillon de données en provenance des organismes d'Assurance Maladie complémentaire (Figure 1.1). Ces différentes bases de données seront décrites dans les paragraphes suivants. Elles peuvent être chaînées sur la base du numéro d'inscription au répertoire (numéro de sécurité sociale) et rendues pseudonimisées par la procédure FOIN (Fonction d'occultation des identifiants nominatifs), on réfère le lecteur à [151] pour une présentation spécifique de cette procédure. L'accès à ces données nécessite des démarches réglementaires qui débutent par une demande d'autorisation auprès du Health Data Hub. Ce dernier adresse au Comité d'Expertise pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé (CEREES) le dossier pour évaluation de la méthodologie scientifique, de la finalité du projet et du périmètre des données demandées. Enfin, la Commission Nationale de l'Informatique et des Libertés (CNIL) autorise ou non l'accès aux données.

Le PMSI Le PMSI a été mis en place en 1983 par Jean de Kervasdoué afin de suivre l'activité médicale des établissements hospitaliers et d'évaluer l'allocation budgétaire en découlant. Ce programme est généralisé en 1996 et est utilisé depuis pour le financement des hôpitaux. Aujourd'hui, il est géré par l'Agence Technique de l'Information sur l'Hospitalisation (ATIH).

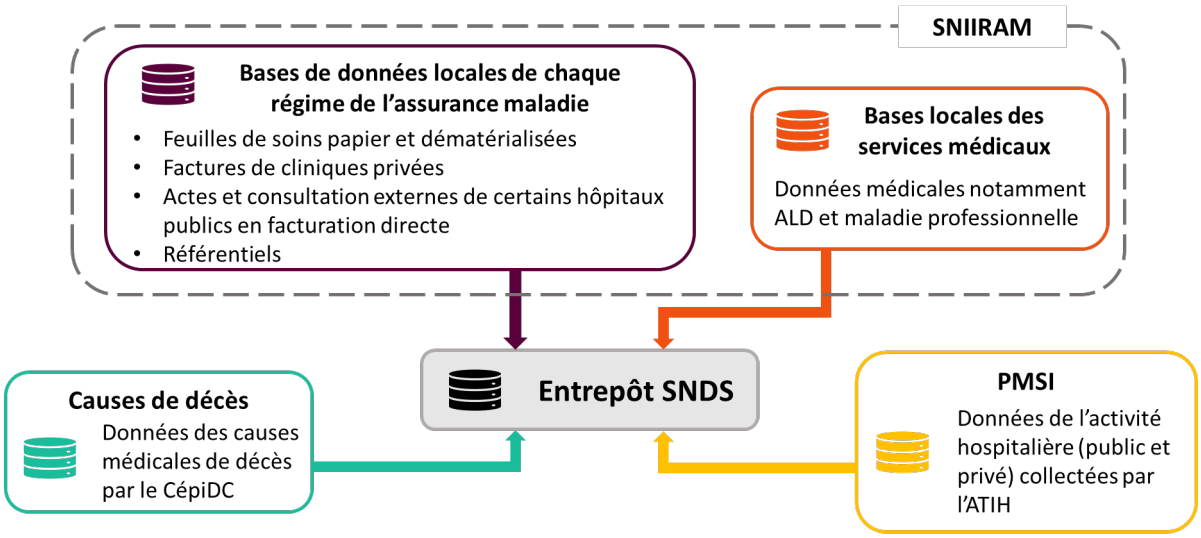

FIGURE 1.1 – Source de données actuellement disponible dans le SNDS

Le champ Médecine-Chirurgie-Obstétrique (MCO) est le premier à avoir bénéficié du PMSI. Chaque séjour dans un établissement publique ou privé en MCO est codifié selon une nomenclature spécifique appelée Groupes Homogènes de Malades (GHM) mise à jour tous les ans. Cette classification trouve sa source dans la classification américaine des Diagnosis Related Groups (DRG) définie par l'équipe du Pr. Fetter [74]. Cette dernière a été construite afin de regrouper des séjours présentant des caractéristiques proches en termes médicaux et économiques.

La classification des GHM est réalisée sur la base des Résumés de Séjour Standardisé (RSS) qui incluent un certain nombre d'informations administratives et médicales. Les informations administratives recueillies concernent le patient, comme l'âge ou le sexe, et le séjour comme sa durée. Les données médicales sont les diagnostics (principal, relié et associé) codés selon la Classification Internationale des Maladies version 10 (CIM-10) et les actes médicaux réalisés codés selon la Classification Commune des Actes Médicaux (CCAM). Ces informations sont recueillies au sein de chaque unité médicale dans lesquelles passent les patients. Un algorithme de groupage est appliqué sur les données recueillies et un GHM en résulte. Il permet ainsi de caractériser l'activité hospitalière réalisée et génère le paiement qui se fait aujourd'hui sur la base de la Tarification A l'Activité (T2A). Les tarifs associés à chaque GHM sont fixés au niveau national et publiés tous les ans par l'ATIH.

Le PMSI pour les autres champs a été mis en place progressivement, Aujourd'hui, il existe pour les soins de suite et de réadaptation, pour les hospitalisations à domicile et pour la psychiatrie.

Le SNIIRAM : Le SNIIRAM a été créé en 1998 [125] dans le but de suivre l'activité des consommations de santé et les parcours de soins du secteur libéral ainsi que leurs niveaux de dépenses et de remboursements afin de contribuer aux politiques de santé au niveau de la mise en œuvre et de l'évaluation. Il est alimenté avec les remboursements de soins par les différents régimes de l'Assurance maladie. Il couvre aujourd'hui près de 98,8% de la population française [23] issues des principaux régimes : Régime général, Régime des indépendants et la Mutualité sociale agricole.

Le SNIIRAM contient des données socio-démographiques sur les assurés, comme la date de naissance,

le sexe, le lieu de résidence (commune et département), la date de décès, le cas échéant, ainsi que le recours à la couverture médicale universelle complémentaire. Il contient aussi l'ensemble des recours aux soins par type de soin. Peuvent être retrouvées les consultations médicales et paramédicales, les délivrances de médicaments et de dispositifs médicaux, la réalisation d'actes techniques et tests biologiques et le recours à des transports médicaux. Chacun de ces soins est décrit entre autres par les dates de réalisation et de prescription, la nature du soin codifiée selon la nomenclature spécifique (comme la Nomenclature des actes de biologie médicale, la Liste des produits et prestations pour les dispositifs médicaux, ou la classification anatomique, thérapeutique et chimique pour les médicaments) et les montants remboursés par l'Assurance Maladie. Le SNIIRAM contient aussi les données du référentiel médicalisé qui recense les exonérations pour différents motifs : Affections de Longue Durée (ALD), invalidités, maladies professionnelles, accidents du travail... Les données de consommation de soins de ville peuvent être chaînées aux données du PMSI sur la base du Numéro d'Inscription au Répertoire ou numéro de sécurité sociale (NIR).

L'Échantillon Généraliste des Bénéficiaires (EGB) : En 2005, l'EGB est créé, il contient un échantillon au 1/97 du SNIIRAM. Les sujets inclus sont tirés au sort sur la base de leur NIR. Cet échantillon est représentatif de la population française pour l'âge et le sexe, cf [51]. Bien que la représentativité le soit au niveau national cette dernière n'est pas assurée au niveau régional ou départemental. La structure et la nature des données sont les mêmes que pour le SNIIRAM à l'exception du PMSI. En effet, seul le PMSI MCO est disponible. Depuis juin 2022, l'Échantillon du Système National des Données de Santé (EGB) remplace l'EGB. Il s'agit d'un échantillon à 2% de la population présente dans les bases du PMSI et SNIIRAM.

1.1.2 . Du remboursement aux analyses épidémiologiques

Comme nous l'avons vu précédemment, les bases du SNDS ont été construites, au cours du temps, dans un objectif de suivi des dépenses et de l'activité. Mais ces bases présentent un fort intérêt pour des études épidémiologiques, de parcours de soin et économiques comme en témoigne l'augmentation du nombre de publications sur ces données. Entre 2007 et 2016, plus de 400 publications sur des données des bases médico-administratives françaises ont été identifiées sur les données de MEDLINE avec une augmentation linéaire du nombre de publications depuis 2009, voir [188] pour plus de détail. L'augmentation de l'intérêt de ces bases se retrouvent dans le nombre de projets déposés au HDH pour un accès aux données du SNDS. Les études portent sur des pathologies très différentes et cherchent principalement à évaluer la prévalence et l'incidence d'une maladie [22, 33], étudier l'usage des traitements en vie réelle [204, 95, 21], étudier les parcours de soin [186] ou étudier les coûts d'une maladie [70].

Les données médico-administratives présentent l'avantage d'être presque exhaustives et individuelles. Par conséquent, l'épidémiologie au niveau national ou régional peut être évaluée, et ce pour des maladies peu fréquentes voir rares [13, 139]. Les patients sont suivis jusqu'à leur décès ou à un changement de régime d'assurance maladie qui ne serait pas présent dans le SNDS (moins de 2% de la population française). Cela réduit considérablement le nombre de perdus de vue qui peut être important dans les études avec un long suivi. Le chaînage des données de consommation hospitalière et de ville permet d'avoir une vision complète des prises en charge en vie réelle. Bien que l'accès aux données nécessite des démarches réglementaires qui représente un temps incompressible (environ 6 mois), ce dernier reste beaucoup moins important que le temps nécessaire à l'inclusion et aux suivis des patients dans une étude prospective ou un essai clinique.

Enfin, la présence du NIR dans les données facilite le chaînage entre le SNDS et d'autres sources de données comme des registres ou des cohortes. Par exemple, la cohorte Constances [88], cohorte épidémiologique de 200 000 français, dispose d'un couplage aux données du SNDS.

Bien que ces bases présentent des avantages certains pour la recherche, un certain nombre de limites sont à noter. La première concerne l'absence de données cliniques. En effet, s'il est possible d'identifier qu'un patient a réalisé une radiographie du poignet, le résultat de cette dernière (poignet cassé ou non) n'est pas renseigné. De la même façon, la délivrance d'un traitement est identifiable mais pas son indication ou sa réelle consommation. Par ailleurs certains facteurs de risque ne sont pas présents comme la consommation d'alcool ou de tabac et sont susceptibles d'introduire un biais important dans l'étude de certaines pathologies comme le cancer. Enfin, la complexité des données rend leur analyse difficile. En effet, une extraction du SNIIRAM comporte plusieurs dizaines de tables comportant elles-mêmes des centaines de variables dont des données temporelles.

L'ensemble des caractéristiques de ces bases de données nécessite de développer des outils d'analyses spécifiques pour répondre aux problématiques classiques de santé publique et d'épidémiologie. Deux thématiques vont être abordées dans cette thèse. La première s'intéresse à l'évaluation de l'efficacité d'un traitement à partir de données de vie réelle à l'aide d'un modèle causal permettant de prendre en compte le côté non aléatoire de l'attribution du traitement. La seconde se focalise sur l'étude des parcours de soins hospitaliers en proposant un procédé d'analyse permettant d'identifier les parcours de soin fréquents et d'évaluer les capacités prédictives de ces parcours sur un évènement d'intérêt, comme les décès ou les réhospitalisations.

1.2 - Thématique 1 : Analyse causale avec des données longitudinales et un critère de jugement dépendant du temps

L'autorisation de mise sur le marché d'un nouveau traitement ou d'un traitement existant pour une nouvelle indication repose en partie sur la preuve que ce traitement est sûr et efficace pour une indication, un profil de patient et une posologie donnée. Ces preuves sont apportées par les résultats obtenus lors d'essais cliniques. Ces derniers sont construits dans le but de comparer l'efficacité du nouveau traitement par rapport au traitement de référence, s'il existe, ou à un placebo. Une méthodologie spécifique est mise en place pour garantir les conditions optimales afin de démontrer l'efficacité du nouveau médicament. En ce sens la plupart de ces essais sont multicentriques (réalisés dans plusieurs hopitaux en parrallèle), en aveugle (le patient et/ou l'équipe médicale ne connaissent pas le traitement administré) et randomisés (à l'inclusion chaque patient a la même probabilité d'être traité par le nouveau traitement).

A l'issue de l'autorisation de mise sur le marché et de leur commercialisation, les médicaments continuent toujours à faire l'objet d'un suivi qui permet d'étudier les effets indésirables sur une plus grande population et sur le long terme, les pratiques et modalités d'utilisation ou encore l'efficacité dans des conditions réelles

d'utilisation et ainsi d'argumenter le maintien de ce traitement dans la stratégie thérapeutique auprès des autorités. Ces suivis sont toujours observationnels et peuvent être réalisés sur des données rétrospectives, c'est à dire déjà recueillies. De par leur construction, ces données présentent des spécificités, notamment des biais qui doivent être pris en compte lors de leur analyse. Des méthodes permettant de mettre en évidence un effet de causalité entre l'utilisation d'un traitement et le critère d'efficacité (décès, rechute, hospitalisation...) doivent être utilisées.

L'exemple d'application ayant motivé dans cette première partie est l'évaluation de l'efficacité des traitements de première ligne de la sclérose en plaques sur la fréquence des poussées.

1.2.1 . Sclérose en Plaques

La Sclérose En Plaques (SEP) est une maladie qui touche 2,8 millions de personnes dans le monde et près de 120 000 personnes en France. Chaque année, on estime que 7 à 9 personnes sur 100 000 sont nouvellement diagnostiquées. Cette maladie est la principale cause d'invalidité grave non traumatique chez les jeunes adultes, voir [118].

Il s'agit d'une maladie inflammatoire dégénérative chronique du système nerveux central d'origine auto-immune. Elle est caractérisée par une infiltration du système nerveux central par des lymphocytes T défaillants, dit autoréactifs, non éliminés par le thymus ou la moelle osseuse. Ces lymphocytes libèrent des cytokines pro-inflammatoires et provoquent des lésions des gaines de myéline des neurones. Les gaines de myéline se situent autour des nerfs et servent à les protéger et assurer une bonne vitesse de propagation de l'afflux nerveux. Les lésions sur ces gaines altèrent la bonne conduction de l'afflux nerveux. Leur localisation et leur importance conduisent à des troubles fonctionnels divers. Les lésions entrainent des déficits neurologiques à un stade précoce de la maladie qui évoluent généralement en poussées de SEP suivies de rémissions avec régression totale ou partielle des symptômes (forme récurrente rémittente SEP-RR) [40]. Toutefois, l'inflammation chronique du système nerveux central et l'accumulation de lésions qui régressent moins avec le temps conduisent, à un stade avancé de la maladie, à une dégénérescence axonale diffuse et progressive et à des déficits neurologiques irréversibles, voir [39]. La maladie peut être progressive dès le début (forme primaire progressive).

Cette maladie se déclare entre 20 et 40 ans dans 70% des cas et est prédominante chez les femmes (Ratio Homme : Femme 1 : 3). La progression et la sévérité de la maladie ne peuvent pas être prédites à l'échelle individuelle [196]. Cependant, les études de cohorte estiment que la maladie commence par des rechutes de SEP dans 85% des cas et par un handicap dans 15% des cas [159, 66]. La plupart des formes récurrentes rémittentes évoluent en une forme secondairement progressive. Les stades de l'invalidité irréversible, tels que mesurés par l'échelle EDSS (Expanded Disability Status Scale) interviendraient dans un délai moyen, respectivement, de 8, 20, 30 ans pour les échelles DSS4 (limitation de la marche), DSS 6 (marche avec canne), DSS 7 (mobilité en fauteuil roulant) [40].

Les traitements de fond actuellement disponibles ne permettent pas de guérir la SEP et ne présentent pas d'efficacité dans les formes évolutives de la maladie mais préviennent les poussées dans les formes récurrentes rémittentes [92]. Il s'agit de traitements immunomodulateurs ou immunosuppresseurs qui sont destinés à réguler la réponse immunitaire responsable des lésions inflammatoires. En ce sens, leur objectif est de réduire la fréquence des poussées et de ralentir la progression de la maladie en diminuant l'apparition de nouvelles lésions. Ces traitements peuvent être classés en 3 lignes :

— Traitements de première ligne qui sont administrés en première intention : traitements administrés par injection (interferons et acétate de glatiramère), traitements administrés par voie orale (teriflunomide et fumarate de diméthyle) ;
— Traitements de deuxième ligne qui sont administrés en cas d'échec de la première ligne ou en première intention en cas de forme grave : traitements administrés par injection (natalizumab, mitoxantrone, ocrelizumab), traitements administrés par voie orale (fingolimod, cladribine) ;
— Traitements de troisième ligne qui sont administrés en cas d'échec de la deuxième ligne ou en première intention en cas de forme grave : traitements injectables (rituximab, alemtuzumab).

1.2.2 . Données disponibles et problématique

Le problème fondamental de l'inférence causale est qu'il est impossible d'observer directement les effets causaux. Cependant, cela ne rend pas l'inférence causale impossible. Certaines hypothèses et modèles permettent de surmonter ce problème de non observation directe. Les projets présentés dans la suite ont été motivés par les spécificités des données que nous avions à notre disposition à savoir des données issues du SNDS portant sur les sujets atteints de sclérose en plaques.

Les traitements de fond de la sclérose en plaques se sont développés depuis une trentaine d'année avec la première utilisation d'interférons dans le traitement de la forme rémittente de la maladie. Ces traitements font, depuis, l'objet d'études visant a réévaluer leur efficacité au regard des nouvelles thérapies disponibles. En plus de l'évaluation de l'efficacité sur une population générale, l'émergeance de la médecine personalisée ou médecine 4P (médecine personnalisée, préventive, prédictive et participative) conduit à s'intéresser à l'efficacité d'un traitement pour chaque patient en fonction de ses caractéristiques propres. Ainsi, chaque patient pourrait bénéficier du traitement qui lui est le plus adapté.

L'étude de l'utilisation des traitements dans des conditions de vie réelle nécessite le suivi d'une population importante sur des périodes de temps suffisamment longues et ce d'autant plus quand la maladie est chronique. Les données du SNDS peuvent apporter des réponses à ces problématiques. C'est, par exemple, sur ces données que le groupe EPI-PHARE (groupement d'intérêt scientifique créé par l'Agence nationale de sécurité du médicament et des produits de santé et la Caisse nationale de l'Assurance Maladie) conduit ses analyses pharmaco-épidémiologiques pour apporter aux services publiques des réponses dans les domaines de la sécurité sanitaire. Ces données présentent notamment l'avantage d'être exhaustives de la population française avec une antériorité de plus de 15 ans.

Une étude des traitements des patients atteints de SEP a été construite sur les données du SNDS. Les patients ont été inclus entre le 1 septembre 2014 et le 31 aout 2016 si ils présentaient les critères d'inclusion suivants :

— Patient présent dans le SNDS entre 2009 et 2017 ;
— Patient âgé de 18 et plus en 2014 ;
— Patient avec une SEP sur la période d'inclusion (du 09/01/2014 au 08/31/2016). La SEP est définie par au moins l'un des critères suivant :
  — Au moins un remboursement au titre de l'affection de longue durée associé à un code diagnostic « G35 : Sclérose en Plaques » ;
  — Au moins une hospitalisation avec un diagnostic principal, relié ou associé de sclérose en plaques entre 2009 et 2013 (code diagnostic « G35 : Sclérose en Plaques »)

— Patient ayant eu une première délivrance d'un traitement de fond de première ligne sur la période d'inclusion
— Patient n'ayant pas eu de délivrance d'un traitement de fond de première ligne entre le 1er septembre 2011 et le 31 août 2014. En l'absence de traitement de fond pendant au moins 3 ans, le patient est considéré comme naïf de traitement.

Tous les patients ont été suivis jusqu'au 31 décembre 2017. Leur exposition aux différents traitements a été relevée ainsi que les différentes poussées de SEP définies par une hospitalisation pour SEP (code de diagnostic principal ou relié G35) ou au moins une délivrance de corticostéroïdes en intraveineuse.

De part sa construction à visée de remboursement, la base du SNDS ne dispose pas de données cliniques. On peut donc observer qu'un patient est traité sur une période donnée par un traitement spécifique sans connaitre les raisons qui ont motivées le choix de ce traitement. En vie réelle, les cliniciens prescrivent à leur patient le traitement qui leur semble présenter la meilleure balance bénéfice/risque. On peut donc supposer que les patients traités avec le traitement A ne présentent pas les mêmes caractéristiques que les autres. Ces caractéristiques peuvent être mesurables ou non dans les données. Par exemple dans le cas de l'analyse du teriflunomide, il est possible d'identifier certaines contre-indications, comme l'insuffisance hépatique grave identifiable par des codes diagnostics, et pas d'autres, comme la baisse importante du taux de protéine dans le sang sachant que les résultats des examens biologiques ne sont pas documentés dans la base. Des méthodes spécifiques d'analyse doivent donc être utilisées pour évaluer l'efficacité des taitements sur ce type de données comme celles permettant de n'étudier l'effet du traitement que chez les sujets traités.

Nous proposons dans ces travaux deux nouveaux algorithmes d'analyse causale pour données longitudinales. Ils permettent d'évaluer l'effet d'un traitement chez les patients traités pour lesquels on dispose de données longitudinales issue des données du SNDS.

1.3 - Thématique 2 : analyse des parcours de soin pour la prédiction d'évènement d'intérêt

L'étude des parcours de soin ou trajectoires de soin est un domaine de recherche émergent [148]. On entend par parcours de soin l'ensemble de la prise en charge d'une pathologie dont bénéficie un patient. Les éléments suivants peuvent constituer les parcours de soin : les consultations médicales, les actes techniques ou biologiques, les traitements, les hospitalisations ou d'autres prises en charges (médico-sociales ou sociales). Pour certaines pathologies, notamment dans le cadre de maladie chronique, les autorités ou collèges d'experts proposent des recommandations de prise en charge qui définissent le meilleur enchaînement et la bonne temporalité des éléments de prise en charge.

L'objectif de cette seconde partie est de proposer une méthode d'analyse des parcours de soin de patients. Elle sera illustrée sur les patients atteints d'insuffisance cardiaque permmettant d'identifier des trajectoires à fort risque de décès.

1.3.1 . Insuffisance cardiaque

En 2021, la Société Européenne de Cardiologie a mis à jour ses précédentes recommandations [150] concernant l'insuffisance cardiaque aiguë et chronique. Parallèlement à ces recommandations, la Haute Autorité de Santé préconise une prise en charge plus personnalisé des patients atteints de maladie chronique [96].

L'Insuffisance Cardiaque (IC) est une maladie qui toucherait près de 2,3% (IC95% : 2,1 - 2,5) de la population française. Sa prévalence augmente avec l'âge pour atteindre 16,3% (IC95% : 13,7 - 18,9) des sujets de 85 ans et plus [50]. Chaque année, plus de 120 000 nouveaux cas sont dénombrés [52].

L'IC est une anomalie structurelle ou fonctionnelle du cœur qui se caractérise par une difficulté à assurer correctement l'apport en oxygène aux différents organes pour répondre à leurs besoins métaboliques qui s'explique notamment par un débit cardiaque trop faible [150]. La quantité insuffisante de sang dans le corps conduit à une accumulation de liquide dans les poumons pour l'insuffisance cardiaque gauche, ou dans les jambes, la jugolaire, les organes digestifs dont le foie pour l'insuffisance droite. Les symptômes les plus fréquents sont l'essouflement, le gonflement des pieds et des jambes, et la fatigue physique.

Les causes de l'IC sont diverses mais cette dernière résulte souvent d'une pathologie cardio-vasculaire antérieure. On retrouve principalement comme causes les cardiopathies ischémiques, l'hypertension, les cardiomyopathies, les cardiopathies valvulaires ou des maladies du rythme cardiaque comme la fibrillation auriculaire. L'insuffisance cardiaque chronique est une cause fréquente d'hospitalisation notamment chez les sujets âgées. On estime que près de 160 000 patients sont hospitalisés par an [81]. Bien que les taux de décès dûs à l'insuffisance cardiaque aient fortement diminués, on estime que plus de 70 000 décès chaque année sont associés à cette pathologie.

L'IC est une maladie qui ne se guérit pas. La prise en charge consiste à traiter les causes et les symptômes par une modification du mode de vie et par traitement médicamenteux. Les traitements les plus fréquents sont les diurétiques qui visent à réduire la rétention d'eau et les œdèmes et les vasodilatateurs. On retrouve aussi l'utilisation des agents inotropes ou des vasopresseurs. Dans le cas d'une insuffisance cardiaque avancée, un traitement chirurgical peut être envisagé comme la pose d'un défibrillateur ou d'un stimulateur cardiaque.

1.3.2 . Données disponibles et problématique

Nous proposons dans ce travail une procédure d'analyse qui combine des méthodes de différents domaines afin d'utiliser les parcours de soin comme variables prédictives du décès. Des méthodes d'analyses de séquences sont utilisées pour identifier les parcours de soin fréquents et quantifier la proximité de deux parcours de soin.

Les maladies de l'appareil circulatoire représente la deuxième cause de décès sur l'ensemble de la population et la première cause de décès chez les femmes [63]. L'insuffisance cardiaque en est la cause majeure. Comme nous l'avons vu précédemment, l'IC a de multiples étiologies et symptômes ce qui rend sa prise en charge complexe et incite à la personnaliser. On cherche dans cette partie à étudier le lien entre le décès et les parcours de soin des patients atteints d'IC. L'identification de certains parcours de soin à risque pourrait permettre aux cliniciens de prendre les décisions optimales au cours de la prise en charge.

Nous avons construit une étude sur les données de l'EGB (voir section 1.1.1) pour répondre à cette problématique. Compte tenu de la prévalence importante de la maladie, l'EGB permettait d'avoir une po-

pulation suffisante pour l'analyse. Ces données ont permis par exemple d'identifier des prédicteurs à des réhospitalisations pour IC [89] et dans d'autres domaines d'identifier des événements inattendus après une chirurgie bariatrique, voir [35]. Les sujets âgés de plus de 18 ont été inclus dans l'étude s'ils avaient été hospitalisés pour insuffisance cardiaque entre le 1er janvier 2010 et le 31 décembre 2016. Les patients hospitalisés en 2008 ou 2009 ont été exclus afin d'identifier les nouveaux patients IC. Nous nous sommes focalisés uniquement sur les parcours de soin hospitaliers ayant eu lieu dans les 2 ans suivants la première hospitalisation pour IC. Une forte diversité de parcours de soin est observée ce qui complexifie leur analyse.

1.4 - Plan du manuscrit

Les travaux présentés dans ce document ont été motivés par les spécificités des données issues des bases médico-administratives françaises et notamment leur caractère longitudinal. Les deux problématiques étudiées traitent de l'analyse du décès ou la survenue d'évènements récurrents comme les réhospitalisations ou les rechutes. Le chapitre 2 présente plus en détail les deux thématiques évoquées précédement. L'analyse de survie dans le cadre de données longitudinales est présentée dans la première partie du chapitre 2 (section 2.1). Les détails méthodologiques des deux thématiques sont présentés successivement. La partie 2.2 (Thématique 1 : Analyse causale avec des données longitudinales et un critère de jugement dépendant du temps) aborde l'analyse causale et les régressions de données longitudinales. La partie 2.3 (Thèmatique 2 : Analyse des parcours de soin pour la prédiction d'évènement d'intérêt) présente les méthodes d'analyse des séquences et les modèles de prédiction dans le cas d'analyse de survie.

La suite du manuscrit se compose de trois chapitres présentant les travaux menés sur les deux thématiques exposées plus haut. Les chapitres 3 et 4 traitent de la première problématique, voir les paragraphes 2.2.1, 2.2.2.4, 2.2.2.2 (dans le chapitre 2) pour plus de détails. Une correction de estimation de l'ATT est présenté tout d'abord dans le cas de facteurs de confusion discrets et dépendants du temps. Dans un second temps, nous présenterons un modèles permettant d'estimer l'ATT corrigé en présence de facteurs de confusion discrets et dépendants du temps. Le chapitre 5 traite de la seconde problématique en proposant une procédure d'analyse des séquences d'hospitalisation pour la prédiction du décès, voir les paragraphes 2.3.1, 2.3.2 (dans le chapitre 2) pour plus de détails. Cette procédure d'analyse est appliquée à l'étude des parcours de soin des patients atteints d'insuffisance cardiaque sur le surrisque de décès.

Un dernier chapitre 6 propose une conclusion et des perspectives de développement sur les travaux présentés.

Chapitre 2

Etat de l'art

Le chapitre précédent a présenté les problématiques liées à l'utilisation des données issues des bases médico-administratives pour l'analyse de l'efficacité des traitements ou l'étude des parcours de soin en vie réelle. Le chapitre qui suit détaille les aspects méthodologiques de ces problématiques. La première partie, partie 2.1, s'inéresse particulièrement aux analyses de durée dans le contexte de données longitudinales en présentant tout d'abord les principales définitions puis deux modèles de régressions utilisés pour ces analyses. Cette partie repose sur les ouvrages de Klein et Martinussen, voir [112, 133]. La seconde partie, partie 2.2, introduit les notions de causalité et les définitions relatives, les méthodes d'analyses causales et en particulier le modèle de Gran, voir [117, 146, 90]. La dernière partie, partie 2.3, présente les notions et méthodes d'analyse utilisées dans l'analyse de séquences, ou trajectoires, ainsi que les méthodes de prédiction dans le contexte de survie (cf [128, 36, 199]).

2.1 - Analyse de durée et données longitudinales

Les études dans le domaine de la santé peuvent être de deux types : transversales ou longitudinales. Les études transversales étudient les sujets à une date donnée. Ce type d'étude est fréquent sur les données du SNDS, par exemple lorsque l'on s'intéresse aux consommations de soin sur une année donnée. Ces données permettent aussi des analyses avec un suivi dans le temps. On parle alors d'étude sur données longitudinales.

Les problématiques des études sur données longitudinales portent souvent sur l'analyse de la survenue d'un ou plusieurs évènements dans le temps. Largement utilisé pour étudier la survenue du décès, ces études peuvent aujourd'hui porter sur d'autres types de critères comme la rechute d'un cancer, les réhospitalisations, l'arrêt d'un traitement... L'une des problématiques principales de ce type d'étude réside dans le fait que tous les sujets ne présentent pas l'évènement d'intérêt sur la période de l'étude. Pour pallier ce problème, des méthodes d'analyse de durée ou analyse de survie ont été développées.

Dans un premier temps, cette partie présentera les définitions et notations nécessaires aux analyses de durée. Ensuite les deux principaux modèles de régression, ceux de Cox et d'Aalen qui sont utilisés dans mes

travaux, seront présentés.

2.1.1 . Définitions et notations

2.1.1.1 . Dates d'intérêt

Lorsqu'une étude s'intéresse à la survenue d'un évènement, les sujets sont nécessairement suivis dans le temps. Pour définir ce suivi, différentes dates sont définies.

- Le suivi du patient débute à la **date d'origine**. Dans un essai clinique, cette date peut correspondre à la date d'inclusion ou à la date de randomisation. Dans une étude observationnelle, elle peut coïncider avec la date de naissance, la date de première identification de la maladie dans la base ou encore à l'initiation d'un traitement.
- La **date de point** correspond à la date de fin de l'étude à partir de laquelle nous ne tiendrons plus compte des informations sur les sujets. Cette date est souvent fixée *a priori* au début de l'étude. Cette dernière peut dépendre de la date d'origine, par exemple lorsqu'on souhaite suivre les sujets pendant un an, ou peut être une date spécifique, par exemple tous les sujets sont suivis jusqu'au 31 décembre 2021. Après cette date, toutes les informations éventuellement disponibles ne seront pas considérées.
- En pratique, il est possible que tous les sujets ne soient pas suivis jusqu'à la date de point. La date de **dernière nouvelle** est donc la date la plus récente où une information est disponible pour le sujet. Dans les bases du SNDS, cette date correspond à la date de la dernière consommation de soin identifiée.
- La ou les **dates de survenue de l'évènement d'intérêt** sont aussi recueillies au cours du temps.
- La **date de fin de suivi** est le minimum entre la date de dernière nouvelle et la date de point.

La figure 2.1 présente les dates jalonnant le suivi de 3 patients dans une étude longitudinale. Ces patients sont inclus entre 2015 et 2017 et sont suivis jusqu'en 2020. Le patient 1 est suivi de son inclusion à la fin du suivi. L'évènement survenant après 2020 n'est pas pris en compte. Le patient 2 est suivi de son inclusion à la survenue de l'évènement. Enfin le patient 3 est suivi de l'inclusion jusqu'à la date de dernière nouvelle. Cette dernière ne coïncidant pas avec la date de fin de suivi, on dit que le patient est perdu de vue.

2.1.1.2 . Censures

Il est possible que la survenue d'un évènement ne puisse pas être observée, on parle alors de censure. C'est par exemple le cas des patients 1 et 3 de la figure 2.1. Différents types de censures existent et sont présentées dans la figure 2.2 :

- La **censure à droite** intervient lorsqu'à la fin du suivi l'évènement ne s'est pas produit. Autrement dit, le temps entre l'évènement, si ce dernier est survenu, et la date d'inclusion est supérieur au temps entre la date d'inclusion et la date de fin de suivi. Si l'on s'intéresse au décès dans les deux ans, un décès se produisant au bout de 3 ans serait censuré.

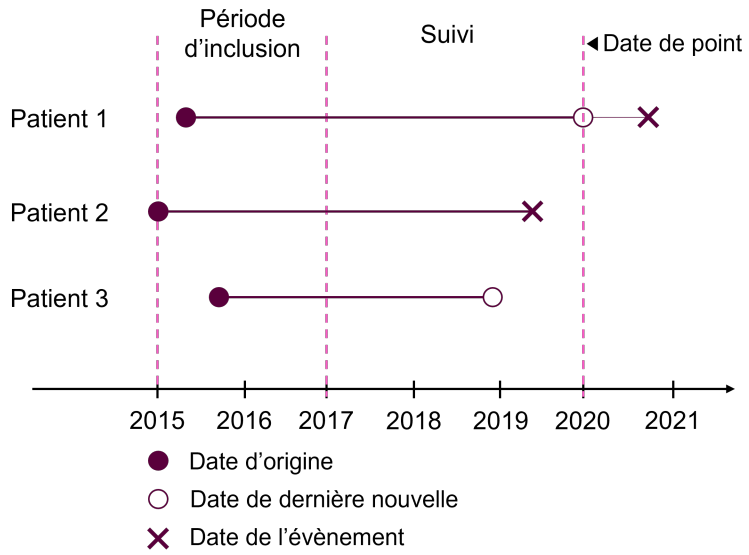

FIGURE 2.1 – Les différentes dates jalonnant le suivi de 3 patients dans une étude longitudinale.

— La **censure à gauche** intervient lorsque le temps entre l'évènement et la date d'inclusion est inférieur ou égal au temps entre la date d'inclusion et la date de fin de suivi.
— La **censure par intervalle** correspond à la survenue d'un évènement sur une période pendant laquelle le sujet n'est pas observé. Dans les bases du SNDS, cela pourrait se produire par exemple si l'évènement survenait alors que le sujet est expatrié. En effet, les Français expatriés ne dépendent plus de la Sécurité Sociale française. Toutes les consommations de soin pendant cette période ne sont pas remontées dans le SNDS.

Dans les données du SNDS, les censures par intervalles sont assez rares compte tenu du nombre de régimes présents dans les bases. Les censures à gauche sont majoritairement évitées par la construction méthodologique des études. Par exemple, on s'assure d'une période d'observation sans délivrance d'un traitement spécifique pour observer l'initiation de ce traitement. Nous nous focaliserons dans ce document uniquement sur les censures à droite.

2.1.1.3 . Evènement d'intérêt

Evènement terminal : Lorsque l'on s'intéresse à un évènement terminal comme le décès ou l'arrêt d'un traitement, on peut définir une durée de survie. Il s'agit du temps écoulé entre la date d'origine et la date de survenue de l'évènement.

On note T^E la variable aléatoire correspondant au temps entre la date d'inclusion et d'observation de l'évènement et T^C la variable aléatoire correspondant au temps de censure entre la date d'inclusion et la date de fin de suivi. La variable aléatoire correspondant aux temps de survie est définie comme suit : $T = min(T^C, T^E)$. La variable binaire suivante est associée à ce temps de survie : $E = 1$ si $T = T^E$ et $E = 0$ si $T = T^C$. Ainsi, on dispose pour tout patient i du couple (T_i, E_i).

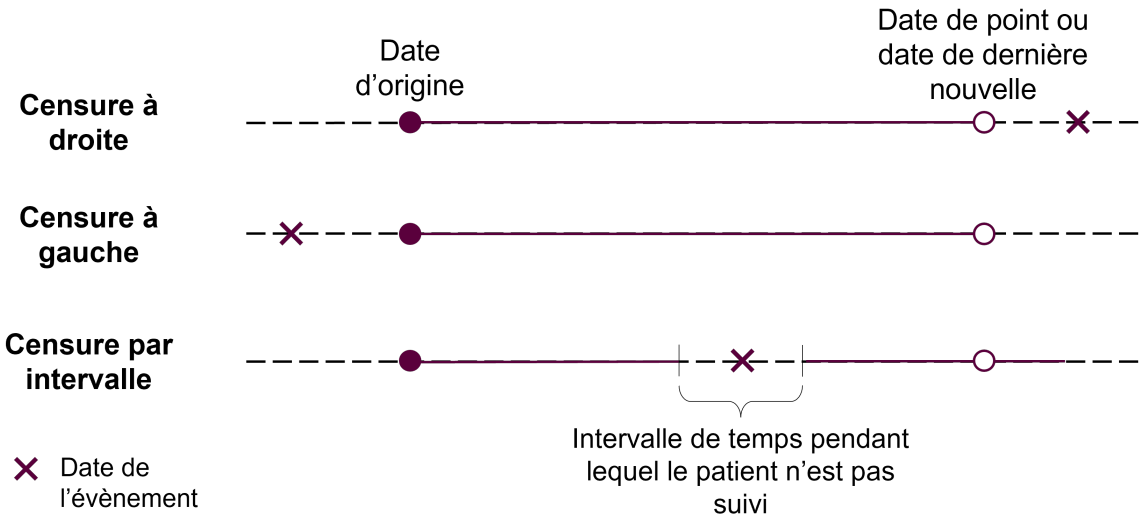

FIGURE 2.2 – Les différents types de censure

Evènements récurrents : En fonction des pathologies et des objectifs d'étude, il peut être plus pertinent d'étudier non pas un évènement mais des évènements qui peuvent se répéter dans le temps. On parle alors d'évènements récurrents. Ces évènements peuvent être par exemple des poussées de sclérose en plaque, des crises d'asthme ou des réhospitalisations pour insuffisance cardiaque. L'analyse des évènements récurrents peut être faite en étudiant le nombre d'événements ayant eu lieu sur une période de temps donnée. Par ce fait, on définit un processus de comptage. $(N(t),\ t \geq 0)$ est un processus stochastique positif qui vérifie les propriétés suivantes :

- $N(t)$ correspond au nombre d'évènements survenus entre les temps 0 et t donc $N(t) \in \mathbb{N},\ \forall t$
- $N(0) = 0$ et $N(t) < \infty$
- $N(t)$ est une fonction en escalier, non décroissante, continue à droite et qui augmente de 1 à chaque saut. La figure 2.3 présente un exemple de processus de comptage.

La filtration naturelle associée à ce processus de comptage est donnée par : $\mathcal{F} = \{\mathcal{F}(t) = \sigma(N(s), s \leq t), t \geq 0\}$. La tribu $\mathcal{F}(t)$ regroupe toutes les informations passées du processus avant le temps t.

L'intensité du processus est définie de la façon suivante, voir [3] :

$$\lambda(t)dt = \mathbb{P}[dN(t) = 1|\mathcal{F}(t-)],$$

avec $dN(t) = N([t, t+dt[) = N((t+dt)-) - N(t-)$, où $t-$ correspond aux temps infinitésimalement plus petits que t.

Le processus de comptage le plus classique est le processus de Poisson. Il s'agit d'un processus de comptage qui vérifie les propriétés suivantes :

- Le nombre d'évènement dans deux intervalles disjoints sont indépendants : $N(]t_1, t_2])$ et $N(]t_3, t_4])$ sont indépendants pour tout $t_1 \leq t_2 \leq t_3 \leq t_4$.

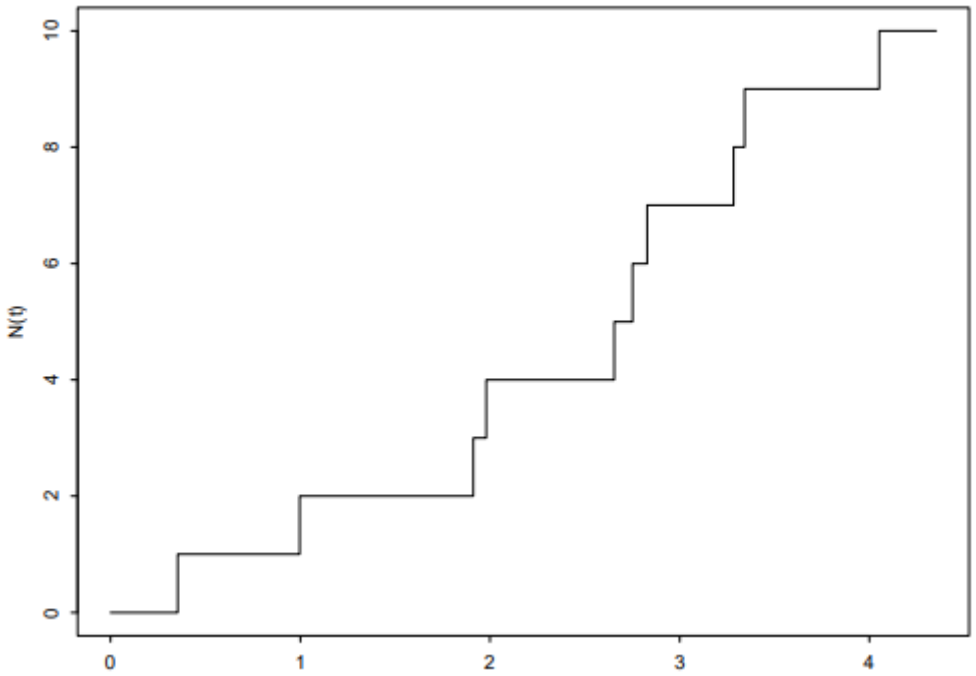

FIGURE 2.3 – Exemple de processus de comptage

— Le nombre d'évènement entre les temps 0 et t suit une loi de Poisson telle que $N(t) \sim \mathcal{P}(\int_0^t \lambda(s)ds)$.

En pratique un processus de comptage n'est observé que de la date d'origine à la date de fin de suivi qu'on notera T^C. Le processus de comptage $\{N(t), t \geq 0\}$ n'est donc pas observé mais le processus censuré l'est $N^C(t) = \{N(t \wedge T^C), t \geq 0\}$. Dans le cas avec au plus un évènement, le processus de comptage peut être noté $N(t) = \mathbb{1}_{T \leq t}$ et l'indicateur d'être à risque $E(t) = \mathbb{1}_{T^C \geq t}$. Le processus de $N^C(t) = \mathbb{1}_{T \leq t, E=1}$. L'intensité du processus censuré N^C est noté $\lambda^C(t) = E(t)\lambda(t)$. Dans la suite, on notera N le processus N^C et λ son intensité.

2.1.1.4 . Fonction de survie et fonction de risque instantané

En revenant au cas où au plus un évènement par patient peut survenir, la **fonction de survie** au temps t correspond à la probabilité que l'évènement n'ai pas été observé avant le temps t et est noté $S(t)$:

$$S(t) = \mathbb{P}[t < T], \ t \geq 0. \tag{2.1}$$

Cette fonction présente les propriétés suivantes :

— $S(0) = 1$;

— $\lim_{t \to +\infty} S(t) = 0$;

— la fonction est décroissante et continue à droite (avec limite à gauche).

La **fonction de risque instantané** peut être interprétée comme la probabilité que l'évènement survienne

dans un petit intervalle de temps juste après t conditionnellement au fait que l'évènement n'ai pas eu lieu jusqu'alors. On la note h et elle est définie de la façon suivante en $t \geq 0$:

$$h(t) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{P}[t \leq T < t + dt | T \geq t].$$

La fonction de survie peut être calculée à partir de la fonction de risque instantané par :

$$S(t) = exp(-\int_0^t h(s)ds).$$

Dans le cas d'un processus de comptage à au plus un évènement, l'intensité du processus censuré peut s'écrire : $\lambda(t)dt = E(t)h(t)dt$.

2.1.1.5 . Covariables

Lors des analyses de durée, deux types de covariables peuvent être utilisées. Nous noterons ces covariables $X(t) = (X_1(t), \cdots, X_p(t))$. On retrouve des variables qui ne dépendent pas du temps. C'est par exemple le cas du sexe, de l'âge à l'inclusion ou de toute autre variable mesurée à l'inclusion. Des variables dépendantes du temps peuvent aussi être utilisées. Il peut s'agir de mesures de biologie ou de l'exposition à un traitement dans le temps. Le traitement des données dépendantes du temps sera présenté dans la section 2.2.2.2.

2.1.2 . Modéles d'estimation

Plusieurs modèles ont été développés pour étudier des critères de jugement dépendant du temps comme la survenue d'un évènement d'intérêt. Ces modèles permettent notamment d'évaluer l'effet des covariables sur la fonction de survie. Ces modèles peuvent être paramétriques. Dans ce cas une hypothèse est faite concernant la distribution du temps T^E. Le modèle de Weibull [203] ou les modèles à temps accélérés [202] sont des exemples de modèles paramétriques.

Nous utiliserons dans la suite deux modèles semi-paramétriques, les modèles de Cox [46] et de Aalen [1, 2]. L'écriture de ces modèles et l'estimation des paramètres sont décrites plus bas.

2.1.2.1 . Le modèle de Cox

Le modèle des risques proportionnels de Cox est le modèle le plus courant pour modéliser les effets des covariables sur la survie. Il a été introduit par Cox [46] pour l'analyse de données de survie, puis étendu par Andersen et Gill [12] dans le cas d'un processus de comptage. Ce modèle propose une modélisation de l'intensité du processus qui s'écrit de la façon suivante :

$$\lambda(t) = Y(t)\lambda_0(t)exp(X^\top(t)\beta_0), \tag{2.2}$$

pour $t \leq 0$ où $X(t) = (X_1(t), \cdots, X_p(t)))$ un vecteur de covariables de dimension p, $Y(t)$ l'indicateur de risque, et $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})$ un vecteur d'inconnus. La fonction λ_0 est l'intensité de base. Elle est appelée ainsi puisqu'elle correspond à l'intensité $\lambda(t)$ lorsque toutes les covariables sont égales à 0. Elle est supposée intégrable : $\int_0^\tau \lambda_0(t)dt < \infty$ ou τ correspond à la date de point. Aucune hypothèse n'est imposée sur la structure de l'intensité de base. Elle n'est pas supposée connue

Les paramètres du modèles de Cox peuvent être estimés en maximisant la vraisemblance partielle du modèle [46, 45]. Pour la suite, supposons que l'on observe n individus entre les temps 0 et τ un temps fini. On dispose pour chacun d'eux des données suivantes : $(N_i(t), Y_i(t), X_i(t)), i = 1, \cdots, n$. On suppose que chaque N_i a pour intensité λ_i modélisée suivant le modèle de Cox 2.2. La vraisemblance partielle du modèle est donnée pour un $\beta \in \mathbb{R}^p$ par :

$$\mathcal{L}(\beta) = \prod_t \prod_{i=1}^n \left(\frac{exp(X_i^\top(t)\beta)}{\sum_{i=1}^n Y_i(t) exp(X_i^\top(t)\beta} \right)^{dN_i(t)}.$$

On peut noter que la vraisemblance partielle ne contient plus le paramètre d'intensité de base λ_0.

La log-vraisemblance pour le paramètre β peut s'écrire :

$$\log(\mathcal{L}(\beta)) = \sum_{i=1}^n \int_0^\tau X_i^\top \beta - \log(\sum_{i=1}^n Y_i(t) exp(X_i^\top \beta)) dN_i(t).$$

L'estimateur de β, noté $\hat{\beta}$ est obtenu comme résultat de l'équation $U(\hat{\beta}) = 0$.

$\hat{\beta}$ est l'estimateur de β obtenu en maximisant la log-vraissemblance partielle. L'estimateur de l'intensité de base cumulée Λ_0 définie par $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ peut être donné par l'estimateur de type Nelson-Aalen :

$$\widehat{\Lambda_0}(t, \beta) = \int_0^t \frac{1}{\sum_{i=1}^n Y_i(t) exp(X_i^\top(t)\beta} dN.(s),$$

avec $N.(t) = \sum_{i=1}^n N_i(t)$. Avec l'équation précédente et l'estimateur de β, l'estimateur de Breslow [31] s'écrit : $\widehat{\Lambda_0}(t) = \widehat{\Lambda_0}(t, \hat{\beta})$.

Le modèle de Cox fait plusieurs hypothèses fortes. Par construction le modèle impose que :

— les risques relatifs ne dépendent pas du temps, on parle d'hypothèse des risques proportionnels. Pour simplifier, considérons un vecteur de covariables non dépendantes du temps, le risque relatif est donné par le ratio suivant :

$$\frac{\lambda(t, (X_1, \cdots, X_k, \cdots, X_p)}{\lambda(t, (X_1, \cdots, X_k - 1, \cdots, X_p))} = exp(\beta_{0k}).$$

— le lien entre l'intensité et les covariables est log-linéaire : $log(\frac{h(t, X_1 = x_1, \cdots, X_k = x_{k_1}, \cdots, X_p = x_p)}{h(t, X_1 = x_1, \cdots, X_k = x_{k_2}, \cdots, X_p = x_p)}) = \beta_0(x_{k_1} - x_{k_2})$.

Comme nous l'avons vu précédemment le modèle fait aussi l'hypothèse de l'indépendance de la cen-

sure avec l'évènement d'intérêt conditionnellement aux covariables. On dit que la censure ne doit pas être informative.

2.1.2.2 . Le modèle de Aalen

Le modèle de Aalen [1, 2] est une alternative au modèle de Cox. Au contraire de ce dernier, qui suppose des effets multiplicatifs des covariables sur l'intensité du processus de comptage, il fait l'hypothèse d'effets additifs des covariables. Dans la suite, nous supposerons l'effet des covariables dépendant du temps comme c'est le cas dans le modèle de Gran présenté section 2.2.2.4. Ainsi, l'intensité du processus de comptage s'écrit sous la forme suivante :

$$\lambda(t) = Y(t)X^\top(t)\beta^*(t), \tag{2.3}$$

avec $Y(t)$ l'indicateur de risque, $X(t)$ le vecteur de covariables de dimension p et $\beta^* : t \to \mathbb{R}^p$ la fonction des coefficients de régression. Les paramètres β^* sont des fonctions pour lesquelles aucune hypothèse n'est faite concernant leur forme. On peut définir un modèle d'Aalen semi-paramétrique $\lambda(t) = Y(t)(\lambda_0(t) + X(t)\beta^*(t))$.

L'estimation directe de la fonction de régression β^* est difficile. Elle est donc faite de manière indirecte en estimant les fonctions de régression cumulées B^*. On définit le vecteur colonne B^* comme le vecteur composé des éléments suivants :

$$B_k^*(t) = \int_0^t \beta_k^*(s)ds, \text{ avec } k = 0, 1, \cdots, p.$$

Les covariables sont réorganisées dans la matrice de design suivante :

$$X(t) = (Y_1(t)X_1(t), \cdots, Y_n(t)X_n(t))^\top.$$

de cette façon, si un sujets i est à risque au temps t, alors la ième ligne de $X(t)$ est $(X_i^0(t), \cdots, X_i^p(t))$. Il s'agit de l'ensemble des mesures des covariables du sujet i au temps t. A l'inverse si le sujet n'est pas à risque de faire l'évènement alors cette ligne est composée uniquement de 0.

On peut montrer que $B^*(t)$ peut être estimé par des régressions linéaires multiples [129] et que la forme intégrable de l'estimateur est la suivante :

$$\hat{B}_k(t) = \int_0^t X^-(s)dN(s),$$

avec $X^-(t)$ l'inverse généralisée de $X(t)$: $X^-(t) = (X^\top(t)W(t)X(t))^{-1}X^\top(t)W(t)))^\top$ avec $W(t)$ une matrice diagonale (voir [133] pour une définition précise). Par convention, $X^-(t)$ est une matrice nulle si l'inverse n'existe pas. Il faut noter que les estimations ne contraignent pas le risque de base à être positif. Une estimation grossière des β_k^* peut être obtenue en regardant la pente de la fonction $\widehat{B}_k$. Des méthodes de lissage par noyau peuvent être utilisées pour obtenir une meilleure estimation des β_k^*.

2.2 - Thématique 1 : Analyse causale avec des données longitudinales et un critère de jugement dépendant du temps

La première problématique à laquelle nous nous intéressons dans ce travail est l'évaluation de l'efficacité d'un traitement à partir de données de vie réelle et notamment des données du SNDS. Comme nous l'avons évoqué dans l'introduction, l'utilisation des données rétrospectives, c'est-à-dire de données déjà recueillie dans un autre cadre que l'étude en question souvent dans les dossiers médicaux, ne permet pas de réaliser une randomisation pour affecter à certains patients le traitement à évaluer et aux autres le traitement de référence ou le placebo. Pourtant cette dernière méthode est considérée comme le « gold standard » lorsqu'il s'agit d'évaluer l'existence d'un lien de causalité entre une intervention, le traitement par exemple, et un effet, comme le décès. Sa force réside notamment dans sa capacité à minimiser les biais de sélection et ainsi s'assurer que la « population contrôle », celle qui n'a paps reçu le traitement, est un portrait fiable de ce qui se serait passé sans l'intervention.

Toutefois, la mise en place d'une étude randomisée ne peut pas être systématiquement appliquée. En effet, cette méthode présente un certain nombre de limites. On peut relever notamment des problématiques de ressources puisque ces études nécessitent du temps et un budget important, de disponibilité des données avant et après l'intervention, d'un effectif suffisant qui peut être difficile à atteindre notamment dans les maladies rares ou encore des problématiques éthiques notamment quand un traitement a montré une efficacité sur des études antérieures [47]. Un pan des statistiques s'est développé pour pallier l'absence de randomisation et permettre l'interprétation des résultats en termes de causalité et non pas seulement de corrélation : c'est l'analyse causale.

Nous présenterons dans la suite les méthodes de régression spécifiques aux données longitudinales qui peuvent être utilisées avec des modèles de survie dans le cadre d'une analyse causale. Ces méthodes sont particulièrement intéressantes lorsque l'on travaille sur l'exposition à un traitement qui peut varier dans le temps. Dans un second temps, nous présenterons les principes de l'analyse causale et les modèles existants pour réaliser des analyses de causalité sur données longitudinales. Nous clôturerons cette partie en présentant le modèle de Gran [90], modèle pour lequel nous avons proposé une correction et une extension aux covariables discrètes.

2.2.1 . Modèles de régression sur des données longitudinales

On se place ici dans le cadre de données suivant. On dispose des données de n individus suivis, au plus, jusqu'au temps τ_i. On s'intéresse à l'effet d'une intervention sur la survenue d'évènements récurrents modélisés par un processus de comptage comme décrit dans la partie 2.1.1.3. Pour chaque sujet, on dispose donc des données suivantes $(N_i(t), E_i(t), X_i(t))$ pour $t \leq \tau$ avec $N_i(t)$ le processus de comptage, $E_i(t)$ l'indicateur de risque, $X_i(t) = (1, X_i^1(t), \cdots, X_i^p(t))$ les covariables, l'une d'elle représentant l'exposition

au traitement.

L'évaluation de l'effet de covariables dépendantes du temps sur le processus de comptage peut être réalisée par les modèles de Cox ou de Aalen présentés dans la partie précédente. La prise en compte des variables dépendantes du temps est permise par le fonctionnement sous-jacent de ces modèles. Ces derniers comparent à chaque temps d'évènement les valeurs actuelles des covariables du sujet ayant fait l'évènement aux valeurs des covariables des autres sujets à risque à ce moment-là. Il est possible de considérer dans les modèles des coefficients de régression indépendant du temps comme présenté pour le modèle de Cox. Autrement dit, on fait l'hypothèse forte que l'effet des covariables est le même à tout temps t. Une extension possible est de considérer des coefficients dépendant du temps comme présenté pour le modèle de Aalen.

La prise en compte des variables dépendantes du temps est facilitée dans les modèles par la construction, en amont de l'analyse, d'une table selon un format spécifique. L'idée est de partitionner la période d'observation de chaque individu en sous-intervalles de temps. Ces intervalles de temps correspondent à des périodes durant lesquelles les covariables sont constantes et au plus un évènement s'est produit au point terminal. Par exemple, on observe pour l'individus i deux covariables dépendantes du temps $(X_i^1(t), X_i^2(t))$ entre 0 et τ_i. La variable X^1 est observée aux temps 0, t_1 et la variable X^2 est observée aux temps 0, t_2 et un évènement a lieu au temps τ_i. La table présentée en figure 2.4 illustre cet exemple.

Sujet	Temps 1	Temps 2	X^1	X^2	Evènement
i	0	t_1	1,3	5,3	0
i	t_1	t_2	2,1	5,3	0
i	t_2	τ_i	2,1	7,5	1

FIGURE 2.4 – Structure d'une table avec des covariables dépendantes du temps.

Nous lisons qu'entre les temps 0 et t_1 la covariable X^1 est de $1,3$ et que cet intervalle de temps ne s'est pas terminé par un évènement. Par convention les intervalles de temps sont considérés ouverts à gauche et fermés à droite. Dans le cas de variables non dépendantes du temps qui serait mesurée au début du suivi comme l'âge à l'inclusion ou le sexe, la mesure sera répétée d'une ligne à l'autre.

Un des points fondamentaux à noter pour l'utilisation de variables dépendantes du temps dans les modèles de durée est que seules les données passées peuvent avoir un effet sur les données au temps t mais que le futur ne peut pas avoir d'impact. Cela implique qu'aucune interpolation ne peut être faite entre deux mesures de temps. Si l'on reprend l'exemple de la table 2.4, on ne peut pas utiliser les mesures de X^2 aux temps 0 et t_2 pour interpoler la mesure au temps t_1.

Plusieurs langages de programmation proposent des fonctions afin de formater les tables selon le schéma présenté précédemment et d'appliquer les modèles de survie sur ces données [182]. En R, le package *timereg* permet d'appliquer les modèles de Cox et de Aalen selon le principe de « start and stop » c'est-à-dire d'utiliser une structure de données où la période de suivie et partitionnée en sous-intervalle de temps comme discuté précédemment. La formulation « start and stop » fait référence aux bornes des sous-intervalles de temps.

2.2.2 . Modèles causaux

En épidémiologie, les questions posées sont plus souvent de nature causale que de la simple association. Les analyses causales cherchent à évaluer ce qui se passerait dans une population si un changement spécifique survenait. On appelle ce changement une intervention qui peut survenir sur toute la population d'analyse ou non. Cette intervention peut être par exemple le traitement, une prise en charge comme un acte chirurgical ou une campagne de prévention.

Dans cette partie nous proposons d'introduire formellement les concepts sur lesquels reposent les analyses causales et de présenter les méthodes pouvant être mises en œuvre pour réaliser de telles analyses.

2.2.2.1 . Notion de causalité

La causalité est à distinguer de la notion d'association ou de corrélation. Cette dernière permet d'établir qu'il existe un motif conjoint entre deux variables. Autrement dit, ces deux variables évoluent conjointement. Il peut exister un lien entre deux variables sans que l'évolution de l'une résulte de l'évolution de l'autre, c'est-à-dire sans lien de cause à effet. Supposons que l'on dispose de données d'une population sur l'activité sportive et les cancers de la peau [43]. On trouve que ces deux variables sont corrélées positivement autrement dit on observe un nombre plus important de cancer de la peau dans la population pratiquant une activité sportive. Une interprétation causale pourrait être de dire que le sport augmente le risque de développer un cancer de la peau. Il est toutefois possible que l'association de ces deux variables résulte d'une autre variable non mesurée dans cette étude qui pourrait être la zone géographique des sujets et notamment le taux d'exposition au soleil. On peut faire l'hypothèse que les gens sont plus nombreux à faire de l'exercice dans les régions ensoleillées et l'on sait que l'exposition au soleil a un effet sur le risque de développer un cancer de la peau.

Trois conditions ont été posées par McDavis et Hawthorn [137] pour pouvoir établir un lien de causalité :

1. **Ordre des évènements :** L'intervention doit avoir lieu avant le résultat observé ;
2. **Corrélation :** Il existe une corrélation entre le résultat et l'intervention ;
3. **Confusion :** Il n'existe pas d'autres facteurs qui pourraient expliquer la corrélation entre le résultat et l'intervention.

Afin de mener à bien des analyses de causalité, les interventions et les paramètres causaux doivent être clairement identifiés. Ces éléments seront présentés par la suite. Si ces différents éléments sont vérifiés, la problématique de l'analyse statistique et celle de la causalité peuvent être traitées séparément.

2.2.2.2 . Définitions et notations

Cadre des données : Nous nous plaçons pour la suite dans le cadre d'une analyse causale basée sur les concepts d'intervention et de contrefactuelles [183, 155, 161, 142] en présence de facteurs de confusion indépendant du temps. L'extension aux facteurs de confusions dépendant du temps sera présenté par la suite. La figure 2.5 est un graphe acyclique dirigé [145, 146] qui présente la relation entre l'intervention, le résultat et les covariables. Les facteurs de confusion sont schématiquement représentés par la flèche en

pointillée dans la figure 2.5. Ils correspondent aux facteurs qui agissent à la fois sur l'intervention et sur le résultat perturbant ainsi l'association et donc l'analyse causale. Pour la suite, nous noterons les variables aléatoires correspondant au traitement, au résultat d'intérêt et aux covariables respectivement D, Y et X. Le résultat d'intérêt peut être une mesure continue, discrète ou le temps de survenue d'un évènement. Par exemple, en oncologie, l'efficacité des traitements est souvent mesurée sur les critères de survie sans progression ou de survie globale. Dans ce cas, le résultat d'intérêt est le temps de survenue du décès. En diabétologie, le taux moyen de glycémie ou le nombre d'épisode d'hypoglycémie sont des critères courant d'évaluation de l'efficacité d'un traitement. Ces deux résutats sont respectivement des mesures continues et discrètes. Nous noterons d, y et x les valeurs observées des variables aléatoires D, Y et X. L'ensemble des régimes de traitement observabés sera noté $\mathcal{D}$, ainsi $d \in \mathcal{D}$.

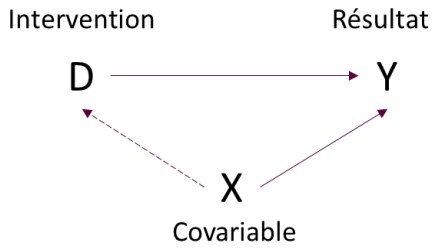

FIGURE 2.5 – Graphe acyclique dirigé présentant les relations entre l'intervention, le résultat et les covariables

Contrefactuelles : L'analyse causale cherche à déterminer si le résultat d'intérêt est le même lorsque le sujet a subi l'intervention ou non. Dans un monde idéal, il faudrait pouvoir observer le résultat pour un même sujet dans deux scénarios : 1. le sujet a subi l'intervention, 2. le sujet n'a pas subi l'intervention. En réalité, seul l'un des scénarios est observé. Les *covariables contrefactuelles* désignent les covariables telles qu'elles auraient été observées dans le scénario qui ne s'est pas produit en réalité. De la même façon le *résultat contrefactuel* désigne le résultat qui se serait produit dans le scénario non observé. Pour la suite nous noterons X^d (respectivement Y^d) les covariables (respectivement le résultat) qui seraient observées si le régime de traitement était $d \in \mathcal{D}$.

Effet causal L'effet causal de D sur Y peut être considéré en étudiant les différences en termes de distribution de Y en fonction des différents régimes de traitement d. Supposons que la distribution de Y^{d^1} et celle de Y^{d^2} ne diffèrent pas quel que soit les régimes de traitement d^1 et $d^2 \in \mathcal{D}$. On peut donc en déduire qu'il n'y a pas de lien causal entre D et Y. Il est nécessaire de bien définir au préalable le régime de traitement d'intérêt et la mesure de l'effet qui sera rapportée. En pratique, on compare le plus souvent deux types de régime. On pourra par exemple comparer le fait d'avoir été traité $d = 1$ au fait de ne jamais avoir été traité $d = 0$.

Etant donné qu'il n'est pas possible d'observer un sujet sous l'ensemble des régimes de traitement, l'estimation de l'effet causal ne peut pas se faire au niveau individuel. Cette estimation devra être faite en moyenne sur une population ou conditionnellement aux covariables. La mesure la plus fréquemment utilisée

est l'Effet total moyen du traitement ou Average Treatment Effect (ATE) qui est donné, lorsque l'on compare les régimes de traitement d^1 et d^2, par :

$$ATE = \mathbb{E}[Y^{d^1}] - \mathbb{E}[Y^{d^2}].$$

On peut interpréter cette mesure comme étant la différence entre l'espérance du résultat si tous les sujets avaient reçu le régime de traitement d^1 et l'espérance du résultat si tous les sujets avaient reçu le régime de traitement d^2. En fonction des problématiques, d'autres indicateurs peuvent être utilisés pour mesurer les effets causaux comme par exemple le risque relatif causal, $\frac{\mathbb{E}[Y^{d^1}]}{\mathbb{E}[Y^{d^2}]}$ ou l'odds ratio causal : $\frac{\mathbb{E}[Y^{d^1}]/1-\mathbb{E}[Y^{d^1}]}{\mathbb{E}[Y^{d^2}]/1-\mathbb{E}[Y^{d^2}]}$. Il peut être aussi intéressant de regarder l'effet moyen du traitement causal dans une sous-population. Nous reviendrons sur ce point dans le paragraphe suivant.

En pratique, il faudrait pouvoir estimer les espérances $\mathbb{E}[Y^{d^1}]$ et $\mathbb{E}[Y^{d^2}]$. Intuitivement, on voudrait pouvoir utiliser les résultats observés chez les patients sous le régime de traitement d^1 pour estimer $\mathbb{E}[Y^{d^1}]$ de même pour le régime d^2. Or en général, $\mathbb{E}[Y^{d^1}] - \mathbb{E}[Y^{d^2}] \neq \mathbb{E}[Y|D = d^1] - \mathbb{E}[Y|D = d^2]$. Afin de pouvoir utiliser les données observées pour relier les résultats observés aux résultats contrefactuels, certaines conditions nécessaires ont été définies. Ainsi il est possible d'identifier la distribution des contrefactuelles et par extension les paramètres de causalité [87, 98], sous les hypothèses suivantes :

- **Consistance** : Cette hypothèse stipule que le résultat potentiel d'un sujet dans un régime de traitement hypothétique qui s'est réalisé est exactement le résultat observé de ce sujet. Mathématiquement, on peut noter : $Y^d = Y$ si $D = d$.
- **Echangeabilité** : Cette hypothèse repose sur l'idée que l'on peut intervertir les sujets traités et non traités et obtenir le même résultat. On peut ainsi écrire que conditionnellement aux covariables le résultat contrefactuel est indépendant du traitement. Dans le cas de deux régimes de traitements, traité ou non traité, cette hypothèse peut être interprétée de la manière suivante : si deux sujets sont parfaitement identiques et que l'un est traité et l'autre non, alors le résultat observé de l'un est le contrefactuel de l'autre (en distribution).
- **Positivité** : Il s'agit de l'hypothèse selon laquelle tout individu a une probabilité positive de recevoir chacun des régimes de traitements différents. Autrement dit, il faut pouvoir identifier dans la population d'étude des patients exposés et d'autres non sans quoi il ne sera pas possible d'avoir d'information sur la distribution du résultat dans le régime non observé.

Si ces hypothèses sont validées, $\mathbb{E}[Y^{d^1}|X = x]$ pourra être estimée par $\mathbb{E}[Y|X = x, D = d]$. Le conditionnement par D est assuré par l'hypothèse de positivité. Les hypothèses de consistance et d'échangeabilité permettent d'écrire :

d'après l'hypothèse de consistance

$$\mathbb{E}[Y|X = x, D = d^1] = \mathbb{E}[Y^{d^1}|X = x, D = d^1]$$

d'après l'hypothèse d'écheangeabilité

$$= \mathbb{E}[Y^{d^1}|X = x].$$

Par intégration par rapport à la loi de X, on obtient alors $\mathbb{E}[Y^{d^1}]$ et donc l'ATE.

ATE *vs* ATT Comme nous l'avons vu précédemment l'indicateur le plus souvent utilisé dans l'analyse de l'effet causal est l'ATE. Ce dernier permet d'évaluer l'effet du traitement dans l'ensemble de la population. Cet indicateur permet de répondre à la question suivante : « Que ce serait-il passé si tous les sujets de la population avait reçu le traitement d'intérêt par rapport à la situation où ils auraient reçu le traitement de référence ou le placebo ». Il correspond à l'effet moyen du traitement dans un monde contrefactuel où toute la population est observée, les sujets ayant été traités comme les sujets non traités. Une autre mesure possible serait d'évaluer l'effet causal du traitement dans une sous-population et notamment chez les patients traités. On parle alors de l'Effet moyen du traitement chez les patients traités ou Average Treatment effect on Treated (ATT). Cette mesure permet de répondre à la question « quelle est l'efficacité du traitement pour ceux qui l'ont reçu ? ». Avec cette méthode, l'effet du traitement n'est pas évalué chez les patients qui n'ont jamais reçu le traitement. La figure 2.6 illustre la différence entre ces deux mesures en terme de populations étudiées.

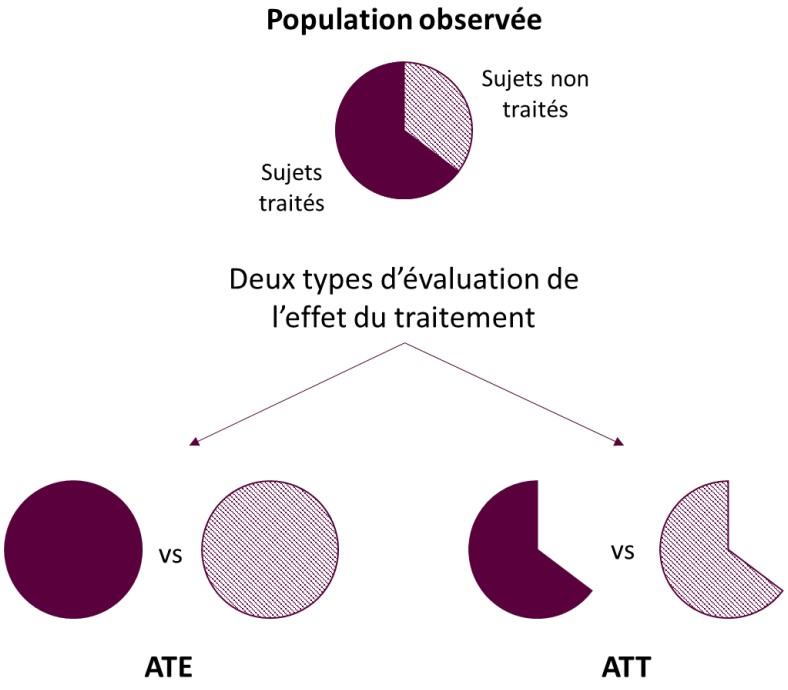

FIGURE 2.6 – Analyse de la population selon que l'on estime l'ATE ou l'ATT [90]

Le choix de la mesure utilisée pour évaluer l'effet du traitement dépend de la problématique étudiée notamment concernant la pathologie et l'intervention. L'ATT est une mesure pertinente de l'effet du traitement lorsqu'il existe une sélection parmi les patients traités [121]. Dans le domaine de la santé, on peut faire l'hypothèse qu'il existe une sélection spécifique des patients traités d'autant plus quand on étudie le traitement dans le temps et en vie réelle. En effet la prescription d'un traitement spécifique est réalisée par un praticien qui a évalué le bénéfice / risque des différents régimes de traitement pour un patient donné. Son choix peut être motivé par exemple par le profil du patient ou ses comorbidités. Plusieurs articles se sont intéressés au choix de l'utilisation de ces deux mesures et ont montré que les résultats obtenus pouvaient être très différents [149, 8]. Ces deux mesures sont complémentaires et peuvent être toutes deux évaluées dans une même étude [201, 115].

Extension aux données longitudinales Les définitions introduitent précédemment peuvent être étendues aux données longitudinales. La figure 2.7 illustre les relations entre l'intervention, le résultat et les covariables au cours du temps que nous allons supposer discret. L'exposition au traitement, les covariables et le résultat sont mesurés en tout temps $k \in \{0, \cdots, K\}$ et sont notés respectivement D_k, X_k et Y_k. Nous noterons respectivement d_k, x_k et y_k les valeurs observées de ces variables aléatoires. Nous utiliserons les barre pour noter l'historique des informations. Ainsi $\bar{D}_k = (D_0, \cdots, D_k)$ correspond à l'historique de l'expotition du traitement au temps k, $\bar{X}_k = (X_0, \cdots, X_k)$ l'historique des covariables et $\bar{Y}_k = (Y_0, \cdots, Y_k)$ l'historique des résultats. Les covariables et les résultats observés au temps k, sous le régime de traitement jusqu'au temps k $\bar{d}_k$, sont notés respectivement $X^{\bar{d}_k}$ et $Y^{\bar{d}_k}$. Prenons, par exemple, les deux régimes suivants : $\bar{d}_k^0$, les patients ne sont pas traités du temps 0 au temps k, et $\bar{d}_k^1$, les patients sont traités en tout temps, du temps 0 au temps k. L'ATE au temps k s'écrit alors : $ATE_k = \mathbb{E}[Y_k^{\bar{d}_k^1}] - \mathbb{E}[Y_k^{\bar{d}_k^0}]$. Les hypothèses de consistence, échangeabilité et positivité peuvent être écrites dans le cas de données longitudinales, voir [157].

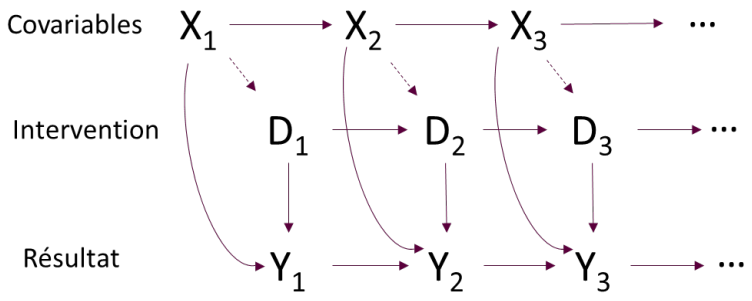

FIGURE 2.7 – Graphe acyclique dirigé présentant les relations entre l'intervention, le résultat et les covariables dans le cas de données longitudinales

2.2.2.3 . Méthodes d'estimation classique

Comme nous l'avons vu précédemment, la problématique de l'évaluation d'un effet causal réside principalement dans la gestion des facteurs de confusion. Cette prise en compte des facteurs de confusion est d'autant plus complexe que l'on s'intéresse à des données longitudinales.

Nous allons voir dans la suite les méthodes communément utilisées pour traiter les facteurs de confusion dépendant du temps puis nous présenterons spécifiquement le modèle développé par Gran [90] qui propose une méthode simple à implémenter pour l'estimation de l'ATT.

Score de propension Le score de propension est défini comme étant la probabilité d'être exposé au traitement d'intérêt conditionnellement aux covariables observées avant l'exposition [160]. La distribution des covariables observées est rendue indépendante de l'exposition au traitement conditionnellement au score de propension. Ainsi un équilibre est créé entre les covariables des sujets traités et non traités.

Les études longitudinales impliquent une exposition au traitement qui peut varier dans le temps. On peut donc définir à chaque temps de l'étude une probabilité d'être traité qui dépendrait des caractéristiques et du régime de traitement observés jusqu'à ce temps. Formellement, le score de propension pour l'individu i au temps k pourrait s'écrire de la manière suivante :

$$sp_k^i = \mathbb{P}[D_k^i = d_k^i | \bar{D}_{k-1}^i = \bar{d}_{k-1}^i, \bar{X}_{k-1}^i = \bar{x}_{k-1}^i]. \tag{2.4}$$

Le score de propension peut être introduit dans des modèles de régression spécifique au résultat étudié de différentes façons :

- **Ajustement** [17] : Cette méthode consiste à utiliser le score de propension estimé comme variable d'ajustement dans un modèle de régression. Ce dernier est donc composé d'au moins deux variables explicatives, l'exposition au traitement et le score de propension. Pour l'utilisation des modèles de type « start and stop » introduits dans la partie 2.2.1, un score de propension peut être calculé pour chaque sous-intervalle de temps. Plusieurs études de simulation ont montré des performances moyennes de cette méthode dans l'estimation des effets marginaux [17, 15].
- **Appariement** [14, 162, 7] : L'appariement consiste à trouver un ou plusieurs sujets non traités ayant des caractéristiques proches de celles du sujet traité en termes de score de propension. L'estimation de l'effet du traitement se fait alors sur la population appariée. Plusieurs méthodes d'appariement peuvent être utilisée dans le cas de données longitudinales [184]. Les deux principales méthodes sont l'appariement séquentiel et l'appariement simultané. Le premier défini sur chaque sous-intervalle de temps un ensemble de sujets exposés et un ensemble de sujets non exposés (même s'ils seront exposés plus tard). L'appariement est réalisé sur chacun de ces intervalles entre les deux ensembles. Les sujets appariés dans l'intervalle de temps $(t-1, t]$ sont exclus du processus d'appariement sur l'intervalle de temps $(t, t+1]$ et des suivants. Dans ce cas l'appariement n'est basé que sur les informations disponibles au moment de l'appariement. L'appariement simultané compare toutes les combinaisons possibles de paires appariées en une seule fois et effectue un seul appariement sur l'ensemble des patients. Cet appariement suppose que les covariables soient exogènes pour assurer une faible association entre les valeurs des covariables futures et l'initiation du traitement au temps k [60]. Le choix du « meilleur témoin »pour un sujet traité est discuté dans différentes publications [20, 32].
- **Pondération** [160, 16] : Cette méthode consiste à pondérer chaque individu par un poids qui est

fonction du score de propension afin d'obtenir des caractéristiques semblables pour les sujets traités et les sujets non traités. Nous allons étudier cette méthode plus en détail dans le paragraphe suivant.

Pondération par la probabilité inverse d'être traité La méthode de Pondération par la probabilité inverse d'être traité ou Inverse Probability Treatment Weighting (IPTW) repose sur le concept de l'analyse d'une *pseudo-population*. Cette pseudo-population est créée afin d'assurer que l'association entre l'exposition et le résultat dans cette population soit causale bien qu'elle ne le soit pas dans la population totale de l'étude à cause d'éventuels facteurs de confusion. Au temps k, l'association est causale si $\mathbb{P}[D_k|\bar{X}_k] = \mathbb{P}[D_k]$. En d'autres termes, la méthode IPTW simule efficacement les données qui auraient été observées si, contrairement à la réalité, l'exposition avait été aléatoire sans condition. Les pseudo-populations créées vérifient que la moyenne de Y_k est identique à celle dans la population totale mais que l'exposition au traitement D_k est indépendante des covariables $\bar{X}_k$.

La méthode IPTW utilise la pondération suivante au temps k : $1/\mathbb{P}[D_k|\bar{X}_k]$. Cela revient à pondérer les sujets traités par l'inverse de la probabilité d'être traité et les sujets non traités par la probabilité d'être non traités. Le score de propension peut être utilisé pour estimer cette pondération. Cette pondération est dite non stabilisée. En effet, les sujets ayant une très faible chance d'être traités, conditionnellement aux covariables observées, auront un poids très élevé. Les analyses seront alors fortement impactées par ces patients.

Des poids dit stabilisés ont été développés et devraient être préférés aux poids non-stabilisés [156]. Ces poids diffèrent uniquement sur le numérateur. Ce dernier est remplacé par la probabilité d'être traité chez les patients traités et peut s'écrire au temps k :

$$w_k^d = \frac{\mathbb{P}[D_k^i = d_k^i|\bar{D}_{k-1}^i]}{\mathbb{P}[D_k^i = d_k^i|\bar{D}_{k-1}^i, \bar{X}_{k-1}^i]}.$$

Une pondération finale peux être obtenue en multipliant l'ensemble des pondérations obtenues sur les sous-intervalles de temps. Cette pondération s'écrit alors :

$$sw_k^d = \prod_{k=1}^{K} \frac{\mathbb{P}[D_k^i = d_k^i|\bar{D}_{k-1}^i]}{\mathbb{P}[D_k^i = d_k^i|\bar{D}_{k-1}^i, \bar{X}_{k-1}^i]}.$$

Il faut noter qu'en cas de censure, un poids de censure peut être introduit :

$$we_k^d = \frac{\mathbb{P}[E_k^i = e_k^i|\bar{D}_{k-1}^i]}{\mathbb{P}[E_k^i = e_k^i|\bar{E}_{k-1}^i, \bar{X}_{k-1}^i]},$$

avec E_k^i l'indicateur de la censure pour le patient i au temps k et $\bar{E}_k^i$ les informations passées concernant les censures. Ces poids de censure peuvent être utilisés comme pondération inverse, on parle de Pondération par la probabilité inverse d'être censuré ou Inverse Probability of Censoring Weighting (IPCW). Un poids final peut être obtenu en considérant le produit à chaque temps du poids de l'exposition w_k^d et du poids de

censure we_k^d.

Le modèle de Cox pondéré par la probabilité inverse est de plus en plus utilisé pour traiter la confusion dépendente du temps dans des contextes d'analyse de durée [177, 156].

G-computation La *g-computation* repose sur le principe de la standardisation. Dans le cadre d'un facteur de risque qui ne dépend pas du temps, X_0, et d'une exposition à l'inclusion, D_0, cette méthode cherche à estimer la valeur attendue de Y^{d_0} c'est-à-dire le résultat qui serait observé si toute la population était sous le régime de traitement d_0 par standardisation. L'expression est la suivante quand X_0 est à valeurs discrètes :

$$\mathbb{E}[Y^{d_0}] = \sum_{x_0 \in \mathcal{X}_,} \mathbb{E}[Y|D_0 = d_0, X_0 = x_0]\mathbb{P}[X_0 = x_0].$$

Le principe de standardisation a été étendu par Robin [183] pour définir la formule de g-computation lorsqu'on est en présence d'une exposition et de covariables dépendantes du temps à valeurs discrères. Cette formule s'écrit alors de la façon suivante :

$$\mathbb{E}[Y^d] = \sum_{\bar{X} \in \bar{\mathcal{X}}} \left[\mathbb{E}[Y|\bar{D} = \bar{d}, \bar{X} = \bar{x}] \prod_{k=0}^{K} \mathbb{P}[X_k = x_k | D_{k-1} = d_{k-1}, X_{k-1} = x_{k-1}] \right],$$

où la somme porte sur toutes les valeurs possible de l'évolution des covariables. En pratique, chacun des termes de la formule peuvent être estimés séparément mais comme il faut estimer la probabilité conditionnelle que $X_k = x_k$, elle n'est pas adaptée à la grande dimension ou dans le cas où les covariables sont continues. En effet, l'estimation d'une densité conditionnelle en grande dimension n'est jamais stable.

G-estimation de modèles structuraux emboîtés Les modèles structurauxs emboîtés sont un ensemble de modèles définis pour chaque temps d'observation $k = 1, \cdots, K$. Le k-ième modèle est une comparaison de deux résultats potentiels. Le premier résultat potentiel serait observé si le sujet avait suivi un certain régime de traitement jusqu'au temps k inclus puis aurait arrété le traitement pour les temps suivant. Le second résultat potentiel serait observé si le sujet avait suivi le même régime de traitement que précédemment mais aurait arrété le traitement un temps plus tôt, c'est-à-dire au temps $k-1$ inclus. Ce sous modèle caractérise donc l'effet final du traitement au temps k. Il peut s'écrire en fonction de l'historique de traitement jusqu'au temps $k-1$, $\bar{d}_{k-1}$, et de l'évolution des covariables, $\bar{x}_k$. Les Modèles structuraux emboîtés ou Structural Nested Model (SNM) sont donc conditionnels aux facteurs de confusion.

Dans le cadre de résultats continus le modèle SNM peut s'écrire sous la forme suivante :

$$\mathbb{E}\left[Y^{(d_0,d_1,\cdots,d_k,0,\cdots,0)}|\bar{D}_{k-1}^i = \bar{d}_{k-1}^i, \bar{X}_k^i = \bar{x}_k^i\right] = \mathbb{E}\left[Y^{(d_0,d_1,\cdots,d_{k-1},0,\cdots,0)}|\bar{D}_{k-1}^i = \bar{d}_{k-1}^i, \bar{X}_k^i = \bar{x}_k^i\right] + \psi_k(\bar{d}_k^i, \bar{x}_k^i, \phi_k^*). \tag{2.5}$$

La fonction ψ_k est fonction du traitement au temps k, et de l'historique de traitement et des covariables jusqu'au temps k avec les paramètres ϕ_k^*. Cette fonction peut prendre plusieurs formes comme par exemple $\psi_k(\bar{d}_k^i, \bar{x}_k^i, \phi_k) = \phi_{k,0}^* d_k$ ou $\psi_k(\bar{d}_k^i, \bar{x}_k^i, \phi_k^*) = (\phi_{k,0}^* + \phi_{k,1}^* d_k + \phi_{k,2}^* x_k) d_k$.

La méthode de g-estimation permet d'estimer les paramètres ϕ_k^* suivant un algorithme itératif. En

pratique l'algorithme teste différentes valeurs de ϕ_k^* pour prédire les résultats contrefactuels. L'objectif est de trouver la valeur telle qu'il n'existe pas d'association entre D et Y conditionnellement à X. Plusieurs papiers présentent plus précisément l'algorithme d'estimation [65, 191]. Si ϕ_k^* est de dimension $d > 2$ alors l'application de l'algorithme est quasiment impossible sur le plan informatique sans hypothèse sur la forme de ψ_k.

2.2.2.4 . Modèle de Gran

Les modèles présentés précédemment permettent d'identifier aussi bien l'ATE que l'effet du traitement chez les patients traités (ATT). Théoriquement, la g-estimation pour les modèles structuraux emboîtés se prête bien à l'estimation de l'ATT. Toutefois ces modèles sont peu utilisés en pratique. Cela peut notamment s'expliquer par la difficulté de sa mise en œuvre et ce d'autant plus dans le cadre de données dépendantes du temps [121, 191]. En 2017, Gran et ses coauteurs [90] ont proposé une nouvelle méthode permettant d'estimer l'ATT de façon relativement simple. La méthode proposée permet à la fois d'estimer l'effet du traitement sur le résultat mais aussi sur l'évolution des covariables dans le temps.

La méthode de Gran se place dans le cadre de données longitudinales comme présenté par le Graphe orienté acyclique ou Directed Acyclic Graph (DAG) de la figure 2.7 avec pour résultats d'intérêt la survie. Pour la suite, nous noterons $S = s$ le temps de l'initiation du traitement. L'hypothèse est faite qu'une fois le traitement initié, le sujet reste traité. Autrement dit en reprenant les notations de la partie 2.2.2.2, $d_k^* = 0$ si $k < s$ et $d_k^* = 1$ si $k \geq s$. Les covariables potentielles seront notées $X^{0|S}$ et $X^{1|S}$ respectivement pour les covariables dans le scénario d'absence de traitement ou le scénario sous traitement. On note T^S le temps de l'évènement observé et $T^{S^{1|*}}$ le temps de l'évènement si le sujet n'avait pas été traité.

La méthode de Gran se décompose en deux étapes. On modélise d'abord les trajectoires des covariables non observées ou contrefactuelles après l'initiation du traitement puis on estime l'effet du traitement par un modèle de Aalen. L'idée est d'estimer l'effet causal du traitement en modélisant l'évolution des covariables si les patients n'avaient jamais été traités. Si les valeurs contrefactuelles des covariables sont substituées aux valeurs réelles observées, on peut alors estimer l'ATT, moyennant certaines hypothèses.

Etape 1 : Modélisation des contrefactuels La première étape consiste donc à modéliser des trajectoires des covariables manquantes après l'initiation du traitement. Autrement dit on cherche à modéliser $X_k^{0|S=s}$ pour $k \geq S$ (Figure 2.8).

Le modèle proposé pour la modélisation reprend le modèle des incréments linéaires de Farewell [68, 59]. Il a été conçu pour l'analyse simple des données manquantes dans les études longitudinales. Avec les incréments, le modèle capture les changements dans l'évolution des covariables [5]. Pour simplifier les notations, notons $\widetilde{X}$ ces covariables. Les incréments s'écrivent de la façon suivante :

$$\Delta \widetilde{X}_k = \widetilde{X}_k - \widetilde{X}_{k-1}.$$

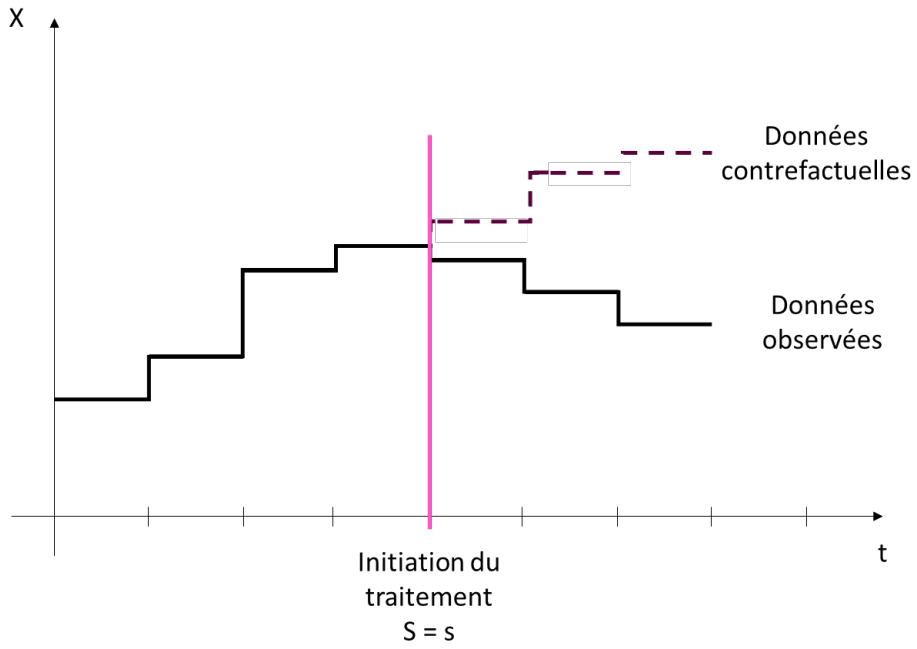

FIGURE 2.8 – Illustration d'une covariable observée et de sa contrefactuelle après l'initiation du traitement.

On suppose que pour tout temps k, les incréments vérifient un modèle linéaire :

$$\Delta\widetilde{X}_t k = \widetilde{X}_{k-1}\beta_k^* + \epsilon_k,$$

avec β_k^* la matrice des paramètres du modèle de taille $p \times p$ et ϵ_k le vecteur d'erreur de taille p. Le lien entre les covariables observées X et l'ensemble des covariables $\widetilde{X}$ peut s'écrire : $\Delta X_k = Q_k^* \Delta\widetilde{X}_k$ avec $Q_k^* = 1$ si l'incrément ΔX_k^* est observé et 0 sinon.

La modélisation des trajectoires suppose qu'il n'y a pas de lien entre $\beta_{k_1}^*$ et $\beta_{k_2}^*$ pour deux temps différents k_1 et k_2. Les paramètres β_k peuvent être estimés par la méthode des moindres carrés.

L'estimation des covariables non observées, c'est-à-dire les contrefactuelles dans l'hypothèse où aucun sujet n'aurait été traité, est réalisée itérativement.

Etape 2 : Estimation de l'effet du traitement Le modèle d'estimation de l'effet du traitement utilisé dans la méthode de Gran est le modèle de Aalen. Son choix a été motivé pour ces propriétés comme la collapsibilité qui assure que l'interprétation du paramètre d'exposition est la même dans les modèles conditionnels et marginaux et l'écriture explicite du processus d'estimation qui sont utiles dans les contextes d'analyse causale [135, 192, 180, 178, 6, 134].

L'effet du traitement peut être estimé par le modèle de Aalen à partir de l'estimation du paramètre de régression correspondant à la covariable indicatrice du traitement. Gran propose deux méthodes d'estimation de l'ATT. Le formalisme mathématique est détaillé dans la publication [90]. Nous présentons ici les grands principes de l'estimation.

La première estimation découle directement du formalisme de la méthode. L'ATT au temps k correspond à l'effet du traitement obtenu avec les covariables observées plus une combinaison linéaire des différences de distribution des covariables contrefactuelles si tous les patients étaient traités *vs* en absence de traitement. L'estimation est réalisée en deux temps. L'estimation dans le modèle de Aalen est d'abord réalisée sur les données observées. Puis l'ATT est estimé à partir des effets du traitement et des covariables issue du modèle de Aalen.

La seconde méthode est une méthode « raccourcie » pour obtenir une approximation de l'estimation de l'ATT. Elle repose sur un seul modèle de régression. L'idée est de réaliser un modèle de Aalen avec pour variables d'ajustement non pas les covariables observées mais directement les covariables contrefactuelles dans le scénario où les sujets n'auraient jamais été traités, modélisées à l'étape 1. L'estimation de l'ATT est obtenue par l'estimation du paramètre de régression du modèle correspondant à la variable de traitement. L'idée est que si les sujets traités ont pour caractéristiques les covariables contrefactuelles en absence de traitement alors la différence observée sur les résultats n'est due qu'au traitement. Cette idée a été utilisée dans d'autres contextes et a bien fonctionné [110, 179].

2.2.3 . Apport de la thèse

Pour rappel, cette première problématique a été motivée par la volonté d'évaluer l'efficacité des traitements dans la sclérose en plaques à partir de données de vie réelle. L'ATT semble être le meilleur indicateur à évaluer à partir de nos données. En effet, nous ne disposons que des données de consommation de soin issues des bases du SNDS des patients traités. La méthode de Gran présente une bonne approche puisqu'elle propose une estimation de l'ATT simple à mettre en œuvre.

Toutefois, nous avons montré dans nos travaux que l'estimation de l'ATT par la méthode de Gran est biaisée, voir le chapitre 3. En effet, cette dernière ne prend pas en compte l'erreur qui est faite lors de la modélisation des contrefactuelles en l'absence de traitement. L'idée repose sur le fait que les contrefactuelles dans le scénario ou les patients ne seraient pas traités sont modélisées. Autrement dit, on ne dispose pas des vraies valeurs des covariables dans ce scénario mais d'une modélisation. Une erreur réside donc entre les données modélisées et les « vraies » données. Or l'estimation de l'ATT repose sur la réalisation d'un modèle de Aalen avec pour variables d'ajustement les contrefactuelles modélisées. L'erreur générée au moment de la modélisation est alors répercutée dans l'estimation de l'ATT. Nous proposons dans ce travail une correction de l'estimateur de l'ATT, voir le chapitre 3.

Les traitements utilisés dans le cadre de la SEP servent à réduire la fréquence des poussées. Leur efficacité est donc évaluée sur la survenue des poussées qui est donc un évènement qui peut se répéter dans le temps. Comme nous l'avons vu dans la section 2.2.2.2, l'utilisation du modèle de Aalen est applicable dans le cas d'analyse de durée avec évènements répétés. Nous avons donc généralisé la méthode de Gran avec un résultat de type processus de comptage et des covariables dépendantes du temps et des covariables à l'inclusion (donc non dépendantes du temps). Nous avons utilisé comme méthode de modélisation des contrefactuelles un modèle vecteur autorégressif (VAR) qui permet de prendre en compte à la fois les erreurs sur les covariables et un horizon passé plus long que le modèle de Farewell qui n'utilise que la valeur précédente.

Enfin nous disposons des données brutes issues des bases du SNDS. Des indicateurs sont donc créés pour décrire et analyser les consommations de soin sur des intervalles de temps prédéfinis. A l'exception des doses de traitement, très peu de ces indicateurs sont des variables continues et dépendantes du temps. La majorité

FIGURE 2.9 – Procédure d'analyse des parcours de soin pour identifier l'association entre ces parcours de soin et le décès.

des indicateurs sont du type de données de comptage comme par exemple le nombre de consultations chez un neurologue par trimestre, le nombre de consultations d'une infirmière par mois ou encore le nombre de délivrances de corticoïdes par mois. La modélisation des contrefactuelles par le modèle VAR n'est donc pas adéquate. Nous proposons l'utilisation du modèle INGARCH [73] pour la modélisation de covariables de comptage et avons étendu ce modèle en dimension p avec $p \geq 1$.

2.3 - Thèmatique 2 : Analyse des parcours de soin pour la prédiction d'évènement d'intérêt

La seconde problématique à laquelle nous nous sommes intéressés dans ce travail est l'étude des parcours de soin afin d'identifier leur association avec un évènement d'intérêt comme le décès. Comme nous l'avons évoqué dans l'introduction, l'identification de parcours de soin associé à un sur-risque de décès permettrait aux praticiens de modifier la prise en charge des patients présentant ce parcours de soin afin de prévenir le décès.

Une procédure d'analyse a été proposée par Pinaire et ses coauteurs [147] pour l'étude des trajectoires hospitalières des patients suite à un infarctus du myocarde. La succession des étapes de cette procédure est présentée Figure 2.9. Cette partie cherche à présenter les différents modèles pouvant être mis en place pour chacune des étapes de la procédure.

Nous présenterons dans la suite comment les parcours de soin peuvent être construits à partir des données du SNDS et particulièrement aux données hospitalières renseignées dans le PMSI. Nous donnerons ensuite un certain nombre de définitions relatives à l'analyse des séquences. Les algorithmes permettant d'identifier les séquences fréquentes et ceux permettant de quantifier la similarité entre deux séquences seront présentés. Enfin, des modèles d'apprentissage statistique permettant de les lier à la survie seront exposés.

2.3.1 . Analyse des trajectoires d'hospitalisation

2.3.1.1 . Les trajectoires de soin dans le SNDS

Comme nous l'avons vu précédemment, le PMSI contient l'ensemble des données issues des hospitalisations dans des établissements publics et privés. Il a été construit dans un objectif de suivi de l'activité et de facturation. Des codages standardisés ont été introduits pour simplifier la tarification. Il s'agit des GHM. Nous allons voir dans cette partie comment sont construits ces codes et comment ils peuvent être utilisés dans l'analyse des trajectoires. Pour la suite nous nous limiterons aux hospitalisations ayant lieux dans des établissements de médecine, chirurgie et obstétrique, relevant ainsi du PMSI MCO.

Lors d'un séjour hospitalier, un patient peut être pris en charge dans une ou plusieurs unités médicales. Un Résumé d'Unité Médicale (RUM) est établi à l'issue d'un passage dans une de ces unités. Ce dernier contient les informations administratives et médicales, codées selon des classifications standardisées pour rendre compte de la prise en charge. Deux informations sont primordiales concernant la prise en charge médicale, les diagnostics et les actes médicaux.

Les diagnostics sont codés selon la classification internationale des maladies, version 10 (CIM-10). Plusieurs types de diagnostic peuvent être codés. Tout d'abord, le Diagnostic Principal (DP) est unique pour chaque RUM et correspond au motif de la prise en charge. Les diagnostics reliés (DR) sont renseignés lorsque le diagnostic principal n'est pas suffisamment informatif. Il intervient principalement lorsque le DP correspond à un code ne codant pas une pathologie mais une prise en charge. Par exemple, le DP peut être « Z511 Séance de chimiothérapie pour tumeur » et le DR précise pour quel cancer, « C64 Tumeur maligne du rein ». Enfin, les Diagnostics Associés (DAS) peuvent être aussi nombreux que nécessaire. Ils codent des pathologies qui pourraient expliquer la majoration de l'effort de soin ou des moyens utilisés.

Les actes médicaux sont codés selon la Classification Commune des Actes Médicaux (CCAM). Trois types d'actes peuvent être identifiés dans un objectif de tarification. Les *actes classants* sont des actes médico-techniques susceptibles de modifier le classement du séjour dans un GHM. On distingue les actes classants opératoires des actes classants non opératoires qui ne nécessitent pas d'être réalisés dans un bloc opératoire. La liste des actes classants est fournie par l'ATIH. Les autres actes sont dits *actes non classants*. Ils sont caractérisés ainsi puisqu'ils n'apportent pas d'informations supplémentaires qui pourraient expliquer une majoration des coûts ou de la durée du séjour. On retrouve par exemple, des actes de radiologie.

L'ensemble des informations des RUM permettent de définir un diagnostic principal au séjour et le GHM. Le diagnostic principal du séjour est déterminé parmi les DP des RUM à partir de la *fonction groupage* qui fonctionne de la façon suivante :

1. Le séjour est constitué d'un seul RUM : Le DP du séjour est le DP du RUM.
2. Le séjour est constitué de plusieurs RUM et il y en a au moins un avec un acte classant chirurgical : Le DP du séjour est le DP du RUM le plus long avec un acte classant chirurgical
3. Le séjour est constitué de plusieurs RUM sans acte classant chirurgical : Le DP du séjour est le DP du RUM qui a la plus longue durée ET qui est le plus proche de la fin du séjour.

L'identification du DP du séjour est primordiale car ce dernier joue un rôle important dans l'affectation d'un GHM.

Un GHM est identifié pour chaque séjour hospitalier par un algorithme de groupage se reposant, en général, d'abord sur le DP du séjour puis sur les actes classants et enfin sur d'autres variables comme l'âge, les DAS ou le mode de sortie (notamment pour les décès). Les GHM sont des codes de 6 caractères.

— Les deux premiers caractères sont un nombre qui représente la catégorie majeure de diagnostic (voir en annexe 7), qui reflète très souvent le domaine de la pathologie.

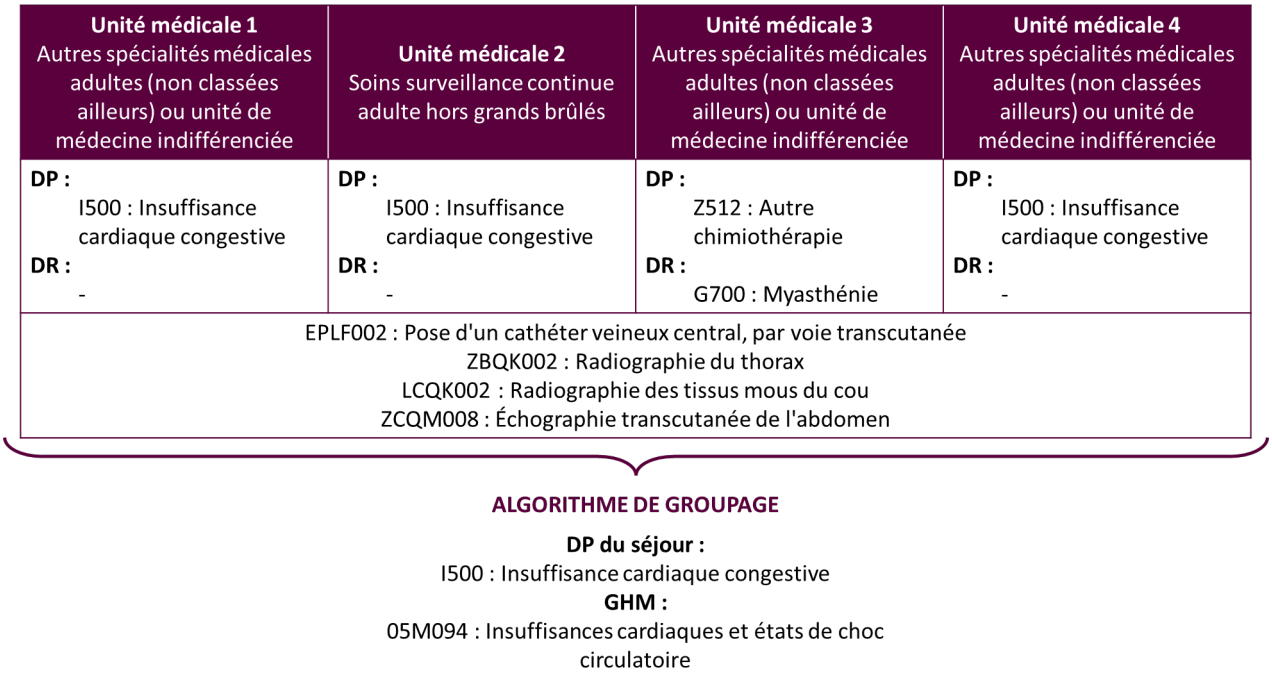

FIGURE 2.10 – Illustration d'un séjour hospitalier multi-RUM.

— Le troisième caractère est une lettre caractérisant le GHM notamment entre GHM chirurgical ou médical (C : au moins un acte classant chirurgical, M au moins un acte classant non chirurgical, K : pas d'actes classant, Z : indifférencié notamment pour les séances de chimiothérapie ou les séances de dialyses).
— Le dernier caractère indique soit la gravité de la maladie (1 à 4 ou A à D), sa durée (T : très courte durée) ou autre (E : décès en cours de séjour, Z : non concerné par les précédents).

L'utilisation des GHM dans l'analyse des parcours de soin présente un fort intérêt notamment puisqu'il est unique pour chaque séjour et que sa classification est standardisée. Le nombre de modalités possibles est beaucoup plus faible que celui du nombre de codes diagnostiques possibles (moins de 200 contre plus de 14 000 pour les diagnostics selon la classification CIM-10). De plus la construction hiérarchisée du code permet de regrouper certains séjours soit par domaine de pathologie, soit par le type de prise en charge (chirurgie *vs* actes médicaux).

2.3.1.2 . Définitions nécessaires à l'étude des trajectoires

L'utilisation des données observationnelles longitudinales peut permettre d'étudier des successions d'état par lesquels passent les sujets. L'analyse du PMSI se prête bien, par exemple, à l'étude des hospitalisations à

Numéro Patient	Numéro d'hospitalisation	Date d'entrée	GHM	Liste des diagnostics
1	1	05/11/2019	05M09	I501
1	2	20/04/2020	05C11	I742, E12
1	3	23/05/2020	05M09	I500, R572, E12
2	1	15/02/2020	05M09	I501, E66
2	2	30/10/2020	05M08	I48.0

FIGURE 2.11 – Table contenant les hospitalisations de deux patients.

la suite d'une chirurgie comme la chirurgie bariatrique [35] afin d'étudier les complications les plus fréquentes. On peut aussi chercher à faire des comparaisons de parcours de soin afin de regrouper des patients avec une prise en charge semblable, identifier des parcours s'éloignant des recommandations ou évaluer le pouvoir pronostic de certains parcours sur un résultat d'intérêt, voir [148, 206].

Les premières méthodes de fouille de données de type trajectoire portaient sur les règles d'association. Ces méthodes ont été popularisées par Agrawal en 1993, voir [9]. Elles ont ensuite été généralisées et ont conduit aux motifs séquentiels [136, 175]. Plusieurs définitions ont besoin d'être introduites ici pour formaliser la notion de motif séquentiel.

La figure 2.11 présente une table fictive regroupant les hospitalisations de deux patients. Pour chacune des hospitalisations, la date d'entrée, le GHM et l'ensemble des codes diagnostiques renseignés sont reportés. Cette table servira d'exemple pour les définitions suivantes.

Eléments ou item : Les éléments ou les items sont les états étudiés, noté e. Un sous-ensemble composé d'un ou plusieurs items est appelé itemset ou ensemble d'item. Un ensemble d'item de dimension m sera noté $\mathcal{E} = (e_1, \cdots, e_m)$. Cette collection est non ordonnée et peut contenir plusieurs fois le même item. La taille d'un ensemble correspond au nombre d'éléments le composant. Dans notre exemple, les éléments sont les GHM survenus lors d'une hospitalisation en MCO. Pour le patient 1 dans notre exemple les élements observés sont les GHM $05M09$ et $05M08$ codés respectivement pour l'hospitalisation 1 et 2.

Transaction : Une transaction, $\mathcal{T}$ est caractérisée par un identifiant, une date de transaction et un ensemble d'items. Dans notre exemple, une transaction peut être assimilée à une hospitalisation. Chaque hospitalisation est définie pour un patient, une date de début et un GHM. Dans ce cas, l'itemset n'est composé que d'un élément étant donné que le GHM est unique pour chaque hospitalisation. Si l'on raisonne non plus sur des éléments qui sont les GHM mais sur les diagnostics, l'ensemble d'items d'une transaction pourrait être de dimension $m \geq 1$ puisque composé du DP, et éventuellement des DR et DAS. Dans notre exemple, deux transactions sont identifiées pour le sujet 2. L'itemset de la première transaction est (05M09) si le GHM est considéré comme évènement ou (I501, E66) si le diagnostic est l'évènement.

Séquence : Les séquences d'ensembles sont définies comme des listes ordonnées d'ensembles d'évènements. Si l'on raisonne sur les transactions, les séquences d'évènement correspondent aux itemsets des transactions ordonnées. Formellement, on peut noter une séquence comme suit : $\langle \mathcal{E}_1, \cdots, \mathcal{E}_l \rangle = \langle (e_{1,1}, \cdots, e_{1,m_1}), \cdots, (e_{l,1}, \cdots, e_{l,m_l}) \rangle$. Dans notre exemple, la séquence d'évènements observée pour le sujet 1 est $\langle$(I501), (I742, E12), (I500, R572, E12)$\rangle$ si les évènements sont les diagnostics et $\langle$(05M09), (05C11), (05M09)$\rangle$ pour les GHM.

Sous-séquence : Soit $S_1 = \langle E_{1,1}, \cdots, E_{1,m_1} \rangle$ et $S_2 = \langle E_{2,1}, \cdots, E_{2,m_2} \rangle$ deux séquences d'ensemble. On

dit que S_1 est une sous-séquence de S_2 si il existe les indices $1 \leq j_1 < \cdots < j_{m_1} \leq m_2$ tels que $E_1, i \subseteq E_{2,j_i}$ pour tout $i = 1, \cdots, m_1$. Autrement dit, pour tout ensemble de S_1 on peut trouver un ensemble de S_2 égal à cet ensemble ou le contenant et que leur ordre d'apparition est le même dans les deux séquences. Par exemple la séquence $\langle (I500), (I500, E12) \rangle$ est une sous-séquence de $\langle (I500), (I742, E12), (I500, R572, E12) \rangle$.

Support : Le support d'une séquence S dans une base de données de séquences $\mathcal{B} = \{S_1, \cdots, \S_b\}$ correspondant à la proportion de séquences de $\mathcal{B}$ ayant S comme sous-séquence. Considérons la liste des diagnostics de la figure 2.11 comme une base de séquences. La séquence $\langle (\text{I501}) \rangle$ a un support de $2/5$.

Motif séquentiel : Un motif séquentiel est une séquence S qui est telle que son support est supérieur ou égal à un seuil, appelé support minimum, $k_\omega > 0$ fixé. On dit aussi que la séquence est fréquente. Dans notre exemple, la séquence $\langle (\text{I501}) \rangle$ est de support $2/5$ et est donc une séquence fréquente au seuil de 30%.

2.3.1.3 . Identification des motifs séquentiels

L'étude des trajectoires d'hospitalisation sur la base du PMSI présente l'avantage de donner une vision globale de la prise en charge en France grâce à l'exhaustivité des séjours hospitaliers et des individus étudiés. Toutefois, chaque sujet étant différent, le nombre de trajectoires distinctes est presque aussi important que le nombre de sujets étudiés. Il est donc nécessaire de restreindre le nombre de séquences à étudier. Une première possibilité est de restreindre la liste des élements d'intérêt. Par exemple, lorsqu'on s'intéresse aux parcours de soin d'une pathologie cardiaque, les hospitalisations pour d'autres motifs comme des fractures peuvent ne pas être pertinents. Dans ce cas, seuls les GHM codant pour une pathologie cardiaque sont utilisés pour construire les trajectoires, voir [147]. Si la sélection des éléments d'intérêt n'est pas pertinentes ou qu'elle ne suffit à restreindre le nombre de trajectoires distinctes, l'utilisation des séquences fréquentes peut permettre cette réduction de dimension. La problématique réside alors dans la représentation et dans l'identification de ces séquences. Deux méthodes sont utilisées pour représenter les trajectoires. La **représentation horizontale** présente pour chaque patient (en ligne) la trajectoire en succession d'items. La **représentation verticale** renseigne pour chaque item (en colonne) les identifiants des patients pour lesquel l'item est observé dans la trajectoire. Les modèles d'extraction des séquences fréquentes sont des algorithmes dit « d'extraction d'éléments fréquents » et de la famille « d'extraction de motifs séquentiels ». Les principaux algorithmes de ces méthodes sont présentés dans la suite.

Extraction d'éléments fréquents (Frequent itemset mining) Les premières méthodes d'extraction d'éléments fréquents ont été proposées au début des années 90 dans le domaine du marketing, voir [9]. Ces méthodes cherchaient à identifier des produits qui étaient fréquemment achetés ensemble. Il est important de noter que dans ces modèles la notion d'ordre dans l'ensemble d'évènements n'est pas prise en compte. Seule la co-occurence est analysée. Depuis, ces algorithmes ont été appliqués à des données de santé dans différents domaines, voir par exemple des études sur la démence [107] ou en cardiologie [105, 84]. Ces algorithmes permettent d'identifier l'ensemble des motifs séquentiels fréquents pour un seuil donné.

L'approche naïve consisterait à regarder l'ensemble des séquences possibles et à identifier leur fréquence dans un jeu de données. Cette approche est infaisable dès lors que le nombre d'évèvements possibles, et par conséquent le nombre de séquence à considérer, est grand. En effet si le nombre d'items est n, l'espace de recherche est de 2^n. L'ensemble des séquences possibles peut être représenté par un arbre. La figure 2.12 présente l'arbre des séquences candidates si l'ensemble possible des items est $\mathcal{E} = \{A, B, C\}$.

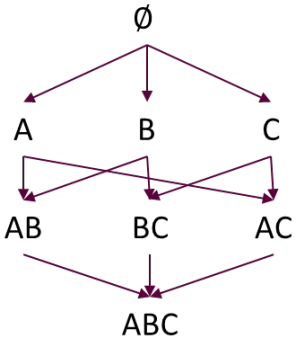

FIGURE 2.12 – Représentation de l'ensemble des séquences possiblement fréquent si les items possibles sont A, B et C

Des algorithmes spécifiques ont été développés pour pallier ce problème. Ces algorithmes reposent sur deux approches différentes. La première consiste à réaliser une recherche en largeur. c'est-à-dire que tous les voisins d'un nœud sont lus avant de passer aux successeurs. Supposons dans notre exemple que l'on commence systématiquement la lecture par le nœud de droite alors l'arbre est exploré de la façon suivante : $\emptyset$, A, B, C, AB, AC, BC, ABC. La seconde approche réalise une recherche en profondeur. c'est-à-dire que tous les successeurs d'un nœud sont explorés avant de lire un voisin. Dans notre exemple la lecture est la suivante : $\emptyset$, A, AB, ABC, AC, B, BC, C. Ces méthodes peuvent être classifiées en trois grandes catégories [10] : les algorithmes basés sur les jointures, les algorithmes basés sur les arbres et les algorithmes de croissance de motifs. Les algorithmes basés sur les jointures appliquent une méthode ascendante pour identifier les éléments fréquents d'un ensemble et les étendre à des ensembles plus grands si leur support est supérieur au seuil. Les algorithmes basés sur des arbres utilisent des concepts d'énumération d'ensembles en construisant un arbre lexicographique. Enfin les algorithmes de croissance de motifs fragmentent la base de données. La recherche de séquences fréquentes est faite sur chaque partition. Les principaux algorithmes de ces 3 catégories sont présentés par la suite.

Le premier algorithme développé est l'algorithme **aPriori** [10]. Il s'agit d'un algorithme basé sur les jointures. Il prend en entrée une représentation horizontale des données et s'assimile à une recherche en largeur. L'objectif de ce dernier est de réduire le nombre de séquences candidates en imposant que l'ensemble de leurs sous-séquences soient des séquences fréquentes. L'algorithme est itératif et teste à chaque étape toutes les séquences de taille k. Chaque itération se compose de deux étapes. La première étape définie l'ensemble des séquences candidates possibles de taille k à partir des séquences fréquentes de taille $k-1$. La seconde étape définit le support de chaque séquence candidate. Les séquences fréquentes de taille k sont alors identifiées. Cet algorithme présente l'avantage d'être facile à comprendre et permet de déterminer toutes les séquences fréquentes. Le principal inconvénient est son temps d'exécution et son utilisation de la mémoire importante. Cela est notamment dû au fait que la base de données est parcourue à chaque itération pour déterminer le support des séquences candidates de taille k [93]. Ainsi cet algorithme n'est pas le plus pertinent lorsque l'on est en présence d'une base de données de grande dimension et d'un grand nombre d'items possibles. La complexité temporelle est de $\mathcal{O}(m^2 n)$ [34] avec m le nombre d'éléments distincts et n le nombre de transactions.

Pour pallier ces défauts de l'algorithme aPriori, l'algorithme **ECLAT** [210] a été développé. Il s'agit d'un algorithme basé sur les arbres. Il prend en entrée une représentation verticale des données et peut être vu comme le parcours en profondeur de l'ensemble des données. Il s'agit d'un algorithme récursif. Pour la suite, nommons « tid-liste » l'ensemble des transactions contenant les items d'intérêt. En prenant l'exemple de la figure 2.11, la « tid-liste » de l'item $I501$ correspond à l'hospitalisation 1 du patient 1 et l'hospitalisation 1 du patient 2. L'algorithme débute par l'identification de la « tid-liste » de chacun des items présents dans la base. L'algorithme poursuit en réalisant l'intersection de la « tid-liste » d'un item avec chacune des « tid-liste » des autres items. De cette façon, une « tid-liste » est récupérée pour tous les couples d'items. Cette liste correspond à l'ensemble des transactions contenant les deux items. Seules les séquences présentant un support supérieur au seuil sont conservées. L'algorithme est répété pour identifier tous les n-uplets à partir du premier item testé puis pour tous les items initiaux. L'algorithme ECLAT ne parcourt la base qu'une seule fois pour identifier les « tid-listes » des séquences composées d'un seul élément. Toutefois, l'intersection de toutes les « tid-listes », qui peuvent être au plus de la taille de la base de données, entre elles peut rendre l'exécution longue.

Enfin l'algorithme **FP-Growth** est un algorithme de croissance de motifs. Il consiste en une recherche en profondeur comme l'algorithme ECLAT mais prend en entrée une représentation horizontale de la base de données. La base est parcourue deux fois. La première fois, le support de chaque séquence de taille 1 est déterminé. Les séquences de chaque transaction sont alors réorganisées selon l'ordre croissant des supports des séquences de taille 1. La base est parcourue une seconde fois pour construire un FP-arbre. Il s'agit d'un arbre composé d'une racine nulle. Les nœuds suivants correspondent au nom de l'item et au nombre d'occurrences de transaction contenant la sous-séquence définie par le chemin de la racine nulle à ce nœud. Cette représentation des données permet une représentation graphique de toutes les séquences présentes dans la base. Les séquences fréquentes peuvent ainsi être identifiées.

D'autres modèles ont été développés au cours du temps. Certain de ces modèles sont présenté figure 2.13. Une revue de la littérature [36] a été réalisée en 2018, elle a permis de mettre en évidence les différences de temps d'exécution de chacun des algorithme.

Extraction de motifs séquentiels (Sequential pattern mining) Les modèles que nous avons présentés dans la partie précédente ne permettent pas de prendre en compte l'ordre de survenue des éléments. Dans le domaine de la santé, cette approche peut être utile pour étudier l'ensemble des diagnostics utilisés lors d'une hospitalisation pour une cause spécifique. Dans ce cas, une transaction correspond à une hospitalisation et l'ensemble d'éléments associé est composé des diagnostics principaux, reliés et associés. Toutefois lorsque l'on s'intéresse aux parcours de soin, l'ordre d'apparition des éléments est important. Être hospitalisé pour insuffisance cardiaque puis pour la pose d'un stent ne décrit pas la même situation pour le sujet que d'être hospitalisé pour la pose d'un stent puis pour insuffisance cardiaque. Dans le premier cas, le stent peut être vu comme un traitement de l'insuffisance cardiaque, dans le second cas, on peut penser à l'échec de la pose. Dans cette problématique, la transaction peut être l'ensemble des hospitalisations ayant eu lieu sur une période de temps. L'ensemble des éléments associés à chacune des transactions correspond aux GHM de ces hospitalisations. Les algorithmes d'extraction des motifs séquentiels permettent de prendre en compte la notion d'ordre. Ces algorithmes ont été développés suivant deux approches, celle des algorithmes aPriori ou ECLAT et celle des algorithmes de croissance de motifs [144].

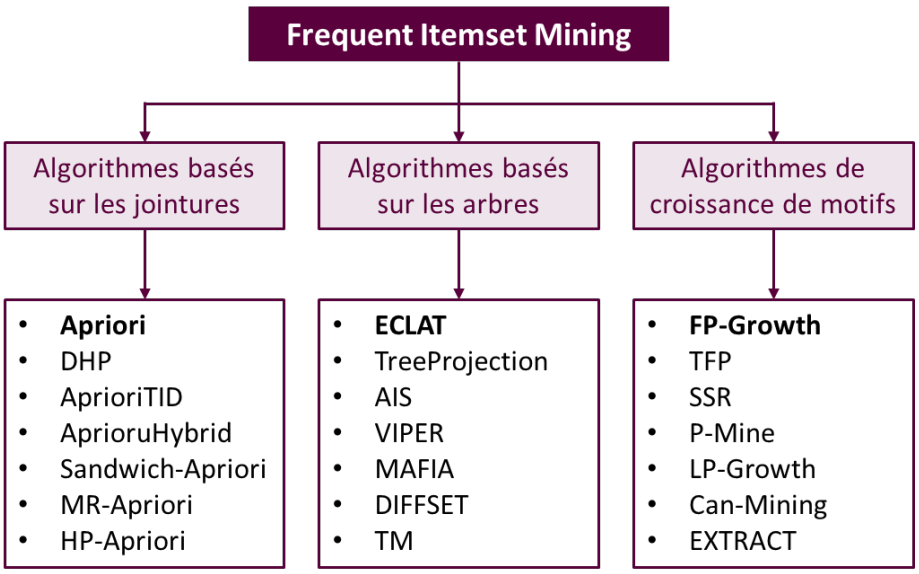

FIGURE 2.13 – Classification des algorithmes d'extraction d'éléments fréquents

Le premier algorithme développé pour l'extraction de motifs séquentiels est l'algorithme **GSP** (Generalized Sequential Patterns), voir [176] pour plus de détail. Cet algorithme est identique à celui aPriori à l'exception de la méthode de génération des séquences candidates. Supposons que l'on travaille avec les diagnostics d'insuffisance cardiaque, codé I50 et d'œdème pulmonaire, codé J81. Dans un premier temps l'algorithme considère les séquences de taille 1 : $\langle I50\rangle$, $\langle J81\rangle$. Les séquences candidates de taille 2 issues des deux précédentes séquence sont les suivantes : $\langle I50, J81\rangle$, $\langle J81, I50\rangle$, $\langle J81, J81\rangle$, $\langle I50, I50\rangle$, et $\langle (I50, J81)\rangle$. De cette façon, l'algorithme explore les séquences de taille de plus en plus grande de manière récursive. Les limites de cet algortihme résident dans la consommation importante de la mémoire due au stockage des séquences imposées par l'exploration en largeur, un temps d'exécution long pour les bases de grande dimension dû à la lecture répétée de la base et la génération de séquences qui n'existent pas dans la base de données [36].

Une alternative à l'algorithme GSP est l'algorithme **SPADE**. Ce dernier repose sur le même principe que le modèle ECLAT. Il utilise une représentation verticale de la base de données ce qui permet de ne garder en mémoire que les séquences fréquentes. Contrairement à l'algorithme ECLAT, la base est lue 3 fois. Les deux premiers balayages permettent de construire les séquences de taille respectivement 1 et 2. La troisième lecture permet de construire toutes les autres séquences. Ces séquences sont formées en joignant les séquences les unes aux autres. L'algorithme s'arrête lorsqu'aucune séquence fréquente ne peut plus être trouvée. Il propose une nouvelle représentation des motifs séquentiels par classe d'équivalence qui sont définies sur les préfixes de covariables. Deux séquences de taille k appartiennent à la même classe d'équivalence si leur préfixe de taille $k-1$ est le même. Par exemple les séquences $\langle 05M09, 05M09, 05K10\rangle$ et $\langle 05M09, 05M09, 05M10\rangle$ sont de classes équivalentes. L'algorithme **SPAM** [18] est une extension de l'algorithme SPADE proposée pour optimiser les jointures qui peuvent être couteuses dans le cas de séquences longues.

Les algorithmes SPAM et SPADE sont couteux à la fois par la génération d'un nombre important de sous-séquences et les opérations de jointure. Pour pallier ce problème, les algorithmes **CM-SPAM** et **CM-**

SPADE ont été proposés [79]. Ils permettent de diminuer l'effet de la génération des séquences candidates sur le temps d'exécution et l'utilisation de la mémoire. Pour ce faire, une matrice appelée « Co-occurence Map (CMAP) » qui contient toutes les sous-séquences fréquentes de taille 2 est définie. Cette matrice permet d'exclure au fur et à mesure certaines sous-séquences dès lors que les deux derniers éléments de cette sous-séquence ne forment pas une sous-séquence fréquente présente dans la matrice CMAP. Les sous-séquences exclues n'interviennent plus dans les jointures. Ces algorithmes sont plus rapides que les algorithmes présentés précédemment [79].

2.3.1.4 . Score de similarité

Nous avons vu, dans la partie précédente, des méthodes d'identification des séquences fréquentes. Pour pouvoir utiliser ces séquences fréquentes à des fins de prédiction, il est nécessaire de quantifier la proximité entre la séquence observée de chaque sujet et l'ensemble des séquences fréquentes identifiées. Ces méthodes sont en partie issues de la bioinformatique. Plusieurs scores existent pour mesurer cette similarité. L'approche la plus classique est celle basée sur les opérations nécessaires pour passer d'une séquence S_1 à une autre séquence S_2 c'est-à-dire aligner la séquence S_1 avec la séquence S_2. Les opérations sont de différentes natures comme la délétion, l'insertion, ou encore la substitution. Ces opérations permettent d'assurer un alignement entre deux séquences.

Le choix du score de similarité utilisé repose sur le type d'alignement qui est considéré [128]. On parle d'alignement global lorsque l'on cherche la similarité maximale entre les deux séquences S_1 et S_2. L'alignement est dit local lorsque cette similarité maximale est cherchée entre deux sous-séquences de S_1 et S_2. Deux notions supplémentaires doivent être prises en compte lors du choix du score de similarité : les tailles des deux séquences et la notion d'ordre des éléments dans la séquence. Dans l'étude des séquences relatives aux parcours de soin hospitalier, la notion d'ordre est primordiale comme nous avons pu en discuter dans la partie précédente. De plus il est légitime de considérer que toutes les séquences n'auront pas la même longueur. En ce sens, les mesures nécessitants des séquences de taille identique ainsi que celles ne prenant pas en compte l'ordre d'apparition des items dans la séquence ne seront pas décrites par la suite. C'est notamment le cas des mesures Cosine [152] ou Hamming [27] qui prennent en entrée des séquences de taille identique ou Jaccard [94] qui ne prend pas en compte l'ordre dans les séquences.

Les méthodes les plus connues permettant de mesurer la similarité globale (ou semi-globale) entre deux séquences sont les méthodes basées sur la plus longue sous-séquence commune. Nous les décrirons ici.

- La première est définie par la longueur de la plus longue sous séquence commune à S_1 et S_2. Plus cette sous-séquence est longue plus les deux séquences sont proches. Au plus la mesure est égale à la taille des séquences S_1 ou S_2, si ces deux séquences sont identiques.
- En 1965, Levenshtein [119] a introduit une distance qui peut être calculée grâce l'algorithme de Wagner et Fischer [197]. Elle repose sur le dénombrement des opérations d'insertion, de suppression ou de substitution d'éléments nécessaires pour convertir S_1 en S_2. Plus cette mesure est faible plus les séquences sont proches. Cette mesure vaut 0 si les deux séquences sont identiques. Plusieurs extensions de cette mesure ont été proposées, voir par exemple [48] pour la mesure de Damerau-Levenshtein qui permet de prendre en compte l'opération de transposition (échange de deux éléments consécutifs) ou des versions normalisées de la distance.
- En 1970, Needleman et Wunsch [140] ont proposé un nouveau score de similarité basé sur la program-

mation dynamique [103]. Cette programmation décompose successivement un problème complexe en parties plus petites pour faciliter sa résolution. Cette dernière permet d'obtenir un alignement maximum basé sur la construction d'une matrice. La mesure de Levenshtein peut être utilisée sur cet alignement.

Le développement de méthodes permettant de prendre en compte les mesures de similarité locale a principalement été motivé par l'étude des séquences de nucléotides ou de protéines où la comparaison de certaines régions n'est pas pertinente notamment si elles sont non codantes. L'étude de l'alignement local pour les parcours de soin peut mettre en avant certaines sous-séquences spécifiques. L'algorithme le plus utilisé pour mesurer la similarité locale est l'agorithme de Smith et Waterman [171]. Il s'agit d'une extension de celui de Needleman et Wunsch et utilise aussi la programmation dynamique. La différence réside dans la construction de la matrice permettant d'obtenir l'alignement optimal. D'autres méthodes peuvent être utilisées pour mesurer la similarité locale [208].

2.3.2 . Modèles de prédiction

Les scores de similarité calculés par les méthodes présentées dans la partie précédente permettent de quantifier la ressemblance entre les séquences fréquentes et les trajectoires de chacun des patients. Les variables constituées par les scores de similarité de chaque séquence fréquente peuvent alors être utilisé dans des modèles multivariés notamment pour l'évaluation de la survie.

Nous avons présenté dans la section 2.1.2 deux modèles d'estimation spécifique aux données de survie. Ces modèles reposent sur un certain nombre d'hypothèses qui peuvent être complexe à vérifier. De plus, l'estimation des paramètres de ces modèles peuvent s'avérer difficile en présence d'un grand nombre de co-variables. Ce cas peut être fréquent lors de l'analyse des trajectoires. En effet, plus le seuil utilisé pour la sélection des trajectoires fréquentes est faible plus le nombre de séquences sélectionnées est grand. Des méthodes d'apprentissage statistique en machine learning spécifiques peuvent être utilisées pour la prédiction de la survie.

Les méthodes de machine learning ont été principalement proposées pour des tâches de classification et de régression [24]. Mais ces méthodes peuvent s'étendre à des données des survie (présentant des censures) [199]. Elles peuvent être classées en quatre grandes classes [174] : les forêts aléatoires, le boosting de gradient, les machines à vecteur de support et les réseaux de neurones. Les trois premières classes sont décrites dans les paragraphes suivants. Nous n'avons pas utilisé de réseau de neurones dans ce travail. Ils ont cependant été étendu à l'analyse de survie, voir [164, 104, 72]. C'est une des extensions possibles de nos travaux.

2.3.2.1 . Les forêts aléatoires

La méthode des forêts aléatoires est un algorithme composite construit par plusieurs arbres de décision simples. La prédiction donnée par cette méthode est souvent la moyenne des résultats des prédictions de chacun des arbres.

Les arbres de décision sont une classe de modèles couramment utilisés en apprentissage automatique. Ils ont l'avantage d'être relativement simple à mettre en œuvre et d'être interprétable. Leur construction repose

sur deux notions : la règle de division et la règle d'arrêt. Un arbre de décision prend un ensemble d'entrée. Cet ensemble est divisé successivement selon la règle de division créant ainsi une série de nœuds. Les divisions sont stoppées suivant la règle d'arrêt. Les nœuds terminaux ainsi obtenus correspondent à une prédiction. Les forêts aléatoires sont souvent préférées aux arbres de décision, ces derniers étant généralement peu robustes aux changement dans les données. Introduites en 1995 par Ho [100], les forêts aléatoires ont été largement étudiées par Breiman et Cutler, voir [30]. Cette méthode repose sur deux principes : le « bagging » et la sélection aléatoire des variables. Le bagging (cf [29]) est une méthode d'ensemble où chaque modèle (arbre de décision dans ce cas) est entrainé sur un ensemble de données rééchantillonnées issues d'un tirage aléatoire avec remise des données observées. Les résultats obtenus sur des ensembles d'apprentissage différents sont ensuite agrégés. La sélection aléatoire des variables intervient au moment du calcul de la règle de division et cela à chaque nœud.

La spécificité des arbres de survie et par extension des forêts de survie aléatoire réside dans le choix de la règle de division et de la nature de la prédiction. Une revue des méthodes d'arbres de survie est proposée dans [28]. La règle de prédiction la plus couramment utilisée est celle basée sur le test d'hypothèse du log-rank, voir [26], et a été développée dans [106]. A chaque nœud la division retenue est celle qui maximise la dissimilarité des deux groupes c'est-à-dire qui maximise la statistique du log-rank. D'autres méthodes de division peuvent être utilisées comme celles basées sur les tests de Kolmogorov-Smirnov modifiés ou Gehan-Wilcoxon [38] ou sur la vraisemblance. Le plus souvent, la prédiction obtenue aux nœuds terminaux est une estimation de la fonction de survie. Cette estimation par des estimateurs non-paramétriques comme les estimateurs de Kaplan-Meier ou de Nelson-Aalen est une mesure naturelle dans les analyses de survie comme peut l'être la moyenne pour des problématiques de régression [102].

2.3.2.2 . Les méthodes de boosting de gradient

Le boosting de gradient est une technique générale qui permet d'optimiser de nombreuses fonctions de perte, fonctions qui quantifient l'écart entre le résultat prédit et le résultat observé (cf [82]). Le choix de la fonction de perte permet de résoudre des problèmes de classification et de régression, voir [154] pour plus de détails. Ces méthodes se basent sur la combinaison de modèles très simples qui sont généralement peu performant, en termes de prédictions, pour obtenir un modèle global meilleur.

Trois grands éléments constituent les méthodes de boosting de gradient : la fonction de perte, définie plus haut, les modèles simples appelés apprenants faibles, et la méthode de combinaison des apprenants faibles. L'utilisation d'une telle méthode dans l'analyse de survie diffère principalement sur le choix de la fonction de perte. En effet, les algorithmes de gradient boosting utilisent principalement des arbres en tant qu'apprenants faibles. Ces arbres sont générés itérativement et séquentiellement afin de prendre en compte les résultats des arbres générés précédemment. Plusieurs fonctions de pertes peuvent être utilisées pour l'analyse de données de survie. Une première fonction de perte repose sur la vraisemblance partielle du modèle de Cox [46, 45]. Il s'agit plus précisément du logarithme négatif de la vraisemblance partielle. L'indice de concordance (C-index) [37], souvent utilisé pour évaluer les performance d'un modèle de survie peut aussi être utilisé. Les prédictions obtenues sont des prédictions de rang.

2.3.2.3 . Machine à vecteurs de support

Les algorithmes de machine à vecteurs de support ou support vector machine (SVM) [44] ont été initialement développés pour des problèmes de classification puis étendus à des problèmes de régression, on réfère le lecteur a [64] pour une présentation générale.

Les premiers développement des SVM dans le cas d'analyse de survie visent à traiter le problème comme un problème de régression, voir [168, 111, 117] avec des ajustements pour prendre en compte la censure. Ces modèles permettent de prédire le temps de survie. D'autres extensions des SVM portent sur la prédiction des rangs relatifs afin d'optimiser directement le pouvoir discriminatoire des modèles [67, 189]. Cette approche considère l'analyse comme un problème de classification avec une variable à prédire qui est ordinale (voir [97]). Cette dernière correspond au rang des individus définis sur la base de leur temps de survie. Une version hybride inspirée de ces deux méthodes est proposée [190]. Elle combine les contraintes des deux tâches de régression et de classement. Cette méthode hybride permet une interprétation du temps de survie tout en optimisant la discrimination. Plusieurs études ont été menées pour comparer ces différentes méthodes [78, 199, 190, 130].

2.3.3 . Apport de la thèse

Pour rappel, cette seconde problématique a été motivée par la volonté d'étudier les trajectoires d'hospitalisation des patients nouvellement diagnostiqués pour une insuffisance cardiaque. Cette pathologie est une cause majeure des décès liés aux maladies de l'appareil circulatoire, deuxième cause de décès en France. L'identification des parcours de soin liés à un sur-risque de décès permettrait aux cliniciens d'être particulièrement attentifs aux patients présentant ces parcours de soin et éventuellement de modifier la prise en charge.

Nous proposons dans ce manuscrit de suivre la procédure d'analyse (Figure 2.9) présentée par Pinaire et ses coauteurs [147] dans le cadre de l'analyse des parcours de soins après un infarctus du myocarde. Toutefois, nous proposons d'utiliser des méthodes d'extraction des motifs séquentiels, d'analyse de la similarité et de prédiction des décès qui répondent plus aux problématiques et à la nature des données de santé. En effet, les algorithmes utilisés permettent de tenir compte de l'importance de l'ordre de survenue des différentes hospitalisations et l'analyse du décès comme une variables dépendante du temps.

Chapitre 3

Facteurs confusion continus et dépendants du temps : estimation débiaisée de l'ATT

Comme nous l'avons vu précédemment, l'estimation de l'efficacité d'un nouveau traitement est une étape incontournable pour sa mise sur le marché. Cette évaluation est toujours réalisée, dans un premier temps, sur des données d'essais cliniques construits dans cet objectif. Toutefois, évaluer l'efficacité de ce traitement dans des conditions de vie réelle présente un intérêt certain pour le maintient de ce dernier dans la stratégie thérapeutique. En effet, les autorités de santé peuvent demander une réévaluation de l'efficacité lors de l'arrivée sur le marché d'une nouvelle molécule, ou le laboratoire peut vouloir disposer de telles données pour alimenter son dossier de prix.

L'estimation de l'effet d'un traitement repose sur des analyses de causalité. En effet, le résultat obtenu doit garantir que la modification du résultat observé est la cause de l'utilisation du traitement. Les notions de causalité basées sur les concepts d'intervention et de contrefactuelle en présence de facteurs de confusion indépendants ou dépendants du temps ont été introduites dans le chapitre précédent (section 2.2.2.2). Des méthodes spécifiques existent pour assurer une interprétation causale des résultats dans des données qui n'ont pas été recueillies dans ce sens (qui ne sont pas des essais randomisés par exemple). Ces méthodes ont été présentées dans la section 2.2.2.3.

L'estimation de l'efficacité du traitement peut passer par l'estimation de l'effet moyen du traitement chez les patients traités. Le modèle proposé par Gran et ses collaborateurs permet une estimation de cet effet, voir la section 2.2.2.4, dans le cas d'une exposition au traitement et de facteurs de confusion dépendants du temps. Nous avons étendu le modèle à l'étude de l'effet d'une intervention sur un résultats qui peut se répéter dans le temps, comme les rechutes de cancer. Nous avons montré que les résultats de ces estimations étaient biaisés et avons proposé une correction. Ces travaux sont présentés dans les parties suivantes et sont issus du manuscrit de l'article « Debiasing the estimate of treatment effect on the treated with time-varying counfounders », en révision au journal Statistical Methods in Medical Research.

3.1 - Introduction

Health data warehouse are increasingly available and could be an alternative to the prospective cohort studies traditionally designed for epidemiological research because data are immediately available and projects are thus less time consuming and expensive to run. In most data sources such as health registries or electronic health records, patients are followed over time, resulting in longitudinal data. Statistical analyses are therefore performed in the presence of time-dependent covariates and time-varying treatment. The development of longitudinal databases makes the study of recurring events easier such as repeated events or events that can be measured using counting processes. For example, they are useful when studying rehospitalizations, relapses in cancer, asthma attacks, or multiple sclerosis relapses.

Time-dependent confounding occurs when a covariate is both affected by past treatment and associated with future treatment choice and outcome. Many longitudinal studies aim to estimate the causal effect of time-varying exposure on the outcome. The main issue of these analyses is to handle such confounding. Specific methods address this issue by taking into account time-confounding and thus avoid biased estimates [49, 109, 184]. The most common method for dealing with time-dependent confunding is the inverse probability weighted marginal structural Cox model [75]. Instrumental variable methods [101] or methods derived from g-methods can also be used.

Two types of treatment effect estimator, the average total treatment effect (ATE) and the average treatment effect on the treated (ATT), can be identified. The most commonly used, the ATE, aims identifying the treatment effect in the whole population, eligible for treatment. It allows to answer the question "What would happen if all the patients had received the evaluated treatment, at every timepoint as compared with they had received its comparator ?". The ATT aims at identifying the average treatment effect only in patients who actually received the treatment of interest. With this method, the effect of treatment is not assessed in patients who never received the treatment. The choice of estimator depends on causal research question of interest and on its availability for the method used. For example, it may be more interesting to evaluate the treatment effect on the population that received that treatment rather than on the whole population. Indeed, in clinical practice, the prescription of a treatment is motivated by a set of indicators, as patients clinical characteristics or the likelihood that the treatment will be effective for that patient. In addition, depending on the data available, only information on treated patients may be available.

The ATT can be estimated mainly by g-estimation of structural nested models [198]. However, this could be complex and prevents from its common use in practice [191]. Gran et al. proposed an easy-to-implement method that allows to estimate the treatment effect in treated patients and to study intermediate time-dependent covariate trajectories [90]. This model was presented for end-point outcome and an unique time-varying covariate. It consists of two steps. The first step is to model time-varying covariates on the time interval where patients have not yet been treated. Then, the causal treatment effect is estimated by modelling the untreated evolution of time-varying covariates. Finally, treatment effect is estimated using intensity regression models. The proposed method is developed for additive intensity regression models [1]. When studying causality in censored data, this method allows a more explicit approach to processes [6] and explicit derivations that are not available for the Cox model [180, 192, 135, 178]. This model seems to be a good alternative to the Cox model [4, 116].

We propose in this paper to generalize the Gran's model for a potentially repeated outcome and in the

presence of multiple time-dependent covariates and baseline covariates. Moreover, we propose a debiased estimate of the ATT based on the error produced during the modelling of the covariates where patients are untreated. Simulation analyses show that the corrected estimator that we propose outperforms Gran's uncorrected estimator.

The paper is organized as follows. Section Model and notations introduces the different notations and assumptions used in the rest of the paper. Section Additive intensity model for the treatment effect estimation presents the additive intensity regression model and the proposed correction for a general time-dependent covariates modelling method. Section Example with a vector autoregressive model VAR(1) specifies the correction in case of vector autoregressive model. Section Simulation study summarises results from simulations. Section Application presents the results of the models on intensive care real-life data from MIMIC-III data base [108].

3.2 - Model and notations

3.2.1 . Model

Consider an individual who suffers from a specific disease with a follow-up over time. At time S, he/she initiates his/her treatment. We assume that once started the treatment is continued, so that the treatment process D is null before S and equals 1 after S. Together with the treatment, we observe d_Z-dimensional baseline covariates $Z = \left(Z^1, \ldots, Z^{dz}\right)$ and time-dependent d_X-dimensional covariates $t \mapsto X(t) = (X^1(t), \ldots, X^{d_X}(t))$. We focus on pathologies for which the outcome of interest can be measured via a counting process N ($N(t)$ being the number of events recorded in $[0, t]$), see examples in Introduction and Application sections. We assume also that N verifies the Aalen additive intensity model [1].

Assumption 1 *With respect $\left(\mathcal{F}_t^\star\right)_{t\geq 0}$, the historical filtration spanned by Z, X, D, and N, we assume the process X is predictable and that N has the following intensity :*

$$\mu^*(t) = \alpha_0^*(t) + \alpha_Z^*(t)Z + \alpha_X^*(t)X(t) + \alpha_D^*(t)D(t).$$

As in [90], we will assume further for simplicity in what follows that independent censoring can occur. Censoring can be assumed to be independent when analyzing data from intensive care unit or administrative databases but not in all pathologies database. Methods allowing dependent censoring such as adjustment using inverse probability of censoring weighting should be used [158].

3.2.2 . Counterfactual quantities, assumptions

We now consider that we can intervene on the treatment decision at time S. Given the definitions introduced above, we have with probability 1

$$D(S) = 1,\ D(s) = 0 \text{ for } s < S \text{ and } \ D(t) = 1 \text{ for } t \geq S.$$

However assuming that the positivity assumption, see e.g. [99] for example, is fulfilled implies that

$$0 < \mathbb{P}\big(D(t) = 1 | D(t-) = 0, \mathcal{F}_{t-}\big) < 1.$$

At any time t before treatment initiation, it can be started or not with a positive probability. This excludes certain pathologies. In practice the counterfactual reasoning requires this positiving assumption which has to be checked on the data.

In the hypothetical situation where the treatment is not initiated at S, we denote by $X^{0|S}$ the covariate process after S. This counterfactual process corresponding to what would have been observed after S. It is important to remark that it cannot be observed, so that its values will have to be estimated (see Section Example with a vector autoregressive model VAR(1)). The potential covariate process in the situation where the treatment is initiated at S is denoted by $X^{1|S}$ (see Figure 3.1).

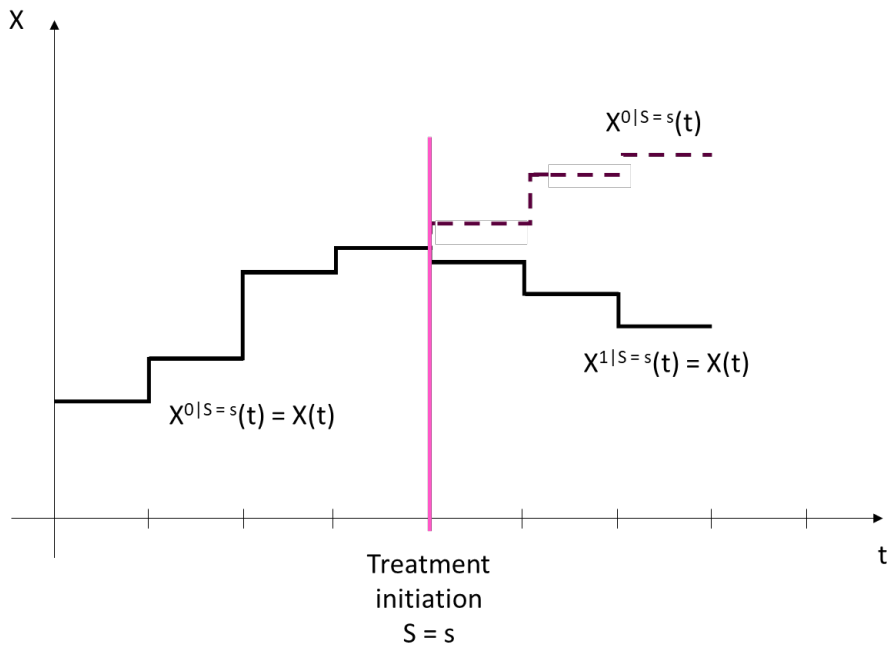

FIGURE 3.1 – Counterfactual process

We also introduce the two potential counting processes $N^{1|S}$ recording the event history after S when

the treatment is actually initiated at S and $N^{0|S}$ when the treatment is not initiated. Their respective intensities are denoted by $\mu^{1|S}$ and $\mu^{0|S}$.

The consistency assumption in this case, see [99], implies that for all $t \geq S = s$

$$X^{1|S=s}(t) = X(t) \text{ and } N(t) = N^{1|S=s}(t).$$

We also assume conditional exchangeability conditionally to $S = s$, for all $a = 0, 1$, all $t \geq s$ and conditionally to the past $\mathcal{F}_{s-}$:

$$\mathbb{E}(dN^{a|S=s}(t)|D(s), \mathcal{F}_{s-}) = \mathbb{E}(dN^{a|S=s}(t)|\mathcal{F}_{s-}) \text{ and}$$
$$\mathbb{E}(X^{a|S=s}(t)|D(s), \mathcal{F}_{s-}) = \mathbb{E}(X^{a|S=s}(t)|\mathcal{F}_{s-}).$$

This means that at time $S = s$ conditionally to the past $\mathcal{F}_{s-}$, if two perfectly similar people (with exactly the same history of treatment and exposure) one receiving the treatment, the other one not receiving the treatment, have a counterfactual outcome which is the other one actual observed outcome. Under Assumption (1), this can be rewritten, for $t \geq S = s$, as

$$\mu^{1|S=s}(t) = \alpha_0^*(t) + \alpha_D^*(t) + \alpha_Z^* Z + \alpha_X^*(t) X^{1|S=s}(t) \text{ and}$$
$$\mu^{0|S=s}(t) = \alpha_0^*(t) + \alpha_Z^* Z + \alpha_X^*(t) X^{0|S=s}(t).$$

The exchangeability assumption can hold only if no unmeasured covariate influences our model. This is a strong assumption in practice but is common in causality reasoning, see e.g. [90, 99] among many others. Following [99], the clinician opinion expertise is key in proposing a list of covariates associated with treatment indication to be included in the baseline covariates Z and the time dependant covariates process X so that this assumption is fulfilled at least approximately conditional on all variables recorded.

A stronger assumption can be found in the literature [99, 116] : conditionally to $S = s$, for all $a = 0, 1$ and conditionally to the past $\mathcal{F}_{s-}$

$$X^{a|S}(t) \perp\!\!\!\perp D(s) \text{ and } N^{a|S}(t) \perp\!\!\!\perp D(s), \text{ for all } t \geq s.$$

3.2.3 . ATT definition and causal estimate

We assume an additive intensity regression model for the outcome processes. Under the model and the assumptions presented in previous paragraph, the time varying causal intensity difference can be then defined

as

$$\begin{aligned} d^*(t,s) &= \mu^{1|S=s}(t) - \mu^{0|S=s}(t) \\ &= \alpha_D^*(t)D(t) + \alpha_X^*(t)(X^{1|S=s}(t) - X^{0|S=s}(t)). \end{aligned}$$

With these definitions, we can rewrite the intensity of N as

$$\mu^*(t) = \alpha_0^*(t) + \alpha_Z^* Z + \alpha_X^*(t)X^{0|S=s}(t) + d_i^*(t,s)D(t),$$

and, for $t \geq S$, the time varying ATT as

$$\text{ATT}(t) = \mathbb{E}[d^*(t,S)|t \geq S] = \alpha_D^*(t) + \alpha_X^*(t)\mathbb{E}\big[X^{1|S}(t) - X^{0|S}(t)\big|t \geq S].$$

This has been described in [90] and strongly relies on the Aalen additive intensity model of Assumption (1). We refer the reader to [4] for more discussions on that matter.

3.3 - Additive intensity model for the treatment effect estimation

Our data consists in the observation for n independent individuals $i = 1, \dots n$ of Z_i, X_i, S_i, D_i and N_i. Each individual is followed between $t = 0$ and $t = \tau_i$ which can be the end time of the study. The intensity of N_i can be written as :

$$\begin{aligned} \mu_i^*(t) &= \alpha_0^*(t) + \alpha_Z^* Z_i + \alpha_X^*(t)X_i^{0|S=S_i}(t) + d_i^*(t,S_i)D_i(t) \\ &= W_i(t)^\top A^*(t). \end{aligned}$$

We can recognize the writing of standard regression where treatment effect is adjusting on treatment, the baseline covariates and the counterfactual covariates, $X_i^{0|S=S_i}(t)$. We remark that in that case the ATT is the treatment regression coefficient.

3.3.1 . Identification of the treatment effect estimate

The aim of these section is to define an estimator of A^* in the hypothetical situation, false in practice, where the trajectories of the counterfactual covariate processes $X_i^{0|S=S_i}$ for $i = 1, \dots, n$ and $t \geq S_i$, or equivalently the processes W_i, would be observed. We simply show that the traditional empirical least-squares risk (see e.g. [83]) for the additive intensity model would lead to a correct estimator of A^*. Let

$$A(t) = (\alpha_0(t), \alpha_Z, \alpha_X(t), d_i(t, S_i))^\top, \forall t \geq 0,$$

be a candidate for the estimation of $A^\star$. The squared risk of $A \in \mathbb{R}^{d_X+d_Z+2}$ is defined as, see [83] :

$$\begin{aligned} r_n(A) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t) W_i(t)^\top A(t) dt \\ - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t) dN_i(t). \end{aligned} \tag{3.1}$$

The fact that

$$\widehat{A} = \arg\min_A r_n(A) \tag{3.2}$$

is a good candidate for the estimation of A^* can be checked by computing the expectation of $r_n(A)$. Calculations presented in Supplementary Materials, section 1.1, allow us to deduce that A^* can be expressed as :

$$A^* = \arg\min_A \mathbb{E}[r_n(A)],$$

which shows that $\widehat{A}$ is an estimator of A^*.

3.3.2 . Estimation of counterfactual covariates

In our situation, the counterfactual trajectories $X_i^{0|S=S_i}$ cannot be observed, as in the data, patient i is treated as time S_i. To estimate $A^\star$, we need to estimated these trajectories. We give an exemple of such estimations in Section Example with a vector autoregressive model VAR(1). We consider for now that we can model the trajectories and estimate the values of $X^{0|S=S_i}(t)$ for $t \geq S_i$ by $\widetilde{X}_i^{0|S=S_i}(t)$. Being an estimation, $\widetilde{X}_i^{0|S=S_i}(t)$ is not equal to $X^{0|S=S_i}(t)$ but we assume that we can write :

$$\widetilde{X}_i^{0|S=S_i}(t) = X_i^{0|S=S_i}(t) + \epsilon_i(t), \text{ for all } t \geq S_i, \tag{3.3}$$

where $\epsilon_i(t)$ is centered and has finite variance. We finally denote by $\widetilde{W_i}(t)$ the vector $W_i(t) + \xi_i(t)$ with $\xi_i(t) = (0, 0_{\mathbb{R}^{d_z}}, \epsilon_i(t), 0)$ for $t \geq S_i$ and $\xi_i(t) = (0, \dots, 0)$ for $t < S_i$.

Mimicking the computation in the above paragraph, let us introduce $\widetilde{R_n}(A)$ the squared risk of A using

counterfactual covariates estimate $\widetilde{W}_i(t)$:

$$\widetilde{R_n}(A) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \widetilde{W}_i(t)\widetilde{W}_i(t)^\top A(t)dt \\ -\frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \widetilde{W}_i(t)dN_i(t).$$

Computations detailed in Equation (1) of Supplementary materials allow to write $\widetilde{R_n}(A)$ in function of $r_n(A)$ using Equation (3.3) as

$$r_n(A) = \widetilde{R}_n(A) \\ + \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} \alpha_X(t)^\top \epsilon_i(t)\epsilon_i(t)^\top \alpha_X(t)dt. \tag{3.4}$$

Using Equation (3.4), the expectation of $\widetilde{R_n}$ is given by the following expression :

$$\mathbb{E}[\widetilde{R_n}(A)] = \mathbb{E}[r_n(A)] \\ - \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} \alpha_X(t)^\top \mathbb{E}[\epsilon_i(t)\epsilon_i(t)^\top]\alpha_X(t)dt. \tag{3.5}$$

Hence $\arg\min_A \mathbb{E}[\widetilde{R}_n(A)] \neq A^*$.

In that sens, the proposal of Gran et al. [90] is biased. We will introduce in the following section a novel estimator of A^* (and the ATT) taking into account the counterfactuals modelling error.

3.3.3 . Debiased treatment effect estimate

We showed that minimizing $\widehat{R_n}$ would had to a biased estimated. We propose in this paragraph to correct this bias. Equation (3.5) implies that

$$\widehat{R_n}(A) = \widetilde{R_n}(A) + \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} \alpha_X(t)^\top \mathbb{E}[\widehat{\epsilon_i(t)\epsilon_i(t)}^\top]\alpha_X(t)dt$$

is the correct loss to estimated A^*. However, as the unknown expectations $\mathbb{E}[\epsilon_i(t)\epsilon_i(t)^\top]$ appear, it cannot be directly used. For some models for the $X_i^{0|S=S_i}$ trajectories, these expectations can however be estimated. This is the case with the vector autoregressive model presented in the next section. Now considering that

we have access to such an estimation $\mathbb{E}[\epsilon_i(t)\epsilon_i(t)^\top]$ for all i and $t \geq S_i$, we can defined our loss as :

$$\widehat{R_n}(A) = \widetilde{R_n}(A) + \widehat{\mathsf{bias}}(A),$$

with

$$\widehat{\mathsf{bias}}(A) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} \alpha_X(t)^\top \mathbb{E}[\epsilon_i(t)\epsilon_i(t)^\top]\alpha_X(t)dt. \tag{3.6}$$

Finally, the debiased estimator of A^* is given by :

$$\begin{aligned}
\widehat{A}_{\mathsf{debiased}} &= \arg\min_A \widehat{R_n}(A) \\
&= \arg\min_A \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)\left[\widetilde{W_i}(t)\widetilde{W_i}(t)^\top \right. \\
&\quad \left. - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \mathbb{E}[\epsilon_i(t)\epsilon_i(t)^\top] & \vdots \\ 0 & \cdots & 0 \end{pmatrix}\right] A(t)dt \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)\widetilde{W_i}(t)dN_i(t).
\end{aligned} \tag{3.7}$$

We now propose a simple model for the continuous time-varying d_x-dimensional covariate. X_i in which the estimation $\widehat{\mathsf{bias}}(A)$ is easily computable.

3.4 - Example with a vector autoregressive model VAR(1)

3.4.1 . Explicit writing of the bias

In the following, the observation period from 0 to $\max_i \tau_i$ is split into observed time intervals defined by following times $t_0 = 0, \ldots, t_k, \ldots t_K = \max_i \tau_i$. We assume that the time varying covariates are constant over these intervals such that for all i, k and $t \in [t_k, t_{k+1})$, $X_i(t) = X_i(t_k)$, and by extension $W_i(t) = W_i(t_k)$.

Considering the case of continuous time-varying covariates, the simplest model is the vector autoregressive model [126]. We explicit the computation in the case of the VAR(1) model :

$$X_i(t_k) = \beta_0^{i\star} + \Pi^\star X_i(t_{k-1}) + \omega_i(t_k),$$

for all i and t, where $\Pi^\star$ is the coefficient matrix and ω_i is an unobservable zero mean white noise vector process with time invariant covariance matrix $\Sigma^\star$.

Following the literature on VAR model, the prediction of $X(t_{k+l})$ using the past up t_k is given iteratively by :

$$\begin{aligned}\widetilde{X}_i(t_{k+l}) &= \beta_0^{i\star} + \Pi^\star \widetilde{X}_i(t_{k+l-1}) \\ &= \sum_{j=0}^{l-1} \Pi^{\star j}\beta_0^{i\star} + \Pi^{\star l} X_i(t_k).\end{aligned}$$

The modelling error can be expressed as follow :

$$X_i(t_{k+l}) - \widetilde{X}_i(t_{k+l}) = \sum_{j=0}^{l-1} \Pi^{\star j}\omega_i(t_{k+l-j}), \tag{3.8}$$

with $\Pi^0 = \mathbf{I}_{d_X}$ the identity matrix of size d_X. The mean squared error (MSE) matrix for $\widetilde{X}_i(t_{k+l})$ is written as :

$$\Sigma(l) = \mathsf{MSE}(X_i(t_{k+l}) - \widetilde{X}_i(t_{k+l})) = \sum_{j=0}^{l-1} \Pi^{\star j}\Sigma{\Pi^\star}^{\top j}. \tag{3.9}$$

The parameters of the VAR(1) process are estimated using multivariate least squares [127]. Estimated parameters is noted $\widehat{\Pi}$ and $\widehat{\beta}_0^i$. The best linear predictor of $X_i(t_{k+l})$ becomes :

$$\widehat{\widetilde{X}}_i(t_{k+l}) = \widehat{\beta}_0^i + \widehat{\Pi}\widehat{\widetilde{X}}_i(t_{k+l-1}). \tag{3.10}$$

Finally, the estimated MSE matrix of the 1-step forecast is computing using equation (3.9) as :

$$\widehat{\Sigma}(l) = \sum_{j=0}^{l-1} \widehat{\Pi}^j\widehat{\Sigma}{\widehat{\Pi}}^{\top j}.$$

3.4.2 . Algorithm

Our algorithm combines two step.

3.4.2.1 . First step : counterfactual estimates

1. After filtering, only observations where patients are **not treated** are kept.
2. For each time-varying covariates j, the matrix in which $t^i_{\max}$ is the last time without treatment for the patient i is designed, i.e the discret observe time before S_i.
3. The linear regression with the design matrix of Equation (3.11) and the vector of predicted outputs $(X^j_1(1), \cdots, X^j_1(t^1_{\max}), \cdots, X^j_n(1), \cdots, X^j_n(t^n_{\max}))$ is computed.
4. Thus we get the estimates of (Σ_{jj}), $(\Pi_{j,q})_{1\leq q\leq d_X+d_Z}$ and β^i_0.
5. Counterfactuals are modelled following Equation (3.10) using estimates of Π and β^i_0, and are denoted $\widehat{\widetilde{X_i}}(t_k)$ for $i = 1, \ldots, n$ and $t_k \geq t_i$.

$$\begin{pmatrix} 1 & Z^1_1 & \cdots & Z^{d_Z}_1 & X^1_1(0) & \cdots & X^{d_X}_1(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z^1_1 & \cdots & Z^{d_Z}_1 & X^1_1(t^1_{\max}-1) & \cdots & X^{d_X}_1(t^1_{\max}-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z^1_n & \cdots & Z^{d_Z}_n & X^1_n(0) & \cdots & X^{d_X}_n(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z^1_n & \cdots & Z^{d_Z}_n & X^1_n(t^n_{\max}-1) & \cdots & X^{d_X}_n(t^n_{\max}-1) \end{pmatrix} \tag{3.11}$$

3.4.2.2 . Second step : ATT estimate with modelled counterfactuals

The ATT estimate is implemented in different analysis softwares such as the `aalen` function of the `timereg` package [133, 163]. We implemented our own method to apply our correction.

— The ATT is estimating by solving Equation (3.2) and (3.7).

3.5 - Simulation study

In this section, ATT estimations with and without correction are compared using simulations in the VAR(1) model. This simulation study is based on the simulation setup proposed in [90] . Data are simulated to mimic a cohort analysis where individuals are under risk of having an event of interest which can be prevented or delayed by treatment. Several covariates are simulated over time. Their values depend on treatment status, and reciprocally, treatment initiation depends on these values, see Figure 3.2. The probability of having the event is then affected by both the treatment status and the covariate values.

3.5.1 . Data generation

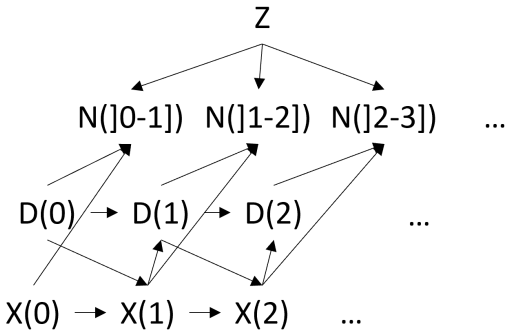

FIGURE 3.2 – Directed Acyclic Graph

Simulation setup is motivated in [90] by an application on the effect of antiretroviral therapy (HAART) on the time to AIDS onset or death. We have extended these simulations with several time-varying covariates and baseline covariates.

The data are simulated on time intervals $[t_k, t_{k+1})$ with pre-specified time points $t_k = 0, 1, ...11$. On each time interval, we define event status $N(t_k)$, treatment status $D(t_k)$, d_X time-varying covariates, $X(t_k) = (X^1(t_k), \ldots, X^{d_X}(t_k))$ and d_Z baseline covariates, Z. Dataset with one, three and six time-varying covariates are simulated.

At initiation, all patient are untreated. Initial time-varying covariates are simulated as uniform distribution. Three baseline covariates are simulated to mimic age as numbers uniformly distributed, gender as a binomial distribution and a comorbidity score as a Poisson distribution.

On the interval where the treatment has not yet been initiated, each covariates $X^j(t_k)$ will decrease steadily according to the following equation $X(t_{k+1}) = \kappa_{D0} X(t_k) + \mathcal{N}(0_{\mathbb{R}^d}, \Sigma)$ with κ_{D0} a $d_X \times d_X$ matrix and $\Sigma = \{\sigma_{jj}\}_{1 \leq j \leq k}$. While from treatment initiations, $X^j(t_k)$ increase over time following the equation $X(t_{k+1}) = X_i(t+1) = X_i(t) + \kappa_{D1}(\sqrt{1000} - X_i(t)) + \mathcal{N}(0_{\mathbb{R}^d}, \Sigma)$ with κ_0 a d_x-dimensional vector and κ_{D1} a $d_X \times d_X$ matrix. Treatment initiation $D(t_k)$ is simulated according to a Bernoulli distribution with parameter $m \sum_{j=1}^{d} \lambda_j e^{\sum_{j=1}^{d} \lambda_j X^j(t_k)})$. We assume that once treatment is initiated, the patient remains on treatment until the end of the follow-up.

Finally, treatment status and time-varying covariates act on the number of experienced events, $N(]t_k, t_{k+1}])$ experienced on $]t_k, t_{k+1}]$. The event times are simulated according to a non-homogeneous Poisson process with intensity function $\delta D(t_k) + \delta_0 + \sum_{j=1}^{d_Z} \delta_{Zj} Z^j + \sum_{j=1}^{d} \delta_{Xj} X^j(t_k)$. This process is generated using acceptance-rejection method called thinning [120]. All patients are followed up until time $t_k = 11$. The relation between covariates, treatment and outcome is summarized by the direct acyclic graph of Figure 3.2.

At the same time, the "true" counterfactual values are simulated by following the evolution of the covariates when $D(t_k) = 0$. The simulation equation is as follows : $\widetilde{X}(t_{k+1}) = \kappa_{D0} \widetilde{X}(t_k) + \mathcal{N}(0_{\mathbb{R}^d}, \Sigma)$ whatever the treatment status. The simulation algorithm is presented in Supplementary Materials, see Section 2.

The number of time-varying covariates varied in the simulated framework. The following results are based on simulations with 1 time-varying covariate affecting D and N, and simulations with three and six covariates with $d/3$ covariates affecting D and N, $d/3$ covariates affecting only D and $d/3$ covariates affecting only N. Finally, the parameter Σ_{jj} was set to the values 0.4, 0.8, 1.2 and 1.6. The values of parameters used in

the simulations are given in the appendix, see Section 8.2. We ran the simulations 100 times with the same parameters on data sets of 1000 individuals. All simulations were done in R version 4.1. Codes used for the simulations are available on gitlab `https://gitlab.com/camille.nevoret/att-estimation`.

3.5.2 . Analysis

As presented in the section Algorithm, the first step in the simulation study is the modelling of counterfactuals. Within the VAR(1) model, we estimated Σ and κ_{D0}, used to calculate the correction of the ATT estimator.

The data set of the 100 simulation repetitions with the true counterfactual values is used to evaluate the "true" ATT value by Monte Carlo. At each time t, we note $d^{MC}(t)$ the "true" ATT. All ATT estimates for each of the 100 simulations are compared to this value. We use the mean integrated square error (MISE) to calculated the error between ATT estimate and the true ATT. For a given estimator $\hat{d}^l(t)$ for the l-th simulation, the MISE is given by :

$$\text{MISE}^l = \mathbb{E}\left[\int_0^{\tau=11} (\widehat{d}^l(t) - d^{MC}(t))^2 dt\right].$$

For each of the 100 simulations, we estimate ATT using five different estimators.

— The first estimator is calculated using the `aalen` function of the R `timereg` package. Counterfactual data use for this estimate are the one simulated as "real counterfactuals". Thus, there are no errors related to the modelling of counterfactuals introduced in the ATT estimation. In this sense, the other estimators, based on modelled counterfactuals cannot make more accurate estimations than this one. It is thus uses as a reference.

— Then, two uncorrected estimators are presented. The first one corresponds to the one presented in [90] using `aalen` function and modelled counterfactuals. The second one is obtain using algorithm presented section Algorithm.

— Finally, two corrected estimators are presented. They use both the algorithm presented section Algorithm and modelled counterfactuals. The first one use the true time invariant covariance matrix, Σ, and the true κ_{D0}. In real life data, they are estimated when modelling the counterfactuals. This is done in the second estimator.

Wilcoxon tests for the difference in distribution are performed between the corrected estimator with estimated parameters and each uncorrected estimators.

3.5.3 . Simulation results

Table 3.1 shows the mean integrated square error for three ATT estimates : the two uncorrected estimators and the corrected estimator using estimate parameters $\widehat{\Sigma}$ and $\widehat{\kappa}_{D0}$. We observe that, whatever the number of covariates, the true mean integrated square error increases with Σ. Data in bold are the lowest MISEs. The symbol * represents a significative difference given by the Wilcoxon test.

For all simulation parameters, ATT estimates using the `aalen` function [133, 163] and using our recoded algorithm are very close. Our corrected estimator is always better than uncorrected estimated. We notice

TABLE 3.1 – Mean MISE $\pm$ standard deviation for corrected and uncorrected estimators (* : Significant Wilcoxon test between corrected estimator with estimated parameters and other estimators, boldface character : the smaller MISE of the corrected and uncorrected estimators)

	Corrected estimator with $\widehat{\Sigma}$	Uncorrected estimator	
		With `aalen` function	Recoded estimate
1 Covariate			
$\Sigma_{k,k} = 0.4$	$\mathbf{0.034 \pm 0.038}$	$0.045^\star \pm 0.047$	$0.044^\star \pm 0.047$
$\Sigma_{k,k} = 0.8$	$\mathbf{0.136 \pm 0.131}$	$0.178^\star \pm 0.155$	$0.177^\star \pm 0.155$
$\Sigma_{k,k} = 1.2$	$\mathbf{0.412 \pm 0.384}$	$0.533^\star \pm 0.445$	$0.527^\star \pm 0.443$
$\Sigma_{k,k} = 1.6$	$\mathbf{0.787 \pm 0.730}$	$1.026^\star \pm 0.853$	$1.014^\star \pm 0.848$
3 Covariates			
$\Sigma_{k,k} = 0.4$	$\mathbf{0.040 \pm 0.046}$	0.045 ± 0.051	0.045 ± 0.051
$\Sigma_{k,k} = 0.8$	$\mathbf{0.224 \pm 0.237}$	0.257 ± 0.265	0.256 ± 0.264
$\Sigma_{k,k} = 1.2$	$\mathbf{0.518 \pm 0.548}$	0.608 ± 0.604	0.604 ± 0.603
$\Sigma_{k,k} = 1.6$	$\mathbf{1.381 \pm 1.299}$	1.598 ± 1.434	1.587 ± 1.429
6 Covariates			
$\Sigma_{k,k} = 0.4$	$\mathbf{0.075 \pm 0.080}$	0.081 ± 0.095	0.081 ± 0.096
$\Sigma_{k,k} = 0.8$	$\mathbf{0.320 \pm 0.360}$	0.338 ± 0.376	0.337 ± 0.377
$\Sigma_{k,k} = 1.2$	$\mathbf{0.901 \pm 1.012}$	0.949 ± 1.043	0.947 ± 1.045
$\Sigma_{k,k} = 1.6$	$\mathbf{2.448 \pm 2.316}$	2.567 ± 2.404	2.560 ± 2.406

significant differences between these estimators, for all simulation scenarios with one covariate.

Table 3.2 allows to compare our corrected ATT estimate with estimated parameters, which is the one that can be used on real data, to corrected the ATT estimate with known parameters and the estimate using the "real counterfactuals". It appears that MISE of corrected estimates are very close which shows that counterfactual modelling lead to good estimates of Σ and κ_{D0}. We can also note large standard deviations of the order of the means of the MISE. These large standard deviations do not seem to be inherent to our estimation method. Indeed, we observe the same type of deviation for the estimate using "real counterfactuals".

Finally, for all simulation parameters tested, the corrected estimators appear to behave like the estimator based on the "real counterfactuals" and perform better than the uncorrected estimators.

3.6 - Application

Now, let us apply method to real life data from the open access MIMIC-III database [108]. This database contains deidentified data of patients admitted to the critical care unit of the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. Over this period, 53,423 distinct admissions of adult patients were identified. Data includes demographics data, laboratory and microbiology test results, diagnosis and procedure codes, bedside monitoring data (vital signs, waveforms, trends...), fluids, medications.

The focus here is on patients with sepsis. These patients are identified according to the algorithm defined

TABLE 3.2 – Mean MISE $\pm$ standard deviation for corrected estimators and estimator obtained using "real counterfactuals"

	Corrected estimator estimator		Estimate using "real counterfactuals"
	with $\widehat{\Sigma}$	with Σ	
1 Covariate			
$\Sigma_{k,k} = 0.4$	0.034 ± 0.038	0.034 ± 0.038	0.034 ± 0.034
$\Sigma_{k,k} = 0.8$	0.136 ± 0.131	0.136 ± 0.131	0.060 ± 0.066
$\Sigma_{k,k} = 1.2$	0.412 ± 0.384	0.412 ± 0.384	0.186 ± 0.168
$\Sigma_{k,k} = 1.6$	0.787 ± 0.730	0.783 ± 0.726	0.315 ± 0.436
3 Covariates			
$\Sigma_{k,k} = 0.4$	0.040 ± 0.046	0.040 ± 0.046	0.041 ± 0.044
$\Sigma_{k,k} = 0.8$	0.224 ± 0.237	0.224 ± 0.237	0.170 ± 0.237
$\Sigma_{k,k} = 1.2$	0.518 ± 0.548	0.516 ± 0.546	0.283 ± 0.379
$\Sigma_{k,k} = 1.6$	1.381 ± 1.299	1.371 ± 1.291	0.652 ± 0.956
6 Covariates			
$\Sigma_{k,k} = 0.4$	0.075 ± 0.08	0.075 ± 0.079	0.071 ± 0.060
$\Sigma_{k,k} = 0.8$	0.320 ± 0.360	0.320 ± 0.359	0.330 ± 0.471
$\Sigma_{k,k} = 1.2$	0.901 ± 1.012	0.898 ± 1.011	0.775 ± 1.033
$\Sigma_{k,k} = 1.6$	2.448 ± 2.316	2.437 ± 2.306	1.563 ± 2.146

on [132], based on specific ICD-9 codes and procedures codes. We try to estimate the effect of vasopressor on death. The following vasopressors, norepinephrine, epinephrine, phenylephrine, vasopressin, dopamine, isuprel, are sought in the fluids administered to patients during their stay. Several prognostic factors are consider, systolic and diastolic blood pressures as time-dependent factors and gender, age, lactate, creatinine and SOFA score [194] at inclusion in intensive care as baseline factors. Blood pressure data are filled in every hour. Patients are followed from entry into intensive care until discharge or death.

Time-dependent covariates are smoothed using moving average of 5 values and standardized to get more stable trajectories. Observed blood pressures at inclusion for each patients are considered as baseline covariates. Counterfactual covariate trajectories under the scenario of no treatment are estimated according to a VAR model as presented in section Example with a vector autoregressive model VAR(1). Estimate starts at treatment time and stops at death or discharge. Baseline variables mentioned above are used as adjusted covariates.

Figures 3.4 and 3.3 shows observed and counterfactual covariates from treatment initiation. The graph is truncated at 200 hours to keep enough patients. The red lines represent counterfactual covariates under the assumption that the patients would not have been treated. These are unobserved trajectories. The grey lines represent observed covariates when patients are treated. Graphically, we see, for treated individuals, an increase in systolic and diastolic blood pressure which is the clinical effect of vasopressors.

As a reminder, the objective here is to evaluate the effect of vasopressors on death. ATT is estimated using Gran's estimator and ours. We include in these models the above mentionned covariates : smoothed standardised systolic and diastolic blood pressure grouped into hourly intervals, gender, age, lactate, creatinine, SOFA score, systolic and diastolic blood pressure on entry into intensive care. These estimations are compared to naive estimation and ATE estimation. Naive estimation is obtained using Aalen model with baseline covariates and observed time-dependent covariates. ATE is estimated using marginal structural

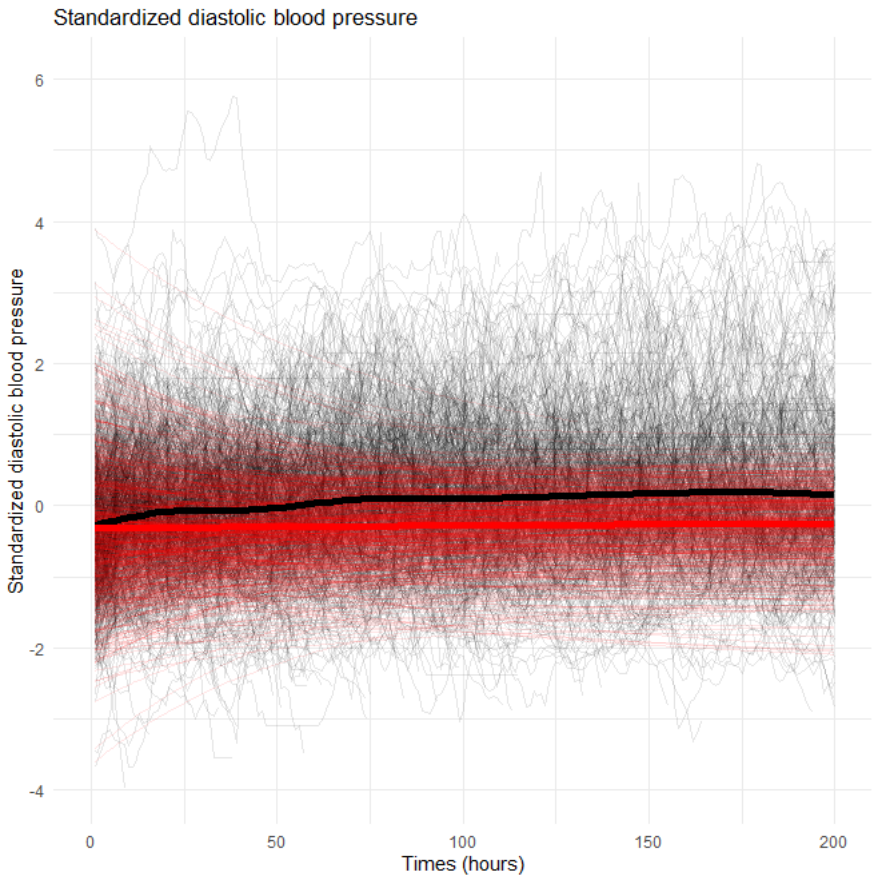

FIGURE 3.3 – Standardized diastolic blood pressure (black : observed trajectories, red : modelled trajectories)

additive model with the same covariates. Time-dependent covariates where replaced by stabilised inverse probability of treatment weights.

Cumulative treatment effect is showed in Figure 3.5 for vasopressor versus no treatment. Treatment has a positive contribution to the intensity (increasing curve) all the more important in the first days. In other words, treatment have not a protective effect neither on the eligible population (ATE) nor on the patients actually treated (ATT). The direction of the effect can mainly be explained because prescription of vasopressor is directly correlated to the severity of septic shock and thus to death. Moreover these methods do not allow to take into account the dose which may have an effect on death.

Finally, we observe that the cumulative effect estimate from the naive model and from the MSM seems to be linear, while this effect seems to stabilize for the Gran's estimators and ours. We observe that the correction has an increasing effect with time, which is explained by the increase in the number of patients treated over time.

3.7 - Discussion

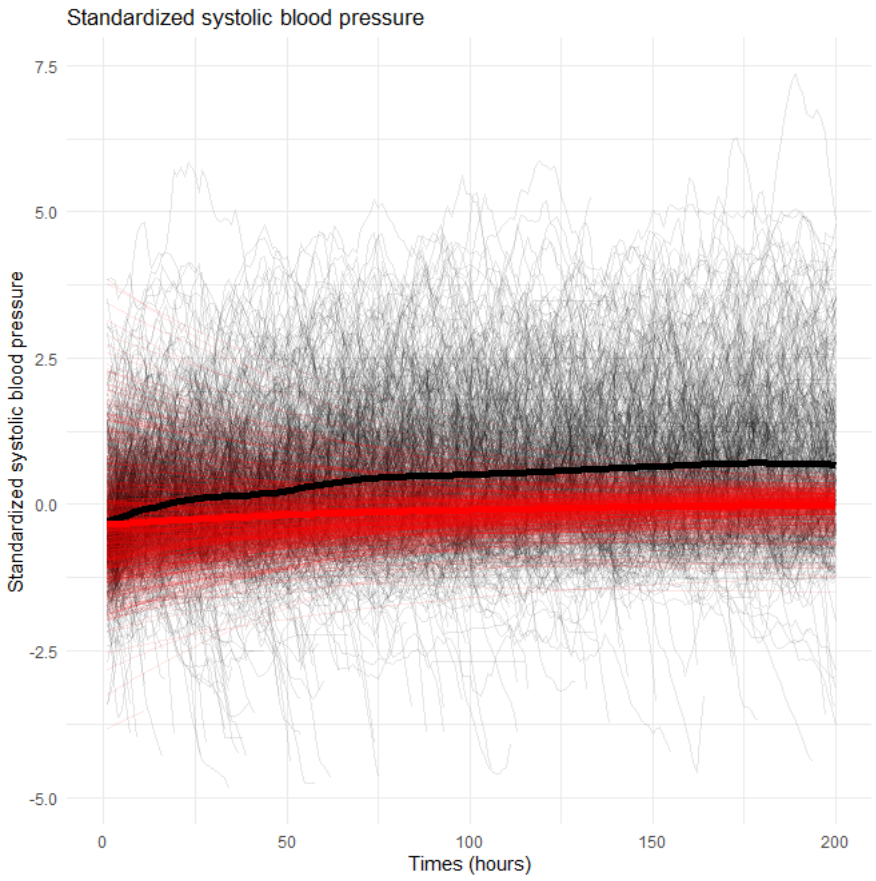

FIGURE 3.4 – Standardized systolic blood pressure (black : observed trajectories, red : modelled trajectories)

In this article, we proposed a corrected estimator of the ATT under time-varying confounders. We were inspired by the method developed by Gran et al. [90]. This method consists of two steps. The first one is the modelling of time-dependent covariates if the patients had not been treated and the second is the estimation of the treatment effect. The correction we propose is based on the error that is made when modelling the counterfactuals. Indeed, as in any modelling, there is a difference between the "real" data and those modelled. Our model allows the study of :

1. terminal outcomes such as death or repeated events/counting processes such as rehospitalizations or asthma attacks ;
2. multiple time varying covariates and baseline covariates.

All simulation scenarios showed that our corrected estimator provides better estimates of the ATT than those obtained with uncorrected estimators.

Estimates of ATT and ATE answer the two different questions. The ATT estimator addresses the causal question "Is the treatment efficient in patients in whom it is actually prescribed ?" rather than "Is the treatment efficient in the whole population" which is addressed by ATE. It may be interesting to estimate these two measures in the same study to get additional information about the effect of the treatment. Their comparison can be used to assess the effect of current treatment policy such as the choice of patients

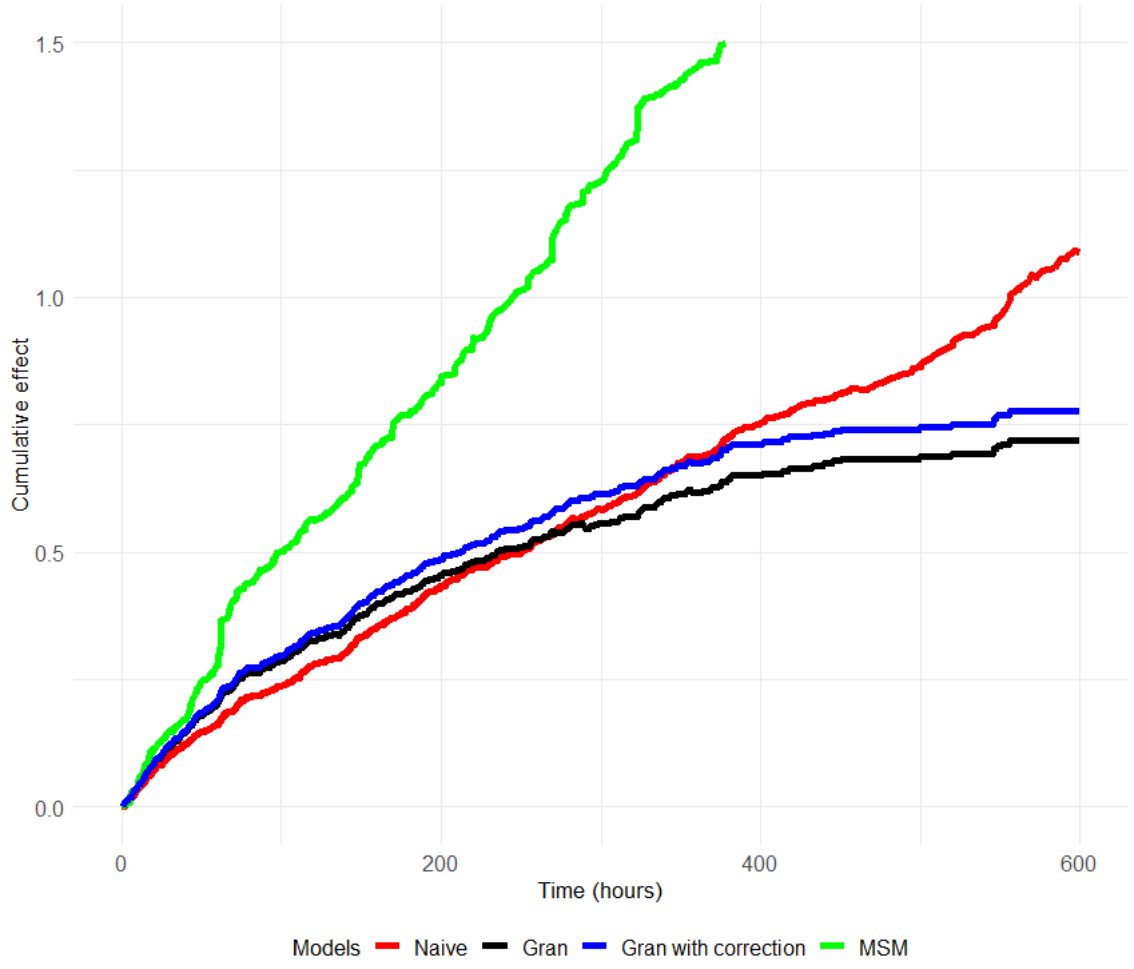

FIGURE 3.5 – Cumulative treatment effect estimated using Naive model, Marginal structural model Gran's estimator with and without correction.

receiving treatment and the treatment initiation time. The methods used to estimate these measures can also be used to select optimal treatment regimes. This is particularly the case for g-estimation method [19], Inverse Probability of Censoring Weighting model [167] or a non-parametric structural equation model and g-computation [58].

Our correction is introduced in the cases where the data are continuous and the counterfactuals are modelled by a VAR(1) model. This model is simple to implement and allows an explicit writing of the correction. Other methods could be used such as models more specific to diseases and time-dependent covariates such as differential equations. However, it should be noted that the more complex the model, the more difficult it will be to make explicit the modelling error and thus the correction of the ATT estimator.

Note that in this paper, only continuous covariates are studied. Depending on the databases analyzed, the time-dependent covariates may be binary or count data such as the number of nursing or general practitioner visits over each time interval studied. Specific methods to model the counterfactuals for this type of variable should be developed. It would also be interesting to be able to consider different types of data in the same model.

Chapitre 4

Facteurs de confusion discrets et dépendants du temps : estimation débiaisée de l'ATT

Le chapitre précédent a présenté un algorithme permettant d'estimer l'effet moyen d'un traitement chez les patients traités en présence de facteurs de confusion continus dépendant du temps. Cet algorithme est une extension de celui proposé par Gran et ses collaborateurs [90] dans le cas où le résultat permettant d'évaluer l'effet peut être modélisé par un processus de comptage. Dans le domaine de la santé, il peut s'agir de réhospitalisations, de crises d'asthme ou de différents évènements cliniques définissant la progression d'une maladie. Cet algorithme peut être aussi appliqué au cas particulier de l'étude d'un évènement terminal. Une estimation débiaisée a été proposée et validée par étude de simulation.

Les facteurs de confusion continus et dépendants du temps sont très courants dans les bases de données cliniques. Il peut s'agir, par exemple, de l'évolution de la taille d'une tumeur cancéreuse, de l'évolution d'une mesure biologique comme la glycémie dans l'étude du diabète ou de taux de CD4 chez les patients atteints du sida. Ces données font très souvent l'objet de recueil dans les essais cliniques ou les entrepôts de données hospitalières.

Toutefois ces données cliniques ne sont pas disponibles dans les bases médico-administratives françaises. Comme présenté en introduction, seules les données relatives aux remboursements sont renseignées. Les données continues présentes dans ce type de base sont les montants payés, remboursés et les posologies des traitements délivrées. Lors de la construction des parcours de soin d'autres variables dépendantes du temps peuvent être générées. C'est notamment le cas des données de comptage. Prenons un exemple. Supposons que l'on s'intéresse aux consultations réalisées avec une infirmière. Le SNDS dispose des dates de chacune de ces consultations. On peut alors s'intéresser à l'évolution du nombre mensuel de recours aux soins infirmiers. Ce type de variable peut être généré pour un grand nombre d'indicateurs comme le nombre de consultation avec chaque type de médecins spécialistes, le nombre de délivrances de certains traitements spécifiques, le nombre d'hospitalisations ambulatoires... La prise en compte de variables de comptage dans les analyses est primordiale dans les données du SNDS.

La suite de ce chapitre va présenter de nouveau les notations et le modèle utilisé pour l'estimation

de l'ATT. Nous introduirons ensuite le modèle VINGARCH, une extension multidimensionnelle du modèle INGARCH (voir [77, 207, 62]). L'écriture de la correction de l'estimation sera donnée dans le cas général d'un modèle VINGARCH et son estimation sera présentée dans le cas particulier d'un modèle VINGARCH(1,0). Une analyse de simulation sera ensuite présentée et une conclusion clôturera cette partie.

4.1 - Model and notations

4.1.1 . Modele

Consider a cohort analysis. For a patient, at cohort entry, d_Z-dimensional baseline covariates $Z = \left(Z^1, \ldots, Z^{d_Z}\right)$ are collected as gender, age, comorbidities... The patient is then followed-up over time and time-dependent d-dimensional covariates $t \mapsto X(t) = (X^1(t), \ldots, X^d(t))$ are observed. His/her treatment is initiated at time S. We assume that once started the treatment is continued, so that the treatment process $t \mapsto D(t)$ is null before S and equals 1 after S. The pathologies of interest are those that have an outcome measured via a counting process N ($N(t)$ being the number of events recorded in $[0, t]$). We also assume that independent censoring can occur and that N verifies the Aalen additive intensity model [1] as studed in the following assumption.

Assumption 2 *With respect $\left(\mathcal{F}_t\right)_{t\geq 0}$, the historical filtration spanned by Z, X, D, and N, we assume the process X to be predictable and that N has the following intensity :*

$$\mu^*(t) = \alpha_0^*(t) + \alpha_Z^*(t)Z + \alpha_X^*(t)X(t) + \alpha_D^*(t)D(t).$$

At time S, we can intervene on the treatment decision. Two notations can be introduced $X^{0|S}$ the covariate process if the treatment had not been initiated at time S and $X^{1|S}$ the covariate process where the treatment is initiated at S. In a similar way, the two potential counting processes recording the event history after S can be denoted $N^{0|S=s}$ and $N^{1|S=s}$. Under Assumption (2), their respective intensities can be written for $t \geq S = s$, as :

$$\mu^{1|S=s}(t) = \alpha_0^*(t) + \alpha_D^*(t) + \alpha_Z^* Z + \alpha_X^*(t)X^{1|S=s}(t) \text{ and}$$

$$\mu^{0|S=s}(t) = \alpha_0^*(t) + \alpha_Z^* Z + \alpha_X^*(t)X^{0|S=s}(t).$$

Three assumptions are made to verify this model : positivity assumption, consistency assumption and exchangeability assumption. Details are presented section 3.2.

4.1.2 . Additive intensity model for the treatment effect estima-

tion

Suppose we have data from n independent individuals, $i = 1, \ldots, n$. These individuals are followed between $t = 0$ and $t = \tau_i$ which can be the end time of the study or the date of censorship. The following data are collected for each individual Z_i, X_i, S_i, D_i and N_i. The intensity of N_i can be written as :

$$\begin{aligned}\mu_i^*(t) &= \alpha_0^*(t) + \alpha_Z^* Z_i + \alpha_X^*(t) X_i^{0|S=S_i}(t) + d_i^*(t, S_i) D_i(t) \\ &= W_i(t)^\top A^*(t),\end{aligned}$$

with $A^*(t) = (\alpha_0^*(t), \alpha_Z^*, \alpha_X^*(t), d_i^*(t, S_i))^\top$ and $W_i(t) = (\mathbb{1}, Z_i, X_i^{0|S=S_i}(t), D_i(t))^\top$. For $t \geq S$, the time varying ATT is defined as $\mathsf{ATT}(t) = \mathbb{E}[d^*(t, S_i)|t \geq S_i]$. It can be estimated by estimating A^*. A good candidate for the estimation of A^* can be $\widehat{A} = \arg\min_A r_n(A)$ with $r_n(A)$ the squared risk of A :

$$\begin{aligned}r_n(A) = &\frac{1}{n} \sum_{i=1}^n \int_0^{\tau_i} A(t)^\top W_i(t) W_i(t)^\top A(t) dt \\ &- \frac{2}{n} \sum_{i=1}^n \int_0^{\tau_i} A(t)^\top W_i(t) dN_i(t).\end{aligned}$$

For $t > S_i$, $X_i^{0|S=S_i}(t)$ is not observed and by extension neither is W_i. To estimate A^*, these counterfactual have to be estimated. We presente an example of such estimation in Section 4.2 with an vector integer-valued generalized autoregressive conditional heteroscedastic vector model (VINGARCH). We denote $\widetilde{X}_i^{0|S=S_i}(t) = X_i^{0|S=S_i}(t) + \epsilon_i(t)$ this estimation with $\epsilon_i(t)$ the center counterfactuals modelling error with finite variance. According to the definition of W_i, we denote $\widetilde{W}_i(t) = (\mathbb{1}, Z_i, \widetilde{X}_i^{0|S=S_i}(t), D_i(t))^\top$. We have showed in 3.3 that the use of modeled counterfactuals in the estimation of A^* introduces the following bias :

$$\mathsf{bias}(A) = \frac{1}{n} \sum_{i=1}^n \int_0^{\tau_i} \alpha_X(t)^\top \mathbb{E}[\epsilon_i(t) \epsilon_i(t)^\top] \alpha_X(t) dt. \tag{4.1}$$

When this type of analysis is performed on real-life data, $\mathbb{E}[\epsilon_i(t)\epsilon_i(t)^\top]$ is not observed. It must be estimated. An explicit expression of this estimation can be given according to the model use for time-varying modeling counterfactuals. The section 4.3.1 presents this estimation in the case of a modeling using a

multivariate INGARCH(p,q) model. Finally, the debiased estimator of A^* is given by :

$$\widehat{A}_{\text{debiased}} = \arg\min_A \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)\left[\widetilde{W}_i(t)\widetilde{W}_i(t)^\top - \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \widehat{\mathbb{E}[\epsilon_i(t)\epsilon_i(t)^\top]} & \vdots \\ 0 & \cdots & 0 \end{pmatrix}\right] A(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)\widetilde{W}_i(t)dN_i(t). \tag{4.2}$$

4.2 - VINGARCH model

In the previous section, we presented a model to estimate the effect of a treatment in treated patients. The following section introduces the linear INGARCH model as a method for modelling discret covariates. An extension of the INGARCH model to a multivariate model called VINGARCH is introduced.

4.2.1 . Univariate linear INGARCH model

For each individual i is observed a discrete covariate process

$$t \mapsto X_i(t) = \big(X_i^1(t), \ldots, X_i^d(t)\big),$$

where each realization of $X_i^j(t)$ (for all $j = 1, \ldots, d$ and $t \geq 0$) is integer-valued. We propose to model these processes via a linear INGARCH model, studied in [77, 207, 62]. We present in this paragraph the model for an univariate integer-valued time series and postpone the presentation of multivariate time series to the next paragraph.

Consider an integer-valued time series $X_i = \{X_i(t), t \in \mathbb{N}\}$, with its natural filtration $\big(\mathcal{F}_t\big)_{t\in\mathbb{N}}$, with conditional mean $\mathbb{E}\big[X_i(t)|\mathcal{F}_{t-1}\big] = \lambda_i(t)$ for all $t \in \mathbb{N}$. The linear INGARCH(p, q) ($p \in \mathbb{N}^*$ and $q \in \mathbb{N}$) imposes the following form for the conditional mean

$$\lambda_i(t) = \beta_0^{*i} + \sum_{k=1}^{p} b_k^* X(t-k) + \sum_{l=1}^{q} c_l^* \lambda(t-l), \tag{4.3}$$

where β_0^{*i} is an intercept, and $b_1^*, \ldots, b_p^*, c_1^*, \ldots, c_q^*$ are positive dependence parameters, see [77]. The basic INGARCH model assumes that $X_i(t)|\mathcal{F}_{t-1}$ is Poisson distributed according to $\mathcal{P}\big(\lambda_i(t)\big)$. Different INGARCH models can be obtained depending on the choice of the conditional mean like negative binomial distribution

NegBin$\big(\lambda_i(t), \phi\big)$ (where, in this parametrization, ϕ is an overdispersion parameter), see [211, 212] for more details. The specific model where $q = 0$ is noted INARCH(p) model.

As we are interested on the effects of treatment switches on the distribution of our integer-valued process, we will follow the lines of [122, 76] to model these. In our case, the treatment is switched from 0 to 1 as time $S = s_i$, we then propose a new form of equation (4.3) for the conditional mean to adapt for this switch :

$$\lambda_i(t) = \beta_0^{*i} + \sum_{j=1}^{p} b_j^* X(t-j) + \sum_{j=1}^{q} c_j^* \lambda_i(t-j) + \nu D_i(t).$$

Notice when the treatment is set to 1, ν corresponds to the shift of the conditional mean.

4.2.2 . A model for multivariate discrete time series : VINGARCH model

An extension of the INGARCH model can be proposed to deal with multivariate discrete processes and take into account possible dependence between the coordinates. Little work has been done to extend the INGARCH model to multivariate problems [205]. A bivariate Poisson INGARCH(1,1) was presented in chapter 4 of [123]. We introduce a generalization of this model, referring as linear Poisson VINGARCH(p,q) model. We assume that the coefficients for the past observations and means are shared between individuals. The conditional mean $\lambda_i \in \mathbb{R}^{+d}$ is written :

$$\lambda_i(t) = \beta_0^{*i} + \sum_{j=1}^{p} B_j^* X_i(t-j) + \sum_{j=1}^{q} C_j^* \lambda_i(t-j) + \nu D_i(t), \tag{4.4}$$

where $\beta_0^{*i} \in \mathbb{R}_+^d$ an intercept and $B_j^* = (b_{kl}^{*j}) \in \mathbb{R}^{d\times d}$ and $C_j^* = (c_{kl}^{*j}) \in \mathbb{R}^{d\times d}$ matrices of size $d \times d$. The distribution of the k component of $X_i(t)|\mathcal{F}_{t-1}$, noted $X_i^k(t)|\mathcal{F}_{t-1}$ is generally assumed to be Poisson $\mathcal{P}\big(\lambda_i^k(t)\big)$, with $\lambda_i^k(t) = \beta_{0,k}^{*i} + \sum_{j=1}^{p} b_{k.}^{*j} X_i(t-j) + \sum_{j=1}^{q} c_{k.}^{*j} \lambda_i(t-j) + \nu^k D_i(t)$.

4.3 - Biais Calculation

Section 4.1.2 highlighted a bias in the estimation of the ATT using the counterfactuals. An unbiased estimator was proposed, without assumptions about the modelling method of the counterfactuals. In the following section, we propose an explicit writing of the corrected estimator, when counterfactuals are modelled using a VINGARCH(p,q) model. We then propose a method for estimating the correction in the case of a VINGARCH(1,0) model.

4.3.1 . Explicit writing of the bias

In the following, we assumed that the observation period from 0 to $\max_i \tau_i$ is partitioned into sub time interval defined by times $t_0 = 0, \ldots, t_k, \ldots t_K = \max_i \tau_i$. Time varying covariates are assumed constant over these sub intervals. For all subject i, $k \in [\![0, K-1]\!]$ and $t \in [t_k, t_{k+1})$, $X_i(t) = X_i(t_k)$. We assumed that discrete time-varying covariates are modelled following a VINGARCH(p,q) as introduced in the previous section.

4.3.1.1 . VINGARCH(2, 2) example

To simplify the calculations, let's start with the particular case of a bivariate time-varying process modelled by a VINGARCH(2,2) model. The modelling of the counterfactuals was carried out under the assumption that the patients would never have started their treatment, the term $\nu D_i(t_k)$ is null. The equation (4.4) can be written as follow :

$$\begin{pmatrix} \lambda_i^1(t_k) \\ \lambda_i^2(t_k) \end{pmatrix} = \begin{pmatrix} \beta_{0,1}^{*i}(t_k) \\ \beta_{0,2}^{*i}(t_k) \end{pmatrix} + \begin{pmatrix} b_{1,1}^{*1} & b_{1,2}^{*1} \\ b_{2,1}^{*1} & b_{2,2}^{*1} \end{pmatrix} \begin{pmatrix} X_i^1(t_{k-1}) \\ X_i^2(t_{k-1}) \end{pmatrix} + \begin{pmatrix} b_{1,1}^{*2} & b_{1,2}^{*2} \\ b_{2,1}^{*2} & b_{2,2}^{*2} \end{pmatrix} \begin{pmatrix} X_i^1(t_{k-2}) \\ X_i^2(t_{k-2}) \end{pmatrix} + \begin{pmatrix} c_{1,1}^{*1} & c_{1,2}^{*1} \\ c_{2,1}^{*1} & c_{2,2}^{*1} \end{pmatrix} \begin{pmatrix} \lambda_i^1(t_{k-1}) \\ \lambda_i^2(t_{k-1}) \end{pmatrix} + \begin{pmatrix} c_{1,1}^{*2} & c_{1,2}^{*2} \\ c_{2,1}^{*2} & c_{2,2}^{*2} \end{pmatrix} \begin{pmatrix} \lambda_i^1(t_{k-2}) \\ \lambda_i^2(t_{k-2}) \end{pmatrix}$$

Let's introduce an zero mean error term, ω_i, to link $X_i(t_k)$ and $\lambda_i(t_k)$, such that $\omega_i(t_k) = X_i(t_k) - \lambda_i(t_k)$. A matrix equation can be formulated to simplify the explicit writing of the bias. It combine the previous equation and the relationship between $X_i(t_k)$ and $\lambda_i(t_k)$. We obtain the following equation :

$$\begin{pmatrix} X_i^1(t_k) \\ X_i^2(t_k) \\ X_i^1(t_{k-1}) \\ X_i^2(t_{k-1}) \\ \lambda_i^1(t_k) \\ \lambda_i^2(t_k) \\ \lambda_i^1(t_{k-1}) \\ \lambda_i^2(t_{k-1}) \end{pmatrix} = \begin{pmatrix} \beta_{0,1}^{*i}(t_k) \\ \beta_{0,2}^{*i}(t_k) \\ 0 \\ 0 \\ \beta_{0,1}^{*i}(t_k) \\ \beta_{0,2}^{*i}(t_k) \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} b_{1,1}^{*1} & b_{1,2}^{*1} & b_{1,1}^{*2} & b_{1,2}^{*2} & c_{1,1}^{*1} & c_{1,2}^{*1} & c_{1,1}^{*2} & c_{1,2}^{*2} \\ b_{2,1}^{*1} & b_{2,2}^{*1} & b_{2,1}^{*2} & b_{2,2}^{*2} & c_{2,1}^{*2} & c_{2,2}^{*2} & c_{2,1}^{*2} & c_{2,2}^{*2} \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ b_{1,1}^{*1} & b_{1,2}^{*1} & b_{1,1}^{*2} & b_{1,2}^{*2} & c_{1,1}^{*1} & c_{1,2}^{*1} & c_{1,1}^{*2} & c_{1,2}^{*2} \\ b_{2,1}^{*1} & b_{2,2}^{*1} & b_{2,1}^{*2} & b_{2,2}^{*2} & c_{2,1}^{*2} & c_{2,2}^{*2} & c_{2,1}^{*2} & c_{2,2}^{*2} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} X_i^1(t_{k-1}) \\ X_i^2(t_{k-1}) \\ X_i^1(t_{k-2}) \\ X_i^2(t_{k-2}) \\ \lambda_i^1(t_{k-1}) \\ \lambda_i^2(t_{k-1}) \\ \lambda_i^1(t_{k-2}) \\ \lambda_i^2(t_{k-2}) \end{pmatrix} + \begin{pmatrix} \omega_1^i(t_k) \\ \omega_2^i(t_k) \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

Block matrix notation can be used to rewrite the above equation as follows :

$$\begin{pmatrix} X_i(t_k) \\ X_i(t_{k-1}) \\ \lambda_i(t_k) \\ \lambda_i(t_{k-1}) \end{pmatrix} = \begin{pmatrix} \beta_0^{*i} \\ 0_2 \\ \beta_0^{*i} \\ 0_2 \end{pmatrix} + \begin{pmatrix} B_1^* & B_2^* & C_1^* & C_2^* \\ Id_2 & 0_{2,2} & 0_{2,2} & 0_{2,2} \\ B_1^* & B_2^* & C_1^* & C_2^* \\ 0_{2,2} & 0_{2,2} & Id_2 & 0_{2,2} \end{pmatrix} \begin{pmatrix} X_i(t_{k-1}) \\ X_i(t_{k-2}) \\ \lambda_i(t_{k-1}) \\ \lambda_i(t_{k-2}) \end{pmatrix} + \begin{pmatrix} \omega^{*i}(t_k) \\ 0_2 \\ 0_2 \\ 0_2 \end{pmatrix}, \tag{4.5}$$

with 0_2 a zero column vector of 2 entries and $0_{2,2}$ a 2×2 zero matrix.

Equation 4.5 can be used iteratively to express $X_i(t_k)$ and $\lambda_i(t_k)$ in terms of the first observations of X_i and λ_i. Let M be the block matrix composed of the matrices B^* and C^*, we obtain the following expression :

$$\begin{aligned} \begin{pmatrix} X_i(t_k) \\ X_i(t_{k-1}) \\ \lambda_i(t_k) \\ \lambda_i(t_{k-1}) \end{pmatrix} &= \begin{pmatrix} \beta_0^{*i} \\ 0_2 \\ \beta_0^{*i} \\ 0_2 \end{pmatrix} + M \left(\begin{pmatrix} \beta_0^{*i} \\ 0_2 \\ \beta_0^{*i} \\ 0_2 \end{pmatrix} + M \begin{pmatrix} X_i(t_{k-2}) \\ X_i(t_{k-3}) \\ \lambda_i(t_{k-2}) \\ \lambda_i(t_{k-3}) \end{pmatrix} + \begin{pmatrix} \omega^{*i}(t_k) \\ 0_2 \\ 0_2 \\ 0_2 \end{pmatrix} \right) + \begin{pmatrix} \omega^{*i}(t_k) \\ 0_2 \\ 0_2 \\ 0_2 \end{pmatrix} \\ &= \sum_{l=0}^{1} M^l \begin{pmatrix} \beta_0^{*i} \\ 0_2 \\ \beta_0^{*i} \\ 0_2 \end{pmatrix} + \sum_{l=0}^{1} M^l \begin{pmatrix} \omega^{*i}(t_k) \\ 0_2 \\ 0_2 \\ 0_2 \end{pmatrix} + M^2 \begin{pmatrix} X_i(t_{k-2}) \\ X_i(t_{k-3}) \\ \lambda_i(t_{k-2}) \\ \lambda_i(t_{k-3}) \end{pmatrix} \\ &= \sum_{l=0}^{k-2} M^l \begin{pmatrix} \beta_0^{*i} \\ 0_2 \\ \beta_0^{*i} \\ 0_2 \end{pmatrix} + \sum_{l=0}^{k-2} M^l \begin{pmatrix} \omega^{*i}(t_k) \\ 0_2 \\ 0_2 \\ 0_2 \end{pmatrix} + M^{k-1} \begin{pmatrix} X_i(t_1) \\ X_i(t_0) \\ \lambda_i(t_1) \\ \lambda_i(t_0) \end{pmatrix}, \end{aligned} \tag{4.6}$$

with M^0 the identity matrix of size 8.

As we saw previously, $X_i^{0|S=s_i}(t_k)$ is not observed for $t_k > s_i$ and need to be modelled. Time-varying modelled counterfactuals are then noted $\widetilde{X}_i^{0|S=s_i}$. The equation (4.6) can be written without the error term. The modelling error can be expressed as the difference between modelled counterfactual $\widetilde{X}_i^{0|S}$ and the true values X_i. This error is the following :

$$\begin{pmatrix} X_i(t_k) \\ X_i(t_{k-1}) \\ \lambda_i(t_k) \\ \lambda_i(t_{k-1}) \end{pmatrix} - \begin{pmatrix} \widetilde{X}_i(t_k) \\ \widetilde{X}_i(t_{k-1}) \\ \widetilde{\lambda}_i(t_k) \\ \widetilde{\lambda}_i(t_{k-1}) \end{pmatrix} = \sum_{l=0}^{k-2} M^l \begin{pmatrix} \omega^{*i}(t_k) \\ 0_2 \\ 0_2 \\ 0_2 \end{pmatrix}, \tag{4.7}$$

As presented in section 4.1.2, the bias depend on the variance of the modelling error. It can be written as :

$$\mathbb{V}(X_i(t_k) - \widetilde{X}_i(t_k)) = \mathbb{V}\left(U \begin{pmatrix} X_i^1(t_k) \\ X_i^2(t_k) \\ X_i^1(t_{k-1}) \\ X_i^2(t_{k-1}) \\ \lambda_i^1(t_k) \\ \lambda_i^2(t_k) \\ \lambda_i^1(t_{k-1}) \\ \lambda_i^2(t_{k-1}) \end{pmatrix} - U \begin{pmatrix} \widetilde{X}_i^1(t_k) \\ \widetilde{X}_i^2(t_k) \\ \widetilde{X}_i^1(t_{k-1}) \\ \widetilde{X}_i^2(t_{k-1}) \\ \widetilde{\lambda}_i^1(t_k) \\ \widetilde{\lambda}_i^2(t_k) \\ \widetilde{\lambda}_i^1(t_{k-1}) \\ \widetilde{\lambda}_i^2(t_{k-1}) \end{pmatrix} \right),$$

with $U = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = (Id_{2,2}|0_{2,6})$. The properties of ω_i allow to write :

$$\mathbb{V}(X_i(t_k) - \widetilde{X}_i(t_k)) = \sum_{l=0}^{k-2} U M^l \mathbb{V}\left(\begin{pmatrix} \omega^{*i}(t_k) \\ 0_2 \\ 0_2 \\ 0_2 \end{pmatrix} \right) {M^l}^\top U^\top. \tag{4.8}$$

The result of equation 4.8 can be introduced in the equation 4.2 to obtain the explicit writing of the debiased ATT estimator in the presence of bivariate covariates modelled by a VINGARCH(2,2).

4.3.1.2 . Generalization

This section is a generalization of the previous one. For each individual i is observed a d-dimensional discrete covariate process : $t \mapsto X_i(t) = \left(X_i^1(t), \ldots, X_i^d(t)\right)$. The counterfactuals are assumed to be modelled according to a VINGARCH(p,q) model.

The block matrix notation introduce in equation 4.5 is generalized in the following equation :

$$\begin{pmatrix} X_i(t_k) \\ \vdots \\ X_i(t_{k-p+1}) \\ \lambda_i(t_k) \\ \vdots \\ \lambda_i(t_{k-q+1}) \end{pmatrix} = \begin{pmatrix} \beta_0^{*i} \\ 0_d \\ 0_d \\ \vdots \\ \beta_0^{*i} \\ \vdots \\ 0_d \\ 0_d \end{pmatrix} + M \begin{pmatrix} X_i(t_{k-1}) \\ \vdots \\ X_i(t_{k-p}) \\ \lambda_i(t_{k-1}) \\ \vdots \\ \lambda_i(t_{k-q}) \end{pmatrix} + \begin{pmatrix} \omega_i(t_k) \\ 0_d \\ \vdots \\ 0_d \\ 0_d \\ \vdots \\ 0_d \\ 0_d \end{pmatrix}, \tag{4.9}$$

with two vectors of size $d(p+q)$, $(X_i(t_k), \cdots, X_i(t_{k-p+1}), \lambda_i(t_k), \cdots, \lambda_i(t_{k-q+1}))^\top$ and $(X_i(t_{k-1}), \cdots, X_i(t_{k-p}), \lambda_i(t_{k-1}), \cdots, \lambda_i(t_{k-q}))^\top$,$0_d$ a zero column vector of d entries, M the following $d(p+q)\times d(p+q)$ matrix

$$\begin{pmatrix} B_1^* & B_2^* & \ldots & B_p^* & C_1^* & C_2^* & \ldots & C_q^* \\ Id_d & 0_{dd} & \ldots & 0_{dd} & 0_{dd} & 0_{dd} & \ldots & 0_{dd} \\ 0_{dd} & Id_d & \ldots & 0_{dd} & 0_{dd} & 0_{dd} & \ldots & 0_{dd} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ B_1^* & B_2^* & \ldots & B_p^* & C_1^* & C_2^* & \ldots & C_q^* \\ 0_{dd} & 0_{dd} & \ldots & 0_{dd} & Id_d & 0_{dd} & \ldots & 0_{dd} \\ 0_{dd} & 0_{dd} & \ldots & 0_{dd} & 0_{dd} & Id_d & \ldots & 0_{dd} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{dd} & 0_{dd} & \ldots & 0_{dd} & 0_{dd} & \ldots & Id_d & 0_{dd} \end{pmatrix},$$

with $B_j^* = (b_{kl}^{*j}) \in \mathbb{R}^{+d\times d}$, $j \in [\![1,p]\!]$, $C_j^* = (c_{kl}^{j*}) \in \mathbb{R}^{d\times d}$, $j \in [\![1,q]\!]$, Id_d the identity matrix of size dd, 0_{dd} the zero matrix of size $d \times d$. The error term w_i is a zero mean error between X_i and λ_i which verifies the following properties : $\mathbb{V}(\omega_i(t_k) = \text{diag}(\mathbb{E}[X_i(t_k)])$ and $Cov(\omega_i(t_k), \omega_i(t_l)) = 0_d$ for all $k \neq l$ and l, $k \geq 0$. These two properties are demonstrated in appendix, section 9.1.1.

By repeating the equation 4.9 iteratively, we obtain a direct link between $X_i(t_k)$ and $\lambda_i(t_k)$ and the p

and q first observations of X_i and λ_i. We obtain the following equation :

$$\begin{pmatrix} X_i(t_k) \\ \vdots \\ X_i(t_{k-p+1}) \\ \lambda_i(t_k) \\ \vdots \\ \lambda_i(t_{k-q+1}) \end{pmatrix} = \sum_{l=0}^{k-p-1} M^l \begin{pmatrix} \beta_0^{*i} \\ 0_d \\ 0_d \\ \vdots \\ \beta_0^{*i} \\ \vdots \\ 0_d \\ 0_d \end{pmatrix} + \sum_{l=0}^{k-p-1} M^l \begin{pmatrix} \omega_i(t_k) \\ 0_d \\ \vdots \\ 0_d \\ 0_d \\ \vdots \\ 0_d \\ 0_d \end{pmatrix} + M^{k-p} \begin{pmatrix} X_i(t_{p-1}) \\ \vdots \\ X_i(t_0) \\ \lambda_i(t_{q-1}) \\ \vdots \\ \lambda_i(t_0) \end{pmatrix}, \tag{4.10}$$

with $M^0 = Id_{d(p+q)}$ the identity matrix of size $d(p+q) \times d(p+q)$.

The modelling error can be expressed as the difference between modelled counterfactual $\widetilde{X}_i^{0|S}$ and the true values X_i. As we saw in the previous section, the equation 4.10 can be written with $\widetilde{X}_i^{0|S}$ and $\widetilde{\lambda}_i^{0|S}$ without the error term. This modelling error is then :

$$\begin{pmatrix} X_i(t_k) \\ \vdots \\ X_i(t_{k-p+1}) \\ \lambda_i(t_k) \\ \vdots \\ \lambda_i(t_{k-q+1}) \end{pmatrix} - \begin{pmatrix} \widetilde{X}_i^{0|S}(t_k) \\ \vdots \\ \widetilde{X}_i^{0|S}(t_{k-p+1}) \\ \widetilde{\lambda}_i(t_k) \\ \vdots \\ \widetilde{\lambda}_i(t_{k-q+1}) \end{pmatrix} = \sum_{l=0}^{k-p-1} M^l \begin{pmatrix} \omega_i(t_k) \\ 0_d \\ \vdots \\ 0_d \\ 0_d \\ \vdots \\ 0_d \\ 0_d \end{pmatrix}, \tag{4.11}$$

The corrected estimator of the ATT is depend on the modelling error and more specifically the variance (see section 4.1.2). From the equation 4.11, the variance is given by :

$$\mathbb{V}(X_i(t_k) - \widetilde{X}_i(t_k)) = \mathbb{V}\left(\sum_{l=0}^{k-p-1} UM^l \begin{pmatrix} \omega_i(t_k) \\ 0_d \\ \vdots \\ 0_d \\ 0_d \\ \vdots \\ 0_d \\ 0_d \end{pmatrix} \right),$$

with $U = (Id_d | 0_{d(p+2q)})$ a block-matrix. The properties of ω_i allow to write :

$$\mathbb{V}(X_i(t_k) - \widetilde{X}_i^{0|S}(t_k)) = \sum_{j=0}^{k-p-1} UM^j \begin{pmatrix} \mathsf{diag}(\mathbb{E}(X_i)) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} {M^j}^\top U^\top. \tag{4.12}$$

Finally, equations 4.1 and 4.12 give the following bias expression in the case of a counterfactual modeling by VINGARCH(p,0) :

$$\mathsf{biais}(A) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau_i} \alpha_X(t)^T \sum_{j=0}^{t-1} UM^j \begin{pmatrix} \mathsf{diag}(\mathbb{E}(X_i)) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} {M^j}^T U^T \alpha_X(t) dt. \tag{4.13}$$

The expression bias can be made more explicit. Let's define the two following applications :

— Vec : $\mathcal{M}_{l,k}(\mathbb{R}) \to \mathbb{R}^{lk}$ the application which map a matrix onto a vector composed of the succession of its columns ;

— $\otimes : \mathcal{M}_{l,k}(\mathbb{R}) \times \mathcal{M}_{m,n}(\mathbb{R}) \to \mathcal{M}_{lm,kn}(\mathbb{R})$ the Kronecker product.

We can demonstrate using properties of Vec (linear property and property on real values) that the equa-

tion 4.13 can be written as follows :

$$\text{biais}(A) = \int_{t=0}^{\tau_i} \alpha_X(t)^\top \otimes \alpha_X(t)^\top U \otimes U \,(Id_{dX} - \big[M \otimes M\big]^\top)(Id_{dX} - M \otimes M)^{-1} dt$$

$$\text{Vec}\Big(\text{diag}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) \\ \vdots \\ 0 \end{pmatrix}\Big),$$

with :

$$-\frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^n (Id_d - \sum_{j=1}^p B_j - \sum_{j=1}^q C_j)^{-1}\beta_0^i = (Id_d - \sum_{j=1}^p B_j - \sum_{j=1}^q C_j)^{-1}\frac{1}{n}\sum_{i=1}^n \beta_0^i.$$

Parameters of VINGARCH(p,q) model are not known. They need to be estimated.

4.3.2 . Estimation

This section proposed a method to estimate parameters of the model VINGARCH(1,0). The last time without treatment for the patient i is noted $t^i_{\max}$. They correspond to the discrete observe time before S_i. To simplify calculations, we assume that β_0^{*i} is a function of baseline covariates. It will be expressed as follows $\beta_0^{*i} = \beta_0^* + \sum_{j=1}^{d_Z} A^* Z_i$ with $\beta_0^* = (\beta_0^{*1}, \cdots, \beta_0^{*d})^\top \in \mathbb{R}_+^d$ and A^* a $d \times d_Z$ matrix composed by the positive elements a^*_{kl}.
Estimating problem of the parameters β_0^*, A^* and B_1^* can be written, for each time-varying covariates j, by the following equation :

$$\begin{pmatrix} X_1^j(1) \\ \vdots \\ X_1^j(t^1_{\max}) \\ \vdots \\ X_n^j(1) \\ \vdots \\ X_n^j(t^n_{\max}) \end{pmatrix} = R \begin{pmatrix} \beta_0^{*j} \\ a^*_{j1} \\ \vdots \\ a^*_{jd_Z} \\ b^{*1}_{j1} \\ \vdots \\ b^{*1}_{jd} \end{pmatrix}, \tag{4.14}$$

with b_{kl}^{*1} element of the B_1^* matrix and R the following matrix of size $\sum_{j=1}^{n} t_{\max}^j \times (1 + d + d_Z)$:

$$R = \begin{pmatrix} 1 & Z_1^1 & \cdots & Z_1^{d_Z} & X_1^1(0) & \cdots & X_1^d(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_1^1 & \cdots & Z_1^{d_Z} & X_1^1(t_{\max}^1-1) & \cdots & X_1^d(t_{\max}^1-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_n^1 & \cdots & Z_n^{d_Z} & X_n^1(0) & \cdots & X_n^d(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_n^1 & \cdots & Z_n^{d_Z} & X_n^1(t_{\max}^n-1) & \cdots & X_n^d(t_{\max}^n-1) \end{pmatrix}, \tag{4.15}$$

These equations can be solve using successive Poison regression with link identity. Another method of solving the estimation problem using non-negative least square model is proposed in the appendix, section 9.1.2.

4.4 - Analyse de simulation

Dans la suite de cette section, les estimations de l'ATT seront calculées avec et sans correction sur des données simulées suivant un modèle VINGARCH(1,0), présenté plus haut. Les données de simulation suivent le même schéma que dans le cas du chapitre précédent (voir 3.5) avec des covariables dépendantes du temps qui dépendent du statut du traitement et réciproquement. La probabilité d'occurence d'un évènement est affectée à la fois par les covariables et par le traitement. La figure 2.7 présente ces relations.

Les données ont été simulées selon le schéma des données disponibles dans le SNDS avec des covariables discrètes qui dépendent du temps, des covariables à l'inclusion et un résultat d'intérêt qui est un évènement répétitif. Ces données sont simulées sur des intervalles de temps prédéfinis $[t_k, t_{k+1})$, $t_k = 0, 1, ...51$ qui peuvent être assimilés à des intervalles de temps mensuel pour un suivi de 4 ans et 3 mois. Sur chaque intervalle de temps le nombre d'évènements $N(t_k)$, le statut du traitement $D(t_k)$, les covariables discrètes de dimension d $X(t_k) = (X^1(t_k), \ldots, X^d(t_k))$ et les d_Z covariables à l'inclusion, Z, sont définis.

Pour les simulations 3 variables dépendantes du temps ont été simulées. Ces variables peuvent être, par exemple, considérées comme le nombre de consultations avec 3 types de spécialistes différents. Elles sont simulées au temps t_{k+1} selon une distribution de Poisson avec pour intensité $\kappa_{D_0} + K_0 X_i(t_k)$ sur les intervalles de temps où les sujets ne sont pas traités et $\kappa_{D_1} + K_1 X_i(t_k)$ dans le scénario où le patient est traité. κ_{D_0} et κ_{D_1} sont deux vecteurs de dimension d et K_0 et K_1 sont deux matrices de taille $d \times d$. Les « vraies » valeurs des contrefactuelles dans le cas où les patients n'auraient jamais été traités sont simulées en suivant l'évolution des covariables quand $D(t_k) = 0$, c'est-à-dire suivant une distribution de Poisson de paramètre $\kappa_{D_0} + K_0 X_i(t_k)$ quel que soit le statut du traitement. Les variables à l'initiation sont simulées afin de représenter le sexe, l'âge à l'inclusion et un score de morbidité.

A l'initiation, aucun patient n'est traité. Pour les temps suivants, le traitement est simulé suivant une distribution de Bernoulli de paramètre $m \sum_{j=1}^{d} \lambda_j e^{\sum_{j=1}^{d} \lambda_j X^j(t_k)})$. Nous faisons l'hypothèse qu'une fois le traitement initié, ce dernier est poursuivi jusqu'à la fin de la période d'observation.

Le processus $N(]t_k, t_{k+1}])$ représentant les évènements d'intérêt ayant eu lieu sur l'intervalle de temps $]t_k, t_{k+1}]$ dépend du statut du traitement et des covariables dépendantes du temps. Le temps de survenue des évènements sont simulés suivant un processus de Poisson non-homogène avec pour intensité la fonction suivante : $\delta D(t_k) + \delta_0 + \sum_{j=1}^{d_Z} \delta_{Zj} Z^j + \sum_{j=1}^{d} \delta_{Xj} X_j(t_k)$. La méthode acceptation-rejet est utilisée pour générer le processus N, se reporter à [120] pour plus d'information. Tous les sujets sont censurés au temps $t_k = 11$.

Le schéma de simulation est résumé en annexe dans la partie 9.2. Le détail des valeurs des paramètres est présenté en annexe 9.2. Ces paramètres ont été fixés afin d'avoir un faible nombre d'évènements par sujets et une variance faible du processus VINGARCH de chacune des covariables. Les simulations ont été répétées 100 fois avec les mêmes paramètres, sur des bases de données de 1000 sujets. Toutes les simulations ont été réalisées avec la version 4.1 de R.

La première étape dans l'analyse de simulation réside dans la modélisation des contrefactuelles dans le scénario de l'absence de traitement grâce à un modèle VINGARCH(1,0). Les paramètres de ce modèle, β_0 et B_1 sont alors estimés.

L'effet du traitement a été estimé par l'ATT. Le « vrai » ATT a été évalué par la méthode de Monte Carlo sur la base des « vraies » contrefactuelles. Il a été estimé selon cinq estimateurs différents. Ce sont les mêmes que ceux décrits dans le chapitre précédent (se reporter au paragraphe 3.5.2) à savoir deux estimateurs non corrigés ; l'un obtenu par l'algorithme présenté dans le paragraphe 4.3.2, l'autre présenté par Gran et estimé par la fonction `aalen` du package R *timereg* ; puis deux estimateurs corrigés obtenus par l'algorithme du paragraphe 4.3.2 avec les vrais paramètres du modèle VINGARCH ou avec les paramètre estimés ; et enfin l'estimateur basé sur les « vraies » contrefactuelles.

Les estimations de l'ATT obtenues pour chacune des 100 répétitions par les 5 estimateurs cités précédemment sont comparées au « vrai » ATT, en tout temps. Les valeurs de MISE (Mean Integrated Square Error) sont calculées pour chaque estimateur de la façon suivante :

$$\text{MISE}^l = \mathbb{E}\left[\int_0^{\tau=51} (\widehat{d}^l(t) - d^{MC}(t))^2 dt\right],$$

avec $\hat{d}^l(t)$ l'estimation de l'ATT pour un estimateur sur la l-ième simulation au temps t et $d^{MC}(t)$ la « vraie » valeur de l'ATT au temps t.

Le tableau 4.1 présente les résultats obtenus dans l'analyse de simulation. La moyenne des MISE sur les 100 répétitions sont reportées et accompagnées de l'écart-type. Les conclusions sont sensiblement les mêmes que celles obtenues dans l'étude de simulation du chapitre 3 (voir 3.5.3). Les résultats obtenus avec la fonction `aalen` du package *timereg* sont semblables à ceux obtenus par notre algorithme. La correction proposée permet d'obtenir une estimation de l'ATT plus proche de la vraie valeur. En effet, les valeurs moyennes de MISE sont plus faibles pour les estimateurs corrigés. Enfin, les MISE des estimateurs corrigés sont semblables, ce qui met en évidence une bonne estimation des paramètres du modèle VINGARCH.

TABLE 4.1 – Moyenne des MISE $\pm$ écart-type pour les différents estimateurs de l'ATT

	Estimation utilisant les vraies contrefactuelles	Estimations non corrigées		Estimations corrigées	
		Fonction `aalen`	Nouvel algorithme	Vrais paramètres	Paramètres estimés
MISE	222.9 ± 130.4	442.7 ± 314.4	449.6 ± 325.4	315.2 ± 210.2	391.5 ± 220.8

Des simulations complémentaires pourraient être réalisées pour étudier l'effet de la durée d'étude mais aussi le nombre moyen d'évènements par sujets ou la variance des modèles VINGARCH sur l'effet de la correction.

4.5 - Conclusion

Nous avons proposé dans cette partie un estimateur corrigé de l'ATT en présence de facteurs de confusion discrets et dépendants du temps, basé sur la méthode de Gran [90]. Nous avons proposé l'utilisation d'un modèle INGARCH multidimensionnel, nommé VINGARCH, pour la modélisation des contrefactuelles dans le scénario où le sujet ne serait pas traité. Comme dans le cas des facteurs de confusion continus présenté dans la partie précédente, la correction de l'estimateur repose sur l'erreur commise lors de la modélisation des contrefactuelles. Une première étude de simulation a montré que pour une petite variance du modèle VINGARCH la correction proposée permet une estimation de l'ATT plus proche que celle sans correction. Des simulations complémentaires pourraient être réalisées pour évaluer l'effet de la variance sur la correction. Le modèle proposé pourra être appliqué à la problématique des patients atteints de sclérose en plaque comme présenté en introduction partie 1.2.

Nous avons explicité la correction de l'estimateur dans le cadre d'une modélisation des contrefactuelles par un modèle VINGARCH(p,0) et proposé une méthode d'estimation et une étude de simulations dans le cas d'un modèle VINGARCH(1,0). Comme dans le cas des modèles VAR, ces modèles sont simples à mettre en œuvre et permettent une écriture explicite de la correction. D'autres méthodes de modélisation pourraient être utilisées, comme les extensions du modèle INGARCH basées sur d'autres distributions que celle de Poisson. La prise en compte d'un taux important de zéro dans les observations pourrait être envisagée en fonction de la distribution des variables, se reporter à [209] pour plus de détail sur les propriétés « zero-inflation » dans les modèles INGARCH. Ce type de modèle serait pertinent dans l'étude des données SNDS notamment pour certains indicateurs. Par exemple, dans le cas de maladies chroniques nécessitant un traitement par intraveineuse, la variable correspondant au nombre mensuel de consultations infirmières sera nul pour presque tous les patients traités. Mais si on s'intéresse au nombre mensuel de consultations avec un médecin généraliste, un taux important de zéro pourrait être observé. A l'image de cet exemple, il est courant d'avoir des variables de différentes natures dans un même jeu de données. En ce sens, il serait intéressant de pouvoir considérer plusieurs types de données dans un même modèle.

Chapitre 5

Analyse des parcours de soin pour la prédiction d'évènement d'intérêt : Application à l'insuffisance cardiaque

Ce chapitre traite de l'étude des parcours de soin pour la prédiction d'un évènement terminal, comme le décès, et en particulier l'identification des parcours de soin qui semblent être liés à un sur-risque de mortalité. Cette identification pourra permettre aux praticiens de porter une attention particulière à ces patients et éventuellement modifier leur prise en charge. Cette étude se base sur la procédure présentée dans le chapitre 2, partie 2.3.

Dans le domaine de la santé, l'étude des parcours de soin ne peut pas être faite sans prendre en compte l'ordre des éléments constituants ce parcours. En effet, il est important de différencier une hospitalisation pour une complication suite à un acte médical, d'un acte médical réalisé suite à une complication. La prise en compte de cette temporalité réside dans le choix des métriques et des méthodes utilisées pour identifier les séquences fréquentes et le calcul des scores de similarité (voir section 2.3.1.3 et 2.3.1.4).

Le développement de la procédure d'analyse citée précédemment prenant en compte l'ordre des éléments des trajectoires a été motivée par l'étude des trajectoires de soin des patients nouvellement diagnostiqués pour une insuffisance cardiaque. Ces travaux sont présentés dans la suite. Il s'agit du manuscrit de l'article « Cardiovascular healthcare trajectories and mortality in heart failure » soumis à l'European heart journal.

5.1 - Introduction

Care pathways are essential monitoring tools, particularly to monitor care for chronic diseases such as diabetes, organ failure or coronary artery and/or acute coronary syndrome (ASC). Model-of-care pathways have been developed and are key in reducing re-admission rates and improving the quality of both ambulatory and in-patient care [148]. They also help construct a rule-based prediction model which provides data concerning the trends seen in patient paths between the different medical units [195]. For instance, in one study, a predictive model of ACS pathways was used to simulate future disease progression in order to anticipate subsequent healthcare needs [147].

Heart failure (HF) is a major public health concern affecting approximately 1-2% of the adult population in developed countries. Prevalence increases with age rising to over 10% among people aged 70 years or more [138]. Despite significant advances in diagnosis and therapy over the past 20 years, HF patients still have poor prognoses. Chronic HF (CHF) is characterized by repeated hospitalizations and high mortality [69] with 41% of cases resulting in either death or hospitalization during a median follow-up of 21 months [195] and a 31% mortality rate in the year after a HF caused hospitalization [50].

Predicting the mortality for HF patients is necessary for assisting clinicians to make optimal decisions during the therapeutic process. In literature, several clinical teams have introduced mortality and hospitalization prediction systems based on real-world data from clinical parameters [69, 200, 42]. Moreover, variables included in mortality models were remarkably different from those found in the HF hospitalization models [195]. A recent observational study evaluated healthcare utilization for people identified with HF and showed the complexity of patient pathways [71]. However, to our knowledge, no study has yet evaluated different healthcare pathways in HF patients and their correlative prognostic value.

Healthcare claims data from the national medical insurance schemes can be used to classify care pathways. This allows for instance the quantification of expected postoperative complications and the identification of unexpected events, as often seen in data complied after bariatric surgery which follows the same pathways [35]. In France, the national healthcare claims database known as the *Programme de Médicalisation des Systèmes d'Information* (PMSI) includes all reimbursed hospitalizations performed in both public and private hospitals. Diagnosis-related group (DRG) sequences are available by amalgamating hospital stays in order to identify the chronological pattern of patient hospitalization (named GHM in France for "Homogeneous Patient Groups"). The objectives of this study were to analyze the nationwide care pathways in the 2 years after a first hospitalization for HF and to identify the association between care pathways associated and risk of mortality in a given population.

5.2 - Method

5.2.1 . Data

5.2.1.1 . Data source and data extraction

TA retrospective study was conducted using data extracted from the EGB (Echantillon Généraliste des Bénéficiaires) database between 2008 and 2018, which is a random representative 1/97th sample of

the French population from national health insurance databases. These databases are organized into a comprehensive digital data source compiling the total consumption of outpatient and inpatient care (from the PMSI database) in both public and privately managed facilities. Insureds are assigned a unique identifier that allows their healthcare utilization to be tracked throughout their lifetimes. It may be worth noting that our study mainly used hospital data and as a result this data are exhaustive and collected as standardized discharge reports.

Each hospital stay entered in the database was categorized into the following fields : i) anonymized patient identification, ii) entry and exit date, iii) length of stay (LOS), iv) primary and associated diagnoses based on the International Classification of Disease (10th edition (ICD-10)) and v) the therapeutic procedures received. All information was incorporated into the DRG. Patient death and date of death included both in-hospital and out-of-hospital occurrences.

5.2.1.2 . Inclusion/exclusion criteria

Eligible patients were over 18 years old with an HF hospitalization identified from January 1, 2010, to December 31, 2016. HF hospitalization was defined as a stay itemized under a specific ICD-10 code of heart failure or a DRG code (see the selection criteria in Table 10.1). The date of the first hospitalization during this period was used as the index date. Patients were followed up from the index date to either the patient's death or the end of the study, which was December 31, 2018. Eligible patients were screened for any previous HF hospitalization occurring between January 1, 2008, and December 31, 2009, and if an occurrence was found, they were excluded. The design of the study is illustrated in Figure 10.1. The whole population thus included was studied in the first step.

5.2.2 . Statistical analyses

Quantitative variables were described using mean and standard deviation (SE) or median and interquartile range. Qualitative variables were described using numbers and percentages (%). Overall survival (OS) data were analyzed using Kaplan-Meier survival curves and compared using log-rank test. Study design and steps of statistical analyses are shown Figure 5.1.

5.2.3 . Hospitalizations sequences analyses (Step 2-5)

The primary outcome was to identify the frequent care pathways defined by recurrent hospitalizations. Hospitalization trajectory data were firstly identified for each patient, then, in order to identify healthcare pathways predictive of mortality, a similarity score was calculated between each patient's hospitalization trajectory and frequency so as to quantify the similarity of sequences. The scores obtained for each sequence were lastly used as covariates in a multivariate model to predict death [147].

Study design	
Step of statistical analyses	**Population**
Step 1: Survival and cardiovascular rehospitalisation analysis	**All patients :** Patients with a first hospitalisation for heart failure between 2010 and 2016
Step 2: Construction of hospitalisation sequences* **Step 3:** Identification of frequent hospitalisation sequences **Step 4:** Calculation of similarity score between patient hospitalisation sequences and each frequent hospitalization sequence **in order to evaluate prognosis value of frequent sequences** **Step 5:** Analysis of the association between frequent hospitalisation sequences and death and identification of the 20 most associated frequent sequences to death and survival	**Only rehospitalised patients** : Patients with a first hospitalisation for heart failure between 2010 and 2016 **AND** with at least one cardiovascular rehospitalisation

FIGURE 5.1 – Study design, step of statistical analyses on the populations concerned

5.2.3.1 . Construction of hospitalisations sequences (step 2)

One hospitalization sequence was defined for each patient from the index date to either patient death or the end of the study. These sequences were identified by the five first digits of each patient's DRG code and were then sorted sequentially according to date. We focused on hospitalizations associated with cardiac disease only. The list of selected DRGs is presented in Table 10.2.

5.2.3.2 . Identification of frequent sequences (step 3)

Sequences were identified from hospitalization data using a pattern mining algorithm which identifies recurrent character strings based on a fixed minimum threshold. Since we had long and dense sequences, a CM-SPAM algorithm was used to identify frequent sequences. This algorithm works by vertical extraction of patterns and makes data analysis faster and less expensive. CM-SPAM was carried out using SPMF (v.2.42) with the support of 1% [80].

5.2.3.3 . Scoring (step 4)

For each patient, a score was calculated to quantify the similarity between their hospitalization sequence and each of the frequency sequences using the Smith-Waterson algorithm. This algorithm allowed us to directly compare the successions of DRGs by considering their order of appearance in the sequence. The score obtained was between 0 and 1. The score was equal to 0 if none of the DRGs of the hospitalization sequence was present in the studied frequency sequence, and to 1 if the two sequences were identical. The `text.alignment` library (V0.1.2) of R (Version 4.0.3) was used in this study.

5.2.3.4 . Prediction model (step 5)

A gradient boosting algorithm for survival analysis with Cox's partial likelihood as the loss function was used to predict survival and identify frequent sequences with lower or high mortality. The event of interest was the delay between first HF hospitalization and death or end of study. At first HF hospitalization, patient age, gender and the similarity score of each frequency sequence were used as input features. This algorithm implements gradient boosting with regression tree base learner. It follows the strength-in-numbers principle by combining the predictions of multiple base learners to obtain a comprehensive overall model. The predictions are combined in a manner in which the addition of each base model improves the overall model. This algorithm intrinsically contains a random part to avoid over-fitting and to ensure that we had robust results, it was run 30 times.

For each repetition, feature importance was calculated. The weighted average of the 20 feature importance calculations was associated at each input feature given that the greater the weight, the greater the impact of the variable in predicting survival. To evaluate the direction of the effect of the covariates, partial dependency graphs were used. Only the results of input features with the same direction of effect in the 30 repetitions were interpreted. Results were ordered according to the average weights from the feature

importance calculations. These analyses were performed using Python (V3.8.6) with *scikit-survival*, *eli5* and *pdp* modules installed.

5.3 - Results

5.3.1 . Population characteristics

Between 2010 and 2016, 12,026 patients were identified with a first HF hospitalization and 480 were ineligible due to no healthcare consumption or unusable data. In total, 11,488 patients were included (Figure 10.2). The mean age of the study population was 78 ± 13 years with 49.4% being male. The follow-up period averaged 2.9 years and an overall survival analysis after the first HF hospitalization showed a mortality rate of 31.7% at one year, 41.5% at two years, 57.2% at three years and 78.8% at five years (Figure 5.2).

During follow-up, 80% of the patients were re-admitted at least one time for any reason and almost half (49.4%) were re-hospitalized for cardiovascular (CV) problems. During the first year following the first HF re-hospitalization, 62.1% of patients had at least one re-hospitalization (for any reason), 33.4% for CV-related issues and 12.5% for HF. On average, patients were re-hospitalized 1.32 times and 2.68 times among the re-hospitalized patients (Figure 5.2). Among the patients who died, most had a previous hospitalization coded 05M which was medical hospitalization for circulatory system disease, however, the reverse was not true (see Figure 5.3). Indeed, after a re-hospitalization for Code 05M, 27.4% of the patients died, whereas 46.6% of the patients re-hospitalized for Code 04M died. Hospitalization for pulmonary edema and respiratory distress (Code 04M) seemed to be the riskiest.

5.3.2 . Care pathway

In regard to the care pathway, the analysis of the hospitalization sequences focused on $5,704$ patients. Fifty-eight patients with a single hospital admission for Code 05M22 (other CMD 05 conditions with death : stays of less than two days) were excluded. In total, $2,222$ distinct sequences were identified. From these sequences, the CM-SPAM algorithm selected 89 recurrent sequences. Similarity scores between the sequence of each patient and the most frequent sequences were calculated. These similarity scores are presented for the twenty most frequent sequences (Figure 5.4).

After running the gradient boosting algorithm 30 times, $1,707$ pathways were identified, of which 21 pathways were identified as good prognosis (see Supplementary Table 3), 15 pathways as bad (Supplementary Table 4) and 53 frequent pathways lacked a clear sense of effect ($N = 43$) or contradictory effects between

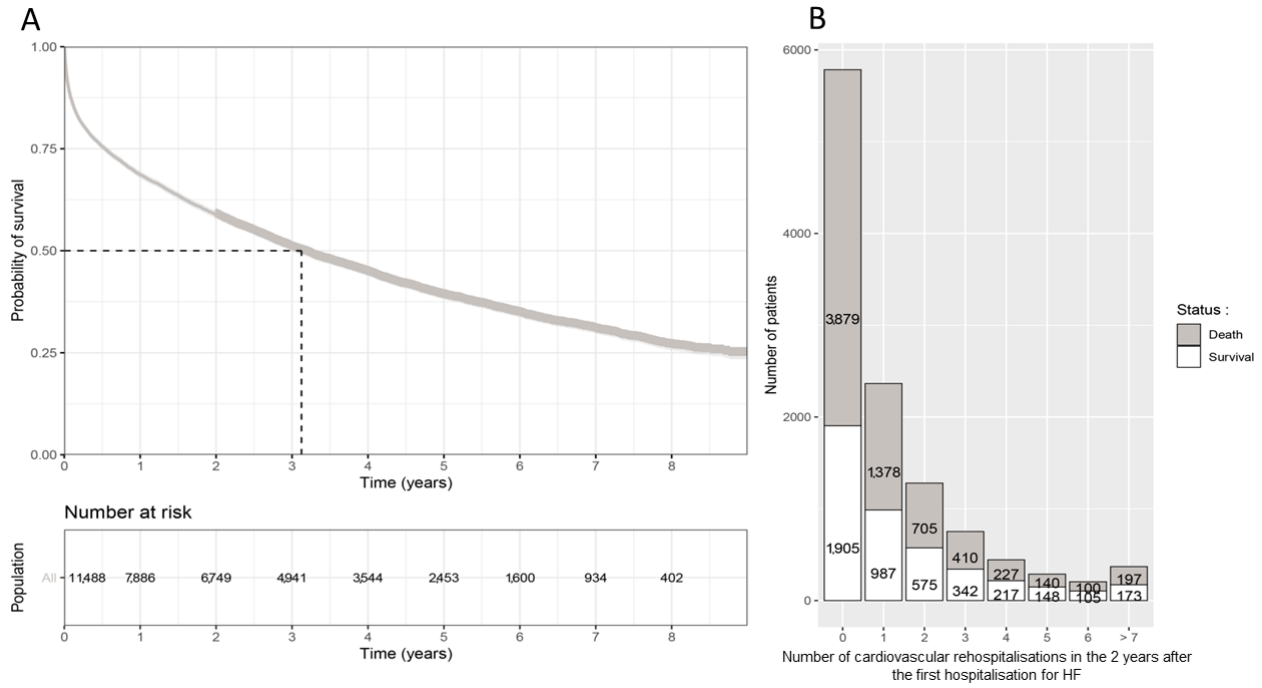

FIGURE 5.2 – **Step 1 :** Death, cardiovascular related re hospitalization and survival on all 11,488 patients (Step 1)
A : Kaplan-Meier analysis of overall survival **B** : Number of cardiovascular rehospitalizations in the 2 years after the first hospitalization for HF depending on death and survival status.

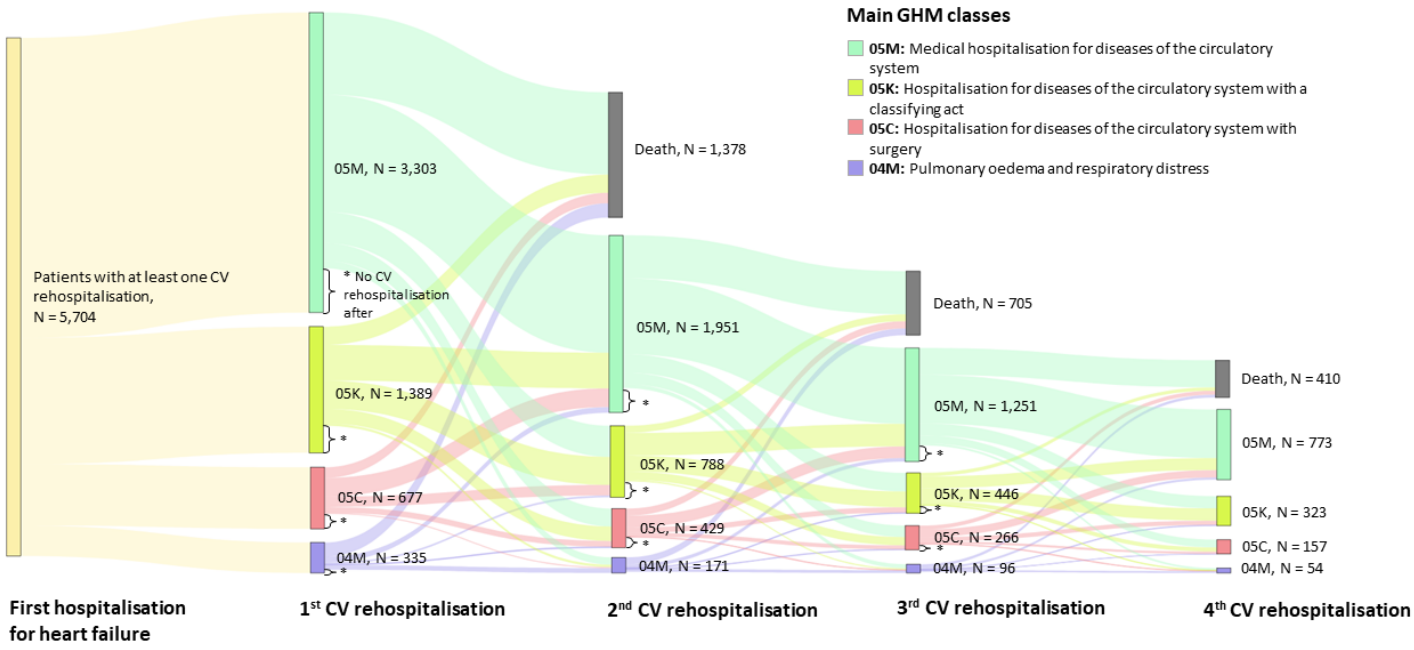

FIGURE 5.3 – Sequential representation of the first four cardiovascular rehospitalizations according to the main diagnosis related group (DRG) classes in patients with at least one cardiovascular rehospitalization (N=5,704) (STEP 2)

repetitions, ($N = 10$) and were not interpreted (Supplementary Table 5). In all models, age and gender were the key predictors of mortality. Patient age at first HF hospitalization was positively associated with mortality (0.09934 ± 0.00079) and being male with a poor prognosis (0.00872 ± 0.00019).

5.3.3 . Pathways associated with a good prognosis

The most significant pathways associated with a good prognosis are shown in Figure 5.5. These pathways involved patients who had received surgical and non-surgical device treatments such as permanent pacemaker placements (without acute myocardial infarction, congestive HF or shock) (0.00467 ± 0.00011), permanent pacemaker replacements ($0.00234 \pm 9.11e-05$), placement of a cardiac defibrillator ($0.00209 \pm 9.63e-05$), vascular stents (without myocardial infarction) followed by vascular diagnostic procedures (0.00224 ± 0.00190), vascular diagnostic procedures and vascular stents (without myocardial infarction) (0.00163 ± 0.00181), cardiac valve bio-prosthetic installation by vascular route ($0.00137 \pm 7.31e-05$), major treatments for arrhythmias by vascular route ($0,00095 \pm 9.73e-05$) and other treatments for arrhythmias by vascular route ($0.00077 \pm 5.43e-05$), valve replacement surgery with extracorporeal circulation (without cardiac

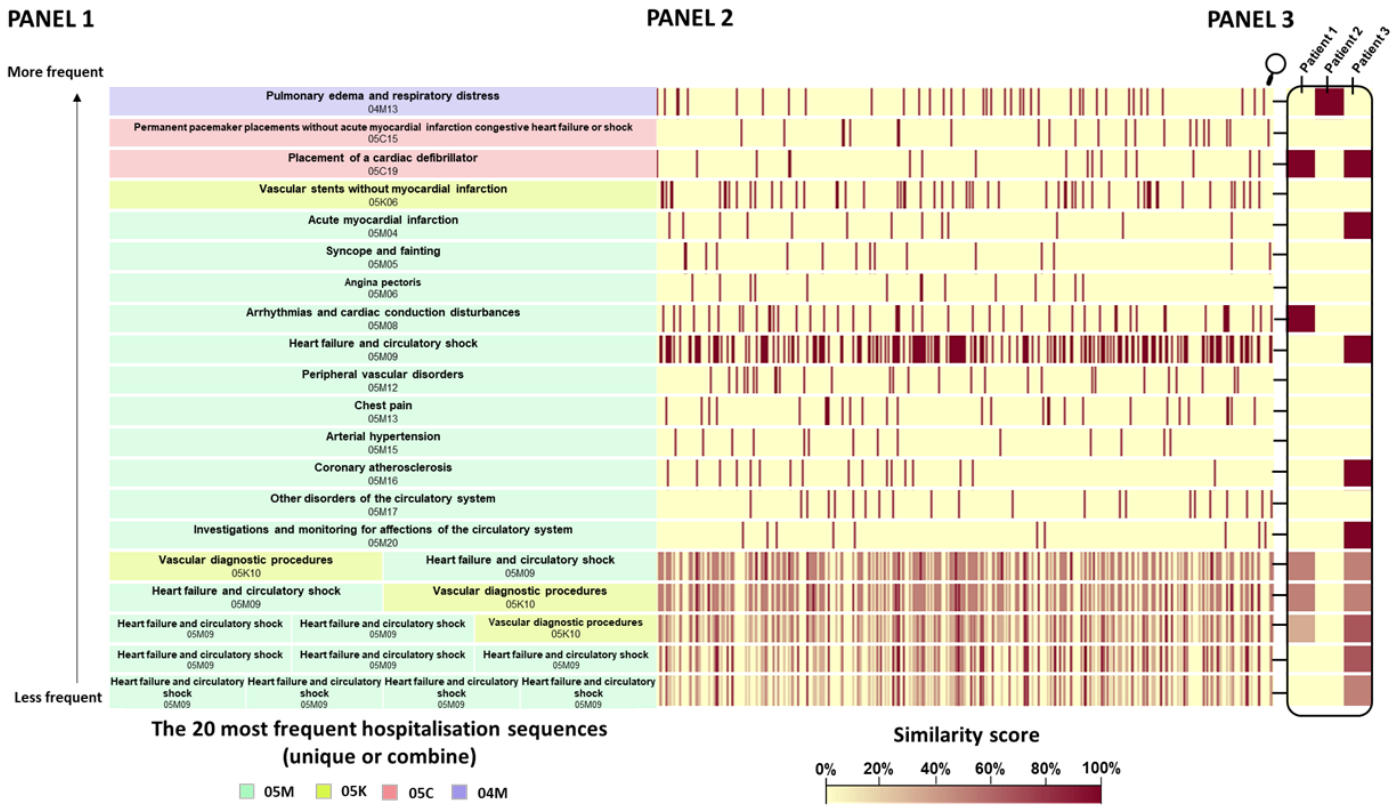

FIGURE 5.4 – **Step 4** :
PANEL 1 : The 20 most frequent hospitalization sequences (unique or combine) in patients with at least one cardiovascular rehospitalization after the first hospitalization for heart failure (HF) (N=5,704)
PANEL 2 : Similarity scores between the sequence of each patient and the sequences presented in PANEL 1. The closer the score is to 100, the more similar the pathways are.
PANEL 3 : Example of similarity scores calculated for three patients.

catheterization or coronary angiography) $(0.00078 \pm 5.44e-05)$, and lastly, coronary artery bypass grafts (without cardiac catheterization or coronary angiography) $(0.00061 \pm 6.76e-05)$.

Other patterns of hospitalization were associated with a good prognosis, namely : i) chest pain (0.00620 ± 0.00017), ii) hypertension $(0.00079 \pm 4.14e-05)$, and iii) syncope/fainting (0.00330 ± 0.00011). All pathways associated with a good prognosis are shown in Table 10.3.

5.3.4 . Pathways associated with bad prognosis

The most prominent bad prognosis pathways are also shown in Figure 5.5 and include care pathways including hospitalization for HF before or after other reasons for hospitalizations, vascular diagnostic procedures followed by HF and circulatory shock $(0.00070 \pm 9.12e-5)$, chest pain followed by HF and circulatory shock

$(0.00016 \pm 3.74e-5)$, coronary atherosclerosis followed by HF and circulatory shock $(0.00014 \pm 3.88e-5)$, heart failure and circulatory shock followed by placement of a cardiac defibrillator $(0.00016 \pm 1.69e-5)$, placement of a cardiac defibrillator followed by HF and circulatory shock $(1.12e-5 \pm 1.52e-9)$, HF and circulatory shock followed by vascular diagnostic procedures $(-1.84e-05 \pm 2.96e-5)$. All bad prognosis pathways are shown in Table 10.4.

5.4 - Discussion

In this study, we showed that a very poor prognosis continues to exist in a given population and within the complex care pathways of patients. Re-hospitalization for HF has a predominant place in the pathways associated with higher mortality. Our cohort was comparable to other similar groups having older patients (mean age of 78 years), more females than males, women older than men [69] and with poor prognosis with high rate of hospitalization and mortality.

In our study, we found that HF hospitalization was 12.6% in patients at the one-year follow-up and rose to 33.5% for all CV-related hospitalizations. These findings are close to other studies that showed a 42% CV re-hospitalization rate at the one-year followup [131] and 13% of HF hospitalization at one year in ambulant patients in registry studies [91]. In contrast, in another French study among living patients at the one year follow-up, 63% were re-admitted and 22% were re-admitted for HF [131]. These higher rates may be explained by the fact that this is the percentage among survivors and concerned a population that likely had HF over a longer period of time. Similarly, higher rates were reported in the Constantinou et al. study using administrative data. In that study, 31.8% were re-admitted at least once for HF over the one-year follow-up [42], however, patients may have been managed for HF before.

A Spanish population-based analysis observed 20.8% of all-cause hospitalizations at one-year follow up [69]. This lower hospitalization rate may be explained by the fact that the diagnosis of HF in that study was probably made on an outpatient basis, which usually included patients with a better prognosis. Lastly, the proportion of overall hospitalizations compared to the proportion of CV-related hospitalizations in our study is consistent with other published observations with most (63%) hospitalizations related to non-CV causes [86].

The mortality in our cohort was also high with 31.6% at one year, 41.5% at two years and 78.7% at five years. These rates are similar to other studies such as Blair et al. which showed 32.2% mortality at one year [25] and 75% at five years covering North America [166]. This mortality rate was also comparable to the UK retrospective population-based study in which the mortality rate after a first hospitalization for HF was 29% at one year and 40% after two years [41].

Similarly, after the first hospitalization for HF in the French population, the two-year survival rate was 60% [187]. However, our mortality was higher than that shown in a Spanish study (11.3% at one year) [69] as well as being higher than 8% in ambulant patients in registry studies [91]. One explanation for this could be the inclusion criteria of our study which was a HF-related hospitalization, an event with an inherently bad prognostic value [114] especially considering that approximately 25% of our HF patients were not previously hospitalized for HF [114]. It is also worth noting that first-time hospital admission is strongly associated with mortality compared to diagnoses formulated in primary care and is probably because the diagnosis during

an acute admission is already associated with high levels of disease severity [114].

In literature, the prognosis of HF could be estimated by the presence of previous hospitalizations for HF. Mortality rate has been shown to be inversely proportional to the time since the last hospitalization [173]. In symptomatic chronic HF patients, HF hospitalization frequency has been shown to be a key predictor of death after discharge [173]. In our study, we did not observe this, however, in our cohort, the risk of mortality was high in the immediate post-discharge period which corresponds to a finding that has already been observed in a study by Solomon et al. [173]. Our results are consistent with the European society of cardiology (ESC) guidelines which indicate that post-discharge one-year mortality can be 25 to 30% with up to more than 45% deaths or re-admission rates [138].

The two strongest predictors of mortality in our cohort were advanced age and being male (independent of age), this result was expected [170]. Moreover, in our cohort, women had less hospital re-admissions for CV-related diseases, which is an observation already described in other studies [143].

Regarding the association of care pathways and prognosis, care pathways and the inclusion of non-surgical device treatments, valve replacements and atrial fibrillation ablation were associated with a better prognosis except in cases where these invasive treatments preceded or followed hospitalization for cardiac decompensation.

Regarding implantable cardioverter-defibrillators (ICDs), our results match two other studies. The first on hospitalization occurring within 30 days before an ICD therapy replacement and its association with death [213] and the second with increasing survival rates in patients at risk of sudden cardiac death due to ventricular tachyarrhythmia ICD therapy [138]. It may also be worth noting that recent meta-analyses that pooled data from all randomized control trials and tested primary-prevention ICD over the past two decades (including the DANISH trial) have confirmed a significant reduction of all-cause mortality associated with ICD use in patients with non-ischemic cardiomyopathy [113].

Previous studies evaluated the factors associated with mortality after ICD implantation [141] however, prior hospitalizations for HF were not included in their study variables. If the benefits of the non-surgical device treatments are well in line with the latest ESC recommendations, then our main results could show the key prognostic value of HF re-hospitalization and the decisive prognostic value compared to other pathways. Indeed, when including HF re-admission, the prognostic value of the trajectory is re-oriented towards a worse prognosis. This result is in agreement with an American observational study which demonstrated that longer HF readmission-free periods are associated with decreased risks of in-hospital mortality [11]. Similarly, our result supports the fact that the number of HF hospitalizations is a strong predictor of mortality in HF patients, whereas in another study, survival inversely correlated with each HF hospitalization episode [165].

5.4.1 . Strengths and limitations

There were both strengths and limitations to our study. Firstly, the results were exploratory and descriptive with common limitations of administrative data coding and a lack of certain clinical information which made it impossible to distinguish the type of HF. Secondly, patients were included based on the primary ICD-10 diagnosis or associated code of HF which are more precise than DRG, which only group together health care acts of the same medical and economic nature. We used this codification to characterize health pathways in order to have a unique code per hospitalization while also taking into account primary diagnostic, any associated diagnoses and the procedures performed. We did not look at hospital admission before

HF diagnosis and we did not compare care pathways between regions as other studies have done which could have posed to be a limitation of this study [153]. The primary strengths of our study were the large number of individuals included and the longer observation period (over 2 years). We did not include data related to the LOS of the re-admitted patients in our analyses as others have done [11] and our study only focused on pathways.

In conclusion, this study was based on over $11,400$ patients with newly diagnosed HF during hospitalization who were culled from a representative French database. We highlighted the value of care pathways on frequent hospitalization pathways, mortality and prognosis and indicated the necessary prognostic value of HF re-hospitalization.

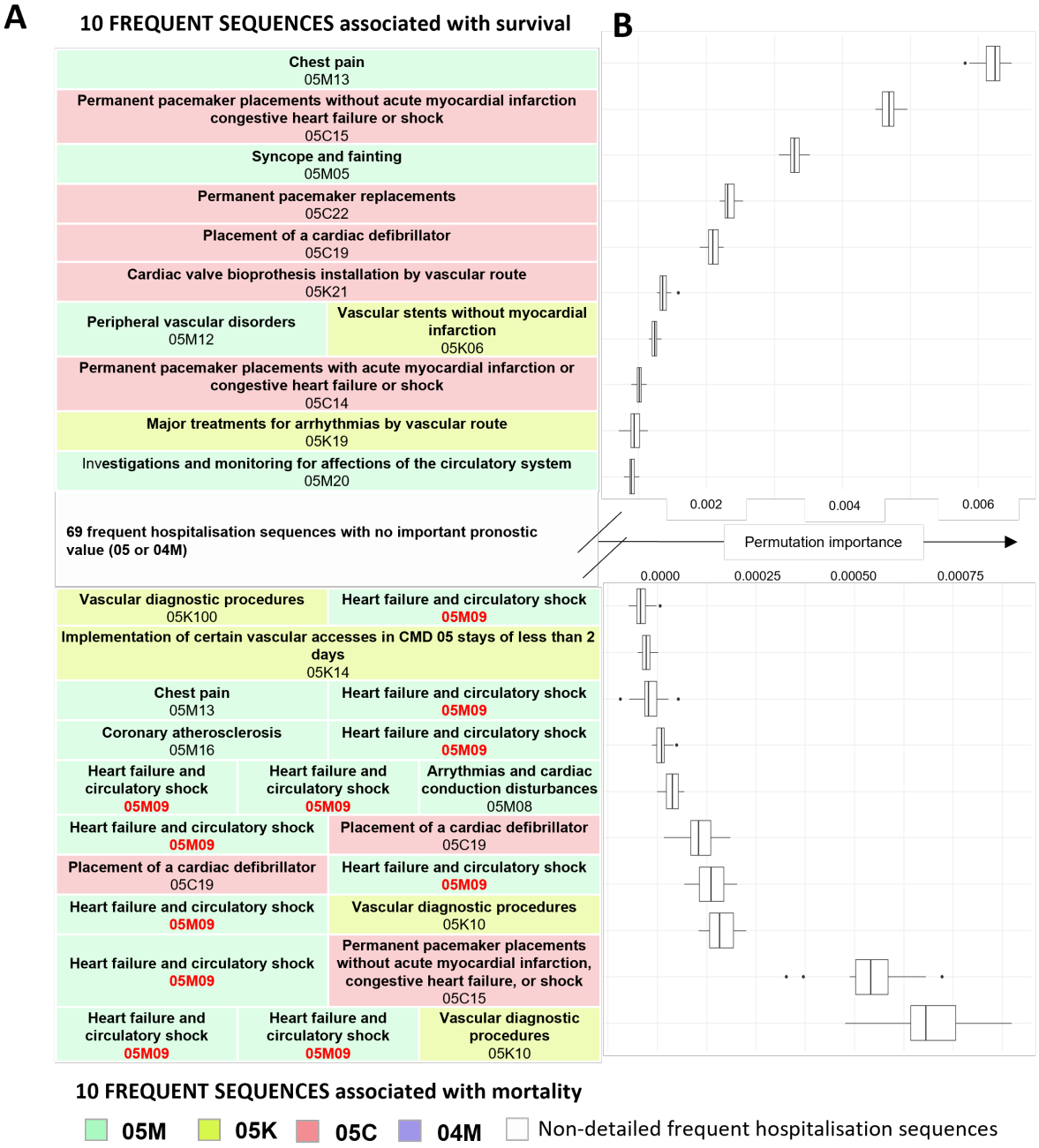

FIGURE 5.5 – Frequent sequences associated with survival or with death (STEP 5)
A. The 20 most frequent hospitalization sequences associated with survival and death
B. Graphical representation of the association between frequent hospitalization sequences presented in A and survival or death using permutation importance (the higher the score, the greater the association)

Chapitre 6

Discussion et Conclusion

Les travaux présentés dans ce manuscrit ont été développés autour des problématiques liées à l'analyse de données longitudinales, et en particulier des données des bases médico-administratives françaises. Ces données sont de plus en plus exploitées pour réaliser des projets de recherche. Entre 2018 et 2021, le nombre de dossiers déposés au Health Data Hub pour accéder aux données du SNDS était de 1176 et a augmenté de 75%, voir [185]. Deux domaines d'analyse ont été exploités ici, l'analyse causale et l'analyse de séquence. Actuellement, peu d'études sur le SNDS traitent de l'efficacité des traitements. Seuls 37 projets sur les 1176 déposés font mention de cette évaluation dans leur titre ou leur description. L'étude des parcours de soin fait, quant à elle, très souvent partie des objectifs des études. En effet, 256 projets indiquent comme objectif l'étude des prises en charge hospitalières et ambulatoires comme proxy des parcours de soin. Toutefois, ces études sont principalement descriptives, voir par exemple [21, 95] pour l'étude des parcours de traitement.

Les travaux présentés précédemment se composent de deux parties :

— La première partie s'intéressaient à l'estimation de l'ATT dans le cas de données longitudinales ;
— La seconde partie s'intéressaient à l'exploitation des parcours de soin comme facteurs prédicteurs d'un décès avec une application sur les données des patients atteints d'insuffisance cardiaque.

La première partie était motivée par le besoin de réaliser des analyses causales sur des bases de données déjà existantes, et notamment sur l'estimation de l'effet d'un traitement en population générale. Nous nous sommes intéressés spécifiquement au modèle développé par Gran et collaborateurs [90] qui permet d'estimer l'effet moyen du traitement chez les patients traités avec une méthodologie facile à implémenter. Dans un premier temps, nous avons généralisé le modèle à l'évaluation de l'effet sur un résultat qui peut être répété dans le temps, comme les rechutes d'un cancer. Ensuite, l'étude de ce modèle pour l'estimation de l'ATT sur la survenue d'un évènement dans le temps et en présence de facteurs de confusion continus et dépendants du temps avec des covariables à l'initiation, nous a permis de mettre en évidence un biais dans l'estimation de l'effet. Ce dernier est induit par les erreurs avec lesquelles sont observées les trajectoires des facteurs de confusion longitudinaux. Nous avons proposé une correction de l'estimateur et l'avons explicitée dans le cas d'un modèle autorégressif vectoriel avec un retard de 1. Nous avons montré que dans les différents scénarios de modélisation, l'estimateur corrigé améliore les performances en estimation.

De par nature, les données du SNDS sont discrètes. L'utilisation du modèle de Gran pour de telles

données nécessitait de trouver des modèles type AR multidimensionnel pour données discrètes. Nous avons ainsi formalisé le modèle INGARCH à plusieurs dimensions, nommé VINGARCH et l'avons utilisé pour modéliser les contrefactuelles dans le cas de facteurs de confusion discret. Nous avons explicité l'estimateur corrigé de Gran dans le cadre général d'un modèle VINGARCH avec un retard de p temps sur les covariables et q temps sur la moyenne du processus. De premières simulations, dans le cas spécifique d'un modèle VINGARCH à un seul temps de retard sur les covariables, ont montré que l'estimation corrigée améliore les performances en estimation.

La seconde partie a consisté à étendre une procédure permettant d'exploiter les parcours de soin comme variables prédictives d'un évènement terminal. Cette procédure combine des méthodes d'analyse de séquences (identification des séquences fréquentes et calcule du score de similarité) ainsi que des méthodes de prédiction permettant de prendre en compte l'ordre de survenue des évènements dans la séquence. Cette procédure a été appliquée à l'étude des séquences de réhospitalisations cardiovasculaires chez des patients atteints d'insuffisance cardiaque. Ce travail a permis de mettre en évidence des parcours de soin associés à un surrisque de décès dans une population de sujets nouvellement diagnostiqués pour insuffisance cardiaque. Cette identification résulte de la mise en œuvre d'une procédure d'analyse prenant en compte l'ordre des réhospitalisations dans le parcours de soin. Nous avons notamment montré un bénéfice des traitements par dispositifs non chirurgicaux ou des remplacements de valves si ces derniers ne sont pas suivis ou précédés d'une décompensation cardiaque.

6.1 - Perspectives et extensions du modèle de Gran

Comme nous l'avons vu précédemment, nous avons, dans ces travaux, proposé une estimation de l'effet d'un traitement, en utilisant l'ATT, sur un résultat qui peut être répété dans le temps, en présence de covariables observées à l'inclusion et de facteurs de confusion soit continus, soit discrets. Une correction de l'estimateur a été proposée quel que soit la méthode de modélisation des contrefactuelles. Une écriture explicite de la correction a été proposée et testée par simulation dans des cas particuliers de modélisation des contrefactuelles : le modèle VAR(1) et le modèle VINGARCH(1,0). Dans un premier temps, l'effet de l'augmentation du retard dans la modélisation des contrefactuelles sur la correction pourrait être étudié.

Une première extension serait d'utiliser d'autres méthodes plus spécifiques aux maladies ou aux covariables dépendantes du temps pour la modélisation des contrefactuelles. Dans le cas de facteurs de confusion continus des modèles non-linéaires pourraient être envisagés, voir par exemple [181, 61] pour les modèles autorégressifs à transition lisse, ou des modélisations par équations différentielles. De leur côté, les facteurs de confusion discrets pourraient être modélisés par des extensions du modèle INGARCH basées sur d'autres distributions que celle de Poisson, voir par exemple [211] pour la distribution négative binomiale ou [169] pour des distributions de Poisson mixtes. Dans certains jeux de données, le taux de 0 peut être particulièrement élevé. Cela peut notamment être le cas dans les données issues du SNDS. En effet, dans l'étude des maladies chroniques, étudier le nombre de consultations chez un médecin généraliste par mois introduira peu de zéro mais le nombre de consultations avec un spécialiste conduira a traiter des covariables avec un taux important de 0. En ce sens, les méthodes basées sur les propriétés « zero-inflation » seraient une bonne alternative au modèle INGARCH (voir [209]). Le choix de la méthode de modélisation des contrefactuelles

devra être fait en fonction de la nature des observations. Il faut toutefois noter que plus le modèle est complexe plus l'erreur de modélisation, et par extension la correction de l'estimateur de l'ATT, sera difficile à écrire explicitement et à calculer.

Une autre extension serait de proposer une méthode de modélisation des contrefactuelles dans un jeu de données contenant des facteurs de confusion de différentes natures, continus et discrets. Une première option est de faire l'hypothèse que les facteurs de confusion discrets sont indépendants de ceux continus. Dans ce cas, les modélisations peuvent être faites indépendamment sur chaque type de covariables. Mais cette hypothèse n'est pas réaliste dans la plus part des applications.

6.2 - Perspectives et extensions à l'étude des parcours de soin

Dans la deuxième partie de ce travail, nous avons étendu une procédure d'étude des trajectoires constituées des réhospitalisations cardiovasculaires pour étudier leur association à un surrisque de mortalité. Nous avons montrer qu'une telle procédure d'analyse est applicable aux données du SNDS. Une analyse complémentaire est en cours de réalisation. Elle se focalise sur les sujets toujours en vie à deux ans. De cette façon, le biais d'immortalité qui peut exister dans les travaux présentés est évité mais les décès précoces ne sont pas pris en compte.

Une première extension résiderait sur les données de l'étude. Nous nous sommes focalisés sur les hospitalisations pour causes cardiovasculaires afin de limiter les trajectoires distinctes. Seuls l'âge et le sexe ont été introduits dans le modèle de prédiction de la survie. Les éventuelles comorbidités n'ont pas été prises en compte ce qui peut avoir un effet sur la mortalité.

Dans la procédure d'analyse, les trajectoires de soin ne sont pas directement incluses dans les modèles de prédiction mais le sont par le biais de leurs scores de similarité avec les séquences fréquentes identifiées. Une extension possible est de se servir des trajectoires pour identifier des sous-groupes de patients homogènes.

Une autre extension possible serait de prendre en compte la temporalité entre deux éléments constituant les trajectoires. En effet, les méthodes utilisées prennent en compte l'ordre de survenue des hospitalisations mais pas le temps qui sépare deux d'entre elles. Or le temps qui sépare deux réhospitalisations peut être informatif. Par exemple, une hospitalisation survenant dans les 30 jours suivants une chirurgie peut être interprétée comme une complication de la chirurgie. Si cette hospitalisation a lieu plus d'un an après, on peut penser qu'il n'y a pas de lien direct de l'une sur l'autre.

Bibliographie

[1] Odd O AALEN. « A linear regression model for the analysis of life times ». In : *Statistics in medicine* 8.8 (1989), p. 907-925.

[2] Odd O AALEN. « Further results on the non-parametric linear regression model in survival analysis ». In : *Statistics in medicine* 12.17 (1993), p. 1569-1588.

[3] Odd O AALEN. « Nonparametric inference for a family of counting processes ». In : *The Annals of Statistics* (1978), p. 701-726.

[4] Odd O AALEN, Richard J COOK et Kjetil RØYSLAND. « Does Cox analysis of a randomized survival study yield a causal treatment effect ? » In : *Lifetime data analysis* 21.4 (2015), p. 579-593.

[5] Odd O AALEN et Nina GUNNES. « A dynamic approach for reconstructing missing longitudinal data using the linear increments model ». In : *Biostatistics* 11.3 (2010), p. 453-472.

[6] Odd O AALEN et al. « Can we believe the DAGs ? A comment on the relationship between causal DAGs and mechanisms ». In : *Statistical methods in medical research* 25.5 (2016), p. 2294-2314.

[7] Alberto ABADIE et Guido W IMBENS. *Matching on the estimated propensity score*. 2. 2016, p. 781-807.

[8] Younathan ABDIA et al. « Propensity scores based methods for estimating average treatment effect and average treatment effect among treated : a comparative study ». In : *Biometrical Journal* 59.5 (2017), p. 967-985.

[9] Rakesh AGRAWAL, Tomasz IMIELIŃSKI et Arun SWAMI. « Mining association rules between sets of items in large databases ». In : (1993), p. 207-216.

[10] Rakesh AGRAWAL, Ramakrishnan SRIKANT et al. « Fast algorithms for mining association rules ». In : *Proc. 20th int. conf. very large data bases, VLDB*. T. 1215. Citeseer. 1994, p. 487-499.

[11] Ahmed M ALTIBI et al. « Readmission-free period and in-hospital mortality at the time of first readmission in acute heart failure patients—NRD-based analysis of 40,000 heart failure readmissions ». In : *Heart failure reviews* 26.1 (2021), p. 57-64.

[12] Per Kragh ANDERSEN et Richard D GILL. « Cox's regression model for counting processes : a large sample study ». In : *The annals of statistics* (1982), p. 1100-1120.

[13] Laurent ARNAUD et al. « Prevalence and incidence of systemic lupus erythematosus in France : a 2010 nation-wide population-based study ». In : *Autoimmunity reviews* 13.11 (2014), p. 1082-1089.

[14] Peter C AUSTIN. « Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples ». In : *Statistics in medicine* 28.25 (2009), p. 3083-3107.

[15] Peter C AUSTIN. « The performance of different propensity score methods for estimating marginal odds ratios ». In : *Statistics in medicine* 26.16 (2007), p. 3078-3094.

[16] Peter C AUSTIN. « The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies ». In : *Statistics in medicine* 29.20 (2010), p. 2137-2148.

[17] Peter C AUSTIN, Paul GROOTENDORST et Geoffrey M ANDERSON. « A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects : a Monte Carlo study ». In : *Statistics in medicine* 26.4 (2007), p. 734-753.

[18] Jay AYRES et al. « Sequential pattern mining using a bitmap representation ». In : *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, p. 429-435.

[19] Jessica K BARRETT, Robin HENDERSON et Susanne ROSTHØJ. « Doubly robust estimation of optimal dynamic treatment regimes ». In : *Statistics in biosciences* 6.2 (2014), p. 244-260.

[20] Onur BASER. « Too much ado about propensity score models ? Comparing methods of propensity score matching ». In : *Value in Health* 9.6 (2006), p. 377-385.

[21] Manon BELHASSEN et al. « Anti-osteoporotic treatments in France : initiation, persistence and switches over 6 years of follow-up ». In : *Osteoporosis International* 28.3 (2017), p. 853-862.

[22] Stephanie BERTHET et al. « Prevalence of diabetes in France and drug use : study based on the French Pharmacovigilance Database ». In : *Therapie* 62.6 (2007), p. 483-488.

[23] Julien BEZIN et al. « The national healthcare system claims databases in France, SNIIRAM and EGB : powerful tools for pharmacoepidemiology ». In : *Pharmacoepidemiology and drug safety* 26.8 (2017), p. 954-962.

[24] Christopher M BISHOP et Nasser M NASRABADI. *Pattern recognition and machine learning*. T. 4. 4. Springer, 2006.

[25] John EA BLAIR, Mark HUFFMAN et Sanjiv J SHAH. « Heart failure in north america ». In : *Current cardiology reviews* 9.2 (2013), p. 128-146.

[26] J Martin BLAND et Douglas G ALTMAN. « The logrank test ». In : *Bmj* 328.7447 (2004), p. 1073.

[27] Abraham BOOKSTEIN, Vladimir A KULYUKIN et Timo RAITA. « Generalized hamming distance ». In : *Information Retrieval* 5.4 (2002), p. 353-375.

[28] Imad BOU-HAMAD, Denis LAROCQUE et Hatem BEN-AMEUR. « A review of survival trees ». In : *Statistics surveys* 5 (2011), p. 44-71.

[29] Leo BREIMAN. « Bagging predictors ». In : *Machine learning* 24.2 (1996), p. 123-140.

[30] Leo BREIMAN. « Random forests ». In : *Machine learning* 45.1 (2001), p. 5-32.

[31] Norman E BRESLOW. « Discussion of Professor Cox's paper ». In : *J Royal Stat Soc B* 34 (1972), p. 216-217.

[32] Marco CALIENDO et Sabine KOPEINIG. « Some practical guidance for the implementation of propensity score matching ». In : *Journal of economic surveys* 22.1 (2008), p. 31-72.

[33] Laure CARCAILLON-BENTATA et al. « Prevalence and incidence of young onset dementia and associations with comorbidities : A study of data from the French national health data system ». In : *PLoS medicine* 18.9 (2021), e1003801.

[34] Joong Hyuk CHANG et Won Suk LEE. « Finding recent frequent itemsets adaptively over online data streams ». In : *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, p. 487-492.

[35] Anaïs CHARLES-NELSON, Andrea LAZZATI et Sandrine KATSAHIAN. « Analysis of trajectories of care after bariatric surgery using data mining method and health administrative information systems ». In : *Obesity Surgery* 30.6 (2020), p. 2206-2216.

[36] Chin-Hoong CHEE et al. « Algorithms for frequent itemset mining : a literature review ». In : *Artificial Intelligence Review* 52.4 (2019), p. 2603-2621.

[37] Yifei CHEN et al. « A gradient boosting algorithm for survival analysis via direct optimization of concordance index ». In : *Computational and mathematical methods in medicine* 2013 (2013).

[38] Antonio CIAMPI et al. « RECPAM : a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features ». In : *Computer methods and programs in biomedicine* 26.3 (1988), p. 239-256.

[39] Alastair COMPSTON et Alasdair COLES. « Multiple sclerosis ». In : *The Lancet* 372.9648 (2008), p. 1502-1517. ISSN : 01406736. DOI : 10.1016/S0140-6736(08)61620-7. URL : https://linkinghub.elsevier.com/retrieve/pii/S0140673608616207 (visité le 04/06/2022).

[40] Christian CONFAVREUX et Sandra VUKUSIC. « The clinical epidemiology of multiple sclerosis ». In : *Neuroimaging Clinics of North America* 18.4 (2008), p. 589-622.

[41] Nathalie CONRAD et al. « Diagnostic tests, drug prescriptions, and follow-up patterns after incident heart failure : a cohort study of 93,000 UK patients ». In : *PLoS medicine* 16.5 (2019), e1002805.

[42] Panayotis CONSTANTINOU et al. « Patient stratification for risk of readmission due to heart failure by using nationwide administrative data ». In : *Journal of Cardiac Failure* 27.3 (2021), p. 266-276.

[43] *Corrélation et causalité*. https://www.jmp.com/fr_fr/statistics-knowledge-portal/what-is-correlation/correlation-vs-causation.html.

[44] Corinna CORTES et Vladimir VAPNIK. « Support-vector networks ». In : *Machine learning* 20.3 (1995), p. 273-297.

[45] David R COX. « Partial likelihood ». In : *Biometrika* 62.2 (1975), p. 269-276.

[46] David R COX. « Regression models and life-tables ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 34.2 (1972), p. 187-202.

[47] Sally CUPITT. « Randomised Controlled Trials – Gold Standard or Fool's Gold ? The Role of Experimental Methods in Voluntary Sector Impact Assessment. » In : The National Council for Voluntary Organi- sations (NCVO) (mai 2015).

[48] Fred J DAMERAU. « A technique for computer detection and correction of spelling errors ». In : *Communications of the ACM* 7.3 (1964), p. 171-176.

[49] Rhian M DANIEL et al. « Methods for dealing with time-dependent confounding ». In : *Statistics in medicine* 32.9 (2013), p. 1584-1618.

[50] Christine DE PERETTI et al. « Prévalences et statut fonctionnel des cardiopathies ischémiques et de l'insuffisance cardiaque dans la population adulte en France : apports des enquêtes déclaratives Handicap-Santé ». In : *Bulletin épidémiologique hebdomadaire* 9-10 (2014), p. 172-181.

[51] Laurence DE ROQUEFEUIL et al. « L'échantillon généraliste de bénéficiaires : représentativité, portée et limites ». In : *Pratiques et organisation des Soins* 40.3 (2009), p. 213-223.

[52] François DELAHAYE et Guy DE GEVIGNEY. « Epidémiologie de l'insuffisance cardiaque ». In : 50.1 (2001), p. 6-11.

[53] *Délibération n° 2017-013 du 19 janvier 2017 autorisant l'Assistance publique – Hôpitaux de Paris à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité un entrepôt de données de santé, dénommé « EDS ». (demande d'autorisation n° 1980120)*.

[54] *Délibération n° 2018-295 du 19 juillet 2018 autorisant le Centre Hospitalier Universitaire de Nantes à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité un entrepôt de données de santé, dénommé « EHOP ». (Demande d'autorisation n° 2129203)*.

[55] *Délibération n° 2019-103 du 5 septembre 2019 autorisant le centre hospitalier universitaire de Lille à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité un entrepôt de données de santé, intitulé « INCLUDE »*.

[56] *Délibération n° 2019-124 du 10 octobre 2019 autorisant le Centre hospitalier universitaire Grenoble Alpes à mettre en œuvre un traitement de données à caractère personnel ayant pour finalité un entrepôt de données de santé dénommé « CHUGA-EDS »*.

[57] *Délibération n° 2020-028 du 27 février 2020 autorisant le Centre hospitalier universitaire de Rennes à mettre en œuvre un traitement de données à caractère personnel ayant pour finalité un entrepôt de données de santé dénommé « eHop Rennes » (Demande d'autorisation n° 2212496)*. URL : https://www.doctrine.fr/d/CNIL/2020/CNILTEXT000042105595.

[58] Iván DÍAZ et al. « Nonparametric causal effects based on longitudinal modified treatment policies ». In : *Journal of the American Statistical Association* (2021), p. 1-16.

[59] Peter DIGGLE, Daniel FAREWELL et Robin HENDERSON. « Analysis of longitudinal data with drop-out : objectives, assumptions and a proposal ». In : *Journal of the royal statistical society : Series C (Applied Statistics)* 56.5 (2007), p. 499-550.

[60] Peter DIGGLE et al. *Analysis of longitudinal data*. Oxford university press, 2002.

[61] Dick van DIJK, Timo TERÄSVIRTA et Philip Hans FRANSES. « Smooth transition autoregressive models—a survey of recent developments ». In : *Econometric reviews* 21.1 (2002), p. 1-47.

[62] Randal DOUC, Paul DOUKHAN et Eric MOULINES. « Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator ». In : *Stochastic Processes and their Applications* 123.7 (2013), p. 2620-2647.

[63] DRESS. *L'état de santé de la population en France, Rapport 2017*. DRESS, 2017, p. 97-109.

[64] Harris DRUCKER et al. « Support vector regression machines ». In : *Advances in neural information processing systems* 9 (1996).

[65] Oliver DUKES et Stijn VANSTEELANDT. « A note on G-estimation of causal risk ratios ». In : *American journal of epidemiology* 187.5 (2018), p. 1079-1084.

[66] Gilles EDAN et Marc COUSTANS. « Évolution et surveillance de la sclérose en plaques : Sclérose en plaques ». In : *La Revue du praticien (Paris)* 49.17 (1999), p. 1866-1871.

[67] Ludger EVERS et Claudia-Martina MESSOW. « Sparse kernel methods for high-dimensional survival data ». In : *Bioinformatics* 24.14 (2008), p. 1632-1638.

[68] Daniel Mark FAREWELL. « Linear models for censored data ». 2006.

[69] Núria FARRÉ et al. « Real world heart failure epidemiology and outcome : A population-based analysis of 88,195 patients ». In : *PloS one* 12.2 (2017), e0172745.

[70] Bruno FAUTREL et al. « Healthcare service utilisation costs attributable to rheumatoid arthritis in France : Analysis of a representative national claims database ». In : *Joint Bone Spine* 83.1 (2016), p. 53-56.

[71] Sarah F FELDMAN et al. « French annual national observational study of 2015 outpatient and inpatient healthcare utilization by approximately half a million patients with previous heart failure diagnosis ». In : *Archives of Cardiovascular Diseases* 114.1 (2021), p. 17-32.

[72] Tommaso FELLIN et Michael HALASSA. *Neuronal Network Analysis : Concepts and Experimental Approaches*. Springer, 2012.

[73] René FERLAND, Alain LATOUR et Driss ORAICHI. « Integer-valued GARCH process ». In : *Journal of time series analysis* 27.6 (2006), p. 923-942.

[74] Robert B FETTER et al. « Case mix definition by diagnosis-related groups ». In : *Medical care* 18.2 (1980), p. i-53.

[75] Zoe FEWELL et al. « Controlling for time-dependent confounding using marginal structural models ». In : *The Stata Journal* 4.4 (2004), p. 402-420.

[76] Konstantinos FOKIANOS et Roland FRIED. « Interventions in log-linear Poisson autoregression ». In : *Statistical Modelling* 12.4 (2012), p. 299-322.

[77] Konstantinos FOKIANOS et Dag TJØSTHEIM. « Log-linear Poisson autoregression ». In : *Journal of Multivariate Analysis* 102.3 (2011), p. 563-578.

[78] Césaire JK FOUODO et al. « Support Vector Machines for Survival Analysis with R. » In : *R Journal* 10.1 (2018).

[79] Philippe FOURNIER-VIGER et al. « Fast vertical mining of sequential patterns using co-occurrence information ». In : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2014, p. 40-52.

[80] Philippe FOURNIER-VIGER et al. « Fast vertical mining of sequential patterns using co-occurrence information ». In : *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2014, p. 40-52.

[81] Santé publique FRANCE. *Insuffisance Cardiaque*. URL : `https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-cardiovasculaires-et-accident-vasculaire-cerebral/insuffisance-cardiaque`.

[82] Jerome H FRIEDMAN. « Greedy function approximation : a gradient boosting machine ». In : *Annals of statistics* (2001), p. 1189-1232.

[83] Stéphane GAÏFFAS et Agathe GUILLOUX. « High-dimensional additive hazards models and the Lasso ». In : *Electronic Journal of Statistics* 6 (2012), p. 522-546.

[84] Dragan GAMBERGER et Nada LAVRAC. « Expert-guided subgroup discovery : Methodology and application ». In : *Journal of Artificial Intelligence Research* 17 (2002), p. 501-527.

[85] Tristant GAUDIAUT. *Le Big Bang du Big Data*. statista. 2021. URL : https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde/.

[86] Yariv GERBER et al. « A contemporary appraisal of the heart failure epidemic in Olmsted County, Minnesota, 2000 to 2010 ». In : *JAMA internal medicine* 175.6 (2015), p. 996-1004.

[87] Richard D GILL et James M ROBINS. « Causal inference for complex longitudinal data : the continuous case ». In : *Annals of Statistics* (2001), p. 1785-1811.

[88] Marcel GOLDBERG et al. « CONSTANCES : a general prospective population-based cohort for occupational and environmental epidemiology : cohort profile ». In : *Occupational and Environmental Medicine* 74.1 (2017), p. 66-71.

[89] Laura A GRAHAM et al. « Exploring trajectories of health care utilization before and after surgery ». In : *Journal of the American College of Surgeons* 228.1 (2019), p. 116-128.

[90] Jon Michael GRAN et al. « Estimating the treatment effect on the treated under time-dependent confounding in an application to the Swiss HIV Cohort Study ». In : *Journal of the Royal Statistical Society : Series C (Applied Statistics)* 67.1 (2018), p. 103-125.

[91] Amy GROENEWEGEN et al. « Epidemiology of heart failure ». In : *European journal of heart failure* 22.8 (2020), p. 1342-1356.

[92] Antoine GUÉGUEN et Olivier GOUT. « New treatments and strategy in multiple sclerosis ». In : *La Revue du Praticien* 66.1 (2016), p. 44-50.

[93] Jiawei HAN, Jian PEI et Hanghang TONG. *Data mining : concepts and techniques*. Morgan kaufmann, 2022.

[94] John M HANCOCK. « Jaccard distance (Jaccard index, Jaccard similarity coefficient) ». In : *Dictionary of Bioinformatics and Computational Biology* (2004).

[95] Brooke HARROW et al. « Patterns of Use and Clinical Outcomes with Long-Acting Somatostatin Analogues for Neuroendocrine Tumors : A Nationwide French Retrospective Cohort Study in the Real-Life Setting ». In : *Advances in Therapy* 39.4 (2022), p. 1754-1771.

[96] HAS. *Promouvoir les parcours de soins personnalisés pour les malades chroniques*. Haute autorité de santé (HAS). 2012. URL : https://www.has-sante.fr/jcms/c_1247611/promouvoir-les-parcours-de-soins-personnalises-pour-les-malades-chroniques.

[97] Ralf HERBRICH, Thore GRAEPEL et Klaus OBERMAYER. « Support vector learning for ordinal regression ». In : (1999).

[98] Miguel HERNÁN et James M. ROBINS. *Causal inference*. Chapman & Hall/CRC monographs on statistics & applied probability. Boca Raton : Chapman & Hall/CRC, 2021. 352 p. ISBN : 978-1-4200-7616-5.

[99] Miguel A HERNÁN et al. « Observation plans in longitudinal studies with time-varying treatments ». In : *Statistical methods in medical research* 18.1 (2009), p. 27-52.

[100] Tin Kam HO. « The random subspace method for constructing decision forests ». In : *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998), p. 832-844.

[101] Joseph W HOGAN et Tony LANCASTER. « Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies ». In : *Statistical Methods in Medical Research* 13.1 (2004), p. 17-48.

[102] Torsten HOTHORN et al. « Bagging survival trees ». In : *Statistics in medicine* 23.1 (2004), p. 77-91.

[103] Ronald A HOWARD. « Dynamic programming ». In : *Management Science* 12.5 (1966), p. 317-348.

[104] Shigao HUANG et al. « Artificial intelligence in cancer diagnosis and prognosis : Opportunities and challenges ». In : *Cancer letters* 471 (2020), p. 61-71.

[105] M ILAYARAJA et T MEYYAPPAN. « Efficient data mining method to predict the risk of heart diseases through frequent itemsets ». In : *Procedia Computer Science* 70 (2015), p. 586-592.

[106] Hemant ISHWARAN et al. « Random survival forests ». In : *The annals of applied statistics* 2.3 (2008), p. 841-860.

[107] Kai-Ming JHANG et al. « Using the apriori algorithm to classify the care needs of patients with different types of dementia ». In : *Patient preference and adherence* 13 (2019), p. 1899.

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

[109] Mohammad Ehsanul KARIM et al. « Comparison of statistical approaches dealing with time-dependent confounding in drug effectiveness studies ». In : *Statistical methods in medical research* 27.6 (2018), p. 1709-1722.

[110] Edward H KENNEDY et al. « The effect of salvage therapy on survival in a longitudinal study with treatment by indication ». In : *Statistics in medicine* 29.25 (2010), p. 2569-2580.

[111] Faisal M KHAN et Valentina Bayer ZUBEK. « Support vector regression for censored data (SVRc) : a novel tool for survival analysis ». In : *2008 Eighth IEEE International Conference on Data Mining*. IEEE. 2008, p. 863-868.

[112] John P. KLEIN et Melvin L. MOESCHBERGER. *Survival analysis : techniques for censored and truncated data*. 2. ed., corr. 3. print. Statistics for biology and health. New York, NY : Springer, 2010. 536 p.

[113] Lars KØBER et al. « Defibrillator implantation in patients with nonischemic systolic heart failure ». In : *New England Journal of Medicine* 375.13 (2016), p. 1221-1230.

[114] Stefan KOUDSTAAL et al. « Prognostic burden of heart failure recorded in primary care, acute hospital admissions, or both : a population-based linked electronic health record cohort study in 2.1 million people ». In : *European journal of heart failure* 19.9 (2017), p. 1119-1127.

[115] Yuvaraj KRISHNAMOORTHY et Tanveer REHMAN. « Impact of antenatal care visits on childhood immunization : a propensity score-matched analysis using nationally representative survey ». In : *Family Practice* 39.4 (2022), p. 603-609.

[116] M. J. van der LAAN et James M. ROBINS. *Unified methods for censored longitudinal data and causality*. Springer series in statistics. New York : Springer, 2003. 397 p. ISBN : 978-0-387-95556-8.

[117] Walker H LAND JR et al. « A new tool for survival analysis : evolutionary programming/evolutionary strategies (EP/ES) support vector regression hybrid using both censored/non-censored (event) data ». In : *Procedia Computer Science* 6 (2011), p. 267-272.

[118] Emmanuelle LERAY et Sandra VUKUSIC. « Épidémiologie de la sclérose en plaques et nouveaux critères diagnostiques ». In : *La France pays à forte prévalence. La Revue du Praticien* 66.1 (2016), p. 32-6.

[119] Vladimir I LEVENSHTEIN et al. « Binary codes capable of correcting deletions, insertions, and reversals ». In : *Soviet physics doklady*. T. 10. 8. Soviet Union. 1966, p. 707-710.

[120] PA W LEWIS et Gerald S SHEDLER. « Simulation of nonhomogeneous Poisson processes by thinning ». In : *Naval research logistics quarterly* 26.3 (1979), p. 403-413.

[121] Yun LI, Douglas E SCHAUBEL et Kevin HE. « Matching methods for obtaining survival functions to estimate the effect of a time-dependent treatment ». In : *Statistics in biosciences* 6.1 (2014), p. 105-126.

[122] Tobias LIBOSCHIK et al. « Modelling interventions in INGARCH processes ». In : *International Journal of Computer Mathematics* 93.4 (2016), p. 640-657.

[123] Heng LIU. *Some models for time series of counts*. Columbia University, 2012.

[124] *Loi de modernisation du système de santé français*. URL : https://www.legifrance.gouv.%20fr/affichTexteArticle.do;jsessionid=8C3901DF701DE7FCC51453E8D105A11D.tpdila20v_1?%20idArticle=JORFARTI000031914480&cidTexte=JORFTEXT000031912641&dateTexte=29990101&%20categorieLien=id.

[125] *Loi n°98-1194 du 23 décembre 1998 de financement de la sécurité sociale pour 1999*.

[126] Helmut LÜTKEPOHL. *Introduction to multiple time series analysis*. Springer Science & Business Media, 2013.

[127] Helmut LÜTKEPOHL. *New introduction to multiple time series analysis*. Berlin Heidelberg : Springer-Verlag GmbH, 2005. ISBN : 978-3-540-27752-1.

[128] Vinh-Trung LUU et al. « A review of alignment based similarity measures for web usage mining ». In : *Artificial Intelligence Review* 53.3 (2020), p. 1529-1551.

[129] Laurent MAGY. « Épidémiologie de la sclérose en plaques ». In : *Rev. prat.* (22 fév. 2022), 72(4).

[130] Hossein MAHJUB et al. « Performance evaluation of support vector regression models for survival analysis : a simulation study ». In : *International Journal of Advanced Computer Science and Applications* 7.6 (2016).

[131] Assurance MALADIE. *Insuffisance cardiaque : caractéristiques, traitements et devenir à deux ans après une première hospitalisation*. URL : 2013-11_insuffisance-cardiaque-parcours-soins_assurance-maladie.pdf.

[132] Greg S MARTIN et al. « The epidemiology of sepsis in the United States from 1979 through 2000 ». In : *New England Journal of Medicine* 348.16 (2003), p. 1546-1554.

[133] Torben MARTINUSSEN et Thomas H. SCHEIKE. *Dynamic regression models for survival data*. Statistics for biology and health. New York : Springer, 2006. ISBN : 978-0-387-33960-3.

[134] Torben MARTINUSSEN et Stijn VANSTEELANDT. « On collapsibility and confounding bias in Cox and Aalen regression models ». In : *Lifetime data analysis* 19.3 (2013), p. 279-296.

[135] Torben MARTINUSSEN et al. « Estimation of direct effects for survival data by using the Aalen additive hazards model ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 73.5 (2011), p. 773-788.

[136] Florent MASSEGLIA. « Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel ». 2002.

[137] James C MCDAVID, Irene HUSE et Laura RL HAWTHORN. *Program evaluation and performance measurement : An introduction to practice*. 2018.

[138] Theresa A MCDONAGH et al. « 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure : Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC ». In : *European heart journal* 42.36 (2021), p. 3599-3726.

[139] Guillaume MOULIS et al. « Epidemiology of incident immune thrombocytopenia : a nationwide population-based study in France ». In : *Blood, The Journal of the American Society of Hematology* 124.22 (2014), p. 3308-3315.

[140] Saul B NEEDLEMAN et Christian D WUNSCH. « A general method applicable to the search for similarities in the amino acid sequence of two proteins ». In : *Journal of molecular biology* 48.3 (1970), p. 443-453.

[141] Roman NEVZOROV et al. « Developing a risk score to predict mortality in the first year after implantable cardioverter defibrillator implantation : Data from the Israeli ICD Registry ». In : *Journal of Cardiovascular Electrophysiology* 29.11 (2018), p. 1540-1547.

[142] Jerzy NEYMAN. *ur les applications de la théorie des probabilités aux expériences agricoles : Essai des principes, réédité en anglais dans la revue*. Statistical Science, Vol. 5, pp. 463-472.

[143] Ayaka NOZAKI et al. « The prognostic impact of gender in patients with acute heart failure–An evaluation of the age of female patients with severely decompensated acute heart failure ». In : *Journal of cardiology* 70.3 (2017), p. 255-262.

[144] Priti K. PATEL et Smit THACKER. « A review of Sequential Pattern Mining ». In : *IJCRT* 5.4 (2017), p. 1969-1974.

[145] Judea PEARL. « Causal diagrams for empirical research ». In : *Biometrika* 82.4 (1995), p. 669-688.

[146] Judea PEARL. *Causality*. Cambridge university press. 2009.

[147] Jessica PINAIRE et al. « Infarctus du myocarde : quelles sont les trajectoires de soins pronostiques du décès à l'hôpital ? » In : (2017), p. 14-25.

[148] Jessica PINAIRE et al. « Patient healthcare trajectory. An essential monitoring tool : a systematic review ». In : *Health information science and systems* 5.1 (2017), p. 1-18.

[149] Romain PIRRACCHIO et al. « Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates ». In : *Statistical methods in medical research* 25.5 (2016), p. 1938-1954.

[150] Piotr PONIKOWSKI et al. « ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure : The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC ». In : *European Heart Journal* 37.27 (14 juill. 2016), p. 2129-2200. ISSN : 0195-668X, 1522-9645. DOI : 10.1093/eurheartj/ehw128. URL : https://academic.oup.com/eurheartj/article-lookup/doi/10.1093/eurheartj/ehw128 (visité le 06/06/2022).

[151] Catherine QUANTIN et al. « Chaînage de bases de données anonymisées pour les études épidémiologiques multicentriques nationales et internationales : proposition d'un algorithme cryptographique ». In : *Revue d'épidémiologie et de santé publique* 57.1 (2009), p. 33-39.

[152] Faisal RAHUTOMO, Teruaki KITASUKA et Masayoshi ARITSUGI. « Semantic cosine similarity ». In : *The 7th international student conference on advanced science and technology ICAST*. T. 4. 1. 2012, p. 1.

[153] Ahsan RAO et al. « Regional variations in trajectories of long-term readmission rates among patients in England with heart failure ». In : *BMC Cardiovascular Disorders* 19.1 (2019), p. 1-11.

[154] Greg RIDGEWAY. « The state of boosting ». In : *Computing science and statistics* (1999), p. 172-181.

[155] James M ROBINS. « Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect" ». In : *Computers & Mathematics with Applications* 14.9-12 (1987), p. 923-945.

[156] James M ROBINS, Miguel Angel HERNAN et Babette BRUMBACK. « Marginal structural models and causal inference in epidemiology ». In : *Epidemiology* 11.5 (2000), p. 550-560.

[157] James M ROBINS et Miguel A HERNÁN. « Estimation of the causal effects of time-varying exposures ». In : *Longitudinal data analysis* 553 (2009), p. 599.

[158] James M ROBINS et Andrea ROTNITZKY. « Recovery of information and adjustment for dependent censoring using surrogate markers ». In : *AIDS epidemiology*. Springer, 1992, p. 297-331.

[159] Stéphanie ROGGERONE et al. « Actualités thérapeutiques dans la sclérose en plaques ». In : *La Revue du praticien (Paris)* 62.8 (2012), p. 1057-1060.

[160] Paul R ROSENBAUM et Donald B RUBIN. « The central role of the propensity score in observational studies for causal effects ». In : *Biometrika* 70.1 (1983), p. 41-55.

[161] Donald B RUBIN. « Estimating causal effects of treatments in randomized and nonrandomized studies. » In : *Journal of educational Psychology* 66.5 (1974), p. 688.

[162] Donald B RUBIN et Neal THOMAS. « Matching using estimated propensity scores : relating theory to practice ». In : *Biometrics* (1996), p. 249-264.

[163] Thomas H SCHEIKE et Mei-Jie ZHANG. « Analyzing competing risk data using the R timereg package ». In : *Journal of statistical software* 38.2 (2011).

[164] Guido SCHWARZER, Werner VACH et Martin SCHUMACHER. « On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology ». In : *Statistics in medicine* 19.4 (2000), p. 541-561.

[165] Soko SETOGUCHI, Lynne Warner STEVENSON et Sebastian SCHNEEWEISS. « Repeated hospitalizations predict mortality in the community population with heart failure ». In : *American heart journal* 154.2 (2007), p. 260-266.

[166] Kevin S SHAH et al. « Heart failure with preserved, borderline, and reduced ejection fraction : 5-year outcomes ». In : *Journal of the American College of Cardiology* 70.20 (2017), p. 2476-2486.

[167] Jincheng SHEN, Lu WANG et Jeremy MG TAYLOR. « Estimation of the optimal regime in treatment of prostate cancer recurrence from observational data using flexible weighting models ». In : *Biometrics* 73.2 (2017), p. 635-645.

[168] Pannagadatta K SHIVASWAMY, Wei CHU et Martin JANSCHE. « A support vector approach to censored targets ». In : *Seventh IEEE international conference on data mining (ICDM 2007)*. IEEE. 2007, p. 655-660.

[169] Rodrigo B SILVA et Wagner BARRETO-SOUZA. « Flexible and robust mixed Poisson INGARCH models ». In : *Journal of Time Series Analysis* 40.5 (2019), p. 788-814.

[170] Joanne SIMPSON et al. « Prognostic models derived in PARADIGM-HF and validated in ATMOSPHERE and the Swedish Heart Failure Registry to predict mortality and morbidity in chronic heart failure ». In : *JAMA cardiology* 5.4 (2020), p. 432-441.

[171] Temple F SMITH et Michael S WATERMAN. « Identification of common molecular subsequences ». In : *Journal of molecular biology* 147.1 (1981), p. 195-197.

[172] Hélène SOL. *Big Data en santé : données concernées, usages, entrepôt bio-hétérogènes et outils d'exploitation*. l'anap (agence nationale de la performance sanitaire et médico-sociale. 2016. URL : https://ressources.anap.fr/numerique/publication/1505.

[173] Scott D SOLOMON et al. « Influence of nonfatal hospitalization for heart failure on subsequent mortality in patients with chronic heart failure ». In : *Circulation* 116.13 (2007), p. 1482-1487.

[174] Raphael Edward Benjamin SONABEND. « A theoretical and methodological framework for machine learning in survival analysis : Enabling transparent and accessible predictive modelling on right-censored time-to-event data ». UCL (University College London), 2021.

[175] Ramakrishnan SRIKANT et Rakesh AGRAWAL. « Mining sequential patterns : Generalizations and performance improvements ». In : *International conference on extending database technology*. Springer. 1996, p. 1-17.

[176] Ramakrishnan SRIKANT et Rakesh AGRAWAL. « Mining sequential patterns : Generalizations and performance improvements ». In : *International conference on extending database technology*. 1996, p. 1-17.

[177] Jonathan AC STERNE et al. « Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death : a prospective cohort study ». In : *The Lancet* 366.9483 (2005), p. 378-384.

[178] Susanne STROHMAIER et al. « Dynamic path analysis–a useful tool to investigate mediation processes in clinical survival trials ». In : *Statistics in medicine* 34.29 (2015), p. 3866-3887.

[179] Jeremy MG TAYLOR et al. « Comparison of methods for estimating the effect of salvage therapy in prostate cancer when treatment is given by indication ». In : *Statistics in medicine* 33.2 (2014), p. 257-274.

[180] Eric J Tchetgen TCHETGEN et al. « Instrumental variable estimation in a survival context ». In : *Epidemiology (Cambridge, Mass.)* 26.3 (2015), p. 402.

[181] Timo TERÄSVIRTA. « Specification, estimation, and evaluation of smooth transition autoregressive models ». In : *Journal of the american Statistical association* 89.425 (1994), p. 208-218.

[182] Terry THERNEAU, Cindy CROWSON et Elizabeth ATKINSON. « Using time dependent covariates and time dependent coefficients in the cox model ». In : *Survival Vignettes* 2.3 (2017), p. 1-25.

[183] Terry THERNEAU, Cindy CROWSON et Elizabeth ATKINSON. « Using time dependent covariates and time dependent coefficients in the cox model ». In : *Survival Vignettes* 2.3 (2017), p. 1-25.

[184] Laine E THOMAS et al. « Matching with time-dependent treatments : A review and look forward ». In : *Statistics in medicine* 39.17 (2020), p. 2350-2370.

[185] *Tous les projets, Health Data Hub*. `https://www.health-data-hub.fr/projets`.

[186] Philippe TUPPIN et al. « Care pathways and healthcare use of stroke survivors six months after admission to an acute-care hospital in France in 2012 ». In : *Revue neurologique* 172.4-5 (2016), p. 295-306.

[187] Philippe TUPPIN et al. « Two-year outcome of patients after a first hospitalization for heart failure : A national observational study ». In : *Archives of Cardiovascular Diseases* 107.3 (2014), p. 158-168.

[188] Philippe TUPPIN et al. « Value of a national administrative database to guide public decisions : From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France ». In : *Revue d'epidemiologie et de sante publique* 65 (2017), S149-S167.

[189] Vanya VAN BELLE et al. « Additive survival least-squares support vector machines ». In : *Statistics in Medicine* 29.2 (2010), p. 296-308.

[190] Vanya VAN BELLE et al. « Support vector methods for survival analysis : a comparison between ranking and regression approaches ». In : *Artificial intelligence in medicine* 53.2 (2011), p. 107-118.

[191] Stijn VANSTEELANDT et Marshall JOFFE. « Structural nested models and G-estimation : the partially realized promise ». In : *Statistical Science* 29.4 (2014), p. 707-731.

[192] Stijn VANSTEELANDT, T MARTINUSSEN et EJ Tchetgen TCHETGEN. « On adjustment for auxiliary covariates in additive hazard models for the analysis of randomized experiments ». In : *Biometrika* 101.1 (2014), p. 237-244.

[193] Cédric VILLANI. *Donner un sens à l'intelligence artificielle, pour une stratégie nationale et européenne*. 2018, p. 194-203.

[194] J-L VINCENT et al. « The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure ». In : (1996).

[195] Adriaan A VOORS et al. « Development and validation of multivariable models to predict mortality and hospitalization in patients with heart failure ». In : *European journal of heart failure* 19.5 (2017), p. 627-634.

[196] Sandra VUKUSIC et Christian CONFAVREUX. « Natural history of multiple sclerosis : risk factors and prognostic indicators ». In : *Current opinion in neurology* 20.3 (2007), p. 269-274.

[197] Robert A WAGNER et Michael J FISCHER. « The string-to-string correction problem ». In : *Journal of the ACM (JACM)* 21.1 (1974), p. 168-173.

[198] Aolin WANG, Roch A NIANOGO et Onyebuchi A ARAH. « G-computation of average treatment effects on the treated and the untreated ». In : *BMC medical research methodology* 17.1 (2017), p. 1-5.

[199] Ping WANG, Yan LI et Chandan K REDDY. « Machine learning for survival analysis : A survey ». In : *ACM Computing Surveys (CSUR)* 51.6 (2019), p. 1-36.

[200] Zhe WANG et al. « Mortality prediction system for heart failure with orthogonal relief and dynamic radius means ». In : *International Journal of medical informatics* 115 (2018), p. 10-17.

[201] Jan WASKOWSKI et al. « Effects of sodium bicarbonate infusion on mortality in medical–surgical ICU patients with metabolic acidosis—A single-center propensity score matched analysis ». In : *Medicina intensiva* (2021).

[202] Lee-Jen WEI. « The accelerated failure time model : a useful alternative to the Cox regression model in survival analysis ». In : *Statistics in medicine* 11.14-15 (1992), p. 1871-1879.

[203] Waloddi WEIBULL. « A statistical distribution function of wide applicability ». In : *Journal of applied mechanics* (1951).

[204] Alain WEILL, Jérémy RUDANT et Joël COSTE. « Utilisation des données de l'assurance maladie française pour étudier l'usage et les effets des médicaments en vie réelle : revue de 216 articles publiés entre 2007 et 2016 ». In : *Revue d'Épidémiologie et de Santé Publique* 65 (2017), S127.

[205] Christian H WEISS. *An introduction to discrete-valued time series*. John Wiley & Sons, 2018. Chap. 4.3, p. 93-94.

[206] Richard WILLIAMS et al. « Using string metrics to identify patient journeys through care pathways ». In : 2014 (2014), p. 1208.

[207] Dawn B WOODARD, David S MATTESON et Shane G HENDERSON. « Stationarity of generalized autoregressive moving average models ». In : *Electronic Journal of Statistics* 5 (2011), p. 800-828.

[208] Renxiang YAN et al. « A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction ». In : *Scientific reports* 3.1 (2013), p. 1-9.

[209] JE YOON et SY HWANG. « Zero-inflated INGARCH using conditional Poisson and negative binomial : data application ». In : *The Korean Journal of Applied Statistics* 28.3 (2015), p. 583-592.

[210] Mohammed Javeed ZAKI. « Scalable algorithms for association mining ». In : *IEEE transactions on knowledge and data engineering* 12.3 (2000), p. 372-390.

[211] Fukang ZHU. « A negative binomial integer-valued GARCH model ». In : *Journal of Time Series Analysis* 32.1 (2011), p. 54-67.

[212] Fukang ZHU. « Modeling time series of counts with COM-Poisson INGARCH models ». In : *Mathematical and Computer Modelling* 56.9-10 (2012), p. 191-203.

[213] Massimo ZONI-BERISSO et al. « Mortality after cardioverter-defibrillator replacement : Results of the DECODE survival score index ». In : *Heart Rhythm* 18.3 (2021), p. 411-418.

Chapitre 7

Annexe : Catégorie majeure de diagnostic

CMD	Libellé
01	Affections du système nerveux
02	Affections de l'œil
03	Affections des oreilles, du nez, de la gorge, de la bouche et des dents
04	Affections de l'appareil respiratoire
05	Affections de l'appareil circulatoire
06	Affections du tube digestif
07	Affections du système hépatobiliaire et du pancréas
08	Affections et traumatismes de l'appareil musculosquelettique et du tissu conjonctif
09	Affections de la peau, des tissus sous-cutanés et des seins
10	Affections endocriniennes, métaboliques et nutritionnelles
11	Affections du rein et des voies urinaires
12	Affections de l'appareil génital masculin
13	Affections de l'appareil génital féminin
14	Grossesses pathologiques, accouchements et affections du post-partum
15	Nouveau-nés, prématurés et affections de la période périnatale
16	Affections du sang et des organes hématopoïétiques
17	Affections myéloprolifératives et tumeurs de siège imprécis ou diffus et/ou CMA
18	Maladies infectieuses et parasitaires
19	Maladies et troubles mentaux
20	Troubles mentaux organiques liés à l'absorption de drogues ou induits par celles-ci
21	Traumatismes, allergies et empoisonnements
22	Brûlures
23	Facteurs influant sur l'état de santé et autres motifs de recours aux services de santé
24	Séjours de moins de 2 jours
25	Maladies dues à une infection par le VIH
26	Traumatismes multiples graves
27	Transplantations d'organes
28	Séances
90	Erreurs et autres séjours inclassables

Chapitre 8

Annexe relative au chapitre 3

8.1 - Computations detailed

8.1.1 . Expectation of $r_n(A)$

$$\begin{aligned}\mathbb{E}[r_n(A)] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)W_i(t)^\top A(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)dN_i(t)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)W_i(t)^\top A(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)\mathbb{E}[dN_i(t)]\end{aligned}$$

Using additive structure of the intensity of counting process N_i : $\mathbb{E}[dN_i(t)] = \mu_i^*(t)dt = W_i(t)^\top A^*(t)dt$, we obtain

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)W_i(t)^\top A(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)W_i(t)^\top A^*(t)dt \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} (A(t)-A^*(t))^\top W_i(t)W_i(t)^\top (A(t)-A^*(t))dt \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A^*(t)^\top W_i(t)W_i(t)^\top A^*(t)dt \\
&= ||A-A^*||_n^2 - \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A^*(t)^\top W_i(t)W_i(t)^\top A^*(t).
\end{aligned}
$$

8.1.2 . $r_n(A)$ in function of R_n

$$
\begin{aligned}
r_n(A) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)W_i(t)^\top A(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top W_i(t)dN_i(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top [\widetilde{W}_i(t)-\xi_i(t)][\widetilde{W}_i(t)-\xi_i(t)]^\top A(t)dt \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top [\widetilde{W}_i(t)-\xi_i(t)]dN_i(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \widetilde{W}_i(t)\widetilde{W}_i(t)^\top A(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \widetilde{W}_i(t)dN_i(t) \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \xi_i(t)\widetilde{W}_i(t)^\top A(t)dt + \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \xi_i(t)\xi_i(t)^\top A(t)dt \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \xi_i(t)dN_i(t) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \widetilde{W}_i(t)\widetilde{W}_i(t)^\top A(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^{\tau_i} A(t)^\top \widetilde{W}_i(t)dN_i(t) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} \alpha_X(t)^\top \epsilon_i(t)\epsilon_i(t)^\top \alpha_X(t)dt \\
&= \widetilde{R}_n(A) + \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau_i} \alpha_X(t)^\top \epsilon_i(t)\epsilon_i(t)^\top \alpha_X(t)dt. \qquad (8.1)
\end{aligned}
$$

8.2 - Simulation algorithm

The following pseudo-algorithm present simulation process. For a patient i, initial values are defined as follow :

- At the initiation :
  - $D_i(0) = 0$;
  - $X_i(0) \sim \mathcal{U}([min_X, max_X])$ with min_X and max_X a k-dimensional vector ;
  - $Z_i^1 \sim \mathcal{U}([min_{Z_1}, max_{Z_1}])$, $Z_i^2 \sim \mathcal{B}er(p_{Z_2})$ and $Z_i^2 \sim \mathcal{P}(\lambda_{Z_3})$
- For $t = 0, ..., 10$
  - $N_i(]t, t+1]) \sim$ Non homogeneous Poisson process with intensity function $\delta D(t) + \delta_0 + \sum_{j=1}^{d_Z} \delta_{Zj} Z^j + \sum_{j=1}^{d} \delta_{Xj} X^j(t)$
    - If $D_i(t) = 0$:
      - $X_i(t+1) = \kappa_{D0} X_i(t) + \epsilon$, with $\epsilon \sim \mathcal{N}(0_{\mathbb{R}^{d_X}}, \Sigma)$
      - $X_i^0(t+1) = X_i(t+1)$
      - $D_i(t) \sim \mathcal{B}er(m \sum_{j=1}^{d_X} \lambda_j e^{\sum_{j=1}^{d_X} \lambda_j X_i^j(t)})$
    - If $D_i(t) = 1$:
      - $X_i(t+1) = X_i(t) + \kappa_{D1}(\sqrt{1000} - X_i(t)) + \epsilon$, with $\epsilon \sim \mathcal{N}(0_{\mathbb{R}^{d_X}}, \Sigma)$
      - $X_i^0(t+1) = \kappa_{D0} X_i(t) + \epsilon$, with $\epsilon \sim \mathcal{N}(0_{\mathbb{R}^{d_X}}, \Sigma)$
      - $D_i(t+1) = 1$

The following parameters were used for the simulations presented in the paper whatever the number of time-varying covariates studied : $min_{Z_1} = -20$,$max_{Z_1} = -10$, $p_{Z_2} = 0.5$, $\lambda_{Z_3} = 0.1$, $\delta = -0.01$, $\delta_0 = 30$, $\delta_Z = (0.1, 0.02, 0.01)$, $m = 1.5$. Table 8.1 presents parameter values depending of the number of simulated time-varying covariates.

Parameters	1 covariate	3 covariates	6 covariates
min_X	0	$(0, 0, 0)$	$(0, 0, 0, 0, 0, 0)$
max_X	10	$(10, 20, 30)$	$(10, 20, 30, 10, 20, 30)$
κ_{D0}	-0.25	$(-0.25, -0.25, -0.25)$	$(-0.25, -0.25, -0.25, -0.25, -0.25, -0.25)$
κ_{D1}	0.25	$(0.25, 0.25, 0.25)$	$(0.25, 0.25, 0.25, 0.25, 0.25, 0.25)$
δ_X	-0.25	$(-0.3, 0, -0.25)$	$(-0.3, -0.2, 0, 0, -0.2, -0.25)$
λ	0.12	$(0.16, 0.14, 0)$	$(0.13, 0.12, 0.13, 0.14, 0, 0)$

TABLE 8.1 – Parameter values depending of the number of simulated time-varying covariates

Chapitre 9

Annexe relative au chapitre 4

9.1 - Computations detailed

9.1.1 . Properties

Properties :

1. $\mathbb{V}(\omega_i(t_k) = \text{diag}(\mathbb{E}[X_i(t_k)])$
2. $Cov(\omega_i(t_k), \omega_i(t_l)) = 0_{d_X}$ for all $k \neq l$ and l, $k \geq 0$

Proof :

i) Let i, $j \in [\![1, d]\!]$, $i \neq j$ and $t \geq 0$,

$$\begin{aligned}\mathbb{E}[\omega_i(t_k)\omega_j(t_k) \mid F_{t_{k-1}}] &= \mathbb{E}[(X_i(t_k) - \lambda_i(t_k))(X_j(t_k) - \lambda_j(t_k)) \mid F_{t_{k-1}}] \\ &= \mathbb{E}[X_i(t_k)X_j(t_k) - X_i(t_k)\lambda_j(t_k) - X_j(t_k)\lambda_i(t_k) + \lambda_i(t_k)\lambda_j(t_k)) \mid F_{t_{k-1}}] \\ &= \mathbb{E}[X_i(t_k)X_j(t_k) \mid F_{t_{k-1}}] - \mathbb{E}[X_i(t_k)\lambda_j(t_k) \mid F_{t_{k-1}}] \\ &\quad - \mathbb{E}[X_j(t_k)\lambda_i(t_k) \mid F_{t_{k-1}}] + \mathbb{E}[\lambda_i(t_k)\lambda_j(t_k) \mid F_{t_{k-1}}].\end{aligned}$$

As $X_i(t_k) \perp\!\!\!\perp X_j(t_k) \mid F_{t_{k-1}}$ and $\lambda(t_k)$ is $F_{t_{k-1}}$-measurable :

$$\begin{aligned}\mathbb{E}[\omega_i(t_k)\omega_j(t_k) \mid F_{t_{k-1}}] &= \lambda_i(t_k)\lambda_j(t_k) - \lambda_j(t_k)\mathbb{E}[X_i(t_k) \mid F_{t_{k-1}}] - \lambda_i(t_k)\mathbb{E}[X_j(t_k) \mid F_{t_{k-1}}] + \lambda_i(t_k)\lambda_j(t_k) \\ &= 2\lambda^i(t_k)\lambda^j(t_k) - 2\lambda^i(t_k)\lambda^j(t_k) = 0.\end{aligned}$$

Then, $\mathbb{V}(\omega_t)$ is a diagonal matrix, and :

$$\mathbb{E}[\omega_i(t_k)^2 \mid F_{t_{k-1}}] = \mathbb{E}[X_i(t_k)^2 \mid F_{t_{k-1}}] - 2\mathbb{E}[X_i(t_k)\lambda_i(t_k) \mid F_{t_{k-1}}] + \mathbb{E}[\lambda_i(t_k)^2) \mid F_{t_{k-1}}]$$

and

$$\begin{aligned}\mathbb{E}[\lambda_i(t_k)^2 \mid F_{t_{k-1}}] &= \mathbb{E}[\mathbb{E}[X_i(t_k) \mid F_{t_{k-1}}]\lambda_i(t_k) \mid F_{t_{k-1}}] = \mathbb{E}[\mathbb{E}[X_i(t_k)\lambda_i(t_k) \mid F_{t_{k-1}}] \mid F_{t_{k-1}}] \\ &= \mathbb{E}[X_i(t_k)\lambda_i(t_k) \mid F_{t_{k-1}}].\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}[\omega_i(t_k)^2 \mid F_{t_{k-1}}] &= \mathbb{E}[X_i(t_k)^2 \mid F_{t_{k-1}}] - \mathbb{E}[X_i(t_k)\lambda_i(t_k) \mid F_{t_{k-1}}] \\ &= \mathbb{V}(X_i(t_k) \mid F_{t_{k-1}}) = \mathbb{E}[X_i(t_k) \mid F_{t_{k-1}}] = \lambda_i(t_k)\end{aligned}$$

as $X_i(t_k) \mid F_{t_{k-1}}$ follows a Poisson distribution.
Finally,

$$\mathbb{E}[\omega_i(t_k)^2] = \mathbb{E}[\mathbb{E}[\omega^i(t_k)^2 \mid F_{t_{k-1}}]] = \mathbb{E}[\lambda_i(t_k)] = \mathbb{E}[X_i(t_k)].$$

Hence,

$$\mathbb{V}(\omega_i(t_k)) = \mathsf{diag}(\mathbb{E}[X_i(t_k)].$$

ii) Let i,j $\in \{1, \ldots, d\}$, and $t_1, t_2 \geq 0$, $t_1 > t_2$, we can write :

$$\begin{aligned}\mathbb{E}[\omega_i(t_k)\omega_j(t_l) \mid F_{t_{k-1}}] &= \mathbb{E}[X_i(t_k)X_j(t_l) \mid F_{t_1-1}] - \mathbb{E}[X_i(t_k)\lambda_j(t_l) \mid F_{t_1-1}] \\ &\quad - \mathbb{E}[\lambda_i(t_k)X_j(t_l) \mid F_{t_1-1}] + \mathbb{E}[\lambda_i(t_k)\lambda_j(t_l)) \mid F_{t_1-1}] \\ &= X_j(t_l)\mathbb{E}[X_i(t_k) \mid F_{t_1-1}] - \lambda_j(t_l)\mathbb{E}[X_i(t_k) \mid F_{t_1}] \\ &\quad - X_j(t_l)\mathbb{E}[\lambda_i(t_k) \mid F_{t_1}] + \lambda_i(t_k)\lambda_j(t_l) \\ &= X_j(t_l)\lambda_i(t_k) - \lambda_j(t_l)\lambda_i(t_k) - X_j(t_l)\lambda_i(t_k) + \lambda_i(t_k)\lambda_j(t_l) = 0.\end{aligned}$$

Hence

$$Cov(\omega_i(t_k), \omega_j(t_l))_{ij} = \mathbb{E}[\omega_i(t_k)\omega_j(t_l)] = 0$$

9.1.2 . Estimation using non-negative least square model

This section proposed a method to estimate parameters of the model VINGARCH(p,0). The last time without treatment for the patient i is noted $t^i_{\max}$. They correspond to the discrete observe time before S_i. To simplify calculations, we assume that β_0^{*i} is a function of baseline covariates. It will be expressed as follows $\beta_0^{*i} = \beta_0^* + \sum_{j=1}^{d_Z} A_j^* Z_i^j$ with $\beta_0^* \in \mathbb{R}_+^{d_X}$ and for $j \in [\![1, d_Z]\!]$, $A_j^* = (a_j^{*1}, \cdots, a_j^{*d_X})^\top \in \mathbb{R}_+^{d_X}$.
With these notations, equation (4.4) can be rewritten for each individuals and observed times as follow :

$$\begin{pmatrix} X_1(1) \\ \vdots \\ X_1(t^1_{\max}) \\ \vdots \\ X_n(1) \\ \vdots \\ X_n(t^n_{\max}) \end{pmatrix} = R \begin{pmatrix} \beta_0^* \\ A_1^* \\ \vdots \\ A_{d_Z}^* \\ b_{11}^{*1} \\ \vdots \\ b_{d_X d_X}^{*1} \\ b_{11}^{*p} \\ \vdots \\ b_{d_X d_X}^{*p} \end{pmatrix} + \begin{pmatrix} \omega_1(1) \\ \vdots \\ \omega_1(t^1_{\max}) \\ \vdots \\ \omega_n(1) \\ \vdots \\ \omega_n(t^n_{\max}) \end{pmatrix}, \tag{9.1}$$

with b_{kl}^{*j} element of the B_j^* matrix and R the following matrix of size $\sum_{j=1}^{n} t^j_{\max} \times d_X(1 + pd_X + d_Z)$:

$$R = \begin{pmatrix} Id_{d_X} & Z_1^1 Id_{d_X} & \dots & Z_1^{d_Z} Id_{d_X} & \mathbb{X}_1(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Id_{d_X} & Z_1^1 Id_{d_X} & \dots & Z_1^{d_Z} Id_{d_X} & \mathbb{X}_1(t^1_{\max}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Id_{d_X} & Z_n^1 Id_{d_X} & \dots & Z_n^{d_Z} Id_{d_X} & \mathbb{X}_n(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Id_{d_X} & Z_n^1 Id_{d_X} & \dots & Z_n^{d_Z} Id_{d_X} & \mathbb{X}_n(t^n_{\max}) \end{pmatrix}, \tag{9.2}$$

and

$$\mathbb{X}_i(t_k) = \begin{pmatrix} X_i(t_k)^\top & 0_{d_X} & \dots & 0_{d_X} \\ 0_{d_X} & X_i(t_k)^\top & \dots & 0_{d_X} \\ \dots & \dots & \ddots & \dots \\ 0_{d_X} & \dots & 0_{d_X} & X_i(t_k)^\top \end{pmatrix}, \tag{9.3}$$

a $d_X \times pd_X$ matrix and 0_{d_X} a zero vector of size d_X.

These notations allow to write the non-negative least square model as followed :

$$\min_{\beta \in \mathbb{R}_+^{d_X(qd_X+d_Z+1)}} ||X - R\boldsymbol{\beta}||, \tag{9.4}$$

with $X = (X_1(1), \cdots, X_1(t^1_{\max}), \cdots, X_n(1), \cdots, X_n(t^n_{\max}))^\top$ and
$\boldsymbol{\beta} = (\beta^*_0, A^*_1, \cdots, A^*_{d_Z}, b^{*1}_{11}, \cdots, b^{*1}_{d_X d_X}, b^{*p}_{11}, \cdots, b^{*p}_{d_X d_X})^\top$.

9.2 - Simulation algorithm

The following pseudo-algorithm present simulation process. For a patient i, initial values are defined as follow :

- At the initiation :
  - $D_i(0) = 0$;
  - $X_i(0) \sim \mathcal{P}(\lambda)$ with λ a d_k-dimensional vector ;
  - $Z_i^1 \sim \mathcal{U}([min_{Z_1}, max_{Z_1}])$, $Z_i^2 \sim \mathcal{B}er(p_{Z_2})$ and $Z_i^2 \sim \mathcal{P}(\lambda_{Z_3})$
- For $t_k = 0, ..., 10$
  - $N_i(]t_k, t_{k+1}]) \sim$ Non homogeneous Poisson process with intensity function $\delta D(t_k) + \delta_0 + \sum_{j=1}^{d_Z} \delta_{Z_j} Z^j + \sum_{j=1}^{d} \delta_{Xj} X_j(t_k)$
    - If $D_i(t_k) = 0$:
      - $X_i(t_{k+1}) \sim \mathcal{P}(\kappa_{D0} + K_0 X_i(t_k))$
      - $X_i^0(t_{k+1}) = X_i(t_{k+1})$
      - $D_i(t_k) \sim \mathcal{B}er(m \sum_{j=1}^{d_X} \lambda_j e^{\sum_{j=1}^{d_X} \lambda_j X_i^j(t_k)})$
    - If $D_i(t)_k = 1$:
      - $X_i(t_{k+1}) \sim \mathcal{P}(\kappa_{D1} + K_1 X_i(t_k))$
      - $X_i^0(t_{k+1}) \sim \mathcal{P}(\kappa_{D0} + K_0 X_i(t_k))$
      - $D_i(t_{k+1}) = 1$

Chapitre 10

Annexe relative au chapitre 5

TABLE 10.1 – Population selection criteria

ICD-10 diagnosis code (primary or associated diagnosis)	
I50	Heart failure
I110	Hypertensive heart disease with (congestive) heart failure
I130	Hypertensive heart and renal disease with (congestive) heart failure
J81	Pulmonary oedema
GHM code	
05M09	Heart failure and circulatory shock

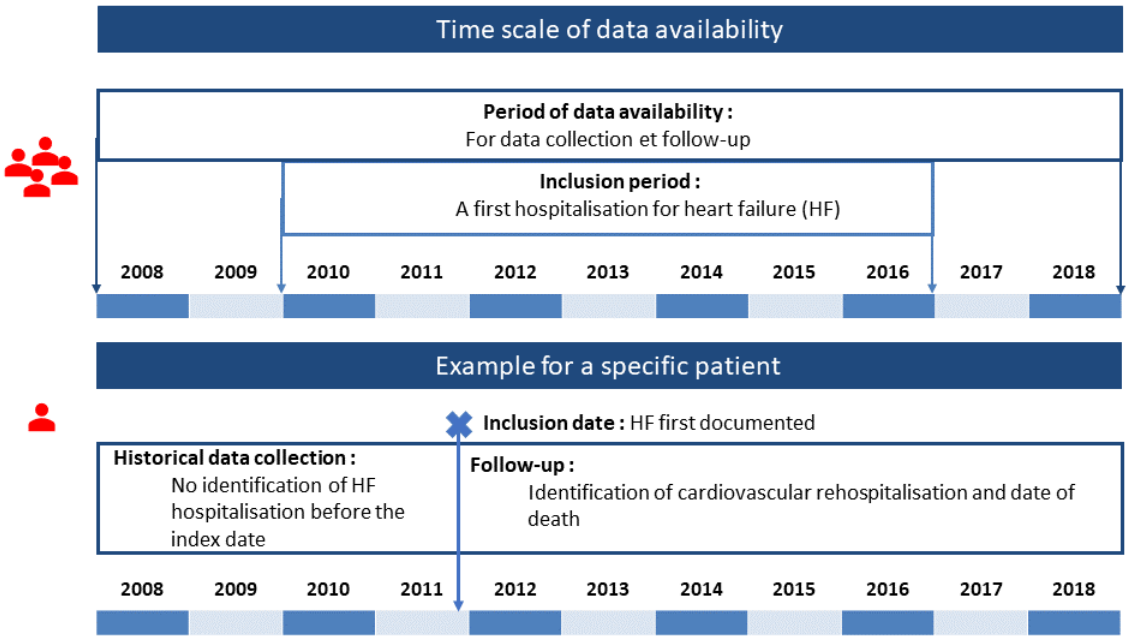

FIGURE 10.1 – Study design : step of statistical analyses on the populations concerned

TABLE 10.2 – Homogeneous Patient Groups (GHM)

Root of the selected GHM	Detail of the selected GHM	Wording og the selected GHM
	05C02	Valve replacement surgery with extracorporeal circulation and with cardiac catheterization or coronary angiography
	05C03	Valve replacement surgery with extracorporeal circulation, without cardiac catheterization or coronary angiography
	05C04	Coronary artery bypass grafts with cardiac catheterization or coronary angiography

Root of the selected GHM	Detail of the selected GHM	Wording og the selected GHM
	05C05	Coronary artery bypass grafts without cardiac catheterization or coronary angiography
	05C06	Other cardiothoracic procedures, age over 1 year, or vascular procedures regardless of age, with extracorporeal circulation
	05C07	Other cardiothoracic procedures, age less than 2 years, with extracorporeal circulation
	05C08	Other cardiothoracic procedures, age over 1 year, or vascular procedures regardless of age, without extracorporeal circulation
	05C09	Other cardiothoracic procedures, age less than 2 years, without extracorporeal circulation
	05C10	Major revascularization surgery
	05C11	Other vascular surgery procedures
	05C12	Amputations of the lower limb, except toes, for circulatory disorders
	05C13	Amputations for circulatory disorders of the upper limb or toes
	05C14	Permanent pacemaker placement with acute myocardial infarction or congestive heart failure or shock
	05C15	Permanent pacemaker placements without acute myocardial infarction, congestive heart failure, or shock
	05C17	Vein ligations and awakenings
	05C18	Other interventions on the circulatory system
	05C19	Placement of a cardiac defibrillator
	05C20	Replacement or surgical removal of electrodes or repositioning of permanent pacemaker boxes
	05C21	Creation and repair of arteriovenous fistulas for conditions of major category of diagnosis “Circulatory system disorders”

Root of the selected GHM	Detail of the selected GHM	Wording og the selected GHM
	05C22	Permanent pacemaker replacements
05K	05K05	Vascular stents with myocardial infarction
	05K06	Vascular stents without myocardial infarction
	05K10	Vascular diagnostic procedures
	05K12	Vascular therapeutic procedures except stents, age less than 18 years
	05K13	Implementation of certain vascular accesses for major category of diagnosis "Circulatory system disorders", except stent
	05K14	Implementation of certain vascular accesses in major category of diagnosis "Circulatory system disorders", stays of less than 2 days
	05K15	Monitoring of heart transplants with vascular diagnostic procedure
	05K17	Cardiovascular conditions without surgery in major category of diagnosis "Circulatory system disorders", with anesthesia
	05K19	Major treatments for arrhythmias by vascular route
	05K20	Other treatments for arrhythmias by vascular route
	05K21	Cardiac valve bioprosthesis installation by vascular route
	05K22	Therapeutic acts by vascular route on the openings of the heart, age over 17 years
	05K23	Removal, repositioning and placement of additional cardiac probes by vascular route, age over 17 years
	05K24	Coronary dilations and other therapeutic acts on the heart by vascular route, age over 17 years
	05K25	Therapeutic acts on the arteries by vascular route, age over 17 years
	05K26	Therapeutic acts on vascular access or veins by vascular route, age over 17 years
	05M04	Acute myocardial infarction

Root of the selected GHM	Detail of the selected GHM	Wording og the selected GHM
05M	05M05	Syncope and fainting
	05M06	Angina pectoris
	05M07	Deep vein thrombophlebitis
	05M08	Arrhythmias and cardiac conduction disturbances
	05M09	Heart failure and circulatory shock
	05M10	Congenital heart disease and valve disease, age less than 18 years
	05M11	Congenital heart disease and valve disease, age over 17 years
	05M12	Peripheral vascular disorders
	05M13	Chest pain
	05M14	Cardiac arrest
	05M15	Arterial hypertension
	05M16	Coronary atherosclerosis
	05M17	Other disorders of the circulatory system
	05M18	Acute and subacute endocarditis
	05M19	Monitoring of heart transplants without diagnostic procedure by vascular route
	05M20	Investigations and monitoring for affections of the circulatory system
	05M23	Symptoms and other health care uses of CMD 05*
04M13	04M13	Pulmonary edema and respiratory distress

TABLE 10.3 – Permutation importance of features associated with a good prognosis

GHM trajectories	Trajectories	Importance feature
05M13	Chest pain	0.00620 $\pm$ 0.00017

GHM trajectories	Trajectories	Importance feature
05C15	Permanent pacemaker placements without acute myocardial infarction, congestive heart failure, or shock	0.00467 ± 0.00011
05M05	Syncope and fainting	0.00330 ± 0.00011
05C22	Permanent pacemaker replacements	0.00234 ± 9.11e-05
05C19	Placement of a cardiac defibrillator	0.00209 ± 9,63e-05
05K21	Cardiac valve bioprosthesis installation by vascular route	0.00137 ± 7.31e-05
05M12_05K06	Peripheral vascular disorders FOLLOWED BY Vascular stents without myocardial infarction	0,00124 ± 4.92e-05
05C14	Permanent pacemaker placement with acute myocardial infarction or congestive heart failure or shock	0,00102 ± 5.53e-05
05K19	Major treatments for arrhythmias by vascular route	0,00095 ± 9.73e-05
05M20	Investigations and monitoring for affections of the circulatory system	0,00091 ± 5.24e-05
05K17	Cardiovascular conditions without surgery in major category of diagnosis "Circulatory system disorders", with anesthesia	0,00089 ± 4.81e-05
05M15	Arterial hypertension	0,00079 ± 4.14e-05
05C03	Valve replacement surgery with extracorporeal circulation, without cardiac catheterization or coronary angiography	0,00078 ± 5.44e-05
05K20	Other treatments for arrhythmias by vascular route	0,00077 ± 5.43e-05
05C05	Coronary artery bypass grafts without cardiac catheterization or coronary angiography	0,00061 ± 6.76e-05
05M16	Coronary atherosclerosis	0,00058 ± 3.56e-05
05M08_05M08	Two arrhythmias and cardiac conduction disturbances	0,00047 ± 3.64e-05
04M13_05M09_05M09	Pulmonary oedema and respiratory distress FOLLOWED BY Two heart failure and circulatory shock	0,00036 ± 1.67e-05
05M23	Symptoms and other health care uses of major category of diagnosis "Circulatory system disorders"	0,00023 ± 2.33e-05

GHM trajectories	Trajectories	Importance feature
05M06	Angina pectoris	0,00020 ± 1.53e-05
05K10_05K10	Two vascular diagnostic procedures	0,00010 ± 2.15e-05

TABLE 10.4 – Permutation importance of features associated with a bad prognosis

GHM trajectories	Trajectories	Importance feature
05K10_05M09	Vascular diagnostic procedures FOLLOWED BY Heart failure and circulatory shock	0.00070 ± 9.12e-5
05K14	Implementation of certain vascular accesses in major category of diagnosis "Circulatory system disorders", stays of less than 2 days	0.00055 ± 8.09e-5
05M13_05M09	Chest pain FOLLOWED BY heart failure and circulatory shock	0.00016 ± 3.74e-5
05M16_05M09	Coronary atherosclerosis FOLLOWED BY heart failure and circulatory shock	0.00014 ± 3.88e-5
05M09_05M09_05M08	Two heart failure and circulatory shock FOLLOWED BY Arrhythmias and cardiac conduction disturbances	0.00011 ± 4.51e-5
05M09_05C19	Heart failure and circulatory shock FOLLOWED BY Placement of a cardiac defibrillator	0.00016 ± 1.69e-5
05C19_05M09	Placement of a cardiac defibrillator FOLLOWED BY heart failure and circulatory shock	1.12e-5 ± 1.52e-9
05M09_05K10	Heart failure and circulatory shock FOLLOWED BY Vascular diagnostic procedures	-1.84e-05 ± 2.96e-5

GHM trajectories	Trajectories	Importance feature
05M09_05C15	Heart failure and circulatory shock FOLLOWED BY Permanent pacemaker placements without acute myocardial infarction,congestive heart failure, or shock	-2.55e-05 ± 1.45e-5
05M09_05M09_05K10	Two heart failure and circulatory shock FOLLOWED BY Vascular diagnostic procedures	-3.97e-05 ± 1.84e-5
05K13	Implementation of certain vascular accesses for major category of diagnosis "Circulatory system disorders" conditions	-5.05e-05 ± 1.94e-5
05M09_05M15	Heart failure and circulatory shock FOLLOWED BY Arterial hypertension	-6.30e-05 ± 1.26e-5
05M09_05M12	Heart failure and circulatory shock FOLLOWED BY Peripheral vascular disorders	-9.52e-05 ± 2.06e-5
04M13_04M13	Two pulmonary edema and respiratory distress	-0.00011 ± 8.87e-5
05M12_05M09	Peripheral vascular disorders FOLLOWED BY Heart failure and circulatory shock	-0.00012 ± 2.63e-5

TABLE 10.5 – Features non interpreted (no effect, n= 43, or contradictory effects between repetitions, n = 10)

Features	Importance
05K06_05K10	Vascular stents without myocardial infarction FOLLOWED BY Vascular diagnostic procedures
05M09_05M09_05K06	Two heart failure and circulatory shock FOLLOWED BY Vascular stents without myocardial infarction
05K10_05K06	Vascular diagnostic procedures FOLLOWED BY Vascular stents without myocardial infarction

Features	Importance
05K10_05M08	Vascular diagnostic procedures FOLLOWED BY Arrhythmias and cardiac conduction disturbances
05M08_05K10	Arrhythmias and cardiac conduction disturbances FOLLOWED BY Vascular diagnostic procedures
05K06_05M09	Vascular stents without myocardial infarction FOLLOWED BY Heart failure and circulatory shock
05M09_05K06_05M09	Heart failure and circulatory shock FOLLOWED BY Vascular stents without myocardial infarction FOLLOWED BY Heart failure and circulatory shock
05M09	Heart failure and circulatory shock
05M09_05M09	Two Heart failure and circulatory shock
05K10	Vascular diagnostic procedures
05K06	Vascular stents without myocardial infarction
05M08	Arrhythmias and cardiac conduction disturbances
05M09_05M09_05M09	Three Heart failure and circulatory shock
05M17	Other disorders of the circulatory system
05M12	Peripheral vascular disorders
05M09_05M09_05M09_05M09	Four Heart failure and circulatory shock
05M04	Acute myocardial infarction
05M09_05M08	Heart failure and circulatory shock FOLLOWED BY Arrhythmias and cardiac conduction disturbances
05M08_05M09	Arrhythmias and cardiac conduction disturbances FOLLOWED BY Heart failure and circulatory shock
05C10	Major revascularization surgery
04M13_05M09	Pulmonary edema and respiratory distress FOLLOWED BY Heart failure and circulatory shock

Features	Importance
05M09_05M09_05M09 _05M09_05M09	Five Heart failure and circulatory shock
05K06_05K06	Two Vascular stents without myocardial infarction
05M09_04M13	Heart failure and circulatory shock FOLLOWED BY Pulmonary edema and respiratory distress
05K05	Vascular stents with myocardial infarction
05K10_05M09_05M09	Vascular diagnostic procedures FOLLOWED BY Two Heart failure and circulatory shock
05M11	Congenital heart disease and valve disease, age less than 18 years
05M09_05K10_05M09	Heart failure and circulatory shock FOLLOWED BY Vascular diagnostic procedures FOLLOWED BY Heart failure and circulatory shock
05K25	Therapeutic acts on the arteries by vascular route, age over 17 years
05C11	Other vascular surgery procedures
05M17_05M09	Other disorders of the circulatory system FOLLOWED BY Heart failure and circulatory shock
05M09_05M17	Heart failure and circulatory shock FOLLOWED BY Other disorders of the circulatory system
05K26	Therapeutic acts on vascular access or veins by vascular route, age over 17 years
05C21	Creation and repair of arteriovenous fistulas for conditions of CMD 05*
05M09_05M09_05M09 _05M09_05M09_05M09	Six Heart failure and circulatory shock
05C12	Amputations of the lower limb, except toes, for circulatory disorders

Features	Importance
05C15_05M09	Permanent pacemaker placements without acute myocardial infarction, congestive heart failure, or shock FOLLOWED BY Heart failure and circulatory shock
05M09_05M13	Heart failure and circulatory shock FOLLOWED BY chest pain
05M15_05M09	Arterial hypertension FOLLOWED BY Heart failure and circulatory shock
05M09_05M09_04M13	Two Heart failure and circulatory shock FOLLOWED BY Pulmonary edema and respiratory distress
05M09_05M04	Heart failure and circulatory shock FOLLOWED BY Infarctus aigu du myocarde
05M09_05M08_05M09	Heart failure and circulatory shock FOLLOWED BY Heart failure and circulatory shock
05M04_05M09	Infarctus aigu du myocarde FOLLOWED BY Heart failure and circulatory shock
05K10_05C19	Vascular diagnostic procedures FOLLOWED BY Placement of a cardiac defibrillator
05M08_05M09_05M09	Arrhythmias and cardiac conduction disturbances FOLLOWED BY Two Heart failure and circulatory shock
05M17_05K10	Other disorders of the circulatory system FOLLOWED BY Vascular diagnostic procedures
05K24	Coronary dilations and other therapeutic acts on the heart by vascular route, age over 17 years
05C19_05K10	Placement of a cardiac defibrillator FOLLOWED BY Vascular diagnostic procedures

Features	Importance
05K10_05M09_05M09_05M09	Vascular diagnostic procedures FOLLOWED BY Three Heart failure and circulatory shock
05M09_05K10_05M09_05M09	Heart failure and circulatory shock FOLLOWED BY Vascular diagnostic procedures FOLLOWED BY Two Heart failure and circulatory shock
05M09_05K06	Heart failure and circulatory shock FOLLOWED BY Vascular stents without myocardial infarction
05K06_05M09_05M09	Vascular stents without myocardial infarction FOLLOWED BY Two Heart failure and circulatory shock
04M13	Pulmonary edema and respiratory distress

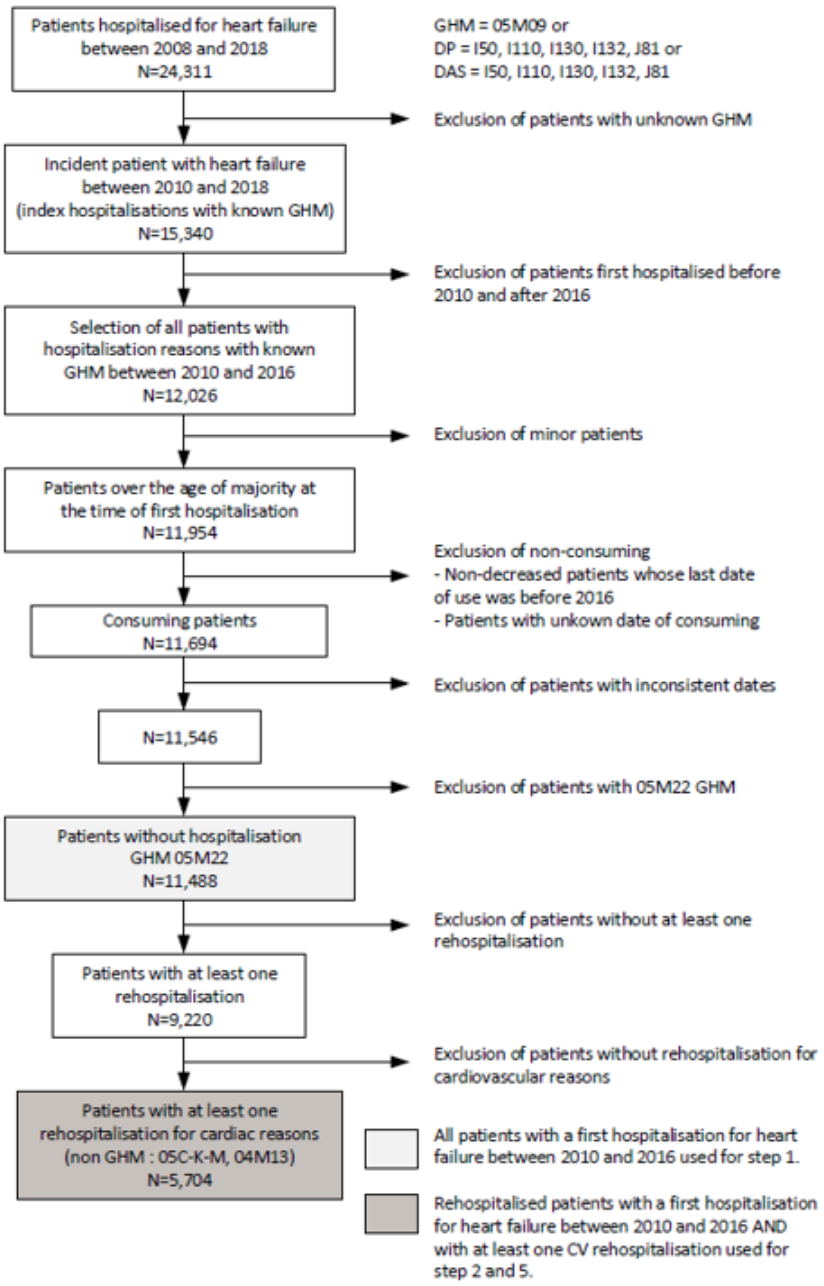

FIGURE 10.2 – FlowChart