Traitements des réhospitalisations par des méthodes statistiques prenant en compte les évènements récurrents

Anaïs Charles-Nelson

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité : Biostatistiques

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS : EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Présentée par :

Anaïs CHARLES-NELSON

Pour obtenir le grade de

DOCTEUR DE SORBONNE UNIVERSITÉ

Traitements des réhospitalisations par des méthodes statistiques prenant en compte les évènements récurrents.

---

qui sera présentée le 24 Juin 2020 devant le jury composé de :

Bruno Falissard	Rapporteur
Agathe Guiloux	Rapporteur
Yassin Mazroui	Examinateur
Audrey Lavenu	Examinateur
Sandrine Katsahian	Directrice de thèse

Sorbonne Université
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à MON : 01 42 34 68 41
Tél. pour les étudiants de MOO à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

Remerciements

Je ne pouvais finir cette thèse sans remercier toutes les personnes qui, de près ou de loin, ont contribué à ce travail, à commencer par ma directrice de thèse. Je remercie Sandrine KATSAHIAN pour m'avoir très tôt transmis sa passion pour la recherche, grâce aux différents projets que nous avons menés ensemble. Je lui suis très reconnaissante pour sa confiance, sa disponibilité et ses précieux conseils qui m'ont fait progresser tout au long de ces années de thèse et qui m'ont permis de la réaliser dans des conditions idéales.

Je tiens à remercier le Professeur Bruno FALISSARD et le Professeur Agathe GUILLOUX qui m'ont fait l'honneur d'accepter d'être les rapporteurs de ma thèse. Je remercie également le Docteur Audrey LAVENU et le Docteur Yassin MAZROUI pour avoir accepté de participer à mon jury de thèse.

Je remercie Anita Burgun pour m'avoir accueillie dans son équipe. Je remercie tous mes collègues actuels et anciens de l'équipe de l'unité de recherche clinique de l'hôpital européen Georges-Pompidou et de l'Unité 1138 pour toutes les intéractions que nous avons pu avoir. Et je remercie tous ceux qui ont relu ma thèse pour leurs corrections.

Je remercie Catherine Schramm. Tu m'as énormément aidée et soutenue tout au long de cette thèse. Sans toi, je n'y serais pas arrivée.

Je remercie ma famille et mes amis pour leur patience, leur soutien et la confiance qu'ils m'ont accordés. Je tiens tout particulièrement à remercier mon frère, ma soeur, ma mère et ma belle-soeur pour m'avoir toujours encouragée et soutenue dans les bons et les moins bons moments. Enfin, mes derniers remerciements, mais non des moindres, sont pour celui qui partage ma vie.

Résumé/Abstract

Résumé

Dans de nombreuses études, les réhospitalisations sont utilisées comme marqueurs de l'efficacité d'un traitement, comme facteurs pronostiques de l'évolution d'une maladie ou encore comme marqueurs de la qualité des soins. Les hospitalisations peuvent survenir plusieurs fois pour un même patient au cours du temps et font partie de la classe des événements appelés événements récurrents. L'étude des réhospitalisations se fait via des modèles statistiques dédiés aux événements récurrents. Cependant, ce type d'étude soulève des problématiques statistiques majeures telles que la prise en compte de la dépendance intra-patient, de l'hétérogénéité et des événements terminaux stoppant de façon permanente le processus de récurrences.
Au-delà du nombre de réhospitalisations et ses facteurs associés, les raisons des réhospitalisations peuvent nous apporter des informations sur les soins prodigués aux patients et ainsi nous permettent de construire des trajectoires de soins. Analyser ces trajectoires permet d'étudier et de mettre en évidence les combinaisons de raisons de réhospitalisations les plus fréquentes. Ceci pourrait aider les médecins à pratiquer une médecine personnalisée en anticipant les soins futurs, ce qui permettrait potentiellement d'éviter des complications sévères.
Cette thèse se compose de deux parties. Nous avons dans un premier temps comparé les différentes méthodes existantes permettant d'analyser les événements récurrents en présence d'un événement terminal. Nous avons étudié les propriétés de chaque modèle sur des données simulées afin de comprendre le comportement des modèles en fonction du nombre de sujets, de la force de la dépendance entre les événements récurrents et la force de la dépendance entre les événements récurrents et l'événement terminal. Cette étude de simulation nous a permis de produire des lignes directrices afin d'aider le statisticien dans le choix de la méthode la plus adaptée à la question posée. Nous avons ensuite développé un modèle permettant d'analyser les événements récurrents en présence de deux événements terminaux. Ce modèle est un modèle joint à fragilités partagées. Il est composé d'un modèle pour les événements récurrents et d'un modèle pour chaque événement terminal. Deux termes de fragilités sont inclus dans le modèle afin de prendre en compte la dépendance entre les événements récurrents et chaque événement terminal. Les paramètres de régression sont estimés par la maximisation de la vraisemblance pénalisée. Les méthodes de vérification de l'adéquation du modèle sont décrites. Les différentes méthodes ont ainsi été appliquées à nos données sur les réhospitalisations après une greffe de foie.
Dans une deuxième partie, nous nous sommes intéresses aux trajectoires de soins des patients. Les trajectoires de soins ont été créées à partir des diagnostics de réhospitalisations dans l'année suivant la chirurgie bariatrique. Les trajectoires ont ensuite été analysées par l'analyse formelle de concept afin de mettre en évidence les combinaisons de diagnostics les plus fréquentes.

Mots clefs : événements récurrents ; événement terminal ; modèle joint ; réhospitalisation ; analyse formelle de concept ; trajectoire de soin

Abstract

In many studies, readmissions are used as indicator of the effectiveness of a treatment, as prognostic factors for the evolution of a disease or as indicator of the quality of care. Readmissions can occur several times for the same patient over time and are part of the class of events called recurrent events. The study of readmissions is done through statistical models dedicated to recurrent events. However, this type of study raises major statistical issues such as the consideration of intra-patient dependence, heterogeneity in the data and terminal events permanently stopping the recurrence process.

Beyond the number of readmissions and its associated factors, the reasons for readmissions can provide information on the care provided to patients and thus allows us to construct trajectories of care. Analyzing these trajectories allows to highlight the most frequent combinations of reasons for readmission. This could help physicians practicing personalized medicine by anticipating future care, potentially avoiding severe complications.

This thesis consists in two parts. We first compare the different existing methods for analyzing recurrent events in the presence of a terminal event. We studied the properties of each model on simulated data in order to understand the performance of the models according to the number of subjects, the strength of the dependence between the recurrent events and the strength of the dependence between the recurrent events and the terminal event. This simulation study allowed us to produce guidelines to help the statistician in choosing the most appropriate method to the question of interest. In a second step, we developed a model to analyze recurrent events in the presence of two terminal events. This model is a shared frailty joint model. It combines a model for recurrent events and a model for each terminal event. Two frailties are included in the model in order to take into account the dependency between the recurrent events and each terminal event. The regression parameters are estimated by maximizing the penalized likelihood. Methods for checking the adequacy of the model are described. The different methods were applied to our data on readmissions after a liver transplant.

In a second part, we became interested in patient care trajectories. The trajectories of care were created from the diagnoses of readmissions in the year following bariatric surgery. The trajectories were then analyzed by the formal concept analysis to highlight the most common diagnostic combinations.

Keywords : recurrent events ; terminal event ; joint model ; readmission ; formal concept analysis ; care trajectories

Table des matières

Table des figures

Liste des tableaux

Production scientifique

Papiers en lien avec la thèse

Papiers acceptés

- **Charles-Nelson A**, Katsahian S, Schramm C. How to analyze and interpret recurrent events data in the presence of a terminal event : An application on readmission after colorectal cancer surgery. *Statistics in Medicine*. 2019 Aug 15. (Chapitre 3)

- **Charles-Nelson A**, Lazzati A, Katsahian S. Analysis of Trajectories of Care After Bariatric Surgery Using Data Mining Method and Health Administrative Information Systems. *Obesity Surgery*. 2020 Feb 6. (Chapitre 6)

Papiers en perspective

- **Charles-Nelson A**, Schramm C, Katsahian S. Analysing recurrent events stopped by terminal event in presence of competing risk : use of joint frailty model. *En préparation*. (Chapitre 4)

Papiers acceptés hors thèse

- Meyer G, Besse B, Doubre H, **Charles-Nelson A**, Aquilanti S, Izadifar A, Azarian R, Monnet I, Lamour C, Descourt R, Oliviero G, Taillade L, Chouaid C, Giraud F, Falcoz PE, Revel MP, Westeel V, Dixmier A, Tredaniel J, Dehette S, Decroisette C, Prevost A, Pichon E, Fabre E, Soria JC, Friard S, Stern JB, Jabot L, Dennewald G, Pavy G, Petitpretz P, Tourani JM, Alifano M, Chatellier G, Girard P. Anti-tumour effect of low molecular weight heparin in localised lung cancer : a phase III clinical trial. Eur Respir J. 2018 Oct 4 ;52(4). pii : 1801220. doi : 10.1183/13993003.01220-2018. Print 2018 Oct.

PubMed PMID : 30262574.

- Dépret F, Aubry A, Fournier A, **Charles-Nelson A**, Katsahian S, Compain F, Mainardi JL, Fernandez-Gerlinger MP. β LACTA testing may not improve treatment decisions made with MALDI-TOF MS-informed antimicrobial stewardship advice for patients with Gram-negative bacteraemia : a prospective comparative study. J Med Microbiol. 2018 Feb ;67(2) :183-189. doi : 10.1099/jmm.0.000665. Epub 2017 Dec 21. PubMed PMID : 29265997.

- Roux M, Pigneur F, Baranes L, Calderaro J, Chiaradia M, Decaens T, Kastahian S, **Charles-Nelson A**, Tselikas L, Costentin C, Laurent A, Azoulay D, Mallat A, Rahmouni A, Luciani A. Differentiating focal nodular hyperplasia from hepatocellular adenoma : Is hepatobiliary phase MRI (HBP-MRI) using linear gadolinium chelates always useful ? Abdom Radiol (NY). 2018 Jul ;43(7) :1670-1681. doi : 10.1007/s00261-017-1377-z. Erratum in : Abdom Radiol (NY). 2017 Dec 11 ; :. PubMed PMID : 29110059.

- Aoun Sebaiti M, Kauv P, **Charles-Nelson A**, Van Der Gucht A, Blanc-Durand P, Itti E, Gherardi RK, Bachoud-Levi AC, Authier FJ. Cognitive dysfunction associated with aluminum hydroxide-induced macrophagic myofasciitis : A reappraisal of neuropsychological profile. J Inorg Biochem. 2018 Apr ;181 :132-138. doi : 10.1016/j.jinorgbio.2017.09.019. Epub 2017 Oct 6. PubMed PMID : 29079320.

- Mas JL, Derumeaux G, Guillon B, Massardier E, Hosseini H, Mechtouff L, Arquizan C, Béjot Y, Vuillier F, Detante O, Guidoux C, Canaple S, Vaduva C, Dequatre-Ponchelle N, Sibon I, Garnier P, Ferrier A, Timsit S, Robinet-Borgomano E, Sablot D, Lacour JC, Zuber M, Favrole P, Pinel JF, Apoil M, Reiner P, Lefebvre C, Guérin P, Piot C, Rossi R, Dubois-Randé JL, Eicher JC, Meneveau N, Lusson JR, Bertrand B, Schleich JM, Godart F, Thambo JB, Leborgne L, Michel P, Pierard L, Turc G, Barthelet M, **Charles-Nelson A**, Weimar C, Moulin T, Juliard JM, Chatellier G ; CLOSE Investigators. Patent Foramen Ovale Closure or Anticoagulation vs. Antiplatelets after Stroke. N Engl J Med. 2017 Sep 14 ;377(11) :1011-1021. doi : 10.1056/NEJMoa1705915. PubMed PMID : 28902593.

- Cholley B, Caruba T, Grosjean S, Amour J, Ouattara A, Villacorta J, Miguet B, Guinet P, Lévy F, Squara P, Aït Hamou N, Carillion A, Boyer J, Boughenou MF, Rosier S, Robin E, Radutoiu M, Durand M, Guidon C, Desebbe O,**Charles-Nelson A**, Menasché P, Rozec B, Girard C, Fellahi JL, Pirracchio R, Chatellier G ; -. Effect of Levosimendan on Low Cardiac Output Syndrome in Patients With Low Ejection Fraction Undergoing Coronary Artery Bypass Grafting With Cardiopulmonary Bypass : The LICORN Randomized Clinical Trial. JAMA. 2017 Aug 8 ;318(6) :548-556. doi : 10.1001/jama.2017.9973. PubMed PMID : 28787507 ; PubMed Central PMCID : PMC5817482.

- Iliou MC, Vergès-Patois B, Pavy B, **Charles-Nelson A**, Monpère C, Richard R, Verdier JC ; on behalf for the CREMS-HF (Cardiac REhabilitation and electrical MyoStimulation-Heart Failure) study group. Effects of combined exercise training and electromyostimulation treatments in chronic heart failure : A prospective multicentre study. Eur J Prev Cardiol. 2017 Aug ;24(12) :1274-1282. doi : 10.1177/2047487317712601. Epub 2017 Jun 1. PubMed PMID : 28569553.

- Commereuc M, Guérot E, **Charles-Nelson A**, Constan A, Katsahian S, Schortgen F. ICU Patients Requiring Renal Replacement Therapy Initiation : Fewer Survivors and More Dialysis Dependents From 80 Years Old. Crit Care Med. 2017 Aug ;45(8) :e772-e781. doi : 10.1097/CCM.0000000000002407. PubMed PMID : 28437374.

- Auclin E, **Charles-Nelson A**, Abbar B, Guérot E, Oudard S, Hauw-Berlemont C, Thibault C, Monnier A, Diehl JL, Katsahian S, Fagon JY, Taieb J, Aissaoui N. Outcomes in elderly patients admitted to the intensive care unit with solid tumors. Ann Intensive Care. 2017 Dec ;7(1) :26. doi : 10.1186/s13613-017-0250-0. Epub 2017 Mar 6. PubMed PMID : 28265980 ; PubMed Central PMCID : PMC5339259.

- Nos C, Clough KB, Bonnier P, Lasry S, Le Bouedec G, Flipo B, Classe JM, Missana MC, Doridot V, Giard S, Charitansky H, **Charles-Nelson A**, Bats AS, Ngo C. Upper outer boundaries of the axillary dissection. Result of the SENTIBRAS protocol : Multicentric protocol using axillary reverse mapping in breast cancer patients requiring axillary dissection. Eur J Surg Oncol. 2016 Dec ;42(12) :1827-1833. doi : 10.1016/j.ejso.2016.07.138. Epub 2016 Aug 26. PubMed PMID : 27769634.

- Eymard F, **Charles-Nelson A**, Katsahian S, Chevalier X, Bercovy M. Predictive Factors of "Forgotten Knee" Acquisition After Total Knee Arthroplasty : Long-Term Follow-Up of a Large Prospective Cohort. J Arthroplasty. 2017 Feb ;32(2) :413-418.e1. doi : 10.1016/j.arth.2016.06.020. Epub 2016 Jun 23. PubMed PMID : 27430181.

- Maitre B, Djibre M, Katsahian S, Habibi A, Stankovic Stojanovic K, Khellaf M, Bourgeon I, Lionnet F, **Charles-Nelson A**, Brochard L, Lemaire F, Galacteros F, Brun-Buisson C, Fartoukh M, Mekontso Dessap A. Inhaled nitric oxide for acute chest syndrome in adult sickle cell patients : a randomized controlled study. Intensive Care Med. 2015 Dec ;41(12) :2121-9. doi : 10.1007/s00134-015-4060-2. Epub 2015 Oct 2. PubMed PMID : 26431718.

- Schortgen F, **Charles-Nelson A**, Bouadma L, Bizouard G, Brochard L, Katsahian S. Respective impact of lowering body temperature and heart rate on mortality in septic shock : mediation analysis of a randomized trial. Intensive Care Med. 2015 Oct ;41(10) :1800-8. doi : 10.1007/s00134-015-3987-7. Epub 2015 Jul 23. PubMed PMID : 26202042.

- Dessap AM, Roche-Campo F, Launay JM, **Charles-Nelson A**, Katsahian S, Brun-Buisson C, Brochard L. Delirium and Circadian Rhythm of Melatonin During Weaning From Mechanical Ventilation : An Ancillary Study of a Weaning Trial. Chest. 2015 Nov ;148(5) :1231-1241. doi : 10.1378/chest.15-0525. PubMed PMID : 26158245.

- Desgranges P, Kobeiter H, Katsahian S, Bouffi M, Gouny P, Favre JP, Alsac JM, Sobocinski J, Julia P, Alimi Y, Steinmetz E, Haulon S, Alric P, Canaud L, Castier Y, Jean-Baptiste E, Hassen-Khodja R, Lermusiaux P, Feugier P, Destrieux-Garnier L, **Charles-Nelson A**, Marzelle J, Majewski M, Bourmaud A, Becquemin JP ; ECAR Investigators. Editor's Choice - ECAR (Endovasculaire ou Chirurgie dans les Anévrysmes aorto-iliaques Rompus) : A French Randomized Controlled Trial of Endovascular Versus Open Surgical Repair of Ruptured Aorto-iliac Aneurysms. Eur J Vasc Endovasc Surg. 2015 Sep ;50(3) :303-10. doi : 10.1016/j.ejvs.2015.03.028. Epub 2015 May 20. PubMed PMID : 26001320.

- Eymard F,**Charles-Nelson A**, Katsahian S, Chevalier X, Bercovy M. "Forgotten knee" after total knee replacement : A pragmatic study from a single-centre cohort. Joint Bone Spine. 2015 May ;82(3) :177-81. doi : 10.1016/j.jbspin.2014.11.006. Epub 2015 Jan 23. PubMed PMID : 25623519.

- Mekontso Dessap A, Contou D, Dandine-Roulland C, Hemery F, Habibi A, **Charles-Nelson A**, Galacteros F, Brun-Buisson C, Maitre B, Katsahian S. Environmental influences on daily emergency admissions in sickle-cell disease patients. Medicine (Baltimore). 2014 Dec ;93(29) :e280. doi : 10.1097/MD.0000000000000280. PubMed PMID : 25546672 ; PubMed Central PMCID : PMC4602624.

- Chiaradia M, Baranes L, Van Nhieu JT, Vignaud A, Laurent A, Decaens T, **Charles-Nelson A**, Brugières P, Katsahian S, Djabbari M, Deux JF, Sobhani I, Karoui M, Rahmouni A, Luciani A. Intravoxel incoherent motion (IVIM) MR imaging of colorectal liver metastases : are we only looking at tumor necrosis ? J Magn Reson Imaging. 2014 Feb ;39(2) :317-25. doi : 10.1002/jmri.24172. Epub 2013 May 30. PubMed PMID : 23723012.

- Caruba T, Katsahian S, Schramm C, **Charles-Nelson A**, Durieux P, Bégué D, Juillière Y, Dubourg O, Danchin N, Sabatier B. Treatment for stable coronary artery disease : a network meta-analysis of cost-effectiveness studies. PLoS One. 2014 Jun 4 ;9(6) :e98371. doi : 10.1371/journal.pone.0098371. eCollection 2014. PubMed PMID : 24896266 ; PubMed Central PMCID : PMC4045726.

- Khellaf M, **Charles-Nelson A**, Fain O, Terriou L, Viallard JF, Cheze S, Graveleau J, Slama B, Audia S, Ebbo M, Le Guenno G, Cliquennois M, Salles G, Bonmati C, Teillet F, Galicier L, Hot A, Lambotte O, Lefrère F, Sacko S, Kengue DK, Bierling P, Roudot-

Thoraval F, Michel M, Godeau B. Safety and efficacy of rituximab in adult immune thrombocytopenia : results from a prospective registry including 248 patients. Blood. 2014 Nov 20 ;124(22) :3228-36. doi : 10.1182/blood-2014-06-582346. Epub 2014 Oct 7. PubMed PMID : 25293768.

- Bridoux A, Drouot X, Sangare A, Al-Ani T, Brignol A, **Charles-Nelson A**, Brugières P, Gouello G, Hosomi K, Lepetit H, Palfi S. Bilateral thalamic stimulation induces insomnia in patients treated for intractable tremor. Sleep. 2015 Mar 1 ;38(3) :473-8. doi : 10.5665/sleep.4512. PubMed PMID : 25515098 ; PubMed Central PMCID : PMC4335528.

- Langlois J, **Charles-Nelson A**, Katsahian S, Beldame J, Lefebvre B, Bercovy M. Predictors of flexion using the rotating concave-convex total knee arthroplasty : preoperative range of motion is not the only determinant. Knee Surg Sports Traumatol Arthrosc. 2015 Jun ;23(6) :1734-40. doi : 10.1007/s00167-014-3479-2. Epub 2014 Dec 23. PubMed PMID : 25533698.

Conférences internationales

Posters

- ISCB 2013 (34th Annual Conference of the International Society for Clinical Biostatistics), Munich, Allemagne, 25-29 Août 2013, Analysing recurrent events stopped by several types of terminal events : use of frailty model .

Liste des abréviations

- AG : Andersen et Gill
- AGA : Anneau Gastrique Ajustable
- CAH : Classification Ascendante Hiérarchique
- CCAM : Classification Communes des Actes Médicaux
- CMS : Centres pour Medicare et Medicaid Services
- CNIL : Commission Nationale de l'Informatique et des Libertés
- DRG : Groupes Homogènes de Diagnostic
- DP : Diagnostic Principal
- EGB : Echantillon Généraliste des Bénéficiaires
- FCA : Analyse Formelle de Concept
- GB : Bypass Gastrique
- GEE : Equations d'Estimation Généralisées
- GHM : Groupes Homogènes de Malades
- HAS : Haute Autorité de Santé
- IMC : Indice de Masse Corporelle

- LWA : Lee, Wei et Amato
- log : logarithme népérien
- MR : Méthodologie de Référence
- PMSI : Programme de Médicalisation des Systèmes d'Information
- PWP : Prentice, William et Peterson
- SAHOS : Syndrome d'Apnées Hypopnées Obstructives du Sommeil
- SG : Sleeve Gastrectomie
- WLW : Wei, Lee et Weissfeld

Introduction générale : les enjeux statistiques de l'utilisation des réhospitalisations comme critère de jugement dans la recherche clinique ou indicateur de résultats en santé publique

En France, le développement des outils de mesure de la qualité et de la sécurité des soins fait partie des objectifs de la *stratégie nationale de santé 2018-2022* [1]. Dans de nombreuses études, les réhospitalisations sont utilisées comme marqueurs de l'efficacité d'un traitement, comme facteurs pronostiques de l'évolution d'une maladie ou encore comme marqueurs de la qualité des soins. Les hospitalisations peuvent survenir plusieurs fois pour un même patient au cours du temps et font partie de la classe des événements appelés événements récurrents. L'étude des réhospitalisations se fait via des modèles statistiques dédiés aux événements récurrents. Cependant, ce type d'étude soulève des problématiques statistiques majeures telles que la prise en compte de la dépendance intra-patient, de l'hétérogénéité et des événements terminaux stoppant de façon permanente le processus de récurrences.

Au-delà du nombre de réhospitalisations et ses facteurs associés, les raisons des réhospitalisations peuvent nous apporter des informations sur les profils des patients et ainsi nous permettent de construire des trajectoires de soins. Analyser ces trajectoires permet d'étudier et de mettre en évidence les combinaisons de raisons de réhospitalisations les plus fréquentes. Ceci permet de mieux planifier l'offre de soins et peut aider les médecins à pratiquer une médecine personnalisée en anticipant les soins futurs, et en diminuant le nombre de complications sévères.

Nous nous sommes intéressés, dans un premier temps, aux événements récurrents en présence de plusieurs types d'événements terminaux dans le but d'identifier des facteurs pronostiques liés aux réhospitalisations après une greffe de foie ainsi qu'au décès dû à la maladie du foie. Dans un deuxième temps, nous nous sommes intéressés aux trajectoires de soins dans la première année suivant une chirurgie bariatrique.

Les réhospitalisations

Les réhospitalisations sont définies comme la réadmission d'un patient à l'hôpital après son hospitalisation index (séjour où ont lieu les soins primaires et à partir duquel les hospitalisations suivantes seront considérées comme étant des réhospitalisations). Les délais de survenue des réhospitalisations ont permis de créer plusieurs indicateurs de vigilance et d'alerte. Le premier est le taux de réhospitalisations dans un délai de 1 à 7 jours (RH7) [2] et concerne les établissements de santé en médecine-chirurgie-obstétrique. Il a pour objectif de s'interroger sur les pratiques organisationnelles et cliniques des équipes. Cet indicateur est calculé à partir du programme de médicalisation des systèmes d'information (PMSI)(Annexe A), et est égal au "nombre de patients réhospitalisés dans un délai de 1 à 7 jours par rapport à celui des patients ayant eu une hospitalisation index (1er séjour de l'année) achevée au cours d'une année (hors transferts et cas particuliers de venues itératives, etc.)" [3]. Le second est le taux de réhospitalisations dans un délai de 1 à 30 jours (RH30) [4] et concerne, quant à lui la problématique de la coordination ville / hôpital. Contrairement au premier indicateur qui est calculé par établissement, celui-ci est calculé par zone géographique et est égal au "nombre de patients réhospitalisés sous 30 jours par rapport à l'ensemble des patients hospitalisés dans l'année en cours (hors transferts et séjours itératifs)" [3]. Un point commun entre ces deux indicateurs est qu'ils ne considèrent que les réhospitalisations non planifiées.

Les réhospitalisations non planifiées sont des réhospitalisations qui auraient pu être évitées si la prise en charge avait été appropriée au moment de l'hospitalisation index. Elles sont considérées comme des événements défavorables pour un patient car elles sont souvent dues par exemple à des complications d'une chirurgie ou encore à des événements cardiovasculaires causés par des traitements, mais elles sont aussi associées à des coûts financiers élevés [5]. En 2009, Les "Centers for Medicare and Medicaid Services" (CMS) ont commencé à rendre public les taux de réhospitalisations non planifiées, qui ont été ajoutés au site Web de comparaison des hôpitaux (https ://www.medicare.gov/hospitalcompare/search.html). Le CMS met en place le programme de réduction des réhospitalisations qui pénalise les hôpitaux lorsque leurs patients sont fréquemment réhospitalisés dans les 30 jours suivant leur hospitalisation index. En conséquence, les cliniciens, les responsables des soins de santé et les décideurs recherchent des moyens de réduire les réhospitalisations non planifiées.

Au cours de ces 60 dernières années, les chercheurs se sont de plus en plus intéressés aux réhospitalisations non planifiées, avec une croissance exponentielle du nombre d'articles publiés sur ce sujet (Figure 1).

Les réhospitalisations sont utilisées comme critère d'évaluation dans de nombreuses spécialités, notamment en cardiologie [6, 7], en politique de santé [8, 9] ou encore en chirurgie [10, 11] et sont alors au coeur de certaines méthodes informatiques et statistiques [12, 13, 14, 15].
Elles permettent de répondre à différents types de questions. Certaines études l'utilisent pour juger de l'efficacité d'un traitement [16] ou d'un dispositif médical [17]. D'autres l'utilisent pour

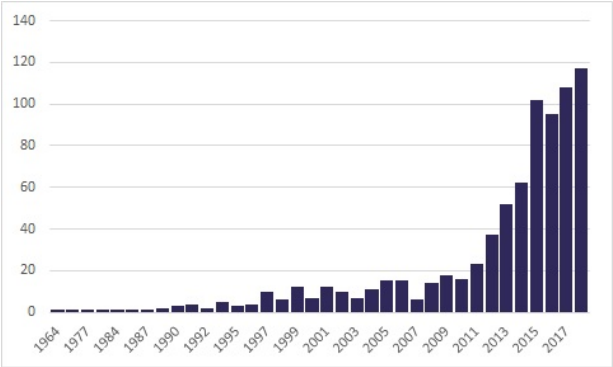


FIGURE 1 – Nombre d'articles publiés sur les réhospitalisations au cours du temps

Les articles sélectionnés sont ceux comportant les mots "hospital readmission" dans le titre. La recherche a été effectuée sur pubmed à la date du 04/04/2019

évaluer les coûts de prise en charge des patients [18, 19].

Un des avantages d'utiliser les réhospitalisations dans les entrepôts de données de santé et dans les bases médico-administratives est la facilité de trouver l'information [20]. En effet, les hôpitaux informatisent de plus en plus les dossiers de malades. En France, des programmes de standardisation des données pour le service public hospitalier et les établissements privés de soin sont mis en place [21], afin d'avoir des données plus fiables et plus complètes. Cependant, lorsque les réhospitalisations sont utilisées comme critère d'évaluation, certains problèmes peuvent se poser. En effet, il est important de pouvoir différencier les réhospitalisations planifiées des non planifiées, pour ne pas surestimer le taux de réhospitalisations. De plus, les patients peuvent être réhospitalisés dans d'autres hôpitaux que celui où ils sont suivis pouvant ainsi provoquer une sous-estimation du taux de réhospitalisations. De plus, les soins donnés après la sortie du patient peuvent influencer ses réhospitalisations [20]. Un autre problème à prendre en compte lors de l'étude des réhospitalisations, est la survenue d'événements en compétition, comme par exemple le décès [20]. En effet, les patients qui décèdent ne peuvent pas être réhospitalisés, et un patient qui décède précocément aura un nombre de réhospitalisations moins important qu'un patient qui vit plus longtemps. Si le décès n'est pas pris en compte, il est possible de conclure à tort à l'efficacité d'un traitement sur le taux de rehospitalisations.

Outre le fait d'avoir été ou non réhospitalisé, une autre information importante sur les réhospitalisations peut être analysée : les causes de réadmission. En effet, elles nous apportent des informations sur le parcours de soins des patients puisqu'ils peuvent être hospitalisés plusieurs fois mais pour des causes différentes (Figure 2). Ainsi identifier les différentes combinaisons, en particulier les combinaisons les plus fréquentes, et décrire les caractéristiques des individus les partageant, pourrait aider à améliorer leur prise en charge, anticiper les problèmes futurs, diminuer le nombre de réadmissions et par conséquent les coûts. Cependant, cette information est bien moins utilisée dans la littérature que le taux de réadmissions. Ceci est probablement dû au nombre élevé de causes de réhospitalisations et donc au nombre encore plus grand des

différentes combinaisons. En effet, ces connaissances sont noyées dans la quantité de données et nécessitent des méthodes spécifiques, adaptées aux données de grandes dimensions ("big data") et à la fouille de données ("data-mining").

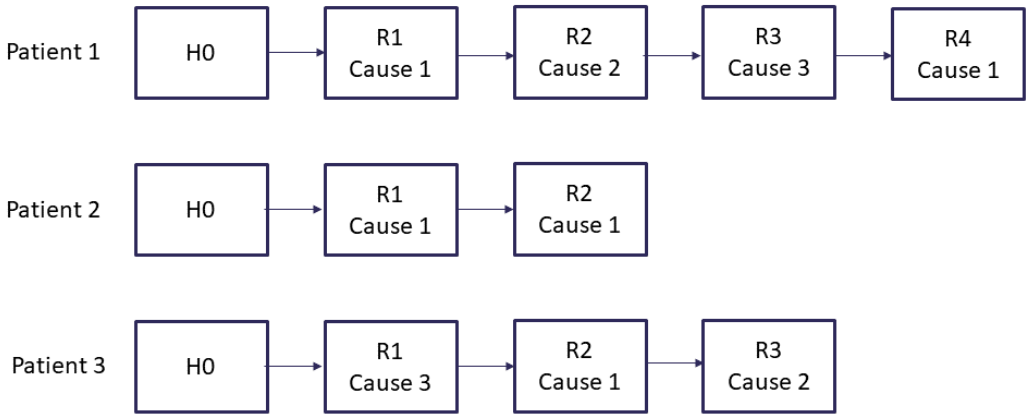


FIGURE 2 – Exemple de parcours de soins

La Figure 2 représente un exemple de construction des parcours de soins des patients où :
H0 représente l'hospitalisation index, ie à partir de laquelle les hospitalisations suivantes sont considérées comme réhospitalisations et prises en compte dans l'analyse ;
Rx sont les réhospitalisations après l'hospitalisation index ;
Cause x sont les raisons pour lesquelles le patient a été réhospitalisé.
On remarque que les patients 1 et 3 ont tous les deux été réhospitalisés pour les mêmes causes : les causes 1,2 et 3 mais aussi que les trois patients ont été réhospitalisés pour la cause 1.

Les problématiques statistiques liées à l'étude des réhospitalisations dans les essais cliniques et la recherche en santé

Les problématiques liées aux réhospitalisations ne seront pas les mêmes selon la question de recherche posée. En effet, si on cherche à évaluer l'association entre les réhospitalisations et les différentes covariables, des analyses statistiques particulières, prenant en compte la récurrence des événements, la dépendance des événements d'un même patient ainsi que d'un potentiel événement terminal (ex : le décés) dépendant des événements récurrents, vont être nécessaires. En revanche, si l'on s'intéresse aux parcours de soins des patients, d'autres problèmes entrent en jeu tels que l'homogénéisation des causes de réadmissions, la grande quantité d'information et par conséquent l'extraction de l'information.

Problèmes statistiques lors de l'évaluation de l'association entre covariables et réhospitalisations

Le choix du modèle

Certaines méthodes statistiques ne sont pas appropriées pour analyser ces données car elles ne tiennent pas compte de toute l'information disponible (Figure 3). Analyser la survenue de la première hospitalisation, qui est une façon simple et naïve d'analyser les réhospitalisations, conduit à une perte d'information [22]. En effet, ces analyses ignorent les réhospitalisations suivantes et ainsi toute l'information connue n'est pas utilisée. Cela peut ainsi amener à des conclusions erronées. De plus, les facteurs associés à la première réhospitalisation peuvent être différents de ceux associés aux réhospitalisations ultérieures [23]. En analysant le nombre total de réhospitalisations, le délai de survenu des réhospitalisations est complétement ignoré [24]. En effet, il est raisonnable de penser que plus les hospitalisations sont rapprochées plus elles peuvent être graves. De plus, un patient suivi plus longtemps qu'un autre peut avoir un nombre de réhospitalisations plus élevé qu'un patient suivi moins longtemps. Il est donc aussi important de prendre en compte le temps de suivi du patient.
Enfin, le problème des covariables dépendantes du temps se pose [22]. Par exemple, le type de traitement peuvt évoluer au cours du temps, ce qui pourrait impacter le risque de réhospitalisation.

(a) Observé

(b) Analyse de la survenue de la 1ère réhospitalisation

(c) Analyse du nombre de réhospitalisations

FIGURE 3 – Analyses inappropriées des événements récurrents

(a) représente trois patients, le premier ayant eu trois réhospitalisations et a mené son suivi à son terme. Le deuxième a une seule réhospitalisation et est perdu de vue pendant son suivi. Enfin le troisième n'a pas eu de réhospitalisation durant son suivi. La Figure (b) montre que l'analyse s'intéresse seulement à la survenue du premier événement et ignore tout ce qui se passe ensuite. La Figure (c) montre que tout se qui se passe durant le suivi est ignoré et seulement le nombre de réhospitalisations est utilisé.

Hétérogénéité et dépendance

Les patients ont des caractéristiques qui leur sont propres et qui les rendent différents les uns des autres. Par exemple, certains vont être réhospitalisés plus tôt ou plus souvent que d'autres. C'est ce qui fait que les populations sont dites hétérogènes. Dans la recherche médicale, cette

hétérogénéité peut être à l'origine des différences d'efficacité d'un traitement d'un patient à l'autre. L'hétérogénéité observée peut être prise en compte dans les modèles de régression en incluant les variables à l'origine de cette hétérogénéité. Cependant, il peut rester de la variabilité. En effet, un traitement peut être efficace pour un patient et pas pour un autre alors que ces deux sujets présentent les mêmes caractéristiques. Ces différences peuvent être alors dues à des facteurs non-observés ou non mesurables tels que des facteurs environnementaux ou encore génétiques. Cette hétérogénéité inter-patient est alors dite non-observée (Figure 4). Dans le contexte des réhospitalisations ou plus généralement la survenue d'événements récurrents, cette hétérogénéité non observée, aussi appelé « fragilité », engendre une corrélation entre les temps de survenue des événements récurrents au sein d'un sujet donné.

La corrélation entre les temps de survenue des événements peut être aussi due à la dépendance entre les événements récurrents. En effet, un patient souvent réhospitalisé aura tendance à l'être de plus en plus. On pourrait en effet penser qu'un patient est réhospitalisé car son état se dégrade et plus son état se dégrade plus il est réhospitalisé. En d'autres termes, la survenue de l'événement modifie (accélère ou ralentit) le risque de survenue d'événements futurs. Ainsi considérer les réhospitalisations comme indépendantes les unes des autres serait une erreur. Une des conséquences de la non prise en compte de cette dépendance serait la diminution de la largeur des intervalles de confiance des paramètres de régression (β) qui impliquerait un risque de rejeter l'hypothèse nulle ($\beta = 0$) à tort plus élevé ie mettre en évidence des facteurs associés aux réhospitalisations à tort.

Les événements terminaux

Les événements terminaux, comme le décès, sont des événements qui stoppent de façon définitive le processus d'événements récurrents. Inversement, les événements récurrents peuvent aussi avoir un impact sur la survenue de l'événement terminal.
Si les événements récurrents ne modifient pas le risque de survenue de l'événement terminal, il est quand même important de le prendre en compte dans l'analyse des événements récurrents car il modifie la population à risque de présenter un événement récurrent. En effet, l'événement terminal peut ainsi modifier artificiellement le nombre d'événements au cours du temps. Ainsi un patient qui décède de façon précoce aura moins d'événements qu'un patient qui décède tardivement car le nombre d'événements augmente au cours du temps. A contrario, si l'on considère que les patients qui meurent tôt dans l'étude sont les patients les plus graves et ont donc un plus haut risque d'être réhospitalisés alors les patients restants sont ceux qui ont moins de risque d'être réhospitalisés. Ainsi au cours du temps, la population à risque est composée de plus en plus de personnes à faible risque de réhospitalisation, mais le risque de réhospitalisation augmente au cours du temps car les patients qui décèdent ne sont plus suivis et ne font plus partie de la population à risque (Figure 5).

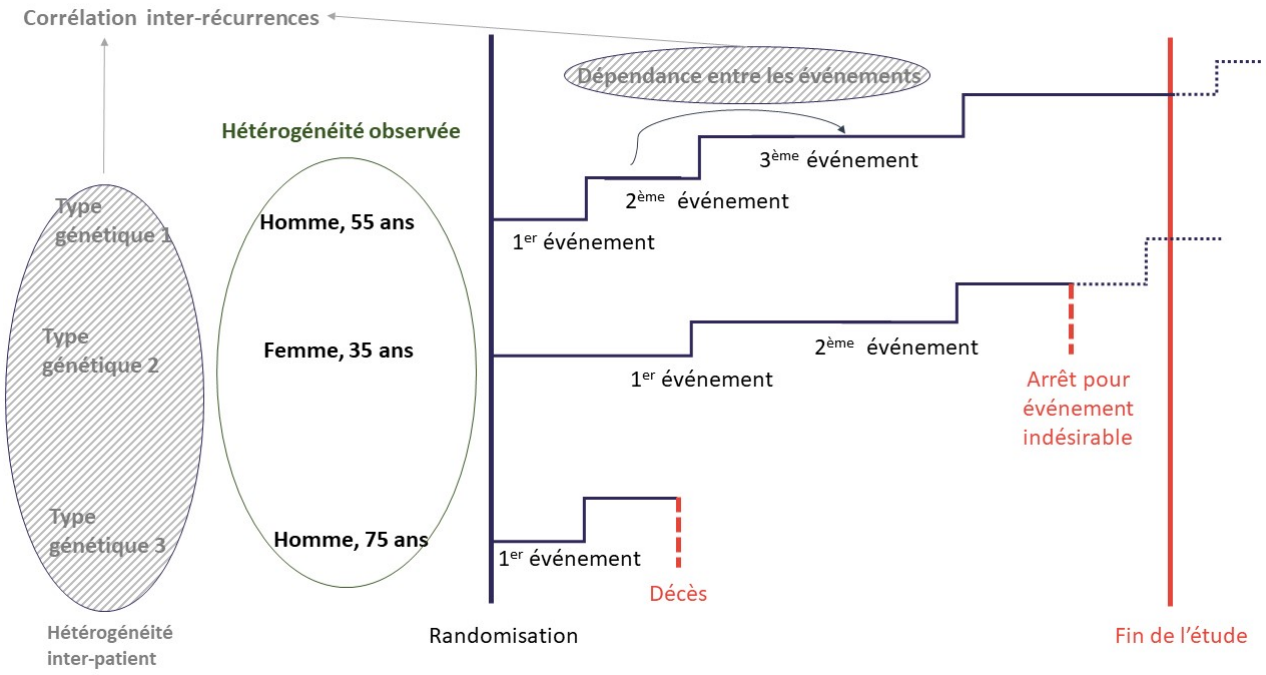


FIGURE 4 – Les différents types d'hétérogénéité

La Figure 4 représente la survenue des réhospitalisations pour trois patients fictifs après une randomisation. Les traits bleus foncés représentent la période de suivi du patient, et les traits bleus clairs en pointillés représentent ce qu'il s'est passé après la fin du suivi du patient et qui n'a pas été observé.
Le patient 1 est un homme de 55 ans ayant été réhospitalisé 3 fois et a été suivi jusqu'à la fin de l'étude mais a été rehospitalisé après. Le patient 2 est une femme de 35 ans ayant été hospitalisée 2 fois pendant son suivi. Elle a dû stopper l'étude à cause d'un événement indésirable et n'a pas donc pu être suivie jusqu'à la fin de l'étude. Elle a été réhospitalisée après la fin de son suivi. Le patient 3 est un homme de 75 ans qui a été réhospitalisé une seule fois et a arrêté l'étude précocement à cause de son décès.
Les différentes caractéristiques au moment de la randomisation (sexe et âge) représentent l'hétérogénéité observée ie des informations qui ont été recueillies au moment de l'inclusion des sujets dans l'étude. Les différents types génétiques de ces patients représentent l'hétérogénéité inter-patients. La dépendance entre les événements ainsi que l'hétérogénéité sont non observés ie cette information existe mais n'a pas été recueillie ou non mesurable et n'est donc pas connue (gris hachuré).

Inversement, l'événement terminal peut être associé au processus de récurrence. En effet, la survenue des événements récurrents peut augmenter le risque de survenue de l'événement terminal. Dans ce cas, l'événement terminal ne peut pas seulement être considéré comme ayant un impact sur la population à risque. Cette dépendance est souvent non-observée et peut être à l'origine d'une hétérogénéité. Il est donc primordial de prendre en compte cette dépendance entre les récurrences et l'événement terminal dans l'analyse. En pratique, dans les essais cliniques, il est commun d'utiliser un critère binaire composite combinant les réhospitalisations et les décès [25, 26, 16]. Cependant, il présente plusieurs inconvénients, puisque d'une part, il ne permet pas de prendre en compte la répétition des réhospitalisations et d'autre part, il considère que le décès et la réhospitalisation sont des événements de même gravité.

Enfin, plusieurs types d'événements terminaux peuvent survenir, même si un patient ne peut avoir qu'un seul type d'événement terminal, par exemple les différentes causes de décès. Cependant, il est fréquent de ne s'intéresser qu'à un seul type d'événement terminal. Dans ce

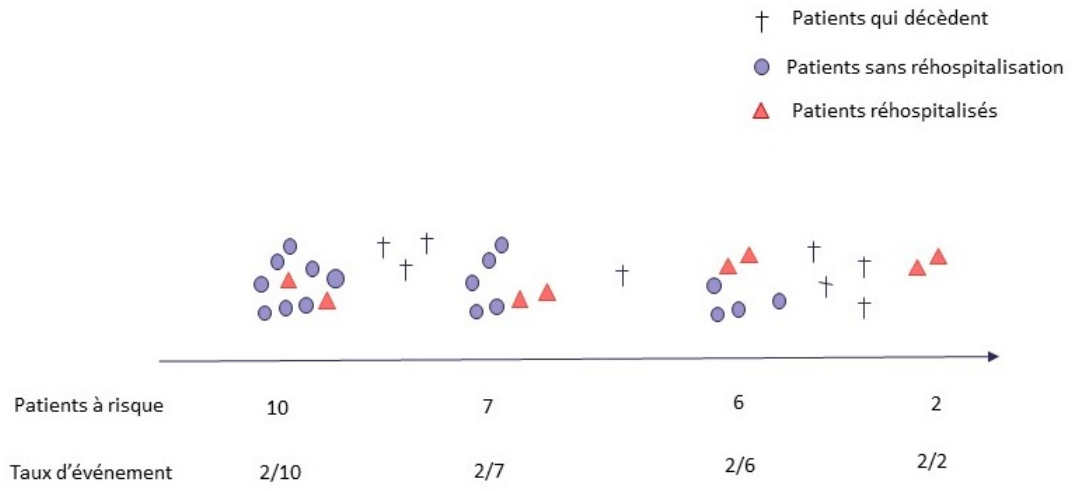


FIGURE 5 – L'évolution du taux de réhospitalisations en fonction de l'événement terminal

La Figure 5 représente l'évolution du taux de réhospitalisations en fonction du nombre de personnes à risque. On remarque que le taux augmente avec le temps alors que le nombre de réhospitalisations reste stable.

cas, les autres types d'événements terminaux sont appelés événements concurrents ou encore risques compétitifs. Un risque compétitif est un événement qui empêche la survenue ou modifie la probabilité de survenue de l'événement d'intérêt. Dans le contexte des réhospitalisations, l'événement terminal peut être vu comme un événement concurrent des réhospitalisations. Ainsi, ignorer les événements concurrents revient à se placer dans un monde fictif dans lequel les individus peuvent expérimenter seulement l'événement d'intérêt et à surestimer l'incidence de l'événement d'interêt [27].

Problèmes statistiques liés à l'identification des parcours de soins des patients

Le parcours de soin des patients peut être défini par les causes de réhospitalisations. Le point primordial pour cette analyse est de définir le niveau de détail des causes. En effet, si elles sont trop détaillées, il est probable que l'on ne trouve que très peu, voire aucune combinaison commune à certains patients. A l'inverse si elles ne le sont pas assez, les différentes combinaisons risquent de ne pas être informatives. La Figure 6 représente les réhospitalisations de trois patients avec un niveau de détail élevé des causes de réhospitalisations. A partir de cet exemple, il est impossible de mettre en évidence un parcours de soin commun à au moins deux patients. La Figure 7 représente les mêmes patients mais avec aucun détail sur les causes de réhospitalisation. Ainsi, on trouve que les trois patients ont été tous réhospitalisés pour hernie et anémie. Cependant, être réhospitalisé pour une hernie d'un disque n'est pas pareil que d'être réhospitalisé pour une hernie inguinale, il en est de même pour une anémie en vitamine B12 et une carence en fer. Avec un codage plus approprié, il est possible d'identifier dans la Figure 8 un parcours de soin commun aux patients 2 et 3, qui ont tous les deux été réhospitalisés pour carence en fer et pour une atteinte des disques intervertébraux.

Une fois l'étape de codage finie, les parcours de soins peuvent être construits. Cependant, une question se pose qui est l'identification et la sélection des parcours de soins pertinents. En effet, certains parcours de soins peuvent être communs à un grand nombre de patients et d'autres spécifiques à un seul patient.

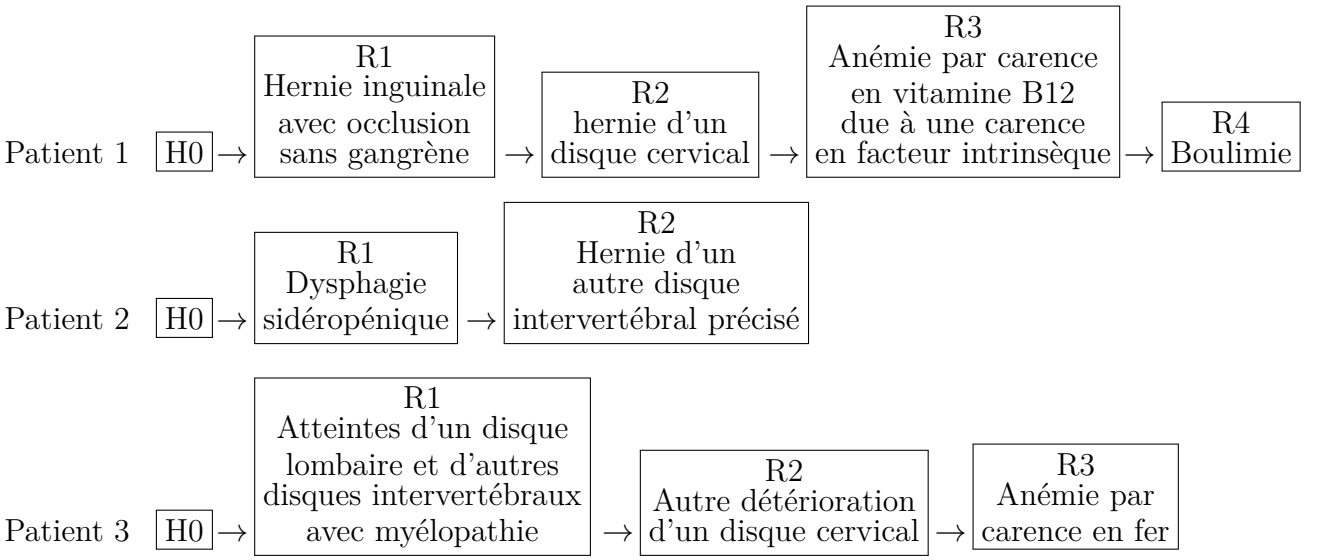


FIGURE 6 – Codage trop détaillé

Exemple de parcours de soins de trois patients où :
H0 représente l'hospitalisation index ;
Rx les réhospitalisations après H0.

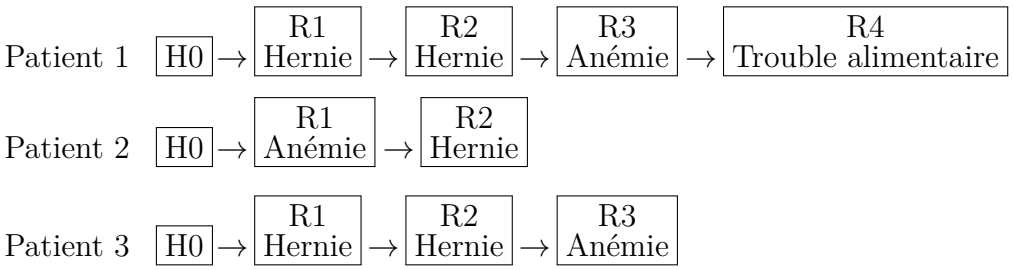


FIGURE 7 – Codage pas assez détaillé

Exemple de parcours de soins de trois patients où :
H0 représente l'hospitalisation index ;
Rx les réhospitalisations après H0.

Description des chapitres

Notre travail a pour but de montrer que l'information issue des réhospitalisations peut être analysée de différentes façons. Le manuscrit se compose de deux grandes parties.
La première partie consiste en l'analyse des réhospitalisations par des modèles de régression prenant en compte les événements récurrents en présence d'un ou plusieurs événements terminaux. Elle commence par un état de l'art de la modélisation de la survenue de la première réhospitalisation (Chapitre 1) et des méthodes d'analyses des événements récurrents (Chapitre 2). Nous avons ensuite comparé par une étude de simulations, les différents modèles de régression existants pour analyser les événements récurrents en présence d'un événement terminal

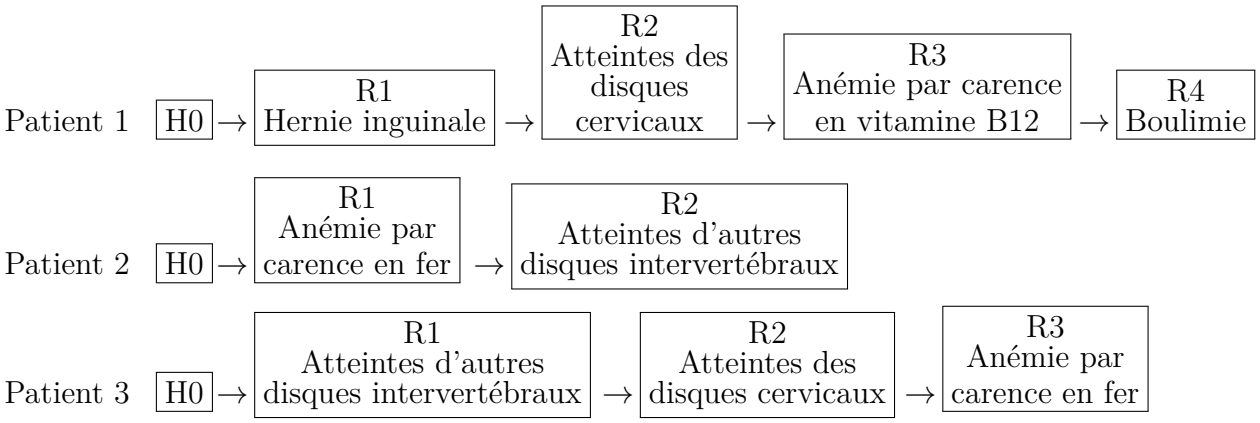


FIGURE 8 – Bon codage

Exemple de parcours de soins de trois patients où :
H0 représente l'hospitalisation index ;
Rx les réhospitalisations après H0.
Cet exemple est pour illustrer le niveau de granularité nécessaire (ni trop, ni pas assez), et ces données ne seront plus utilisées dans le reste du manuscrit.

(Chapitre3). Puis nous avons développé un modèle de régression permettant d'analyser les événements récurrents et un événement terminal en présence d'un risque compétitif sur l'événement terminal (Chapitre 4). Nous avons ainsi étudié les propriétés de ce modèle dans différents contextes à partir d'une étude de simulations. Le développement de ce modèle a été motivé par les données sur les réhospitalisations, survenant à n'importe quel moment, suivant une greffe de foie. De ce fait, nous avons alors appliqué ce modèle à ces données afin d'évaluer l'effet de l'âge et du sexe sur les réhospitalisations et aux décès liés aux maladies du foie après une greffe de foie (chapitre 4). Dans cette première partie, nous nous sommes intéressés aux réhospitalisations non planifiées.

La seconde partie concerne l'identification des trajectoires de soins la première année suivant une chirurgie bariatrique. Les données permettant de construire les trajectoires sont issues des raisons des réhospitalisations. Elle commence par décrire les différentes méthodes de fouille de données avant de décrire le principe de l'analyse formelle de concept (Chapitre 5). Nous avons ensuite appliqué la méthode d'analyse formelle de concept aux données de réhospitalisations la première année suivant la chirurgie bariatrique (Chapitre 6). Pour la construction des trajectoires de soins, nous avons considéré les réhospitalisations non planifiées mais aussi les réhospitalisations planifiées.

Première partie

Modèles de régression pour analyser les réhospitalisations

Chapitre 1

Analyse de la survenue de la première réhospitalisation

De nombreuses études s'intéressent au délai de survenue d'un événement en particulier (par exemple le décès ou la première réhospitalisation) (Figure 1.1). Pour cela, les sujets sont suivis sur une période de temps. Cependant, pour diverses raisons, l'événement peut ne pas être observé pour certains sujets. Par exemple, le sujet a été perdu de vue, et est donc sorti précocement de l'étude, ou encore car le sujet n'a pas présenté l'événement avant la fin de l'étude. C'est ce que l'on appelle des données censurées.
Une façon simple et naïve d'analyser cette information est de le faire de façon binaire « le sujet a t'il eu l'événement entre son inclusion et la fin de l'étude ? ». Cependant, si le sujet n'a pas poursuivi le suivi jusqu'à son terme, on ne peut pas répondre à la question et le sujet ne peut pas être pris en compte dans l'analyse, ce qui conduit à une perte de puissance. Afin de pallier ce problème, des méthodes ont été développées afin d'éviter l'exclusion des patients censurés de l'analyse. Ce sont les méthodes appelées méthodes de survie.

FIGURE 1.1 – Analyse de la survenue du premier événement

Ce chapitre décrit les méthodes les plus utilisées pour l'analyse de survie. Dans un premier temps, nous définirons les notions nécessaires à la compréhension des différentes méthodes. Puis, nous décrirons les méthodes d'estimation et de comparaison des fonctions de survie et enfin nous présenterons le modèle de Cox.

1.1 Définitions

1.1.1 Les différentes dates

L'analyse de survie s'intéresse à un délai jusqu'à la survenue d'un événement. Le calcul du délai nécessite de définir une date de début et une date de fin.

- La **date d'origine** est la date à partir de laquelle on a débuté l'observation. Elle peut correspondre par exemple à la date de naissance, la date de diagnostic, ou même la date de randomisation.
- La **date de dernière nouvelle** est la date la plus récente où le sujet a été revu.
- La **date de point** est fixée à priori (dans le protocole) et correspond à la date du bilan de l'étude, au-delà de cette date, on ne tient plus compte des informations éventuellement recueillies.

1.1.2 Censures

Dans les analyses longitudinales, la survenue de l'événement d'intérêt peut ne pas être observée et donc la date exacte de survenue de l'événement peut ne pas être connue, c'est ce qu'on appelle la censure. On distingue trois types de censures (Figure 1.2) :

- La censure à droite : l'événement apparaît après la fin de la période de suivi du sujet soit après la date de point ou de dernière nouvelle.
- La censure à gauche : l'événement apparaît avant le début de la période de suivi du sujet soit avant la date d'origine.
- La censure par intervalle : l'événement apparaît entre deux observations (visites par exemple).

Dans les recherches biomédicales la censure la plus rencontrée est la censure à droite. Elle peut être due à la fin de l'étude ou au fait que le sujet soit perdu de vue.
Dans la suite du manuscrit, on ne parlera que des censures à droite.

1.1.3 Durée de survie

La durée de survie désigne le temps écoulé jusqu'à la survenue d'un événement précis depuis la date d'origine.
Soit T une variable aléatoire définit par $T = C \wedge X$ (a$\wedge$b=min(a,b)) où C et X sont respectivement les variables aléatoires correspondant aux temps de censure et de survenue de l'événement. La variable aléatoire T est non négative et on suppose dans ce qui suit que sa loi est continue. Sa distribution est souvent asymétrique à droite.

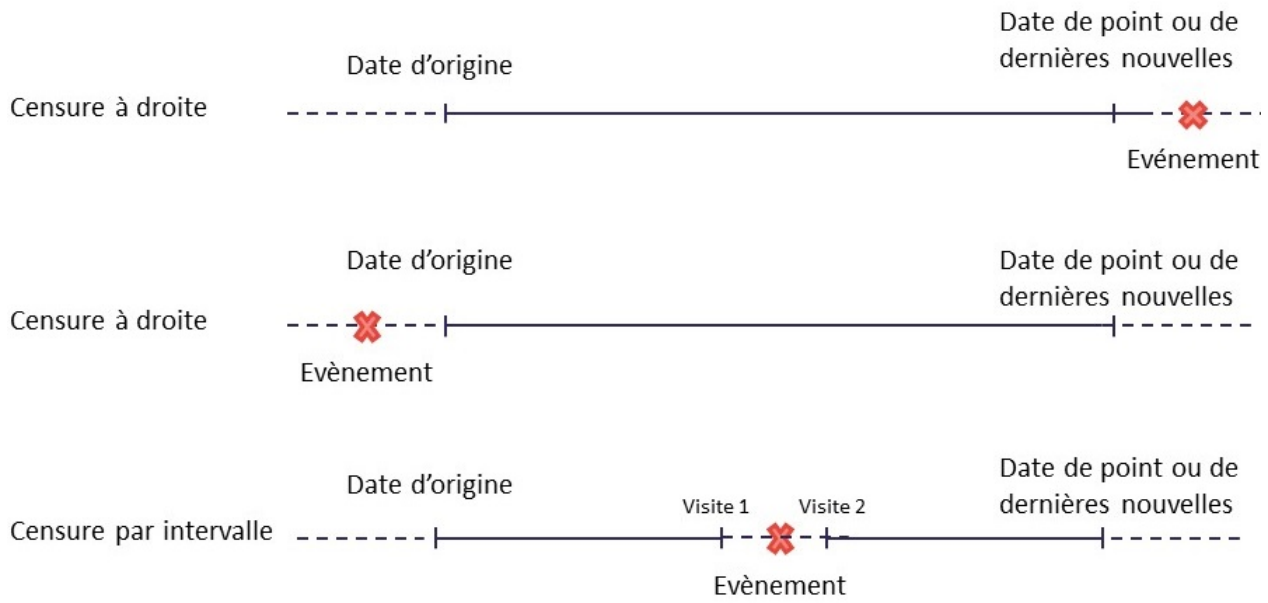


FIGURE 1.2 – Censure

1.1.4 Fonction de survie S et fonction de répartition F

$\forall t \in \mathbb{R}^+$, la **fonction de survie** est la probabilité de ne pas présenter l'événement d'intérêt avant t et est notée $S(t)$:

$$S(t) = \mathbb{P}(t < T), \quad t \geq 0. \tag{1.1}$$

La fonction de survie est monotone décroissante et continue : $S(0) = 1$ et $\lim_{t \to +\infty} S(t) = 0$

La **fonction de répartition** est la probabilité de survenue de l'événement d'intérêt avant le temps t et est notée $F(t)$:

$$F(t) = \mathbb{P}(T \leq t) = 1 - S(t). \tag{1.2}$$

La **densité de probabilité** est la fonction positive telle que pour tout $t \geq 0$:

$$F(t) = \int_0^t f(u)du. \tag{1.3}$$

Si la fonction de répartition F admet une dérivée au point t alors :

$$f(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt)}{dt} = F'(t) = -S'(t). \tag{1.4}$$

La densité représente la probabilité instantanée de présenter l'événement d'interêt dans un petit intervalle de temps après t.

1.1.5 Fonctions de risque

La **fonction de risque instantané** est la probabilité instantanée de présenter l'événement d'intérêt dans un petit intervalle de temps juste après t sachant que l'on n'a pas fait l'événement avant t. Elle est notée $h(t)$:

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt | T \geq t)}{dt}. \tag{1.5}$$

D'après le théorème de Bayes, elle est égale à :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log[S(t)]. \tag{1.6}$$

Ce qui donne :

$$S(t) = \exp\left(-\int_0^t h(u)du\right). \tag{1.7}$$

La fonction de risque cumulé est notée $H(t)$ et est égale à :

$$H(t) = \int_0^t h(u)du. \tag{1.8}$$

On obtient alors les relations suivantes :

$$S(t) = \exp\left(-H(t)\right) \tag{1.9}$$

$$f(t) = h(t)S(t) = h(t)\exp\left(-H(t)\right). \tag{1.10}$$

1.1.6 La fonction de vraisemblance

Soit une censure aléatoire à droite C indépendante du temps de survenue de l'événement d'intérêt X de densités f_C et f et de survies S_C et S respectivement. Soit la durée T_i où $T_i = \min(X_i, C_i)$ et $\delta_i = \mathbf{1}_{\{X_i < C_i\}}$ qui indique si le patient i a présenté l'événement. La fonction de vraisemblance [1] complète s'écrit alors :

$$L(\beta) = \prod_{i=1}^{n} (\mathbb{P}(T_i \in [t_i, t_i + dt[, T_i = X_i | \beta)^{\delta_i} (\mathbb{P}(T_i \in [t_i, t_i + dt[, T_i = C_i | \beta)^{1-\delta_i} \tag{1.11}$$

où t_i est la valeur de la durée observée pour le patient i et β le vecteur de paramètres de régression. La première partie de la vraisemblance $\mathbb{P}(T_i \in [t_i, t_i + dt[, T_i = X_i | \beta)$ est la

1. La vraisemblance est une expression mathématique qui décrit les probabilités conjointes d'obtenir les données observées des sujets inclus dans l'étude, comme étant une fonction des paramètres inconnus du modèle considéré

probabilité que le temps t_i soit un événement et pas une censure ce qui revient à :

$$\begin{aligned} \mathbb{P}(T_i \in [t_i, t_i + dt[, T_i = X_i|\beta) &= \mathbb{P}(X_i \in [t_i, t_i + dt[, \delta_i = 1|\beta) && (1.12) \\ &= \mathbb{P}(X_i \in [t_i, t_i + dt[, C_i > t_i|\beta). && (1.13) \end{aligned}$$

On fait l'hypothèse de la censure indépendante i.e. X_i est indépendante de C_i conditionnellement aux covariables, ainsi que l'hypothèse de censure non informative i.e. la loi de la censure ne dépend pas du paramètre β, on obtient alors :

$$\begin{aligned} \mathbb{P}(X_i \in [t_i, t_i + dt[, C_i > t_i|\beta) &= \mathbb{P}(X_i \in [t_i, t_i + dt[|\beta)\mathbb{P}(C_i > t_i) && (1.14) \\ &= f(t_i|\beta)S_C(t_i). && (1.15) \end{aligned}$$

De la même façon on a :

$$\begin{aligned} \mathbb{P}(T_i \in [t_i, t_i + dt[, T_i = C_i|\beta) &= \mathbb{P}(C_i \in [t_i, t_i + dt[, \delta_i = 0|\beta) && (1.16) \\ &= \mathbb{P}(C_i \in [t_i, t_i + dt[, X_i > t_i|\beta) && (1.17) \\ &= \mathbb{P}(C_i \in [t_i, t_i + dt[)\mathbb{P}(X_i > t_i|\beta) && (1.18) \\ &= f_C(t_i)S(t_i|\beta). && (1.19) \end{aligned}$$

On a alors :

$$L(\beta) = \prod_{i=1}^{n} \left(f(t_i|\beta)S_C(t_i)\right)^{\delta_i} \left(f_C(t_i)S(t_i|\beta)\right)^{1-\delta_i}. \quad (1.20)$$

D'après l'hypothèse de censure non informative et indépendante, les paramètres β n'apparaissent pas dans la loi de la censure, et la vraisemblance L est alors proportionnelle à :

$$L(\beta) \propto \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \quad (1.21)$$

1.2 Estimation et test de la fonction de survie S(t)

Il existe plusieurs méthodes pour estimer la fonction de survie. Certaines sont paramétriques et font donc l'hypothèse d'une distribution prédéfinie de la fonction de survie. D'autres sont non paramétriques comme la méthode actuarielle [28] pour temps discrets ou encore la méthode de Kaplan-Meier [29] pour les temps continus (Figure 1.3). Après avoir estimé la fonction de survie, il est d'usage de vouloir comparer deux ou plusieurs fonctions de survie, par exemple comparer les fonctions de survie de deux groupes de traitement. Pour cela, plusieurs méthodes ont été développées. Une de ces méthodes consiste à comparer les médianes de chaque courbe de survie, pour laquelle plusieurs tests ont été développés [30, 31, 32, 33]. Une autre permet de comparer la globalité des courbes entre elles grâce au test du Logrank [34]. La suite de cette

section n'abordera que la méthode de Kaplan-Meier [29], méthode d'estimation et le test du Logrank [34] méthode pour la comparaison des fonctions de survie les plus utilisées.

1.2.1 Méthode de Kaplan-Meier

L'estimation de Kaplan-Meier [29] est aussi appelée Produit-Limite et est un estimateur du maximum de vraisemblance (eq 1.21). La construction de l'estimateur repose sur l'idée que ne pas présenter l'événement après un temps t, c'est ne pas l'avoir présenté juste avant t et de ne pas le présenter au temps t (Table 1.1). Soit deux temps de survie t_1 et t_2 tels que $t_1 < t_2$ alors :

$$\mathbb{P}(T > t_2) = \mathbb{P}(T > t_2 | T > t_1)\mathbb{P}(T > t_1) \tag{1.22}$$

$\mathbb{P}(T > t_2|T > t_1)$ peut être estimé par $\frac{n_{t_2} - m_{t_2}}{n_{t_2}}$ où n_{t_2} est le nombre de personnes à risque au temps t_2, m_{t_2} est le nombre d'événements au temps t_2. De façon générale, on obtient la récurrence suivante :

$$\hat{S}(t_j) = \hat{S}(t_{j-1})\frac{n_j - m_j}{n_j} \tag{1.23}$$

où

— m_j est le nombre d'événements survenus entre les temps t_{j-1} et t_j

— $S(t_{j-1})$ est le taux de survie au temps t_{j-1}

— n_j est le nombre de sujets à risque au temps t_j et est égal à $n_{j-1} - c_{j-1} - m_{j-1}$ où c_{j-1} est le nombre de censures ayant eu lieu entre t_{j-1} et t_{j-2}

En pratique :

TABLE 1.1 – Estimation de la fonction de survie par la méthode de Kaplan-Meier

Temps d'événements ordonnés	Personnes à risque	Nombre d'événements	Nombre de personnes censurées	$\hat{S}(t)$
$T_0 = 0$	N	0	0	1
T_1	N	m_1	c_1	$1 \times \frac{N - m_1}{N} = \hat{S}(T_1)$
T_2	$n_2 = N - m_1 - c_1$	m_2	c_2	$\hat{S}(T_1) \times \frac{n_2 - m_2}{n_2}$
...				

La fonction de survie est une fonction en escalier, décroissante, constante par intervalle. Ci-dessous une représentation graphique de la fonction de survie (Figure 1.3). Les données sont issues du package *survival* du logiciel R. Elles contiennent les temps de survie (décès ou censure) de patients atteints de leucémie myéloïde aiguë.

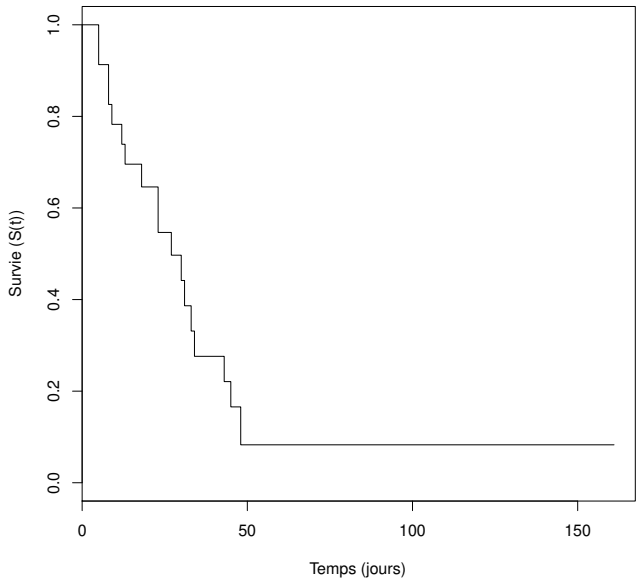


FIGURE 1.3 – Exemple de courbe de survie

1.2.2 Test du LogRank

Le test du LogRank [34] est le test le plus utilisé pour la comparaison de deux fonctions de survie. C'est une généralisation du test du χ^2 stratifié de Mantel-Haenszel [35]. Les hypothèses du test :

H_0 : Les deux fonctions de survie sont égales
H_1 : Les fonctions de survie diffèrent l'une de l'autre.

Il est basé sur des tableaux de contingence 2x2 à chaque temps d'événement. Il calcule ainsi à chaque temps de survenue d'un événement les effectifs théoriques et observés de la même façon que le test du χ^2 [35] (Table 1.2).

TABLE 1.2 – Calcul de la statistique de test du Logrank au temps T_j

	Nombre d'événements	Nombre de sujets sans événement	Total
Groupe 1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
Groupe 2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
	d_j	$n_j - d_j$	n_j

La statistique du LogRank est :

$$\chi^2 = \frac{(\sum_{j=1}^{k} d_{1j} - E_{1j})^2}{\sum_{j=1}^{k} V_{1j}} \sim \chi^2_{1ddl} \tag{1.24}$$

Où

— k le nombre total d'événements

— E_{1j} est le nombre d'événements attendus sous l'hypothèse d'égalité dans le groupe 1 au temps j. $E_{1j} = \frac{n_{1j}d_j}{n_j}$

— $V_{1j} = \frac{n_{1j}n_{2j}(n_j - d_j)d_j}{n_j^2(n_j-1)}$

Une condition est nécéssaire pour appliquer le test du Logrank. Il faut que le risque dans un groupe soit toujours supérieur au risque de l'autre groupe tout au long du suivi. En d'autres termes, il ne faut pas que les courbes se croisent.
Le test du Logrank [34] peut être généralisé afin de comparer plus de deux fonctions de survie.

Le test du Logrank [34] donne le même poids à tous les événements. Cependant, d'autres tests, reposant sur le même principe que le test du Logrank [34], ont été développés, la différence étant les poids donnés aux événements. Le test de Wilcoxon [36, 37] (ou test de Gehan ou de Breslow-Gehan-Wilcoxon) donne un poids plus important aux événements survenus précocément. Pour cela il pondère la statistique par le nombre d'individus encore à risque avant l'événement. Le test de Prentice-Peto-Peto [38, 39] pondère par la survie.

1.3 Modèles de régression

La fonction de survie permet de visualiser si un ou plusieurs groupes ont des fonctions de survie différentes. Cependant, cette méthode ne permet pas de quantifier l'effet de la covariable sur la fonction de survie, or il est souvent intéressant de chercher à établir un lien entre les deux. Cela permet d'autant plus d'améliorer la qualité de la modélisation en utilisant de l'information supplémentaire, comme l'ajout de variables d'ajustement. Les modèles de régression permettent de répondre à cette question. Plusieurs types de modèles ont été développés :

— Les modèles paramétriques qui font une hypothèse sur la distribution de la survie. Il y a les modèles à temps accélérés [40], qui modélisent directement les temps de survenue des événements ou encore le modèle de Weibull [41], qui modélise quant à lui la fonction de risque instantané.

— Les modèles semi-parametriques comme le modèle de Cox [42] qui modèlisent comme le modèle de Weibull [41] la fonction de risque instantané.

Dans la section suivante, nous nous focaliserons uniquement sur le modèle de Cox. Nous décrirons plus en détail l'écriture du modèle de Cox ainsi que l'estimation et les tests de ses paramètres.

1.3.1 Le modèle de Cox

Il permet d'estimer l'effet de variables sur la fonction de risque instantané et s'écrit de la façon suivante :

$$h(t|Z) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j Z_j\right) \tag{1.25}$$

où β sont les coefficients de régression et Z_j les p covariables et $Z = (Z_1, ..., Z_p)$. Le modèle de Cox est composé de deux parties, le risque instantané de base noté $h_0(t)$ et la partie exponentielle. Le risque instantané de base est dépendant du temps et pas des covariables et correspond au risque de faire l'événement lorsque toutes les covariables sont à 0. C'est une fonction non-specifiée, c'est pour cette raison que l'on dit que le modèle de Cox est un modèle semi-paramétrique. La partie exponentielle est quant à elle indépendante du temps et correspond à l'exponentielle d'une combinaison linéaire. Cette écriture permet de garantir un risque instantané positif.

1.3.2 L'estimation des paramètres du modèle de Cox

1.3.2.1 Vraisemblance et vecteur score

Les paramètres du modèle de Cox sont estimés par la méthode du maximum de vraisemblance. En reprenant l'équation 1.21 et en remplaçant $f(T_i|\beta)$ par $h(T_i|\beta)S(T_i|\beta)$(1.10) on obtient :

$$\begin{aligned}
L_p(\beta) &= \prod_{i=1}^{n} f(T_i|\beta)^{\delta_i} S(T_i|\beta)^{1-\delta_i} = \prod_{i=1}^{n} h(T_i|\beta)^{\delta_i} S(T_i|\beta) && (1.26)\\
&= \prod_{i=1}^{n} \left(h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j Z_{ij}\right)\right)^{\delta_i} \exp\left(-\int_0^{T_i} h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j Z_{ij}\right) dt\right) && (1.27)\\
&= \prod_{i=1}^{n} \left(\frac{h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j Z_{ij}\right)}{\sum_{k=1}^{n} I_{(T_k \geq T_i)} h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)}\right)^{\delta_i} \left(\sum_{k=1}^{n} I_{(T_k \geq T_i)} h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)\right)^{\delta_i} && (1.28)\\
&\quad \times \exp\left(-\int_0^{T_i} h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j Z_{ij}\right) dt\right) && (1.29)
\end{aligned}$$

où Z_{ij} est la valeur de la covariable Z_j du patient i. D'après Cox (1972), les intervalles de temps ne contenant aucun événement n'apportent aucune information sur les paramètres de régression β (vecteur de paramètres de dimension p) car on peut concevoir que l'estimation de $h_0(t)$ est nulle dans ces intervalles là.

La vraisemblance se simplifie (si on se restreint aux fonctions constantes par morceaux pour le risque de base cumulé), en utilisant la vraisemblance partielle.

$$L_p(\beta) = \prod_{i=1}^{n} \left(\frac{\exp\left(\sum_{j=1}^{p} \beta_j Z_{ij}\right)}{\sum_{k=1}^{n} I_{(T_k \geq T_i)} \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)} \right)^{\delta_i} \tag{1.30}$$

$\frac{\exp\left(\sum_{j=1}^{p} \beta_j Z_{ij}\right)}{\sum_{k=1}^{n} I_{(T_k \geq T_i)} \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)}$ peut s'interpréter comme étant la probabilité conditionnelle que le sujet i présente l'événement au temps T_i sachant que T_i est un temps d'événement observé.

La vraisemblance du modèle de Cox (1.30) est alors appelée vraisemblance partielle. En effet, elle ne considère seulement que les probabilités pour les sujets qui ont présenté l'événement d'interêt et pas celles des sujets censurés. De plus, elle ne nécessite pas l'estimation de la fonction de risque de base qui est donc considérée comme un paramètre de nuisance.

La log-vraisemblance pour le paramètre β s'écrit de la façon suivante :

$$LL_p(\beta) = \log\left(L_p(\beta)\right) = \sum_{i=1}^{n} \delta_i \left(\sum_{j=1}^{p} \beta_j Z_{ij} - \log\left(\sum_{k=1}^{n} I_{(T_k \geq T_i)} \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj} \right) \right) \right). \tag{1.31}$$

Le vecteur score de la log-vraisemblance est égal à :

$$\begin{aligned} U(\beta_j) &= \left(\frac{\partial LL_p(\beta)}{\partial \beta_j} \right)_{j \in 1,..,p} \\ &= \left(\sum_{i=1}^{n} \delta_i \left(Z_{i1} - \frac{\sum_{k=1}^{n} I_{(T_k \geq T_i)} Z_{k1} \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)}{\sum_{k=1}^{n} I_{(T_k \geq T_i)} \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)} \right), ..., \sum_{i=1}^{n} \delta_i \left(Z_{ip} - \frac{\sum_{k=1}^{n} I_{(T_k \geq T_i)} Z_{kp} \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)}{\sum_{k=1}^{n} I_{(T_k \geq T_i)} \exp\left(\sum_{j=1}^{p} \beta_j Z_{kj}\right)} \right) \right). \end{aligned}$$

et vaut 0 en $\hat{\beta}$. Il n'y a pas de solution exacte à ce problème. L'algorithme de Newton-Raphson est souvent utilisé par les logiciels pour obtenir une solution approchée.

L'estimateur du risque cumulé de base $H_0(t)$ est de type Breslow [37] :

$$\hat{H}_0(t) = \sum_{i=1}^{n} \frac{\delta_i I_{(t \geq T_i)}}{\sum_{k=1}^{n} I_{(T_k \geq T_i)} \exp\left(\sum_{j=1}^{p} \hat{\beta}_j Z_{kj}\right)}. \tag{1.32}$$

1.3.2.2 Tests pour les paramètres de régression

Il existe trois tests permettant de tester si les paramètres de régression sont différents de 0. Les hypothèses sont les suivantes :

$$H_0 : \beta_1 = ... = \beta_p = 0$$
$$H_1 : \text{Il existe } j \in 1, .., p \text{ tel que } \beta_j \neq 0$$

— Test de Wald (maximum de vraisemblance)

$$(\hat{\beta} - 0)^T I(\hat{\beta})^{-1}(\hat{\beta} - 0) \xrightarrow{\mathcal{L}} \chi^2 \text{ à } p \text{ degrés de liberté}$$

— Rapport de vraisemblance

$$2(LL_p(\hat{\beta}) - LL_p(0)) \xrightarrow{\mathcal{L}} \chi^2 \text{ à } p \text{ degrés de liberté}$$

— Test du Score

$$U^T(0)I(0)^{-1}U(0) \xrightarrow{\mathcal{L}} \chi^2 \text{ à } p \text{ degrés de liberté}$$

où $\hat{\beta}$ est l'estimateur du maximum de vraisemblance de dimension p, où p est le nombre de paramètres à tester et I est la matrice d'information de Fisher.

1.3.3 Les hypothèses du modèle

Le modèle de Cox nécessite la vérification de trois hypothèses importantes :

— La censure doit être indépendante de la variable d'interêt conditionnellement aux covariables et non informative ie la loi de la censure ne dépend pas des paramètres du modèle.

— La log linéarité pour les variables continues : $\log(\frac{h(t|Z_j=z_1,...,Z_j=z_j)}{h(t|Z_j=z_0,...,Z_j=z_j)}) = \beta_j(z_1 - z_0)$, le log du taux est une fonction linéaire des covariables, ie le risque instantané est constant pour une augmentation d'une unité quelque soit la valeur de la covariable.

— La proportionnalité des risques :
$$\frac{h(t|Z_1,..,Z_j,..,Z_p)}{h(t|Z_1,..,0,..,Z_p)} = \exp(\beta_j Z_j)$$
le taux est constant au cours du temps (indépendant du temps).

En pratique, la censure non informative est une hypothèse difficile à vérifier. La log-linéarité se vérifie par l'analyse graphique des résidus de martingales. Enfin, différentes méthodes existent pour la vérification de la proportionnalité des risques :

— La méthode graphique : elle consiste à tracer soit :

  — le -log(-log($\hat{S}$(t))) pour différentes combinaisons de covariables [43]. $\hat{S}$(t) est obtenue à partir l'estimateur de Kaplan-Meier (1.2.1). La proportionnalié des risques est vérifiée si les courbes sont parallèles.

  — les courbes de survies observées et prédites [43]. La courbe de survie observée est obtenue à partir de l'estimateur de Kaplan-Meier et la courbe de survie prédite est obtenue à partir du modéle de Cox. Si les deux courbes sont proches alors on considère que la proportionnalité des risques est respectée.

— La méthode d'adéquation du modèle basée sur les résidus de Schoenfeld [44]. Cette méthode donne une p-valeur pour chaque covariable incluse dans le modèle, qui s'interprète de la façon suivante : si la p-valeur est inférieure à 0,05 alors la proportionnalité des risques n'est pas respectée pour la covariable concernée, sinon la proportionnalité des risques est respectée.

— La méthode utilisant une covariable dépendante du temps : elle consiste à introduire dans le modèle une interaction entre la covariable et une fonction dépendante du temps. Si l'interaction est significative (Test de Wald ou test du rapport de vraisemblance) alors la proportionnalité des risques n'est pas respectée.

1.4 Conclusion

De nombreuses méthodes permettent d'analyser la survenue d'un événement en présence de censures. Dans un premier temps, l'estimation de la courbe de survie permet de visualiser l'effet d'une covariable discrète. De plus, les modèles de régression permettent de quantifier cet effet. Cependant, dans certains cas, l'événement étudié peut survenir plusieurs fois chez un même sujet, par exemple les réhospitalisations. C'est ce que l'on appelle les événements récurrents. Une façon d'analyser ce type de données est de s'intérésser à la survenue du premier événement. Cependant, cela conduit à une perte d'information et potentiellement à des résultats biaisés. Des méthodes ont alors été développées afin d'utiliser toute l'information. Ces méthodes sont décrites dans le chapitre suivant.

Chapitre 2

Modèles pour l'analyse des réhospitalisations en prenant en compte la récurrence des événements (état de l'art)

2.1 Introduction

Les événements récurrents sont des événements qui peuvent survenir plusieurs fois chez un même sujet (Figure 2.1). Ces événements peuvent être du même type (crises d'epilepsie, infarctus du myocarde) ou de types différents comme par exemple les localisations de métastases. Il est courant d'analyser ce genre de données en ne considérant la survenue que du premier événement par les méthodes classiques adaptées aux données censurées (Chapitre 1). Cependant, ces méthodes sont inefficaces car elles n'utilisent pas toute l'information disponible et donc peuvent produire des résultats biaisés.

De plus, des problèmes inhérents aux événements récurrents se posent. En effet, ils engendrent de la dépendance intra-sujet, c'est-à-dire que la survenue de plusieurs événements chez un même sujet peut affaiblir l'état du sujet et ainsi la probabilité de faire un autre événement peut être modifiée (augmentée ou diminuée). Un deuxième problème est l'hétérogénéité entre les différents sujets. En effet, un plus grand nombre d'événements peut être observé chez les sujets suivis plus longtemps que les sujets suivis moins longtemps.

Plusieurs méthodes ont été développées pour analyser les événements récurrents en prenant en compte les différents problèmes posés. Ces méthodes peuvent êtres classifiées en trois catégories : **i) les modèles conditionnels qui modélisent l'intensité ; ii) les modèles marginaux qui modélisent la fonction de taux ou la fonction moyenne et iii) les modèles à fragilité qui utilisent un terme de fragilité**. Le choix de la méthode va dépendre de plusieurs paramètres tels que l'intervalle de temps, les événements passés, la fonction que l'on souhaite modéliser et de la question posée.

Dans un premier temps nous introduirons quelques notations et définitions. Ensuite nous présenterons les modèles conditionnels, marginaux et à fragilité.

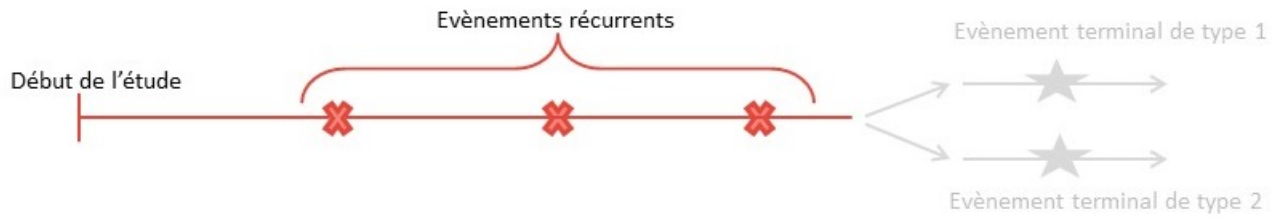


FIGURE 2.1 – Evénements récurrents

2.2 Notations et définitions

2.2.1 Notations

Soit X_{ij} le temps observé du j$^{\text{ème}}$ événement du sujet i définis par $X_{ij} = T_{ij} \wedge C_i$, $j = 1..n_i$ du sujet i $(i = 1..n)$, où $a \wedge b = \min(a, b)$, n est le nombre de sujets, n_i est le nombre d'événements récurrents pour le sujet i. T_{ij} est le temps du $j^{ème}$ événement récurrent pour le sujet i, C_i est le temps de censure qui peut être observé ou non. L'indicatrice $\delta_{ik} = \mathbb{1}_{(T_{ik} \leq C_{ik})}$ indique si le temps T_{ik} est le temps d'un événement ou de censure. Soit $N_i^R(t)$ un processus de comptage correspondant au nombre d'événements récurrents observés jusqu'au temps t pour le sujet i, $dN_i^R(t) = N_i^R((t + dt)^-) - N_i^r(t^-)$ où t^- correspond aux temps infinitésimalement plus petits que t, Z_j^R les p covariables et $Z^R = (Z_1^R, ..., Z_p^R)$ le vecteur de covariables. Le processus $Y_i(t) = \mathbb{1}_{(t<C_i)}$ indique si le sujet i est toujours suivi au temps t. La filtration $(\mathcal{H}_i(t))_{t \geq 0}$ où $\mathcal{H}_i(t) = \sigma\left\{N_i(u), Y_i(u), Z^R(u), 0 < u < t\right\}$ contient l'information accumulée jusqu'au temps t interprétée comme l'histoire du sujet i .

2.2.2 Définitions

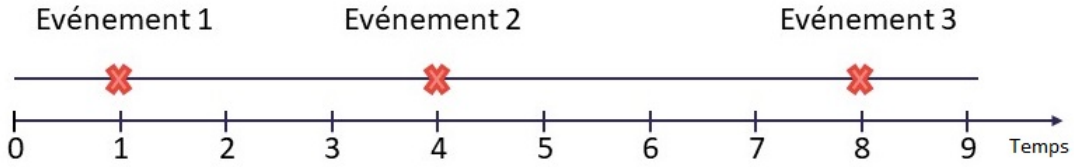
2.2.2.1 Intervalles de temps

L'intervalle de temps définit la période à risque pour chaque sujet. Il existe trois intervalles de temps différents : le *temps entre deux événements consécutifs*, l'*échelle de temps calendaire* et le l'échelle définie par le *processus de comptage* (Figure 2.2). Le choix de l'échelle de temps à utiliser n'est pas toujours évident, sachant que :

— L'intervalle **de temps entre deux événements consécutifs** se construit par une borne inférieure de 0 et la borne supérieure correspond au temps entre deux événements successifs. Cette définition de l'intervalle est utile lorsque l'on considère qu'il y a un renouveau ou que l'état du sujet après l'apparition de chaque événement revient à son état de départ. Très utilisée dans l'industrie, cette hypothèse n'est cependant que très peu probable dans le domaine médical, puisque le sujet est considéré comme "guéri" après la survenue d'un événement. Elle peut cependant être utilisée lorsque le nombre d'événements est faible.

— **L'échelle de temps calendaire** est le temps entre le début de l'étude et le temps jusqu'à l'apparition de l'événement. Comme pour l'intervalle de temps entre deux événements consécutifs, la borne inférieure est égale à 0 mais la borne supérieure est le temps entre le début de l'étude et le temps jusqu'à l'apparition de l'événement.

— Le **processus de comptage** est construit comme pour l'échelle de temps calendaire cependant, il reconnaît qu'un sujet peut entrer tardivement dans l'étude ou qu'il peut être censuré. Le sujet ne peut être à risque de faire le $j^{ème}$ événement que s'il a présenté le $(j-1)^{ème}$ événement. Cette échelle de temps se justifie lorsque la maladie évolue ou progresse au cours du temps. La borne inférieure est égale au temps d'apparition de l'événement précédent et la borne supérieure correspond au temps d'apparition de l'événement.



Evénement	Temps entre deux événements successifs	Temps calendaire	Processus de comptage
1	[0-1]	[0-1]	[0-1]
2	[0-3]	[0-4]	[1-4]
3	[0-4]	[0-8]	[4-8]

FIGURE 2.2 – Intervalles de temps

Pour le temps entre deux événements successifs, la borne inférieure de chaque intervalle est égal à 0 et la borne supérieure correspond au délai entre deux événements successifs. Pour le temps calendaire, la borne inférieure de chaque intervalle est aussi égale à 0 alors que la borne supérieure correspond au temps de survenue de l'événement. Pour le processus de comptage, la borne inférieure correpond au temps de survenue de l'événement précedent et la borne supérieure correspond au temps de survenue de l'événement.

2.2.2.2 Définition de l'ensemble à risques

L'ensemble à risque pour le $k^{ème}$ événement contient les intervalles de temps des sujets considérés comme pouvant présenter cet événement. Il existe plusieurs définitions de l'ensemble à risque, qui sont influencées par le type de risque instantané de base qui peut être commun à tous les événements ou spécifique à chaque événement :

— l'ensemble sans restriction : tous les intervalles de temps de chaque sujet peut être à risque pour n'importe quel événement quel que soit le nombre d'événements présentés par chaque sujet.

— l'ensemble restreint contient seulement les intervalles de temps pour le $k^{ème}$ événement des sujets ayant déjà présenté $k-1$ événements.

— l'ensemble semi-restreint est associé à un risque instantané de base spécifique à chaque événement c'est à dire que l'ensemble semi-restreint contient pour le $k^{ème}$ événement les sujets ayant présentés $k-1$ ou moins événements.

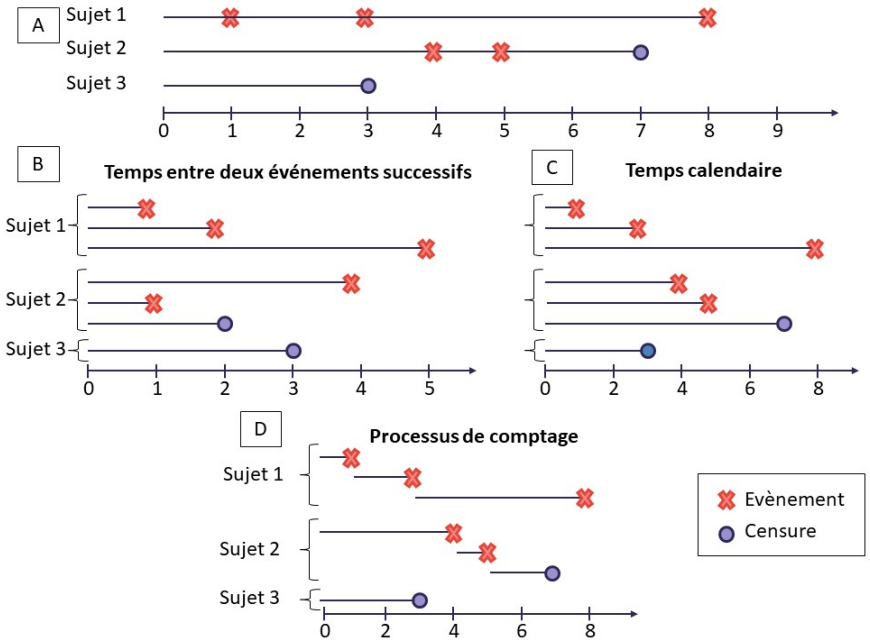


FIGURE 2.3 – Illustration des intervalles de temps

La Figure 2.3 représente l'exemple de trois sujets ayant ou non des événements selon les différentes échelles de temps. Quant à la Table 2.1, elle contient les personnes à risque de faire l'événement au temps 5 du sujet 2 pour chaque échelle de temps et selon le type d'ensemble.

2.2.2.3 Les différentes fonctions et le processus de Poisson

La fonction d'intensité

Un processus de comptage est entièrement déterminé par son intensité. Elle correspond à la probabilité instantanée de présenter un événement dans un petit intervalle de temps conditionnellement aux événements passés. La probabilité conditionnelle de présenter un événement est proportionnelle à la longueur de l'intervalle. En supposant que deux événements ne peuvent pas survenir en même temps et que le temps est continu, l'intensité donne la probabilité instantanée de survenue d'un événement au temps t sachant le passé et s'écrit comme suit :

$$r(t) = \lim_{dt \to 0} \frac{\mathbb{P}(dN^R(t) = 1 | \mathcal{H}(t))}{dt} \tag{2.1}$$

TABLE 2.1 – Exemple de personnes à risque au temps 5

	Ensemble sans restriction	Ensemble restreint	Ensemble semi-restreint
Temps entre deux événements successifs (Figure 2.3 B)	Les trois événements des sujets 1 et 2 et le premier événement du sujet 3	Le deuxième événement du sujet 1 et le deuxième événement du sujet 2.	Le deuxième événement du sujet 1, le deuxième événement du sujet 2 et le premier événement du sujet 3
Temps calendaire (Figure 2.3 C)	Le troisième événement du sujet 1, le deuxième et le troisième événement du sujet 2 et aucun pour le sujet 3.	Le deuxième événement du sujet 2	Le deuxième événement du sujet 2
Processus de comptage (Figure 2.3 D)	Le troisième pour le sujet 1 le deuxième pour le sujet 2 et aucun pour le sujet 3	Le deuxième événement du sujet 2	Le deuxième événement du sujet 2

Ensemble de personnes à risque pour l'événement survenant au temps 5 du sujet 2 (Figure 2.3 A).

où $\mathcal{H}(t)$ est l'histoire jusqu'au temps t qui inclut le nombre d'événements survenus jusqu'au temps t.

La fonction de taux

Lorsque l'intensité du processus de comptage ne dépend pas de l'histoire $H_i(t)$, cela revient au taux marginal. La fonction de taux est la probabilité instantanée marginale de survenue d'un événement au temps t et est définie par :

$$\rho(t) = \lim_{dt \to 0} \frac{\mathbb{P}(dN^R(t) = 1)}{dt} \tag{2.2}$$

Ce qui revient à $\rho(t)dt = E[dN^R(t)]$. Contrairement à l'intensité, la fonction de taux ne dépend pas des événements passés.

Fonction moyenne

La fonction moyenne $\mu(t)$ est le nombre marginal d'événements survenus jusqu'au temps t.

$$\mu(t) = E[N^R(t)] = \int_0^t \rho(s)ds \tag{2.3}$$

Le processus de Poisson

Un processus de Poisson N est un processus de comptage classique i.e. une suite de variables aléatoires réelles $(N(t))_{t \geq 0}$ telles que :

— $N(0) = 0$
— $\forall$ t ≥ 0, $N(t) \in \mathbb{N}$
— t$\rightarrow$ N(t) est croissance

En plus des propriétés des processus de comptage, le processus de Poisson, défini par son intensité $\lambda > 0$, vérifie les trois conditions suivantes :

— Les nombres d'occurrences dans des intervalles de temps disjoints sont indépendants.
— Le nombre d'occurrence dans un intervalle de temps de longueur $t \geq 0$ suit une loi de Poisson de paramètre λt.
— La probabilité qu'il y ait plus d'une occurrence dans un petit intervalle de temps est négligeable.

Un processus de Poisson est à accroissements stationnaires (aussi dit homogène) si la loi de probabilité de $N(t)-N(s)$ ne dépend ni de t ni de s mais seulement de la longueur de l'intervalle de temps $]s, t[$.

2.3 Différentes approches pour l'analyse des événements récurrents

2.3.1 Les modèles conditionnels : Le modèle d'Andersen-Gill (AG) et de Prentice, William et Peterson (PWP)

Les modèles conditionnels modélisent l'intensité et donc prennent en compte les événements passés.

Les modèles AG [45] et PWP [46] sont des extensions du modèle de Cox [42].
Le modèle AG utilise le processus de comptage comme échelle de temps et l'ensemble des personnes à risque est non restreint. L'intensité du modèle s'écrit pour l'individu i :

$$r_i(t) = Y_i(t) r_0(t) \exp\left(\beta Z_i^R(t)\right) \tag{2.4}$$

où $Y_i(t)$ indique si le sujet est à risque, $r_0(t)$ est la fonction d'intensité de base. Le modèle AG utilise un risque de base non stratifié impliquant que le risque de base d'un événement (risque de faire un événement lorsque toutes les covariables sont à zéro) est le même que ce soit le premier, deuxième...
La vraisemblance partielle du modèle est :

$$L_p(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{n_i} \left(\frac{\exp\left(\beta Z_{ik}^R(X_{ik})\right)}{\sum_{j=1}^{n} \sum_{l=1}^{n_i} Y_{jl}(X_{ik}) \exp\left(\beta Z_{jl}^R(X_{ik})\right)} \right)^{\delta_{ik}} \tag{2.5}$$

où $Y_{jl}(t) = I_{(X_{j(l-1)} < t \leq X_{jl})}$. La log-vraisemblance partielle est de la forme :

$$LL_p(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{n_i} \delta_{ik}\left(\beta Z_{ik}^R(X_{ik}) - \log\left(\sum_{j=1}^{n}\sum_{l=1}^{n_i} Y_{jl}(X_{ik})\exp\left(\beta Z_{jl}^R(X_{ik})\right)\right)\right). \tag{2.6}$$

Les paramètres sont estimés par la méthode du maximum de vraisemblance. Le modèle AG fait l'hypothèse que les événements d'un même sujet sont indépendants et que la dépendance entre les événements d'un même sujet est expliquée par les covariables. Ainsi la dépendance peut être capturée par des variables dépendantes du temps comme le nombre d'événements passés. Cependant, si cette hypothèse ne tient pas, il est possible d'utiliser une variance robuste de type sandwich qui utilise l'estimation Jackknife [47]. De ce fait, le modèle AG est approprié lorsque l'on souhaite évaluer l'effet global d'un traitement sur l'intensité des événements récurrents lorsque la dépendance entre les événements est due aux variables dépendantes du temps.

Le modèle PWP [46] est un modèle AG stratifié, ainsi le risque de base d'un événement n'est pas le même et utilise un ensemble à risque restreint. Ce modèle permet d'identifier si le traitement retarde la survenue du $k^{ème}$ événement depuis le $k-1^{ème}$ événement. Il peut aussi être utilisé avec l'intervalle de l'échelle de temps entre deux événements successifs. Le PWP est préféré à l'AG si les effets des covariables varient d'un événement à l'autre. Le modèle s'écrit :

$$r_{ij}(t) = Y_i(t) r_{0j}(t)\exp\left(\beta Z_i^R(t)\right) \tag{2.7}$$

où $j = 1..n_{max}$ représente le nombre d'événements récurrents et n_{max} correspond au plus grand nombre d'événements récurrents rencontrés par un sujet. En pratique, il faut se limiter à un nombre maximal d'événements car le nombre de sujets dans chaque strate diminue. La vraisemblance partielle du modèle est :

$$L_p(\beta) = \prod_{i=1}^{n}\prod_{k=1}^{n_i}\left(\frac{\exp\left(\beta Z_{ik}^R(X_{ik})\right)}{\sum_{j=1}^{n} Y_{jk}(X_{ik})\exp\left(\beta Z_{jk}^R(X_{ik})\right)}\right)^{\delta_{ik}} \tag{2.8}$$

où $Y_{jk}(t) = I_{(X_{j(k-1)}<t\leq X_{jk})}$, un sujet ne peut être à risque pour le $k^{ème}$ événement s'il n'a pas fait le $k-1^{ème}$ événement. Les paramètres sont estimés par la méthode du maximum de vraisemblance. Ces deux modèles ne permettent pas de prendre en compte la dépendance intra-récurrences sauf par l'inclusion de covariables dans le modèle, cependant l'utilisation de la variance robuste permet de corriger ce problème.

2.3.2 Modèles Marginaux

Ce qui différencie les modèles marginaux des modèles conditionnels est qu'ils ne prennent pas en compte l'histoire du patient et que la dépendance entre les événements n'est pas spécifiée, mêmes s'ils n'excluent pas qu'elle puisse exister. Ils modélisent alors le taux et non plus l'intensité. Il existe deux types de modèles marginaux, les extensions du modèle de Cox comme

les modèles Lee, Wei et Amato (LWA) [48] et de Wei, Lee et Weissfeld (WLW) [49], et les équations d'estimation généralisées (GEE) [50].

Le modèle LWA utilise un risque de base commun à tous les événements avec un intervalle de temps calendaire et une définition des personnes à risque non restreint et s'écrit :

$$\rho_i(t) = \rho_0(t) \exp\left(\beta Z_i^R(t)\right) \tag{2.9}$$

La vraisemblance partielle du modèle est :

$$L_p(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{n_i} \left(\frac{\exp\left(\beta Z_{ik}^R(X_{ik})\right)}{\sum_{j=1}^{n} \sum_{l=1}^{n_i} Y_{jl}(X_{ik}) \exp\left(\beta Z_{jl}^R(X_{ik})\right)} \right)^{\delta_{ik}} \tag{2.10}$$

où $Y_{jk}(t) = I_{(X_{jk} \geq t)}$, un sujet peut être à risque plusieurs fois pour un événement. De ce fait, il est préferable d'utiliser ce modèle non pas pour les événements récurrents mais pour les données en grappe (*cluster*). Les coefficients de régressions sont estimés par la maximisation de la vraisemblance partielle.

Le modèle WLW [49] utilise un risque de base spécifique pour chaque événement avec un intervalle de temps calendaire et une définition des personnes à risque semi-restreint et s'écrit :

$$\rho_i(t) = \rho_{0k}(t) \exp\left(\beta Z_i^R(t)\right) \tag{2.11}$$

La vraisemblance partielle du modèle est :

$$L_p(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{n_i} \left(\frac{\exp\left(\beta Z_{ik}^R(X_{ik})\right)}{\sum_{j=1}^{n} Y_{jk}(X_{ik}) \exp\left(\beta Z_{jk}^R(X_{ik})\right)} \right)^{\delta_{ik}} \tag{2.12}$$

où $Y_{jk}(t) = I_{(X_{jk} \geq t)}$, un sujet peut être à risque pour le $k^{ème}$ événement même s'il n'a pas fait le $(k-1)^{ème}$événement. En pratique, si par exemple le nombre maximal d'événements présentés par un sujet est cinq alors, chaque sujet sera potentiellement à risque pour les cinq événements. Dans ce cas, des données seront ajoutées dans la base afin que chaque patient ait cinq lignes. De ce fait, il est préferable d'utiliser ce modèle lorsqu'il y a plusieurs types d'événements. Les coefficients de régression sont estimés par la maximisation de la vraisemblance partielle.

Les modèles GEE peuvent être considérés comme une extension des modèles linéaires généralisés pour l'analyse des données corrélées. Ces modèles sont ainsi utiles lorsque l'on s'intéresse au nombre moyen d'événements et lorsque l'ordre des événements n'est pas important. La théorie des martingales ne s'applique plus pour ces modèles et la théorie des processus empirique

est alors utilisée. Le modèle s'écrit :

$$\log(\mu) = \beta Z^R$$

où μ est le nombre moyen d'événements récurrents. Les paramètres sont estimés en résolvant les équations d'estimations de U(β) proposés par [50].

Lors de la construction du modèle, une hypothèse concernant la structure de dépendance entre les événements récurrents au sein d'un même patient est faite, comme par exemple, l'indépendance, la dépendance non-structurée ou encore auto-régressive. La dépendance n'est alors pas modélisée mais est considérée comme un paramètre de nuisance. De ce fait, les écarts-types sont sous-estimés et peuvent amener à des conclusions erronées. Ainsi, l'estimateur de type sandwich pour obtenir une estimation robuste de la variance des paramètres et pour la construction des tests et des intervalles de confiance associés aux paramètres du modèle est souvent utilisé. Un inconvénient majeur est que le délai de survenue d'événement n'est pas pris en compte dans cette analyse.

2.3.3 Le modèle à fragilité

Les modèles précédents supposent que la corrélation entre les temps événements est expliquée par les variables incluses dans les modèles et ne font aucune hypothèse sur la structure de la corrélation. Cependant la corrélation intra-récurrence peut être due à des facteurs pertinents non mesurés ou à la dépendance entre les événements. Dans ce cas là, l'inclusion d'un terme de fragilité permet de prendre en compte ces deux paramètres [51]. Le modèle s'écrit :

$$r_i(t) = r_0(t) u_i \exp\left(\beta Z_i^R(t)\right) \tag{2.13}$$

où u_i est le terme de fragilité qui va prendre en compte l'hétérogénéité non observée due aux facteurs non observés et à la dépendance entre les événements récurrents au sein d'un même patient. En effet, l'effet aléatoire remplace les facteurs pertinents non observés qui ont un effet sur le risque de survenue des événements récurrents. Généralement, u_i suit une loi Gamma de moyenne 1 et de variance θ pour plus de flexibilité. Cependant d'autres distributions comme la loi normale peuvent être utilisées. Les coefficients sont estimés par la maximisation de la log-vraisemblance :

$$L(\beta) = \log\left(\prod_{i=1}^{n} \int_{-\infty}^{+\infty} \prod_{k=1}^{n_i} \left(r_0(t) u_i \exp\left(\beta Z_i^R(t)\right)\right)^{\delta_i(T_{ik})} \exp\left(\int_0^{T_{ik}} r_0(t) u_i y_i(t) \exp\left(\beta Z_i^R(t)\right)\right) f(u_i) du_i\right)$$

où $T_{ik} = X_{ik} \wedge C_i$ est le temps de survenue du $k^{\text{ème}}$ événement du sujet i et $y_i(t)$ indique si le sujet i est à risque au temps t. L'estimation des coefficients pour chaque événement varie lorsque le terme de fragilité est significatif.

L'inclusion d'un terme de fragilité dans le modèle va modifier l'interprétation des coefficients.

En effet, l'effet d'un facteur est conditionnel à l'effet aléatoire et fait donc l'hypothèse que l'effet aléatoire et les covariables sont indépendantes [52]. Cependant, cette hypothèse est une limite des modèles à fragilité puisque rien ne laisse supposer que les facteurs non observés sont indépendants des facteurs observés. De ce fait, l'effet aléatoire ne prend en compte que l'hétérogénéité due aux facteurs pertinents non observés et indépendants des facteurs observés.

2.4 Conclusion et discussion

Les événements récurrents peuvent être analysés selon deux méthodologies différentes : i) les méthodes de survie (AG, PWP, WLW et les modèles de fragilité) ; ii) ou par les modèles pour les données de comptages (modèles de Poisson et modèle binomial négatif) avec GEE pour prendre en compte la dépendance intra-sujet (Table 2.2). Ces modèles peuvent être classifiés en modèles conditionnels, qui modélisent la fonction d'intensité et prennent en compte les événements passés et les modèles marginaux qui modélisent la fonction de taux et les modèles à fragilité qui permettent de prendre en compte l'hétérogénéité non-observée.
Une hypothèse majeure de ces modèles est que la censure est non-informative. Or dans certains cas, un événement terminal, tel que le décès par exemple, peut stopper le processus de récurrence. Ainsi un sujet ayant présenté l'événement terminal de façon précoce aura tendance à avoir moins d'événements récurrents qu'un sujet sans événement terminal. Dans ce cas, le traiter comme une censure viole l'hypothèse de non-information et donne des résultats biaisés. De nombreuses méthodes ont été développées pour prendre en compte ce phénomène.

TABLE 2.2 – Les modèles conditionnels et marginaux

	Modèle conditionnels		Modèles marginaux		
Modèle	AG	PWP	LWA	WLW	GEE
Extension de	Cox	Cox	Cox	Cox	Modèle linéaire généralisé
Fonction modélisée	Intensité	Intensité	Taux	Taux	Nombre moyen
Intervalle de temps	Processus de comptage	Processus de comptage	Temps calendaire	Temps calendaire	
Personnes à risque	Non restreint	Restreint	Non restreint	Semi-restreint	
Interprétation	Effet global des covariables	Effet des covariables sur chaque événement	Effet global des covariables	Effet des covariables sur chaque événement	Effet des covariables sur nombre moyen
Inconvénient	Dépendance totalement expliquée par les covariables	Difficile d'estimer l'effet global des covariables	Pas de prise en compte de la dépendance entre les événements		
			Pas adapté au événements récurrents mais aux données en cluster	Pas adapté au événements récurrents mais à plusieurs types d'événements	Ne prend pas en compte le délai de survenue d'événement

Chapitre 3

Analyse des réhospitalisations : Événements récurrents en présence d'un événement terminal, revue et comparaison des différentes méthodes

Dans les chapitres précédents, nous avons décrit les différentes méthodes pour analyser la survenue de la première réhospitalisation ainsi que les méthodes pour prendre en compte la récurrence de celles-ci. Cependant, un événement terminal peut survenir. Ignorer cet événement peut produire des résultats biaisés. En effet, les patients qui présentent l'événement terminal précocement, ont moins d'événements récurrents que ceux qui le présentent tardivement. De plus, les événements récurrents et l'événement terminal peuvent être dépendants (Figure 3.1).

Cependant dans la littérature et notamment dans le domaine des réhospitalisations, l'événement terminal, par exemple le décès, est analysé indépendamment des réhospitalisations. Pourtant, de très nombreux modèles ont été développés, chacun répondant à des questions différentes, prenant en compte ou non la dépendance entre les événements récurrents et considérant l'événement terminal comme une nuisance ou comme étant un paramètre d'intérêt. Ce sont des modèles souvent complexes et donc difficiles à comprendre et à interpréter. Le choix du modèle à utiliser reste ainsi compliqué. De plus, le manque d'accessibilité aux différentes méthodes développées peut être un frein à leur utilisation car très peu de ces modèles sont implémentés dans les logiciels.

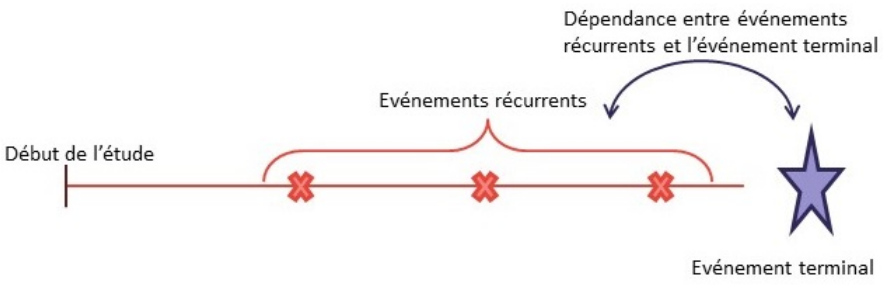


FIGURE 3.1 – Illustration des événements récurrents en présence d'un événement terminal

Nous avons donc voulu répertorier les différentes méthodes permettant d'analyser les évé-

nements récurrents en présence d'un événement terminal. Nous avons ensuite fait une étude de simulation afin d'étudier le comportement de ces modèles en fonction du nombre de sujets, de la dépendance entre les événements récurrents et la dépendance entre les événements récurrents et l'événement terminal et de produire une aide pour choisir le modèle le plus adapté à la question posée.

3.1 Article "How to analyze and interpret recurrent events data in the presence of a terminal event : An application on readmission after colorectal cancer surgery"

Received: 28 June 2018 | Revised: 27 March 2019 | Accepted: 27 March 2019

DOI: 10.1002/sim.8168

TUTORIAL IN BIOSTATISTICS

WILEY Statistics in Medicine

How to analyze and interpret recurrent events data in the presence of a terminal event: An application on readmission after colorectal cancer surgery

Anaïs Charles-Nelson[1,2,3,4] | Sandrine Katsahian[2,3,4] | Catherine Schramm[1,2,3]

[1]Sorbonne Universités, UPMC Univ Paris 06, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France

[2]INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France

[3]Université Paris Descartes, Sorbonne Paris Cité, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France

[4]Assistance Publique Hôpitaux de Paris, Hôpital Européen Georges-Pompidou, Unité d'Épidémiologie et de Recherche Clinique, INSERM, Centre d'Investigation Clinique 1418, Module Épidémiologie Clinique, Paris, France

Correspondence
Anaïs Charles-Nelson, Sorbonne Universités, UPMC Univ Paris 06, UMRS 1138, Centre de Recherche des Cordeliers, 75006 Paris, France; or INSERM, UMRS 1138, Centre de Recherche des Cordeliers, 75006 Paris, France; or Université Paris Descartes, Sorbonne Paris Cité, UMRS 1138, Centre de Recherche des Cordeliers, 75006 Paris, France; or Assistance Publique Hôpitaux de Paris, Hôpital Européen Georges-Pompidou, Unité d'Épidémiologie et de Recherche Clinique, INSERM, Centre d'Investigation Clinique 1418, Module Épidémiologie Clinique, Paris, France.
Email: anais.charles.nelson@gmail.com

Recurrent events arise when an event occurs many times for a subject. Many models have been developed to analyze these kind of data: the Andersen-Gill's model is one of them as well as the Prentice-William and the Peterson's model, the Wei Lee and Weissfeld's model, or even frailty models, all assuming an independent and noninformative censoring. However, in practice, these assumptions may be violated by the existence of a terminal event that permanently stops the recurrent process (eg, death). Indeed, a patient who experiences an early terminal event is more likely to have a lower number of recurrent events than a patient who experiences a terminal event later. Thus, ignoring terminal events in the analysis may lead to biased results. Many methods have been developed to handle terminal events. In this paper, we describe the existing methods classifying into conditional or marginal methods and compare them in a simulation study to highlight bias in results if an inappropriate method is used, when recurrent events and terminal event are correlated. In addition, we apply the different models on a real dataset to show how results should be interpreted. Finally, we provide recommendations for choosing the appropriate method for analyzing recurrent events in the presence of a terminal event.

KEYWORDS

conditional models, marginal models, recurrent events, review, terminal event

1 | INTRODUCTION

Recurrent events have become increasingly of interest in biomedical research studies since 1990s. Recurrent events may occur more than once for the same subject during a follow-up period. They can be of the same type (eg, epileptic seizures) or of different types (eg, locoregional and metastatic relapses). Various methods have been developed to analyze recurrent events and have been classified as conditional (subject-specific) or marginal models (population average).[1-5]

All models assume independent right censoring but, in practice, this assumption may be violated if a terminal event permanently stops the recurrent process (eg, the death). Indeed, the terminal event may impact the recurrent event

process, ie, a patient experiencing earlier than others the terminal event tends to have a lower number of recurrent events. Inversely, the occurrence of recurrent events may also impact the occurrence of the terminal event. Thus, ignoring terminal events in the analysis may lead to biased results. Thereby, the choice of the appropriate method to analyze recurrent events depends on four main factors: the dependence between recurrent events, the dependence between recurrent and the terminal events, the question of interest, and the biological process of the recurrent events.

Many models have been well developed to handle this type of data, and Sinha et al[6] already reviewed Bayesian methodologies, which are useful in case of small amount of data. However, it requires an assumption of the a priori distribution which may be complicated to estimate in practice. No comprehensive review is available for frequentist approaches. The conditional and marginal frequentist approaches can be split into the following: (i) nonjoint models focusing on recurrent events after taking into account terminal event, estimations are made in two steps and it supposes that recurrent events do not impact the terminal event; (ii) joint models that estimate simultaneously the recurrent and terminal events as well as their dependence supposing that both event types may have an impact on the other; and (iii) models using a composite endpoint where recurrent and terminal events are combined into a single endpoint.

The aim of this paper is to present frequentist approaches to model recurrent events in the presence of a terminal event and to provide guidelines on the use of these methods. The remainder of this article focuses on recurrent events data with known time of event occurrences and assumes that two events cannot happen at the same time (continuous time). In the next section, important notations are introduced. Conditional and marginal approaches are then described in Sections 3 and 4. A simulation study has been conducted in Section 5. Finally, models are illustrated on a real dataset concerning readmission after colorectal cancer in Section 6. A discussion and guidelines on the choice of the appropriate model are provided in Section 7.

2 | NOTATIONS

All along the article, the term terminal event refers to an event what permanently stop the process of recurrent event, like death, and dependent censoring refers to an event what precludes the observation of recurrent event. This means that a subject may have subsequent recurrent event after a dependent censoring. Dependent censoring may be the end of a study of withdrawal of subject.

Let T_{ij} be the jth observed event times of subject i defined as $T_{ij} = X_{ij} \wedge C_i \wedge D_i$ (a$\wedge$b=min(a,b)), $j = 1 \dots n_i$ and $i = 1 \dots n$, where n is the number of subjects and n_i is the number of observed recurrent events for subject i. X_{ij} is the time of the jth recurrent event for subject i, and C_i and D_i are the censoring and terminal event times respectively for the subject i. The terminal event may or may not be observed. Let denote $\Delta_i = \mathbb{1}(D_i \leq C_i)$, where $\mathbb{1}$(condition) is the indicator function taking the value of 1 if condition is true and 0 otherwise. We generally observe only the minimum times of the terminal event and censoring defined as $T_i^* = C_i \wedge D_i$. $Y_i(t) = \mathbb{1}(T_i^* \geq t)$ defines if a patient is at risk to experience a recurrence at time t, ie, if he has not been censored or experienced a terminal event yet. Let $N_i^R(t)$ be the number of recurrent events up to t and $N_i^D(t) = \mathbb{1}(D_i \leq t)$ indicating whether a terminal event occurred before t, for subject i. We denote $dN_i^R(t) = N_i^R((t+dt)^-) - N_i^R(t^-)$ where t^- denotes times infinitesimally smaller than t. Recurrent event process may continue to increment after censoring but stop after terminal event occurrence. Let Z^R and $Z^R(t)$ be the time-independent and time-varying vectors of covariates for recurrent events and Z^D, $Z^D(t)$ be the time-independent and time-varying vectors of covariates for terminal events.

Subjects in the study who have not experienced the terminal event yet are called "survivors." The term "deceased" corresponds to subjects who have experienced the terminal event.

The recurrent events process is modeled through the intensity/rate/mean function. Let denote $r(t)$ the intensity function which is the instantaneous probability of an event occurring in a short interval of time dt conditional upon the history and provides the instantaneous probability of an event occurring at time t[7] and is defined as follows:

$$r(t) = \lim_{dt \to 0} \frac{P(dN^R(t) = 1 | H(t))}{dt}, \tag{1}$$

where $H(t)$ is the history up to time t. It includes the number of events occurring up to time t. The intensity function for the terminal event is defined similarly as (1) and is noted $\lambda(t)$.

The intensity function among survivors at time t is conditional upon D and is computed only for subjects who have not experienced a terminal event at time t:

$$r_s(t) = \lim_{dt \to 0} \frac{P(dN^R(t) = 1 | H(t), D \geq t)}{dt}. \tag{2}$$

Let denote $\rho(t)$ the rate function which is the marginal instantaneous probability of an event occurring at time t and is defined as

$$\rho(t) = \lim_{dt \to 0} \frac{P(dN^R(t) = 1)}{dt}. \tag{3}$$

This leads to $\rho(t)dt = E[dN(t)]$. Contrary to the intensity function, the rate function does not depend on the event history.

The mean function $\mu(t)$ is the marginal number of events over time up to t.

$$\mu(t) = E[N(t)] = \int_0^t \rho(s)ds \tag{4}$$

In practice, recurrent events induce heterogeneity between subjects, which may be taken into account when modeling a intensity/rate/mean function. Indeed, the subject may be more frail to experience an event, even after controlling the measured covariates because some unobserved factors are unknown. One way to deal with it consists in including a frailty parameter that allows to account for unobserved heterogeneity in baseline in survival analyses. It is an unobserved random factor that models these unknown factors and modifies multiplicatively the hazard/rate/intensity function of an individual. In case of recurrent events, events of an individual are assumed to share the same frailty. These models are called "shared frailty models".

Recurrent events process and terminal event process could be analyzed by a two-step method (nonjoint) or joint modeling. The two-step method consists in modeling the terminal event process and using fitted values as weights or parameters in the recurrent events model. Obviously, this kind of method supposes that the recurrent process does not impact the occurrence of the terminal event. The joint modeling approach consists in using shared random effects and estimate parameters associated with both processes simultaneously. In survival analysis, these random effects are called frailties if they are parametric. If no assumption of distribution is made, the random effects are called latent variables.

3 | CONDITIONAL APPROACH

In medical studies, the time to the occurrence of a recurrent event may be measured in two different scales, the time between two consecutive events (eg, relapses) (gap times) or the time elapsed since the beginning of a study until an event occurs (calendar time); gap time is preferred when the interest lies in the time between two consecutive recurrent events or, when a "renewal" happens after each recurrence, the calendar scale is appropriate when the time since the beginning of the study is of interest or for making individual predictions. This implies that the order of events is important. In practice, recurrent events within the same subject are often correlated to each other. Indeed, a patient experiencing a recurrent event at time t is more prone to experience the next one than a patient with no recurrences at time t. In other words, occurrence of the kth event modifies the probability of the $k + 1$th event to occur and is modeled using intensity function (conditionally to the past). Thus, conditional approach models are suitable in these situations as they model the intensity function. Recurrent events are then considered as survival data.

Miloslavsky et al[8] has extended the nonjoint conditional Andersen-Gill[9] model for recurrent events in calendar time to assess the effect of covariates on recurrent events taking into account the existence of a terminal event. This model supposes that a subject cannot be at risk for the kth event if he/she has not experienced the $k - 1$th event. This model assumes that recurrent events do not influence the terminal event, and thus, the terminal event is then modeled separately. Moreover, the model considers that the dependence between recurrent events is completely due to observed covariates.

However, in many cases, recurrent events have an impact on the occurrence of terminal event. Then, joint models are required. Liu et al[10] developed a method to model two intensity functions, one for each endpoint and for which covariates could differ. It considers that dependence between recurrent events and terminal event processes can be due to unobserved covariates.

The first subsection describes Miloslavsky's extension of the Andersen-Gill's model and the second Liu's joint model as well as several extensions.

3.1 | Conditional nonjoint model: Miloslavsky's extension of the Andersen-Gill's model

The Miloslavsky's model is a semiparametric multiplicative proportional hazard model of the following form:

$$r(t|Z^R, f(t), Y_c(t)) = Y_c(t) r_0(t) e^{\beta Z^R(t) + f(t)}, \tag{5}$$

where $r_0(t)$ is the common baseline function for all events, $Y_c(t) = \mathbb{1}(t < C)$ is the at-risk indicator as defined in the Andersen Gill's model[9] and $f(t)$ is a time-dependent function capturing the dependence between recurrent events. For example, it may be the number of events at time t. β is the vector of regression for the covariate of interest for which the effect is assumed to be linear and constant over time. The log partial likelihood of the model is of the following form:

$$\log(L) = \int_0^\infty \log\left(Y_c(t)r_0(t)e^{\beta Z^R(t)}\right) d(N(t)) - \int_0^\infty Y_c(t)r_0(t)e^{\beta Z^R(t)}d(t). \tag{6}$$

Because the independent censoring time is not observed for subjects experiencing the terminal event inducing dependent censoring, the inverse probability weighted censoring procedure (IPCW) is used. The weighted procedure consists of artificially creating censoring time[11] for deceased subjects such as subjects experiencing a terminal event have a weight ($w_i(t)$) at time t equal to the probability of being uncensored at time t given that they were uncensored when they experienced the terminal event. Thus, the weight decreases with time and subjects who experience the terminal event remain in the risk set but their contribution decreases with the increasing value of t. This procedure relaxes the assumption of independent censoring and death is considered as a competing risk. Thus, the dependence between recurrent events and the terminal event is treated as a nuisance parameter. Even though it relaxes the noninformative censoring assumption, the procedure requires an assumption on the censoring distribution. For example, the Cox model with censoring as event of interest (alternative models such as accelerated failure models) may be considered to model censoring:

$$\lambda^C(t|Z^C(t)) = \lambda_0^C(t)e^{\gamma_C Z^C(t)}, \tag{7}$$

where $Z^C(t)$ are covariates associated with the censoring time which may be different from Z^R and Z^D. The weights are then estimated as $w_i(t) = \mathbb{1}(C_i \geq D_i \wedge t)\frac{\hat{G}(t|Z_i^C(t))}{\hat{G}(D_i \wedge C_i \wedge t)}$, where $\hat{G}(t|Z^C(t)) = \int_0^t \hat{\lambda}_0^C(t)e^{\hat{\gamma}_C Z^C(t)}$. $\hat{\lambda}_0^C(t)$ is the Breslow estimator of $\lambda_0^C(t)$ and $\hat{\gamma}_C$ is the maximum partial likelihood estimators of γ_C. Weights are then incorporated in the score function as follows:

$$U(\beta) = \int_0^\tau Z_i^R(t) - \frac{\sum_{j=1}^n w_i(t)Z_j^R(t)e^{\beta Z_j^R(t)}}{\sum_{j=1}^n w_i(t)e^{\beta Z_j^R(t)}} w_i(t)dN_i(t), \tag{8}$$

where τ is the end of the study time. $w_i(t)$ decrease with increasing values of t, echoing that they are less and less likely to remain uncensored as time goes by and hence contribute less and less to the score function.

According to this estimation procedure, coefficients are interpreted as the rate of events among survivors and deceased patients, ie, they are interpreted as the average number of events over time.

Only one type of recurrent events and only one type of terminal event are allowed in the model. This model also assumes that the dependence between recurrent event times is completely explained by the measured covariates. This implies that the time increments between events are conditionally uncorrelated given the observed covariates. In practice, an appropriate time-dependent function ($f(t)$) that captures the dependence between recurrent events can be included in the model, for example, the number of events at time t. If the assumption of a time-dependent function $f(t)$ that completely captures the dependence between recurrent events does not hold, a robust sandwich covariance matrix can be used to avoid estimation bias. It provides robust standard errors for the regression coefficients.[12]

3.2 | Conditional joint model: Liu's joint frailty model

Assuming that recurrent events do not influence the occurrence of the terminal event that does not hold in many cases. Then, it is necessary to model conjointly both processes. The Liu's model[10] is a joint model for both recurrent events and terminal events. The two processes as well as their dependence are evaluated at once. Both are semiparametric multiplicative proportional hazard models which share a frailty taking into account intrarecurrence dependence and dependence between recurrent events and the terminal event for each subject. It takes the following form:

$$\begin{cases} r(t|a, Z^R(t)) = ar_0(t)e^{\beta Z^R(t)} \\ \lambda(t|a, Z^D(t)) = a^\gamma \lambda_0(t)e^{\alpha Z^D(t)}, \end{cases} \tag{9}$$

where $r(t)$ and $r_0(t)$ (respectively, $\lambda(t)$ and $\lambda_0(t)$) are intensity and baseline intensity functions for recurrent events (respectively, terminal event); $Z^R(t)$ and β (respectively, $Z^D(t)$and α) are covariates and their regression coefficients for recurrent (respectively, terminal) event(s) process. α and β are assumed to be linear, ie, effect of covariates increases/decreases for a α one unit change of $Z^R(t)$ (respectively, $Z^D(t)$) and is constant over time (effect of covariates does not vary over time) and

they have a log-linear effect on the intensity function. The model allows only one type of recurrent event and one type of terminal event but can account for both gap time and calendar time. It allows different sets of covariates for recurrent and terminal events processes and can include time-dependent covariates. a_i is the frailty parameter associated with subject i and follows a Gamma distribution with mean 1 and variance θ. In practice, it is complicated to check the distribution of the frailty even though a violation of the distribution may provide bias results. a_i reflects the dependence between recurrent and terminal events as well as intrarecurrences dependence due to unobserved variables and γ indicates the direction of the dependence between recurrent and terminal events (positively or negatively correlated). $\gamma > 0$ signifies that the higher the risk of recurrence, the higher the risk of the terminal event. In contrast, $\gamma < 0$ corresponds to a negative correlation and indicates that the lower the risk of recurrence the higher the risk of the terminal event. $\gamma = 0$ implies that $\lambda(t)$ is noninformative for the recurrent events rate, ie, recurrent and terminal events are conditionally independent of covariates. Note that γ can only be interpreted if θ is significantly different from zero in other words if the times of recurrent and terminal events are significantly correlated. Furthermore, the model assumes that recurrent events and the terminal event are not independent, even conditional upon frailty, meaning that the intensity function of recurrent events becomes 0 after the terminal event, assuming that experiencing the terminal event stops further occurrences of recurrent events and that censoring does not. This assumption reflects the real life. Unlike the Miloslavsky's model, censoring times are considered to be independent. Indeed, the distribution of censoring times is required to be independent of frailty and of recurrent and terminal events. Parameters are estimated from the conditional likelihood for subject i is

$$L_i = exp^{-\int_0^\infty Y_i(t)a_i e^{\beta Z_i^R(t)} dr_0(t)} \prod_j \left[a_i e^{\beta Z_i^R(t)} dr_0(t_{ij})\right]^{\delta_{ij}} e^{-\int_0^\infty Y_i(t)a_i^\gamma e^{\alpha Z_i^D(t)} \lambda_0(t)} \times \left[a_i^\gamma e^{\alpha Z_i^D(t)} d\lambda_0(T^*)\right]^{\Delta_i}. \quad (10)$$

Baseline intensity for the recurrences and the terminal event take the form of Breslow's estimate.[13] Parameters β and α are estimated by maximizing the likelihood through a Monte Carlo expectation-maximization algorithm with the Metropolis-Hasting sampler in the E-step. Rondeau et al[14] have proposed an alternative method of estimation using the maximum penalized likelihood. This method has the advantage of being less time consuming and is appropriate for nonparametric estimations of continuous baseline hazard function (eg, spline), whereas the Breslow estimator provides a piecewise-constant baseline hazard function. Moreover, this estimation approach is available in the R package "frailtypack."[15] Coefficients have to be interpreted with caution. Indeed, coefficients correspond to the intensity among survivors, conditionally upon the frailty. The main differences between the Liu's and the Miloslavsky's models are the following: (i) the intensity function is conditional upon the patient being alive; (ii) the dependence between recurrent and terminal events is estimated instead of being considered as a nuisance parameter; and (iii) the effect of treatment on the terminal event is estimated taking into account recurrent events, whereas in the previous model, the terminal event is modeled independently of recurrent events.

Extensions of the Liu's shared frailty model:

Zhangsheng and Liu[16] proposed a model in which the effects of covariates was nonparametric allowing nonlinear and unspecified effects. Liu et al[17] developed a model to account for zero-inflated recurrences (high number of patients with no recurrence) combining a cure model[18] and Liu's model.[10] This model is suitable for medical studies looking to evaluate whether treatment will result in a higher fraction of cured subjects (ie, patients with no recurrence at all). Mazroui et al[19] extended the joint frailty model to two types of recurrent events. Two frailties are included in the model, one for each recurrent events process, each taking into account both the dependence of recurrent events of the same type and their dependence with the terminal event. Belot et al[20] developed a model that includes excess mortality to separate death due to the disease of interest from death due to other causes. Mazroui et al[21] proposed a shared frailty model that includes two Gamma-distributed frailties (u_i and v_i) to distinguish dependence within recurrences (u_i) from dependence between recurrent and terminal events (v_i). In contrast, both types of dependence are taken into account in Liu's model but are included in a single frailty. Only the models of Mazroui et al[21] have been implemented in the R software within the "frailtypack" package.

The models described above assume that recurrent and terminal events are not independent even conditional upon frailty. Thus, the intensity is equal to 0 after the terminal event occurs. In contrast, the models developed by Huang and Liu,[22,23] and Zeng and Lin[24] assume that conditional upon frailty recurrent events and terminal event are independent implying that the intensity function is not equal to 0 after the terminal event has occurred. Huang and Liu[22] developed a stratified model based on gap time. This model assesses the effect of treatment on the ith gap time rather than the overall effect by stratification. It is useful if the question of interest is the distribution of time to the next event. Liu and Huang[23] developed a joint frailty model to simultaneously analyze recurrent events, the terminal event, and repeated measurements (eg, measure of blood pressure at different time points). Zeng and Lin[24] proposed a class of transformation

function, including proportional hazard/intensity and proportional odds models. In practice, however, this hypothesis corresponds to a nonnatural situation. Thus, the terminal event is considered to be a dependent censoring.

None of these models have been implemented in any software.

4 | MARGINAL APPROACH

The intensity function is a difficult quantity to understand. Then, in practice, the number of events is preferred by physicians. The interest in the number of events implies that the event's past history and the order of events are ignored. Thus, this outcome is favorised for identifying treatment effects or risk factors. However, it is very complex to model this quantity from a conditional model. Thus, marginal models are highly applicable in these situations.

Indeed, marginal approaches model the probability to experience an event at time t unconditionally from previous events (rate function). Thus, in these models, the intrarecurrence dependence is not specified. However, it is possible to take this into account using a sandwich estimator of the variance-covariance matrix. The sandwich estimator does not require specification of the correlation structure among observations.

Similarly to the Miloslavsky's model, Ghosh and Lin[25] studied the mean function when recurrent events process does not influence the terminal event and proposed two approaches to analyze it: (i) among survivors and deceased patients, ie, in the hypothetical case where there is no terminal event that stops the recurrence, and (ii) among survivors. However, when recurrent events influence the terminal event, joint models are preferred and were developed by Huang and Wang[26] for the marginal approach.

However, in many situation, the total number of events may be very low. Thus, it may be reasonable to combine recurrent and terminal events in a single outcome to increase the power. Mao and Lin[27] developed a model to deal with this combined endpoint. The models of Gosh, Huang, and Mao are all detailed as follows.

4.1 | Marginal nonjoint model: Ghosh's model

Ghosh and Lin[25] proposed a semiparametric proportional rate model for recurrent events of the following form:

$$\rho(t|Z^R(t)) = \rho_0(t)e^{\beta Z^R(t)}. \tag{11}$$

Time-varying covariates are allowed in this model. The terminal event is taken into account in the estimation procedure by using weighting procedures. The score function is as follows:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} Z_i^R(t) - \frac{\sum_{j=1}^{n} w_i(t) Z_j^R(t) e^{\beta Z_j^R(t)}}{\sum_{j=1}^{n} w_i(t) e^{\beta Z_j^R(t)}} w_i(t) dN_i(t), \tag{12}$$

where $w_i(t)$ represents the weight of subject i at time t. The way in which the terminal event is taken into account depends on the method used in the weighting computation. The first method uses the inverse probability of censoring weighting (IPCW) as described in Section 3.1 and, the second, the inverse probability of survival weighting (IPSW) described in Section 4.1.2.

4.1.1 | Among survivors and deceased patients: IPCW method

The IPCW procedure is computed in the same way as in Miloslavky's model presented in Section 3.1 to calculate weights and is incorporated into the model (Equation (12)). This procedure uses a nonparametric estimator (eg, Kaplan-Meier estimator) and is more suitable if recurrent events are of primary interest and censoring is independent from covariates.[25]

Coefficients are interpreted as the average number of events among survivors and deceased patients. In other words, they can be interpreted as the average number of events over time in the hypothetical case where no terminal event stops the occurrence of events.

4.1.2 | Among surviving patients (also called partial marginal): IPSW method

The IPSW method requires modeling the terminal event using the Cox model (or an alternative method such as accelerated failure time model). This procedure is preferred when survival is also of interest

$$\lambda^D(t|Z^D(t)) = \lambda_0^D(t)e^{\gamma_D Z^D(t)}, \tag{13}$$

where $\hat{\lambda}_0^D(t)$ is Breslow estimator of $\lambda_0^D(t)$ and $\hat{\gamma}_D$ is maximum partial likelihood estimators of γ_D. The weights are then defined as $w_i(t) = \frac{\mathbb{1}(D_i>t)}{S(t|Z_i^D(t))}$, where $S(t|Z_i^D(t)) = P(D_i \geq t|Z_i^D(t)) = exp(-\int_0^t \lambda^D(t|Z^D(t))dt)$. Contrary to the IPCW method, the risk set for recurrent events in the IPSW method only includes alive patients. Thus, coefficients should be interpreted as the estimated average number of events among survivors.

Ghosh and Lin[25] proposed a multiplicative semiparametric model with a time-invariant coefficient. However, additive models are more appropriate when there are continuous covariates. Zhao et al[28] developed an additive semiparametric model using the IPSW method. In an additive model, coefficients are expressed as the absolute mean difference. Zhao et al proposed semiparametric transformation models with time-varying coefficients.[29] Additive and proportional model can be modeled with the identity or exponential transformations. Cook and Lawless[30] and Lin et al[31] described two alternative manners to assess the overall effect of treatment among survivors. The first one consists of performing a separate model among patients who have survived up to a prespecified time point. The second one uses a landmark analysis what is quite similar than the method proposed by Cook. Indeed, landmark analysis consists in performing a separate model on patient still at risk at each chosen time points.

A variation of these approaches is the use of a latent parameter to account for dependence between the two processes. It supposes that this dependence is due to unmeasurable variables. Wang et al[32] have included a latent parameter in a proportional model, Sun and Kang[33] in an additive-multiplicative rate model, and Zhao et al[34] in an additive model that takes into account excess of zero when many patients experience no recurrent events. The coefficients are interpreted conditionally to the latent parameter in these models.

4.2 | Marginal joint model: Huang's model

Huang's model[26] uses two semiparametric multiplicative proportional models, one for recurrent events and one for the terminal event, which both share the same latent parameter that takes into account the correlation between recurrent events and the terminal event for each subject. The model is of the following form:

$$\begin{cases} \rho(t|a, Z^R) = E[dN(t)|Z^R, a] = a\rho_0(t)e^{\beta Z^R} \\ \lambda(t|a, Z^D) = a\lambda_0(t)e^{\alpha Z^D} \end{cases}, \tag{14}$$

where $\rho(t)$ and $\lambda(t)$ are the rate functions for recurrent events and the hazard function for terminal event, respectively. $\rho_0(t)$ and $\lambda_0(t)$ are the common baseline rate and hazard functions for recurrent and terminal events, respectively, Z^R and Z^D are covariates associated with the recurrent event process (respectively, the terminal event process). β and α are assumed to be linear (the effect of the covariate increases/decreases of α for one unit change of Z^R (respectively, Z^D) and constant over time (the effect of the covariates does not vary over time) and have a log-linear effect on the rate function. a_i is the nonparametric frailty (latent) parameter associated with subject i and is considered to be a nuisance parameter. It reflects the dependence between recurrent and terminal events, as well as intrarecurrences dependence due to unobservable variables. The model assumes that, conditionally upon frailty, recurrent and terminal events are independent. Thus, the rate function of the recurrent events continues to increment after the terminal event. This implies that the assumption of independent censoring is relaxed both for recurrent and terminal events. Huang's model allows only one type of recurrent event and one type of terminal event. It does not deal with time-dependent covariates and only positive correlation between recurrent and terminal events is considered. However, it allows for different sets of covariates for recurrent and terminal events. Estimation of the regression parameters is a three-step estimation procedure called the borrow-strength method (see the work of Huang and Wang[26] for more details). Coefficients are interpreted as rate and hazard function respectively for recurrent and terminal events conditionally upon the latent parameter.

This latter model has been extended to time-dependent covariates by Huang et al.[35] Models with and without time-dependent covariates have been extended for different types of recurrent events.[36,37] Ye et al[38] proposed a joint model which assumes that frailty is Gamma distributed. Contrary to Huang's model, Ye's model assumes that even conditional on the frailty recurrent and terminal events is dependent leading to a modeling of the mean function among survivors only. Moreover, a Gamma distributed frailty parameter is included in the model and dependence between both processes can be estimated.

Chen et al[39] proposed a joint frailty model based on Ye's idea, for multiple types of recurrent events (K). $K + 1$ independent Gamma distributed frailties are included in the proposed model to differentiate interrecurrence dependence from the dependence between recurrences and terminal event. These models are unable to detect time-varying effects over time that is why Chen et al[40] attempted to address this issue by extending the partly Aalen model[41] to analyze recurrent

and terminal events. In this model, recurrent events are modeled through an additive partly Aalen model and the terminal event via a proportional hazard model with an unspecified baseline hazard. Parameters for recurrent events are interpreted as the absolute effect of the covariate, which is the expected change in the rate of events at time t for a unit change in the covariate given the frailty.

None of these models are available in any software.

4.3 | Combined endpoint: Mao's model

Mao and Lin[27] proposed a semiparametric proportional mean model in which recurrent and terminal events are combined in a unique endpoint weighted by their degree of severity. The model has the following form:

$$\rho(t|Z(t))dt = E[dN^*(t)|Z(t)] = \rho_0(t)e^{\beta Z(t)}, \tag{15}$$

where $Z(t)$ are the covariates associated to the combined endpoint which may be $Z^R(t)$ and/or $Z^D(t)$ or different from $Z^R(t)$ and $Z^D(t)$. In this model, $N^*(t)$ indicates the total number of events between 0 and t and is equal to $N^*(t) = \sum_{k=1}^{K} c_k N_k^*(t)$ where K represents different types of events including the terminal event and $N_k^*(t)$ denotes the cumulative number of events of the kth type at time t. c_k is the weight associated with each type of event according to their severity, the more severe the event, the higher the weight. The weight should be chosen *a priori* by the clinician. Mao et al indicated that one may choose the weights in a data-adaptive manner such that the weight for the terminal event depends on how many patients have experienced the terminal event. Censoring is assumed to be independent of recurrent and terminal events conditional on covariates. Coefficients are assumed to be linear and constant over time but the model can deal with time-varying covariates. Parameters are estimated using the weighting score function with an IPCW technique. Thus, deceased patients remain in the risk set. The score function is

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} Z_i(t) - \frac{\sum_{j=1}^{n} w_j(t) Z_j(t) e^{\beta Z_j(t)}}{\sum_{j=1}^{n} w_j(t) e^{\beta Z_j(t)}} dN_i(t), \tag{16}$$

where $w_j(t)$ are the weights of subject j according to the IPCW method and τ denotes the end of the study. The model does not require the Poisson assumption for recurrent events. The baseline mean function is estimated using a weighted version of the Breslow estimator.[13]

Mao et al estimate the overall effect of treatment, but it may be of interest to estimate the effect of the treatment on the kth event. Li and Lagakos[42] proposed to address the terminal event in the marginal approach of Wei et al[43] by considering the failure time for each recurrence as the first occurrence of the recurrent events or the terminal event, whichever came first. Contrary to Mao's model, the latter approach estimates the treatment effect among survivors only.

5 | SIMULATION STUDY

5.1 | Generating data

Simulation studies were performed to assess properties of the models described in section (III) and (IV). Event times were generated from the following models:

$$\begin{cases} r_i(t|u_{i1}, u_{i2}) = u_{i1}u_{i2}r_0(t)e^{(\beta_1 Z_1 + \beta_2 Z_2)} & (1) \\ \lambda_1(t|u_{i1}) = u_{i1}\lambda_0(t)e^{(\beta_3 Z_1)} & (2), \end{cases}$$

where Z_1 and Z_2 are binary covariates and were generated from a Bernoulli distribution with a probability of 0.5. We set $\beta_1 = 1$, $\beta_2 = -0.5$ and considered constant baseline functions for both processes with $\lambda_0(t) = 1$ and $r_0(t) = 2$.

u_1 and u_2 are two frailty parameters representing the dependence between recurrences and the terminal event and the intrarecurrence dependence, respectively. They both follow a Gamma distribution with mean 1 and variance θ_1 and θ_2 respectively. The higher the value of θ_1 (θ_2), the stronger the correlation (larger the intersubjects heterogeneity) according to event recurrences and terminal events. For each patient, data were simulated in two steps:

- The first step is the generation of the time to terminal events from an exponential distribution with intensity $\lambda_i(t|u_{i1})$. The time for each terminal event is generated from an exponential distribution under model (2). Independent censoring time follows a uniform distribution from 0 up to 2. The terminal event time is defined as the minimum of censoring and terminal event time.

- The second step is generating recurrent events times: gap times are generated from an exponential distribution with intensity $r_i(t|u_{i1}, u_{i2})$. Total time for the jth event is defined as the sum of the $k = 1$ to $j - 1$ gap time. If the total time for the jth event do not exceed the terminal event time, we observed a recurrent event. If the total time for the jth event exceeds the terminal event time, then the event was not observed and was censored at the terminal event time. We set up a maximum number of events of 8.

Two different settings have been simulated for the terminal event. The first one considered that the terminal event and recurrent events are both associated to the covariate Z_1, we set $\beta_3 = 0.7$. In the second setting $\beta_3 = 0$, this means that the covariate Z_1 is associated to recurrent events only.

For each setting, we considered different value of $\theta_1(\theta_1 = 0, \theta_1 = 1, \theta_1 = 5)$. A value of 0 for θ_1 means that recurrences and the terminal event are independent. Similarly, we considered different value for $\theta_2(\theta_2 = 0, \theta_2 = 1, \theta_2 = 5)$. A value of 0 for θ_2 means that recurrences are not correlated to each other for the same subject.

The sample size n was set to 50, 100, and 500. For each setting, 500 simulations were generated. We have then applied the Miloslavsky's model, the Liu's model using penalized likelihood estimation, the Ghosh's model, the Huang's model, and the Mao's model on the simulated databases. For the Miloslavsky's model, we have performed a model including the cumulative number of recurrent events as a time-dependent function ($f(t)$) to account for the intrarecurrence dependence and a model without time-dependent variable. For the IPCW method, censoring time was modeled using proportional hazard model including only an intercept term, assuming that the censoring time is not dependent on any covariates. The model was fitted using last recording for each patient (censure or terminal event). Censoring was considered as the variable of interest and thus terminal event was considered as the censoring. For the IPSW method, the terminal event is also modeled using a proportional hazard model including Z_1 as a covariate.

5.2 | Simulation results

The mean number of events per subject and the percentage of subject experiencing a terminal events are presented in Table 1. Table 2 shows the convergence rate based on the first 500 simulated data for each approach and each scenario. Some convergence issues were encountered when applying the Liu's model probably due to the fact that Liu's model is parsimonious (the percentage of convergence was from 22% to 99%); see Table 2. The number of iteration was 1000 and the number of simulations was increased to 1000. The first 500 convergent models were used. In contrast, the other models reached the convergence criterion for all settings.

Results for the first setting ($\beta_3 = 0.7$) are presented in Table 3 for conditional approaches and in Table 4 for the marginal approaches.

Table 3 shows that the Miloslasvky's model seems to provide strong biased results in all scenarios. It can be due to the misspecification of how the intrarecurrence dependence is accounted for. Indeed, in the Miloslavsky's model, intrarecurrence dependence is taken into account by the cumulative number of event at each time (time-dependent covariate), while a random effect is used in the simulation of data. Indeed, the time dependent covariates seem to not capture well the intrarecurrence dependence. Moreover, in the simulations, the covariate Z_1 is associated with both recurrences and the terminal event. However, the Miloslavsky's model only modeled the censoring distribution and do not account for the association between the covariate Z_1 and the terminal event. This means that if a covariate is associated with both processes, the Miloslavsky's model will provide strong biased results. Thus, the Miloslavsky's model seems not appropriated when recurrent and terminal events are associated to the same covariates.

TABLE 1 Results of the simulation study: description of the simulated data

θ_1	θ_2	Mean Number of Recurrent Events Per Subject	Percentage of Subjects Experiencing the Terminal Event
0	0	1.275	66.6%
	1	1.277	66.7%
	5	1.277	66.6%
1	0	1.013	52.7%
	1	1.022	52.7%
	5	1.046	53.1%
5	0	0.603	31.1%
	1	0.589	31.3%
	5	0.591	31.4%

θ_1 and θ_2 represent the dependence between the terminal event and the recurrent events and the intrarecurrence dependence, respectively.

TABLE 2 Convergence rate of each model based on 500 simulated data

n	θ_1	θ_2	Liu* ($\beta_3 = 0.7$)	Liu* ($\beta_3 = 0$)
50	0	0	84%	76%
		1	85%	78%
		5	82%	76%
	1	0	72%	60%
		1	71%	63%
		5	67%	60%
	5	0	62%	55%
		1	65%	60%
		5	63%	57%
100	0	0	80%	75%
		1	82%	80%
		5	84%	79%
	1	0	53%	41%
		1	55%	41%
		5	62%	51%
	5	0	78%	69%
		1	79%	70%
		5	82%	66%
500	0	0	47%	27%
		1	62%	52%
		5	66%	55%
	1	0	60%	22%
		1	63%	27%
		5	61%	35%
	5	0	77%	99%
		1	72%	97%
		5	78%	91%

θ_1 and θ_2 represent the dependence between the terminal event and the recurrent events and the intrarecurrence dependence, respectively.
*The maximum number of iteration was 1000 for the Liu model.

When recurrent events are independent as well as recurrent and terminal events, ie, $\theta_1 = \theta_2 = 0$, the Ghosh's, the Huang's (Table 4), and the Liu's models provide similar results in term of bias. This shows that joint models work well even when the recurrences and the terminal events are independent. We can see that the θ parameter in the Liu's model is quite close to the value 0 when the terminal and recurrences are independent ($\theta_1 = 0$). However, the Ghosh's model provide lower coverage probability than joint models in this case but it is more parsimonious than the joint models. However, the higher the dependence between recurrent and terminal events, the more biased will the results from the Ghosh's model be. Indeed, in the Ghosh's model, the terminal event is modeled independently from the recurrent events. It first models the terminal event as an independent model and then includes weights in the model for recurrences. Thus, it does not account for the dependence between these two processes. However, when the dependence between the terminal event and recurrences is strong, joint models are preferred. The higher the intrarecurrence dependence, the more appropriate will the Liu's model be compared to the Huang's model for the sample size $n = 50, 100$. This is due to the fact that the Liu's model accounts for the intrarecurrence dependence and the event's past history. This can be due to the estimation method for the Huang's model. Indeed, it first estimates parameters for the recurrences without taking into account the intrarecurrence dependence, then it estimates the latent variable and includes the value of the latent variable in the model for the terminal event. Furthermore, the Huang's model supposes that the recurrences follow a nonhomogeneous Poisson process but recurrences were simulated following an exponential distribution. Thus, misspecification of the recurrences distribution may also biased the results. However, the coverage probability for the Liu's model is lower than the Huang's model but remains reasonable. When the sample size is high ($n = 500$), the Huang's model seems better than the Liu's model.

The Mao's model provides biased results for all the scenarios. Indeed, the recurrent and terminal events are combined in a single endpoint.

Results for the second setting ($\beta_3 = 0$) are presented in Table 5 for conditional approaches and in Table 6 for the marginal approaches.

TABLE 3 Results of the simulation study: conditional approach for the first setting ($\beta_3 = 0.7$)

			$\beta_1 = 1, \beta_2 = -0.5, \beta_3 = 0.7$															
			Miloslavsky With Time-Dependent Covariate				Miloslavsky Without Time-Dependent Covariate				Liu							
n	θ_1	θ_2	β_1(SE)	CP	β_2(SE)	CP	β_1(SE)	CP	β_2(SE)	CP	β_1(SE)	CP	β_2(SE)	CP	β_3(SE)	CP	θ(SE)	α(SE)
50	0	0	0.16 (0.23)	5.6%	−0.18 (0.23)	65.1%	0.56 (0.36)	51.2%	−0.49 (0.37)	76%	0.94 (0.33)	62.8%	−0.49 (0.3)	70.6%	0.56 (0.54)	71.8%	0.05 (0.12)	6.65 (4.7)
		1	0.16 (0.34)	17%	−0.21 (0.32)	67.4%	0.55 (0.52)	52.2%	−0.48 (0.56)	63.8%	1 (0.44)	81.4%	−0.47 (0.48)	82.8%	0.78 (0.47)	78.2%	0.71 (0.3)	0.47 (2.01)
		5	0.26 (0.98)	35.6%	−0.25 (0.59)	64.2%	0.67 (1.34)	47.8%	−0.5 (0.94)	52%	1.07 (0.92)	64.4%	−0.45 (0.79)	77.2%	0.8 (0.41)	81.4%	1.33 (0.2)	0.1 (0.23)
50	1	0	0.2 (0.28)	16.4%	−0.22 (0.28)	75.2%	0.56 (0.42)	61.4%	−0.49 (0.45)	81.4%	1.02 (0.45)	86.2%	−0.47 (0.43)	89.6%	0.82 (0.81)	79.2%	0.78 (0.22)	1.54 (1.21)
		1	0.19 (0.95)	24.7%	−0.24 (0.41)	73%	0.52 (1.07)	58%	−0.49 (0.66)	65%	0.99 (1.01)	79.8%	−0.47 (0.6)	81%	0.73 (0.66)	81.6%	1.06 (0.21)	0.82 (0.86)
		5	0.28 (1.34)	38.4%	−0.39 (1.56)	70.6%	0.62 (1.7)	47.6%	−0.68 (1.92)	51.8%	0.95 (1.44)	68.8%	−0.57 (2.19)	69.8%	0.68 (0.6)	80.6%	1.4 (0.29)	0.53 (0.66)
50	5	0	0.32 (1.51)	42%	−0.25 (0.45)	88%	0.69 (1.64)	69.2%	−0.6 (0.7)	78.2%	0.88 (0.91)	78.2%	−0.6 (0.68)	84%	0.85 (1.33)	85.4%	1.3 (0.16)	2.39 (1.53)
		1	0.18 (1.31)	50.2%	−0.24 (2.18)	81.8%	0.46 (1.59)	59.8%	−0.5 (2.53)	67.8%	0.86 (1.29)	74.6%	−0.53 (0.94)	74.6%	0.7 (1.14)	87.2%	1.36 (0.21)	1.63 (1.24)
		5	1.07 (4.2)	52.1%	−0.41 (3.48)	75.55%	1.58 (5.56)	53%	−0.7 (4.38)	57.2%	0.51 (3.08)	62%	−0.97 (3.44)	65%	0.65 (0.95)	85.2%	1.47 (0.34)	1.22 (1.14)
100	0	0	0.17 (0.18)	1%	−0.18 (0.15)	41.6%	0.58 (0.25)	31%	−0.5 (0.25)	79.2%	0.97 (0.24)	63.8%	−0.5 (0.21)	71.6%	0.54 (0.45)	71.2%	0.03 (0.07)	6.9 (4.09)
		1	0.18 (0.28)	7%	−0.21 (0.23)	50.2%	0.56 (0.39)	38.4%	−0.47 (0.36)	66.4%	1.04 (0.32)	83.8%	−0.44 (0.31)	87%	0.76 (0.29)	79.8%	0.81 (0.17)	0.06 (0.43)
		5	0.21 (0.44)	18%	−0.22 (0.44)	53.6%	0.62 (0.63)	43.8%	−0.46 (0.69)	43%	1.06 (0.52)	71.8%	−0.43 (0.55)	71.2%	0.77 (0.27)	84.2%	1.39 (0.16)	0.13 (0.42)
100	1	0	0.23 (0.22)	6.61%	−0.22 (0.19)	60.72%	0.59 (0.32)	43.6%	−0.49 (0.3)	80.8%	1.03 (0.35)	84.2%	−0.47 (0.27)	92.6%	0.84 (0.54)	81.4%	0.83 (0.14)	1.29 (0.68)
		1	0.27 (0.3)	14%	−0.24 (0.28)	62.6%	0.58 (0.41)	49%	−0.5 (0.42)	68.2%	1.07 (0.44)	79.6%	−0.49 (0.43)	82.6%	0.73 (0.34)	85.8%	1.13 (0.13)	0.66 (0.28)
		5	0.29 (0.5)	27.2%	−0.3 (0.5)	61.4%	0.62 (0.73)	47%	−0.53 (0.77)	43%	0.98 (0.65)	67.6%	−0.51 (1.02)	65.6%	0.67 (0.33)	83.6%	1.52 (0.2)	0.39 (0.17)
100	5	0	0.22 (0.31)	22.4%	−0.2 (0.3)	72.6%	0.57 (0.46)	59.2%	−0.48 (0.47)	78%	0.89 (0.56)	78.8%	−0.5 (0.45)	85.4%	0.83 (0.82)	90.4%	1.33 (0.11)	1.96 (1.06)
		1	0.28 (0.4)	31%	−0.23 (0.41)	70.2%	0.58 (0.61)	58.4%	−0.47 (0.63)	65.6%	0.87 (1.27)	74.8%	−0.45 (0.65)	72.4%	0.66 (0.65)	89.6%	1.41 (0.17)	1.29 (0.71)
		5	0.35 (0.73)	43.6%	−0.34 (1.04)	67.4%	0.62 (1.04)	49%	−0.58 (1.36)	51.4%	0.89 (1.03)	60.2%	−0.5 (1.02)	58.6%	0.54 (0.52)	87.4%	1.58 (0.29)	0.83 (0.41)
500	0	0	0.19 (0.08)	0%	−0.19 (0.07)	1.42%	0.59 (0.11)	0.4%	−0.5 (0.11)	77.8%	0.95 (0.16)	80%	−0.46 (0.18)	65.65%	0.49 (0.38)	69.57%	0.04 (0.15)	6.7 (3.48)
		1	0.24 (0.14)	0%	−0.24 (0.13)	21%	0.58 (0.16)	4.08%	−0.49 (0.17)	63.47%	1.02 (0.15)	84.2%	−0.45 (0.19)	86.8%	0.74 (0.12)	80.2%	0.86 (0.06)	0.03 (0.16)
		5	0.25 (0.25)	2.2%	−0.27 (0.22)	33.8%	0.57 (0.29)	18.4%	−0.49 (0.29)	43.8%	1.01 (0.23)	71.2%	−0.44 (0.26)	74.6%	0.75 (0.12)	78%	1.46 (0.1)	0.12 (0.13)
500	1	0	0.25 (0.1)	0%	−0.22 (0.1)	11.2%	0.58 (0.13)	4%	−0.49 (0.13)	81.6%	0.99 (0.14)	93.8%	−0.49 (0.13)	95.2%	0.73 (0.18)	94%	0.87 (0.06)	1.09 (0.21)
		1	0.31 (0.16)	0.2%	−0.28 (0.15)	34.4%	0.59 (0.19)	9.8%	−0.51 (0.18)	68.2%	0.99 (0.2)	84.8%	−0.49 (0.18)	91%	0.64 (0.15)	86.8%	1.17 (0.08)	0.57 (0.11)
		5	0.36 (0.28)	4.9%	−0.32 (0.26)	43.33%	0.61 (0.34)	29.8%	−0.51 (0.36)	42%	0.96 (0.32)	65.6%	−0.49 (0.31)	64.8%	0.61 (0.15)	82.8%	1.59 (0.11)	0.36 (0.09)
500	5	0	0.23 (0.16)	0%	−0.2 (0.15)	29.2%	0.54 (0.2)	14.8%	−0.48 (0.21)	76.2%	0.83 (0.23)	74.6%	−0.51 (0.2)	85.4%	0.59 (0.26)	90.8%	1.36 (0.05)	1.48 (0.22)
		1	0.32 (0.24)	5.6%	−0.29 (0.22)	51.4%	0.55 (0.29)	23.4%	−0.5 (0.28)	63.4%	0.83 (0.32)	66.6%	−0.51 (0.27)	74.6%	0.53 (0.22)	84.2%	1.5 (0.09)	0.97 (0.13)
		5	0.34 (0.33)	11.8%	−0.33 (0.34)	50.6%	0.54 (0.45)	31.2%	−0.53 (0.43)	47.8%	0.78 (0.47)	51%	−0.55 (0.41)	59.4%	0.49 (0.2)	78%	1.73 (0.17)	0.67 (0.08)

θ_1 and θ_2 represent the dependence between the terminal event and the recurrent events and the intrarecurrence dependence respectively. CP, coverage probability; SE, standard error.

TABLE 4 Results of the simulation study: marginal approach for the first setting ($\beta_3 = 0.7$)

			$\beta_1 = 1, \beta_2 = -0.5, \beta_3 = 0.7$													
			Ghosh (IPSW)				Huang						Mao			
n	θ_1	θ_2	β_1(SE)	CP	β_2(SE)	CP	β_1(SE)	CP	β_2(SE)	CP	β_3(SE)	CP	β_1(SE)	CP	β_2(SE)	CP
50	0	0	1 (0.33)	89.2%	−0.51 (0.32)	87%	1.04 (0.47)	95%	−0.53 (0.51)	93.4%	0.8(0.55)	97.8%	0.48 (0.2)	26.45%	−0.32 (0.21)	84.8%
		1	0.95 (0.55)	85.2%	−0.5 (0.58)	82.6%	1 (0.61)	94.8%	−0.53 (0.62)	95.8%	0.81(0.72)	97.4%	0.46 (0.31)	55.2%	−0.3 (0.32)	84.2%
		5	1.03 (1.41)	76.8%	−0.54 (1.07)	79.2%	1.06 (1.23)	93.8%	−0.5 (0.94)	95.4%	2.75(24.77)	97.6%	0.5 (0.55)	70.6%	−0.29 (0.55)	85.6%
50	1	0	0.74 (0.46)	87.6%	−0.49 (0.49)	91.2%	0.98 (0.62)	96.2%	−0.53 (0.67)	94.2%	0.82(2.54)	98.2%	0.46 (0.28)	49.8%	−0.31 (0.28)	89.2%
		1	0.71 (1.12)	85.7%	−0.47 (0.7)	86.2%	1.05 (1.23)	93.8%	−0.59 (0.95)	91.8%	0.85(1.98)	97%	0.47 (0.39)	65.3%	−0.3 (0.41)	86%
		5	0.78 (1.73)	79%	−0.7 (1.95)	83%	1.09 (1.75)	95%	−0.7 (1.94)	95.6%	0.81(55.6)	97.8%	0.48 (0.6)	71.2%	−0.33 (0.62)	85.6%
50	5	0	0.74 (1.65)	86.4%	−0.61 (0.74)	91.4%	1.14 (1.96)	93%	−0.62 (1.18)	92.8%	−0.35(20.07)	98%	0.45 (0.5)	77.2%	−0.37 (0.49)	91.2%
		1	0.52 (1.62)	82%	−0.52 (2.56)	85%	0.88 (1.83)	93.6%	60.59 (2.61)	92.2%	−1.31(42.3)	98.6%	0.41 (0.63)	73.8%	−0.33 (0.63)	88.2%
		5	1.6 (5.52)	73%	−0.67 (4.31)	80.6%	1.8 (4.86)	92.6%	−0.68 (4.1)	94%	−10.78(374.17)	94%	0.51 (0.86)	75.6%	−0.24 (0.82)	83.2%
100	0	0	1.02 (0.23)	91.6%	−0.51 (0.23)	89%	1 (0.33)	96.6%	−0.5 (0.34)	94.8%	0.72(0.38)	96.2%	0.49 (0.14)	7%	−0.33 (0.14)	77.6%
		1	0.95 (0.42)	85.4%	−0.47 (0.41)	89%	0.99 (0.45)	95.2%	−0.5 (0.45)	94.8%	0.75(0.48)	97.4%	0.47 (0.23)	33.6%	−0.3 (0.22)	83.2%
		5	0.97 (0.75)	84.6%	−0.47 (0.82)	82.4%	1.04 (0.64)	94.2%	−0.48 (0.66)	92.6%	1.35(12.94)	96.8%	0.51 (0.30)	62%	−0.29 (0.42)	84.6%
100	1	0	0.76 (0.35)	81.56%	−0.5 (0.31)	94.39%	1.02 (0.49)	94.6%	−0.52 (0.49)	94.6%	0.71(0.52)	96.39%	0.48 (0.22)	29.7%	−0.32 (0.19)	84.6%
		1	0.75 (0.45)	87.8%	−0.5 (0.46)	92.6%	1.01 (0.56)	93.2%	−0.52 (0.57)	94.8%	0.74(0.61)	95.8%	0.49 (0.27)	47.4%	−0.31 (0.27)	85.2%
		5	0.78 (0.77)	86.6%	−0.53 (0.83)	84.8%	1.05 (0.81)	92%	−0.54 (0.9)	89.2%	−0.06(28.96)	97.8%	0.5 (0.44)	69.6%	−0.33 (0.47)	86.6%
100	5	0	0.63 (0.48)	86%	−0.48 (0.5)	92.2%	1.04 (0.86)	91.8%	−0.46 (0.84)	91.6%	0.73(0.92)	95.4%	0.46 (0.34)	61.8%	−0.3 (0.35)	89.4%
		1	0.64 (0.63)	85%	−0.47 (0.66)	90.4%	1.1 (0.95)	92%	−0.46 (1.01)	90%	0.73(1.3)	95.8%	0.45 (0.42)	66.8%	−0.3 (0.43)	87.8%
		5	0.67 (1.06)	81.4%	−0.59 (1.38)	84.2%	1.04 (1.41)	92%	−0.48 (1.71)	90.8%	−0.5(33.15)	97.6%	0.44 (0.61)	71.2%	−0.32 (0.62)	83.2%
500	0	0	1.01 (0.11)	92.89%	−0.49 (0.12)	92.28%	1.02 (0.19)	96.34%	−0.49 (0.22)	94.72%	0.72(0.21)	93.9%	0.52 (0.07)	0%	−0.33 (0.07)	31.1%
		1	0.98 (0.2)	91.8%	−0.48 (0.23)	91%	1.01 (0.2)	94%	−0.51 (0.2)	95.2%	0.73(0.21)	93.8%	0.51 (0.1)	0%	−0.32 (0.1)	58.4%
		5	0.96 (0.38)	90.4%	−0.49 (0.42)	89%	1.01 (0.26)	94.8%	−0.5 (0.28)	96%	0.75(0.27)	92.2%	0.51 (0.18)	22.8%	−0.31 (0.18)	81%
500	1	0	0.75 (0.14)	56.2%	−0.49 (0.15)	94.8%	1 (0.22)	94.8%	−0.5 (0.24)	94.8%	0.7(0.24)	94.6%	0.51 (0.09)	0%	−0.32 (0.09)	48.8%
		1	0.75 (0.2)	77.8%	−0.51 (0.21)	95%	1.01 (0.27)	95.6%	−0.52 (0.29)	92.6%	0.72(0.27)	93.8%	0.51 (0.12)	2.8%	−0.33 (0.12)	70.6%
		5	0.77 (0.37)	85.29%	−0.51 (0.39)	90.2%	1 (0.4)	92.75%	−0.51 (0.4)	93.33%	0.7(0.38)	93.33%	0.52 (0.22)	33.53%	−0.33 (0.23)	82.94%
500	5	0	0.59 (0.21)	49.2%	−0.49 (0.22)	94.2%	1 (0.39)	95%	−0.52 (0.42)	92.4%	0.69(0.42)	95.2%	0.46 (0.15)	6.6%	−0.31 (0.16)	76.4%
		1	0.6 (0.3)	66.2%	−0.5 (0.29)	93.2%	1.05 (0.46)	92.8%	−0.49 (0.49)	93%	0.75(0.47)	93.2%	0.47 (0.21)	22.4%	−0.32 (0.19)	80%
		5	0.59 (0.45)	79.6%	−0.53 (0.45)	92%	1 (0.61)	92.6%	−0.53 (0.66)	92.8%	0.74(0.61)	93.8%	0.45 (0.3)	50.4%	−0.34 (0.29)	88.6%

θ_1 and θ_2 represent the dependence between the terminal event and the recurrent events and the intrarecurrence dependence respectively. CP, coverage probability; IPSW, inverse probability of survival weighting; SE, standard error.

TABLE 5 Results of the simulation study: conditional approach for the second setting ($\beta_3 = 0$)

			$\beta_1 = 1, \beta_2 = -0.5, \beta_3 = 0$															
			Miloslavsky With Time-Dependent Covariate				**Miloslavsky Without Time-Dependent Covariate**				**Liu**							
n	θ_1	θ_2	β_1(SE)	CP	β_2(SE)	CP	β_1(SE)	CP	β_2(SE)	CP	β_1(SE)	CP	β_2(SE)	CP	β_3(SE)	CP	θ(SE)	α(SE)
50	0	0	0.47 (0.26)	28.8%	−0.18 (0.21)	56.4%	0.97 (0.33)	81.2%	−0.46 (0.32)	74.2%	0.94 (0.28)	60.6%	−0.46 (0.27)	71.8%	−0.15 (0.55)	72%	0.04 (0.12)	6.91 (4.73)
		1	0.5 (0.38)	40.4%	−0.2 (0.94)	61.8%	1 (0.5)	69%	−0.45 (1.09)	59.2%	1.04 (0.44)	79.6%	−0.47 (0.48)	82%	0 (0.46)	81.4%	0.76 (0.26)	0.42 (1.86)
		5	0.57 (1.11)	45.2%	−0.28 (0.6)	63%	1.09 (1.28)	49%	−0.53 (0.94)	48.8%	1 (0.82)	62.4%	−0.47 (0.79)	71.8%	0 (0.42)	86.2%	1.41 (0.22)	0.12 (0.26)
50	1	0	0.53 (0.33)	47%	−0.23 (0.28)	71.2%	1 (0.41)	81.6%	−0.47 (0.46)	74.4%	1.04 (0.43)	81.6%	−0.46 (0.43)	85.2%	0.01 (0.65)	83.4%	0.78 (0.24)	1.5 (1.5)
		1	0.52 (0.47)	45.6%	−0.27 (0.38)	71.8%	0.97 (0.6)	70%	−0.51 (0.62)	62.8%	1.03 (0.58)	76.2%	−0.46 (0.54)	84.4%	0.03 (0.59)	82.6%	1.11 (0.2)	0.68 (0.43)
		5	0.63 (1.17)	47.6%	−0.29 (0.68)	65%	1.08 (1.43)	52.6%	−0.55 (1.11)	44.6%	0.94 (1.58)	63.6%	−0.55 (1)	66.2%	0.01 (0.56)	82.4%	1.5 (0.29)	0.44 (0.45)
50	5	0	0.56 (0.93)	61.8%	−0.3 (0.96)	79.4%	1.06 (1.1)	79.4%	−0.6 (1.11)	72.6%	1.08 (0.89)	75.4%	−0.63 (1.05)	80.4%	−0.01 (1.37)	82.4%	1.39 (0.17)	2.23 (1.47)
		1	0.73 (1.99)	61.6%	−0.23 (1.37)	76.2%	1.15 (2.14)	65.6%	−0.45 (1.67)	61.2%	1.05 (1.03)	67%	−0.54 (1.01)	71.8%	0.04 (1.25)	85.8%	1.47 (0.23)	1.51 (1.24)
		5	1.43 (4.3)	56.11%	−0.45 (2.84)	68.34%	2.11 (5.34)	50.6%	−0.67 (3.66)	50%	0.79 (2.35)	59.4%	−0.93 (3.47)	62%	0.03 (1.21)	85.2%	1.58 (0.36)	1.06 (0.93)
100	0	0	0.51 (0.19)	15.8%	−0.2 (0.15)	37.6%	0.99 (0.23)	81.2%	−0.5 (0.23)	76%	0.94 (0.23)	58.2%	−0.48 (0.21)	68.2%	−0.2 (0.46)	69.8%	0.04 (0.11)	7.28 (4.12)
		1	0.53 (0.28)	27.6%	−0.21 (0.22)	45.2%	0.99 (0.35)	68%	−0.45 (0.35)	64.6%	1.03 (0.29)	84.4%	−0.44 (0.29)	89.4%	0.01 (0.29)	84%	0.85 (0.14)	0.02 (0.27)
		5	0.56 (0.47)	39%	−0.23 (0.44)	49.6%	1.04 (0.61)	50%	−0.45 (0.69)	43%	1.01 (0.5)	67.6%	−0.43 (0.55)	72.4%	0.02 (0.29)	85.4%	1.48 (0.15)	0.11 (0.13)
100	1	0	0.55 (0.27)	29.8%	−0.23 (0.19)	53.8%	1 (0.32)	76%	−0.48 (0.3)	75.2%	1.07 (0.32)	80.26%	−0.46 (0.28)	90.24%	0.04 (0.43)	84.6%	0.84 (0.15)	1.21 (0.54)
		1	0.57 (0.35)	37.6%	−0.26 (0.27)	61.8%	0.98 (0.43)	66%	−0.5 (0.42)	65.8%	1.05 (0.4)	77.53%	−0.46 (0.38)	83.37%	0.06 (0.34)	86.07%	1.18 (0.13)	0.59 (0.23)
		5	0.61 (0.58)	40%	−0.31 (0.52)	54.8%	1.05 (0.75)	50.2%	−0.54 (0.8)	43.8%	1.06 (0.68)	61.2%	−0.47 (0.68)	62.8%	0.09 (0.35)	84.6%	1.62 (0.2)	0.35 (0.15)
100	5	0	0.55 (0.37)	52%	−0.23 (0.33)	67.8%	1.03 (0.46)	78.2%	−0.49 (0.48)	72.6%	1.13 (0.57)	77.6%	−0.49 (0.46)	81.2%	0.11 (0.74)	89.2%	1.42 (0.12)	1.65 (0.79)
		1	0.59 (0.48)	54%	−0.27 (0.42)	68.2%	1.03 (0.62)	67.2%	−0.48 (0.64)	60.2%	1.08 (0.65)	74.8%	−0.47 (0.6)	75.2%	0.1 (0.66)	89.8%	1.52 (0.19)	1.18 (0.58)
		5	0.63 (0.73)	53.8%	−0.35 (1.08)	61%	1.03 (0.99)	48.6%	−0.53 (1.39)	48%	1.05 (0.92)	60.8%	−0.52 (1.01)	58.8%	0.12 (0.53)	86.8%	1.71 (0.31)	0.77 (0.35)
500	0	0	0.54 (0.08)	0%	−0.21 (0.07)	80%	1 (0.1)	79.59%	−0.51 (0.1)	73.47%	0.91 (0.17)	50.8%	−0.47 (0.17)	62.4%	−0.23 (0.42)	64.4%	0.04 (0.14)	5.85 (3.62)
		1	0.6 (0.15)	5.80%	−0.24 (0.12)	14.4%	0.99 (0.15)	66.2%	−0.49 (0.16)	61.8%	1.01 (0.15)	84.15%	−0.44 (0.21)	86.18%	0.05 (0.15)	80.49%	0.88 (0.06)	0.03 (0.19)
		5	0.62 (0.26)	18.0%	−0.27 (0.22)	32.8%	0.99 (0.28)	47.6%	−0.48 (0.29)	41%	0.98 (0.23)	68.2%	−0.43 (0.26)	72%	0.05 (0.14)	82.6%	1.54 (0.1)	0.12 (0.14)
500	1	0	0.6 (0.12)	4.20%	−0.24 (0.1)	11.2%	1 (0.13)	77.4%	−0.49 (0.13)	77.4%	1 (0.19)	81.4%	−0.42 (0.28)	86.6%	0.07 (0.21)	83.6%	0.88 (0.06)	1.04 (0.17)
		1	0.68 (0.18)	20.4%	−0.31 (0.15)	36.8%	1 (0.18)	67.2%	−0.51 (0.19)	61.2%	1.02 (0.21)	78.4%	−0.45 (0.26)	80.6%	0.1 (0.18)	77.8%	1.2 (0.08)	0.55 (0.12)
		5	0.7 (0.31)	32.2%	−0.31 (0.27)	39.60%	1.01 (0.35)	39%	−0.5 (0.37)	36.8%	0.99 (0.31)	63%	−0.43 (0.34)	60.6%	0.13 (0.17)	72%	1.67 (0.15)	0.35 (0.13)
500	5	0	0.61 (0.2)	21.2%	−0.24 (0.16)	33.4%	1 (0.2)	77.4%	−0.49 (0.21)	72%	1.05 (0.23)	83.6%	−0.51 (0.2)	83.2%	0.02 (0.26)	91.6%	1.45 (0.05)	1.34 (0.15)
		1	0.69 (0.28)	34.6%	−0.31 (0.22)	50.8%	1.01 (0.3)	59%	−0.5 (0.29)	56.4%	1.07 (0.32)	68.6%	−0.52 (0.27)	73%	0.08 (0.23)	89.2%	1.61 (0.09)	0.91 (0.1)
		5	0.69 (0.39)	35.6%	−0.34 (0.35)	44.2%	0.99 (0.46)	43%	−0.54 (0.46)	43.8%	1.02 (0.46)	54.4%	−0.51 (0.42)	55%	0.13 (0.21)	88.6%	1.88 (0.17)	0.63 (0.07)

θ_1 and θ_2 represent the dependence between the terminal event and the recurrent events and the intrarecurrence dependence, respectively. CP, coverage probability; SE, standard error.

TABLE 6 Results of the simulation study: marginal approach for the second setting ($\beta_3 = 0$)

			$\beta_1 = 1, \beta_2 = -0.5, \beta_3 = 0$													
			Ghosh (IPSW)				Huang						Mao			
n	θ_1	θ_2	β_1(SE)	CP	β_2(SE)	CP	β_1(SE)	CP	β_2(SE)	CP	β_3(SE)	CP	β_1(SE)	CP	β_2(SE)	CP
50	0	0	0.98 (0.3)	89%	−0.47 (0.27)	88.8%	1 (0.46)	94.6%	−0.49 (0.45)	93.6%	−0.01 (0.59)	94.6%	0.69 (0.19)	66.8%	−0.34 (0.21)	85.4%
		1	0.99 (0.51)	88.2%	−0.48 (1.04)	88.2%	1.05 (0.53)	94%	−0.49 (1.34)	95%	0 (0.63)	94%	0.7 (0.32)	78.2%	−0.36 (0.37)	85.8%
		5	1.08 (1.34)	82%	−0.55 (1.03)	78.8%	1.12 (1.16)	92.8%	−0.52 (0.88)	94.4%	−4.15 (39.64)	97.6%	0.69 (0.59)	80.2%	−0.36 (0.63)	86%
50	1	0	1 (0.45)	91.6%	−0.49 (0.49)	89.4%	1.04 (0.63)	95.4%	−0.47 (0.64)	91.6%	−0.03 (0.74)	98%	0.7 (0.3)	79.4%	−0.35 (0.33)	88.8%
		1	0.98 (0.64)	90%	−0.51 (0.65)	88%	1 (0.81)	93.2%	−0.52 (0.81)	92%	−0.78 (10.14)	99%	0.68 (0.42)	82.4%	−0.35 (0.45)	89.4%
		5	1.08 (1.46)	81.2%	−0.57 (1.14)	80.8%	1.17 (1.45)	93.2%	−0.58 (1.17)	94.8%	−5.48 (36.52)	97%	0.68 (0.7)	80.6%	−0.36 (0.71)	85%
50	5	0	1.06 (1.11)	92.4%	−0.6 (1.13)	90.8%	1.15 (1.36)	93%	−0.58 (1.32)	92.4%	−3.77 (38.78)	97.8%	0.71 (0.53)	89%	−0.41 (0.56)	92.8%
		1	1.16 (2.17)	85.4%	−0.46 (1.68)	83.4%	1.23 (2.18)	93%	−0.49 (1.89)	93.6%	−6.25 (39.99)	98.4%	0.67 (0.69)	81.8%	−0.36 (0.73)	84%
		5	2.09 (5.29)	75.55%	−0.64 (3.59)	78.56%	2.1 (5.39)	91.78%	−0.62 (3.37)	96.39%	−13.86 (111.09)	94.19%	0.73 (0.96)	81.96%	−0.29 (0.99)	81.76%
100	0	0	1.01 (0.19)	92.6%	−0.5 (0.18)	92.8%	1.02 (0.32)	95.8%	−0.5 (0.32)	95.4%	0.02 (0.39)	95.2%	0.71 (0.14)	45.8%	−0.37 (0.15)	87.2%
		1	1 (0.36)	91.6%	−0.46 (0.37)	90%	1.01 (0.4)	93.8%	−0.48 (0.39)	92.8%	0 (0.47)	95%	0.71 (0.22)	73.6%	−0.33 (0.25)	87%
		5	1.04 (0.69)	86.8%	−0.46 (0.74)	84.8%	1.05 (0.6)	92.2%	−0.45 (0.63)	93%	−0.06 (0.64)	98.8%	0.72 (0.42)	82.4%	−0.32 (0.48)	86%
100	1	0	1 (0.35)	90%	−0.49 (0.31)	92.4%	1.04 (0.5)	93.2%	−0.47 (0.46)	93.4%	−0.01 (0.53)	96.4%	0.71 (0.23)	68.6%	−0.35 (0.22)	89.2%
		1	0.99 (0.46)	90.6%	−0.49 (0.44)	92.2%	1.05 (0.55)	92%	−0.51 (0.53)	94.4%	0.04 (0.56)	96.6%	0.7 (0.3)	76.2%	−0.36 (0.31)	89.6%
		5	1.04 (0.78)	87%	−0.54 (0.83)	85.4%	1.08 (0.84)	91.2%	−0.54 (0.88)	91.2%	−0.14 (5.28)	98.2%	0.72 (0.5)	83.2%	−0.39 (0.57)	86.8%
100	5	0	1.04 (0.49)	94%	−0.48 (0.5)	93.6%	1.13 (0.72)	92.8%	−0.46 (0.75)	93.4%	0.07 (0.92)	95.6%	0.73 (0.36)	85.8%	−0.35 (0.39)	91.6%
		1	1.03 (0.64)	91%	−0.48 (0.67)	87.8%	1.1 (0.89)	90%	−0.52 (0.97)	90.4%	−0.01 (1)	95.6%	0.71 (0.46)	87.4%	−0.35 (0.49)	88%
		5	1.03 (1)	86.2%	−0.53 (1.4)	81.4%	1.06 (1.31)	93.2%	−0.53 (1.6)	92.2%	−4.36 (38.52)	97.2%	0.66 (0.66)	83%	−0.34 (0.74)	83.4%
500	0	0	1 (0.09)	94.8%	−0.51 (0.08)	94.4%	1.01 (0.18)	92.6%	−0.51 (0.18)	93.2%	0.01 (0.2)	95.2%	0.72 (0.06)	0.80%	−0.38 (0.07)	56.6%
		1	0.99 (0.17)	92.8%	−0.49 (0.17)	95.4%	1 (0.22)	95.0%	−0.51 (0.2)	95.40%	0.01 (0.25)	92.40%	0.71 (0.11)	22.20%	−0.36 (0.11)	76.4%
		5	0.99 (0.32)	92.8%	−0.48 (0.35)	91.2%	1.02 (0.27)	93.2%	−0.5 (0.28)	94.6%	0.03 (0.27)	93.2%	0.71 (0.19)	65.0%	−0.36 (0.21)	87.8%
500	1	0	1 (0.14)	94.4%	−0.49 (0.14)	93.6%	1.01 (0.22)	95.2%	−0.51 (0.22)	94.6%	0 (0.24)	94.2%	0.72 (0.1)	19.2%	−0.36 (0.1)	70.6%
		1	1 (0.2)	94.8%	−0.51 (0.21)	94.6%	1.02 (0.25)	95.2%	−0.52 (0.25)	94.2%	0.01 (0.27)	94.8%	0.72 (0.14)	46.6%	−0.38 (0.15)	86.0%
		5	1.01 (0.37)	90.2%	−0.5 (0.39)	89.4%	0.99 (0.38)	92.8%	−0.5 (0.38)	93.0%	−0.02 (0.36)	94.0%	0.72 (0.25)	71.4%	−0.37 (0.27)	87.8%
500	5	0	1 (0.21)	95.0%	−0.49 (0.23)	94.0%	1.01 (0.35)	93.8%	−0.5 (0.36)	94.2%	−0.01 (0.38)	94.4%	0.71 (0.16)	57.4%	−0.36 (0.17)	85.2%
		1	1.01 (0.31)	92.0%	−0.5 (0.31)	93.8%	1.02 (0.43)	93.4%	−0.49 (0.41)	95.4%	0 (0.43)	94.6%	0.73 (0.23)	70.0%	−0.37 (0.23)	88.0%
		5	0.99 (0.46)	91.8%	−0.53 (0.48)	92.6%	1.01 (0.57)	92.4%	−0.53 (0.58)	94.2%	0.02 (0.54)	94.4%	0.71 (0.33)	79.0%	−0.39 (0.35)	90.8%

θ_1 and θ_2 represent the dependence between the terminal event and the recurrent events and the intrarecurrence dependence, respectively. CP, coverage probability; IPSW, inverse probability of survival weighting; SE, standard error.

In this setting, the Miloslavsky and the Ghosh models provide unbiased estimations of parameters compared to the first setting ($\beta_3 = 0.7$). Indeed, these two models only consider that the terminal event is independent from recurrent events even given the covariates. However, when the sample size is small ($n = 50$) and the two dependences are high ($\theta_i = 1$ or 5, $i = 1, 2$), results are biased, but the bias decreases when the sample size increase. The Miloslavsky's model without time-dependent variable provide unbiased results except when the dependence between the terminal and recurrent events is high and the sample size is low ($n = 50$). However, when the time-dependent covariate is included in the Miloslavsky's model, coefficients are underestimated. This is due to the misspecification of the intrarecurrence dependence. For the Ghosh's model, results are unbiased and the coverage probability is close to the nominal value.

For the joints models, the Liu and the Huang models provide similar results for the parameters β_1 and β_2, ie, estimations for these two parameters are good. However, for the Huang's model, the estimation of the parameter β_3 is highly biased when the sample size is small ($n = 50$) and when dependences is high ($\theta_i = 1$ or 5, $i = 1, 2$). However, the bias decreases when the sample size increases. For the Liu's model, the bias for the parameter β_3 is reasonable. However, the coverage probability is better and close to the nominal value for the marginal approach than for the conditional approach.

The Mao's model provides similar results as in the first setting. However, the bias is smaller.

6 | APPLICATION

6.1 | Readmission data

Dataset is available in the frailtypack package[15] of the R software. They were obtained from a prospective cohort study at the hospital of Bellvitge of Barcelona.[44] A total of 403 patients with a new diagnosis of colorectal cancer who had a surgery were identified from January 1996 to December 1998. The database contains rehospitalization times after surgery. The aim of the study was to assess the effects of gender and chemotherapy on hospital readmission after surgery. This cohort included 239 (59%) men and 217 (54%) patients treated with chemotherapy. More details of the dataset form are available in the Appendix.

Not taking into account the terminal event in the analysis of recurrence events is equivalent to focus on the cumulative number of events only as displayed in Figure 1A and to conclude that treated patients have lower events than nontreated patients, ie, to conclude for the effectiveness of the treatment. However, if we compare the cumulative number of events for survivors only (Figure 1B), the difference between treated and nontreated patients becomes lower. Indeed, these curves do not take into account the deceased patients that, by definition, cannot experience the event after their death. If the cumulative number of events is similar between survivors and deceased patients in control group, it is not the case in the treated group with very low events within the treated patients who died during the study. This is explained by Figure 1C, showing that, in mean, the treated patients died earlier during the study than those in the control group and hence experienced less rehospitalization.

We have applied the models described above on the readmission dataset and provided the R code in the Appendix. The Miloslavsky's model can be implemented using the `coxph` function in the R software (R code available in the Appendix). To model the dependence between recurrent events, we include the number of events before time t (enum) as a covariate in the model. We applied Liu's model using the estimation of Rondeau et al based on readmission data using the R code (see the Appendix). In this example, we used the gap time scale. The R code for fitting the Ghosh's model using the IPSW weighting method is detailed in the Appendix. We applied Huang's model (available in the `reReg` package) on readmission data (R code available in the Appendix). As standard errors are estimated using the bootstrap method, it is preferable to use the `set.seed` option to consistently obtain the same results. In the Mao's model, the first step is to create a composite endpoint with a value of 2 if the observation is a recurrent event, 1 if it is a terminal event, and 0 if it is a censoring. This function requires the removal of missing data (R code available in the Appendix). We applied this model on the readmission data by attributing the same weight equal to 1 to recurrent and terminal event.

6.2 | Results

Impact of the event past history, the heterogeneity, and the risk set on results

Table 7 shows that the size of the treatment effect and conclusions for some parameters differ according to the method used.

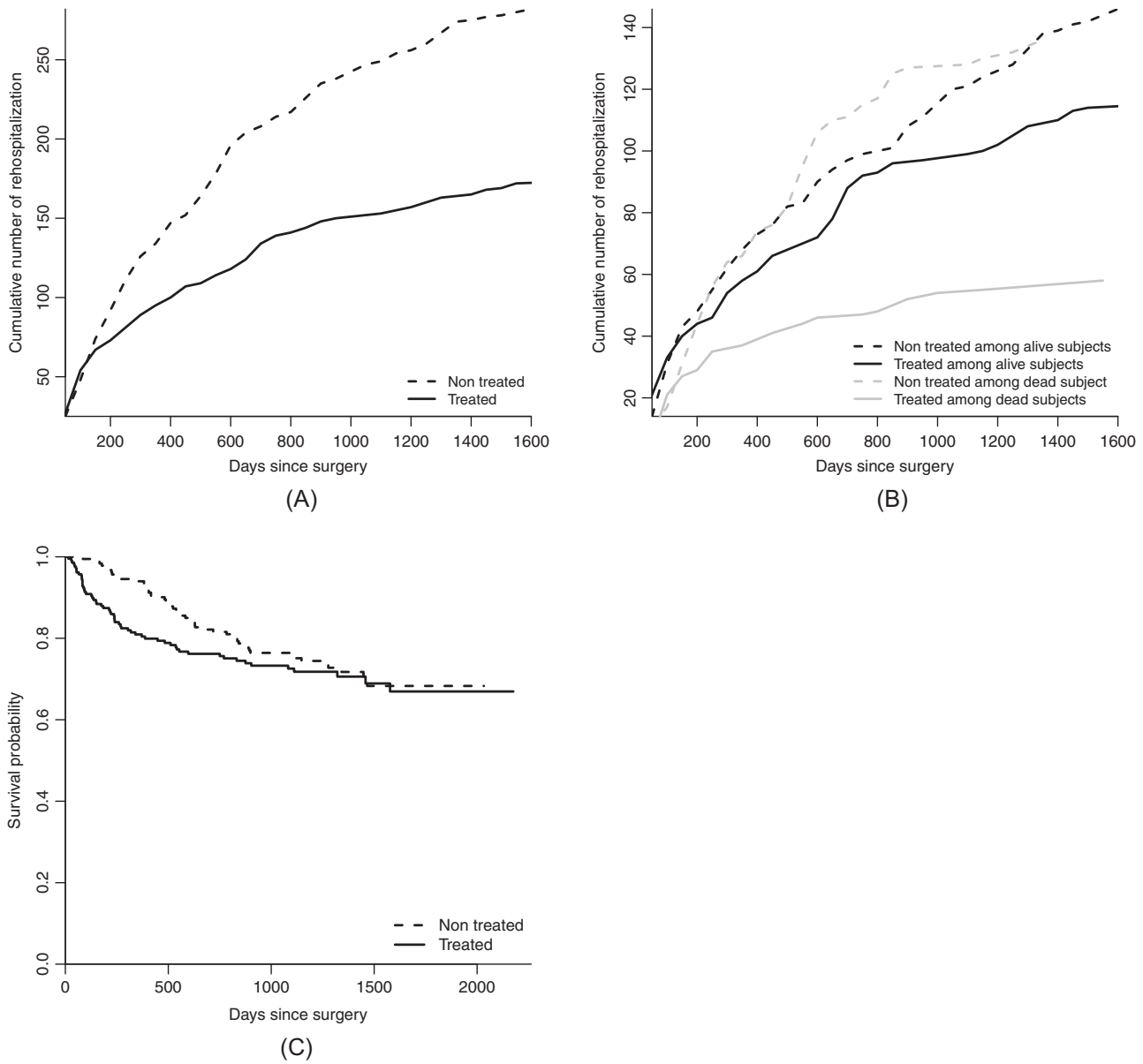


FIGURE 1 Cumulative number of rehospitalizations and survival curves. A, Cumulative number of rehospitalization between treated and nontreated subjects; B, Cumulative number of rehospitalization between treated and nontreated subjects according to death status; C, Survival curve

We can first notice that conditional models provide lower covariates effects than the marginal models, probably because of the event's past history taken into account in the conditional model and not in the marginal models. Indeed, the occurrence of an event is due to both inherent characteristics of the subject, such as event past history and external covariates. Thus, taking into account past history should reduce the effect of covariates. The Huang's model provides higher covariates effects than the Liu's model; this could be due to the event's past history. Indeed, the coefficients in the Liu's model are estimated simultaneously from the complete likelihood function and then include all information, including the event's past history. On the contrary, the Huang's model estimates the coefficients in three steps (see the work of Huang and Wang[26] for more details).

Ignoring dependence between recurrent events and the terminal event may lead to wrong conclusions. In Miloslavsky's model, chemotherapy significantly decreased the risk of readmission but it did not in Liu's model. This may be due to the dependence caused by unobserved variables between readmission and death, as θ was significantly different from zero; γ (noted `alpha` in the R output). These parameters were both significantly different from 0 meaning that readmission times and death time were positively ($\gamma > 0$) correlated. Thus, an unobserved variable similarly affects both the intensity of

TABLE 7 Results

	Conditional Models						Marginal Models								
	Miloslavsky			Liu			Ghosh (IPSW)			Huang			Mao		
	β	Se	p-value	β	Se	p-value	β	Se	p-value	β	Se	p-value	β	Se	p-value
Recurrences															
Gender	−0.27	0.08	0.06	−0.41	0.15	0.005	−0.49	0.09	0.007	−0.77	0.32	0.016	-	-	-
Chemotherapy	−0.34	0.08	0.017	−0.21	0.14	0.14	−0.49	0.09	0.005	−0.26	0.30	0.384	-	-	-
Number of previous events	0.18	0.009	<0.0001	-	-	-	-	-	-	-	-	-	-	-	-
Terminal Event															
Gender	-	-	-	−0.11	0.28	0.69	-	-	-	−0.66	0.42	0.229	-	-	-
Chemotherapy	-	-	-	0.63	0.30	0.033	-	-	-	0.24	0.50	0.012	-	-	-
θ	-	-	-	0.93	0.01	0	-	-	-	-	-	-	-	-	-
γ	-	-	-	1.84	0.33	<0.0001	-	-	-	-	-	-	-	-	-
Composite Endpoint															
Gender	-	-	-	-	-	-	-	-	-	-	-	-	−0.41	0.15	0.005
Chemotherapy	-	-	-	-	-	-	-	-	-	-	-	-	−0.43	0.14	0.003

θ represents the dependence between recurrent and terminal events. IPSW, inverse probability of survival weighting.

readmission and death. Joint models provide a nonsignificant effect of "chemo" on recurrent events but are significant for death. This is due to the fact that recurrent events and terminal events are correlated. In the joint frailty model, standard errors are higher than in the Miloslavsky's and the Ghosh's model because the two further capture heterogeneity due to unobserved covariates, whereas the two latter consider observations independent conditionally upon covariates. One way to correct it is the use of a robust variance (see R code in the Appendix).

Keeping deceased patients in the risk set (IPCW procedure) provides a lower covariates effect than when they are removed (IPSW procedure).

The Milosavsky's model shows a smaller effect of chemotherapy and gender than the Ghosh's model because patients who die remain in the risk set in the Miloslasky's model, whereas they are removed from the risk set in the Ghosh's model using the IPSW procedure. The fact that subjects who died remain in the risk set tend to decrease the covariates effects in the time.

Interpretation of the intensity and the mean/rate functions

Coefficients's interpretation differs from one method to the other one.

With the Miloslavsky's model, we interpret the coefficient in terms of risk. For example, women tended to have a lower risk of readmission than men (hazard ratio (HR) = 0.76, 95% CI = [0.57 − 1.01], p = 0.062) among all patients (survivors and deceased subjects), but for the Liu's model, coefficients are interpretable in terms of the instantaneous probability of recurrent events and the terminal event conditional upon the subject's event history among survivors conditionally on the frailty level. For the same level of frailty, women had a lower instantaneous probability of experiencing a new event than men (HR = 0.66, 95% CI = [0.50 − 0.88], p = 0.0049).

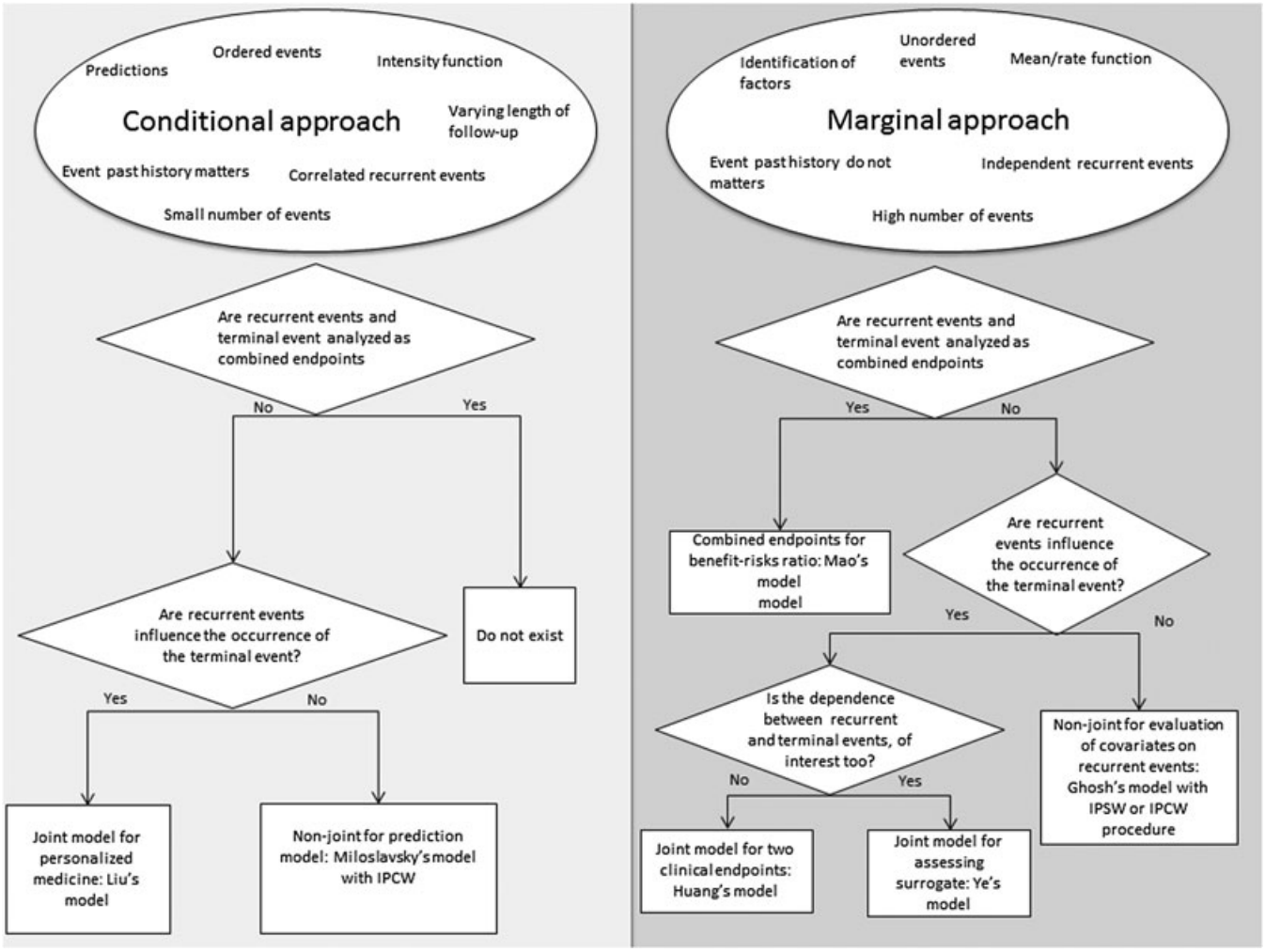


FIGURE 2 Choice of the adequate model for the analysis of recurrent and terminal events

With the marginal approach, coefficients represent differences of mean number of event in the Ghosh's model. Women had a reduction of 40% of the mean number of readmission than men ($p = 0.0066$) among survivors. Chemotherapy reduces the mean of readmission by 40% ($p = 0.0053$). In the Huang's model, coefficients are interpreted in term of rate and are conditional upon the frailty parameter. For patients with a similar dependence between recurrent and terminal events (same frailty value), women had a lower rate of experiencing a recurrent event than men (Rate Ratio (RR) = 0.46, 95% CI = [0.25 − 0.86], $p = 0.023$). Considering the link between the rate and mean function, this is equivalent to say that women have a lower number of recurrences than men. Nontreated patients survived longer than treated patients (HR = 3.55, 95% CI = [1.32 − 9.54], $p = 0.013$).

In the Mao's model, coefficients are also interpreted as mean number of the weighted composite event, women had 0.66 (95%CI = [0.50-0.88], $p = 0.0047$) times less events (recurrent or terminal events) than men, ie, the mean number of events for women was 44% that of the mean number of events for men.

7 | DISCUSSION

We have reviewed the different approaches to analyze recurrent events in the presence of a terminal event. Models can be split into four main categories: (i) nonjoint conditional approach (Miloslavsky's model); (ii) joint conditional approach (Liu's model); (iii) nonjoint marginal approach (Ghosh's model); (vi) joint marginal approach (Huang's model). As for the standard analysis of recurrent events, here, the choice of the appropriate methodology also depends on the biological process of recurrent events, the dependence between recurrent and terminal events, and the question of interest (Figure 2).

Conditional approaches (Table 8) are recommended when the order of the occurrence of the recurrent events is primordial or when events are correlated. Indeed, the conditional approach models the intensity function which accounts for the event's past history and considers data as survival data. The two-step and the joint models have been described in the previous sections. If recurrent events do not influence the occurrence of the terminal event, then the two-step conditional model is preferred (Miloslavsky's model). It consists in modeling the censoring distribution, incorporate it in the

TABLE 8 Pros and cons of conditional approach

	Conditional Approach	
	Terminal event (TE) impacted by recurrent events (RE)	TE not impacted by RE
Function	$\begin{cases} r_s(t) \\ \lambda(t) \end{cases}$	$r(t)$
When to use it		
	•To make prediction or for personalized medicine •Small number of events •Past event history is important •Correlation between RE and TE of interest	•**Correlation between RE and TE not of interest** •To make prediction •Small number of recurrent events •Past event history is important
Interpretation	Subject's instantaneous probability of occurrence of recurrent event conditional on the past event history and frailty **among survivors given frailty**	**Average** instantaneous probability of occurrence of recurrent event conditional on the past event history among survivors and deceased patients
Pros		
	•Past event history is taken into account •Estimation of correlation between TE and RE	•**Do not require assumption of independent censoring** •Past event history is taken into account
Cons		
	•**Population changes over time** •Complicated to achieve marginal functions which is often of interest •Not suitable for clinical research because of past history •Interpretation nonnatural •Require independent censoring assumption	•**Need to model censoring distribution to apply IPCW method** •Not suitable for clinical research because of past history •Interpretation nonnatural

TABLE 9 Pros and cons of marginal approach

	Marginal Approach			
	Terminal event (TE) impacted by recurrent events (RE)	TE not impacted by RE		
	Assess simultaneously effect of covariates on RE and ET (Huang's model)	RE of interest only (Ghosh's model with IPSW)	RE of interest (Ghosh's model with IPCW)	RE and TE of interest in a combined endpoint (Mao's model)
Function	$\begin{cases}\rho_2(t)Y(t)) \\ \lambda(t)\end{cases}$	$\rho_2(t)Y(t)$	$\rho_1(t)$	$\rho_1(t)$
When to use it				
	•Frequent events	•Frequent recurrent events	•Frequent recurrent events	•Small number of recurrent events
	•Correlation between ER and ET not of interest	•Clinical trial with two primary endpoints		
	•identification of factor and treatment effect			
Interpretation	Subject's marginal effect on the rate/mean function over time **given frailty**	Average marginal effect on the conditional recurrent event rate/mean **among survivors**	Average marginal effect on the marginal recurrent event rate/mean **over time**	Average marginal effect on the marginal event (recurrent+terminal events) rate/mean **over time**
Pros				
	•Easy interpretation	•Easy interpretation	**•Do not need assumption of independent censoring**	**•Do not need assumption of independent censoring because of the use of IPCW method**
	•Suitable for clinical research	•Suitable for clinical research	**•Robust to the type of underlying recurrent process**	•Easy interpretation
	•Do not require independent censoring assumption		•Easy interpretation	•Suitable for clinical research
			•Suitable for clinical research	
Cons				
	•Cannot be used to predict future events given past event history	•Cannot be used to predict future events given past event history	•Cannot be used to predict future events given past event history	**•Not possible to distinguish effect of the treatment on recurrent event and on the terminal event**
	•Recurrent events process continue to increment after death	•Require independent censoring assumption	**•Do not require independent censoring assumption but need to model censoring distribution to apply IPCW method**	•Cannot be used to predict future events given past event history
			•Reduce efficiency	

Abbreviation: IPSW, inverse probability of survival weighting.

recurrent events model. Otherwise, joint conditional models are recommended (Liu's model). Recurrent and terminal events are simultaneously modeled.

The two-step conditional model (Miloslavsky's model) is appropriate when the question of interest lies in the overall effect of covariates on the time to the occurrence of recurrent events. However, the Miloslavsky's model makes a strong assumption about intrarecurrence dependence. Indeed, it has to be reasonable to assume that intrarecurrence dependence is completely due to observed covariates. This means that time increments are independent conditionally on covariates. A solution is to introduce previous events as a time-dependent covariate in the model. In practice, omission of an important covariate will bias the results. The Miloslavsky's model is preferred when recurrent and the terminal events are not associated to the same covariates or in case of dependent censoring (ie, the processus of recurrence can continue to increment after censoring time). Indeed, patients remain in the risk set event after the occurrence of the terminal event.

The joint conditional model (Liu's model) has a great appealing for personalized medicine or for making predictions. In fact, personalized medicine aims to personalize care in order to give the best treatment according to the response rate of each patient. Thus, the Liu's model is appropriate for individual prediction because all available information about the patient is used in the model (the previous events for example) but also information about the occurrence of the terminal event. Interpretation is subject-specific. It accounts for intrasubject correlation using a parametric frailty parameter and allows assessing the dependence between recurrent and terminal events. However, the model is less parsimonious than the two-step model and can be time consuming in case of a high amount of data. Moreover, coefficients's interpretation is not straightforward. Indeed, it is conditioned upon the frailty level. This means that patients who are most frail (high value of the frailty) would experience an event earlier.

Unlike the conditional approach, marginal models (Table 9) are preferred when recurrent events are unordered or frequent or independent. Indeed, the mean or rate functions are modeled in this approach and ignore the event past history. Three main types of marginal model have been developed, a two-step and joint model as in the conditional approach and a composite endpoint combining recurrent and terminal events in a single endpoint. When the recurrent events do not influence the occurrence of the terminal event, then the marginal nonjoint or combined endpoint is preferred. Otherwise, joint marginal models are recommended.

The two-step marginal model, as the Ghosh's model, is appropriate when the interest lies in assessing the population average effect of covariates on the mean/rate functions. The Ghosh's model with the IPCW procedure is preferred in case of dependent censoring and the Ghosh's model with the IPSW procedure, in case of a terminal event. These models are sensitive to the censoring or survival distribution, as the models use them for the weights procedure. Indeed, it assumes that recurrent events do not increase or decrease the risk of the occurrence of the terminal event. Thus, caution has to be taken in modeling censoring/survival.

The joint marginal model with latent parameter (nonparametric frailty) like the Huang's model is very appealing in case of two primary endpoints in clinical trial. This model captures the dependence between recurrent and terminal events. However, it considers it as a nuisance parameter so it is not possible to assess dependence between these two processes. Moreover, this model assumes that the recurrent event process continues to increment event after the terminal event has occurred, which is a nonnatural assumption. Marginal joint models with frailty parameter, such as the Ye's model, are useful to check whether recurrent events are surrogates of the terminal event. Indeed, it allows checking the three conditions of validation of a surrogate, the evaluation effect of covariate on the recurrent events and on terminal event, and assessing dependence between these two processes.

Finally, the use of a combine endpoint is require in case of a few number of events and when the interest may lie in the benefit risk ratio, which may be evaluated by combining “good” and “bad” events. However, they must be used with caution. Indeed, they are appropriate only if the terminal and recurrent events are explained by the same covariates because they do not differentiate the effect of the covariates on recurrent or the terminal events.

An alternative way to analyze recurrent events in this context would be to directly model the time to recurrent events using accelerated failure models.[45,46] The main advantage of these models is their easier interpretation because the parameters measure the effect of the covariates on survival time instead of the hazard. However, these models require the user to specify the survival distribution.

Moreover, only models for which the exact times of event occurrence are known and assuming continuous time are discussed in this article. However, subjects may be observed at discrete time points only (eg, annual visit), these types of data are called panel count data and only the number of events occurring between two time points is known (panel count data). Zhao et al[47] proposed a model to analyze panel count data in the presence of a terminal event.

In conclusion, models for recurrent events in the presence of terminal events should be routinely used in the analyses of clinical trial or in epidemiology studies when required. If the intensity function is of interest or if recurrent events are correlated, then conditional models are required. However, if the cumulative mean function or the recurrent events are independent, then marginal models are recommended. In case of a strong dependence between recurrent and terminal events, it is recommended to use a joint model.[48] However, when a model for recurrent and terminal event is used, cautious interpretation of the coefficients must be made as the coefficients of the conditional models are conditional upon the past event history and upon the frailty parameter in models including a frailty parameter.

ORCID


Anaïs Charles-Nelson https://orcid.org/0000-0001-6437-7059

REFERENCES

1. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer Science+Business Media; 2000.
2. Kalbfleisch JD, Prentice RL. Relative risk (Cox) regression models. In: *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc; 2002:95-147.
3. Kelly PJ, Lim LL-Y. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statist Med*. 2000;19(1):13-33.
4. Cook RJ, Lawless JF. Analysis of repeated events. *Stat Methods Med Res*. 2002;11(2):141-166.
5. Cai J, Schaubel DE. Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Anal*. 2004;10(2):121-138.
6. Sinha D, Maiti T, Ibrahim JG, Ouyang B. Current methods for recurrent events data with dependent termination: a Bayesian perspective. *J Am Stat Assoc*. 2008;103(482):866-878.
7. Cook RJ, Lawless J. *The Statistical Analysis of Recurrent Events*. New York, NY: Springer Science+Business Media; 2007.
8. Miloslavsky M, Keles S, van der Laan MJ, Butler S. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *J Royal Stat Soc Stat Methodol Ser B*. 2004;66(1):239-257.
9. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat*. 1982;10(4):1100-1120.
10. Liu L, Wolfe RA, Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics*. 2004;60(3):747-756.
11. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*. 2000;56(3):779-788.
12. Cox DR. Regression models and life-tables. *J Royal Stat Soc Stat Methodol Ser B*. 1972;34(2):187-202.
13. Breslow N. Covariance analysis of censored survival data. *Biometrics*. 1974;30(1):89-99.
14. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*. 2007;8(4):708-721.
15. Rondeau V, Mazroui Y, Gonzalez J. Frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw*. 2012;47(4):1-28.
16. Zhangsheng Y, Liu L. A joint model of recurrent events and a terminal event with a nonparametric covariate function. *Statist Med*. 2011;30(22):2683-2695.
17. Liu L, Huang X, Yaroshinsky A, Cormier JN. Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics*. 2016;72(1):204-214.
18. Rondeau V, Schaffner E, Corbiere F, Gonzalez JR, Mathoulin-Pelissier S. Cure frailty models for survival data: application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Stat Methods Med Res*. 2013;22(3):243-260.
19. Mazroui Y, Mathoulin-Pelissier S, MacGrogan G, Brouste V, Rondeau V. Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data. *Biometrical Journal*. 2013;55(6):866-884.
20. Belot A, Rondeau V, Remontet L, Giorgi R, CENSUR working survival group. A joint frailty model to estimate the recurrence process and the disease-specific mortality process without needing the cause of death. *Statist Med*. 2014;33(18):3147-3166.
21. Mazroui Y, Mathoulin-Pelissier S, Soubeyran P, Rondeau V. General joint frailty model for recurrent event data with a dependent terminal event: application to follicular lymphoma data. *Statist Med*. 2012;31(11-12):1162-1176.
22. Huang X, Liu L. A joint frailty model for survival and gap times between recurrent events. *Biometrics*. 2007;63(2):389-397.
23. Liu L, Huang X. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *J Royal Stat Soc Appl Stat Ser C*. 2009;58(1):65-81.
24. Zeng D, Lin DY. Semiparametric transformation models with random effects for joint analysis of recurrent and terminal events. *Biometrics*. 2009;65(3):746-752.
25. Ghosh D, Lin DY. Marginal regression models for recurrent and terminal events. *Statistica Sina*. 2002;12(3):663-688.
26. Huang C-Y, Wang M-C. Joint modeling and estimation for recurrent event processes and failure time data. *J Am Stat Assoc*. 2004;99(468):1153-1165.
27. Mao L, Lin DY. Semiparametric regression for the weighted composite endpoint of recurrent and terminal events. *Biostatistics*. 2016;17(2):390-403.
28. Zhao H, Zhou J, Sun L. A marginal additive rates model for recurrent event data with a terminal event. *J Commun Stat Theory Methods*. 2013;42(14):2567-2583.
29. Zhao X, Zhou J, Sun L. Semiparametric transformation models with time-varying coefficients for recurrent and terminal events. *Biometrics*. 2011;67(2):404-414.
30. Cook RJ, Lawless JF. Marginal analysis of recurrent events and a terminating event. *Statist Med*. 1997;16(8):911-924.
31. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *J Royal Stat Soc Stat Methodol Ser B*. 2000;62(4):711-730.
32. Wang M-C, Qin J, Chiang C-T. Analyzing recurrent event data with informative censoring. *J Am Stat Assoc*. 2001;96(455):1057-1065.
33. Sun L, Kang F. An additive-multiplicative rates model for recurrent event data with informative terminal event. *Lifetime Data Anal*. 2013;19(1):117-137.
34. Zhao XB, Zhou X, Wang JL. Semiparametric model for recurrent event data with excess zeros and informative censoring. *J Stat Plan Inference*. 2012;142(1):289-300.
35. Huang C-Y, Qin J, Wang M-C. Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics*. 2010;66(1):39-49.

36. Zhu L, Sun J, Tong X, Srivastava DK. Regression analysis of multivariate recurrent event data with a dependent terminal event. *Lifetime Data Anal*. 2010;16(4):478-490.
37. Zhao X, Liu L, Liu Y, Xu W. Analysis of multivariate recurrent event data with time-dependent covariates and informative censoring. *Biometrical Journal*. 2012;54(5):585-599.
38. Ye Y, Kalbfleisch JD, Schaubel DE. Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*. 2007;63(1):78-87.
39. Chen C-M, Chuang Y-W, Shen P-S. Two-stage estimation for multivariate recurrent event data with a dependent terminal event. *Biometrical Journal*. 2015;57(2):215-233.
40. Chen C-M, Shen P-S, Chuang Y-W. The partly Aalen's model for recurrent event data with a dependent terminal event. *Statist Med*. 2016;35(2):268-281.
41. Aalen O. A model for nonparametric regression analysis of counting processes. In: Klonecki W, Kozek A, Rosiński W, eds. *Mathematical Statistics and Probability Theory*. New York, NY: Springer; 1980:1-25.
42. Li QH, Lagakos SW. Use of the Wei–Lin–Weissfeld method for the analysis of a recurring and a terminating event. *Statist Med*. 1997;16(8):925-940.
43. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc*. 1989;84(408):1065-1073.
44. González JR, Fernandez E, Moreno V, et al. Sex differences in hospital readmission among colorectal cancer patients. *J Epidemiol Community Health*. 2005;59(6):506-511.
45. Ghosh D, Lin DY. Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics*. 2003;59(4):877-885.
46. Hsieh J-J, Ding AA, Wang W. Regression analysis for recurrent events data under dependent censoring. *Biometrics*. 2011;67(3):719-729.
47. Zhao HZ, Li Y, Sun J. Analyzing panel count data with a dependent observation process and a terminal event. *Can J Stat*. 2013;41(1):174-191.
48. Rogers JK, Yaroshinsky A, Pocock SJ, Stokar D, Pogoda J. Analysis of recurrent events with an associated informative dropout time: application of the joint frailty model. *Statist Med*. 2016;35(13):2195-2205.

How to cite this article: Charles-Nelson A, Katsahian S, Schramm C. How to analyze and interpret recurrent events data in the presence of a terminal event: An application on readmission after colorectal cancer surgery. *Statistics in Medicine*. 2019;1–27. https://doi.org/10.1002/sim.8168

APPENDIX

A.1 | Description of the dataset

Form of the dataset included in the frailtypack and used in the application

	id	enum	t.start	t.stop	time	event	chemo	sex	death
1	1	1	0	24	24	1	Treated	Female	0
2	1	2	24	457	433	1	Treated	Female	0
3	1	3	457	1037	580	0	Treated	Female	0
6	3	1	0	15	15	1	NonTreated	Male	0
7	3	2	15	783	768	0	NonTreated	Male	1
23	8	1	0	1466	1466	0	NonTreated	Female	1

"id" is the subject's number. "enum" is the cumulative number of lines for each patient. "t.start" is the time of the previous event (0 is the entrance into the study). "t.stop" is the time of occurrence for the current event and could be recurrence (readmission) or a terminal event (death) or censoring (lost to follow up independent from death and covariates). "event" is whether the subject had a new readmission (recurrence) at the time "t.stop" ("event" = 1). "time" is the number of days between two events (gap time). "sex" is the gender and "chemo" is whether the subject received the treatment (covariates). "death" = 1 indicates that the subject was dead at the time "t.stop" and the event is considered to be the terminal event. In this dataset, a patient with no readmission has only one line (patient id 8). The others have the number of readmission+1 lines, the last line corresponding to the time of death if "death" = 1 (eg, id = 3) or censoring if "death" = 0 (eg, id = 1). The complete dataset is available in the frailtypack package in the R software.

A.2 | R code for Miloslavsky's model

The second step is to introduce weight to the coxph function. First, we modeled the censoring mechanism and setting the database in the required format. We did not include covariates in the model because we assume that censoring is independent of covariates:

```
require(rms)
require(survival)
require(stringr)

#Calculation of cumulative sum of event per patients
readmission$nb_event_cum<-unlist(tapply(readmission$event,readmission$id,cumsum))
# 1: Modeling censoring time
#extracts the last line for each subject
last<-readmission[!duplicated(readmission$id,fromLast = T),]
#creates censoring indicator
last$cens<-ifelse(last$death==0,1,0)
#model for censoring
km.cens <- cph(Surv(t.stop,cens)~1, last, conf.type="none",surv=TRUE)

# 2: Computing weights Gj
#subset of patients experiencing the terminal event
last2<-subset(last,death==1,select=c("id","cens","t.stop","chemo","sex","nb_event_cum","t
    .stop"))
#order survival time from
last2<-last2[order(last2$t.stop),]
#estimate survival at time points
last2$Gj <-survest(km.cens,times=last2[ , "t.stop"],se.fit=FALSE)$surv[1,]
```

```
# Creating a database with only subjects who died, for each of these individuals the
    interval from time of death until the longest time of event or censoring in the
    dataset is divided in subintervals. The boundaries of these intervals are the event
    and censoring times of all subjects in the dataset beyond the death time.
last2$t.start<-last2$time_cens
last2$t.stop1<-max(readmission$t.stop)
last2_long<-survSplit(last2, cut=sort(unique(readmission$t.stop)), end="t.stop1",start="t
    .stop", event="cens",id = "id1")

# Computing the weights for patients experiencing the terminal event
last2_long$Gt <-survest(km.cens,times=last2_long[ , "t.stop1"],se.fit=FALSE)$surv[1,]
last2_long$weights <- last2_long$Gt / last2_long$Gj
last2_long$event<-0
last2_long<-last2_long[,c("id","t.stop","t.stop1","event","chemo","sex","nb_event_cum","
    weights")]
names(last2_long)<-c("id","t.start","t.stop","event","chemo","sex","nb_event_cum","
    weights")
# Merging the original data with the data containing weights for patients experiencing
    the terminal event for coxph
readmission$weights=1
data.final <- rbind(readmission[ , c("id","t.start","t.stop","event","chemo","sex","
    nb_event_cum","weights")], last2_long)

#3 Model for recurrent events
ag<-coxph(Surv(t.start,t.stop,event)~sex+chemo+nb_event_cum,data=data.final,weights=
    weights,robust=T)
```

The cluster term is used to perform a sandwich variance to take into account the dependence between recurrent events and avoid estimation bias.

```
                coef exp(coef) se(coef) robust se     z       p
sexFemale    -0.2747    0.7598   0.1031    0.1013 -2.71 0.00667
chemoTreated -0.3401    0.7117   0.1003    0.0999 -3.40 0.00066
nb_event_cum  0.1781    1.1949   0.0090    0.0112 15.84 < 2e-16
```

A.3 | R code for the Liu's model

```
frailtyPenal(Surv(time, event) ~ cluster(id) + sex +
    chemo + terminal(death), formula.terminalEvent = ~sex + chemo,
    data = readmission, recurrentAG = FALSE, n.knots = 14, kappa = c(9.55e+09,1.41e+12))
```

Kappa's parameters are the smoothing parameters of the penalized likelihood. They were provided by fitting two independent models for recurrent events and terminal event processes using the function `frailtyPenal` with the cross validation method (`cross-validation method=T`).[15] The `recurrentAG` option indicates the chosen time scale, if `TRUE`, the calendar time scale is used and the required program is `Surv(t.start,t.stop,event)`. In this example, we used the gap time (`recurrentAG=FALSE`; hence, the variable "time" (time between two successive recurrent events) is used. `n.knots` corresponds to the number of knots to use (see the work of Rondeau et al[15]).

```
Recurrences:
-------------
                  coef exp(coef) SE coef (H) SE coef (HIH)        z
sexFemale    -0.413080  0.661609    0.146792      0.146792 -2.81404
chemoTreated -0.207344  0.812740    0.141787      0.141787 -1.46236
                     p
sexFemale    0.0048922
chemoTreated 0.1436400

Terminal event:
----------------
                  coef exp(coef) SE coef (H) SE coef (HIH)         z
sexFemale    -0.113227  0.892948    0.282761      0.282761 -0.400434
chemoTreated  0.631288  1.880030    0.296598      0.296598  2.128432
                    p
sexFemale    0.688840
chemoTreated 0.033301

 Frailty parameters:
   theta (variance of Frailties, w): 0.936374 (SE (H): 0.0979091 ) p = 0
   alpha (w^alpha for terminal event): 1.84028 (SE (H): 0.329016 ) p <0.0001
```

A.4 | R code for the Ghosh's model: IPSW procedure

```
#1: Modeling the death
#Compute a cox for death
ghosh<-readmission[!duplicated(readmission$id,fromLast = T),]
coxsurv<-cph(Surv(t.stop,death)~chemo+sex, ghosh,surv=TRUE, conf.type="none")
#2: Computation of weights
#setting up the dataset for IPSW procedure: For each individual in the data set the
    interval from time origin until time to event or censoring is divided in subintervals.
     The boundaries of these intervals are the event and censoring times of all subjects
    in the dataset
dfrm1<-survSplit(readmission, cut=sort(unique(readmission$t.stop)), end="t.stop",start="t
    .start", event="event",id = "id1")
```

```
#create a dataset with all combinations of covariates
newdata=expand.grid(chemo=levels(factor(readmission$chemo)),sex=levels(factor(
    readmission$sex)))

#create a database from all #Estimate survival probability at each event time for each %
    combinations of the covariates
tempo<-survest(coxsurv,newdata=newdata,times=sort(unique(readmission$t.stop)))$surv

#calculate survival time for all combination at each time
tempo_names<-colnames(tempo)  #colnames are event/censor times
tempo<-data.frame(tempo)
names(tempo)<-str_trim(tempo_names)   #remove spaces
```

```
#calculation of wi according to covariates values: for each patient and each time,
    extract the corresponding survival probability acoording to covariates values

for(i in 1:dim(dfrm1)[1])
{
place<-which(newdata[,"chemo"]==dfrm1[i,"chemo"]&newdata[,"sex"]==dfrm1[i,"sex"]) #find
    the row number corresponding to the combination of covariates for patient i
if(dfrm1[i,"t.stop"]==0){dfrm1[i,"surv"]<-NA}
{column<-which(names(tempo)==dfrm1[i,"t.stop"])   #find the column number corresponding
    to the time when the survival have to be calculated for patient i
dfrm1[i,"surv"]<-tempo[place,column]}
}

dfrm1$wi<-1/dfrm1$surv     #calculation of weights

#3: Incorporation of weights in the model
ghosh<-coxph(Surv(t.start,t.stop,event)~sex+chemo+cluster(id),data=dfrm1,weights=wi)
```

```
                coef exp(coef) se(coef) robust se     z      p
sexFemale    -0.4869    0.6145   0.0935    0.1792 -2.72 0.0066
chemoTreated -0.4882    0.6137   0.0894    0.1751 -2.79 0.0053
```

A.5 | R code for the Huang's model

```
require(reReg)
data(readmission)
set.seed(123)
summary(reReg(reSurv(t.start, id, death, t.stop)~sex+chemo,data= readmission,method="HW")
```

```
Method: Huang-Wang method

Coefficients (rate):
             Estimate StdErr z.value p.value
sexFemale      -0.774  0.320  -2.417   0.016 *
chemoTreated   -0.264  0.303  -0.871   0.384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Coefficients (hazard):
             Estimate StdErr z.value p.value
sexFemale       0.515  0.428   1.204   0.229
chemoTreated    1.268  0.504   2.514   0.012 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A.6 | R code for the Mao's model

```
#Create the composite variable
readmission$event_comp<-ifelse(readmission$death==1,1,ifelse(readmission$event==1,2,0))
#Remove missing data
readmission_ss_na<-na.omit(readmission[,c("id","t.stop","event_comp","sex_num","chemo_num
    ")])
readmission_ss_na$sex<-readmission_ss_na$sex_num
readmission_ss_na$chemo<-readmission_ss_na$chemo_num
id=readmission_ss_na$id
time=readmission_ss_na$t.stop
status=readmission_ss_na$event_comp
Z=as.matrix(readmission_ss_na[,c("sex","chemo")])
```

```
CompoML(id = id, time = time, status = status, Z = Z, w = c(1,
    1), ep = 1e-04)

Estimates for Regression parameters:

      Estimate       se z.value  p.value
sex   -0.41456  0.14647 -2.8303 0.004650 **
chemo -0.42706  0.14409 -2.9638 0.003038 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Point and interval estimates for mean ratios:

      Mean Ratio 95% lower CL 95% higher CL
sex    0.6606308    0.4957729     0.8803084
chemo  0.6524236    0.4919028     0.8653267
```

3.2 Simulations des données

Les données de l'article "How to analyze and interpret recurrent events data in the presence of a terminal event : An application on readmission after colorectal cancer surgery" ont été simulées en deux étapes et suivant un processus de Poisson homogène : i) simulation du temps de survie ; ii) simulation des temps des événements récurrents.

3.2.1 Simulation de l'événement terminal

L'événement terminal est simulé à partir du modèle de Cox [42] :

$$h(t|Z) = h_0(t)\omega exp\left(\sum_{j=1}^{p}\beta_j Z_j\right)$$

où $Z_1, .., Z_p$ sont les covariables, $\beta_1, .., \beta_p$ sont les coefficients de régression associés et $h_0(t)$ est la fonction de risque instantané de base dépendant du temps et $h(t|Z)$ est la fonction de risque instantané. Afin d'introduire de la dépendance entre les événements récurrents et le temps de l'événement terminal, nous avons introduit dans le modèle de Cox un effet aléatoire ω suivant une loi Gamma de moyenne 1 et de variance θ_1.

La fonction de distribution $H(t|Z)$ du modèle est :

$$H(t|Z) = 1 - exp\left(-H_0(t)\omega exp\left(\sum_{j=1}^{p}\beta_j Z_j\right)\right)$$

où $H_0(t)$ est la fonction de risque instantané de base cumulée.
Soit Y une variable aléatoire avec H comme fonction de distribution, on a alors $U = H(Y)$ qui suit une loi uniforme sur l'intervalle $[0, 1]$ et donc (1-U) suit aussi une loi uniforme sur l'intervalle $[0, 1]$.
Soit T le temps de survie du modèle de Cox, on en déduit alors que :

$$U = 1 - exp\left(-H_0(T)\omega exp\left(\sum_{j=1}^{p}\beta_j Z_j\right)\right) \sim U[0, 1]$$

Si pour tout t, $h_0(t) > 0$, on obtient :

$$T = H_0^{-1}\left[-log(1-U)\omega^{-1}exp\left(-\sum_{j=1}^{p}\beta_j Z_j\right)\right]$$

où U est une variable aléatoire suivant une loi uniforme sur l'intervalle $[0, 1]$. Dans cette expression, $log(1-U)$ peut être remplacée par $log(U)$. On obtient alors :

$$T = H_0^{-1}\left[-log(U)\omega^{-1}exp\left(-\sum_{j=1}^{p}\beta_j Z_j\right)\right]$$

Cette méthode de simulation est appelée la méthode d'inversion [53].

Nous avons choisi une distribution exponentielle de paramètre λ pour la fonction de risque de base afin d'avoir un risque instantané de base constant au cours du temps :

$$h_0(t) = \lambda, \lambda > 0$$

On obtient donc :

$$H_0(t) = \lambda t \quad \text{et} \quad H_0^{-1}(t) = \lambda^{-1} t$$

Ainsi :

$$T = -\frac{log(U)}{\lambda \omega exp\left(\sum_{j=1}^{p} \beta_j Z_j\right)}$$

où $U \sim U[0,1]$.
D'autres distributions telles que Weibull [41] ou Gompertz [54] aurait pu être utilisées pour la fonction de risque instantané de base, dans ces cas la fonction de risque instantané de base serait dépendante du temps.

Le temps de censure C est simulé à partir d'une loi uniforme sur l'intervalle [0,2] afin de respecter l'hypothèse de censure indépendante et non informative. Le temps de survie est alors $X = min(T, C)$.

3.2.2 Simulation des temps des événements récurrents

Les temps des événements récurrents sont simulés à partir de modèles multiplicatifs basés sur l'intensité et sont simulés selon l'échelle de temps calendaire, ie, temps entre le début de l'étude et l'événement.

Le modèle multiplicatif à intensité proportionnelle s'écrit alors :

$$r(t|Z') = r_0(t)\omega_1\omega_2 exp\left(\sum_{j=1}^{q} \beta'_j Z'_j\right)$$

où q est le nombre de covariables incluses dans le modèle, β' et Z' sont respectivement les paramètres de régression et les covariables incluses dans le modèle. Les covariables incluses dans le modèle peuvent être différentes de celles incluses dans le modèle utilisé pour simuler les temps de l'événement terminal. Afin de simuler de la dépendance entre les événements récurrents et l'événement terminal ainsi que de la dépendance intra-récurrence, nous avons inclus deux effets aléatoires ω_1 et ω_2 suivant chacun une loi Gamma de moyenne 1 et de variance θ_1 et θ_2 respectivement.

La fonction d'intensité de base choisie constante, le processus d'événement récurrent peut être considéré comme un processus de Poisson homogène. Pour générer les différents temps des événements récurrents, nous avons utilisé la méthode d'inversion, décrite dans le paragraphe

précédent. Les temps des événements récurrents sont simulés à partir de l'algorithme suivant :

1. On pose T_0=0
2. Pour l'événement récurrent k, simulation de $V_k \sim U[0,1]$.
3. Soit $W_k = F_k^{-1}(V_k)$ et $T_k = T_{k-1} + W_k$
4. Si $T_k < X$ alors on réitère les étapes 2, 3 et 4 pour l'événement suivant. Sinon, le $k^{ème}$ temps est censuré et $T_k = X$

où W_k est le temps entre les événements k et $k-1$ et X est le temps de survie simulé dans l'étape précédente. $F_k(.)$ est la fonction de distribution de la variable aléatoire W_k.

Nous avons choisi une distribution exponentielle de paramètre λ' pour la fonction de base de l'intensité cumulée. Comme le processus est de Poisson est homogène, on obtient alors :

$$T_k = T_{k-1} + W_k = T_{k-1} - \frac{log(V_k)}{\lambda' \omega_1 \omega_2 exp\left(\sum_{j=1}^{q} \beta'_j Z'_j\right)}$$

où $V_k \sim U[0,1]$.

3.2.3 Simulations supplémentaires

3.2.3.1 Simulation d'un processus de Poisson non-homogène

Le processus de Poisson homogène suppose que la fonction de risque instantané de base est constante au cours du temps, c'est à dire que la probabilité de faire un événement lorsque toutes les covariables sont égales à 0 est la même quel que soit t. En réalité cette probabilité peut varier au cours du temps.
Comme pour les simulations d'un processus de Poisson homogène, la première étape consiste à simuler le temps de l'événement terminal (cf paragraphe 3.2.1).
Le modèle multiplicatif pour les événements récurrents à intensité proportionnelle utilisé pour les simulations suivantes est :

$$r_{NH}(t|Z') = r_{0NH}(t)\omega_1\omega_2 exp\left(\sum_{j=1}^{q} \beta'_j Z'_j\right)$$

Les temps d'événements récurrents sont simulés par la méthode d'inversion et selon l'échelle de temps calendaire. On peut ainsi écrire le temps de l'événement k en fonction des temps entre les événements : $W_k = T_k - T_{k-1}$ avec $T_0 = 0$ et $T_k = \sum_{l=1}^{k} W_l$. La distribution des W_k dépend du temps de l'événement précédant T_{k-1} puisque le risque de l'événement dépend du temps total. La fonction de répartition $F_k(.)$ du temps entre les événements récurrents W_k s'écrit :

$$\begin{aligned} F_k(t) = & \quad 1 - exp\left(- \int_{T_{k-1}}^{T_{k-1}+w} r_{NH}(u|Z')du \right) \\ = & \quad 1 - exp\left(- \int_0^w r_{0NH}(T_{k-1}+t|Z')\omega_1\omega_2 exp\left(\sum_{j=1}^q \beta'_j Z'_j \right) dt \right) \\ = & \quad 1 - exp\left(-\omega_1\omega_2 exp\left(\sum_{j=1}^q \beta'_j Z'_j \right) (R_{0NH}(T_{k-1}+w|Z') - R_{0NH}(T_{k-1}|Z')) \right) \end{aligned}$$

où $R_{0NH}(t)$ est la fonction instantanée de base cumulée.

Soit T_k le temps du $k^{ème}$ événement, on en déduit alors que :

$$U_k = 1 - exp\left(-\omega_1\omega_2 exp\left(\sum_{j=1}^q \beta'_j Z'_j \right) (R_{0NH}(T_{k-1}+w|Z') - R_{0NH}(T_{k-1}|Z')) \right) \sim U[0,1]$$

Si pour tout t, $r_{0NH}(t) > 0$, on obtient :

$$T_k = T_{k-1} + w = R_{0NH}^{-1}\left[-log(1-U_k)\omega_1^{-1}\omega_2^{-1} exp\left(-\sum_{j=1}^p \beta'_j Z'_j \right) + R_{0NH}(T_{k-1}|Z') \right]$$

où U_k est une variable aléatoire suivant une loi uniforme sur l'intervalle $[0,1]$. Dans cette expression, $log(1-U_k)$ peut être remplacée par $log(U_k)$. On obtient alors :

$$T_k = R_{0NH}^{-1}\left[-log(U_k)\omega_1^{-1}\omega_2^{-1} exp\left(-\sum_{j=1}^p \beta'_j Z'_j \right) + R_{0NH}(T_{k-1}|Z') \right]$$

L'algorithme de calcul des temps des événements récurrents est le suivant :

1. On pose T_0=0
2. Pour l'événement récurrent k, simulation de $U_k \sim U[0,1]$.
3. $T_k = R_{0NH}^{-1}\left[-log(U_k)\omega_1^{-1}\omega_2^{-1} exp\left(-\sum_{j=1}^p \beta'_j Z'_j \right) + R_{0NH}(T_{k-1}|Z') \right]$
4. Si $T_k < X$ alors on réitère les étapes 2, 3 et 4 pour l'événement suivant. Sinon, le $k^{ème}$ temps est censuré et $T_k = X$

On pose : $r_{0NH}(t) = rt$, on obtient : $R_{0NH}(t) = \frac{r}{2}t^2$ et $R_{0NH}^{-1}(t) = \sqrt{\frac{1}{\frac{r}{2}}t}$. Ainsi

$$T_k = \sqrt{\frac{1}{\frac{r}{2}}\left(-log(U_k)\omega_1^{-1}\omega_2^{-1} exp\left(-\sum_{j=1}^p \beta'_j Z'_j \right) + \frac{r}{2}T_{k-1}^2 \right)}$$

Pour la fonction de risque de base pour les événements récurrents, nous avons choisi $r = 4$, les autres paramètres sont fixés comme dans l'article présenté dans la section 3.1.

Résultats

La Table 3.1 montre que le modèle de Miloslasvky [55] semble fournir des résultats fortement biaisés dans tous les scénarios. Cela peut être dû à une mauvaise spécification de la façon dont la dépendance intra-récidive est prise en compte. En effet, dans le modèle de Miloslavsky [55], la dépendance intra-récurrence est expliquée par le nombre cumulé d'événements à chaque instant (covariable temporelle) tandis qu'un effet aléatoire est utilisé dans la simulation des données et donc les covariables dépendantes du temps semblent ne pas bien saisir la dépendance intra-récurrence. De plus, dans les simulations, la covariable Z_1 est associée à la fois aux récurrences et à l'événement terminal. Cependant, le modèle de Miloslavsky [55] n'a modélisé que la distribution de censure et ne tient pas compte de l'association entre la covariable Z_1 et l'événement terminal. Cela signifie que si une covariable est associée aux deux processus, le modèle de Miloslavsky [55] fournit des résultats fortement biaisés. Il n'est donc pas approprié pour analyser les événements récurrents en présence d'événements terminaux.

Lorsque les événements récurrents sont indépendants entre eux et qu'ils sont aussi indépendants des événements terminaux, les modèles de Ghosh [56] et de Huang [57] fournissent des résultats similaires en termes de biais pour les coefficients β_1 et β_2. Cela montre que le modèle conjoint de Huang [57] fonctionne bien même lorsque les récurrences et les événements terminaux sont indépendants. Cependant, le modèle de Ghosh [56] offre une probabilité de couverture plus faible que le modèle de Huang [57] dans ce cas, mais il est plus parcimonieux que les modèles conjoints. De plus, lorsque la dépendance entre les événements récurrents et terminaux est plus élevée, les résultats sont plus biaisés pour le modèle de Ghosh [56]. En effet, dans ce modèle, l'événement terminal est modélisé indépendamment des événements récurrents. Il modélise d'abord l'événement terminal comme un modèle indépendant, puis inclut des poids dans le modèle pour les récurrences, il ne tient donc pas compte de la dépendance entre ces deux processus. Il en est de même lorsque la dépendance intra-récurrence augmente puisque le modèle de Ghosh [56] ne prend pas en compte cette information dans l'estimation des coefficients.

Le modèle de Huang [57] fournit des résultats moins biaisés pour les coefficients β_1 et β_2 que le modèle de Liu [58][59]. En effet, le modèle de Huang [57] utilise la vraisemblance partielle pour estimer les coefficients associés aux événements récurrents, ce qui ne nécessite pas l'estimation de la fonction de base. Au contraire, le modèle de Liu [58][59] estime les paramètres à partir de la vraisemblance complète et requière ainsi l'estimation de celle-ci. Pour cela, la fonction de risque instantané de base dans le modèle de Liu [58] est estimée par la méthode de Breslow [37] mais nous avons utilisé la fonction de risque de base approximée par la méthode des splines [59]. Les biais d'estimation du modèle de Liu, avec la fonction de risque de base estimée par la méthode des splines, peuvent être expliqués par le fait que le nombre de noeuds (6 noeuds) est élevé par rapport au nombre de patients. En effet, le modèle nécessite d'estimer plus de paramètres que les autres modèles. Cependant, nous voyons qu'augmenter le nombre de patients réduit le biais. Une autre solution pourrait être de réduire le nombre de noeuds

lorsqu' augmenter le nombre de patients n'est pas possible.

Lorsque le nombre de sujets est faible, le coefficient β_3 est très surestimé par le modèle de Huang [57] mais le biais diminue lorsque le nombre de sujets augmente. Cela peut être dû à la méthode d'estimation du modèle de Huang [57]. En effet, il estime d'abord les paramètres des récidives sans tenir compte de la dépendance intra-récidives puis estime la variable latente et intègre la valeur de la variable latente dans le modèle de l'événement terminal. Concernant, l'estimation du coefficient β_3 dans le modèle de Liu [58][59], il semble bien estimé avec un biais faible lorsqu'il n'y a pas de dépendance entre les événements récurrents et terminal. Cependant, lorsque la dépendance entre les deux processus ainsi que la dépendance intra-récurrence augmente, le coefficient semble sous-estimé. Cela est certainement dû au fait que les coefficients sont estimés simultanément à partir de la maximisation de la vraisemblance et que la méthode des splines ne semble pas bien approximer la fonction de base, ce qui impacte les coefficients associés aux récurrences mais aussi ceux associés à l'événement terminal lorsque les deux processus sont dépendants.

Le modèle de Mao [60] fournit des résultats biaisés pour tous les scénarios. En effet, les événements récurrents et terminaux sont combinés en un seul point final.

3.3 Conclusion et discussion

De nombreuses méthodes ont été développées pour analyser les événements récurrents en présence d'un événement terminal. Nous avons ainsi vu dans ce chapitre qu'ils peuvent être classifiés en quatre catégories : i) modèles conditionnels non joints ; ii) modèles conditionnels joints ; iii) modèles marginaux non joints et iv) modèles marginaux joints. Le choix de l'utilisation du modèle dépend de la question posée, du processus biologique des événements récurrents et de la dépendance entre les événements récurrents et l'événement terminal. Cependant, dans certains cas, l'événement terminal peut ne pas être observé du fait de la survenue d'un autre événement, appelé risque compétitif. En effet, il est possible que plusieurs types d'événements terminaux puissent survenir, mais seulement la survenue du premier événement est observée. Par exemple, en cancérologie, il est courant de ne s'intéresser qu'aux décès liés au cancer. Or certains patients peuvent décéder d'autre chose. Dans ce cas l'événement terminal d'intérêt ne peut pas être observé. Il est donc important, de pouvoir prendre en compte la compétition sur l'événement terminal lorsque celle-ci existe.

TABLE 3.1 – Résultats des simulations à partir d'un processus de Poisson non-homogène : approches conditionnelles

			$\beta_1 = 1$, $\beta_2 = -0.5$, $\beta_3 = 0.7$															
			Miloslavsky avec covariable dépendante du temps				**Miloslavsky sans covariable dépendante du temps**				**Liu**							
n	θ_1	θ_2	β_1**(SE)**	**cp**	β_2**(SE)**	**cp**	β_1**(SE)**	**cp**	β_2**(SE)**	**cp**	β_1**(SE)**	**cp**	β_2**(SE)**	**cp**	β_3**(SE)**	**cp**	θ**(SE)**	α**(SE)**
50	0	0	-0,09 (0,33)	5,20%	-0,13 (0,33)	61,20%	0,24 (0,59)	32,20%	-0,49 (0,57)	65,00%	1,18 (0,30)	76,85%	-0,37 (0,33)	89,20%	0,74 (0,40)	93,21%	0,04 (0,13)	10,10 (4,22)
		1	-0,03 (0,46)	14,60%	-0,21 (0,52)	68,00%	0,25 (0,77)	36,80%	-0,48 (0,85)	53,00%	1,13 (0,59)	75,39%	-0,36 (0,63)	80,47%	0,69 (0,47)	80,08%	0,63 (0,36)	1,88 (4,30)
		5	-0,02 (1,94)	29,40%	-0,33 (1,28)	65,20%	0,27 (2,85)	40,40%	-0,68 (2,18)	44,40%	0,91 (1,83)	69,59%	-0,60 (1,80)	73,73%	0,65 (0,45)	79,26%	1,17 (0,21)	0,22 (0,92)
	1	0	0,13 (0,92)	18,20%	-0,24 (0,95)	73,60%	0,37 (1,05)	45,80%	-0,52 (1,09)	69,60%	0,90 (0,58)	76,92%	-0,33 (0,62)	78,70%	0,52 (0,84)	71,01%	0,66 (0,36)	3,26 (4,77)
		1	0,15 (1,58)	25,20%	-0,28 (0,96)	78,40%	0,34 (1,73)	46,40%	-0,55 (1,25)	61,00%	0,85 (1,62)	75,00%	-0,32 (0,75)	77,94%	0,51 (0,77)	69,85%	0,94 (0,26)	1,16 (2,29)
		5	0,22 (2,65)	42,00%	-0,58 (2,39)	72,40%	0,45 (3,16)	42,40%	-0,86 (3,13)	50,20%	0,66 (1,82)	60,14%	-1,00 (3,25)	69,59%	0,46 (0,62)	74,32%	1,26 (0,24)	0,50 (0,38)
250	0	0	-0,09 (0,15)	0,00%	-0,17 (0,16)	23,40%	0,28 (0,23)	2,00%	-0,52 (0,24)	67,00%	1,08 (0,12)	80,31%	-0,46 (0,14)	91,88%	0,70 (0,17)	94,69%	0,01 (0,03)	8,90 (2,30)
		1	-0,03 (0,23)	0,00%	-0,24 (0,24)	41,00%	0,25 (0,35)	8,20%	-0,51 (0,35)	50,60%	1,10 (0,20)	89,39%	-0,38 (0,23)	86,03%	0,71 (0,18)	78,21%	0,84 (0,10)	-0,02 (0,18)
		5	-0,03 (0,39)	3,60%	-0,27 (0,43)	44,00%	0,24 (0,57)	15,60%	-0,49 (0,61)	34,00%	1,11 (0,35)	72,20%	-0,36 (0,41)	67,63%	0,73 (0,17)	85,89%	1,27 (0,09)	0,18 (0,09)
	1	0	0,10 (0,18)	0,00%	-0,20 (0,18)	38,80%	0,35 (0,27)	8,20%	-0,49 (0,25)	73,20%	1,09 (0,29)	80,56%	-0,40 (0,25)	83,33%	0,72 (0,26)	93,89%	0,83 (0,09)	0,96 (0,34)
		1	0,12 (0,25)	2,60%	-0,24 (0,26)	47,60%	0,33 (0,36)	13,60%	-0,49 (0,37)	59,40%	1,04 (0,32)	82,72%	-0,35 (0,30)	80,10%	0,66 (0,20)	94,24%	1,05 (0,08)	0,55 (0,15)
		5	0,17 (0,41)	10,80%	-0,30 (0,41)	48,60%	0,37 (0,63)	27,20%	-0,51 (0,60)	40,20%	1,01 (0,49)	63,55%	-0,39 (0,44)	70,56%	0,60 (0,20)	89,25%	1,37 (0,12)	0,38 (0,08)

θ_1 et θ_2 représentent respectivement la dépendance entre l'événement terminal et les événements récurrents et la dépendance intra-récurrences, cp= probabilité de couverture, SE= écart-type,

TABLE 3.2 – Résultats des simulations à partir d'un processus de Poisson non-homogène : approches marginales

			$\beta_1 = 1$, $\beta_2 = -0,5$, $\beta_3 = 0,7$													
			Ghosh (IPSW)				**Huang**						**Mao**			
n	θ_1	θ_2	β_1**(SE)**	**cp**	β_2**(SE)**	**cp**	β_1**(SE)**	**cp**	β_2**(SE)**	**cp**	β_3**(SE)**	**cp**	β_1**(SE)**	**cp**	β_2**(SE)**	**cp**
50	0	0	1,02 (0,40)	85,40%	-0,51 (0,39)	86,60%	1,03 (0,84)	91,60%	-0,58 (0,88)	91,40%	0,66 (3,96)	98,40%	0,28 (0,27)	22,20%	-0,28 (0,28)	84,00%
		1	0,98 (0,77)	75,80%	-0,50 (0,79)	76,00%	1,00 (1,02)	90,20%	-0,50 (0,96)	93,60%	3,19 (23,37)	97,80%	0,28 (0,39)	42,20%	-0,25 (0,40)	81,20%
		5	0,95 (2,96)	73,40%	-0,70 (2,40)	71,40%	0,99 (2,68)	95,80%	-0,65 (2,05)	95,00%	4,67 (31,23)	97,60%	0,27 (0,60)	52,80%	-0,25 (0,61)	78,40%
	1	0	0,67 (1,05)	78,00%	-0,53 (1,09)	85,80%	0,94 (1,36)	89,80%	-0,62 (1,39)	91,40%	1,28 (21,90)	95,80%	0,31 (0,30)	36,00%	-0,26 (0,32)	84,00%
		1	0,66 (1,23)	81,20%	0,88 (1,86)	93,60%	-0,57 (1,43)	90,20%	1,56 (26,35)	96,80%	0,31 (0,43)	50,00%	-0,27 (0,43)	83,40%		
		5	0,71 (3,18)	72,00%	-0,86 (3,12)	77,60%	1,00 (3,04)	94,20%	-0,81 (3,00)	94,20%	4,23 (49,16)	95,80%	0,31 (0,66)	57,40%	-0,28 (0,63)	81,20%
250	0	0	1,00 (0,16)	90,20%	-0,52 (0,17)	91,20%	0,99 (0,49)	96,00%	-0,50 (0,55)	93,00%	0,68 (0,43)	94,40%	0,30 (0,12)	0,00%	-0,30 (0,12)	65,60%
		1	0,94 (0,38)	84,80%	-0,50 (0,40)	85,80%	1,02 (0,49)	93,00%	-0,50 (0,54)	92,80%	0,74 (0,44)	91,00%	0,28 (0,19)	5,80%	-0,29 (0,19)	75,80%
		5	0,89 (0,72)	81,20%	-0,48 (0,80)	80,40%	1,00 (0,62)	94,20%	-0,50 (0,69)	92,40%	0,76 (0,54)	96,40%	0,29 (0,31)	29,00%	-0,28 (0,34)	78,00%
	1	0	0,65 (0,28)	68,00%	-0,49 (0,26)	94,60%	0,94 (0,65)	92,20%	-0,49 (0,70)	90,40%	0,65 (0,53)	92,80%	0,33 (0,15)	0,40%	-0,29 (0,14)	69,00%
		1	0,62 (0,37)	79,00%	-0,50 (0,40)	91,40%	0,96 (0,72)	91,00%	-0,52 (0,73)	90,80%	0,70 (0,61)	93,80%	0,32 (0,19)	7,40%	-0,28 (0,21)	73,40%
		5	0,66 (0,68)	83,80%	-0,53 (0,66)	88,40%	1,01 (0,87)	89,00%	-0,50 (0,92)	89,20%	0,72 (0,68)	95,00%	0,34 (0,36)	38,40%	-0,29 (0,34)	83,60%

θ_1 et θ_2 représentent respectivement la dépendance entre l'événement terminal et les événements récurrents et la dépendance intra-récurrences. CP= probabilité de couvertu, SE= écart-type.

Chapitre 4

Analyse des réhospitalisations suivant une greffe de foie en présence de deux types d'evénements terminaux

De nombreux articles s'intéressent aux réhospitalisations après une greffe de foie avec un nombre croissant d'articles depuis les années 80 (Figure 4.1).

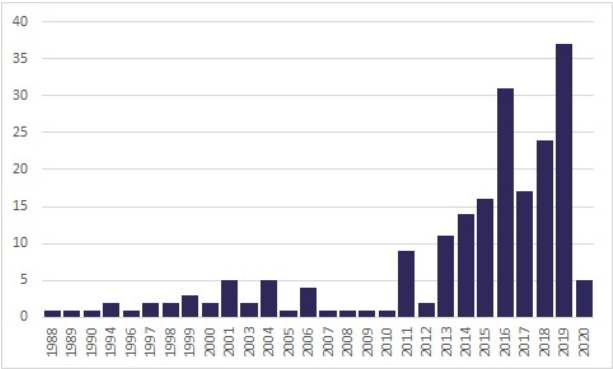


FIGURE 4.1 – Nombre d'articles traitant des réhospitalisations après une greffe de foie au cours du temps

Les articles sélectionnés sont ceux répondant à la requête PUBMED "readmission liver graft". La recherche a été effectuée le 28/02/2020

La quasi-totalité de ces articles (>99%) n'ont analysé que le délai de survenue de la première réhospitalisation ou de façon binaire (hospitalisation oui/non), très peu ont utilisés des méthodes adaptées aux événements récurrents (<0,5%) et aucun n'a pris en compte l'événement terminal.

Dans le chapitre précédent, nous avons vu les différentes méthodes pour analyser les événements récurrents en présence d'un événement terminal. Or dans la recherche médicale, il est possible que la survenue d'un autre événement empêche la survenue de l'événement terminal d'intérêt (Figure 4.2), c'est ce qu'on appelle un risque compétitif. Dans ce cas, le patient ne peut présenter qu'un seul des différents événements terminaux, et on s'intéresse alors à la survenue du premier événement terminal. Par exemple, en oncologie, il est de coutume d'évaluer

l'effet du traitement sur les rechutes mais aussi de regarder l'effet du traitement sur le décès par cancer. Mais dans ce cas, les patients peuvent décéder d'une cause autre que le décès par cancer, notamment d'un effet indésirable de la chimiothérapie par exemple. Analyser ce type de données en censurant les patients décédés d'une cause autre que le cancer violerait l'hypothèse de censure non-informative nécessaire à l'application des modèles présentés dans le chapitre précédent. Zeng [61] a développé un modèle joint permettant d'analyser plusieurs types d'événements récurrents en présence de plusieurs types dévénements terminaux. Cependant, il n'est pas possible avec son modèle d'évaluer la dépendance entre les événements récurrents et chaque type d'événement terminal qui est un élément important lorsque l'on souhaite évaluer les mécanismes d'une maladie. Le but de ce chapitre est de présenter le modèle que nous avons développé afin d'analyser les événements récurrents en présence de risques compétitifs sur l'événement terminal. Nous avons pour cela étendu le modèle proposé par Rondeau [59]. C'est un modèle joint à fragilité partagée permettant d'analyser simultanément l'effet des covariables sur les événements récurrents et les événements terminaux tout en permettant d'évaluer la dépendance entre les événements récurrents et chaque type d'événement terminal.

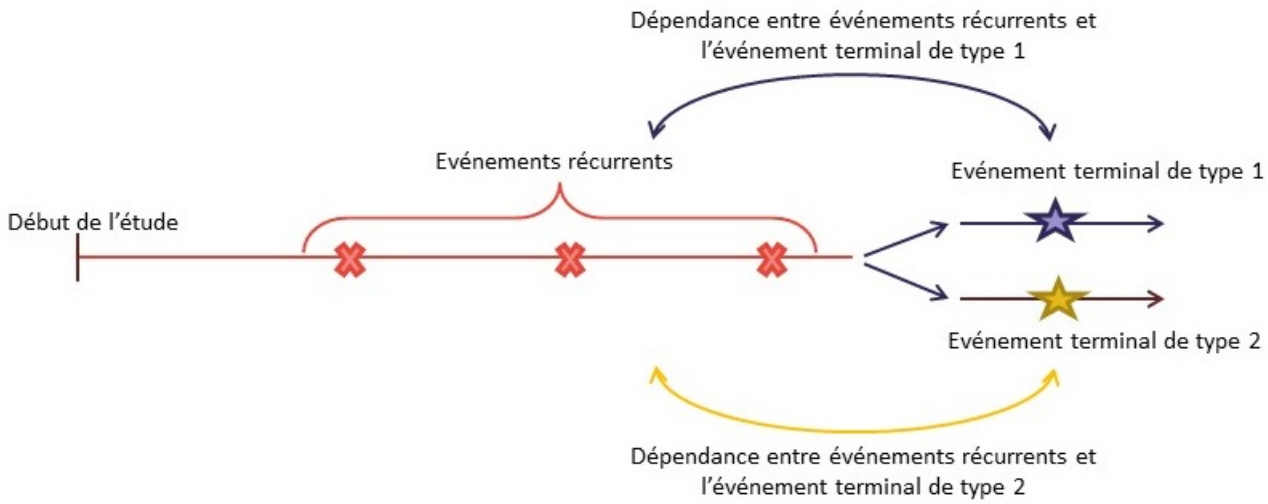


FIGURE 4.2 – Schéma d'événements récurrents en présence de deux types d'événements terminaux

Le développement de ce modèle a été motivé par les données des patients ayant subi une greffe de foie. L'objectif est alors d'évaluer l'effet du sexe et de l'âge au moment de la greffe sur les réhospitalisations ainsi que sur les décès liés aux maladies du foie.

4.1 La greffe de foie

Il existe plusieurs pathologies du foie plus ou moins graves qui peuvent augmenter le risque de cancer et ainsi nécessiter une greffe de foie. Ainsi, lorsque le foie n'est plus fonctionnel, la greffe de foie est possible.

La greffe d'un foie est une opération longue et complexe, pouvant durer jusqu'à une quinzaine d'heures, et se déroule en plusieurs étapes distinctes (Figure 4.3) : l'ablation du foie malade, prélèvement du foie chez le donneur pour ensuite le greffer chez le receveur et finalement s'assurer qu'il fonctionne normalement.

Lorsque l'ensemble du foie malade est retiré, la greffe du foie sain peut débuter. Le chirurgien place d'abord le greffon dans l'abdomen et relie en priorité les vaisseaux sanguins pour permettre au foie d'être à nouveau alimenté en sang [62]. Ensuite, les différents vaisseaux transportant la bile sont raccordés. Une fois la greffe finie, le chirurgien s'assure que la circulation du sang et de la bile est bien établie. Ces deux étapes doivent être parfaitement synchronisées, car le foie greffé ne doit pas rester trop longtemps privé de sang pour pouvoir à nouveau fonctionner normalement. Une fois greffés, les patients peuvent mener une vie normale, sous réserve d'éviter

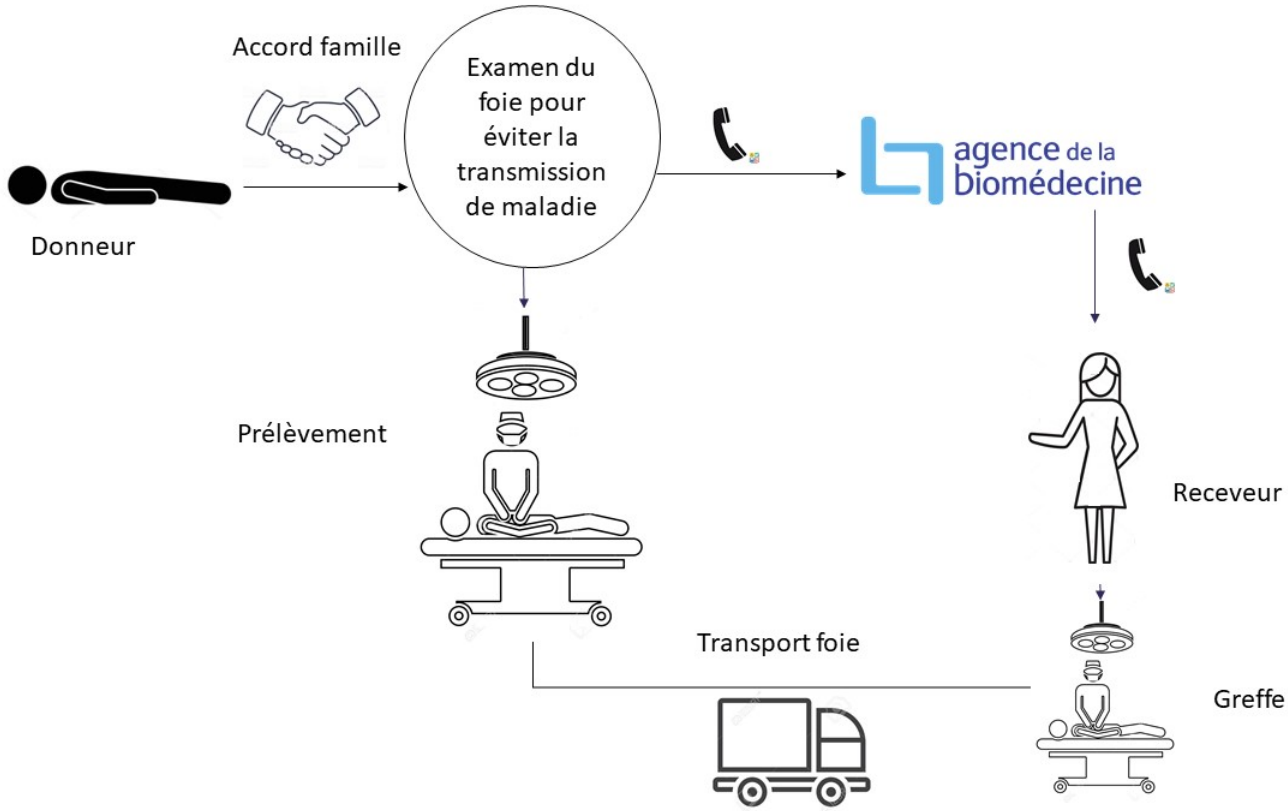


FIGURE 4.3 – Les différentes étapes de la greffe de foie

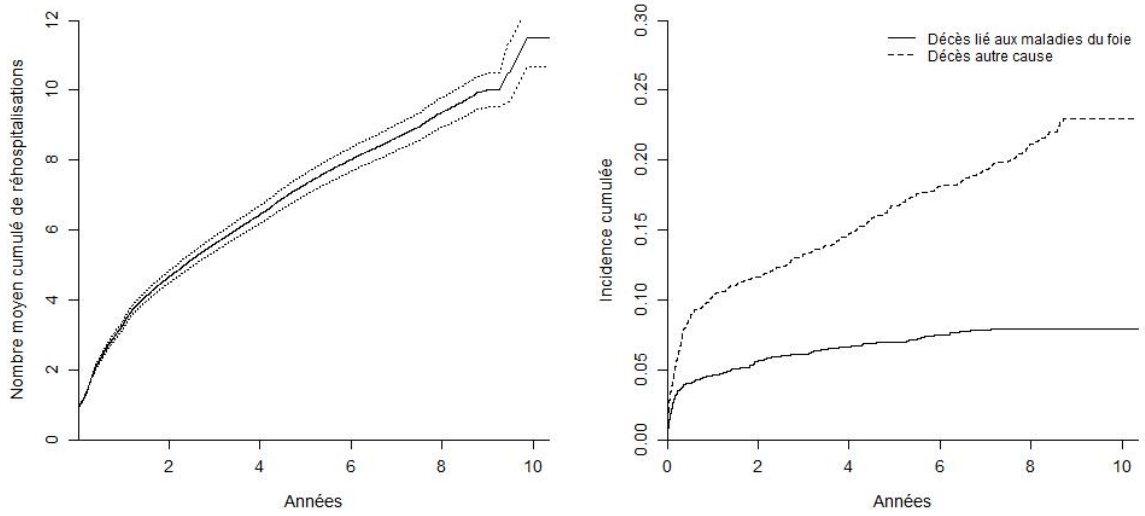
les boissons alcoolisées[63]. Mais, afin d'éviter un rejet tardif du greffon, ces patients doivent prendre un traitement immunosuppresseur à vie, même s'il peut être allégé après six mois ou un an. Ce type de traitement a pour effet de diminuer les défenses immunitaires de l'organisme. Les patients greffés sont alors plus à risque d'avoir des infections virales et de développer certaines tumeurs. Il se peut aussi qu'il ne soit pas bien supporté par le rein et augmente ainsi les risques cardiovasculaires. Toutes ces complications amènent souvent les patients à être réhospitalisés. Ainsi le taux de réhospitalisations peut être vu comme un indicateur de l'état de santé du patient après la greffe et de la qualité de la greffe.

Extraction et sélection des données

Les patients greffés du foie en 2008 ou 2009 ont été identifiés à l'aide des codes CCAM HLEA001 et HLEA002, qui correspondent à une greffe de foie total et à une greffe de foie réduit. La réhospitalisation de ces patients de 2008 à 2016 ainsi que le sexe, l'âge au moment de la transplantation et les informations sur le décès au cours de la réhospitalisation ont été extraites. Comme seul le décès à l'hôpital est enregistré dans le PMSI, le taux de mortalité est sous-estimé.
La réhospitalisation avec une durée de zéro jour, la réhospitalisation avant la greffe et la réhospitalisation pour dialyse ou traitement contre le cancer (chimiothérapie ou irradiation) sont retirées de la base de données et les patients en vie sont censurés à la date du 12/12/2016.

Description des données

Un total de 1933 patients, enregistrés dans le PMSI ont subi une greffe du foie en 2008 ou en 2009. Il y a 1361 hommes (70,4%) et l'âge moyen lors de la greffe est de 49 ± 16 ans. Au total, 12748 réhospitalisations ont été enregistrées avec une moyenne de 6,6 ± 7,8 réhospitalisations par patient allant de 0 à 109. 135 patients (7 %) sont décédés des suites d'une maladie du foie et 384 (20%) sont décédés d'autres causes.



(a) Fonction de moyenne cumulée des réhospitalisations

(b) Incidence cumulée des événements terminaux

FIGURE 4.4 – Description des données

Les décès, liés ou non aux maladies du foie, surviennent en grande partie lors des six premiers mois suivant la greffe de foie. Au delà, le nombre de décès liés aux maladies du foie semble se stabiliser alors que le nombre augmente de façon linéaire pour les autres causes de décès (Figure 4.4). Il en est de même pour les réhospitalisations. En effet, le nombre de réhospitalisations est élevé l'année suivant la greffe et augmente de façon linéaire ensuite.

4.2 Article "Analysing recurrent events stopped by several types of terminal events : use of joint frailty model"

Analysing recurrent events stopped by multivariate terminal events: use of joint frailty model.

Anaïs Charles-Nelson, Msc[1,2,3,4] Catherine Schramm, phD[1,2,3] Sandrine Katsahian, MD, phD[2,3,4]

1. Sorbonne Universités, UPMC Univ Paris 06, UMRS 1138, Centre de Recherche des Cordeliers, F75006, Paris,France
2. INSERM, UMRS 1138, Centre de Recherche des Cordeliers, F75006, Paris, France
3. Université Paris Descartes, Sorbonne Paris Cité, UMRS 1138, Centre de Recherche des Cordeliers, F75006, Paris, France
4. Assistance Publique Hôpitaux de Paris, Hôpital européen Georges Pompidou, Unité d'Épidémiologie et de Recherche Clinique, INSERM, Centre d'Investigation Clinique 1418, module Épidémiologie Clinique, HEGP, F75015 Paris, France

Correspondance: Charles-Nelson Anais, Paris, France anais.charles.nelson@gmail.com

Keywords: Recurrent events, terminal events, conditional models, joint model, competing risks, penalized likelihood

Abstract The study of recurrent events becomes increasingly of interest for biomedical research as they provide longitudinal information about patients condition and treatment efficacy. To properly model a recurrent events process, censoring should be carefully defined as dependent or independent. Indeed, in many situations recurrent event process may be permanently stopped by a non-independent terminal event. Joint shared Gamma-distributed frailty models were developed to handle this problematic. However, no such a model exists when several terminal events are of interest in a competing risk situation. In this article, we have extended the joint shared Gamma-distributed frailty model to simultaneously analyze recurrent events in the presence of competing terminal events. Intensities of recurrent events and terminal events are modeled using semiparametric models. Dependence between recurrences and each terminal events are taking into account through two independent Gamma-distributed frailties. Cubic M-splines are used to approximate baseline hazard functions. Parameters are estimated by maximizing the penalized log-likelihood function through nonlinear and constrained generalized estimating equations. Penalization parameters are estimated by maximizing the sum of leave-one-out likelihood of each subject. Performances of our model have been assessed through a simulation study. We finally illustrated the model on a real dataset to assess simultaneously the effect of gender and age on rehospitalization after a liver transplant, and the effect of gender on both liver disease related death and other causes of death.

1 | INTRODUCTION

Recurrent events arising over time are longitudinally observed and provide meaningful information about patient's condition and treatment efficacy that may be directly measured on first or subsequent occurrences of these events. Methods to study recurrent events are numerous and of increasing interest in biomedical research studies, with a growing number of articles dealing with it since the 1990's[1,2]. Most popular methods focus on rate of events by modeling the instantaneous risk of event with an adaptation of the well-known Cox's model to the recurrence process[3,4]. The correlation between events is modeled through frailty parameters[5].

These methods assume independent censoring, an assumption often violated in medical studies. First, in a longitudinal study, a patient may be lost to follow-up due to the study condition or to the treatment removing the random nature of the censoring. Secondly, the recurrent event process may be permanently stopped by a terminal event (e.g., death) such that both processes are correlated. Indeed, the terminal event may be a consequence of a large number of repeated events. Inversely, a patient experiencing early the terminal event may have a lower number of recurrent events than a patient experiencing later the terminal event. In that case, considering the terminal event as random censoring may lead to biased results. The shared frailty model has been extended to handle the terminal event in recurrent events analyses through the joint shared frailty model[6,7,8].

The terminal event of interest could itself not be observed because of another event preventing its occurrence. This phenomenon is known as competing risk[9]. Zeng[10] extended the Liu's joint frailty semi-parametric model to handle multivariate recurrent and terminal events. Likelihood maximization for parameter estimation is based on expectation-maximization procedure, which may be time-consuming. The baseline intensity functions are approximated using Breslow's estimates, leading to a piecewise-constant baseline hazard function or unspecified baseline hazard function[11]. As an alternative, we propose to extend the Rondeau's joint frailty nonparametric model[11] to the case of competing risk. For example, it may be of interest to assess the effect of treatment on relapse and effect of treatment on both death by cancer and death by other causes linked to the disease. Here, parameters are estimated by maximizing the penalized likelihood, and baseline intensity function is approximated by a smooth function. To impose a continuous hazard function makes sense in clinical studies because baseline intensity often has a meaningful interpretation.

This article is organized as follows: first, we present our joint shared frailty model for recurrent events in presence of competing terminal events; in a second part, we assess performance of our model in a simulation study; then, we applied the model to a cohort of patients rehospitalized after liver transplant. A discussion finishes this article.

2 | NOTATIONS AND MODEL

Our model extends the joint frailty model developed by Rondeau[11] to the case of competing risks. Thus, we kept their notations and follow the structure of the article of Mazroui et al[12]. The proposed model is called the RESTE model (Recurrent Events stopped by Several Terminal Events).

2.1 | Notations

Let T_{ij} be the $j^{th} \in \left\{1, ..., n_i\right\}$ observation time of a patient $i \in \{1, ..., n\}$ defined as $T_{ij} = \min(X_{ij}, C_i, D_i)$ where X_{ij}, C_i and D_i are respectively the j^{th} recurrent event time, the right censoring time and the terminal event time for the patient i. We assume continuous time such that terminal, censoring and recurrent events can not occur at the same time. Right censoring is non-informative and independent of both terminal and recurrent processes. Terminal and recurrent events are dependent events conditional upon frailty.
Only the first event occurring among C_i and D_i is observed and noted $T_i = \min(C_i, D_i)$. In the presence of competing risk, only the first terminal event that occurs may be observed (if censure has not occurred yet) such that $D_i = \min_k(D_{ik})$ with D_{ik}, the terminal event time of type $k \in \{1, ..., K\}$ for patient i. In this article, we consider the case $K = 2$.
If $I_{\{.\}}$ is the indicator function, we denote $\delta_{ij} = I_{\{T_{ij}=X_{ij}\}}$ and $\delta_i^{D_k} = I_{\{D_{ik}\leq C_i\}}$ the recurrent event indicator and the terminal event indicator of type k respectively. A patient is "at-risk" for recurrent events if he has not experienced any type of the terminal event or is not censured yet. This "at-risk" process for the patient i is defined by $Y_i(t) = I_{\{t\leq T_i\}}$.

The number of recurrent events for patient i over the interval $[0, t]$ is denoted $N_i^{R^*}(t)$ and is not observed due to censoring. The observed number of events over the interval $[0, t]$ is defined as $N_i^R(t) = N^{R^*}\left(\min(T_i, t)\right)$. We denote $\Delta N_i^{R^*}(t) = N_i^{R^*}((t + \Delta t)^-) - N_i^{R^*}(t^-)$ where t^- denotes times infinitesimally smaller than t and $\Delta N_i^R(t) = Y_i(t)\Delta N_i^{R*}(t)$, the number of respectively actual and observed recurrent events over $[t, t + \Delta t[$. Recurrent event process may continue to increment after censoring but stops after terminal event occurrence.
Similarly, $N_i^{D_k}(t) = I_{\{D_{ik}\leq t, \delta_i^{D_k}=1\}}$ and $N_i^{D_k^*}(t) = I_{\{D_{ik}\leq t\}}$ correspond respectively to the observed and the actual indicator indicating that the terminal event of type k was observed before t.
Finally, we denote $Z_i^R(t)$, $Z_i^{D_1}(t)$ and $Z_i^{D_2}(t)$ the covariates at time t for patient i that may be time-dependent and affecting the recurrent and terminal event processes of type 1 and 2 respectively. Thus the observations for patient i at time t is $O_i(t) = \left\{ Y_i(u), N_i^R(u), N_i^{D_1}(u), N_i^{D_2}(u), Z_i^R(u), Z_i^{D_1}(u), Z_i^{D_2}(u) \right\}_{0\leq u\leq t}$. Of note, $O_i(t)$ defines the history of the process for the patient i up to t.

We assume that recurrent events and terminal events of a same patient are correlated and that they depend also on unobserved covariates. To take into account all these aspects, we define two independent gamma-distributed unobserved frailty parameters u_{i1} and u_{i2} such that u_{ik} measures the dependence between recurrences and terminal event k. Of note, observations and unobserved frailty parameters define the filtration $F_{it^-} = \sigma\{O_i(t), u_{i1}, u_{i2}\}_{i=1..n}$.

2.2 | Modeling

We model the intensity functions of subject i for recurrent $r_i(t)$ and terminal events $\alpha_i^{(k)}(t)$ $k = 1, 2$ as:

$$\begin{aligned} r_i(t) &= lim_{\Delta t\to 0}\frac{P(\Delta N_i^{R^*}(t)=1|Z^R(t),u_{i1},u_{i2},D_i\geq t)}{dt} \\ \alpha_i^{(k)}(t) &= lim_{\Delta t\to 0}\frac{P(\Delta N_i^{D_k^*}(t)=1|Z_i^{D_k}(t),u_{ik},D_i\geq t)}{\Delta t} \end{aligned} \quad ([1])$$

Of note, intensities are not observed and only $Y_i(t)r_i(t)$ and $Y_i(t)\alpha_i^{(k)}(t)$ are observed and may be expressed as:

$$\begin{aligned} Y_i(t)r_i(t) &= lim_{\Delta t\to 0}\frac{P(\Delta N_i^R(t)=1|F_{it^-})}{\Delta t} \\ Y_i(t)\alpha_i^{(k)}(t) &= lim_{\Delta t\to 0}\frac{P(\Delta N_i^{D_k}(t)=1|F_{it^-})}{\Delta t} \end{aligned} \quad ([2])$$

We define the following model where recurrent process and both terminal events are modeled through proportional hazard functions:

$$\begin{cases} r_i(t \mid u_{i1}, u_{i2}) = u_{i1}u_{i2}r_0(t)\exp\left(\beta_R Z_i^R(t)\right) & \text{(recurrences)} \\ \alpha_i^{(1)}(t \mid u_{i1}) = u_{i1}\alpha_0^{(1)}(t)\exp\left(\beta_{D_1} Z_i^{D_1}(t)\right) & \text{(terminal event 1)} \\ \alpha_i^{(2)}(t \mid u_{i2}) = u_{i2}\alpha_0^{(2)}(t)\exp\left(\beta_{D_2} Z_i^{D_2}(t)\right) & \text{(terminal event 2)} \end{cases} \quad ([3])$$

where u_{i1} and u_{i2} are frailty parameters taking into account the dependence between recurrence and terminal event of type 1 and 2 respectively. They are independent and identically Gamma-distributed with $E(u_k) = 1$ and $Var(u_k) = \theta_k$ for $k = 1, 2$. A variance of frailties $\theta_k = 0$ means that the dependence is completely due to the observed covariates. This model assumes that the effect of the frailty u_k is identical for the recurrent events and the terminal event k. $r_0(.)$ is the baseline intensity function for recurrence process, and $\alpha_0^{(1)}(.)$ and $\alpha_0^{(2)}(.)$ are baseline hazard functions for terminal event of type 1 and 2 respectively. β_R, β_{D_1} and β_{D_2} are time-invariant regression coefficients.

Let's denote $\omega = \left\{ r_0(.), \alpha_0^{(1)}(.), \alpha_0^{(2)}(.), \theta_1, \theta_2, \beta_R, \beta_{D_1}, \beta_{D_2} \right\}$ all the parameters that should be estimated.

3 | THE LOG-LIKELIHOOD AND THE ESTIMATION PROCEDURE

3.1 | Penalized log-likelihood

Parameters are estimated by maximizing the marginal penalized log-likelihood. Penalization lets us approximate the three baseline hazard functions $r_0(.)$, $\alpha_0^{(1)}(.)$ and $\alpha_0^{(2)}(.)$ with smooth continuous functions. Here we give some key steps in the definition of the penalized log-likelihood. More details are available in Appendix A.

The conditional likelihood for the patient i for recurrent events is as follows:

$$\begin{aligned} L_i^R\left(\omega \mid O_i, u_{i1}, u_{i2}\right) = & \prod_{j=1}^{n_i} \left(r_0(T_{ij})u_{i1}u_{i2}\exp\left(\beta_R Z_i^R(T_{ij})\right)\right)^{\delta_{ij}} \\ & \times \exp\left\{ -u_{i1}u_{i2} \sum_{j=1}^{n_i+1} \int_{T_{i(j-1)}}^{T_{ij}} Y_i(t) r_0(t)\exp\left(\beta_i Z_i^R(t)\right) dt \right\} \end{aligned} \quad ([4])$$

For all patients, we set $T_{i0} = 0$ as in medical research, we are interested in the time to event since the diagnosis or from randomization. The conditional likelihood functions for each type k of terminal event for the patient i is as follows:

$$\begin{aligned} L_i^{D_k}\left(\omega \mid O_i, u_{ik}\right) = & \left\{ \alpha_0^{(k)}(T_i)u_{ik}\exp\left(\beta_{D_k} Z_i^{D_k}(t)\right) \right\}^{\delta_i^{D_k}} \\ & \times \exp\left\{ -\int_0^{T_i} Y_i(t)\alpha_0^{(k)}u_{ik}\exp\left(\beta_{D_k} Z_i^{D_k}(t)\right) dt \right\} \end{aligned} \quad ([5])$$

The conditional likelihood function for the patient i is the product of the likelihood functions corresponding to the three intensity functions we model:

$$L_i\left(\omega \mid O_i, u_{i1}, u_{i2}\right) = L_i^R\left(\omega \mid O_i, u_{i1}, u_{i2}\right) \times L_i^{D_1}\left(\omega \mid O_i, u_{i1}\right) \times L_i^{D_2}\left(\omega \mid O_i, u_{i2}\right) \quad ([6])$$

The marginal likelihood $L_i(\omega \mid O_i)$ is obtained by integrating $L_i(\omega \mid O_i, u_{i1}, u_{i2})$ according to u_{i1} and u_{i2}.

$$L_i(\omega \mid O_i) = \int_0^\infty \int_0^\infty L_i(\omega | O_i, u_{i1}, u_{i2}) f(u_{i1}) f(u_{i2}) du_{i1} du_{i2} \quad ([7])$$

where $f(.)$ is the density function of Gamma-distribution. This double integral is complex and does not have an analytical form. Only

integration according to one of the frailty parameter has an analytical form. We choose to first integrate according to u_{i2} (see more detail of the calculation in the appendix A). Then, the contribution of patient i to the marginal log-likelihood based on observed data is:

$$\begin{aligned} ll_i(\omega \mid O_i) &= log(L_i(\omega \mid O_i)) \\ &= \delta_i^{D_1}\left\{log\left(\alpha_i^{(1)}(T_i)\right)\right\} + \delta_i^{D_2}\left\{log\left(\alpha_i^{(2)}(T_i)\right)\right\} \\ &\quad + \sum_{j=1}^{n_i}\delta_{ij}\left\{log\left(r_i(T_{ij})\right)\right\} - \tfrac{1}{\theta_1}log(\theta_1) \\ &\quad -log\left(\Gamma(\tfrac{1}{\theta_1})\right) - \tfrac{1}{\theta_2}log(\theta_2) - log\left(\Gamma(\tfrac{1}{\theta_2})\right) + log\left(\Gamma(n_i + \tfrac{1}{\theta 2} + \delta_i^{D_2})\right) \\ &\quad +log\left(\int_0^{\infty} g(u_{i1})\right) du_{i1} \end{aligned} \tag{[8]}$$

$$g(u_{i1}) = \frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\theta_1}-1} exp\left\{-u_{i1}\left(\int_0^{T_i} Y_i(t)\alpha_0^{(1)}(t)exp\left(\beta_{D_1}Z_i^{D_1}(t)\right)dt + \frac{1}{\theta_1}\right)\right\}}{\left\{u_{i1}\sum_{j=1}^{n_i+1}\int_{T_{i(j-1)}}^{T_{ij}} Y_i(t)r_0(t)exp\left(\beta_R Z_i^R(t)\right)dt + \frac{1}{\theta_2} + \int_0^{T_i} Y_i(t)\alpha_0^{(2)}(t)exp\left(\beta_{D_2}Z_i^{D_2}(t)\right)dt\right\}^{\frac{1}{\theta_2}+\delta_i^{D_2}+n_i}}$$

To overcome the problem of overfitting leading to local fluctuation due to the estimation of the baseline functions, a penalized log-likelihood is used. As in Mazroui [13], we penalized the curvature of each baseline hazard functions [14] with ξ_R, ξ_{D_1} and ξ_{D_2} the tuning parameters corresponding to the recurrent events and the terminal events 1 and 2. Thus, the penalized log-likelihood $pll(\omega|O)$ is:

$$pll(\omega|O) = \sum_{i=1}^{n} ll_i\left(\omega|O_i\right) - \xi_R\int_0^{\infty}[r_0''(t)]^2dt - \xi_{D_1}\int_0^{\infty}[\alpha_0^{''(1)}(t)]^2dt - \xi_{D_2}\int_0^{\infty}[\alpha_0^{''(2)}(t)]^2dt \tag{[9]}$$

where $O = \{O_1, ...O_n\}$.

3.2 | Approximation of baseline functions

Baseline hazard functions $r_0(.)$, $\alpha_0^{(1)}(.)$ and $\alpha_0^{(2)}(.)$ are approximated by non-negative M-splines of order $d = 4$ (cubic splines) allowing enough flexibility to capture complex hazard functions. M-Splines are smooth piecewise polynomial functions defined on knot sequences $x = \left\{x_1, ..., x_{p+2d}\right\}$, where p is the number of internal knots and bounded between 0 and $\max_{i=1..n}(T_i)$ such as $x_1 = ... = x_d = 0$ and $x_{p+d+1} = ... = x_{p+2d} = \max_{i=1..n}(T_i)$.

It results in $p + d$ M-spline functions of order d that are positive on the interval $[x_m, x_{m+d}]$ and equal zero elsewhere. They are defined recursively [15] as follows:

- for $c = 1$ and $\forall m \in \{1, ..., p + 2d - 1\}$
$$M_m(t|c, x) = \begin{cases} \frac{1}{(x_{m+1}-x_m)} & \text{if } x_m \le t \le x_{m+1} \\ 0 & \text{otherwise} \end{cases}$$

- $\forall c \in \{2, .., d\}$ and $\forall m \in \{1, ..., p + 2d - c\}$
$$M_m(t|c, x) = \begin{cases} \frac{c\left\{(t-x_m)M_m(t|c-1,x)+(x_{m+c}-t)M_{m+1}(t|c-1,x)\right\}}{(c-1)(x_{m+c}-x_m)} & \text{if } x_m \le t \le x_{m+c} \\ 0 & \text{otherwise} \end{cases}$$

The three baseline hazard functions are thus estimated using a linear combination of the same basis of M-splines of order 4. Only the coefficients are different.

$$\begin{cases} \tilde{r}_0(t) = \gamma_1^R M_1(t|4, x) + \gamma_2^R M_2(t|4, x) + ... + \gamma_{p+d}^R M_{p+d}(t|4, x) & \text{(recurrences)} \\ \tilde{\alpha}_0^{(1)}(t) = \gamma_1^{(1)} M_1(t|4, x) + \gamma_2^{(1)} M_2(t|4, x) + ... + \gamma_{(p+d)}^{(1)} M_{(p+d)}(t|4, x) & \text{(terminal event 1)} \\ \tilde{\alpha}_0^{(2)}(t) = \gamma_2^{(2)} M_1(t|4, x) + \gamma_2^{(2)} M_2(t|4, x) + ... + \gamma_{(p+d)}^{(2)} M_{p+d}(t|4, x) & \text{(terminal event 2)} \end{cases}$$

where $\gamma = \{\gamma_1^R, ..., \gamma_{p+d}^R, \gamma_1^{(1)}, ..., \gamma_{(p+d)}^{(1)}, \gamma_2^{(2)}, ..., \gamma_{(p+d)}^{(2)}\}$ are the coefficients associated to the splines.

The number of estimated parameters for each of the three cubic M-splines are the number of internal knots plus the order of the spline ($p + d$). The higher the number of knots, better the approximation. However, if the number of knots is high, local fluctuations could appear, but the penalization handles this problem. The first and second derivatives of the M-splines are given in the Appendix B. Variance of the baseline hazard functions are estimated as follows[13].

$$Var\left(\tilde{r_0}(.)\right) = \mathbf{M}(.)^{\mathbf{T}}\mathbf{I}_{\boldsymbol{\gamma}}^{-\mathbf{1}}\mathbf{M}(.),\ Var\left(\tilde{\alpha_0}^{(1)}(.)\right) = \mathbf{M}(.)^{\mathbf{T}}\mathbf{I}_{\boldsymbol{\gamma}_{1.}}^{-\mathbf{1}}\mathbf{M}(.),\ Var\left(\tilde{\alpha_0}^{(2)}(.)\right) = \mathbf{M}(.)^{\mathbf{T}}\mathbf{I}_{\boldsymbol{\gamma}_{2.}}^{-\mathbf{1}}\mathbf{M}(.)$$

where $M(.) = (M_1(.), ..., M_{p+d}(.))$ is the M-splines vector and $I_\gamma = \frac{\partial^2 pll(\omega|O_i)}{\partial \gamma^2}$, $I_{\gamma_{1.}} = \frac{\partial^2 pll(\omega|O_i)}{\partial \gamma_{1.}^2}$, $I_{\gamma_{2.}} = \frac{\partial^2 pll(\omega|O_i)}{\partial \gamma_{2.}^2}$ are subsets of the Hessian matrix corresponding to the variance-covariance matrices associated with vectors of parameters γ, γ_1 and γ_2 respectively.
Thus 95% pointwise confidence intervals are

$$\tilde{r_0}(.) \pm Z_{1-\frac{\alpha}{2}} * \sqrt{Var\left(\tilde{r_0}(.)\right)},\ \tilde{\alpha_0}^{(1)}(.) \pm Z_{1-\frac{\alpha}{2}} * \sqrt{Var\left(\tilde{\alpha_0}^{(1)}(.)\right)},\ \tilde{\alpha_0}^{(2)}(.) \pm Z_{1-\frac{\alpha}{2}} * \sqrt{Var\left(\tilde{\alpha_0}^{(2)}(.)\right)}$$

3.3 | Estimation procedure

Maximization of the penalized log-likelihood requires estimation of integrals that cannot be directly calculated. We approximated them using the Gauss-Laguerre quadrature for infinite integrals[16] and Gauss-Legendre quadrature for finite integrals[16].
The estimation procedure follows two steps. First, tuning parameters (ξ_R, ξ_{D_1}, ξ_{D_2}) are obtained by fitting three independent models, a shared frailty model for recurrent events and one model for each terminal event. For the three models, we used a penalized likelihood where the tuning parameter is a computed using cross-validation method. The cross-validation method maximizes the sum of the leave-one-out likelihood of each subject[17 18].
Second, the obtained tuning parameter values are incorporated in the penalized likelihood. the parameters (β,θ_1,θ_2) as well as the γ parameters for the three baseline hazards are estimated by minimizing the negative of the penalized likelihood ([9]) using nonlinear and constrained generalized estimating equations[19 20]. The estimation procedure uses a bounds constrained quasi-Newton method[21]. To obtain positive hazard functions, splines coefficients are imposed to be positive in the estimation procedure using the square transformation. Moreover, to overcome the positivity constraint of the variance of frailty parameters θ_1 and θ_2, we used the exponential transformation. Standard errors of the regression parameters are estimated as the square root of the diagonal of the inverse of the Hessian matrix and using the delta method[22] for the θ_1 and θ_2 parameters.

3.4 | Implementation

The model was implemented using R software. First, tuning parameters (ξ_R, ξ_{D_1}, ξ_{D_2}) are obtained by fitting three models based on penalized likelihood, one shared frailty model for recurrent events and one for each terminal event using the `frailtyPenal` function with the `cross.validation=TRUE` option available in the `frailtypack` package[23]. The splines and its second derivatives were calculated using the `mSpline` function from the `splines2` package[24]. The implementation of the optimization procedure was done using the `nlminb` function available in the `stats` package[25]. Initial parameter values for the β, γ, and for the frailty parameters θ_1 and θ_2 are all set up at the value 0.1. To obtain positive hazard functions, spline coefficients were imposed to be positive in the estimation procedure using the `lower` option of the `nlminb` function. Hessian matrix were obtained using the `hessian` function included in the `numDeriv` package[26].

3.5 | Goodness of fit

3.5.1 | Martingale residuals

The RESTE model is a composite of three parts and the responses in these three parts may not be comparable thus the goodness of fit for the RESTE model is checked for the three parts separately. The goodness of fit is assessed using the martingale residuals. Indeed,

they can be used to checked whether the model predicts correctly the number of observed events as well as the fonctional form of covariates[27].

$$\begin{cases} M^R(t) = N^R(t) - Y_i(t)u_{i1}u_{i2}\int_0^t \hat{r}_i(s)ds & \text{(recurrences)} \\ M^{D_1}(t) = N^{D_1}(t) - Y_i(t)u_{i1}\int_0^t \hat{\alpha}_i^{(1)}(s)ds & \text{(terminal event 1)} \\ M^{D_2}(t) = N^{D_2}(t) - Y_i(t)u_{i2}\int_0^t \hat{\alpha}_i^{(2)}(s)ds & \text{(terminal event 2)} \end{cases} \quad ([10])$$

The parameters $(\beta, \theta_1, \theta_2, \gamma)$ used to estimates the martingales residuals are those estimated by the maximization of the penalized likelihood. The individual estimates of the frailty effects $\hat{u}_{i1}$ and $\hat{u}_{i2}$ are required to calculate the martingale residuals. Thus,https://fr.overleaf.com/project/5d0cb73587fdf51152e35c0athe empirical Bayesian estimator has been used and is equal to the value maximizing the posterior probability density.

$$f(u_{i1}, u_{i2}|O_i, \hat{\omega}) \propto f(O_i|u_{i1}, \hat{\omega})f(u_{i1}, u_{i2}|\hat{\omega})$$

where $f(u_{i1}, u_{i2}|O_i, \hat{\omega})$ is the a-posteriori density functiosn of the frailtty parameters, $f(O_i|u_{i1}, \hat{\omega})$ is the likelihood for the subject i given $\hat{\omega}$, u_{i1} and u_{i2} and $f(u_{i1}, u_{i2}|\hat{\omega})$ is the density function of the frailty parameters. Thus,

$$f(u_{i1}, u_{i2}|O_i, \hat{\omega}) \propto \frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\hat{\theta}_1}-1} exp\left\{-u_{i1}\int_0^{T_i} Y_i(t)\hat{\alpha}^{(1)}(t)dt\right\} exp\left\{-\frac{u_{i1}}{\hat{\theta}_1}\right\}}{\hat{\theta}_1^{\frac{1}{\hat{\theta}_1}}\Gamma(\frac{1}{\hat{\theta}_1})} \times \frac{u_{i2}^{n_i+\delta_i^{D_2}+\frac{1}{\hat{\theta}_2}-1} exp\left\{-u_{i2}(u_{i1}\sum_{j=1}^{n_i+1}\int_{T_{i(j-1)}}^{T_{ij}} Y_i(t)\hat{r}(t)dt + \frac{1}{\hat{\theta}_2}+\int_0^{T_i} Y_i(t)\hat{\alpha}^{(2)}(t)dt)\right\}}{\hat{\theta}_2^{\frac{1}{\hat{\theta}_2}}\Gamma(\frac{1}{\hat{\theta}_2})}$$

The Marquardt algorithm was used to estimate the mode of the posterior density function.

3.6 | Model selection

Because of the penalized likelihood, basic selection criterion like the Akaike information criterion or the Bayesian information criterion are not appropriate. Liquet et al[28] have discussed that the likelihood cross-validation (LCV) could be used for choosing between two semi-parametric models. Commenges et al have proposed the approximated likelihood cross validation criterion (LCVa) that is valid in any penalized likelihood[29].

$$LCVa = \frac{Tr(\hat{H_{pl}^{-1}}I) - l(\hat{\omega})}{\sum_{i=1}^n n_i + 1}$$

where $\hat{H_{pl}^{-1}}$ is the Hessian matrix of the penalized likelihood, I is the Fisher information matrix and $n_i + 1$ is the number of observations of the subject i.

4 | SIMULATION STUDIES

4.1 | Generating data

We generated 500 replicated databases from the following model:
$\forall i \in [1, ...n]$

$$, \begin{cases} r_i(t|u_{i1}, u_{i2}) = u_{i1}u_{i2}r_0(t)exp\left(\beta_1 Z_{1i} + \beta_2 Z_{2i}\right) & (1) \\ \alpha_i^{(1)}(t|u_{i1}) = u_{i1}\alpha_0^{(1)}(t)exp\left(\beta_3 Z_{1i}\right) & (2) \\ \alpha_i^{(2)}(t|u_{i2}) = u_{i2}\alpha_0^{(2)}(t)exp\left(\beta_4 Z_{1i}\right) & (3) \end{cases}$$

where Z_{1i} and Z_{2i} are binary covariates, both generated from a Bernoulli distribution with a probability of 0.5. We set $\beta_1 = 1$, $\beta_2 = -0.5$, $\beta_3 = 0.7$ and $\beta_4 = -1$. The frailty parameters were generated from a Gamma distribution with mean 1. The variance θ_1 was set to 0.5 whereas θ_2 varies across scenarios. The baseline hazard functions were simulated as constant for the terminal events and were set at $\forall t$

$\alpha_0^{(1)}(t) = 1.5$, $\alpha_0^{(2)}(t) = 2$ for the terminal event of cause 1, of cause 2, and for recurrent events respectively while baseline hazards for the recurrent events vary accross scenario. For each patient, data were simulated in two steps:

- The first step was the generation of the time to terminal events. Time for terminal event was generated from an exponential distribution under models (2) and (3)[30]. Censoring time was generated from a uniform distribution over [0;1] and we assumed a proportion of censored data about one third. The last observed time was defined as the minimum of censoring and terminal event times. Index function for the last observed time was equal to 0, 1 or 2 if the minimum time equals the censoring time, the time for the terminal event of cause 1 or the time for the terminal event of cause 2, respectively.

- The second step was the generation of recurrent event times. Gap times were generated from an exponential distribution under the model (1). We took into account only recurrent events with total time less than or equal to the last observed time defined in the first step.

In the first setting, the variance of θ_2 was set to 0.5. This suggest a significant dependence between the recurrent events process and the terminal event of type 2. In this setting, expected values of variance of the frailty in both the Rondeau's model ($\tilde{\theta}$) and the Zeng's model (ξ) were unknown as well as the value of flexibility terms in these two models denoted α for the Rondeau's model and ϕ_1 and ϕ_2 for the Zeng's model. The baseline hazards for recurrent events were first simulated :

- as a constant baseline $r_o(t) = 8$
- as a linear function $r_o(t) = 32t$
- using a Weibull function with parameters $\lambda = 8$ and $\nu = 0.75$.

In the second setting the value for the parameter θ_2 was set to 0.01. This scenario mimics the case of independence between the recurrent event process and the terminal event process of type 2. Thus in this setting, the variance of frailty parameter for the Rondeau's model ($\tilde{\theta}$) was expected to be $\tilde{\theta} \sim \theta_1 = 0.5$ and the expected value of the flexibility parameter was expected to be $\alpha \sim 1$. However, the expected value of the frailty and of the flexibility parameters in the Zeng's model remained unknown. In this setting, the baseline function is considered as constant $r_o(t) = 8$.

For all scenario, the sample size n was set to 500.

4.2 | Simulation results

For each scenario, we have fitted the developed model (RESTE model) and compared it to the model developed by Rondeau[11] and the one developed by Zeng[10]. In the Rondeau's model, the terminal event of type 2 was censored. For the Rondeau and the RESTE model, the boundary knots for splines were setup as 0 and the last event time, inner knots were located at the tercile of the sample distribution of event times. Results of the simulation study are presented in Table 1 and in Table 2.
In the first setting ($\theta_2 = 0.5$), the mean number of events per subjects is ranged from 1.32 to 2.02. The terminal event rate of cause 1 is 41.5% and the terminal event of cause 2 is 26.5%. The model including two gamma frailties performs better than the Rondeau's model and the Zeng's model in terms of estimation of parameters. Indeed, the regression parameters are well estimated for the RESTE model with the two gamma distributed frailties. In all scenario, all parameters are well estimated. The Zeng's model also provide unbiased estimates for β_1 and β_2, but β_3 is underestimated and the bias remains the same no matter the shape of the baseline hazard. The Rondeau's model seems to provide slighly biased estimation for the parameter β_1 and as the Zeng's model underestimate the parameter β_3. In the second setting ($\theta_2 = 0.01$), the mean number of events per subjects is 2.11. The terminal event rate of cause 1 is 54.5 % and the terminal event of cause 2 is 35.5%. All regression coefficients are well estimated with the three models. The Rondeau's model provides unbiased estimation for the frailty parameters and, as expected, the value of α parameter is close to 1. In this scenario, the Zeng's model provides good estimations for all parameters but β_3 is overestimated. However, we cannot assess performances of this model for the estimation of frailty parameters.

TABLE 1 Simulation studies results for the second setting ($\theta_2 = 0.5$)

Baseline	Parameters	The RESTE model Estimates (se**)	Rondeau Joint frailty model Estimates (se)	Zeng model Estimates (se)
$r_0(t) = 8$	$\beta_1 = 1$	0.99 (0.15)	1.06 (0.13)	1.00 (0.13)
	$\beta_2 = -0.5$	-0.50 (0.122)	-0.50 (0.15)	-0.50 (0.13)
	$\beta_3 = 0.7$	0.69 (0.25)	0.64 (0.16)	0.62 (0.17)
	$\beta_4 = -1$	-0.98 (0.26)		-0.92 (0.21)
	$\theta_1 = 0.5$	0.53 (0.13)		
	$\theta_2 = 0.5$	0.51 (0.13)		
	$\tilde{\theta} \sim 0.5$		0.90 (0.06)	
	$\alpha = 1$		0.48 (0.12)	
	$\xi = ?$			0.65 (0.10)
	$\phi_1 = ?$			0.84 (0.37)
	$\phi_2 = ?$			0.83 (0.08)
$r_0(t) = 32t$	$\beta_1 = 1$	1.00 (0.20)	1.10 (0.18)	0.99 (0.18)
	$\beta_2 = -0.5$	-0.49 (0.15)	-0.49 (0.20)	-0.5 (0.16)
	$\beta_3 = 0.7$	0.69 (0.20)	0.64 (0.16)	0.62 (0.18)
	$\beta_4 = -1$	-1.03 (0.24)		-0.93 (0.22)
	$\theta_1 = 0.5$	0.46 (0.15)		
	$\theta_2 = 0.5$	0.50 (0.16)		
	$\tilde{\theta} \sim ?$		0.86 (0.07)	
	$\alpha = ?$		0.48 (0.16)	
	$\xi = ?$			0.73 (0.18)
	$\phi_1 = ?$			0.79 (0.26)
	$\phi_2 = ?$			0.81 (0.32)
$r_0(t) = 8\gamma t^{\gamma-1}$, $\gamma = 0.75$	$\beta_1 = 1$	0.97 (0.16)	1.06 (0.15)	1.00 (0.14)
	$\beta_2 = -0.5$	-0.52 (0.13)	-0.49 (0.17)	-0.51 (0.14)
	$\beta_3 = 0.7$	0.69 (0.24)	0.64 (0.16)	0.62 (0.18)
	$\beta_4 = -1$	-1.00 (0.25)		-0.92 (0.22)
	$\theta_1 = 0.5$	0.53 (0.14)		
	$\theta_2 = 0.5$	0.52 (0.13)		
	$\tilde{\theta} \sim ?$		0.89 (0.06)	
	$\alpha = ?$		0.49 (0.13)	
	$\xi = ?$			0.64 (0.10)
	$\phi_1 = ?$			0.85 (0.28)
	$\phi_2 = ?$			0.85 (0.31)
Lognormale $r_0(t) = 2$	$\beta_1 = 1$	1.02 (0.13)	1.04 (0.12)	1.01 (0.11)
	$\beta_2 = -0.5$	-0.48 (0.11)	-0.48 (0.15)	-0.50 (0.11)
	$\beta_3 = 0.7$	0.72 (0.15)	0.66 (0.13)	0.65 (0.14)
	$\beta_4 = -1$	-0.96 (0.19)		-0.94 (0.19)
	$\theta_1 = 0.5$	0.26 (0.06)		
	$\theta_2 = 0.5$	0.26 (0.06)		
	$\tilde{\theta} \sim ?$		0.48 (0.10)	
	$\alpha = ?$		0.46 (0.20)	
	$\xi = ?$			0.29 (0.05)
	$\phi_1 = ?$			0.88 (0.31)
	$\phi_2 = ?$			0.91 (0.31)

$\tilde{\theta}$ is the variance of the frailty parameter in the Rondeau's model capturing the dependence between recurrences and the terminal event of type 1. α is a flexibility parameter in the Rondeau's model determining the direction of the association between recurrences and the terminal event hazards. ξ is the variance of the random effect shared by all the events (recurrences and terminal events) that captures the dependence between recurrences and terminal events. ϕ_1 and ϕ_2 are flexibility parameters in the Zeng's model giving the direction of the association between recurrences and the terminal events hazards.

TABLE 2 Simulation studies results for the second setting ($\theta_2 = 0.01$)

Baseline	Parameters	The RESTE model Estimates (se**)	Rondeau Joint frailty model Estimates (se)	Zeng model Estimates (se)
$r_0(t) = 8\ \theta_2$	$\beta_1 = 1$	0.98 (0.12)	1.00 (0.11)	1.00 (0.11)
	$\beta_2 = -0.5$	-0.50 (0.12)	-0.50 (0.11)	-0.50 (0.10)
	$\beta_3 = 0.7$	0.70 (0.26)	0.70 (0.17)	0.80 (0.21)
	$\beta_2 = -1$	-0.98 (0.22)		-1.01 (0.18)
	$\theta_1 = 0.5$	0.47 (0.13)		
	$\theta_2 = 0.01$	0.09 (0.13)		
	$\tilde{\theta} \sim 0.5$		0.54 (0.09)	
	$\alpha = 1$		0.98 (0.21)	
	$\xi =?$			0.36 (0.06)
	$\phi_1 =?$			1.72 (0.31)
	$\phi_2 =?$			0.01 (0.36)

$\tilde{\theta}$ is the variance of the frailty parameter in the Rondeau's model capturing the dependence between recurrences and the terminal event of type 1.
α is a flexibility parameter in the Rondeau's model determining the direction of the association between recurrences and the terminal event hazards.
ξ is the variance of the random effect shared by all the events (recurrences and terminal events) that captures the dependence between recurrences and terminal events.
ϕ_1 and ϕ_2 are flexibility parameters in the Zeng's model giving the direction of the association between recurrences and the terminal events hazards.

5 | APPLICATION

Rehospitalization as a measure of the cost-effectiveness and quality of health has been largely discussed in the literature[31 32]. However, they are often wrongly analyzed by focusing only on the first rehospitalization time[33 34]. The Figure 1 presents the schematic of the application. The RESTE model has been applied to assess the effect of gender and age on rehospitalization after a liver transplant, and on liver disease related death. Results were compared to ones obtained with Zeng's model. The Rondeau model has been also applied twice on these data. In the first model, death from liver disease has been considered as the terminal event whereas death from other causes have been considered as censure. In the second model, death from other causes have been considered as the terminal event and death from liver disease has been considered as censure.

Data are extracted from the PMSI (Information Systems Medicalization Program) which summarize all hospital stay in France. It records the main diagnosis corresponding to the reason why the patient is admitted to the hospital and which procedures/cares the patient has received during the hospital stay, as well as some other information. Patients with liver transplant in 2008 or 2009 were identified using the codes HLEA001 and HLEA002. Rehospitalization from 2008 to 2016 of these patients as well as gender, age at transplant, and information about death during rehospitalization were extracted. As only death at the hospital is recorded in the PMSI, the death rate is underestimated. Rehospitalization with a duration of zero day, rehospitalization before the transplant, and rehospitalization for dialysis or treatment for cancer (chemotherapy or irradiation) are removed from the database and alive patients are censored at the date of 12/12/2016.
A total of 1933 patients had a liver transplant in 2008 or 2009. There are 1361 (70.4%) males and the mean age at transplant is 49 $\pm$ 16 years old. A total of 12,793 recurrences have been recorded with an average of 6.6 $\pm$7.1 rehospitalizations per subject ranging from 0 to 109. 135 patients (7%) died from liver disease and 384 (20%) died from other causes. The covariates included in the model are gender (1 = male, 0 = female) and age (1= patients older than 53 years old, 0 = patients younger than 53 years old). Results are summarized in Table 5.

FIGURE 1 Schematic of the application

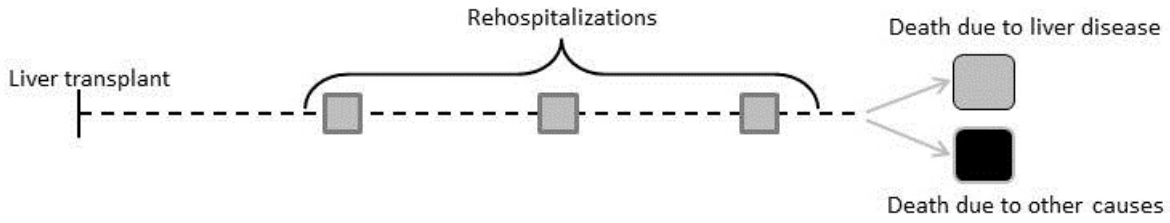


TABLE 3 Analysis of rehospitalizations and death after liver transplant

Covariate	The RESTE model		Rondeau's model				Zeng's model	
			TE=Liver disease related death		TE=Other causes of death			
	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value
Rehospitalizations								
Gender (men vs women)	1.12 [1.07;1.18]	<0.0001	1.09 [1.00;1.18]	0.052	1.10 [1.02;1.20]	0.02	1.09 [1.01;1.17]	0.02
Age (>53 vs ≤53)	1.04 [0.98;1.11]	0.17	1.03 [0.96;1.11]	0.41	0.99 [0.92;1.07]	0.92	1.02 [0.95;1.10]	0.53
Liver disease related death								
Gender (men vs women)	1.45 [1.20;1.76]	<0.0001	1.78 [1.15;2.76]	0.01	-	-	1.76 [1.16;2.67]	0.01
Other causes of death								
Gender (men vs women)	1.54 [1.37;1.74]	<0.0001	-	-	1.70 [1.31;2.22]	<0.0001	1.63 [1.27;2.10]	0.0001
θ_1	0.16	<0.0001	-	-	-	-	-	-
θ_2	0.21	<0.0001	-	-	-	-	-	-
$\tilde{\theta}$	-	-	0.51	<0.0001	0.52	<0.0001	-	-
α	-	-	0.84	<0.0001	1.17	<0.0001	-	-
ξ	-	-	-	-	-	-	0.28	0.11
ϕ_1	-	-	-	-	-	-	0.05	0.88
ϕ_2	-	-	-	-	-	-	1.18	0.10

TE is terminal event. The reference category for gender is "women" and for age is "≤ 53". $\tilde{\theta}$ is the variance of the frailty parameter in the Rondeau's model capturing the dependence between recurrences and the terminal event. α is a flexibility parameter in the Rondeau's model determining the direction of the association between recurrences and the terminal event hazards. If $\alpha < 0$, hazards of recurrences and the terminal event are negatively correlated. If $\alpha = 0$, the terminal event hazard is non informative for the recurrent event hazard. Finally if $\alpha > 0$, hazards of recurrences and of the terminal event are positively correlated. ξ is the variance of the random effect shared by all the events (recurrences and terminal events) capturing the dependence between recurrences and terminal events in the Zeng's model. ϕ_1 and ϕ_2 are flexibility parameters in the Zeng's model giving the direction of the association between recurrences and the terminal events hazards.

Table 5 shows that according to the RESTE model, to be a man significantly increases the intensity of rehospitalization (HR=1.12, 95%CI= [1.07; 1.18]) and intensity of death due to liver disease (HR=1.45, 95%CI= [1.20; 1.76]), as well as death due to other causes (HR=1.54, 95%CI= [1.37; 1.74]). In contrast, age does not have any significant effect on rehospitalization (HR=1.04, 95%CI= [0.98; 1.11]). Dependencies between rehospitalization and the two terminal events are significant ($\theta_1 = 0.16$ and $\theta_2 = 0.21$) suggesting that unobservable factors increase intensities of rehospitalization and intensities of the two types of terminal events.
According to the Rondeau's model, gender also has a significant effect on rehospitalization (HR=1.10, 95%CI= [1.02; 1.20]) when analyzing death due to other causes, but not when analyzing death due to liver disease (HR=1.09, 95%CI= [1.00; 1.18]). We note that the hazard ratio for the two terminal events, HR=1.78, 95%CI= [1.15; 2.76] and HR=1.70, 95%CI= [1.31; 2.22] for death due to liver disease and death due to other causes with the Rondeau's model respectively, are higher than the hazard ratios with the RESTE model. The Rondeau's model detects a significant dependence between rehospitalization and terminal events ($\theta_1 = 0.51$ and $\theta_2 = 0.52$ for death due to cancer and death due to other causes, respectively), but as shown in the simulation study the values of the variance of the frailty parameter are higher than values provided by the RESTE model. The significant association between rehospitalization and both terminal events may explain why the value of HR for terminal events is higher in the Rondeau's model than in the RESTE model. The Rondeau's model indicates that the rehospitalization and the two terminal events are positively correlated ($\alpha > 0$).
The Zeng's model provides a similar effect of covariates on rehospitalization (HR and 95%CI for gender= 1.09 [1.01; 1.17], and HR and 95%CI for age= 1.02 [0.95; 1.10]) and similar conclusion than the RESTE model in term of significance for rehospitalization and terminal events. But the dependence between recurrences and the terminal events is not significant (p-value=0.11), contrary to the RESTE and the Rondeau's models. However, HR of terminal events are higher than for the RESTE model probably for the same reason as for the Rondeau's model. For the Zeng's model, the flexibility parameters ϕ_1 and ϕ_2, giving the direction of the association between recurrences and the terminal event hazards, also show positive dependence between recurrences and both terminal events ($\phi_1 = 0.05$ and $\phi_2 = 1.18$), but they are not significantly different from zero contrary to the Rondeau's model.

6 | DISCUSSION

In this article, we proposed a joint shared frailty model to simultaneously analyze recurrent events with two competing terminal events. Indeed, the RESTE model allows to study the joint evolution over time of recurrent events and two types of terminal events, which has a great interest in practice. The first interest of this model is that dependence between recurrent events and each type of terminal events is measured by two independent frailty parameters from a Gamma distribution, allowing greater flexilibity. The second interest is that covariates are allowed to be different for recurrent events and both type of terminal events. They may be time-dependent or time-invariant. The third advantage is that baseline hazard functions are approximated by cubic M-splines which provides smooth estimates for the three hazard functions. This allows to estimate incidence of recurrent events and mortality rates what are often of interest in epidemiology. In this article, we performed a simulation study and have shown that the RESTE model provides unbiased coefficients for the regression parameters and reasonable bias for variance of the frailty parameters.
This model is appropriate to predict occurrence of a recurrent event and/or terminal events given event past history.
This model assumes that recurrent events and each type of terminal event are positively associated. However, in some situations it may be necessary to allow for negative association between these processes. Our model may be extended to this case following the procedure used by Liu (2004)[7]. Moreover, a third frailty parameter could be added to the model to measure specifically the intra-dependence of recurrent events which could not be assessed by the model proposed in this article. Here, we only focused on two types of terminal events, but it can be easily extended to more than two different terminal events. Similarly, the RESTE model could also be extended for multivariate recurrent events. Although we used intensity and proportional hazard models, alternative models can be used. In this model, intra-dependence recurrence cannot be assessed but a solution could be to include a third frailty parameter. Finally the RESTE model assumes that parameters are time-invariant, but it could be extended to the case of covariate effects on intensity functions that change over time and thus induce a temporal effect.

References

1. Saad F, Gleason DM, Murray R et al. Long-term efficacy of zoledronic acid for the prevention of skeletal complications in patients with metastatic hormone-refractory prostate cancer. *J Natl Cancer Inst* 2004; 96(11): 879–882.

2. Xia H, Ebben J, Ma JZ et al. Hematocrit levels and hospitalization risks in hemodialysis patients. *J Am Soc Nephrol* 1999; 10(6): 1309–1316.

3. Andersen PK and Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* 1982; 10(4): 1100–1120.

4. Prentice RL, Williams BJ and Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika* 1981; 68(2): 373–379.

5. Therneau TM and Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health, New York: Springer-Verlag, 2000.

6. Liu L, Huang X. The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Stat Med* 2008; 27: 2665–2683.

7. Liu L, Wolfe RA and Huang X. Shared frailty models for recurrent events and a terminal event. *Biometrics* 2004; 60(3): 747–756.

8. Huang CY and Wang MC. Joint Modeling and Estimation for Recurrent Event Processes and Failure Time Data. *J Am Stat Assoc* 2004; 99(468): 1153–1165.

9. Fine JP and Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 1999; 94(446): 496–509.

10. Zeng D, Ibrahim JG, Chen MH et al. Multivariate recurrent events in the presence of multivariate informative censoring with applications to bleeding and transfusion events in myelodysplastic syndrome. *J Biopharm Stat* 2014; 24(2): 429–442.

11. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H et al. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics* 2007; 8(4): 708–721.

12. Mazroui Y, Mathoulin-Pelissier S, MacGrogan G, et al. Multivariate frailty models for two types of recurrent events with a dependent terminal event: Application to breast cancer data. *Biometrical Journal* 2013; 55(6): 866—884.

13. Mazroui Y, Mathoulin-Pelissier S, Soubeyran P et al. General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. *Stat Med* 2012; 31(11-12): 1162–1176.

14. Devarajan K and Ebrahimi N. On penalized likelihood estimation for a non-proportional hazards regression model. *Stat Probab Lett* 2013; 83(7): 1703–1710.

15. Ramsay JO. Monotone Regression Splines in Action. *Statist Sci* 1988; 3(4): 425–441.

16. Abramowitz M and Stegun IA. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. Government Printing Office, 1970.

17. Joly P, Commenges D and Letenneur L. A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics* 1998; 54(1): 185–194.

18. Marquardt D. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 1963; 11(2): 431–441.

19. Dennis JE Jr, Gay DM and Welsch RE. Algorithm 573: NL2sol An Adaptive Nonlinear Least-Squares Algorithm [E4]. *ACM Trans Math Softw* 1981; 7(3): 369–383.

20. M Gay D. ALGORITHM 61 Subroutines for Unconstrained Minimization Using a Model/Trust-Region Approach. *ACM Trans Math Softw* 1983; 9: 503–524.

21. Byrd R, Lu P, Nocedal J et al. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J Sci Comput* 1995; 16(5): 1190–1208.

22. Knight K. *Mathematical Statistics*. CRC Press, 1999.

23. Rondeau V, Marzroui Y and Gonzalez JR. frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *Journal of Statistical Software* 2012; 47(1): 1–28.

24. Wang W and Yan J. *splines2: Regression Spline Functions and Classes*, 2017. URL https://CRAN.R-project.org/package=splines2. R package version 0.2.6.

25. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

26. Gilbert P and Varadhan R. *numDeriv: Accurate Numerical Derivatives*, 2015. URL https://CRAN.R-project.org/package=numDeriv. R package version 2014.2-1.

27. Therneau TM, Grambsch PM. Martingale-based residuals for survival models. *Biometrika* 1990; 77(1): 147–160.

28. Liquet B, Commenges D. Estimating the expectation of the log-likelihood with censored data for estimator selection. *Lifetime Data Anal* 2004;10, 351—367.

29. Commenges D, Joly P, Gégout-Petit A., Liquet B Choice between semi-parametric estimators of Markov and non-Markov multi-state models from generally coarsened observations. *Scand. J. Statist* 2007;.34: 33-–52.

30. Austin P. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine* 2012; 31(29): 3946–3958.

31. Jha AK. To Fix the Hospital Readmissions Program, Prioritize What Matters. *JAMA* 2018; 319(5): 431–433.

32. Benbassat J and Taragin M. Hospital readmissions as a measure of quality of health care: advantages and limitations. *Arch Intern Med* 2000; 160(8): 1074–1081.

33. Shankar N, Marotta P, Wall W et al. Defining readmission risk factors for liver transplantation recipients. *Gastroenterol Hepatol (N Y)* 2011; 7(9): 585–590.

34. Steinman MA, Zullo AR, Lee Y et al. Association of β-Blockers With Functional Outcomes, Death, and Rehospitalization in Older Nursing Home Residents After Acute Myocardial Infarction. *JAMA Intern Med* 2017; 177(2): 254–262.

7 | APPENDIX

7.1 | A. The likelihood

The conditional contribution of the patient i to the likelihood is:

$$
\begin{aligned}
L_i(\omega|O_i,u_{i1},u_{i2}) = & \\
& \prod_{j=1}^{n_i}\left\{r_0(T_{ij})u_{i1}u_{i2}\exp\left(\beta_R Z_i^R(T_{ij})\right)\right\}^{\delta_{ij}}\exp\left\{-u_{i1}u_{i2}\sum_{j=1}^{n_i+1}\int_{T_{i(j-1)}}^{T_{ij}} Y_i(t)r_0(t)\exp\left(\beta_R Z_i^R(t)\right)dt\right\} \\
& \times\left\{\alpha_0^{(1)}(T_i)u_{i1}\exp\left(\beta_{D_1}Z_i^{D_1}(t)\right)\right\}^{\delta_i^{D_1}}\exp\left\{-\int_0^{T_i} Y_i(t)\alpha_0^{(1)}u_{i1}\exp\left(\beta_{D_1}Z_i^{D_1}(t)\right)dt\right\} \\
& \times\left\{\alpha_0^{(2)}(T_i)u_{i2}\exp\left(\beta_{D_2}Z_i^{D_2}(t)\right)\right\}^{\delta_i^{D_2}}\exp\left\{-\int_0^{T_i} Y_i(t)\alpha_0^{(2)}u_{i2}\exp\left(\beta_{D_2}Z_i^{D_2}(t)\right)dt\right\} \\
= & \left\{\alpha_0^{(1)}(T_i)exp\left(\beta_{D_1}Z_i^{D_1}\right)\right\}^{\delta_i^{D_1}}\left\{\alpha_0^{(2)}(T_i)exp\left(\beta_{D_2}Z_i^{D_2}\right)\right\}^{\delta_i^{D_2}}\times\prod_{j=1}^{n_i}\left\{r_0(T_{ij}))exp\left(\beta_R Z_i^R(T_{ij})\right)\right\}^{\delta_{ij}} \\
& \times u_{i1}^{n_i+\delta_i^{D_1}}u_{i2}^{n_i+\delta_i^{D_2}}exp\left\{-u_{i1}\int_0^{T_i} Y_i(t)\alpha_0^{(1)}(t)exp\left(\beta_{D_1}Z_i^{D_1}(t)\right)dt\right\} \\
& \times exp\left\{-u_{i2}(u_{i1}\sum_{j=1}^{n_i+1}\int_{T_{i(j-1)}}^{T_{ij}} Y_i(t)r_0(t)exp\left(\beta_R Z_i^R(t)\right)dt+\int_0^{T_i} Y_i(t)\alpha_0^{(2)}(t)exp\left(\beta_{D_2}Z_i^{D_2}(t)\right)dt)\right\}
\end{aligned}
$$

The marginal likelihood $L_i(\omega|O_i,u_{i1},u_{i2})$ is obtained by integrating $L_i(\omega|O_i,u_{i1},u_{i2})$ according to u_{i1} and u_{i2}.

$$L_i(\omega\mid O_i)=\int_0^{\infty}\int_0^{\infty} L_i(\omega|O_i,u_{i1},u_{i2})f(u_{i1})f(u_{i2})du_{i1}du_{i2}$$

The frailties are mutually independent and Gamma distributed with the density function f defined by:

$$f(u_{ik})=\frac{u_{ik}^{\frac{1}{\theta_k}-1}exp\left(-\frac{u_{ik}}{\theta_k}\right)}{\theta_k^{\frac{1}{\theta_c}}\Gamma(\frac{1}{\theta_k})},k=1,2$$

Thus

$$
\begin{aligned}
L_i(\omega\mid O_i)= & \\
& \left\{\alpha_0^{(1)}(T_i)exp\left(\beta_{D_1}Z_i^{D_1}\right)\right\}^{\delta_i^{D_1}}\left\{\alpha_0^{(2)}(T_i)exp\left(\beta_{D_2}Z_i^{D_2}\right)\right\}^{\delta_i^{D_2}} \\
& \times\prod_{j=1}^{n_i}\left\{r_0(T_{ij}))exp\left(\beta_R Z_i^R(T_{ij})\right)\right\}^{\delta_{ij}} \\
& \times\int_0^{\infty}\frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\theta_1}-1}exp\left\{-u_{i1}\int_0^{T_i} Y_i(t)\alpha_0^{(1)}(t)exp\left(\beta_{D_1}Z_i^{D_1}(t)\right)dt\right\}exp\left\{-\frac{u_{i1}}{\theta_1}\right\}}{\theta 1^{\frac{1}{\theta_1}}\Gamma(\frac{1}{\theta_1})} \\
& \times\int_0^{\infty}\frac{u_{i2}^{n_i+\delta_i^{D_2}+\frac{1}{\theta_2}-1}exp\left\{-u_{i2}(u_{i1}\sum_{j=1}^{n_i+1}\int_{T_{i(j-1)}}^{T_{ij}} Y_i(t)r_0(t)exp\left(\beta_R Z_i^R(t)\right)dt+\frac{1}{\theta_2}+\int_0^{T_i} Y_i(t)\alpha_0^{(2)}(t)exp\left(\beta_{D_2}Z_i^{D_2}(t)\right)dt)\right\}}{\theta_2^{\frac{1}{\theta_2}}\Gamma(\frac{1}{\theta_2})}du_{i1}du_{u2}
\end{aligned}
$$

Thanks to the properties of moment of order d of a Gamma-distribution, we have:

$$\forall x\geq 0,\forall d\in\mathbb{N},\forall a>0,\forall\lambda>0\int_0^{\infty}x^{d+a-1}\frac{\lambda^a}{\Gamma(a)}exp\left(-\lambda x\right)dx=\frac{\Gamma(d+a)}{\lambda^d\Gamma(a)}$$

Using this property with:

$d=n_i$

$a=\frac{1}{\theta_2}+\delta_i^{D_2}$

$x=u_{i2}$

$$\lambda = (u_{i1} \sum_{k=1}^{n_i+1} \int_{T_{ik}}^{T_{i(k+1)}} Y_i(t) r_0(t) exp\left(\beta_R Z_i^R(t)\right) dt + \frac{1}{\theta_2} + \int_0^{T_i} Y_i(t) \alpha_0^{(2)}(t) exp\left(\beta_{D_2} Z_i^{D_2}(t)\right) dt)$$

The marginal contribution of the patient i becomes:

$$\begin{aligned} L_i(\omega \mid O_i) = & \\ & = \left\{\alpha_0^{(1)}(T_i) exp\left(\beta_{D_1} Z_i^{D_1}\right)\right\}^{\delta_i^{D_1}} \left\{\alpha_0^{(2)}(T_i) exp\left(\beta_{D_2} Z_i^{D_2}\right)\right\}^{\delta_i^{D_2}} \\ & \times \prod_{j=1}^{n_i} \left\{r_0(T_{ij}) exp\left(\beta_R Z_i^R(T_{ij})\right)\right\}^{\delta_{ij}} \\ & \times \int_0^{\infty} \frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\theta_1}-1} exp\left\{-u_{i1}\left(\int_0^{T_i} Y_i(t)\alpha_0^{(1)}(t) exp\left(\beta_{D_1} Z_i^{D_1}(t)\right) dt + \frac{1}{\theta_1}\right)\right\}}{\theta_1^{\frac{1}{\theta_1}} \Gamma(\frac{1}{\theta_1}) \theta_2^{\frac{1}{\theta_2}} \Gamma(\frac{1}{\theta_2})} \\ & \times \frac{\Gamma\left(n_i + \frac{1}{\theta 2} + \delta_i^{D_2}\right)}{\left\{u_{i1} \sum_{j=1}^{n_i+1} \int_{T_{i(j-1)}}^{T_{ij}} Y_i(t) r_0(t) exp\left(\beta_R Z_i^R(t)\right) dt + \frac{1}{\theta_2} + \int_0^{T_i} Y_i(t)\alpha_0^{(2)}(t) exp\left(\beta_{D_2} Z_i^{D_2}(t)\right) dt\right\}^{\frac{1}{\theta 2}+\delta_i^{D_2}+n_i}} du_{i1} \end{aligned}$$

The integral has a complex form and cannot be analytically computed. Thus we approximated it using the Gauss Laguerre quadrature. The marginal log-likelihood for the patient i is :

$$\begin{aligned} ll_i\left(\omega \mid O_i\right) = & \delta_i^{D_1}\left\{log\left(\alpha_i^{(1)}(T_i)\right)\right\} + \delta_i^{D_2}\left\{log\left(\alpha_i^{(2)}(T_i)\right)\right\} \\ & + \sum_{j=1}^{n_i} \delta_{ij}\left\{log\left(r_i(T_{ij})\right)\right\} - \frac{1}{\theta_1} log(\theta_1) \\ & - log\left(\Gamma(\frac{1}{\theta_1})\right) - \frac{1}{\theta_2} log(\theta_2) - log\left(\Gamma(\frac{1}{\theta_2})\right) + log\left(\Gamma(n_i + \frac{1}{\theta 2} + \delta_i^{D_2})\right) \\ & + log\left(\int_0^{\infty} g(u_{i1}) du_{i1}\right) \end{aligned}$$

where

$$g(u_{i1}) = \frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\theta_1}-1} exp\left\{-u_{i1}\left(\int_0^{T_i} Y_i(t)\alpha_0^{(1)}(t) exp\left(\beta_{D_1} Z_i^{D_1}(t)\right) dt + \frac{1}{\theta_1}\right)\right\}}{\left\{u_{i1} \sum_{j=1}^{n_i+1} \int_{T_{i(j-1)}}^{T_{ij}} Y_i(t) r_0(t) exp\left(\beta_R Z_R(t)\right) dt + \frac{1}{\theta_2} + \int_0^{T_i} Y_i(t)\alpha_0^{(2)}(t) exp\left(\beta_{D_2} Z_i^{D_2}(t)\right) dt\right\}^{\frac{1}{\theta_2}+\delta_i^{D_2}+n_i}}$$

Finally the overall marginal log-likelihood is $ll(\omega|O) = \sum_{i=1}^{n} ll_i\left(\omega|O_i\right)$

7.2 | B. First and second derivatives of the M-spline

For $c \in \{2, .., l\}$ and $m \in \{1, ..., p + 2d - c\}$, the first derivative of $M_m(t|c, x)$ according to t is:

$$M_m^{(')}(t|c,x) = \begin{cases} \frac{c\left\{(t-x_m)M_m'(t|c-1,x)+M_m(t|c-1,x)+(x_{m+c}-t)M_{m+1}'(t|c-1,x)-M_{m+1}(t|c-1,x)\right\}}{(c-1)(x_{m+c}-x_m)} & \text{if } x_m \le t \le x_{m+c} \\ 0 & \text{otherwise} \end{cases}$$

and the second derivarive is:

$$M_m''(t|c,x) = \begin{cases} \frac{c\left\{(t-x_m)M_m''(t|c-1,x)+2M_m'(t|c-1,x)+(x_{m+c}-t)M_{m+1}''(t|c-1,x)-2M_{m+1}'(t|c-1,x)\right\}}{(c-1)(x_{m+c}-x_m)} & \text{if } x_m \le t \le x_{m+c} \\ 0 & \text{otherwise} \end{cases}$$

□

4.3 Fonction de risque de base approchée par des M-Splines

Le terme "spline" désigne une large classe de fonctions qui sont utilisées dans des applications nécessitant une interpolation et / ou un lissage des données. C'est une fonction polynomiale par morceaux et une combinaison linéaire de splines peut ainsi approcher une fonction inconnue. Dans les problèmes d'interpolation, la méthode des splines donnent des résultats similaires à l'interpolation polynomiale même lors de l'utilisation de polynômes de faible degré, tout en évitant le phénomène de Runge [64] pour les degrés supérieurs. C'est pour cela qu'elle est généralement préférée à la méthode de l'interpolation polynomiale.
Dans cette thèse, nous avons choisi d'approcher la fonction de risque de base par des M-splines, version normalisée des B-splines, car ce sont des fonctions positives ou nulles. Elles ont aussi l'avantage de pouvoir approcher la fonction de risque de base cumulée par des I-splines (primitive des M-splines), fonctions monotones et croissantes, nécessaires dans l'estimation de la fonction de survie. Ces deux types de splines ont été présentés et décrits dans Ramsay [65] et Joly [66]. L'utilisation de la méthode de M-splines nécessite de choisir l'ordre k des M-splines, et de définir une séquence de noeuds noté $x = \{x_1, ..., x_{p+2d}\}$, où p est le nombre de noeuds internes et borné entre 0 and $\max_{i=1..n}(T_i)$, où $T_i = min(C_i, D_i)$ où C_i est le temps de censure et D_i est le temps de l'événement terminal survenant en premier entre l'événement terminal d'intérêt et le risque compétitif, tel que $x_1 = ... = x_d = 0$ et $x_{p+d+1} = ... = x_{p+2d} = \max_{i=1..n}(T_i)$. Les M-splines sont donc des polynômes de degrés $k-1$. Le calcul des différents polynômes m se fait à partir de la séquence de noeuds et de façon récursive.

— Pour $c = 1$ et $\forall m \in \{1, ..., p+2d-1\}$

$$M_m(t|c,x) = \begin{cases} \frac{1}{(x_{m+1}-x_m)} & \text{si } x_m \leq t \leq x_{m+1} \\ 0 & \text{sinon} \end{cases}$$

— $\forall c \in \{2, .., d\}$ et $\forall m \in \{1, ..., p+2d-c\}$

$$M_m(t|c,x) = \begin{cases} \frac{c\{(t-x_m)M_m(t|c-1,x)+(x_{m+c}-t)M_{m+1}(t|c-1,x)\}}{(c-1)(x_{m+c}-x_m)} & \text{si } x_m \leq t \leq x_{m+c} \\ 0 & \text{sinon} \end{cases}$$

Les trois fonctions de risque de base sont donc estimées en utilisant une combinaison linéaire de la même base de M-splines d'ordre k. Seuls les coefficients sont différents.

$$\begin{cases} \tilde{r}_0(t) = \gamma_1^R M_1(t|4,x) + \gamma_2^R M_2(t|4,x) + ... + \gamma_{p+d}^R M_{p+d}(t|4,x) & \text{(récurrences)} \\ \tilde{\alpha}_0^{(1)}(t) = \gamma_1^{(1)} M_1(t|4,x) + \gamma_2^{(1)} M_2(t|4,x) + ... + \gamma_{(p+d)}^{(1)} M_{(p+d)}(t|4,x) & \text{(événement terminal 1)} \\ \tilde{\alpha}_0^{(2)}(t) = \gamma_2^{(2)} M_1(t|4,x) + \gamma_2^{(2)} M_2(t|4,x) + ... + \gamma_{(p+d)}^{(2)} M_{p+d}(t|4,x) & \text{(événement terminal 2)} \end{cases}$$

où $\gamma = \{\gamma_1^R, ..., \gamma_{p+d}^R, \gamma_1^{(1)}, ..., \gamma_{(p+d)}^{(1)}, \gamma_2^{(2)}, ..., \gamma_{(p+d)}^{(2)}\}$ sont les coefficients associés aux M-splines.

Le nombre de paramètres estimés pour chacune des trois M-splines cubiques est le nombre de nœuds internes plus l'ordre des splines $(p+d)$. Plus le nombre de nœuds est élevé, meilleure est l'approximation. Cependant, si le nombre de nœuds est élevé, des fluctuations locales peuvent apparaître, mais la pénalisation gère ce problème. Nous avons donc choisi une pénalisation (section 4.4) par la dérivée seconde de la fonction de risque [65], de ce fait nous avons choisi des M-splines d'ordre 4 soit des polynômes de degré 3.
La variance des fonctions de risque de base est estimée comme suit [67] :

$$Var\left(\tilde{r_0}(.)\right) = \mathbf{M}(.)^{\mathbf{T}}\mathbf{I}_{\gamma}^{-\mathbf{1}}\mathbf{M}(.), \quad Var\left(\tilde{\alpha_0}^{(1)}(.)\right) = \mathbf{M}(.)^{\mathbf{T}}\mathbf{I}_{\gamma_{\mathbf{1}.}}^{-\mathbf{1}}\mathbf{M}(.), \quad Var\left(\tilde{\alpha_0}^{(2)}(.)\right) = \mathbf{M}(.)^{\mathbf{T}}\mathbf{I}_{\gamma_{\mathbf{2}.}}^{-\mathbf{1}}\mathbf{M}(.)$$

où $M(.) = (M_1(.), ..., M_{p+d}(.))$ est le vecteur de M-splines et $I_\gamma = \frac{\partial^2 pll(\omega|O_i)}{\partial\gamma^2}$, $I_{\gamma_{1.}} = \frac{\partial^2 pll(\omega|O_i)}{\partial\gamma_{1.}^2}$, $I_{\gamma_{2.}} = \frac{\partial^2 pll(\omega|O_i)}{\partial\gamma_{2.}^2}$ est un sous-ensemble de la matrice Hessienne correspondant à la matrice de variance-covariance associée aux vecteurs des paramètres γ, γ_1 et γ_2 respectivement.
Alors l'intervalle de confiance à 95% s'écrit :

$$\tilde{r_0}(.) \pm Z_{1-\frac{\alpha}{2}} \times\sqrt{Var\left(\tilde{r_0}(.)\right)}, \quad \tilde{\alpha_0}^{(1)}(.) \pm Z_{1-\frac{\alpha}{2}} \times\sqrt{Var\left(\tilde{\alpha_0}^{(1)}(.)\right)}, \quad \tilde{\alpha_0}^{(2)}(.) \pm Z_{1-\frac{\alpha}{2}} \times\sqrt{Var\left(\tilde{\alpha_0}^{(2)}(.)\right)}$$

4.4 Vraisemblance pénalisée, paramètres de lissage et algorithme de maximisation

4.4.1 La vraisemblance pénalisée

Comme énoncé dans le paragraphe précédent, nous avons choisi de pénaliser la vraisemblance par la norme L_2 de dérivée seconde de la fonction à estimer. Pour lisser les trois fonctions de risque de base et les rendre continues, nous avons ajouté trois termes de lissage ξ_R, ξ_{D_1} et ξ_{D_2}. La log vraisemblance pénalisée s'écrit alors :

$$pll(\omega|O) = ll\left(\omega|O\right) - \xi_R \int_0^\tau [r_0^{''}(t)]^2 dt - \xi_{D_1} \int_0^\tau [\alpha_0^{''(1)}(t)]^2 dt - \xi_{D_2} \int_0^\tau [\alpha_0^{''(2)}(t)]^2 dt \tag{4.1}$$

où $ll(\omega)$ est la log-vraisemblance, r_0, $\alpha_0^{(1)}$ et $\alpha_0^{(2)}$ sont les fonctions de risque de base pour les événements récurrents, et les événements terminaux respectivement. La fonction de risque de base étant approximée par une combinaison de M-splines, nous obtenons alors

$$\xi_R \int_0^\tau [r_0^{''}(u)]^2 du = \int_0^\tau \xi_R^T M''(u) M''(u) \xi_R du$$

$$\xi_{D_1} \int_0^\tau [\alpha_0^{''(1)}(u)]^2 du = \int_0^\tau \xi_{D_1}^T M''(u) M''(u) \xi_{D_1} du$$

$$\xi_{D_2} \int_0^\tau [\alpha_0^{''(2)}(u)]^2 du = \int_0^\tau \xi_{D_2}^T M''(u) M''(u) \xi_{D_2} du$$

où $M''(.)$ est la matrice des dérivées secondes des M-splines et ξ_R, ξ_{D_1} et ξ_{D_2} les vecteurs de coefficients pour les événements récurrents et chaque type d'événement terminal. Les dérivées première et seconde des M-splines sont données dans l'annexe B.3.

4.4.2 Les paramètres de lissage

Le choix du paramètre de lissage est important puisque le sur-lissage ou le sous-lissage aboutissent tous les deux à des estimateurs de mauvaise qualité. Plusieurs méthodes existent pour choisir le paramètre de lissage. Certaines méthodes, comme le "choix subjectif" et la méthode du "graphe test" nécessitent de tracer des graphes à partir de plusieurs estimations du paramètre de lissage et de choisir la plus réaliste. Nous avons choisi d'utiliser une méthode de validation croisée qui est une méthode automatique. Introduite par Duin [68] et Habbema [69], elle consiste à maximiser le critère de validation croisée (CV). Le principe de la méthode de validation croisée "leave-one-out", consiste à estimer les paramètres d'intérêt sur un échantillon d'apprentissage, étant l'échantillon total auquel une observation a été retirée. La fonction d'intérêt est ensuite calculée à partir des paramètres estimés précédemment sur l'échantillon test, constitué seulement de l'observation exclue de l'échantillon d'apprentissage. Le principe est réitéré pour chaque observation. Le critère de validation croisée s'écrit alors pour le paramètre de lissage pour les événements récurrents :

$$CV(\xi_R) = \frac{1}{n} \sum_{i=1}^{n} ll_i(\hat{\gamma_{\xi_R}^{R}}^{-i})$$

où $ll_i(.)$ est la contribution à la log-vraisemblance du sujet i, n est le nombre de sujets dans l'échantillon d'apprentissage, $\hat{\gamma_{\xi_R}^{R}}^{-i}$ est l'estimateur du maximum de vraisemblance pénalisée calculé sur l'échantillon en excluant le sujet i. Cette procédure étant chronophage, O'Sullivan a proposé en 1988, une approximation du critère de la validation croisée pour le modèle de Cox [70] :

$$CV_{approchée}(\xi_R) = \frac{1}{n} \left(- \sum_{i=1}^{n} ll_i(\hat{\gamma_{\xi_R}^{R}}^{-i}) + trace \left[I(\hat{\gamma^R}) + 2\xi_R \Omega \right]^{-1} I(\hat{\gamma^R}) \right)$$

où $I(\hat{\gamma^R})$ est la matrice d'information de Fisher et $\Omega = \int_0^\tau M''(u) M''(u) du$.

En pratique, le paramètre de lissage est calculé pour les événements récurrents et chaque événement terminal séparément. Pour cela, nous avons utilisé un modèle de Cox avec validation croisée pour chaque événement terminal, à partir de la fonction `frailtyPenal` du package `frailtypack` du logiciel `R`, et un modèle à fragilité avec validation croisée pour les événements récurrents à partir de la même fonction.

4.4.3 L'algorithme de maximisation

Pour maximiser la vraisemblance pénalisée, nous avons utilisé un algorithme quasi-Newton, une variante de la méthode Broyden-Fletcher-Goldfarb-Shanno (BFGS) [71][72][73][74]. L'objectif est le même que celui de la méthode de Newton qui cherche pour une fonction donnée $f : \mathbf{R}^n \to \mathbf{R}^n$ les solutions $\hat{\omega}$ tel que $f(\hat{\omega}) = 0$.
L'estimateur du maximum de vraisemblance pénalisée des paramètres $\hat{\omega}$ est estimé par la récurrence suivante :

$$x_{k+1} = x_k - \rho_k \; B_k \cdot f(x_k)$$

où ρ_k est un paramètre réel choisi pour optimiser la convergence, et B_k est une matrice, qui cherche à approximer l'inverse de la matrice Hessienne de f en x_k. Elle est mise jour à chaque itération par la méthode de la sécante [75]. La méthode quasi-Newton est préférée à la méthode de Newton car cette méthode est plus rapide. En effet, la méthode de Newton requiert le calcul de la matrice jacobienne, dont le calcul est chronophage si la dimension n du système est grande.

4.5 Le test des effets aléatoires

Les effets aléatoires, inclus dans le modèle permettent de mesurer la dépendance entre les événements récurrents et chaque événement terminal. Une question qui se pose alors est de savoir si cette dépendance est statistiquement significative, ie si la variance de l'effet aléatoire est significativement différente de 0. La variance d'un effet aléatoire, notée σ^2, étant positive ou nulle, il est nécessaire de faire un test unilatéral dont les hypothèses sont :

— Hypothèse nulle $H_0 : \sigma^2 = 0$
— Hypothèse alternative H_1 est $\sigma^2 > 0$.

Le test est fait à la limite inférieure du domaine de définition de σ^2. Nous pouvons cependant, utiliser un test de Wald unilatéral dont le quantile à 5% est de 1.64 [76].

4.6 Sélection et Vérification de la qualité du modèle

4.6.1 Vérification de l'adéquation du modèle

Les résidus de martingales sont un bon outil d'évaluation de l'adéquation pour les modèles ajustés issus d'un processus de comptage. Ils représentent les erreurs entre le modèle et les données au cours du temps et sont interprétés comme étant la différence entre le nombre observé et attendu d'événements pour chaque sujet i. Le calcul des résidus de martingales se déduit de la décomposition de Doob-Meier : $M_i(t) = N_i(t) - \Lambda_i(t)$, où $N_i(t)$ est le vecteur composé par le nombre d'événements observés sur l'intervalle $[0, t]$ pour le sujet i, et $\Lambda_i(t)$ est le vecteur des compensateurs associés. Dans le modèle proposé, les résidus de martingales pour chaque type

d'événement s'écrivent :

$$\begin{cases} M_i^R(t) = N_i^R(t) - Y_i(t)u_{i1}u_{i2}\int_0^t \hat{r}_i(s)ds & \text{(récurrences)} \\ M_i^{D_1}(t) = N_i^{D_1}(t) - Y_i(t)u_{i1}\int_0^t \hat{\alpha}_i^{(1)}(s)ds & \text{(événement terminal 1)} \\ M_i^{D_2}(t) = N_i^{D_2}(t) - Y_i(t)u_{i2}\int_0^t \hat{\alpha}_i^{(2)}(s)ds & \text{(événement terminal 2)} \end{cases} \tag{4.2}$$

où $Y_i(t)$ est l'indicatrice précisant si le sujet i est à risque au temps t. Les paramètres $(\beta, \theta_1, \theta_2, \gamma)$ utilisés pour estimer les résidus des martingales sont ceux estimés par la maximisation de la vraisemblance pénalisée. Cependant, dans le modèle proposé, seules les variances des effets aléatoires sont estimées alors que les estimations individuelles des effets de fragilité $\hat{u}_{i1}$ et $\hat{u}_{i2}$ sont nécessaires pour calculer les résidus de martingales. Ainsi, un estimateur bayésien empirique, le mode de la densité *a posteriori*, a été utilisé et est égal à la valeur maximisant $\arg\min_{u_{i1},u_{i2}} f(u_{i1}, u_{i2}|O_i, \hat{\omega})$. La densité *a posteriori* des effets aléatoires s'obtient par :

$$f(u_{i1}, u_{i2}|O_i, \hat{\omega}) \propto f(O_i|u_{i1}, u_{i2}, \hat{\omega}) f(u_{i1}, u_{i2}|\hat{\omega})$$

où $f(u_{i1}, u_{i2}|O_i, \hat{\omega})$ est la fonction de densité *a posteriori* des paramètres de fragilité, $f(O_i|u_{i1}, u_{i2}, \hat{\omega})$ est la contribution individuelle à la vraisemblance conditionnellement aux effets aléatoires pour le sujet i sachant $\hat{\omega}$, u_{i1} et u_{i2} et $f(u_{i1}, u_{i2}|\hat{\omega})$ est la fonction de densité des paramètres de fragilité. Donc,

$$f(u_{i1}, u_{i2}|O_i, \hat{\omega}) \propto \frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\hat{\theta}_1}-1} exp\left\{-u_{i1}\int_0^{T_i} Y_i(t)\hat{\alpha}^{(1)}(t)dt\right\} exp\left\{-\frac{u_{i1}}{\hat{\theta}_1}\right\}}{\hat{\theta}_1^{\frac{1}{\hat{\theta}_1}}\Gamma(\frac{1}{\hat{\theta}_1})} \times \frac{u_{i2}^{n_i+\delta_i^{D_2}+\frac{1}{\hat{\theta}_2}-1} exp\left\{-u_{i2}(u_{i1}\sum_{j=1}^{n_i+1}\int_{T_{i(j-1)}}^{T_{ij}} Y_i(t)\hat{r}(t)dt + \frac{1}{\hat{\theta}_2} + \int_0^{T_i} Y_i(t)\hat{\alpha}^{(2)}(t)dt)\right\}}{\hat{\theta}_2^{\frac{1}{\hat{\theta}_2}}\Gamma(\frac{1}{\hat{\theta}_2})}$$

L'algorithme de Marquardt [77] a été utilisé pour estimer le mode de la fonction de densité postérieure. Les résidus de martingales ont en théorie une espérance nulle. L'adéquation du modèle est bonne si lors du tracé des résidus en fonction du temps pour chaque covariable, la moyenne des résidus de martingales est égale à 0, idéalement ils sont centrés autour de l'axe des abscisses.

4.6.2 Sélection du modèle

Du fait de la pénalisation de la vraisemblance, les critères de sélection classiques comme le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC) ne sont pas appropriés. Liquet et al [78] ont discuté de l'utilisation du critère de validation croisée de la vraisemblance (LCV) pour choisir entre deux modèles semi-paramétriques. Commenges et al

[79] ont proposé le critère de validation croisée de vraisemblance approximative (LCVa) qui permet de comparer des modèles multi-états ou issus de processus de comptage et estimés par maximisation de vraisemblance pénalisée ou non. Le LCVa s'écrit :

$$LCVa = \frac{trace(\hat{H_{pl}^{-1}} I) - l(\hat{\omega})}{\sum_{i=1}^{n} n_i}$$

où $\hat{H_{pl}^{-1}} I$ est le produit de la matrice hessienne et de la matrice d'information de Fisher, n le nombre de sujets et n_i le nombre d'observations pour le sujet i.

4.7 Simulations complémentaires

4.7.1 Simulations des données

Dans l'article, nous avons évalué le comportement du modèle en faisant varier la forme de la fonction de risque de base pour les événements récurrents, mais aussi en prenant l'événement en compétition indépendant des événements récurrents.
Dans le modèle proposé, les effets aléatoires sont modélisés par une loi Gamma. Le choix de cette loi s'est surtout basé sur ses propriétés mathématiques puisqu'aucun argument clinique ne permet d'indiquer les distributions les plus adaptées. Cependant, en pratique, il est difficile de vérifier cette hypothèse. Dans un premier temps, nous avons donc simulé des effets aléatoires de distribution Gamma, puis nous avons aussi voulu voir le comportement du modèle lorsque les effets aléatoires sont issus d'une loi log-normale.
Pour cela, nous avons simulé 500 bases de données à partir du modèle suivant :
$\forall i \in [1, ...n]$

$$, \begin{cases} r_i(t|u_{i1}, u_{i2}) = u_{i1}u_{i2}r_0(t)exp\left(\beta_1 Z_{1i} + \beta_2 Z_{2i}\right) & (1) \\ \alpha_i^{(1)}(t|u_{i1}) = u_{i1}\alpha_0^{(1)}(t)exp\left(\beta_3 Z_{1i}\right) & (2) \\ \alpha_i^{(2)}(t|u_{i2}) = u_{i2}\alpha_0^{(2)}(t)exp\left(\beta_4 Z_{1i}\right) & (3) \end{cases}$$

où Z_{1i} et Z_{2i} sont des variables binaires, toutes deux générées à partir d'une loi de Bernoulli avec une probabilité de 0,5. Les paramètres sont $\beta_1 = 1$, $\beta_2 = -0,5$, $\beta_3 = 0,7$ et $\beta_4 = -1$. Les fonctions de risques de bases sont constantes avec $r_0(t) = 8$, $\alpha^{(1)}(t) = 1.5$ et $\alpha^{(2)}(t) = 2$. Pour chaque patient les données sont simulées en deux étapes de façon similaire à ce qui a été décrit dans la section 3.2 :

— La première étape a été la génération du temps pour les événements terminaux. Le temps de l'événement terminal a été généré à partir d'une distribution exponentielle selon les modèles (2) et (3). Le temps de censure a été généré à partir d'une distribution uniforme sur [0 ; 1] et nous avons supposé une proportion de données censurées d'environ un tiers. Le dernier temps observé a été défini comme le minimum entre les temps de censure et l'événement terminal. La fonction indicatrice pour le dernier temps observé est égale à

0, 1 ou 2 si le temps minimum est égal au temps de censure, au temps pour l'événement terminal de cause 1 ou au temps pour l'événement terminal de cause 2, respectivement.

— La deuxième étape a été la génération de temps d'événements récurrents. Les temps entre deux événements récurrents successifs ont été générés à partir d'une distribution exponentielle suivant le modèle (1). Nous n'avons pris en compte que les événements récurrents avec un temps total inférieur ou égal au dernier temps observé défini à la première étape.

Dans notre modèle nous supposons que les effets aléatoires sont issus d'une loi Gamma, nous voulons voir comment se comporte le modèle lorsque la distribution des effets aléatoires n'est pas respectée. Ainsi, dans ce scénario, nous avons utilisé deux lois log-normales de moyennes 1 et de variances $\theta_1 = 0.5$ et $\theta_2 = 0.5$ pour la dépendance entre les événements récurrents et l'événement terminal d'intérêt et en compétition respectivement.

Dans l'article, nous avons simulé les données avec 500 patients et avons montré que le modèle estime bien les paramètres. Nous avons alors voulu voir le comportement du modèle lorsque la taille de l'échantillon est plus petite. Nous avons ainsi simulé tous les scénarios présentés dans l'article avec une taille d'échantillon $N = 350$.

4.7.2 Résultats

Le nombre d'événements récurrents moyen par patient varie de 1,3 à 2,0 en fonction des scénarios. La proportion de patients présentant l'événement terminal 1 varie de 41,4% à 54,5% et de 26,5% à 35,5% pour l'événement terminal 2 (Table 4.1).

Nous avons vu dans l'article que le modèle RESTE estimait de façon non biaisé les paramètres de régression quel que soit le scénario, lorsque la taille de l'échantillon est N=500. Nous avons reproduit les mêmes scénarios en diminuant la taille de l'échantillon à N=350 (Table 4.2). Les résultats montrent que le modèle RESTE produit des estimations non biaisées ou très faiblement biaisées quel que soit la forme de la fonction de base. Les paramètres sont aussi sans biais lorsqu'il n'y a pas de dépendance entre les événements récurrents et l'événement terminal de type 2. Dans ce dernier scénario, le paramètre θ_2 est quant à lui surestimé. L'analyse de sensibilité sur la distribution des effets aléatoires, ie la distribution des effets aléatoires ne suivent pas une loi Gamma, les estimations des paramètres du modèle RESTE ne semblent pas affectées.

Nous avons comparé les résultats des simulations du modèle RESTE avec les modèles de Zeng et de Rondeau. Les résultats des simulations montrent que le modèle de Zeng estime sans biais les paramètres β_1 et β_2 mais sous-estime les paramètres β_3 et β_4 pour tous les scenarios sauf lorsque la dépendance entre les événements récurrents et l'événement terminal de type 2 est nulle. Dans ce scénario, les paramètres β_3 et β_4 sont surestimés et comme attendu, le paramètre ϕ_2 est très proche de 0. Cependant lorsque les effets aléatoires sont distribués suivant une

TABLE 4.1 – Description des simulations

	Nombre d'événements récurrents moyen	Evénement terminal 1 (%)	Evénement terminal 2 (%)
$r_0(t) = 8$	1,99	41,46	26,51
$r_0(t) = 32t$	1,32	41,46	26,51
$r_0(t) = \lambda\gamma t^{\gamma-1}$	1,49	41,46	26,51
$r_0(t) = 8$ et $\theta_2 = 0,01$	2,11	41,43	28,44
$r_0(t) = 2$ et Loi Log-normale	2,02	54,50	35,47

loi log-normale, le modèle de Zeng semble fournir des estimations moins biaisées. Ceci est certainement dû au fait que dans le modèle de Zeng les effets aléatoires sont issus d'une loi normale.

Le modèle de Rondeau surestime légèrement le paramètre β_1 lorsque la fonction de risque de base est constante ou affine. Comme pour le modèle de Zeng, le paramètre β_3 est sous-estimé probablement parce que les patients ayant eu l'événement terminal 2 ont été censurés dans cette analyse. Tout comme le modèle RESTE, les estimations ne sont pas sensibles au fait que les effets aléatoires ne suivent pas une loi gamma. Et comme attendu, les estimations sont sans biais lorsque la dépendance entre les événements récurrents et l'événement terminal de type 2 est nulle.

De façon globale, les paramètres β_1 et β_2 sont généralement bien estimés dans tous les scénarios par les trois modèles, même si le modèle de Zeng estime ces paramètres quasiment sans biais sûrement du fait que celui-ci est plus parcimonieux que les modèles RESTE et de Rondeau. Pour les paramètres β_3 et β_4, le modèle RESTE fournit des estimations meilleures que les modèles de Rondeau et de Zeng.

TABLE 4.2 – Résultats des simulations

Paramètres des simulations	Paramètres	RESTE Estimation (se$*$)	Liu Estimation (se$*$)	Zeng Estimation (se$*$)
$r_0(t) = 8$	$\beta_1 = 1$	0,96 (0,18)	1,06 (0,23)	1,00 (0,16)
	$\beta_2 = -0,5$	-0,50 (0,15)	-0,47 (0,28)	-0,49 (0,16)

Paramètres des simulations	Paramètres	RESTE Estimation (se∗)	Rondeau Estimation (se∗)	Zeng Estimation (se∗)
	$\beta_3 = 0,7$	0,66 (0,31)	0,64 (0,19)	0,62 (0,21)
	$\beta_4 = -1$	-1,02 (0,30)		-0,93 (0,24)
	$\theta_1 = 0,5$	0,53 (0,16)		
	$\theta_2 = 0,5$	0,52 (0,16)		
	$\tilde{\theta} \sim ?$		0,85 (0,08)	
	$\alpha =?$		0,49 (0,21)	
	$\xi =?$			0,67 (0,14)
	$\phi_1 =?$			0,81 (0,24)
	$\phi_2 =?$			0,79 (0,29)
$r_0(t) = 32t$	$\beta_1 = 1$	0,99 (0,26)	1,04 (0,20)	0,99 (0,23)
	$\beta_2 = -0,5$	-0,46 (0,19)	-0,48 (0,23)	-0,47 (0,20)
	$\beta_3 = 0,7$	0,66 (0,28)	0,65 (0,19)	0,62 (0,21)
	$\beta_4 = -1$	-1,04 (0,29)		-0,95 (0,26)
	$\theta_1 = 0,5$	0,47 (0,19)		
	$\theta_2 = 0,5$	0,50 (0,20)		
	$\tilde{\theta} \sim ?$		0,88 (0,07)	
	$\alpha =?$		0,49 (0,17)	
	$\xi =?$			0,75 (0,20)
	$\phi_1 =?$			0,73 (0,33)
	$\phi_2 =?$			0,72 (0,39)
$r_0(t) = 8\gamma t^{\gamma-1}$, $\gamma = 0,75$	$\beta_1 = 1$	0,95 (0,20)	1,05 (0,15)	1,00 (0,17)
	$\beta_2 = -0,5$	-0,50 (0,16)	-0,48 (0,20)	-0,50 (0,17)
	$\beta_3 = 0,7$	0,66 (0,29)	0,67 (0,16)	0,62 (0,21)
	$\beta_4 = -1$	-1,03 (0,30)		-0,93 (0,25)
	$\theta_1 = 0,5$	0,53 (0,17)		

Paramètres des simulations	Paramètres	RESTE Estimation (se∗)	Rondeau Estimation (se∗)	Zeng Estimation (se∗)
	$\theta_2 = 0,5$	0,50 (0,17)		
	$\tilde{\theta} \sim ?$		0,47 (0,12)	
	$\alpha =?$		0,48 (0,23)	
	$\xi =?$			0,64 (0,13)
	$\phi_1 =?$			0,85 (0,26)
	$\phi_2 =?$			0,79 (0,34)
Loi Log-normale	$\beta_1 = 1$	1,01 (0,15)	0,99 (0,13)	0,99 (0,14)
$r_0(t) = 2$	$\beta_2 = -0,5$	-0,47 (0,14)	-0,49 (0,16)	-0,50 (0,13)
	$\beta_3 = 0,7$	0,71 (0,19)	0,71 (0,22)	0,64 (0,17)
	$\beta_4 = -1$	-0,97 (0,23)		-0,97 (0,23)
	$\theta_1 = 0,5$	0,25 (0,07)		
	$\theta_2 = 0,5$	0,26 (0,07)		
	$\tilde{\theta} \sim ?$		0,53 (0,10)	
	$\alpha =?$		1,05 (0,30)	
	$\xi =?$			0,29 (0,06)
	$\phi_1 =?$			0,88 (0,33)
	$\phi_2 =?$			0,85 (0,46)
$r_0(t) = 8\ \theta_2$	$\beta_1 = 1$	0,98 (0,15)	0,99 (0,13)	1,00 (0,13)
	$\beta_2 = -0,5$	-0,50 (0,11)	-0,49 (0,16)	-0,50 (0,12)
	$\beta_3 = 0,7$	0,69 (0,28)	0,71 (0,22)	0,79 (0,23)
	$\beta_4 = -1$	-1,01 (0,24)		-1,03 (0,24)
	$\theta_1 = 0,5$	0,47 (0,11)		
	$\theta_2 = 0,01$	0,09 (0,06)		
	$\tilde{\theta} \sim$0,5		0,53 (0,10)	
	$\alpha = 1$		1,05 (0,30)	

Paramètres des simulations	Paramètres	RESTE Estimation (se∗)	Rondeau Estimation (se∗)	Zeng Estimation (se∗)
	ξ =?			0,37 (0,09)
	ϕ_1 =?			1,66 (0,36)
	ϕ_2 =?			0,003 (0,44)

∗ Erreur standard de la moyenne de l'estimateur. $\tilde{\theta}$ est la variance du paramètre de fragilité dans le modèle de Rondeau capturant la dépendance entre les événements récurrents et l'événement terminal de type 1. α est un paramètre de flexibilité dans le modèle de Rondeau déterminant la direction de l'association entre les événements récurrents et la fonction de risque d'événements terminaux. ξ est la variance de l'effet aléatoire partagée par tous les événements (récurrents et événements terminaux) qui capture la dépendance entre les événements récurrents et les événements terminaux. ϕ_1 and ϕ_2 sont des paramètres de flexibilité dans le modèle de Zeng donnant la direction de l'association entre les événements récurrents et les événements terminaux.

4.8 Conclusion et discussion

Le modèle présenté dans ce chapitre permet d'analyser l'effet des covariables simultanément sur les événements récurrents et sur l'événement terminal en présence d'un risque compétitif sur l'événement terminal. De plus ce modèle permet d'évaluer la dépendance entre les événements récurrents et l'événement terminal ainsi que sur le risque compétitif. Les covariables peuvent être dépendantes du temps ou non, et les covariables incluses peuvent être différentes pour les événements récurrents et les événements terminaux.

Cependant, de nombreuses extensions de ce modèle peuvent être développées. En effet, nous nous sommes intéressés à un seul type d'événement récurrent. Il est cependant possible de s'intéresser à plusieurs types d'événements récurrents. De plus, les coefficients de régression sont constants au cours du temps. Or en pratique, l'effet d'une covariable peut varier au cours du temps. Une extension alors possible serait d'inclure des coefficients de régression dépendants du temps. Une autre extension possible serait d'inclure un autre terme de fragilité pour différencier la dépendance intra-récurrence et la dépendance entre les événements récurrents et les événements terminaux. Dans ces chapitres, nous avons utilisé des modèles de régressions pour analyser les réhospitalisations en présence d'un ou plusieurs événements terminaux. Cependant, ces modèles ne s'intéressent qu'au fait de savoir si le patient a été réhospitalisé ou non et dans quels délais. Le motif des réhospitalisations nous offre une information supplémentaire que l'on peut aussi utiliser.

Deuxième partie

Trajectoire de soins des patients après une chirurgie bariatrique : utilisation de méthodes de fouille de données

Les motifs de réhospitalisations sont une source d'information importante. Les analyser peut permettre d'identifier des trajectoires communes à plusieurs sujets. En effet, selon les maladies, les patients peuvent être amenés à être hospitalisés plusieurs fois, soit pour le même motif soit pour un motif différent. Trouver des combinaisons de motifs fréquents permettrait de mieux prendre en charge les soins des patients et d'anticiper de futures réhospitalisations.

Ainsi, nous nous sommes intéressés aux trajectoires de soins de patients dans l'année suivant la chirurgie bariatrique. Les données sont issues du PMSI et contiennent par conséquent une grande quantité d'information. Afin de pouvoir extraire de l'information pertinente de ces données et de construire les trajectoires de soins, nous avons utilisé une méthode de fouille de données.

Chapitre 5

La fouille de données et l'analyse formelle de concept (état de l'art)

5.1 La fouille de données et la découverte de connaissances à partir de données

Actuellement, de plus en plus de données de différents types s'accumulent et constituent une source d'information conséquente. Due à une grande quantité de données, l'information contenue est difficilement exploitable et interprétable en tant que telle. Le besoin de moyens pour extraire des informations intéressantes a donné naissance à la fouille de données. La fouille de données ou encore data-mining, exploration des données, ou même forage de données, consiste à extraire de l'information non triviale, utile, compréhensible et interprétable en découvrant des relations entre les données à partir d'une grande quantité de données. Selon Fayyad [80], la fouille de données est une étape cruciale du processus de découverte de connaissances dans les bases de données collectées. Ce processus inclut cinq différentes étapes indispensables après avoir bien identifié et énoncé les besoins (Figure 5.1). La première étape est de sélectionner des données représentatives du problème. Une fois les données sélectionnées, il faut les préparer. En effet les données peuvent contenir des informations manquantes ou incorrectes. Cette étape de préparation des données consiste alors à enlever les valeurs aberrantes et traiter les données manquantes afin d'enlever le bruit. Ensuite vient l'étape de transformation, qui est définie comme le processus de transformation des données en une forme appropriée requise par la procédure d'exploration. Une fois les données prêtes, l'étape de la fouille de données peut commencer. Elle consiste à utiliser la technique la plus pertinente pour extraire les motifs (structures caractéristiques communes à plusieurs individus) les plus utiles, ainsi que l'algorithme le plus approprié afin de répondre au mieux à la problématique. Enfin vient l'étape d'interprétation et d'évaluation afin de visualiser les résultats fournis lors de l'étape de fouille de données.

Le fondement de la fouille de données repose sur trois disciplines scientifiques telles que les

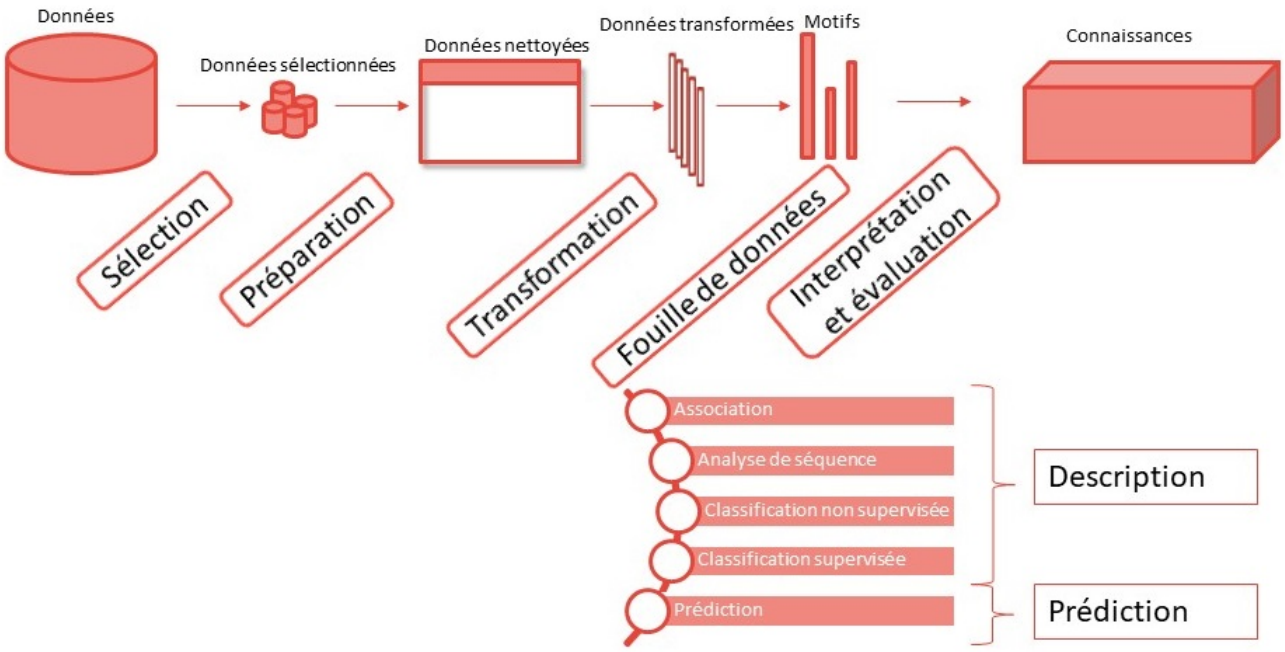


FIGURE 5.1 – Processus de découverte de connaissances à partir des méthodes de fouille des données

statistiques, l'intelligence artificielle et l'apprentissage automatique (machine learning). Du fait de la grande quantité de données, des méthodes automatiques ou semi-automatiques, reposant sur des algorithmes complexes et sophistiqués permettant de segmenter les données, sont utilisées dans la fouille de données. La fouille de données comporte cinq techniques principales qui peuvent être catégorisées en deux grandes classes : les méthodes dites de description et les méthodes de prédiction. Les méthodes de description incluent les techniques d'associations, l'analyse de séquences, la classification supervisée et non supervisée (clustering). Elles ont pour but de mettre en évidence des informations cachées dans une grande quantité de données. Les méthodes de prédiction quant à elles cherchent à extrapoler l'information présente à de l'information future.

5.2 Les méthodes de description de la fouille de données

5.2.1 Les méthodes d'association et d'analyses de séquence

Les méthodes d'association consistent à mettre en évidence des règles d'association en cherchant une corrélation entre deux ou plusieurs éléments en identifiant le motif caché dans le jeu de données. Les relations non couvertes peuvent être représentées sous forme de règles d'association ou d'ensembles d'éléments fréquents. Par exemple :

Identifiant	Combinaison d'objets
1	{ A,B }
2	{A,C,D,E}
3	{B,C,D,F}
4	{A,B,C,D}

On peut extraire par exemple la règle suivante : {B, C} →{D}, c'est à dire que si les individus ont $\{B, C\}$ alors ils ont aussi l'objet $\{D\}$

Les méthodes utilisées consistent en l'identification des motifs les plus fréquents qui sont appelés des règles d'association. Elles sont composées de deux parties, la première est appelée l'antécédent, c'est une combinaison d'objets trouvée dans les données. Et la deuxième est la conséquence qui est une combinaison de l'antécédent et d'un ou plusieurs autres objets. Deux notions permettent de définir la pertinence d'une règle :

- Le **Support** d'une règle (R) noté sup(R) est un indicateur de fiabilité et correspond au nombre de fois où la règle est observée dans l'échantillon.
- La **confiance** d'une règle (R) notée conf(R)=$\frac{sup(R)}{sup(\text{antécédent de R})}$ est un indicateur de précision et correspond au nombre de fois où la règle est observée chez les sujets présentant l'antécédent de la règle.

Si on reprend l'exemple précédent :

Identifiant	Combinaison d'objets
1	{ A,B }
2	{A,C,D,E}
3	{B,C,D,F}
4	{A,B,C,D}
5	{A,B,C,F}

Si on s'intéresse à la règle

R=Si {B, C} alors {D} aussi notée R={B, C} → {D}

On obtient :

- Sup(R)=2 car deux sujets présentent cette règle. On peut aussi l'exprimer en termes relatifs, dans ce cas, Sup(R)=$\frac{2}{5}$=0,4 soit 40%.
- sup({B, C})=3 alors conf(R)=$\frac{2}{3}$

Une "bonne" règle est une règle avec un support et une confiance élevés. Pour qu'une règle soit sélectionnée, elle doit vérifier deux conditions, le support et la confiance de la règle doivent être supérieurs aux supports et la confiance minimum, qui sont fixés par l'utilisateur. La construction des règles se fait en deux temps, d'abord repérer les combinaisons fréquentes de plusieurs objets ou éléments vérifiant les deux conditions puis construire les règles à partir de ces combinaisons fréquentes. Pour cela, différents algorithmes ont été développés.

L'algorithme *Apriori* qui construit les combinaisons fréquentes de façon itérative. Chaque itération est composée de deux étapes : i) la jointure qui consiste à construire l'ensemble des combinaisons d'objets dont la taille augmente à chaque itération ; ii) l'élagage qui ne sélectionne que les combinaisons fréquentes. Dans l'exemple présenté dans la Figure 5.2, nous avons choisi de ne sélectionner que les combinaisons d'objets ayant un support minimum de 0,5.

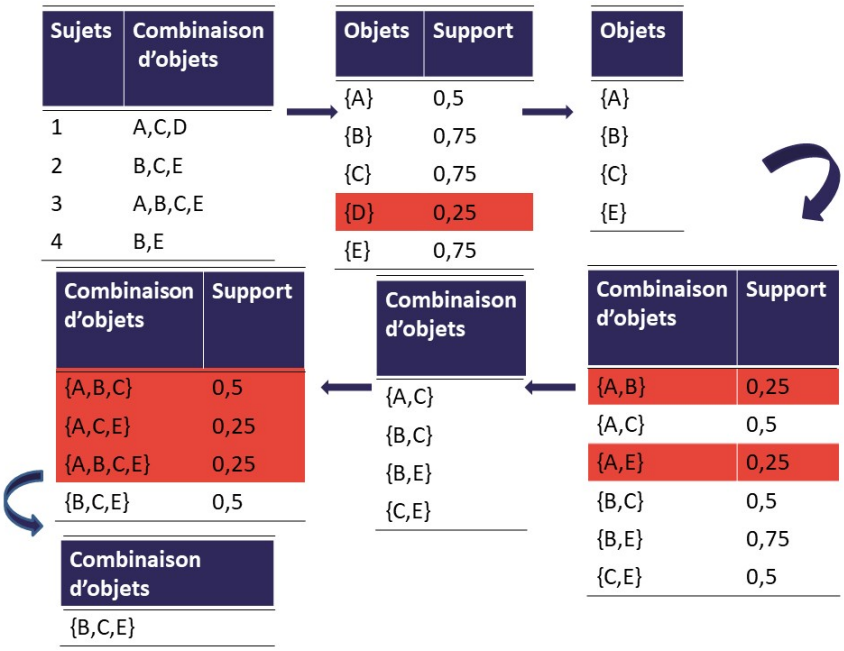
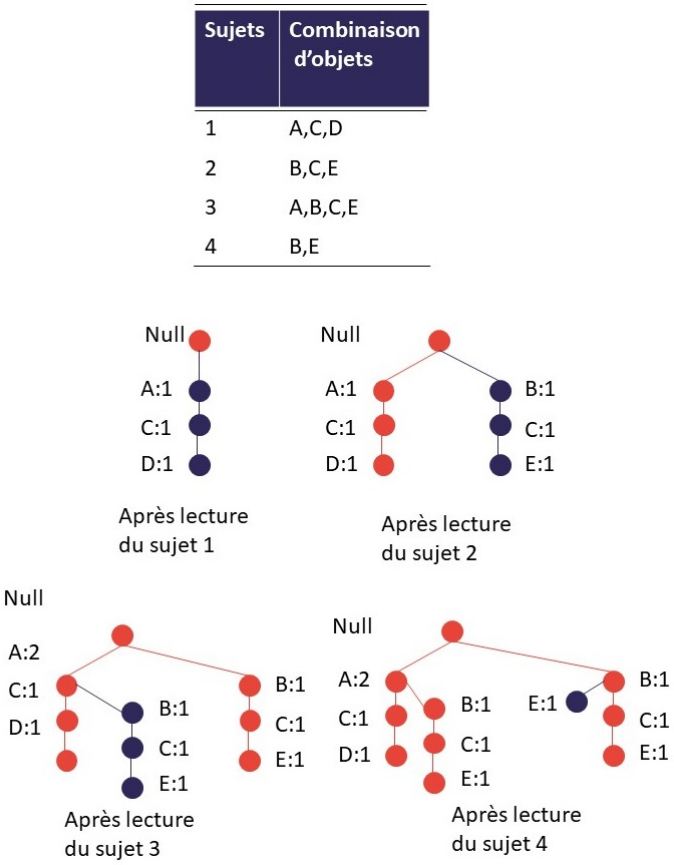


FIGURE 5.2 – Exemple pour l'algorithme Apriori

Les lignes surlignées en rouge sont les lignes ayant un support inférieur au support minimum (égal à 0,5) qui ne sont pas gardées pour l'itération suivante.



L'algorithme *Frequent Pattern-Growth* (FP-Growth) a pour but de construire un arbre à motifs fréquents. Chaque nœud de l'arbre représente un objet. Le nœud racine représente "null". Chaque branche représente alors une combinaison d'objets fréquents (Figure 5.3). L'algorithme consiste d'abord à calculer le support de chaque objet et à ne garder que les objets fréquents, c'est-à-dire, pour chaque individu, les objets non fréquents sont enlevés de sa combinaison d'objets. La construction de la structure de l'arbre commence par la création de l'élément "Racine" de l'arbre. Ensuite, une branche est créée pour chaque individu, mais les individus ayant le même début de combinaison d'objets partageront le même début de branche.

FIGURE 5.3 – Exemple pour l'algorithme FP-Growth

L'algorithme *Eclat* est un algorithme récursif qui consiste à trouver les combinaisons d'objets

fréquents par intersection de combinaisons d'objets (Figure 5.4). Il se décompose en plusieurs étapes :

1. Il ne garde que les objets ayant le support minimum (choisi par l'utilisateur).
2. Il les combine pour faire des combinaisons de deux objets et ne garde que les combinaisons ayant un support supérieur au support minimum.
3. Il regarde les intersections des combinaisons deux à deux, trouvées à l'étape précédente, et ne garde que les combinaisons de trois objets ayant un support supérieur au support minimum.
4. Ces trois étapes sont répétées jusqu'à ne plus pouvoir faire d'intersection.

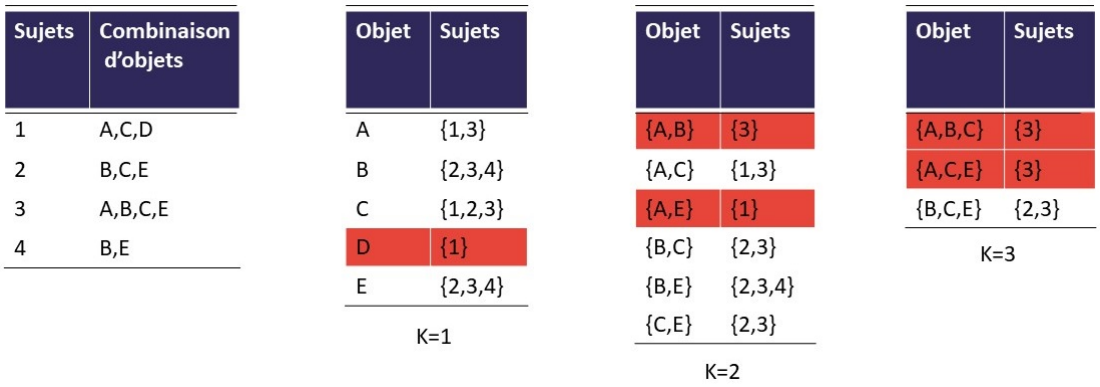


FIGURE 5.4 – Exemple pour l'algorithme Eclat

Les lignes surlignées en rouge sont les lignes ayant un support inférieur au support minimum (égal à 0.5) qui ne sont pas gardées pour l'itération suivante.

Les algorithmes *Apriori* et *FRP-growth* utilisent des données au format horizontal alors que l'algorithme *Eclat* utilise les données au format vertical.

Les méthodes d'analyse de séquence sont des extensions des méthodes de règles d'associations en prenant en compte le coté temporel. Elles sont utilisées pour identifier les motifs qui se produisent fréquemment au cours d'une certaine période. Une séquence est une liste ordonnée d'objets, par exemple :

Identifiant	Combinaison d'objets
1	{A,B}
2	{A,C,D,E}
3	{B,C,D,F}
4	{A,B,C,D}
5	{A,B,C,F}

Motif	Support
A	4
A,B	3
A,B,C	2
A,B,C,D	1
A,B,C,F	1
A,C	3

Le premier tableau représente un exemple de base de séquences. Le deuxième contient dans la première colonne un sous-ensemble des différents motifs contenus dans la base. La deuxième colonne représente le nombre de fois où le pattern est retrouvé dans notre base, encore appelé support. Par exemple, le motif "A,B" est retrouvé chez les patients 1,4,5.

5.2.2 Les méthodes de classifications supervisées

La classification supervisée consiste à affecter un nouvel individu à une classe prédéfinie en fonction de ses caractéristiques. Les méthodes de classifications supervisées cherchent à définir des règles permettant de classer des individus dans des groupes prédéfinis à l'avance à partir de variables quantitatives ou qualitatives. Elles font parties des méthodes dites de Machine Learning. Quelle que soit la méthode de classification appliquée, la méthodologie reste la même. On applique la méthode sur un échantillon de départ, appelé échantillon d'apprentissage pour l'apprentissage des règles. Une fois les règles définies, on les valide sur un échantillon test afin de les évaluer. De nombreuses méthodes ont été dévelopées.

Arbre de décision

Le principe repose sur la construction d'un arbre de façon récursive par partitionnement selon des règles sur les variables explicatives. Pour cela, il faut définir une suite de noeuds permettant de faire une partition des objets en 2 groupes sur la base d'une des variables explicatives (Figure 5.5). La sélection des meilleurs noeuds se fait selon le critère de sélection choisi (par exemple l'entropie, critère de Gini [57],...). Ensuite, vient l'étape de découpage de l'arbre par la définition du noeud terminal qui correspondra à une classe. Lorsque le nombre de noeuds est trop grand, l'étape d'élagage de l'arbre devient nécessaire. Elle consiste à sélectionner un sous-arbre optimal à partir de l'arbre maximal. Par extension, les forêts aléatoires consistent à construire des arbres de décision sur des sous-échantillons afin de construire un arbre "moyen".

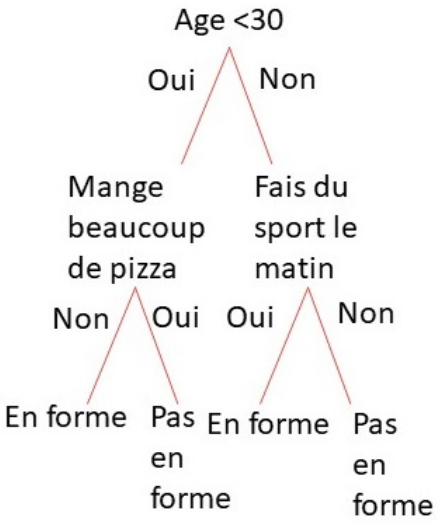


FIGURE 5.5 – Exemple d'arbre de décision

Réseau de neurones

Cette méthode est caractérisée par des signaux d'entrée (variables explicatives) et une fonction d'activation (f) souvent sigmoïdale. Les neurones sont ensuite associés en couche. Une couche d'entrée lit les signaux entrant, un neurone par entrée, une couche en sortie fournit la réponse du système. Un neurone d'une couche cachée est connecté en entrée à chacun des neurones de la couche précédente et en sortie à chaque neurone de la couche suivante [81] (Figure 5.6).

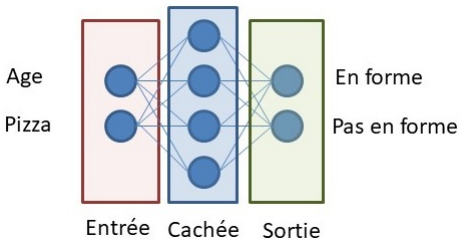


FIGURE 5.6 – Exemple de réseau de neurones

Machine à support vectoriel

Le principe de cette méthode est de se placer dans un espace de plus grande dimension dans lequel se trouve un hyperplan qui sépare au mieux nos classes (Figure 5.7). L'hyperplan optimal est celui qui maximise la marge entre l'échantillon et l'hyperplan séparateur.

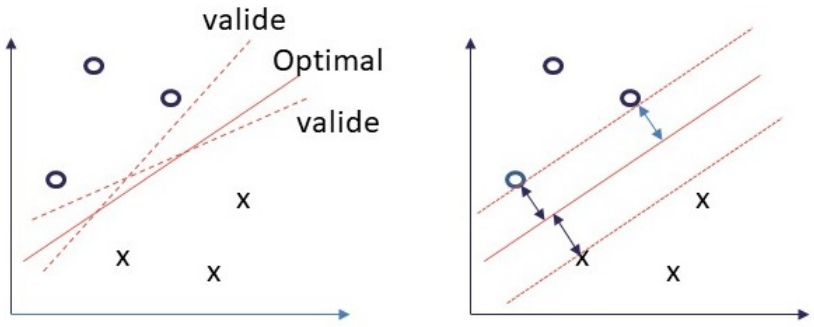


FIGURE 5.7 – Exemple de machine à support vectoriel

Méthodes de régression

De nombreuses méthodes reposent sur les régressions comme par exemple la régression logistique, mais aussi les régressions multiples, Ridge [82], Lasso [83], moindres carrés partiels (PLS)...Certaines de ces méthodes (Ridge [82] et Lasso [83]) permettent de faire de la sélection de variables.

5.2.3 Les méthodes de classifications non supervisées

La classification non supervisée consiste à segmenter une population hétérogène en groupes d'individus homogènes sans *a priori* et seulement basée sur les caractéristiques des individus. L'objectif est de regrouper les individus en groupes homogènes sans *a priori* sur les classes. C'est une forme de classification comme précédemment décrite mais elle est dite non supervisée. La différence entre la classification supervisée et non supervisée est que dans la classification supervisée, les intitulés des groupes sont prédéfinis alors que dans la classification non supervisée les intitulés des groupes ne sont pas prédéfinis à l'avance. Si on prend l'exemple de la Figure 5.8, pour la classification supervisée, on va demander à l'algorithme de classer les objets dans les classes prédéfinies telles que rond, ovale, carré et rectangle en fonction

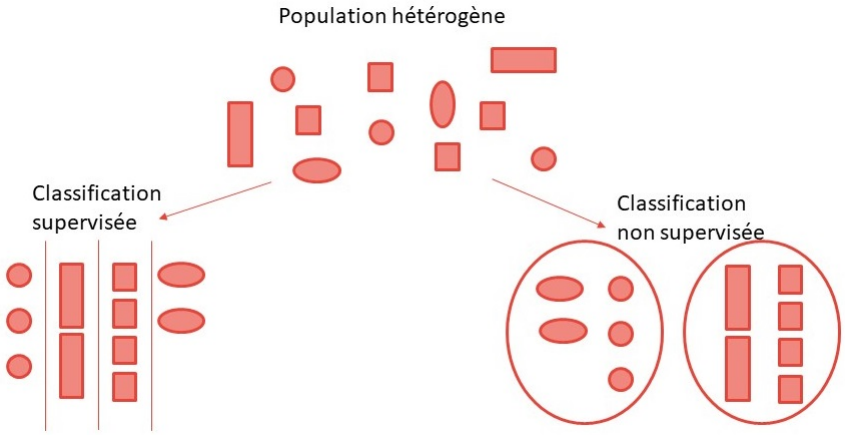


FIGURE 5.8 – Différence entre la classification supervisée et non supervisée

des caractéristiques des objets. On peut ensuite vérifier que les objets sont bien classés par l'algorithme. A l'opposé, pour la classification non supervisée, on va laisser l'algorithme choisir les objets composant les différents groupes, ainsi l'algorithme peut choisir de mettre les objets en fonction de leur caractéristique dans un même groupe. C'est en décrivant les caractéristiques des objets à l'intérieur du groupe qu'on peut lui attribuer un nom. Ainsi dans l'exemple, on voit que l'algorithme a classé les formes circulaires dans un groupe et les formes rectangulaires dans un autre.

Les algorithmes de clustering se basent sur des mesures de similarité ou de dissimilarité entre deux observations. Différentes méthodes existent, les méthodes non paramétriques regroupant les algorithmes hiérarchiques et les algorithmes à partitionnement, et les algorithmes paramétriques.

Les algorithmes hiérarchiques

La classification Ascendante Hiérarchique (CAH) [84] est un algorithme déterministe qui part d'un état où il y a n groupes, chacun étant une observation d'un échantillon pour arriver à un état où il n'y a qu'un seul groupe, l'échantillon lui même. L'algorithme regroupe au fur et à mesure les deux groupes les plus proches jusqu'à n'en former qu'un seul. Toutes les méthodes de CAH peuvent se transposer à la classification descendante hiérarchique, sauf que l'algorithme part d'un état où il n'y a qu'un seul groupe, l'échantillon lui même, pour arriver à l'état où chaque observation représente un groupe.

Les algorithmes de partitionnement

Les algorithmes à partitionnement centroïde, le plus connu étant les K-moyennes [85], consistent à partitionner l'échantillon en un nombre prédéfini k de groupes où chaque observation appartient à un et un seul groupe. Ce sont des méthodes itératives. Dans un premier temps, les observations sont séparées en k groupes de façon aléatoire puis redéfinies en attribuant chaque observation au groupe le plus proche.

Les algorithmes paramétriques

Les algorithmes paramétriques se basent sur l'écriture de modèles dont peuvent être issues les observations de l'échantillon. Les modèles de mélange [86] fini considèrent que les observations forment des groupes chacun ayant une distribution de probabilité différente. Les distributions peuvent ne pas appartenir à la même famille, ou appartenir à la même famille mais différer dans les valeurs des paramètres. Chaque élément a alors une probabilité d'appartenir à chaque groupe.

5.3 L'analyse formelle de concept

L'analyse formelle de concept (FCA) fait partie de la théorie des treillis (lattice) appliquée dont l'objectif est d'étudier comment les objets peuvent être groupés de façon hiérarchique en fonction de leur attributs. Elle a été introduite au début des années 1980 par Rudolf Wille qui utilise l'interprétation philosophique du concept comme une unité de connaissance comprenant un ensemble d'objets et un ensemble d'attributs communs. Elle est devenue une technique populaire dans de nombreux domaines de recherche d'information notamment dans l'apprentissage automatique ou encore machine learning, la découverte des connaissances, la fouille de texte ou text mining ou encore la fouille de données. En effet, la FCA est étroitement liée aux règles d'associations de la fouille de données puisque de nombreux algorithmes sont basés sur cette technique.

5.3.1 Définitions et mesures de la FCA

Définition 1 : Un contexte formel F est un triplet noté $F = (O, A, R)$ où O représente l'ensemble des objets, A l'ensemble des attributs et R une relation binaire entre O et A. En d'autres termes, R signifie que l'objet o possède l'attribut a, où $o \in O$ et $a \in A$. F peut être vu comme une table reliant des objets et leurs attributs.

Définition 2 : Pour un contexte formel $F = (O, A, R)$, on définit $^{\uparrow}$ et $^{\downarrow}$ deux opérateurs de formation de concept définis par :

$^{\uparrow} : 2^O \to 2^A$ tel que $\forall X \subseteq O$, $X^{\uparrow} = \{a \in A | \forall o \in X : (o, a) \in R\}$, $X^{\uparrow}$ correspond à l'ensemble des attributs communs à tous les objets de X.

$^{\downarrow} : 2^A \to 2^O$ tel que $\forall Y \subseteq A$, $Y^{\downarrow} = \{o \in O | \forall a \in Y : (o, a) \in R\}$, $Y^{\downarrow}$ correspond à l'ensemble des objets ayant tous les attributs de Y

Définition 3 : Un concept formel de F est une paire (X, Y) avec $X \subseteq O$ et $Y \subseteq A$ telle que $X^{\uparrow} = Y$ et $Y^{\downarrow} = X$. Ainsi, (X, Y) est un concept formel si et seulement si l'ensemble X contient uniquement les objets qui présentent tous les attributs de Y, et l'ensemble Y contient tous les attributs communs à tous les objets de X. Alors, X et Y sont respectivement appelés l'intention et l'extension du concept.

Définition 4 : Soit deux concepts formels (X_1, Y_1) et (X_2, Y_2) d'un contexte formel F alors $(X_1, Y_1) \leq (X_2, Y_2)$ si et seulement si $X_1 \subseteq X_2$ et $Y_2 \subseteq Y_1$. La relation $\leq$ est définie comme étant la relation d'ordre partiel entre deux concepts.

Définition 5 : Un treillis de concept d'un contexte formel F est une structure $\langle Y(O, A, R), \leq \rangle$ où $Y(O, A, R)$ est une collection de tous les concepts formels, $Y(O, A, R) = \{(X, Y) \in 2^O \times 2^A | X^{\uparrow} = Y, Y^{\downarrow} = X\}$ et $\leq$ est la relation d'ordre partiel.

TABLE 5.1 – Exemple d'un concept formel

Patients	DP1	DP2	DP3	DP4
1	x	x	x	
2		x	x	x
3			x	x
4	x		x	

DP=diagnostic principal

Un exemple d'un contexte formel est présenté dans la table 5.1, les objets sont représentés par les patients et les attributs sont représentés par les diagnostics principaux (DP). La relation R de cet exemple indique que le patient 1 a présenté les diagnostics principaux DP1 et DP2 alors que le patient 2 a présenté les diagnostics 2, 3 et 4.
(2,3),(DP3,DP4) est un concept du contexte formel présenté dans cet exemple.

Un treillis de concept peut être représenté par un diagramme (Figure 5.9), qui peut s'avérer très intéressant dans le domaine de la fouille de données afin de mieux comprendre les relations entre les données. Cependant, le nombre de concepts formels augmente lorsque la taille du contexte formel augmente. C'est pour cela que des mesures ont été proposées afin de réduire la complexité du treillis de concept. Nous nous intéresserons seulement au support d'un concept (X, Y) défini par le nombre d'objets dans son extension : $support(X, Y) = |X|$.

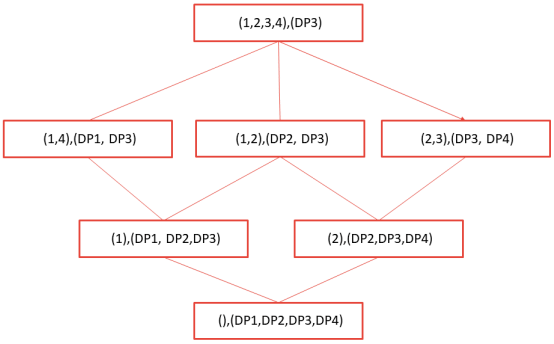


FIGURE 5.9 – Exemple d'illustration d'un treillis de concept

Chapitre 6

Trajectoires de soins de patients après la chirurgie bariatrique

L'obésité est un enjeu majeur de santé publique et la chirurgie bariatrique connaît un essor important ces dernières années. De nombreuses études se sont intéressées aux complications à la suite d'une chirurgie bariatrique mais aucune ne s'est intéressée aux trajectoires de soins de patients. Nous nous sommes donc penchés sur les parcours de soins des patients après une chirurgie bariatrique. Pour cela nous avons utilisé les données issues du PMSI et plus précisément les raisons de réhospitalisations des patients, décrites par les diagnostics principaux.

6.1 L'obésité, la chirurgie bariatrique et l'extraction des données

L'obésité est définie par un indice de masse corporelle (IMC) supérieur à 30 kg/m^2. C'est un problème de santé publique majeur dont la prévalence a considérablement augmenté au cours des 25 dernières années. En effet, le nombre de personnes obèses a presque triplé depuis 1975, avec plus de 1,9 milliard d'adultes en surpoids et environ 650 millions d'obèses dans le monde [87]. L'obésité est connue pour être un facteur de risque majeur pour certaines maladies chroniques telles que les maladies cardiovasculaires, le diabète de type 2, la dyslipidémie, le syndrome d'apnées obstructives du sommeil, les troubles musculo-squelettiques, en particulier l'arthrose et certains cancers (endomètre, sein, ovaire, prostate, foie, vésicule biliaire, rein et côlon). Des études ont montré que les thérapies comportementales et pharmacologiques ont une efficacité limitée à long terme sur la perte de poids chez les patients souffrant d'obésité sévère [88]. La chirurgie bariatrique est devenue un traitement complémentaire fiable pour l'obésité. Au cours des deux dernières décennies, un développement et une amélioration majeurs des procédures bariatriques ont été observés, avec environ 500 000 procédures effectuées chaque année dans le monde depuis 2013 (3). Cependant, certaines conditions doivent être remplies pour pouvoir prétendre à une chirurgie bariatrique (HAS 2009) :

— Patients avec un IMC $\geq$ 40 kg/m^2 ou bien avec un IMC $\geq$ 35 kg/m^2 associé à au moins

une comorbidité susceptible d'être améliorée après la chirurgie (notamment hypertension artérielle, syndrome d'apnées hypopnées obstructives du sommeil (SAHOS) et autres troubles respiratoires sévères, désordres métaboliques sévères, en particulier diabète de type 2, maladies ostéo-articulaires invalidantes, stéatohépatite non alcoolique)
— En deuxième intention après échec d'un traitement médical, nutritionnel, diététique et psychothérapeutique bien conduit pendant 6-12 mois.
— En l'absence de perte de poids suffisante ou en l'absence de maintien de la perte de poids.
— Patients bien informés au préalable, ayant bénéficié d'une évaluation et d'une prise en charge préopératoires pluridisciplinaires.
— Patients ayant compris et accepté la nécessité d'un suivi médical et chirurgical à long terme.
— Risque opératoire acceptable.

La décision finale se fait par décision collégiale, après concertation pluridisciplinaire.

Les différents types de chirurgie bariatrique

Plusieurs techniques ont été développées et peuvent être classées en techniques restrictives, malabsorptives ou mixtes. En France, l'anneau gastrique ajustable (AGA), le bypass (GB) et la sleeve gastrectomie (SG) sont les techniques les plus utilisées.

L'anneau gastrique ajustable (Figure 6.1, figure issue de "https ://www.sleeve-gastrique-en-tunisie.com/anneau-gastrique-en-tunisie-html") est un anneau réglable en silicone posé autour de la jonction entre l'oesophage et l'estomac permettant de ralentir le passage des aliments. Le principe est analogue à celui du sablier. L'anneau permet donc de créer une petite poche et les aliments vont s'écouler doucement. L'anneau est relié à un petit boîtier sous-cutané permettant de serrer ou desserrer l'anneau par l'injection d'un liquide à travers la peau. Cette technique à l'avantage de ne pas être définitive.

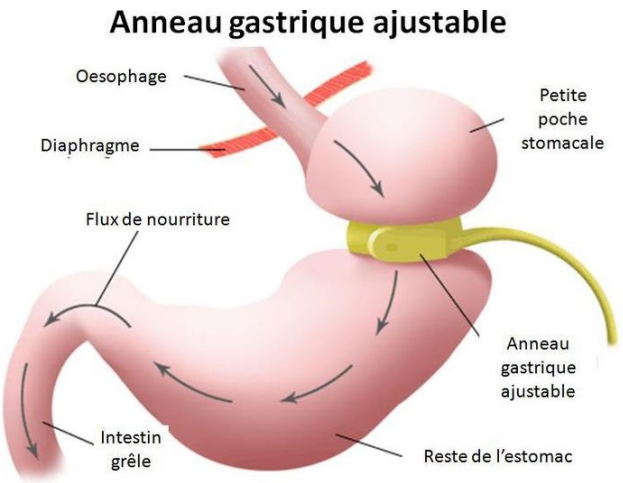


FIGURE 6.1 – L'anneau gastrique

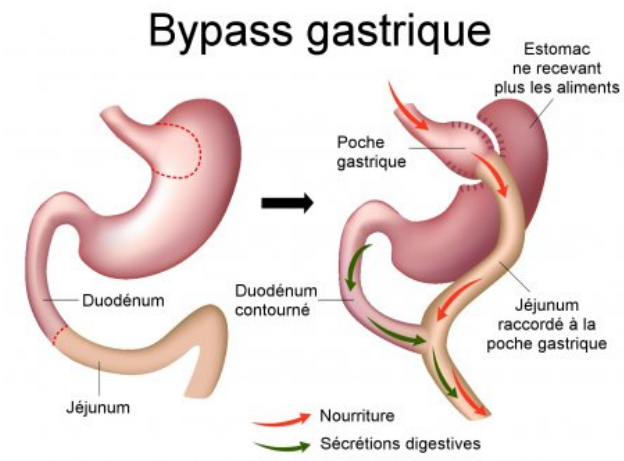


FIGURE 6.2 – Le bypass gastrique

Le bypass (Figure 6.2, figure issue "https ://www.copaix.fr/techniques-chirurgie-obesite/bypass-gastrique-chirurgie-obesite-aix-en-provence.html") fait partie des techniques mixtes et consiste à créer un court-circuit de l'estomac en reliant le bas de l'oesophage et une petite partie du pôle supérieur de l'estomac directement au jéjunum, anse de l'intestin grêle. Cette technique, comme l'anneau, ralentit le passage des aliments, mais diminue aussi l'appétit, provoque une malabsorption ainsi qu'un dumping syndrome (sensation de malaise en cas d'absorption d'aliments très sucrés en trop grande quantité). Contrairement à l'anneau ajustable, le bypass est une opération irréversible.

La sleeve (Figure 6.3) fait partie des techniques dites restrictives et consiste à retirer 2/3 de l'estomac qui est donc réduit à un tube vertical. La partie enlevée contient les cellules sécrétant la ghréline, hormone qui stimule l'appétit. Par conséquent, la sleeve diminue la sensation de faim et les aliments passent rapidement à l'intestin sans que cela ne perturbe la digestion. Elle modifie aussi la flore bactérienne de l'estomac, provoquant un changement de la perception des goûts. Tout comme le bypass, cette technique est irréversible.

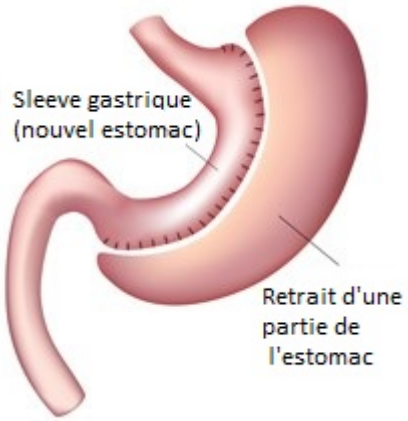


FIGURE 6.3 – La sleeve

Extraction des données

Nous avons inclus tous les patients adultes (âgés de 18 ans et plus et de moins de 60 ans) présentant un diagnostic principal d'obésité et ayant subi l'une des procédures bariatriques les plus courantes pratiquées en France : AGA, GB ou la SG. Nous avons extrait les informations de la base de données PMSI du 1er janvier 2013 au 31 décembre 2018 et avons inclus les patients opérés avant 2018 afin de disposer d'un an de suivi potentiel pour chaque patient. Les patients ont été sélectionnés en respectant les codes de la CCAM pour les procédures bariatriques et le code d'obésité de la CIM-10. Les codes HFMA009 et HFMC007 ont été utilisés pour identifier les AGA, les codes HFCA001 et HFCC003 pour les GB et les codes HFFA011 et HFFC018 pour les SG. Les codes E66x ont été utilisés pour identifier l'obésité. Nous avons exclu les patients dont les procédures étaient codées de manière ambiguë et ceux présentant une erreur de codage ou une valeur manquante dans les informations de diagnostic principales.

Description des données

Nous avons identifié 240 821 chirurgies bariatriques entre 2013 et 2017. Parmi elles, 42 432 (17,6%) ne répondaient pas aux critères d'inclusion et ont été exclues : 13 012 (5,4%) ont été pratiquées sur des patients adolescents ou adultes, 22 352 (9,3%) ont présenté une erreur de codage sur le diagnostic principal ou sur le type d'opération ou ont présenté une ambiguïté dans le codage, 3 905 (1,6%) étaient des procédures bariatriques autres que AGA, GB ou SG et 3 163 (1,3%) n'avaient pas de code de diagnostic principal d'obésité. Ainsi, 198 389 procédures bariatriques effectuées sur 196 323 patients ont été incluses dans l'analyse. Parmi les interventions chirurgicales, 13 744 (6,9%) étaient des AGA, 55 945 (28,2%) étaient des GB et 128 700 (64,9%) étaient des SG. La plupart des patients (90%) ont subi une seule intervention chirurgicale. Les patients sont principalement des femmes (80,7%) atteintes d'obésité morbide (70,0% avec un IMC supérieur à 40 kg / m^2). L'indice de comorbidité de Charlson était nul pour 85,4% des patients. Enfin, 30,7% avaient un syndrome d'apnées obstructives du sommeil (Table 6.1). L'AGA était principalement utilisé chez les patients plus jeunes (62,6% âgés de 18 à 39 ans) et avec un IMC inférieur à 40 kg / m^2 (48,0%). Une GB et une SG ont été réalisées chez des patients âgés (42,3% et 51,2% d'entre eux âgés respectivement de 18 et 39 ans) atteints d'obésité morbide (respectivement 73,9% et 70,3% avaient un IMC supérieur à 40 kg / m^2) (Table 6.1).

Table 6.1 – Description de la population

Covariable		Population totale (N=198389)	AGA (N=13744)	GB (N=55945)	SG (N=128700)
Sexe	Female	160178 (80,7%)	11663 (84,9%)	46143 (82,5%)	102372 (79,5%)
Age (années)	18-29	42007 (21,2%)	4542 (33,1%)	8730 (15,6%)	28735 (22,3%)
	30-39	56092 (28,3%)	4067 (29,6%)	14922 (26,7%)	37103 (28,8%)
	40-49	57661 (29,1%)	3277 (23,8%)	17801 (31,8%)	36583 (28,4%)
	50-60	42629 (21,5%)	1858 (13,52%)	14492 (25,9%)	26279 (20,4%)
IMC (kg/m^2)	30-40	55151 (30,0%)	6290 (48,0%)	13627 (26,1%)	35234 (29,7%)
	40-50	109025 (59,3%)	6169 (47,0%)	32601 (62,5%)	70255 (59,2%)
	>=50	19815 (10,8%)	654 (5,0%)	5956 (11,4%)	13205 (11,1%)
	NA	14398	631	3761	10006
Index de comorbidités de Charlson	0	169492 (85,4%)	12760 (92,8%)	46416 (83,0%)	110316 (85,7%)
	1-2	25281 (12,7%)	860 (6,3%)	8099 (14,5%)	16322 (12,7%)
	>=3	3616 (1,8%)	124 (0,9%)	1430 (2,6%)	2062 (1,6%)
Syndrome d'apnée du sommeil		60832 (30,7%)	1882 (13,7%)	18517 (33,1%)	40433 (31,4%)
Approche chirurgicale	laparoscopie	197266 (99,4%)	13701 (99,7%)	55 587 (99,4%)	127978 (99,4%)
	ouvert	1123 (0,6%)	43 (0,3%)	358 (0,6%)	722 (0,6%)
Hospital ownership	Private, for profit	120662 (61,3%)	11121 (81,6%)	32921 (59,6%)	76620 (59,9%)
	Private, nonprofit	10190 (5,2%)	342 (2,5%)	2701 (4,9%)	7147 (5,6%)
	Public	65943 (33,5%)	2160 (15,9%)	19603 (35,5%)	44180 (34,5%)
	NA	1594	121	720	753
Durée de séjours (jours)	0-1	11055 (5,6%)	5758 (41,9%)	1718 (3,1%)	3 579 (2,8%)
	2-7	176470 (89,0%)	7965 (58,0%)	50215 (89,8%)	118290 (91,9%)
	>7	10864 (5,5%)	21 (0,2%)	4012 (7,2%)	6831 (5,3%)
Année de la chirurgie	2013	35541 (17,9%)	4614 (33,6%)	10513 (18,8%)	20 414 (15,9%)
	2014	39339 (19,8%)	3678 (26,8%)	11286 (20,2%)	24375 (18,9%)
	2015	39965 (20,1%)	2373 (17,3%)	11360 (20,3%)	26232 (20,4%)
	2016	43184 (21,8%)	1880 (13,7%)	11849 (21,2%)	29455 (22,9%)
	2017	40360 (20,3%)	1199 (8,7%)	10937 (19,6%)	28224 (21,9%)
Réopération	Oui	12839 (6,5 %)	74 (0,5 %)	6414 (11,5 %)	6351 (4,9 %)

6.2 Article "Analyse of trajectories of care after bariatric surgery using data-mining method and health administrative information systems."

Obesity Surgery
https://doi.org/10.1007/s11695-020-04430-6

ORIGINAL CONTRIBUTIONS

Analysis of Trajectories of Care After Bariatric Surgery Using Data Mining Method and Health Administrative Information Systems

Anaïs Charles-Nelson[1,2,3,4] · Andrea Lazzati[3,5] · Sandrine Katsahian[2,3,4]

Abstract

Context The 30-day readmission rate after bariatric surgery is considered an important metric of the quality of hospital care. However, readmission rate beyond 30 days is rarely reported and does not provide any information about trajectories of care which would be of great interest for healthcare planning. The aim of this study was to analyze trajectories of care during the first year after bariatric surgery on a nationwide basis using data mining methods.

Method This was a retrospective descriptive study on the trajectories of care within the first year after bariatric surgery. Data were extracted from a national administrative claims database (the PMSI database) and trajectories were defined as principal diagnosis of successive readmissions. Formal Concept Analysis was performed to find common concepts of trajectories of care.

Results We included for analysis 198,389 bariatric procedures performed on 196,323 patients. Twelve main concepts were selected. About one third of patients (32.4%) were readmitted in the first year after surgery. Most common trajectories were as follows: regular follow-up (14.9%), cholelithiasis (2.2%), abdominal pain (1.9%), and abdominal sepsis (1.3%). Important differences were found in trajectories among different bariatric procedures: 1.8% of gastric banding patients had pregnancy-related events (delivery or medical abortion), while we observed a readmission rate for abdominal sepsis in 2.7% and 5.1% of patients operated of gastric bypass and sleeve gastrectomy respectively.

Conclusion Administrative claim data can be analyzed through Formal Concept Analysis in order to classify trajectories of care. This approach permits to quantify expected postoperative complications and to identify unexpected events.

Keywords Data mining · Formal Concept Analysis · Claim data · Trajectory of care · Bariatric surgery

✉ Anaïs Charles-Nelson
anais.charles.nelson@gmail.com

Andrea Lazzati
andrea.Lazzati@chicreteil.fr

Sandrine Katsahian
sandrine.katsahian@aphp.fr

[1] Sorbonne Universités, UPMC Univ Paris 06, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France

[2] INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France

[3] Université Paris Descartes, Sorbonne Paris Cité, UMRS 1138, Centre de Recherche des Cordeliers, F75006 Paris, France

[4] Assistance Publique Hôpitaux de Paris, Hôpital européen Georges Pompidou, Unité d'Épidémiologie et de Recherche Clinique, INSERM, Centre d'Investigation Clinique1418, module Épidémiologie Clinique, HEGP, Paris, France

[5] Department of General Surgery, Centre Hospitalier Intercommunal de Créteil, 40 avenue de Verdun, 94000 Créteil, France

Introduction

Obesity is a major public health problem whose prevalence has dramatically increased in the last 25 years. Indeed, the number of obese patients has almost tripled since 1975 with more than 1.9 billion overweighed adults and about 650 million of obese [1]. Obesity is known for being a major risk factor for some chronic diseases such as cardiovascular disease, type 2 diabetes, dyslipidemia, obstructive sleep apnea syndrome, musculoskeletal disorders, especially osteoarthritis, and some cancers (endometrium, breast, ovarian, prostate, liver, gallbladder, kidney, and colon). Studies have shown that behavioral and pharmacological therapies have a limited long-term efficacy on weight loss for patients with severe obesity [2]. Bariatric surgery has become a reliable complementary treatment for obesity. In the last two decades, a major development and improvement in bariatric procedures have been observed with about 500,000 procedures per year performed worldwide since 2013 [3].

Published online: 06 February 2020

It is accepted that bariatric surgery has a rate of early complications between 5% and 10% [4, 5].

Several factors have been associated to the risk of major postoperative complications: the type of procedure [6], surgical skills, and presence of chronic health problems [7, 8].

The 30-day readmission rate is considered as an important metric of the quality of hospital care. This indicator is of clinical interest for the patient and economical for the institutions. Thus, hospital readmission rates have been widely studied [9–11]. However, readmission rate beyond 30 days are less frequently reported. Doumouras et al. reported a readmission rate of 6.1% over a period of 3 years [10], and Saunders et al. reported 1-year readmission rate of 18% [12]. However, readmission rate provides no information about trajectories of care even though they are of great interest for healthcare planning. Indeed, understanding trajectories of care may help the physician to anticipate future care and, thus, may potentially anticipate severe complications, leading to a more personalized medicine. Trajectories of care have been studied in different domain such as breast cancer [13] and prenatal care [14]. In the field of obesity, however, only medical weight loss trajectories have been studied [15–17], but no study analyzed trajectories of care after bariatric surgery. Thus, as readmission is considered as a metric of quality of hospital care, a better understanding of trajectories may participate to enhance the quality itself.

In the last few years, several studies have used national registries or administrative databases to monitor their outcomes [18–20]. In France, the national administrative claims database (*Programme de Médicalisation des Systèmes d'Information* PMSI) includes all reimbursed surgical interventions performed in France in every hospital. It was introduced in the mid-1980s and was first presented as an epidemiological tool before being used for budget allocation. It is also allowed the promotion exchange between hospital partners: doctors, nurses, and administrators [21]. The collection and processing of data for the short stay are directly inspired by the American Classification of Diagnosis Related Groups (DRG) [22]. This classification classifies hospital stays in groups with a double homogeneity in medical and economic terms. The main advantage of the PMSI database is that data collected gives comprehensive information from all surgical centers and can be used to follow any single patient across different hospitals. Because of its exhaustiveness, the PMSI database is considered a big dimension dataset. A major challenge concerning big datasets is the identification of pertinent information. In the last few years, data mining methods have been recognized as an efficient tool to treat large volume data [23]. Data mining involves drilling, exploring, or delving into data. Unlike conventional analyses, Data Mining allows to explore associations and relationships between data, which are hidden or not obvious due to large volumes of data. Thus, it provides a summary of big dataset and makes the information usable and understandable for decision-making.

The aim of this study was to analyze trajectories of care within the first year after bariatric surgery on a nationwide basis using data mining methods.

Material and Method

Data Source

This was a retrospective descriptive study on the trajectories of care within the first year after bariatric surgery. Data were extracted from the PMSI database. Data are exhaustive as the database includes all reimbursed surgical procedures performed in France, in any hospital regardless of their academic affiliation or ownership (public and private for-profit and private nonprofit). In the PMSI database, data are collected as standardized discharge reports and include patient demographic data (age, gender, zip code, entry, and release dates); primary and associated diagnoses based on the International Classification of Disease, 10th edition (ICD-10); and therapeutic procedures based on the Common Classification of Medical Acts (Classification Commune des Actes Médicaux, CCAM, 11th edition), which is a national standardized classification of medical procedures [24]. A unique and anonymous identifier is assigned to each patient, thus making it possible to identify all of his or her hospital stays planned or not in any hospital in France. Patient consent is not required as the individual information is anonymous and publicly available.

Participants

We included all adult patients (≥ 18 years and age ≤ 60 years) with a principal diagnosis of obesity who underwent one of the most common bariatric procedure performed in France: adjustable gastric banding (AGB), gastric bypass (GB), or sleeve gastrectomy (SG). We retrieved information from the PMSI database from January 1, 2013, to December 31, 2018, and included patients operated before 2018 in order to have 1 year of potential follow-up for each patient.

Patients were selected matching the CCAM codes for bariatric procedures and the ICD-10 code of obesity. The codes HFMA009 and HFMC007 were used to identify AGB, the codes HFCA001 and HFCC003 for GB, and the codes HFFA011 and HFFC018 for SG. The codes E66x were used to identify obesity.

We excluded patients that had ambiguous procedures coding and patients presenting a coding error or a missing value on the principal diagnosis information.

Outcomes

We were interested in the trajectories of care using the primary diagnosis within 1 year from discharge after bariatric surgery took place. Trajectories are defined as a sequence of hospitalizations starting within 1 year after the bariatric surgery, which is the first element of the trajectory [13] irrespective of the length of the initial stay.

Coding of Variables

Length of stay was categorized into three classes: 0 to 1 day (capturing ambulatory and outpatient surgery patients), 2 to 7 days, and > 7 days. One day of hospital stay means that the patient spent one night at the hospital irrespective of the total number of hours. The cutoff of 7 days was selected according to the American Society for Metabolic and Bariatric Surgery (ASMBS) definition [25] which considered a length of stay greater than 7 days as a major complication. However, in our study, this cutoff was only used to categorize the length of stay into three categories.

Demographic data included age and gender. The body mass index (BMI) is classified into three categories 30 to 40 kg/m^2, from 40 to 50 kg/m^2, > 50 kg/m^2, and patients with BMI unspecified were excluded from the analysis. Comorbidities were assessed using the Charlson Comorbidity Index, using the version of Quan and colleagues [26]. The final score was the categorized into three groups (0, 1–2, and ≥ 3). Obstructive sleep apnea syndrome (OSAS) was included as a separate covariate. Missing data about comorbidities or OSAS at the hospital stay where the bariatric took place were imputed using the last observation carried forward method, as previously reported [27]. Primary diagnoses are coded using the first 3 digits of ICD-10 code with a few exceptions. Firstly, the codes starting with the digit "Z," which indicates the admission for regular follow-up, were grouped in one category. Secondly, the code K316 (fistula of stomach and duodenum) was kept in its full length and grouped with the code K65 (peritonitis) for better consistency. We will refer to this group as "abdominal sepsis." Finally, we also kept four digits for the code K910 (Vomiting following gastrointestinal surgery) as other K91x diagnoses were poorly consistent (Table 1).

Statistical Analyses

Descriptive Analysis

Continuous demographic and clinical characteristics are presented as mean ± one standard deviation if the parameter follows a normal distribution and median with interquartile range if the distribution is not normal for quantitative parameters. Qualitative variables are presented as numbers and proportions.

Formal Concept Analysis

Formal Concept Analysis (FCA) was used to hierarchically group objects according to their common attributes [28]. A group of objects sharing some common set of attributes is called a concept. Thus, the FCA method finds and represents all concepts and dependencies in the tabular input data [29]. Input data are in a Boolean matrix form, called formal context, where each row represents an object, and each column represents one of the defined attributes [29]. It is a matrix fulfilled with 0 and 1, where 1 means that an object presents an attribute. In this study, bariatric surgeries and principal diagnoses (PD) of readmission are objects and attributes respectively. Table 2 shows an example of a Boolean matrix adapted to this study.

FCA method transforms a formal concept into a mathematical structure called concept lattice. It represents all concepts of a formal context with the order relation, representing all possible combinations of diagnoses observed in patient trajectories. Figure 1 a shows a subset of an example of a concept lattice. It is illustrated using a line diagram, representing hierarchical relationship of all the found concepts and a list of all found attributes implications in the formal context [29]. Figure 1 b represents the line diagram of the concept lattice of the example.

Due to the high number of diagnosis resulting in a high number of concepts, they are filtered according to proportion of own objects and following a stepwise method. Due to a lack of consensus, thresholds for proportion of objects to select concepts are arbitrary chosen. The first step selected concepts combining PD of obesity with a second PD. Concepts were selected if it represented at least 0.4% of the number of surgeries. The second step consists in selecting concepts and combining concepts previously selected with another PD. Concepts were selected if it represented at least 0.2% of the number of surgeries. Finally, concepts combining obesity and three other principal diagnoses were selected during the last step. Combinations were selected if they represented at least 0.1% of the number of surgeries.

Concepts observed in trajectories were represented using an organigram where each level corresponds to a different number of combinations of principal diagnosis. The top of the diagram is the diagnosis of obesity. The second level corresponds to diagnosis of obesity combined with one other principal diagnosis. The third level corresponds to combination of obesity with two other principal diagnoses. And the last level corresponds to concept combining obesity with three other PD.

Table 1 Main principal diagnoses codes

Grouped code	Codes	Designation
E66	E660, E6600, E6601, E6602, E6609, E661, E6610, E6611, E6612, E6619, E662, E6620, E6621, E6622, E6629, E668, E6680, E6681, E6682, E6689, E669, E6690, E6691, E6692, E6699, E6603, E6613, E6683, E6693, E6604, E6605, E6606, E6607, E6614, E6615, E6616, E6617, E6624, E6625, E6626, E6627, E6684, E6685, E6686, E6687, E6694, E6695, E6696, E6697	Obesity
I83	I830, I831, I832, I839	Varicose veins of lower extremities
K22	K220, K221, K222, K223, K224, K225, K226, K227, K228 K229	Other diseases of esophagus
K25	K250, K251, K252, K253, K24, K255, K256, K257, K259	Gastric ulcer
K28	K280, K281, K282, K283, K284, K285, K286, K287, K289	Gastrojejunal ulcer
K29	K290, K291, K292, K293, K294, K295, K296, K297, K298, K299	Acute hemorrhagic gastritis
K316-K65	K316, K650, K658, K659	Fistula of stomach and duodenum and Acute peritonitis
K43	K430, K431, K432, K433, K434, K435, K436, K437, K439	Ventral hernia
K56	K560, K561, K562, K563, K564, K565, K566, K567	Paralytic ileus and intestinal obstruction without hernia
K63	K630, K631, K632, K633, K634, K635, K635 + 0, K635 + 8, K638, K639	Other diseases of intestine
K80	K800, K801, K802, K803, K804, K805, K808	Cholelithiasis
K910		Vomiting following gastrointestinal surgery
K92	K920, K921, K922, K928, K929	Other diseases of digestive system
L02	L020, L021, L022, L023, L024, L028, L029	Cutaneous abscess, furuncle, and carbuncle
N20	N200, N201, N202, N209	Calculus of kidney and ureter
N23		Unspecified renal colic
O04	O04, O040, O041, O042, O043, O044, O045, O046, O047, O048, O049, O0400, O0401, O0402, O0403, O0410, O0411, O0412, O0413, O0420, O0421, O0422, O0423, O0430, O0431, O0432, O0433, O0440, O0441, O0442, O0443, O0450, O0451, O0452, O0453, O0460, O0461, O0462, O0463, O0470, O0471, O0472, O0473, O0480, O0481, O0482, O0483, O0490, O0491, O0492, O0493,	Medical abortion
O80	O800, O801, O802, O808, O809	Single spontaneous delivery
R10	R100, R101, R102, R103, R104	Abdominal and pelvic pain
R11		Nausea and vomiting
R13	T858, T859	Dysphagia
T85	T850, T851, T852, T853, T854, T855, T8550, T8558, T856, T857, T858, T859	Complications of other internal prosthetic devices, implants, and grafts
Zx	Z015, Z018, Z048, Z080, Z081, Z090, Z092, Z097, Z098, Z099, Z121, Z131, Z132, Z138, Z301, Z302, Z305, Z308, Z312, Z313, Z349, Z358, Z359, Z361, Z380, Z411, Z420, Z422, Z431, Z432, Z433, Z434, Z438, Z448, Z450, Z452, Z458, Z4580, Z4588, Z459, Z462, Z465, Z466, Z468, Z470, Z480, Z488, Z489, Z491, Z502, Z503, Z508, Z5100, Z5101, Z511, Z512, Z514, Z518, Z5188, Z530, Z538, Z539, Z540, Z547, Z566, Z713, Z718, Z751, Z800, Z850, Z860, Z931, Z940, Z944, Z968, Z978, Z980, Z988, Z991, Z000, Z005, Z006, Z008, Z010, Z013, Z014, Z016, Z017, Z019, Z022, Z028, Z030, Z031, Z033, Z034, Z035, Z036, Z038, Z039, Z041, Z043, Z044, Z045, Z082, Z087, Z088, Z089, Z091, Z093, Z094, Z108, Z110, Z111, Z112, Z118, Z120, Z122, Z125, Z128, Z129, Z130, Z133, Z136, Z202, Z206, Z21, Z258, Z290, Z291, Z300, Z304, Z309, Z310, Z311, Z314, Z318, Z319, Z321, Z33, Z340, Z348, Z350, Z351, Z352, Z353, Z354, Z356, Z357, Z360, Z362, Z363, Z364, Z368, Z381, Z383, Z390, Z391, Z392, Z400, Z408, Z412, Z418, Z4180, Z4188, Z421, Z423, Z424, Z428, Z429, Z430, Z436, Z449, Z451, Z453, Z4581, Z4582, Z4583, Z4584, Z463, Z467, Z478, Z490, Z500, Z501, Z507, Z509, Z513, Z5130, Z5131, Z515, Z516, Z5180, Z520, Z521, Z523, Z524, Z528, Z531, Z532, Z590, Z597, Z599, Z601, Z602, Z608, Z609, Z629, Z630,Z640, Z653, Z659, Z711, Z717, Z720, Z724, Z740, Z741, Z742, Z743, Z750, Z752, Z753, Z754, Z7588,Z762, Z763, Z765, Z768, Z827, Z837, Z848, Z853, Z855, Z858, Z863, Z866, Z867, Z871, Z874, Z875, Z877, Z878, Z880, Z888, Z889, Z900, Z903, Z904, Z905, Z915, Z9150, Z924, Z941, Z942, Z945, Z946, Z947, Z9480, Z94800, Z94801, Z952, Z955, Z958, Z960, Z964, Z967, Z969, Z975, Z991 + 0, Z991 + 1, Z991 + 8, Z992 + 0, Z993, Z998, Z003, Z004, Z023, Z029, Z042, Z04800, Z04801, Z04802, Z04880, Z115, Z201, Z243, Z292, Z298, Z299, Z303, Z355, Z369, Z409, Z441, Z443, Z4920,	Regular follow-up

Table 1 (continued)

Grouped code	Codes	Designation
	Z4921, Z548, Z578, Z596, Z598, Z603, Z634, Z637, Z638, Z641, Z651, Z749, Z760, Z764, Z8000, Z8009, Z804, Z80880, Z809, Z820, Z824, Z8370, Z8379, Z8480, Z8500, Z8509, Z851, Z85802, Z859, Z86000, Z86001, Z86080, Z862, Z865, Z8660, Z8661, Z8671, Z8701, Z8711, Z8712, Z8719, Z872, Z8740, Z881, Z884, Z886, Z901, Z908, Z913, Z929, Z932, Z933, Z934, Z950, Z951, Z025, Z027, Z032, Z040, Z102, Z365, Z435, Z479, Z504, Z529, Z544, Z594, Z632, Z639, Z712, Z714, Z721, Z730, Z769, Z833, Z834, Z8670, Z8710, Z911, Z961, Z962, Z981, Z103, Z124, Z208, Z225, Z3900, Z3908, Z461, Z636, Z658, Z716, Z76880, Z803, Z8520, Z86005, Z861, Z864, Z8709, Z873, Z8741, Z935, Z9488, Z137, Z268, Z320,Z410, Z413, Z437, Z5280, Z5288, Z584, Z633, Z710, Z723, Z735, Z748, Z759, Z8001, Z813, Z8501, Z854, Z857, Z85803, Z86002, Z894, Z899, Z910, Z921, Z94802, Z94803, Z953, Z992, Z223, Z316, Z4780, Z4788, Z560, Z715, Z80800, Z823, Z85805, Z85880, Z8700, Z8781, Z8788, Z902, Z930, Z9580, Z966, Z209, Z595, Z731, Z733, Z8422, Z85800, Z8662, Z8780, Z918, Z94804, Z954, Z9588, Z963, Z4000, Z4001, Z4008, Z505, Z506, Z5912, Z5968, Z922, Z9481, Z994, Z220, Z315, Z4002, Z492, Z526, Z5910, Z5918, Z6020, Z605, Z80805, Z811, Z85804, Z943, Z631, Z755, Z8488, Z907, Z920, Z926, Z94809, Z982, Z464, Z5978, Z728, Z76800, Z86003, Z8742, Z917, Z992 + 1	

Softwares

Trajectories were analyzed using the Coron System (available at http://coron.loria.fr). Other analyses were performed using R version 3.4.4 (R172 Foundation for Statistical Computing, Vienna, Austria). Data are reported according to the Reporting of studies Conducted using Observational Routinely collected Data statement (RECORD statement) [30].

Results

We identified 240,821 bariatric surgeries between 2013 and 2017. Among these 42,432 (17.6%) did not meet inclusion criteria and were excluded: 13,012 (5.4%) were performed on adolescents or elderly patients; 22,352 (9.3%) presented a coding error on the principal diagnosis or the type of surgery or coding ambiguity; 3905 (1.6%) were bariatric procedures other than AGB, GB, or SG; and 3163 (1.3%) did not have a main diagnosis code of obesity. Hence, 198,389 bariatric procedures, performed on 196,323 patients, were included for analysis. Among surgical procedures, 13,744 (6.9%) were AGB, 55,945 (28.2%) were GB, and 128,700 (64.9%) were SG. Most of patients (90%) had only one surgery. Patients are mostly woman (80.7%) with morbid obesity (70.0% with a BMI greater than 40 kg/m^2). The Charlson Comorbidity Index was null for 85.4% of patients. Finally, 30.7% had an obstructive sleep apnea syndrome (Table 3). AGB was mostly used on younger patients (62.6% aged between 18 and 39 years) and with a BMI lower than 40 kg/m^2 (48.0%). GB and SG were performed on older patients (42.3% and 51.2% are aged between 18 and 39 years, respectively) with morbid obesity (73.9% and 70.3% had a BMI greater than 40 kg/m^2, respectively) (Table 3).

Table 2 Example of Boolean matrix in the event of a bariatric surgery

Surgery	PD*1	PD2	PD3
1	0	0	1
2	1	0	0
3	0	1	0
4	0	1	0
5	1	0	1
6	1	1	1

*PD, principal diagnosis

Main Concepts

The formal concept had 1059 attributes (PD) and the lattice results had 14,617 concepts. Twelve main concepts were selected using the algorithm described above (Fig. 2a). About one third of patients (64,350) have been readmitted at least once in the first year after surgery, among whom 45.8% have been readmitted for regular follow-up, 6.8% for cholelithiasis, 6.0% for abdominal pain, 4.0% for abdominal sepsis, 2.3% for medical abortion, 2.2% for hernias of the anterior abdominal wall, 1.9% for gastritis and duodenitis, and 1.3% for dysphagia. Among patients readmitted for regular follow-up, 2.0% were successively readmitted for cholelithiasis, 2.8% for abdominal pain, and 4.2% for abdominal sepsis (gastric leak or peritonitis).

According to the type of surgery, patients who had an AGB had the lowest readmission rate (23.4%) compared to the GB (38.2%) and SG (30.9%) patients.

The lattice result has 1209 concepts for AGB, 6726 concepts for GB procedure, and 8429 concepts for SG. The

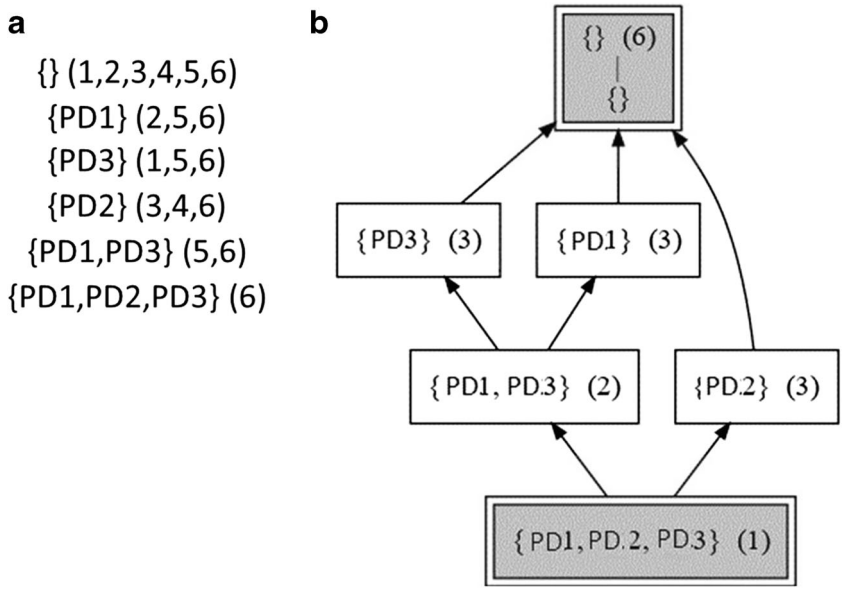


Fig. 1 Example of formal concept analysis results. **a** A subset of concept lattice. **b** Line diagram

number of selected concepts differs according to the type of procedure with 7 concepts for AGB, 10 concepts for the sleeve, and 20 for the bypass procedures (Fig. 2b–d).

Four concepts are common to all types of surgery (Fig. 2). Among readmitted patients, patients operated of GB and SG were more frequently readmitted for regular follow-up, corresponding to the concept (E66, Zx) found in 37.5%, 47.1%, and 45.8% for the AGB, GB, and SG, respectively. Patients with a GB were more commonly readmitted for abdominal pain (8.6%), corresponding to the concept (E66, R10), than other procedures, 3.9% for the AGB and 4.7% for SG.

Among readmitted patients, the concept (E66, K80), corresponding to cholelithiasis, was also present for the three types of surgeries but more common for SG (7.4%) compared to the AGB (3.6%) and GB (6.2%). The rate of medical abortion, corresponding to the concept (E66, O04), was higher in patients with the AGB (4.9%).

The concepts (E66, K43, abdominal wall hernia; E66, K29, gastro-duodenitis; and E66, K316-K65, abdominal sepsis) seem to be more specific to both GB and SG, among readmitted patients. However, patients were more readmitted for hernias of the anterior abdominal wall (K43) after GB (2.7%) than after SG (2.0%), while they were more readmitted for abdominal sepsis (K316-K65) for SG (5.0%) than GB (2.7%).

Some concepts were selected by the algorithm for only one type of surgery. For instance, the concepts (E66, O80, delivery; E66, T85, mechanical complication of a device) have been selected only in patients with an AGB device.

Patient Profile

Patients readmitted for regular follow-up had more comorbidities than the total sample with a Charlson index >3 of 5.1% versus 1.8% and OSAS of 49.3% versus 30.1% (Table 4). Patients in the abdominal sepsis concept (codes K316-K65) have a longer length of stay than other concepts. Patients with the code K43 (abdominal wall hernia) have the highest proportion of open procedures (3.6%), and a higher rate of reoperation during the initial bariatric procedure (5.7%).

Discussion

We have applied the FCA method on data extracted from the PMSI database to analyze trajectories of care within 1 year after bariatric surgery. The use of the PMSI database has several advantages. The first is the access to information of readmissions of the entire population in all health French centers leading to reduction of the selection bias. The second is the ability to identify readmissions of patients in the non-index hospital. Finally, an official guide of data coding [31] is used to record and code data, which provides consistency, quality, and homogeneity of the data.

Formal Concept Analysis has been applied in many fields such as medicine [12], biology [32], genetics [33], and sociology [34] and even to visualize radicalization trajectories over time [35]. In this study, FCA has been used to find clusters of patients having similarities in trajectories of care. Indeed, a better understanding of trajectories of care may help the clinician to anticipate future complications and thus provide better and earlier medical care. The FCA method is simple to implement and only requires a binary table. The PMSI provides massive amount of data and the FCA method can deal with these kinds of database. Furthermore, being an unsupervised method, it does not need a training sample. The main advantage of the FCA is that the formal concept can be visualized. However, one limitation of the FCA method is that

Table 3 Baseline characteristics at the time of the bariatric surgeries

Covariate		Whole sample ($N = 198,389$)	AGB ($N = 13,744$)	GB ($N = 55,945$)	SG ($N = 128,700$)
Gender	Female	160,178 (80.7%)	11,663 (84.9%)	46,143 (82.5%)	102,372 (79.5%)
Age (years)	18–29	42,007 (21.2%)	4542 (33.1%)	8730 (15.6%)	28,735 (22.3%)
	30–39	56,092 (28.3%)	4067 (29.6%)	14,922 (26.7%)	37,103 (28.8%)
	40–49	57,661 (29.1%)	3277 (23.8%)	17,801 (31.8%)	36,583 (28.4%)
	50–60	42,629 (21.5%)	1858 (13.52%)	14,492 (25.9%)	26,279 (20.4%)
BMI (kg/m^2)	30–40	55,151 (30.0%)	6290 (48.0%)	13,627 (26.1%)	35,234 (29.7%)
	40–50	109,025 (59.3%)	6169 (47.0%)	32,601 (62.5%)	70,255 (59.2%)
	≥ 50	19,815 (10.8%)	654 (5.0%)	5956 (11.4%)	13,205 (11.1%)
	NA	14,398	631	3761	10,006
Charlson Comorbidity Index	0	169,492 (85.4%)	12,760 (92.8%)	46,416 (83.0%)	110,316 (85.7%)
	1–2	25,281 (12.7%)	860 (6.3%)	8099 (14.5%)	16,322 (12.7%)
	≥ 3	3616 (1.8%)	124 (0.9%)	1430 (2.6%)	2062 (1.6%)
OSAS*		60,832 (30.7%)	1882 (13.7%)	18,517 (33.1%)	40,433 (31.4%)
Surgical approach	Laparoscopy	197,266 (99.4%)	13,701 (99.7%)	55,587 (99.4%)	127,978 (99.4%)
	Open	1123 (0.6%)	43 (0.3%)	358 (0.6%)	722 (0.6%)
Hospital ownership	Private, for profit	120,662 (61.3%)	11,121 (81.6%)	32,921 (59.6%)	76,620 (59.9%)
	Private, nonprofit	10,190 (5.2%)	342 (2.5%)	2701 (4.9%)	7147 (5.6%)
	Public	65,943 (33.5%)	2160 (15.9%)	19,603 (35.5%)	44,180 (34.5%)
	NA	1594	121	720	753
LOS † (days)	0–1	11,055 (5.6%)	5758 (41.9%)	1718 (3.1%)	3579 (2.8%)
	2–7	176,470 (89.0%)	7965 (58.0%)	50,215 (89.8%)	118,290 (91.9%)
	> 7	10,864 (5.5%)	21 (0.2%)	4012 (7.2%)	6831 (5.3%)
Year of surgery	2013	35,541 (17.9%)	4614 (33.6%)	10,513 (18.8%)	20,414 (15.9%)
	2014	39,339 (19.8%)	3678 (26.8%)	11,286 (20.2%)	24,375 (18.9%)
	2015	39,965 (20.1%)	2373 (17.3%)	11,360 (20.3%)	26,232 (20.4%)
	2016	43,184 (21.8%)	1880 (13.7%)	11,849 (21.2%)	29,455 (22.9%)
	2017	40,360 (20.3%)	1199 (8.7%)	10,937 (19.6%)	28,224 (21.9%)
Reoperation	Yes	3039 (1.5%)	11 (0.1%)	1230 (2.2%)	1798 (1.4%)

*OSAS, obstructive sleep apnea syndrome; BMI, body mass index

† LOS, length of stay

it provides hierarchical concepts. Thus, concepts are non-disjoint, and a patient can belong to several concepts.

In the analysis of trajectories of care after bariatric surgery, the first result is that almost one third of patients are readmitted in the first postoperative year. Even if we included planned and unplanned readmissions, this rate is considerably higher than previously reported [19, 20]. This observation suggests that complications requiring hospital care occur during the entire first year after surgery and not only in the immediate postoperative period.

We also observed some expected trajectories, as the concept (K316-K65), which included gastric leak and peritonitis, which are typical postoperative complications for GB and SG. Also, the concept (R10), abdominal and pelvic pain, is found for the three bariatric procedures, and it has already been reported as a common cause for readmission [27].

The FCA allowed outlining also a few unexpected trajectories, such as the concept (O04), medical abortion; the concept (K80), cholelithiasis; and the concept (K43), abdominal wall hernia. The issue around pregnancy raises questions on the preoperative care of female candidates to bariatric surgery. In fact, national guidelines for bariatric surgery include a birth control planning before bariatric surgery in order to prevent pregnancy at least during the first year after surgery [36]. When we consider any PD related to pregnancy, we observe about 2.1% of pregnancies in the first year after bariatric surgery. This information indicates the need to reinforce the strategy of information about contraception in the preoperative care.

The problem of gallstones (concept K80) is well known as obesity itself and rapid weight loss have been identified as risk factors for gallstone formation [37, 38]. Several meta-analyses also suggest that administration of ursodeoxycholic acid after bariatric surgery prevents gallstones formation [39, 40]. In this study, we found that more than 2% of patients needed a hospital stay for cholelithiasis, being the first and second reason

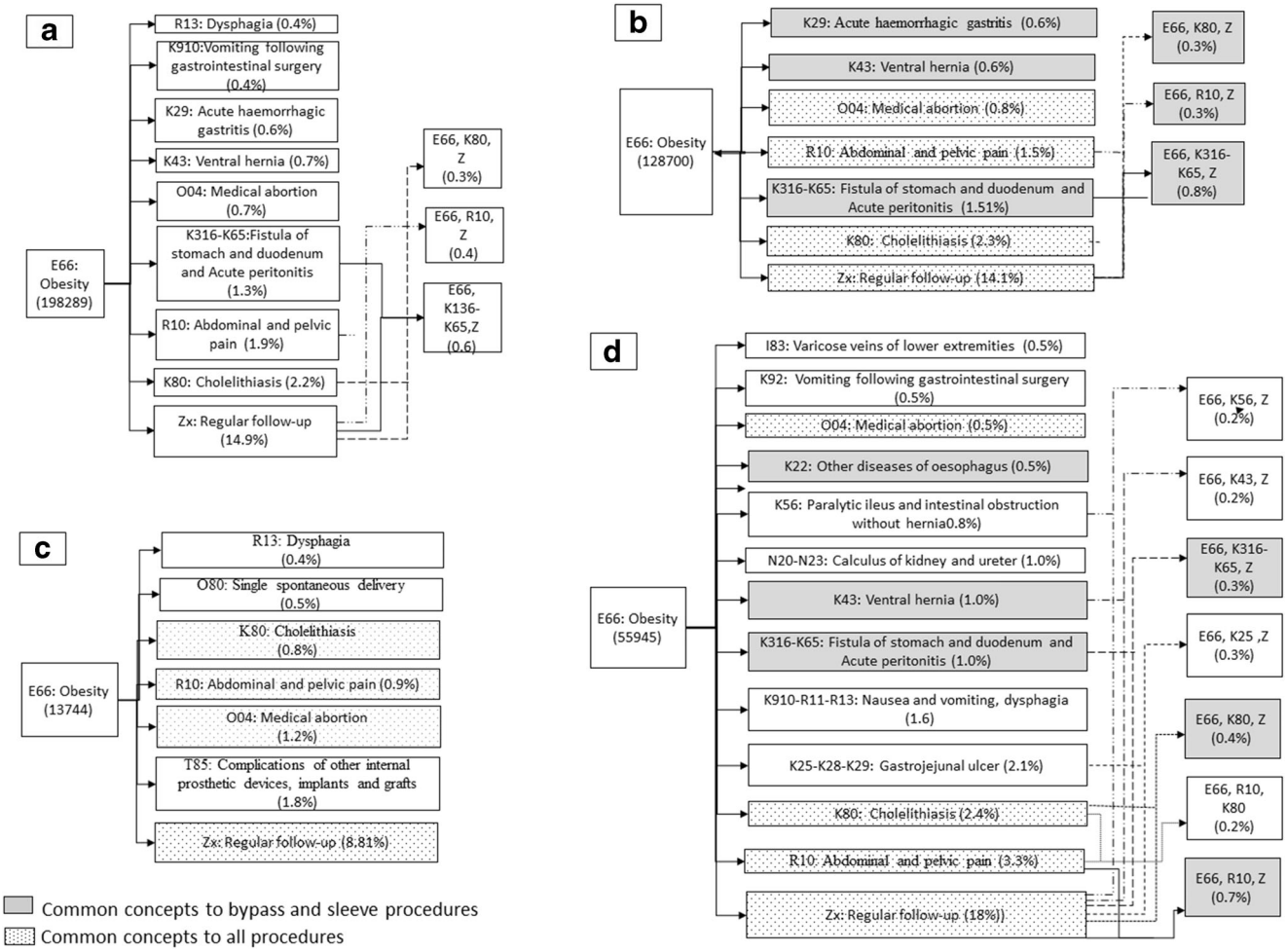


Fig. 2 Main concepts: **a** all type of surgery, **b** sleeve gastrectomy, **c** adjustable gastric banding, **d** gastric bypass

for unplanned readmission after SG and GB, respectively. Even if that goes beyond the aim of present study, it would be of great interest to assess the severity of the disease in this cohort as it has been reported that about 25% of gallstone carriers develop complications [41].

Our analysis also found that about 1% of patients are readmitted for an abdominal wall hernia related problem. This topic, which is quite common in bariatric population, has been recently discussed by the ASMBS with a consensus guideline [42], which conclude that in "patients with severe obesity and ventral hernia, and both being amenable to laparoscopic repair, combined hernia repair and bariatric surgery may be safe and associated with good short-term outcomes and low risk of infection." Authors also outline that at present, "there is a relative lack of evidence, however, about the use of synthetic mesh in this setting."

Limitations

This study has several limitations. Coding errors or coding bias on the principal diagnostic of the readmission is encountered in the database. Moreover, the principal diagnosis has been used to determine the reason of the readmission. The principal diagnosis implies only one reason, but cause of readmission may be more complex and multifactorial.

In total, 9.3% of surgeries were removed from the analysis because of coding error. Surgeries removed from the analysis were comparable to surgeries that were included except for the type of hospital. Indeed, it seems that there are more errors in private than in public hospitals; hence, 7.3%, 0.2%, and 1.7% were removed from private for-profit hospitals, private not-for-profit hospitals, and public hospitals, respectively.

Conclusion

In this study, we assessed the trajectories of care after bariatric surgery through a Formal Concept Analysis. This method permitted to identify trajectories associated either to postoperative complications (as cholelithiasis or abdominal wall hernia) or unexpected events like pregnancy. This approach could help health professional to improve the clinical pathway before and after bariatric surgery.

Table 4 Patients characteristics according to major concepts

Covariate		Whole sample (N = 198,389) {E66}	Regular follow-up (*N* = 29,497) {E66, Z}	Cholelithiasis (*N* = 4382) {E66, K80}	Abdominal pain (*N* = 3841) {E66, R10}	Abdominal sepsis (*N* = 2562) {E66, K316-K65}	Medical Abortion (*N* = 1457) {E66, O04}	Abdominal wall hernia (*N* = 1403) {E66, K43}	Acute hemorrhagic gastritis (*N* = 1192) {E66, K29}	Dysphagia (*N* = 810) {E66, R13}	Vomiting (*N* = 829) {E66, K910}
Gender	Female	160,178 (80.7%)	23,413 (79.4%)	3893 (88.8%)	3334 (86.8%)	2015 (78.7%)	1457 (100%)	1083 (77.2%)	1014 (85.1%)	685 (84.6%)	728 (87.8%)
Age (years)	18–29	42,007 (21.2%)	4551 (15.4%)	1308 (29.9%)	1035 (27.0%)	491 (19.2%)	750 (51.5%)	73 (5.2%)	242 (20.3%)	201 (24.8%)	213 (25.7%)
	30–39	56,092 (28.3%)	7783 (26.4%)	1342 (30.6%)	1137 (29.6%)	728 (28.4%)	631 (43.3%)	298 (21.2%)	296 (24.8%)	210 (25.9%)	216 (26.1%)
	40–49	57,661 (29.1%)	9211 (31.2%)	1029 (23.5%)	1033 (26.9%)	771 (30.1%)	76 (5.2%)	547 (39.0%)	352 (29.5%)	208 (25.7%)	210 (25.3%)
	50–60	42,629 (21.5%)	7952 (27.0%)	703 (16.0%)	636 (16.6%)	572 (22.3%)	0 (0%)	485 (34.6%)	302 (25.3%)	191 (23.6%)	190 (23.0%)
BMI	30–40	55,151 (30.0%)	6172 (22.5%)	1120 (27.5%)	1071 (30.2%)	662 (28.1%)	468 (34.1%)	379 (29.1%)	429 (38.7%)	250 (33.4%)	220 (28.8%)
(kg/m^2)	40–50	109,025 (59.3%)	17,117 (62.3%)	2556 (62.9%)	2155 (60.8%)	1421 (60.4%)	833 (60.8%)	763 (58.6%)	597 (53.9%)	420 (56.1%)	468 (61.2%)
	≥ 50	19,815 (10.8%)	4193 (15.3%)	391 (9.6%)	318 (9.0%)	270 (11.5%)	70 (5.1%)	161 (12.4%)	82 (7.4%)	79 (10.6%)	77 (10.1%)
Charlson comorbidity index	0	169,492 (85.4%)	21,073 (71.4%)	3896 (88.9%)	3196 (83.2%)	2119 (82.7%)	1337 (91.8%)	1099 (78.3%)	973 (81.6%)	666 (82.2%)	653 (78.8%)
	1–2	25,281 (12.7%)	6934 (23.5%)	443 (10.1%)	559 (14.6%)	382 (14.9%)	113 (7.8%)	258 (18.4%)	188 (15.8%)	120 (14.8%)	136 (16.4%)
	≥ 3	3616 (1.8%)	1490 (5.1%)	43 (1.0%)	86 (2.2%)	61 (2.4%)	7 (0.5%)	46 (3.3%)	31 (2.6%)	24 (3.0%)	40 (4.8%)
OSAS*		60,832 (30.7%)	14,552 (49.3%)	1115 (25.5%)	1173 (30.5%)	860 (33.6%)	218 (15.0%)	562 (40.1%)	407 (34.1%)	273 (33.7%)	295 (35.6%)
Surgical approach	Laparoscopic	197,266 (99.4%)	29,250 (99.2%)	4368 (99.7%)	3807 (99.1%)	2518 (98.3%)	1450 (99.5%)	1352 (96.4%)	1181 (99.1%)	805 (99.4%)	817 (98.6%)
	Open	1123 (0.6%)	247 (0.8%)	14 (0.3%)	34 (0.9%)	44 (1.7%)	7 (0.5%)	51 (3.6%)	11 (0.9%)	5 (0.6%)	12 (1.5%)
Hospital ownership	Private, for-profit	120,662 (61.3%)	9486 (32.3%)	2779 (63.8%)	2235 (58.8%)	1450 (57.3%)	961 (66.5%)	810 (58.1%)	647 (55.3%)	415 (51.3%)	388 (47.0%)
	Private, nonprofit	10,190 (5.2%)	1821 (6.2%)	222 (5.1%)	207 (5.5%)	148 (5.9%)	54 (3.7%)	54 (3.9%)	125 (10.7%)	55 (6.8%)	43 (5.2%)
	Public	65,943 (33.5%)	18,081 (61.5%)	1352 (31.1%)	1359 (35.8%)	934 (36.9%)	431 (29.8%)	531 (38.1%)	399 (34.1%)	339 (41.9%)	395 (47.8%)
LOS † (days)	0–1	11,055 (5.6%)	1205 (4.1%)	185 (4.2%)	151 (3.9%)	79 (3.1%)	96 (6.6%)	37 (2.6%)	49 (4.1%)	45 (5.6%)	36 (4.3%)
	2–7	176,470 (89.0%)	25,716 (87.2%)	3964 (90.5%)	3377 (87.9%)	1952 (76.2%)	1307 (89.7%)	1172 (83.5%)	1059 (88.8%)	676 (83.5%)	702 (84.7%)
	> 7	10,864 (5.5%)	2576 (8.7%)	233 (5.3%)	313 (8.2%)	531 (20.7%)	54 (3.7%)	194 (13.8%)	84 (7.1%)	89 (11.0%)	91 (11.0%)
Reoperation		3039 (1.5%)	865 (2.9%)	69 (1.6%)	121 (3.2%)	402 (15.7%)	18 (1.2%)	80 (5.7%)	31 (2.6%)	23 (2.8%)	32 (3.9%)

*OSAS, obstructive sleep apnea syndrome; BMI, body mass index

† LOS, length of stay

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Consent Statement This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Obésité et surpoids [Internet]. [cited 2019 Jul 22]. Available from: https://www.who.int/fr/news-room/fact-sheets/detail/obesity-and-overweight
2. Hall KD, Kahan S. Maintenance of lost weight and long-term management of obesity. Med Clin North Am. 2018;102(1):183–97.
3. Angrisani L, Santonicola A, Iovino P, et al. Bariatric surgery and endoluminal procedures: IFSO worldwide survey 2014. Obes Surg. 2017;27(9):2279–89.
4. Longitudinal Assessment of Bariatric Surgery (LABS) Consortium, Flum DR, Belle SH, et al. Perioperative safety in the longitudinal assessment of bariatric surgery. N Engl J Med. 2009;361(5):445–54.
5. Encinosa WE, Bernard DM, Du D, et al. Recent improvements in bariatric surgery outcomes. Med Care. 2009;47(5):531–5.
6. Regan JP, Inabnet WB, Gagner M, et al. Early experience with two-stage laparoscopic Roux-en-Y gastric bypass as an alternative in the super-super obese patient. Obes Surg. 2003;13(6):861–4.
7. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013;369(15):1434–42.
8. Buchwald H, Estok R, Fahrbach K, et al. Weight and type 2 diabetes after bariatric surgery: systematic review and meta-analysis. Am J Med. 2009;122(3):248–256.e5.
9. García-Ruiz-de-Gordejuela A, Madrazo-González Z, Casajoana-Badia A, et al. Evaluation of bariatric surgery patients at the emergency department of a tertiary referral hospital. Rev Espanola Enfermedades Dig Organo Of Soc Espanola Patol Dig. 2015;107(1):23–8.
10. Doumouras AG, Saleh F, Hong D. 30-day readmission after bariatric surgery in a publicly funded regionalized center of excellence system. Surg Endosc. 2016;30(5):2066–72.
11. Rosenthal RJ, Montorfano L, Abdemur A, et al. Readmission rates of bariatric procedures. J Am Coll Surg. 2015;221:e47–8.
12. Saunders J, Ballantyne GH, Belsley S, et al. One-year readmission rates at a high volume bariatric surgery center: laparoscopic adjustable gastric banding, laparoscopic gastric bypass, and vertical banded gastroplasty-Roux-en-Y gastric bypass. Obes Surg. 2008;18(10):1233–40.
13. Jay N, Nuemi G, Gadreau M, et al. A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer. BMC Med Inform Decis Mak. 2013;13:130.
14. Le Meur N, Gao F, & Bayat S. Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. BMC Health Serv Res. (2015);15:200 https://doi.org/10.1186/s12913-015-0857-5. Accessed Aug 2019
15. Fennig U, Snir A, Halifa-Kurzman I, et al. Pre-surgical weight loss predicts post-surgical weight loss trajectories in adolescents enrolled in a bariatric program. Obes Surg. 2019;29(4):1154–63.
16. Lent MR, Hu Y, Benotti PN, et al. Demographic, clinical, and behavioral determinants of 7-year weight change trajectories in Roux-en-Y gastric bypass patients. Surg Obes Relat Dis Off J Am Soc Bariatr Surg. 2018;14(11):1680–5.
17. Pinto-Bastos A, de Lourdes M, Brandão I, et al. Weight loss trajectories and psychobehavioral predictors of outcome of primary and reoperative bariatric surgery: a 2-year longitudinal study. Surg Obes Relat Dis Off J Am Soc Bariatr Surg. 2019;
18. Berger ER, Huffman KM, Fraker T, et al. Prevalence and risk factors for bariatric surgery readmissions: findings from 130,007 admissions in the metabolic and bariatric surgery accreditation and quality improvement program. Ann Surg. 2018;267(1):122–31.
19. Bruze G, Ottosson J, Neovius M, et al. Hospital admission after gastric bypass: a nationwide cohort study with up to 6 years follow-up. Surg Obes Relat Dis Off J Am Soc Bariatr Surg. 2017;13(6):962–9.
20. Telem DA, Yang J, Altieri M, et al. Rates and risk factors for unplanned emergency department utilization and hospital readmission following bariatric surgery. Ann Surg. 2016;263(5):956–60.
21. PMSI | Fédération hospitalière de France [Internet]. [cited 2019 Apr 5]. Available from: https://www.fhf.fr/gestion-hospitaliere/pmsi.html. Accessed Aug 2019
22. Fetter RB, Shin Y, Freeman JL, et al. Case mix definition by diagnosis-related groups. Med Care. 1980;18(2 Suppl):iii. 1–53
23. Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. J Integr Bioinform. 2018;15(3):20170030. Published 2018 May 10. https://doi.org/10.1515/jib-2017-0030. Accessed Aug 2019
24. Moulis G, Lapeyre-Mestre M, Palmaro A, et al. French health insurance databases: What interest for medical research? Rev Med Interne. 2015;36(6):411–7.
25. Brethauer SA, Kim J, el Chaar M, et al. Standardized outcomes reporting in metabolic and bariatric surgery. Surg Obes Relat Dis Off J Am Soc Bariatr Surg. 2015;11(3):489–506.
26. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol. 2011;173(6):676–82.
27. Lazzati A, Chatellier G, & Katsahian S. Readmissions After Bariatric Surgery in France, 2013–2016: a Nationwide Study on Administrative Data. OBES SURG. (2019);29, 3680–3689. https://doi.org/10.1007/s11695-019-04053-6
28. Ignatov DI. Introduction to Formal Concept Analysis and its applications in information retrieval and related fields. ArXiv170302819 Cs Stat. 2015;505:42–141.
29. Škopljanac-Mačina F, Blašković B. Formal Concept Analysis – overview and applications. Procedia Eng. 2014;69:1258–67.
30. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12(10):e1001885.
31. Guide méthodologique MCO 2018 | Publication ATIH [Internet]. [cited 2019 Aug 15]. Available from: https://www.atih.sante.fr/guide-methodologique-mco-2018. Accessed Aug 2019.
32. Lounkine E, Auer J, Bajorath J. Formal concept analysis for the identification of molecular fragment combinations specific for active and highly potent compounds. J Med Chem. 2008;51(17):5342–8.
33. Gebert J, Motameny S, Faigle U, et al. Identifying genes of gene regulatory networks using formal concept analysis. J Comput Biol. 2008;15(2):185–94.
34. Hao F, Min G, Pei Z, et al. K-clique community detection in social networks based on formal concept analysis. IEEE Syst J. 2017;11(1):250–9.
35. Elzinga P, Poelmans J, Viaene S, Dedene G, Morsing S. Terrorist threat assessment with formal concept analysis. In: 2010 IEEE International Conference on Intelligence and Security Informatics. 2010. p. 77–82. https://doi.org/10.1109/ISI.2010.5484773

36. Ciangura C, Nocca D, Lindecker V. Guidelines for clinical practice for bariatric surgery. Presse Medicale Paris Fr 1983. 2010;39(9): 953–9.
37. Stampfer MJ, Maclure KM, Colditz GA, et al. Risk of symptomatic gallstones in women with severe obesity. Am J Clin Nutr. 1992;55(3):652–8.
38. Yang H, Petersen GM, Roth MP, et al. Risk factors for gallstone formation during rapid loss of weight. Dig Dis Sci. 1992;37(6): 912–8.
39. Magouliotis DE, Tasiopoulou VS, Svokos AA, et al. Ursodeoxycholic acid in the prevention of gallstone formation after bariatric surgery: an updated systematic review and meta-analysis. Obes Surg. 2017;27(11):3021–30.
40. Stokes CS, Gluud LL, Casper M, et al. Ursodeoxycholic acid and diets higher in fat prevent gallbladder stones during weight loss: a meta-analysis of randomized controlled trials. Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc. 2014;12(7): 1090–1100.e2. quiz e61
41. Friedman GD. Natural history of asymptomatic and symptomatic gallstones. Am J Surg. 1993;165(4):399–404.
42. Menzo EL, Hinojosa M, Carbonell A, et al. American Society for Metabolic and Bariatric Surgery and American Hernia Society consensus guideline on bariatric surgery and hernia surgery. Surg Obes Relat Dis Off J Am Soc Bariatr Surg. 2018;14(9):1221–32.

Discussion et conclusion générale

Discussion et limites

Ce travail de thèse a été développé autour de la problématique des réhospitalisations. De nombreuses études ont déjà été menées sur ce sujet puisque près de 1000 articles (recherche pubmed du 05/04/2020 avec "hospital readmission" dans le titre) ont été publiés. Seulement, la plupart d'entre elles, ne se sont intéressées qu'aux facteurs associés à la survenue de la première réhospitalisation [89][90]. Les réhospitalisations peuvent survenir plusieurs fois chez un même patient et ignorer l'information au-delà de la première réhospitalisation peut biaiser les résultats. De plus, d'autres problématiques inhérentes aux événements récurrents se posent comme la dépendance des réhospitalisations d'un même patient ainsi que les potentiels événements terminaux. Au-delà du nombre de réhospitalisations et ses facteurs associés, les raisons des réhospitalisations peuvent nous apporter des informations sur les soins prodigués aux patients et ainsi nous permettre de construire des trajectoires de soins. Des trajectoires de soins ont déjà été étudiées dans le domaine du cancer du sein et dans les soins prénataux. Dans le domaine de l'obésité, seulement les trajectoires de perte de poids ont été étudiées. Ici nous avons proposé d'étudier les trajectoires dans le contexte des réhospitalisations.

Le travail effectué dans cette thèse se compose de deux parties :

— La première partie a consisté à évaluer l'effet de l'âge et du sexe sur les réhospitalisations et le décès liés aux maladies du foie après une greffe de foie. Pour cela, nous avons dans un premier temps fait une revue de la littérature sur les différentes méthodes permettant d'analyser les événements récurrents en présence d'un événement terminal. Dans un second temps, nous avons développé un modèle permettant d'analyser les événements récurrents en présence de risques compétitifs sur l'événement terminal afin de pouvoir analyser simultanément l'effet de l'âge et du sexe sur les réhospitalisations et le décès liés aux maladies du foie après une greffe de foie.

— La deuxième partie a consisté à trouver des trajectoires de soins des patients après la chirurgie bariatrique. Pour cela, nous avons utilisé les diagnostics principaux des réhospitalisations dans la première année suivant la chirurgie. Ces trajectoires ont été construites à partir d'une méthode de data-mining et qui est la méthode d'analyse formelle de concept.

La revue de la littérature des différents modèles développés pour analyser les événements récurrents en présence d'un événement terminal nous a permis de classifier les modèles en quatre classes : les modèles conditionnels non joints, les modèles conditionnels joints, les modèles marginaux non joints et les modèles marginaux joints. La comparaison des différents modèles nous a permis ensuite de produire la démarche à suivre pour choisir le modèle le plus adapté aux données et à la question posée. Le choix du modèle va ainsi dépendre de plusieurs paramètres tels que le processus biologique des événements récurrents, la dépendance entre les événements récurrents et la dépendance entre les événements récurrents et terminaux ainsi que la question posée. Ainsi les modèles conditionnels sont préférés lorsque l'ordre des événements est important ou quand les temps d'événements sont corrélés ou que l'intensité du processus est la quantité d'intérêt. Au contraire, les modèles marginaux sont préférés lorsque les scientifiques s'intéressent au nombre moyen d'événements ou quand les événements sont indépendants. Enfin, l'utilisation d'un modèle joint va dépendre de si la présence de l'association entre les covariables et l'événement terminal est aussi d'intérêt.

Nous avons ensuite montré que le modèle développé pour analyser les événements récurrents et un événement terminal en présence de risque compétitif, estimait bien les paramètres mêmes lorsque la distribution des effets aléatoires était mal spécifiée. L'application de ce modèle sur les données de réhospitalisations après une greffe de foie, a permis de mettre en évidence une association significative entre le sexe et les réhospitalisations ainsi que sur les différentes causes de décès. Nous avons aussi montré une dépendance significative entre les réhospitalisations et les différentes causes de décès.

Enfin, la deuxième partie de ce travail de thèse a permis d'identifier différents trajectoires de soins en fonction du type de chirurgie bariatrique. Le premier résultat est que près d'un tiers des patients sont réadmis dans la première année postopératoire. Nous avons également observé certaines trajectoires attendues, comme le concept (K316-K65), qui comprenait une fuite gastrique et une péritonite, qui sont des complications postopératoires typiques pour GB et SG. Le concept (R10), douleur abdominale et pelvienne, se retrouve également pour les trois procédures bariatriques, et il a déjà été signalé comme une cause fréquente de réadmission.

Les limites liées aux données issues du PMSI

Tout au long de cette thèse, nous nous sommes intéressés aux réhospitalisations. Pour cela, nous avons utilisé des données issues du PMSI [21], qui résument tous les séjours hospitaliers en France par collecte automatique et standardisée des informations. Ces données permettent de récolter une grande quantité de données, appelées données massives ou "big data". Ayant été introduites dans les années 1980, ces données existent depuis longtemps mais n'étaient pas accessibles et la capacité de stockage des ordinateurs et des logiciels les rendait difficilement exploitables. Depuis que la CNIL a homologué, le 7 juin et publié au journal officiel le 13 juillet 2018, les méthodologies de référence MR005 (Annexe C.1) et MR006 (Annexe C.2), l'accès aux données du PMSI a été facilité et est gratuit pour les établissements de santé et des fédérations hospitalières, et fourni contre rémunération pour les prestataires ou laboratoires de recherche

disposant d'un numéro Siret.

Malgré la richesse de l'information contenue dans le PMSI et son accessibilité, ces données peuvent être complétées. En effet, lorsque nous nous sommes intéressés aux réhospitalisations et au décès à la suite d'une greffe de foie, seuls les décès survenus au cours d'une hospitalisation sont enregistrés, ainsi l'incidence des décès est sous-estimée dans le PMSI. De plus, des facteurs de confusion tels que par exemple les données du secteur médico-social ou encore les données génomiques, ou les traitements pris par les patients qui pourraient avoir un impact sur les réhospitalisations et/ou le décès ne sont pas incluses dans le PMSI. Ainsi des données complémentaires telles que :

- les données issues de l'échantillon généraliste des bénéficiaires (EGB), contenant des informations anonymes sur les caractéristiques sociodémographiques (âge, sexe et lieu de résidence) et médicales des bénéficiaires et les prestations qu'ils ont perçues (https ://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/sniiram/structure-du-sniiram.php).
- les données issues du CépiDC contenant les causes de décès (https ://www.cepidc.inserm.fr/)
- les données génomiques proviennent de plateformes de séquençage.

permettraient d'enrichir les connaissances sur la santé. Cependant, ces données sont stockées sur différentes structures d'hébergement qui ne sont pas forcément connectées rendant difficiles le croisement entre les différentes sources de données ce qui peut impacter la qualité des données en générant des doublons. Le Health data Hub (https ://www.health-data-hub.fr/), mis en place durant le premier trimestre 2019, est une plateforme permettant de remédier à ce problème en regroupant toutes les données à un seul endroit et mettant en relation ces données afin de favoriser les études, les recherches ou évaluations présentant un caractère d'intérêt public. Cependant, cette plateforme a été ouverte début 2020. Une problématique importante engendrée par ces données massives est la difficulté dégager de l'information pertinente.

Les limites liées à la fouille de données (data-mining)

Dans le cadre de cette thèse, nous nous sommes intéressés aux trajectoires de soins des patients la première année suivant la chirurgie bariatrique. Pour cela, nous avons utilisé la méthode d'analyse formelle de contexte (FCA), qui fait partie des méthodes de data-mining. La FCA a permis de mettre en évidence des combinaisons de diagnostics à l'origine de réhospitalisations. Cependant, la chronologie des réhospitalisations n'a pas été prise en compte. En effet, si par exemple, un patient est réhospitalisé pour dysphagie, vomissement et douleur abdominale, et qu'un autre patient est réhospitalisé pour vomissement, douleur abdominale et dysphagie, ces deux patients appartiendront à la même classe. Vu le nombre important de diagnostics observés, très peu de trajectoires de soins communes à plusieurs patients auraient été mises en évidence si la chronologie avait été prise en compte.

Les méthodes de data-mining sont des méthodes permettant d'extraire, de décrire et de visualiser de l'information provenant de base de données massives. Elles ne permettent pas de montrer des associations ou des dépendances entre les variables mais peuvent aider à trouver des nouveaux signaux. A part les méthodes de classifications supervisées, ce sont des méthodes

non supervisées et sans *a priori*.

Les limites liées au développement du modèle joint

Contrairement aux méthodes de fouille de données, qui sont des méthodes exploratoires permettant de trouver de nouveaux signaux, le modèle proposé dans la première partie de cette thèse, afin d'analyser simultanément les événements récurrents et un événement terminal en présence de risque compétitif, s'inscrit dans la recherche hypothético-déductive. Cette problématique a été motivée par les données de la greffe de foie. L'objectif était d'évaluer les effets de l'âge et du sexe sur les réhospitalisations après une greffe de foie ainsi que sur les décès liés aux maladies du foie. Cependant, de nombreux décès autres que ceux liés aux maladies du foie ont été trouvés dans la base de données, et ont ainsi été considérés comme un risque compétitif.

Le modèle développé est un modèle joint incluant deux termes de fragilité, permettant d'évaluer la dépendance entre les événements récurrents et chaque type d'événement terminal. Les termes de fragilité sont issus d'une loi Gamma. Les fonctions de risque de base sont quant à elles modélisées de façon faiblement paramétriques à partir de M-splines.

Nous avons choisi la distribution Gamma pour les termes de fragilité surtout pour ses propriétés mathématiques et pour sa flexibilité. En effet, aucune étude clinique ne s'est intéressée à la distribution de ces paramètres et ainsi aucun argument clinique ou médical ne permet de choisir la distribution la plus adaptée. Nous avons donc voulu étudier le comportement du modèle dans le cas où la distribution de la dépendance ne suivait pas une loi Gamma. Les résultats des simulations ont donc montré que la mauvaise spécification de la loi n'influence pas les résultats des coefficients des covariables.

La fonction de risque de base, si elle est mal spécifiée, peut elle aussi influencer les résultats. Dans la littérature, la fonction de risque de base est souvent modélisée soit par une fonction constante par morceau [42][58] ou de type Weibull [41], toutes deux étant des méthodes paramétriques. Les fonctions de risque de base ne sont pas toutes connues par les cliniciens qui s'attendent généralement à avoir une fonction de risque de base continue ne suivant pas forcément une loi de Weibull. La fonction de risque de base pouvant prendre n'importe quelle forme, nous avons choisi de la modéliser par des M-splines d'ordre 4. Les M-splines sont faiblement paramétriques. Nous avons étudié le comportement du modèle en faisant varier la forme de la fonction de risque de base pour les événements récurrents. Les résultats ont montré que l'ensemble des coefficients sont bien estimés.

Cependant, ce modèle a quelques limites. En effet, le modèle fait l'hypothèse de proportionnalité des risques, c'est à dire que l'effet des covariables sur les fonctions d'incidence des événements (récurrent ou terminal) est constant au cours du temps. Or après une greffe de foie, certaines covariables peuvent avoir un effet sur la survie et/ou la réhospitalisation qui varient au cours du temps. En effet, Ren et al [91] ont comparé les modèles de Cox et de Gray avec des coefficients dépendant du temps afin d'évaluer les facteurs associés à la survie d'enfants ayant une maladie hépatique de stade terminal après une transplantation de foie. Ils ont alors montré que le fait que les donneurs aient un groupe sanguin de type AB et le sexe du receveur ainsi

que l'utilisation de ventilateur au moment de la transplantation sont significativement associés à la survie des patients après la transplantation de foie, et que leur effet varient au cours du temps.

Une autre limite du modèle est que l'on suppose une dépendance positive entre les événements récurrents et chaque événement terminal. Cependant, dans de nombreuses disciplines, il est rare que les événements récurrents et terminaux soient négativement dépendants.

Conclusion et perspectives

Dans la littérature, les événements récurrents et terminaux sont analysés séparément alors que ceux-ci peuvent être dépendants. Le travail sur les modèles récurrents en présence d'un ou plusieurs événements terminaux a permis de mieux comprendre les différents modèles existants et ainsi de produire des indications pour le choix du modèle le plus adapté à la problématique. En effet, ces modèles permettent d'utiliser la totalité de l'information présente dans la base de données et d'avoir des analyses plus puissantes. Dans le cadre d'essais cliniques analysant les réhospitalisations, ces modèles pourront être ainsi utilisés par la plupart des statisticien(ne)s.

La deuxième partie de cette thèse a permis de montrer que des informations autres que le nombre de réhospitalisations peuvent être utilisées. En effet, la raison des réhospitalisations est une information importante qui permet d'identifier des trajectoires de soins communes à plusieurs sujets. Identifier ces trajectoires permet d'anticiper les soins et peut être utilisé dans les essais cliniques (par exemple identification de sous-groupes de patients en fonction des trajectoires de soins).

Les extensions du modèle développé

Nous avons vu que le modèle développé avait quelques limites. Des extensions peuvent alors être envisagées afin de pallier ces problèmes.

Dans un premier temps, il serait possible d'ajouter deux termes de flexibilité afin de permettre une dépendance positive et négative entre les événements récurrents et chaque événement terminal. En effet, le modèle développé ne permet seulement que des dépendances positives entre les différents processus d'événements.

Une autre extension possible serait d'introduire des coefficients de régression dépendant du temps. En effet, l'effet d'une covariable peut évoluer à court et long terme. Par exemple, Ren [91] a montré que certaines covariables avaient des effets qui varient au cours du temps sur la survie d'enfants atteints de maladies hépatiques de stade terminal. Pour cela, des splines peuvent être utilisés pour estimer les coefficients dépendant du temps.

Dans le modèle proposé, nous avons inclus deux termes de fragilité permettant d'évaluer la dépendance entre les événements récurrents et chaque type d'événement terminal. Cependant, il peut être d'intérêt de différencier la dépendance intra-récurrence et la dépendance entre les événement récurrents et les événements terminaux. Pour cela, on pourrait inclure un troisième terme de fragilité afin de différencier les différences sources d'hétérogénéité.

Nous avons proposé un modèle où un seul type d'événement récurrent et deux événements

terminaux sont pris en compte. Cependant, il serait facile d'étendre ce modèle à plusieurs types d'événements récurrents et rajouter d'autres risques en compétition.

Utilisation du modèle dans les protocoles d'essais cliniques

Lors de l'élaboration des essais cliniques, il est primordial de bien décrire le critère de jugement principal. En effet, c'est sur ce critère là que l'on pourra conclure à l'efficacité d'un traitement. Pour cela, il est nécessaire de calculer la taille d'échantillon qui permettra d'observer la différence que l'on souhaite mettre en évidence. Ce calcul doit se faire à partir de l'analyse statistique qui sera utilisée pour analyser le critère de jugement principal. Pour pouvoir utiliser le modèle développé, il est donc primordial de développer une méthode de calcul du nombre de sujets nécessaires pour mettre en évidence une différence. Un des avantages d'utiliser ce modèle est qu'il permettra de répondre à plusieurs questions en seulement une seule analyse, puisque habituellement, le critère de jugement principal est unique.

Stratégie d'analyse combinant les méthodes d'exploration et des méthodes hypothético-déductives

Les méthodes d'explorations comme décrites dans le chapitre 5 sont des méthodes sans *a priori*. Elles permettent d'extraire de l'information pertinente d'une grande base de données où l'information utile est noyée dans la quantité d'information. Ces méthodes permettent alors d'émettre des hypothèses de recherche qui peuvent ensuite être validées par des méthodes hypothético-déductives. Ainsi, il serait intéressant de pouvoir coupler les méthodes d'identifications de trajectoires pour ensuite pouvoir les analyser par les méthodes statistiques. Par, exemple, les trajectoires d'hospitalisations les plus fréquentes après une greffe de foie pourraient être identifiées par la méthode de FCA dans un premier temps. Ensuite dans une étude de cohorte ou dans un essai clinique, l'objectif pourrait être d'évaluer l'effet des covariables sur les raisons de réhospitalisations trouvées lors de l'étude des trajectoires. Ceci permettrait d'identifier les variables associées aux réhospitalisations les plus fréquentes et de pouvoir à terme mieux prendre en charge ces patients afin d'éviter de nouvelles réhospitalisations.

On pourrait aussi penser, que l'effet des covariables pourrait être différent selon les trajectoires de soins. En effet, l'identification des trajectoires de soins pourrait permettre de déterminer des sous-groupes de gravité de patients. L'interaction entre les sous groupes de gravité et les covariables pourrait être ainsi testée à l'aide du modèle développé.

Annexes

Annexe A

Source des données : le programme de médicalisation des systèmes d'information

Les données sont extraites du programme de médicalisation des systèmes d'information (PMSI)[21] qui résume tous les séjours hospitaliers en France. Le PMSI a d'abord été introduit dans les années 1980 et a été présenté comme un outil épidémiologique avant de devenir un outil d'allocation budgétaire. C'est un dispositif inclus dans la réforme du système de santé français et qui a pour objectif de réduire les inégalités de ressources entre les établissements de santé. Afin de répondre à ces objectifs, les informations ont été quantifiées et standardisées. En effet, le PMSI a été élaboré en s'inspirant du modèle américain développé par le Pr Fetter dans les années 70 [114, 113] , basé sur la création des groupes homogènes de diagnostic (DRG) qui a ensuite été adapté par le système français en groupes homogènes de malades (GHM). L'homogénéisation de ces groupes, à la fois médicale et financière, a permis d'estimer les coûts des hospitalisations.

Informations contenues dans le PMSI

Différents types d'informations sont contenus dans le PMSI :

— Les informations administratives relatives à l'établissement et au patient (âge, sexe, code postal, modalité d'entrée et de sortie) et au séjour (durée du séjour, mois et année du séjour).

— Les information médicales ; les diagnostics codés à partir de la classification internationale des maladies et recours aux services de santé n°10 (CIM-10)[104, 103], et actes diagnostiques ou thérapeutiques classés selon la Classification Communes des Actes Médicaux (CCAM).

— Les informations de groupage indiquant le GHM et la catégorie majeure de diagnostic dans lesquels le séjour est classé ainsi que le tarif associé.

Annexe B

Calcul de la vraisemblance

B.1 Notions nécessaires pour la construction de la fonction de densité pour l'écriture de la vraisemblance

B.1.1 Le "product integral"

Le "*product integral*" d'une fonction continue et intégrable $g(u)$ sur un intervalle $[a, b]$, où $a = u_0 < u_1 < ... < u_R = b$ est une partition de $[a, b]$, $\Delta u_r = u_{r+1} - u_r$, $r = 0, 1, ..., R$ et $u_{R+1} = u_R^+$, est définie par :

$$\prod_{[a,b]} (1 + g(u)du) = \lim_{R\to\infty} \prod_{r=0}^{R} (1 + g(u_r)\Delta u_r) \tag{B.1}$$

où quand $R \to \infty$, $max(\Delta u_r)$ tend vers 0. Le développement limité d'ordre 1 de $log\,(1 + g(u)\Delta u) = g(u)\Delta u + o(\Delta u)$. On peut donc voir que le log de B.1 tend vers l'intégrale de Riemann $\int_a^b g(u)du$ et donc :

$$\prod_{[a,b]} (1 + g(u)du) = exp\left\{\int_a^b g(u)du\right\}$$

Ce développement est aussi applicable si g a un nombre fini de discontinuités sur l'intervalle $[a, b]$. Il est facile de voir que :

$$\prod_{[a,b]} (1 + g(u)du + o(du)) = exp\left\{\int_a^b g(u)du\right\} \tag{B.2}$$

B.1.2 L'intégrale de Riemann-Stieltjes

Soit G(t) une fonction non-décroissante, continue à droite avec une limite à gauche et un nombre fini de discontinuités sur n'importe quel intervalle fini. Supposons que $g(t) = G'(t)$ existe sauf aux points de discontinuité de $G(.)$, et qu'en ces points de discontinuité t_j, nous ayons $G(t_j) - G(t_j^-) = g_j$. L'intégrale de Riemann-Stieltjes de $dG(.)$ (différentielle de G) sur

l'intervalle $[a, b]$ est alors définie par :

$$\int_a^b dG(u) = \int_a^b g(u)du + \sum_{j:a\leq t_j\leq b} g_j$$

où $\int_a^b g(u)du$ est une intégrale de Riemann. L'intégrale de Riemann-Stieltjes est alors une intégrale de Riemann lorsque $G(t)$ est continue, et est réduite à une somme si $G(t)$ est une fonction en escaliers avec des sauts g_j en un nombre de points fini t_j.

B.2 La vraisemblance

B.2.1 La fonction de densité de probabilité

Nous pouvons maintenant adapter les notions ci-dessus au processus d'événements observé sur l'intervalle de temps $[\tau_0, \tau]$, conditionnellement à $H(\tau_0)$, où $H(t) = \{N(s) : 0 \leq s \leq t\}$ est l'histoire du patient. La densité de probabilité du critère "survenue de n événements aux temps $t_1 < .... < t_n \leq \tau$" d'événements peut être obtenue en considérant la partition $\tau_0 = u_0 < u_1 < ... < u_R = \tau$ et en prenant la limite. La distribution de probabilité de $N(u_1), ..N(u_R)$ sachant $H(u_0)$ est :

$$\prod_{r=0}^{R} \mathbb{P}\left(N(u_r)|H(u_r)\right) = \prod_{r=0}^{R} \mathbb{P}\left(\Delta N(u_r)|H(u_r)\right)$$

où $\Delta N(u_r)$ est le nombre d'événements sur l'intervalle $[u_r, u_{r+1})$. En supposant que deux événements ne peuvent pas survenir en même temps on obtient :

$$\begin{aligned}
\prod_{r=0}^{R} \mathbb{P}\left(\Delta N(u_r) = 0|H(u_r)\right) &= 1 - \lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r) \\
\prod_{r=0}^{R} \mathbb{P}\left(\Delta N(u_r) = 1|H(u_r)\right) &= \lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)
\end{aligned}$$

où $o(\Delta u_r) = \mathbb{P}\left(\Delta N(u_r) \geq 2|H(u_r)\right)$ et $\lambda(t|H(t)$ est la fonction d'intensité du processus d'événements et s'écrit :

$$\lambda(t|H(t)) = \lim_{\Delta t\downarrow 0} \frac{\mathbb{P}\left(\Delta N(t) = 1|H(t)\right)}{\Delta t} \tag{B.3}$$

On a alors :

$$\begin{aligned}
\prod_{r=0}^{R} \mathbb{P}\left(N(u_r)|H(u_r)\right) = \quad & \prod_{r=0}^{R} \left\{\lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)\right\}^{\Delta N(u_r)} \\
& \left\{1 - \lambda(u_r|H(u_r))\Delta u_r + o(\Delta u_r)\right\}^{1-\Delta N(u_r)}
\end{aligned} \tag{B.4}$$

En augmentant R, la taille des termes Δu_r tend vers 0, ainsi les n intervalles contenant les temps d'événements $t_1, ..., t_n$ ont $\Delta N(u_r) = 1$ et pour tous les autres intervalles $\Delta N(u_r) = 0$. En divisant B.4 par $\prod_{r=0}^{R} (\Delta u_r)^{\Delta N(u_r)}$ et quand $R \to \infty$, on obtient que conditionnellement à $H(\tau_0)$, la densité de probabilité de "la survenue de n événements aux temps $t_1 < ... < t_n$" quand $n \geq 0$, pour un processus avec une intensité B.3 sur l'intervalle $[\tau_0, \tau]$ est :

$$\prod_{j=1}^{n} \lambda(t_j|H(t_j)) \times exp\left\{\int_{\tau_0}^{\tau} -\lambda(u|H(u))du\right\}$$

Le terme exponentiel est obtenu par le résultat du "product integral" de l'équation B.2 en posant $g(u) = -\lambda(u|H(u))$.

Un autre résultat important est que, pour un processus d'événement avec une intensité B.3 intégrable :

$$\mathbb{P}(N(s,t) = 0|H(s^+)) = exp\left\{-\int_{s}^{t} \lambda(u|H(u))du\right\}$$

Le modèle s'écrit :

$$\begin{cases} r_i(t \mid u_{i1}, u_{i2}) = u_{i1}u_{i2}r_0(t)\exp\left(\beta_R Z_i^R + \beta_{R'} Z_i^R\right) & \text{(recurrences)} \\ \alpha_i^{(1)}(t \mid u_{i1}) = u_{i1}\alpha_0^{(1)}(t)\exp\left(\beta_{D_1} Z_i^{D_1} + \beta_{D_1'} Z_i^{D_1}\right) & \text{(terminal event 1)} \\ \alpha_i^{(2)}(t \mid u_{i2}) = u_{i2}\alpha_0^{(2)}(t)\exp\left(\beta_{D_2} Z_i^{D_2} + \beta_{D_2'} Z_i^{D_2}\right) & \text{(terminal event 2)} \end{cases} \tag{B.5}$$

La vraisemblance totale est le produit des vraisemblances pour les événements récurrents et pour chaque événement terminal.

B.2.2 La vraisemblance pour les événements récurrents

En posant ω l'ensemble des paramètres à estimer $\phi = \{r_0(.), \alpha_0^{(1)}(.), \alpha_0^{(2)}(.), \beta, \theta_1, \theta_2\}$, la vraisemblance conditionnelle pour les événements récurrents pour le patient i est :

$$\begin{aligned} V_i^R(\omega|u_{i1}, u_{i2}) = & \prod_{j=1}^{n_i} \mathbb{P}\left(\Delta N_i^R(T_{ij}) = 1|H_i(T_{ij})\right)^{\Delta N_i^R(T_{ij})} \times \mathbb{P}\left(\Delta N_i^R(T_{ij}) = 0|H_i(T_{ij})\right) \\ & \times \mathbb{P}\left(\Delta N_i^R(T_{n_i+1}) = 0|H_i(T_{n_i+1})\right) \end{aligned}$$

Le dernier terme correspond au fait qu'après le dernier évènement récurrent il peut y avoir du suivi et donc être censuré.

$$\begin{aligned} V_i^R(\omega|u_{i1}, u_{i2}) &= \prod_{j=1}^{n_i} \mathbb{P}\left(\Delta N_i^R(T_{ij}) = 1|H_i(T_{ij})\right)^{\Delta N_i^R(T_{ij})} \times \left(1 - \mathbb{P}\left(\Delta N_i^R(T_{ij}) = 0|H_i(T_{ij})\right)\right) \\ &\times \left(1 - \mathbb{P}\left(\Delta N_i^R(T_{n_i+1}) = 1|H_i(T_{n_i+1})\right)\right) \\ &= \prod_{j=1}^{n_i} r_i(T_{ij}|u_{i1}, u_{i2})^{\Delta N_i^R(T_{ij})} \\ &\times exp\left\{-\int_{T_{i(j-1)}}^{T_{ij}} r_i(u|u_{i1}, u_{i2})du\right\} \times exp\left\{-\int_{T_{n_i}}^{T_{i(n_i+1)}} r_i(u|u_{i1}, u_{i2})du\right\} \\ &= \prod_{j=1}^{n_i} r_i(T_{ij}|u_{i1}, u_{i2})^{\Delta N_i^R(T_{ij})} \\ &\times exp\left\{-\sum_{k=1}^{n_i} \int_{T_{i(k-1)}}^{T_{ik}} r_i(u|u_{i1}, u_{i2})du\right\} \times exp\left\{-\int_{T_{n_i}}^{T_{i(n_i+1)}} r_i(u|u_{i1}, u_{i2})du\right\} \\ &= \prod_{j=1}^{n_i} r_i(T_{ij}|u_{i1}, u_{i2})^{\Delta N_i^R(T_{ij})} \times exp\left\{-\sum_{k=1}^{n_i+1} \int_{T_{i(k-1)}}^{T_{ik}} r_i(u|u_{i1}, u_{i2})du\right\} \end{aligned}$$

B.2.3 La vraisemblance pour les événements terminaux

La vraisemblance conditionnelle pour les événements terminaux pour le sujet i :

$$\begin{aligned} V_i^{(1)}(\omega|u_{i1}) = & \left(\alpha_i^{(1)}(T_i^*|u_{i1})\right)^{\Delta N_i^{D_1}(T_i^*)} \times exp\left\{-\int_0^{T_i^*} \alpha_i^{(1)}(u|u_{i1})du\right\} \\ V_i^{(2)}(\omega|u_{i2}) = & \left(\alpha_i^{(2)}(T_i^*|u_{i1})\right)^{\Delta N_i^{D_2}(T_i^*)} \times exp\left\{-\int_0^{T_i^*} \alpha_i^{(2)}(u|u_{i1})du\right\} \end{aligned}$$

B.2.4 La vraisemblance

La vraisemblance totale pour le patient i est le produit des vraisemblances pour les événements récurrents et pour chaque type d'événement terminal.

$$V_i(\omega|u_{i1}, u_{i2}) = V_i^R(\phi|u_{i1}, u_{i2}) \times V_i^{(1)}(\phi|u_{i1}) \times V_i^{(2)}(\phi|u_{i2})$$

La vraisemblance marginale pour le patient i, s'obtient en intégrant la vraisemblance conditionnelle par rapport à u_{i1} et u_{i2} :

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} V_i(\phi|u_{i1}, u_{i2}) f(u_{i1} f(u_{i2}) du_{i2} du_{i1}$$

Les effets aléatoires suivent une loi Gamma de paramètre θ_c, $u_{ic} \sim \Gamma\left(\frac{1}{\theta_c}, \frac{1}{\theta_c}\right)$, de densité $f(.)$, $f(u_{ic}) = \frac{u_{ic}^{\frac{1}{\theta_c}-1} exp\left\{-\frac{u_{ic}}{\theta_c}\right\}}{\theta_c^{\frac{1}{\theta_c}}\Gamma\left(\frac{1}{\theta_c}\right)}$, $c = 1, 2$. Loi Gamma étant définie sur $[0, +\infty[$, les effets aléatoires étant mutuellement indépendants et identiquement distribués, on obtient :

$$\begin{aligned} V_i(\omega) = & \left(\alpha_0^{(1)}(T_i^*) exp\left\{\beta_{D_1} Z_i^{D_1}(T_i^*)\right\}\right)^{\delta_i^{D_1}} \left(\alpha_0^{(2)}(T_i^*) exp\left\{\beta_{D_2} Z_i^{D_2}(T_i^*)\right\}\right)^{\delta_i^{D_2}} \\ & \times \prod_{j=1}^{n_i} \left(r_0(T_{ij})) exp\left\{\beta_R Z_i^R(T_{ij})\right\}\right)^{\delta_{ij}} \frac{1}{\theta_1^{\frac{1}{\theta_1}}\tau(\frac{1}{\theta_1})} \frac{1}{\theta_2^{\frac{1}{\theta_2}}\tau(\frac{1}{\theta_2})} \\ & \times \int_0^{\infty} u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\theta_1}-1} exp\left\{-u_{i1}\left(\int_0^{T_i^*} Y_i(t)\alpha_0^{(1)}(t) exp\left\{\beta_{D_1} Z_i^{D_1}(t)\right\} dt + \frac{1}{\theta_1}\right)\right\} \\ & \times \int_0^{\infty} u_{i2}^{n_i+\delta_i^{D_2}+\frac{1}{\theta_2}-1} exp\left\{-u_{i2}\left(u_{i1} \sum_{k=1}^{n_i+1} \int_{T_{i(k-1)}}^{T_{ik}} Y_i(t) r_0(t) exp\left\{\beta_R Z_i^R(t)\right\} dt + \frac{1}{\theta_2}\right.\right. \\ & \left.\left. + \int_0^{T_i^*} Y_i(t)\alpha_0^{(2)}(t) exp\left\{\beta_{D_2} Z_i^{D_2}(t)\right\} dt\right)\right\} du_{i2} du_{i1} \end{aligned}$$

En utilisant les moments d'ordre n de la loi Gamma :

$$\int_0^{\infty} x^{n+a-1} \frac{\lambda^a}{\Gamma(a)} exp\left\{-\lambda x\right\} dx = \frac{\Gamma(n+a)}{\lambda^n \Gamma(a)}$$

et en posant :

$n = n_i,$
$a = \frac{1}{\theta 2} + \Delta_i N^{D2}(T_i^*),$
$\lambda = \left(u_{i1} \sum_{k=1}^{n_i+1} \int_{T_{i(k-1)}}^{T_{ik}} Y_i(t) r_0(t) exp\left\{\beta_R Z_i^R(t)\right\} dt + \frac{1}{\theta_2} + \int_0^{T_i^*} Y_i(t)\alpha_0^{(2)}(t) exp\left\{\beta_{D_2} Z_i^{D_2}(t)\right\} dt \right)$

On obtient :

$$\begin{aligned} V_i(\omega) &= \left(\alpha_0^{(1)}(T_i^*) exp\left\{\beta_{D_1} Z_i^{D_1}(T_i^*)\right\}\right)^{\delta_i^{D_1}} \left(\alpha_0^{(2)}(T_i^*) exp\left\{\beta_{D_2} Z_i^{D_2}(T_i^*)\right\}\right)^{\delta_i^{D_2}} \\ &\times \prod_{j=1}^{n_i} \left(r_0(T_{ij})) exp\left\{\beta_R Z_i^R(T_{ij})\right\}\right)^{\delta_{ij}} \frac{1}{\theta_1^{\frac{1}{\theta_1}} \tau(\frac{1}{\theta_1})} \frac{1}{\theta_2^{\frac{1}{\theta_2}} \tau(\frac{1}{\theta_2})} \Gamma(n_i + \frac{1}{\theta_2} + \delta_i^{D_2}) \\ &\times \int_0^{\infty} \frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\theta_1}-1} exp\left\{-u_{i1}\left(\int_0^{T_i^*} Y_i(t)\alpha_0^{(1)}(t) exp\left\{\beta_{D_1} Z_i^{D_1}(t)\right\} dt + \frac{1}{\theta_1}\right)\right\}}{\left(u_{i1} \sum_{k=1}^{n_i+1} \int_{T_{i(k-1)}}^{T_{ik}} Y_i(t) r_0(t) exp\left\{\beta_R Z_i^R(t)\right\} dt + \frac{1}{\theta_2} + \int_0^{T_i^*} Y_i(t)\alpha_0^{(2)}(t) exp\left\{\beta_{D_2} Z_i^{D_2}(t)\right\} dt\right)^{\frac{1}{\theta_2}+\delta_i^{D_2}+n_i}} du_{i1} \end{aligned}$$

En prenant ensuite le logarithme de la vraisemblance, on a :

$$\begin{aligned} log(V_i(\omega)) &= \delta_i^{D_1} log\left(\alpha_0^{(1)}(T_i^*) exp\left\{\beta_{D_1} Z_i^{D_1}(T_i^*)\right\}\right) + \delta_i^{D_2} log\left(\alpha_0^{(2)}(T_i^*) exp\left\{\beta_{D_2} Z_i^{D_2}(T_i^*)\right\}\right) \\ &+ \sum_{j=1}^{n_i} \delta_{ij} log\left(r_0(T_{ij})) exp\left\{\beta_R Z_i^R(T_{ij})\right\}\right) - \frac{1}{\theta_1} log(\theta_1) - log\left(\tau(\frac{1}{\theta_1})\right) - \frac{1}{\theta_2} log(\theta_2) - log\left(\tau(\frac{1}{\theta_2})\right) \\ &+ log\left(\int_0^{+\infty} g(u_{i1} d_u i1\right) \end{aligned}$$

où $g(u_{i1}) = \int_0^{\infty} \frac{u_{i1}^{n_i+\delta_i^{D_1}+\frac{1}{\theta_1}-1} exp\left\{-u_{i1}\left(\int_0^{T_i^*} Y_i(t)\alpha_0^{(1)}(t) exp\left\{\beta_{D_1} Z_i^{D_1}(t)\right\} dt + \frac{1}{\theta_1}\right)\right\}}{\left(u_{i1} \sum_{k=1}^{n_i+1} \int_{T_{i(k-1)}}^{T_{ik}} Y_i(t) r_0(t) exp\left\{\beta_R Z_i^R(t)\right\} dt + \frac{1}{\theta_2} + \int_0^{T_i^*} Y_i(t)\alpha_0^{(2)}(t) exp\left\{\beta_{D_2} Z_i^{D_2}(t)\right\} dt\right)^{\frac{1}{\theta_2}+\delta_i^{D_2}+n_i}} du_{i1}$

B.3 Dérivée première et seconde des M-splines

Pour $c \in \{2, .., l\}$ et $m \in \{1, ..., p + 2d - c\}$, la dérivée première de $M_m(t|c, x)$ en fonction de t est :

$$M_m^{(\prime)}(t|c,x) = \begin{cases} \frac{c\left\{(t-x_m) M'_m(t|c-1,x) + M_m(t|c-1,x) + (x_{m+c}-t) M'_{m+1}(t|c-1,x) - M_{m+1}(t|c-1,x)\right\}}{(c-1)(x_{m+c}-x_m)} & \text{si } x_m \leq t \leq x_{m+c} \\ 0 & \text{sinon} \end{cases}$$

et la dérivée seconde est :

$$M_m''(t|c,x) = \begin{cases} \frac{c\Big\{(t-x_m)M_m''(t|c-1,x)+2M_m'(t|c-1,x)+(x_{m+c}-t)M_{m+1}''(t|c-1,x)-2M_{m+1}'(t|c-1,x)\Big\}}{(c-1)(x_{m+c}-x_m)} & \text{si } x_m \leq t \leq x_{m+c} \\ 0 & \text{sinon} \end{cases}$$

Annexe C

Les méthodologies de référence

C.1 MR005

DECLARATION

20/02/2020

MR 05
Études nécessitant l’accès aux données du PMSI et/ou des RPU par les établissements de santé et les fédérations hospitalières

ÉTUDES NÉCESSITANT L'ACCÈS AUX DONNÉES DU PMSI ET/OU DES RPU PAR LES ÉTABLISSEMENTS DE SANTÉ ET LES FÉDÉRATIONS HOSPITALIÈRES

(Déclaration N° 05)

La méthodologie de référence MR-005 encadre l'accès par des établissements de santé et des fédérations hospitalières aux données du Programme de médicalisation des systèmes d'information (PMSI) et aux RPU (Résumé de passage aux urgences) mises à disposition sur la plateforme sécurisée de l'Agence technique de l'information sur l'hospitalisation (ATIH). Les responsables de traitement ont l'obligation de documenter les projets menés dans le registre des activités de traitement. Les études menées doivent présenter un caractère d'intérêt public et aucun appariement avec d'autres données à caractère personnel n'est autorisé. Les responsables de traitement doivent enregistrer leurs traitements auprès d'un répertoire public tenu par l'INDS.

TEXTE OFFICIEL

Délibération n° 2018-256 du 7 juin 2018

RESPONSABLES DE TRAITEMENT CONCERNES

- Les établissements de santé (publics ou privés, à but lucratif ou à but non lucratif);
- La Fédération hospitalière de France (FHF) ;
- La Fédération de l'hospitalisation privée (FHP) ;
- La Fédération des établissements hospitaliers et d'aide à la personne privés non lucratifs (FEHAP) ;
- La Fédération Unicancer ;
- La Fédération nationale des établissements d'hospitalisation à domicile (FNEHAD).

Le responsable de traitement désigne un délégué à la protection des données et tient à jour, au sein du registre des activités de traitement, la liste des études mises en œuvre dans le cadre de la méthodologie de référence.

OBJECTIF(S) POURSUIVI(S) PAR LE TRAITEMENT (FINALITES)

Le traitement des données doit présenter un caractère d'intérêt public.

Les finalités couvertes par la méthodologie de référence sont la planification et la valorisation de l'offre de soins ainsi que les études épidémiologiques et les études médico-économiques.

Une obligation de transparence incombe aux responsables de traitement. Elle se traduit par l'enregistrement de chaque étude conforme à la MR 005 dans un répertoire public tenu par l'INDS et accessible sur son site internet (www.indsante.fr). En pratique, un protocole (incluant la justification d'intérêt public), un résumé de l'étude selon un format arrêté par l'INDS et une déclaration des intérêts devront être soumis à l'Institut. Une fois l'étude terminée, les résultats obtenus devront aussi être communiqués dans un délai raisonnable conformément aux dispositions du code de la santé publique relatives au SNDS.

UTILISATION(S) EXCLUE(S) DU CHAMP DE LA NORME

La méthodologie de référence n'est pas applicable aux traitements :

- nécessitant un export des données à caractère personnel en dehors de la plateforme sécurisée ;
- nécessitant un appariement à des données à caractère personnel autres que celles mises à disposition par l'ATIH.

Au-delà de l'interdiction de réidentification des patients, deux finalités sont expressément interdites :

1. La promotion des produits de santé en direction des professionnels de santé ou d'établissements de santé ;
2. L'exclusion de garanties des contrats d'assurance et la modification de cotisations ou de primes d'assurance d'un individu ou d'un groupe d'individus présentant un même risque.

DONNEES PERSONNELLES CONCERNEES

Le responsable de traitement s'engage à ne traiter que les données pertinentes, adéquates et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées.

La nécessité de leur traitement doit être justifiée dans le protocole de recherche.

Les catégories de données à caractère personnel pouvant faire l'objet du traitement proviennent exclusivement des bases de données constituées par ATIH au titre du PMSI et des RPU. Ces données sont centralisées et mises à disposition sur la plateforme sécurisée de l'ATIH, en particulier :

- Les fichiers dans les champs de la médecine, la chirurgie, l'obstétrique et l'odontologie (MCO) ;
- Les fichiers concernant les soins de suite et de réadaptation (SSR) ;
- Le recueil d'Information Médicalisée en Psychiatrie (RIM-P) ;
- Les fichiers relatifs aux hospitalisations à domicile (HAD) ;
- Le fichier ANO qui permet de relier toutes les données du PMSI d'un même patient.

Par ailleurs, les données des résumés de passage aux urgences (RPU), mises à disposition par l'ATIH dans les mêmes conditions, sont également incluses dans le champ de la méthodologie de référence.

Les traitements inclus dans le cadre de la présente méthodologie de référence portent sur les données nationales du PMSI et des RPU dont la profondeur historique maximale est de neuf ans plus l'année en cours. La zone géographique concernée ainsi que la profondeur historique des données consultées sont justifiées dans le protocole.

DUREE DE CONSERVATION DES DONNEES

La durée d'accès aux données dans la plateforme sécurisée doit être limitée à la durée nécessaire à la mise en œuvre du traitement.

Lorsque le responsable de traitement en justifie, l'accès aux données peut être maintenu à l'issue de l'étude, dans la limite de deux ans à compter de la dernière publication relative aux résultats.

DESTINATAIRES DES DONNEES

Les données de l'ATIH sont mises à disposition du responsable de traitement sur une plateforme sécurisée. Aucune exportation de données à caractère personnel ne peut être réalisée.

Seul le personnel habilité par le responsable de traitement peut accéder aux données.

Ces personnes sont soumises au secret professionnel dans les conditions définies par les articles 226-13 et 226-14 du code pénal.

La qualification des personnes habilitées et leurs droits d'accès doivent être régulièrement réévalués, conformément aux modalités décrites dans la procédure d'habilitation établie par le responsable de traitement.

INFORMATION DES PERSONNES ET RESPECT DES DROITS "INFORMATIQUE ET LIBERTES"

La MR-005 n'impose pas d'information individuelle des personnes concernées. Cependant, elle prévoit que les responsables de traitement indiquent sur leur site internet qu'ils réalisent des projets à partir des données du PMSI et qu'ils rappellent que les personnes ont des droits d'accès, de rectification et d'opposition qui s'exercent auprès du directeur de l'organisme gestionnaire du régime d'assurance maladie obligatoire auquel elles sont rattachées.

SECURITE ET CONFIDENTIALITE

La mise en œuvre des traitements de données à caractère personnel intervenant dans le cadre de l'étude s'effectue sous la responsabilité du responsable du traitement, y compris chez des tiers agissant pour son compte, dans le respect du règlement général sur la protection des données et de l'arrêté du 22 mars 2017 relatif au référentiel de sécurité applicable au SNDS.

La sécurité des traitements est assurée par la mise à disposition des données par l'ATIH sur une plateforme sécurisée et homologuée au référentiel de sécurité applicable au SNDS.

Un espace de travail sur la plateforme est fourni par l'ATIH afin que les utilisateurs puissent consulter les données. Seules des statistiques agrégées de telle sorte que l'identification des personnes est impossible peuvent être extraites de la plateforme.

C.2 MR006

DECLARATION

20/02/2020

MR 06
Études nécessitant l’accès aux données du PMSI par les industriels de santé

ÉTUDES NÉCESSITANT L'ACCÈS AUX DONNÉES DU PMSI PAR LES INDUSTRIELS DE SANTÉ

(Déclaration N° 06)

La MR 006 encadre l'accès par des industriels de santé aux données du Programme de médicalisation des systèmes d'information (PMSI) de l'Agence technique de l'information sur l'hospitalisation (ATIH) mises à disposition via une solution sécurisée. Les responsables de traitement ont l'obligation de documenter les projets menés dans le registre des activités de traitement. Les études menées doivent présenter un caractère d'intérêt public et aucun appariement avec d'autres données à caractère personnel n'est possible. Les responsables de traitement doivent enregistrer leurs traitements auprès d'un répertoire public tenu par l'INDS. Les industriels devront recourir à un bureau d'études/laboratoires de recherches ayant réalisé un engagement de conformité au référentiel fixé par l'arrêté du 17 juillet 2017 auprès de la CNIL. Ils devront également faire réaliser un audit indépendant tous les 3 ans sur l'utilisation des données et le respect de l'interdiction des finalités interdites.

TEXTE OFFICIEL

Délibération n° 2018-257 du 7 juin 2018 portant homologation d'une méthodologie de référence relative aux traitements de données nécessitant l'accès pour le compte des personnes produisant ou commercialisant des produits mentionnés au II de l'article L. 5311-1 du code de la santé publique aux donnée

RESPONSABLES DE TRAITEMENT CONCERNES

Les personnes produisant ou commercialisant des produits mentionnés au II de l'article L. 5311-1 du Code de la santé publique (par exemple : laboratoires pharmaceutiques, fabricants de dispositifs médicaux etc.).

Le responsable de traitement désigne un délégué à la protection des données et tient à jour, au sein du registre des activités de traitement, la liste des études mises en œuvre dans le cadre de la méthodologie de référence.

OBJECTIF(S) POURSUIVI(S) PAR LE TRAITEMENT (FINALITES)

Le traitement des données doit présenter un caractère d'intérêt public.

La MR 006 pose une présomption d'intérêt public des finalités suivantes :

- La préparation de dossiers de discussions et réunions avec les autorités et comités compétents ;
- La réalisation d'études en conditions réelles d'utilisation à destination ou à la demande des autorités ;
- Le ciblage des centres et/ou la réalisation d'études de faisabilité dans le cadre d'une recherche impliquant ou n'impliquant pas la personne humaine ;
- La réalisation d'études dans le cadre de la vigilance et de la surveillance après commercialisation.

La MR 006 n'exclut pas la possibilité de réaliser des études poursuivant d'autres finalités : il appartient alors au responsable de traitement de justifier que ces études répondent à l'ensemble des obligations de la méthodologie de référence, notamment qu'elles présentent un caractère d'intérêt public.

Une obligation de transparence incombe aux responsables de traitement qui se traduit par l'enregistrement de chaque étude conforme à la MR 006 dans un répertoire public tenu par l'INDS et accessible sur son site internet (www.indsante.fr). En pratique, un protocole (incluant la justification d'intérêt public), un résumé de l'étude selon un format arrêté par l'INDS et une déclaration des intérêts devront être soumis à l'Institut. Une fois l'étude terminée, les résultats obtenus devront aussi être communiqués dans un délai raisonnable conformément aux dispositions du code de la santé publique relatives au SNDS.

UTILISATION(S) EXCLUE(S) DU CHAMP DE LA NORME

La méthodologie de référence n'est pas applicable aux traitements :

- nécessitant un export des données à caractère personnel en dehors de la solution sécurisée ;
- nécessitant un appariement à des données à caractère personnel autres que celles mises à disposition par l'ATIH.

Au-delà de l'interdiction de réidentification des patients, deux finalités sont expressément interdites :

1. La promotion des produits de santé en direction des professionnels de santé ou d'établissements de santé ;
2. L'exclusion de garanties des contrats d'assurance et la modification de cotisations ou de primes d'assurance d'un individu ou d'un groupe d'individus présentant un même risque.

Tous les trois ans, le responsable de traitement doit faire réaliser un audit externe indépendant en vue de s'assurer des finalités poursuivies et de l'utilisation des résultats des études réalisées. Le rapport d'audit devra être transmis au président du comité d'audit du SNDS, prévu à l'article 65 de la loi « informatique et libertés ».

DONNEES PERSONNELLES CONCERNEES

Le responsable de traitement s'engage à ne collecter que les données pertinentes, adéquates et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées.

La nécessité de leur traitement doit être justifiée dans le protocole de recherche.

Les données doivent provenir exclusivement des bases de données constituées par l'ATIH au titre du PMSI.

Les catégories de données à caractère personnel pouvant faire l'objet du traitement sont les données centralisées et mises à disposition sur la plateforme sécurisée de l'Agence technique de l'information sur l'hospitalisation (ATIH), en particulier :

- Les fichiers dans les champs de la médecine, la chirurgie, l'obstétrique et l'odontologie (MCO) ;
- Les fichiers concernant les soins de suite et de réadaptation (SSR) ;
- Le recueil d'Information Médicalisée en Psychiatrie (RIM-P) ;
- Les fichiers relatifs aux hospitalisations à domicile (HAD) ;
- Le fichier ANO qui permet de relier toutes les données du PMSI d'un même patient.

Les traitements inclus dans le cadre de la présente méthodologie de référence portent sur les données nationales du PMSI dont la profondeur historique maximale est de neuf ans plus l'année en cours. La zone géographique concernée ainsi que la profondeur historique des données consultées sont justifiées dans le protocole.

Les données de l'ATIH sont mises à disposition du responsable de traitement par l'intermédiaire d'une solution sécurisée (voir § « sécurité »).

DUREE DE CONSERVATION DES DONNEES

La durée d'accès aux données dans la plateforme sécurisée doit être limitée à la durée nécessaire à la mise en œuvre du traitement.

Lorsque la solution sécurisée est détenue par le laboratoire de recherche ou le bureau d'études, la durée d'accès aux données et la durée de conservation dans la solution sécurisée doivent être limitées à la durée nécessaire à la mise en œuvre du traitement.

Lorsque le responsable de traitement en justifie, l'accès aux données, et le cas échéant, leur conservation peuvent être maintenus à l'issue de l'étude, dans la limite de deux ans à compter de la dernière publication relative aux résultats.

DESTINATAIRES DES DONNEES

Les données de l'ATIH sont mises à disposition du laboratoire de recherche ou bureau d'études par l'intermédiaire d'une solution sécurisée. Aucune exportation de données à caractère personnel ne peut être réalisée en dehors de la solution sécurisée utilisée.

Seul le personnel du laboratoire de recherche et du bureau d'études peut accéder aux données.

Ces personnes sont soumises au secret professionnel dans les conditions définies par les articles 226-13 et 226-14 du code pénal.

La qualification des personnes habilitées et leurs droits d'accès doivent être régulièrement réévalués, par le laboratoire de recherche ou bureau d'études conformément aux modalités décrites dans la procédure d'habilitation qu'il a établie.

INFORMATION DES PERSONNES ET RESPECT DES DROITS "INFORMATIQUE ET LIBERTES"

La MR 006 n'impose pas d'information individuelle des personnes concernées. Cependant, elle prévoit que les responsables de traitement indiquent sur leur site internet qu'ils réalisent des projets à partir des données du PMSI et qu'ils rappellent que les personnes ont des droits d'accès, de rectification et d'opposition qui s'exercent auprès du directeur de l'organisme gestionnaire du régime d'assurance maladie obligatoire auquel elles sont rattachées.

SECURITE ET CONFIDENTIALITE

La mise en œuvre des traitements de données à caractère personnel intervenant dans le cadre de l'étude s'effectue sous la responsabilité du responsable du traitement, et du laboratoire de recherche ou bureau d'études agissant pour son compte, dans le respect des dispositions des articles 24, 25, 28, 32 à 35 du RGPD, ainsi que de l'arrêté du 22 mars 2017 relatif au référentiel de sécurité applicable au SNDS et de l'arrêté du 17 juillet 2017 relatif au référentiel déterminant les critères de confidentialité, d'expertise et d'indépendance pour les laboratoires de recherche et bureaux d'études.

Les systèmes mettant à disposition les données du PMSI doivent ainsi être conformes au référentiel de sécurité applicable au SNDS précité ; deux modalités de mise à disposition sont incluses dans le cadre de la méthodologie de référence :

- Les données sont mises à disposition du laboratoire de recherche ou bureau d'études par l'intermédiaire du prestataire d'accès sécurisé désigné par l'ATIH ; (actuellement : le CASD – Centre d'accès sécurisé aux données) ;
- Les données sont exportées vers un laboratoire de recherche ou un bureau d'études disposant d'une solution sécurisée et ayant conclu une convention avec l'ATIH (« bulles sécurisées »).

Bibliographie

[1] https ://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/strategie-nationale-de-sante/.

[2] https ://www.scansante.fr/sites/default/files/content/346/fiche_scansante_n8_rehospitalisation7j.p

[3] https ://www.solidarites-sante.gouv.fr/soins-et-maladies/qualite-des-soins-et-pratiques/qualite/les-indicateurs/article/les-indicateurs-de-rehospitalisation-et-de-coordination.

[4] https ://www.scansante.fr/sites/default/files/content/396/notice_taux_rh30.pdf.

[5] McIlvennan Colleen K. , Eapen Zubin J. , Allen Larry A. . Hospital Readmissions Reduction Program. *Circulation.* 2015 ;131(20) :1796–1803.

[6] Abdul-Aziz Ahmad A., Hayward Rodney A., Aaronson Keith D., Hummel Scott L.. Association Between Medicare Hospital Readmission Penalties and 30-Day Combined Excess Readmission and Mortality. *JAMA cardiology.* 2017 ;2(2) :200–203.

[7] Babayan Zaruhi V., McNamara Robert L., Nagajothi Nagaprasad, et al. Predictors of cause-specific hospital readmission in patients with heart failure. *Clinical Cardiology.* 2003 ;26(9) :411–418.

[8] Joynt Karen E., Figueroa Jose E., Oray John, Jha Ashish K.. Opinions on the Hospital Readmission Reduction Program : results of a national survey of hospital leaders. *The American Journal of Managed Care.* 2016 ;22(8) :e287–294.

[9] Lemieux Jeff, Sennett Carry, Wang Ray, Mulligan Teresa, Bumbaugh Jon. Hospital readmission rates in Medicare Advantage plans. *The American Journal of Managed Care.* 2012 ;18(2) :96–104.

[10] Almussallam Basem, Joyce Maurice, Marcello Peter W., et al. What Factors Predict Hospital Readmission after Colorectal Surgery ?. *The American Surgeon.* 2016 ;82(5) :433–438.

[11] Abraham Christa R., Werter Christopher R., Ata Ashar, et al. Predictors of Hospital Readmission after Bariatric Surgery. *Journal of the American College of Surgeons.* 2015 ;221(1) :220–227.

[12] Shameer Khader, Johnson Kipp W., Yahi Alexandre, et al. PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING : A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.* 2017 ;22 :276–287.

[13] Lee Kyu Ha, Haneuse Sebastien, Schrag Deborah, Dominici Francesca. Bayesian Semiparametric Analysis of Semi-competing Risks Data : Investigating Hospital Readmission after a Pancreatic Cancer Diagnosis. *Journal of the Royal Statistical Society. Series C, Applied Statistics.* 2015 ;64(2) :253–273.

[14] Urach Christoph, Zauner Günther, Wahlbeck Kristian, Haaramo Peija, Popper Niki. Statistical methods and modelling techniques for analysing hospital readmission of discharged psychiatric patients : a systematic literature review. *BMC Psychiatry.* 2016 ;16(1) :413.

[15] Bartolomeo Nicola, Trerotoli Paolo, Moretti Annamaria, Serio Gabriella. A Markov model to evaluate hospital readmission. *BMC medical research methodology.* 2008 ;8 :23.

[16] Murphy Patrick B., Rehal Sunita, Arbane Gill, et al. Effect of Home Noninvasive Ventilation With Oxygen Therapy vs Oxygen Therapy Alone on Hospital Readmission or Death After an Acute COPD Exacerbation : A Randomized Clinical Trial. *JAMA.* 2017 ;317(21) :2177–2186.

[17] Arai Takahide, Yashima Fumiaki, Yanagisawa Ryo, et al. Hospital readmission following transcatheter aortic valve implantation in the real world. *International Journal of Cardiology.* 2018 ;269 :56–60.

[18] Law Amy, Cyhaniuk Anissa, Krebs Blake. Comparison of health care costs and hospital readmission rates associated with negative pressure wound therapies. *Wounds : A Compendium of Clinical Research and Practice.* 2015 ;27(3) :63–72.

[19] Cox James C., Sadiraj Vjollca, Schnier Kurt E., Sweeney John F.. Incentivizing Cost-Effective Reductions in Hospital Readmission Rates. *Journal of Economic Behavior & Organization.* 2016 ;131(B) :24–35.

[20] Fischer Claudia, Lingsma Hester F., Mheen Perla J. Marang-van de, Kringos Dionne S., Klazinga Niek S., Steyerberg Ewout W.. Is the Readmission Rate a Valid Quality Indicator ? A Review of the Evidence. *PLOS ONE.* 2014 ;9(11) :e112282.

[21] *PMSI \textbackslashtextbar Fédération hospitalière de France.*

[22] Yadav Chander Prakash, V Sreenivas, Ma Khan, Rm Pandey. An Overview of Statistical Models for Recurrent Events Analysis : A Review. *Epidemiology : Open Access.* 2018 ;08(04).

[23] Yang Wei, Jepson Christopher, Xie Dawei, et al. Statistical Methods for Recurrent Event Analysis in Cohort Studies of CKD. *Clinical Journal of the American Society of Nephrology : CJASN.* 2017 ;12(12) :2066–2073.

[24] Amorim Leila DAF, Cai Jianwen. Modelling recurrent events : a tutorial for analysis in epidemiology. *International Journal of Epidemiology.* 2015 ;44(1) :324–333.

[25] Olson Daiwai M., Cox Margueritte, Pan Wenqin, et al. Death and rehospitalization after transient ischemic attack or acute ischemic stroke : one-year outcomes from the adherence evaluation of acute ischemic stroke-longitudinal registry. *Journal of Stroke and Cerebrovascular Diseases : The Official Journal of National Stroke Association.* 2013 ;22(7) :e181–188.

[26] Pacho Cristina, Domingo Mar, Núñez Raquel, et al. Predictive biomarkers for death and rehospitalization in comorbid frail elderly heart failure patients. *BMC geriatrics.* 2018 ;18(1) :109.

[27] Koller Michael T, Raatz Heike, Steyerberg Ewout W, Wolbers Marcel. Competing risks and the clinical community : irrelevance or ignorance ?. *Statistics in Medicine.* 2012 ;31(11-12) :1089–1097.

[28] D'actuaires Congrès International. *Septième Congrès International D'actuaires, Vol. 1 : Sous Le Haut Patronage De S. A. R. Le Prince des Pays-Bas, Duc De Mecklembourg ; Amsterdam, 2 Au 7 ... Thèmes A Discuter.* Forgotten Books ; 2018.

[29] Kaplan E. L., Meier Paul. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association.* 1958 ;53(282) :457.

[30] Brookmeyer Ron, Crowley John. A k-Sample Median Test for Censored Data. *Journal of the American Statistical Association.* 1982 ;77(378) :433–440.

[31] Chen Zhongxue. A nonparametric approach to detecting the difference of survival medians. *Communications in Statistics - Simulation and Computation.* 2017 ;46(1) :395–403.

[32] Chen Zhongxue. Extension of Mood's median test for survival data. *Statistics & Probability Letters.* 2014 ;95 :77–84.

[33] Tang Shaowu, Jeong Jong-Hyeon. Median Tests for Censored Survival Data ; a Contingency Table Approach. *Biometrics.* 2012 ;68(3) :983–989.

[34] Mantel N.. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports.* 1966 ;50(3) :163–170.

[35] Mantel Nathan, Haenszel William. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI : Journal of the National Cancer Institute.* 1959 ;22(4) :719–748.

[36] Gehan Edmund A.. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. *Biometrika.* 1965 ;52(1/2) :203–223.

[37] Breslow Norman. A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship. *Biometrika.* 1970 ;57(3) :579–594.

[38] Peto Richard, Peto Julian. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General).* 1972 ;135(2) :185–207.

[39] Prentice R. L.. Linear Rank Tests with Right Censored Data. *Biometrika.* 1978 ;65(1) :167–179.

[40] Wei L. J.. The accelerated failure time model : A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine.* 1992 ;11 :1871–1879.

[41] Weibull Waloddi, Sweden Sics. A Statistical Distribution Function of Wide Applicability. In : ; 1951.

[42] Cox D. R.. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological).* 1972 ;34(2) :187–220.

[43] Kleinbaum David G, Klein Mitchel. *Survival analysis.* Springer ; 2010.

[44] Schoenfeld David. Partial residuals for the proportional hazards regression model. *Biometrika.* 1982 ;69(1) :239–241.

[45] Andersen P. K., Gill R. D.. Cox's Regression Model for Counting Processes : A Large Sample Study. *The Annals of Statistics.* 1982 ;10(4) :1100–1120.

[46] Prentice R. L., Williams B. J., Peterson A. V.. On the regression analysis of multivariate failure time data. *Biometrika.* 1981 ;68(2) :373–379.

[47] Quenouille Maurice H. Approximate tests of correlation in time-series 3. In : :483–484Cambridge University Press ; 1949.

[48] Lee Eric W., Wei L. J., Amato David A., Leurgans Sue. Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations. In : Klein John P., Goel Prem K., eds. *Survival Analysis : State of the Art*, Nato Science. Dordrecht : Springer Netherlands 1992 (pp. 237–247).

[49] Wei L. J., Lin D. Y., Weissfeld L.. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association.* 1989 ;84(408) :1065–1073.

[50] Liang Kung-Yee, Zeger Scott L.. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986 ;73(1) :13–22.

[51] Vaupel J. W., Manton K. G., Stallard E.. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography.* 1979 ;16(3) :439–454.

[52] Wienke A. *Frailty Models in Survival Analysis.* Chapman and Hall CRC ; 2009.

[53] Pénichoux Juliette, Moreau Thierry, Latouche Aurélien. Simulating recurrent events that mimic actual data : a review of the literature with emphasis on event-dependence 2014.

[54] Jiménez Fernando, Jodrá Pedro. A Note on the Moments and Computer Generation of the Shifted Gompertz Distribution. *Communications in Statistics - Theory and Methods.* 2008 ;38(1) :75-89.

[55] Miloslavsky Maja, Keleş Sündüz, Laan Mark J., Butler Steve. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society : Series B (Statistical Methodology).* 2004 ;66(1) :239-257.

[56] Ghosh Debashis, Lin D. Y.. Marginal regression models for recurrent and terminal events. *Stat Sin.* 2002 ;12(3) :663–688.

[57] Huang Chiung-Yu, Wang Mei-Cheng. Joint Modeling and Estimation for Recurrent Event Processes and Failure Time Data. *J Am Stat Assoc.* 2004 ;99(468) :1153–1165.

[58] Liu Lei, Wolfe Robert A., Huang Xuelin. Shared Frailty Models for Recurrent Events and a Terminal Event. *Biometrics.* 2004 ;60(3) :747–756.

[59] Rondeau Virginie, Mathoulin-Pelissier Simone, Jacqmin-Gadda Hélène, Brouste Véronique, Soubeyran Pierre. Joint frailty models for recurring events and death using maximum penalized likelihood estimation : application on cancer events. *Biostatistics.* 2007 ;8(4) :708–721.

[60] Mao Lu, Lin D. Y.. Semiparametric regression for the weighted composite endpoint of recurrent and terminal events. *Biostatistics.* 2015 ; :kxv050.

[61] Zeng Donglin, Ibrahim Joseph G, Chen Ming-Hui, Hu Kuolung, Jia Catherine. Multivariate recurrent events in the presence of multivariate informative censoring with applications to bleeding and transfusion events in myelodysplastic syndrome. *Journal of biopharmaceutical statistics.* 2014 ;24(2) :429–442.

[62] *Déroulement de la greffe - Greffe de foie.*

[63] *Foie Greffes transplantation - Transplantation hépatique : une greffe bien maîtrisée - Doctissimo.*

[64] Runge Carl. Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten. *Zeitschrift für Mathematik und Physik.* 1901 ;46(224-243) :20.

[65] Ramsay James O, others . Monotone regression splines in action. *Statistical science.* 1988 ;3(4) :425–441.

[66] Joly Pierre. Estimation de la fonction de risque dans un contexte général de troncature et de censure : application à l'estimation de l'incidence de la démence. PhD thesisBordeaux 21996.

[67] Mazroui Yassin, Mathoulin-Pelissier Simone, Soubeyran Pierre, Rondeau Virginie. General joint frailty model for recurrent event data with a dependent terminal event : application to follicular lymphoma data. *Statistics in medicine.* 2012 ;31(11-12) :1162–1176.

[68] Duin Robert P. W.. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers.* 1976 ;(11) :1175–1179.

[69] Habbema JDF, Hermans J, Broek K. A stepwise discrimination program using density estimation. In : :100–110Physica Verlag Vienna ; 1974.

[70] O'Sullivan Finbarr. Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators. *SIAM Journal on Scientific and Statistical Computing.* 1988 ;9(2) :363-379.

[71] Broyden Charles George. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics.* 1970 ;6(1) :76–90.

[72] Fletcher Roger. A new approach to variable metric algorithms. *The computer journal.* 1970 ;13(3) :317–322.

[73] Goldfarb Donald. A family of variable-metric methods derived by variational means. *Mathematics of computation.* 1970 ;24(109) :23–26.

[74] Shanno David F. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation.* 1970 ;24(111) :647–656.

[75] Dennis Jr John E, Gay David M, Walsh Roy E. An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software (TOMS).* 1981 ;7(3) :348–368.

[76] Rondeau Virginie, Schaffner Emmanuel, Corbiere Fabien, Gonzalez Juan R, Mathoulin-Pélissier Simone. Cure frailty models for survival data : application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Statistical methods in medical research.* 2013 ;22(3) :243–260.

[77] Marquardt Donald W. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics.* 1963 ;11(2) :431–441.

[78] Liquet Benoit, Commenges Daniel. Estimating the expectation of the log-likelihood with censored data for estimator selection. *Lifetime Data Analysis.* 2004 ;10(4) :351–367.

[79] Commenges Daniel, Joly Pierre, Gégout-Petit Anne, Liquet Benoit. Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics.* 2007 ;34(1) :33–52.

[80] Fayyad Usama., Piatetsky-Shapiro Gregory, Smyth Padhraic. From data mining to knowledge discovery in databases. *AI Mag.* 1996 ;17(3) :37–54.

[81] McCulloch W.S., Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics.* 1943 ; :115–133.

[82] Hoerl Arthur E., Kennard Robert W.. Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics.* 1970 ;12(1) :55-67.

[83] Tibshirani Robert. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological).* 1996 ;58(1) :267–288.

[84] Day W.H.E., Edelsbrunner H.. Efficient algorithms for agglomerative hierarchical clustering methods.. *Journal of Classification.* 1984 ;1(1) :7–24.

[85] Hartigan J. A., Wong M. A.. Algorithm AS 136 : A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics).* 1979 ;28(1) :100–108.

[86] Banfield Jeffrey D., Raftery Adrian E.. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics.* 1993 ;49(3) :803–821.

[87] https ://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[88] Hall Kevin D, Kahan Scott. Maintenance of lost weight and long-term management of obesity. *Medical Clinics.* 2018 ;102(1) :183–197.

[89] Alexopoulos Sophoclis P, Matsuoka Lea, Hafberg Einar, et al. Liver Transplantation for Propionic Acidemia : A Multicenter-linked Database Analysis. *Journal of pediatric gastroenterology and nutrition.* 2020 ;70(2) :178–182.

[90] Hasan Shaakir, Abel Stephen, Uemura Tadahiro, et al. Liver transplant mortality and morbidity following preoperative radiotherapy for hepatocellular carcinoma. *HPB.* 2019 ;.

[91] Yi Ren Gabriel L. Zenarosa Heather E. Tomko Drew Michael S. Donnell Hyung-joo Kang Mark S. Roberts, Bryce Cindy L.. Gray's Time-Varying Coefficients Model for Post-transplant Survival of Pediatric Liver Transplant Recipients with a Diagnosis of Cancer. *Computational and Mathematical Methods in Medicine.* 2013 ;.

[92] Bender Ralf, Augustin Thomas, Blettner Maria. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005 ;24(11) :1713–1723.

[93] Jahn-Eimermacher Antje, Ingel Katharina, Ozga Ann-Kathrin, Preussler Stella, Binder Harald. Simulating recurrent event data with hazard functions defined on a total time scale. *BMC Medical Research Methodology.* 2015 ;15.

[94] Jessup Mariell , Greenberg Barry , Mancini Donna , et al. Calcium Upregulation by Percutaneous Administration of Gene Therapy in Cardiac Disease (CUPID). *Circulation.* 2011 ;124(3) :304–313.

[95] Kianifard F., Gallo P. P.. Poisson regression analysis in clinical research. *Journal of Biopharmaceutical Statistics.* 1995 ;5(1) :115–129.

[96] Ismail Noriszura, Jemain Abdul Aziz. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. 2007 ; :56.

[97] Kennedy Byron S., Kasl Stanislav V., Vaccarino Viola. Repeated Hospitalizations and Self-rated Health among the Elderly : A Multivariate Failure Time Analysis. *American Journal of Epidemiology.* 2001 ;153(3) :232–241.

[98] Kennedy Byron S.. Does race predict stroke readmission ? An analysis using the truncated negative binomial model. *Journal of the National Medical Association.* 2005 ;97(5) :699–713.

[99] House Chad M., Anstadt Mary A., Stuck Logan H., Nelson William B.. The Association Between Cardiac Rehabilitation Attendance and Hospital Readmission. *American Journal of Lifestyle Medicine.* 2018 ;12(6) :513–520.

[100] Wang Hao, Johnson Carol, Robinson Richard D., et al. Roles of disease severity and post-discharge outpatient visits as predictors of hospital readmissions. *BMC Health Services Research.* 2016 ;16.

[101] Seraj Siamak M, Campbell Emily J, Argyropoulos Sarah K, Wegermann Kara, Chung Raymond T, Richter James M. Hospital readmissions in decompensated cirrhotics : Factors pointing toward a prevention strategy. *World Journal of Gastroenterology.* 2017 ;23(37) :6868–6876.

[102] Seibt Silvia, Gilchrist Catherine A., Reed Peter W., et al. Hospital readmissions with acute infectious diseases in New Zealand children \textbackslashtextless 2 years of age. *BMC Pediatrics.* 2018 ;18.

[103] *CIM-10 Version :2008.*

[104] *WHO \textbackslashtextbar International Classification of Diseases, 11th Revision (ICD-11).*

[105] *CIM-10 FR 2019 à usage PMSI \textbackslashtextbar Publication ATIH.*

[106] Amin Ahmad, Chitsazan Mitra, Shiukhi Ahmad Abad Fatemeh, Taghavi Sepideh, Naderi Nasim. On admission serum sodium and uric acid levels predict 30 day rehospitalization or death in patients with acute decompensated heart failure. *ESC heart failure.* 2017 ;4(2) :162–168.

[107] *Foie - Anatomie, Physiologie, Pathologies, Soin.* 2016.

[108] work(s) : Karl Pearson Reviewed. Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society of London. A.* 1895 ;18 :343– 414.

[109] James Ian R.. Tests for location with k samples and censored data. *Biometrika.* 1987 ;74(3) :599–607.

[110] Fischer Claudia, Lingsma Hester F., Mheen Perla J., Kringos Dionne S., Klazinga Niek S., Steyerberg Ewout W.. Is the Readmission Rate a Valid Quality Indicator ? A Review of the Evidence. *PLoS ONE.* 2014 ;9(11).

[111] Milne R, Clarke A. Can readmission rates be used as an outcome indicator ?. *BMJ : British Medical Journal.* 1990 ;301(6761) :1139–1140.

[112] Clarke A.. Readmission to hospital : a measure of quality or outcome ?. *BMJ Quality & Safety.* 2004 ;13(1) :10–11.

[113] Fetter Robert B., Freeman Jean L.. Diagnosis Related Groups : Product Line Management within Hospitals. *The Academy of Management Review.* 1986 ;11(1) :41–54.

[114] Fetter Robert B., Brand Donald A.. *DRGs : their design and development.* Health Administration Press ; 1991. Google-Books-ID : fV4QAQAAMAAJ.

[115] Fayyad Usama., Piatetsky-Shapiro Gregory, Smyth Padhraic. From data mining to knowledge discovery in databases. *AI Mag.* 1996 ;17(3) :37–54.

[116] Rondeau Virginie, Mazroui Yassin, Gonzalez Juan. frailtypack : An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. *J Stat Softw.* 2012 ;47(1) :1–28.

[117] C. Gini. Measurement of inequality of income. *the economic journal.* 1921 ; :22-43.

[118] Pénichoux Juliette, Moreau Thierry, Latouche Aurélien. Simulating recurrent events that mimic actual data : a review of the literature with emphasis on event-dependence. *arXiv preprint arXiv :1503.05798.* 2015 ;.