



UNIVERSITÉ PARIS CITÉ

École Doctorale Pierre Louis de Santé Publique - ED393

Centre de Recherche Inria Paris, HeKA

Centre de Recherche des Cordeliers, UMR-S 1138 Inserm

Prédiction d'événements récurrents en survie : développement de méthode d'apprentissage et application en oncologie

par Juliette MURRIS

Thèse de doctorat de Biostatistiques et Biomathématiques

Dirigée par Pr. Sandrine KATSAHIAN

Et co-dirigée par Audrey LAVENU

Présentée et soutenue publiquement le 18/10/2024

Devant un jury composé de :

Catherine LEGRAND	Prof., Université catholique de Louvain	Rapportrice
Robin GENUER	MCU-HDR, Université de Bordeaux	Rapporteur
Cécile PROUST-LIMA	DR, Université de Bordeaux	Examinateuse
Viet-Thi TRAN	PU-PH, Université Paris Cité	Examinateur
Sandrine KATSAHIAN	PU-PH, Université Paris Cité	Directrice de thèse
Audrey LAVENU	MCU-HDR, Université de Rennes	Co-directrice de thèse
Guillaume DESACHY	Pierre Fabre	Membre invité
Abir TADMOURI-SELLIER	Pierre Fabre	Membre invitée

Résumé

Titre : Prédiction d'événements récurrents en survie : développement de méthode d'apprentissage et application en oncologie

Mots clefs : Survie ; Événements récurrents ; Apprentissage ; Forêts aléatoires

Résumé : En oncologie, la rechute de cancer ou encore la progression de la tumeur sont souvent utilisées pour mesurer l'effet d'un traitement ou d'une intervention. Les méthodes classiques visent à modéliser le temps d'apparition de la première rechute, ou de la première progression, comme l'estimateur de Kaplan-Meier ou le modèle de Cox. Or, un événement clinique d'intérêt peut se produire plusieurs fois pour un patient. L'analyse basée sur le temps jusqu'au premier événement ne peut pas être utilisée pour examiner l'effet des facteurs de risque sur le nombre de récidives au fil du temps. Il existe alors de nombreuses méthodes d'analyse des événements récurrents, soit conditionnelles soit marginales. Pour répondre à ce besoin, plusieurs modèles statistiques ont été développés, mais un consensus sur les approches d'apprentissage pour les données à haute dimension demeure insatisfaisant. Dans le cadre de ce travail de thèse, une revue systématique de la littérature a été réalisée pour synthétiser les méthodologies les plus avancées et comparer entre elles les méthodes existantes, révélant ainsi un manque important dans la modélisation des événements récurrents avec des techniques d'apprentissage automatique. Ce projet de recherche introduit une version améliorée de l'algorithme des forêts de survie aléatoires, spécialement conçue pour les données d'événements récurrents. Cette méthode

combine les approches statistiques non paramétriques et semi-paramétriques avec le principe des méthodes d'ensemble. Nous avons adapté les forêts de survie aléatoires en modifiant les règles de division à chaque nœud et les estimations des nœuds terminaux afin de prendre en compte les événements récurrents. L'objectif de cette méthode, appelée RecForest est d'estimer le nombre attendu d'événements récurrents pour chaque individu au fil du temps. Nous avons également développé des métriques de performance adaptées, notamment des extensions de l'indice de concordance et de l'erreur quadratique moyenne, qui sont applicables à divers modèles de survie traitant des événements récurrents. De plus, cette thèse explore le rôle essentiel de l'interprétabilité et de l'explicabilité des algorithmes, en particulier lorsqu'ils sont intégrés dans des dispositifs médicaux. Nous formulons des recommandations pour garantir la conformité avec les normes établies par les autorités de santé, en soulignant la nécessité de disposer de dispositifs médicaux basés sur l'IA qui soient transparents et responsables. Ces propositions sont destinées à améliorer l'évaluation de la performance des algorithmes, en favorisant la confiance et la fiabilité des outils de prise de décision médicale. Enfin, ce travail recherche offre des horizons prometteurs pour améliorer les prédictions de survie et la gestion des événements récurrents en oncologie.

Title : Predicting recurrent events in a survival framework : development of a machine learning approach and an application in oncology

Keywords : Survival analysis ; Recurrent events ; Machine learning ; Random forests

Abstract : In oncology, cancer relapse or tumour progression are often used to assess treatment effect. Traditional methods aim to model the time to the first relapse or progression, such as the Kaplan and Meier estimator or the Cox model. However, a clinical event of interest may occur several times for a patient. Analysis based on the time to the first event is not suitable anymore for examining the effect of risk factors on the number of recurrences over time. Many statistical methods therefore exist for analysing recurrent events, either conditional or marginal. To address this, several statistical models have been developed, yet a consensus on learning approaches for high-dimensional data remains elusive. A systematic literature review was conducted to synthesize state-of-the-art methodologies and compare existing methods, revealing a significant gap in modeling recurrent events with machine learning techniques. In response, this research project introduces an enhanced version of the Random Survival Forest algorithm, specifically designed for recurrent event data.

This extension, named RecForest, leverages survival analysis principles and ensemble learning. The splitting rule has been refined to accommodate recurrent events, and the ensemble estimate is derived by aggregating the expected number of events. We have also developed pertinent performance metrics, including extensions of the concordance index and mean square error, which are applicable across various survival models handling recurrent events. Furthermore, this thesis addresses the critical need for interpretability and explainability in algorithms, particularly when integrated into medical devices. Recommendations are provided to ensure compliance with health authority standards, underscoring the necessity for transparent and accountable AI-based medical devices. These guidelines aim to enhance the evaluation of algorithmic performance, fostering trust and reliability in medical decision-making tools. In conclusion, this PhD project offers promising avenues for improving survival predictions and managing recurrent events in oncology.

*L'espérance,
c'est l'idée de croire
que ce que nous faisons
peut avoir une importance.*

Rebecca Solnit

A Thibault,

Remerciements

En premier lieu, je tiens à remercier mes rapporteurs, Pr Catherine Legrand et Robin Genuer, d'avoir accepté d'évaluer cette thèse. Je remercie également Cécile Proust-Lima et Pr Viet-Thi Tran d'avoir accepté d'être les examinateurs de cette thèse. Je mesure pleinement la chance que j'ai de bénéficier de l'expertise et du temps de chacun d'entre vous.

Je remercie mes directrices de thèse pour leur confiance et leur soutien tout au long de cette thèse. Je remercie Audrey pour sa patience et sa présence à chacune de mes communications pour me soutenir. Sandrine, merci de m'avoir donné ce goût de la recherche, de l'entraide et des défis, et aussi de m'avoir fait découvrir le merveilleux monde de l'enseignement. J'en profite pour adresser ici un message de remerciement à Yann de Rycke pour m'avoir présentée à Sandrine en 2020 et introduite au monde des données censurées.

Je souhaite exprimer ma profonde gratitude à Pr Anita Burgun et à Sarah Zohar pour leur accueil, initialement au sein de l'équipe 1138 du Centre de recherche des Cordeliers, puis au sein de la dynamique équipe HeKA à l'Inria. Mes remerciements s'étendent également à Meriem Guemair pour sa bienveillance, à Philippe Gesnouin pour son optimisme à l'occasion de chaque nouveau séminaire que je lui ai proposé d'organiser, ainsi qu'à Sabrina Deveaux pour sa bonne humeur contagieuse et son aide précieuse dans la préparation de ma soutenance. De plus, je ne manquerai pas de témoigner ma reconnaissance à Isabelle Ricard, gestionnaire UPC dédiée à notre équipe.

Ce projet de thèse Cifre n'aurait jamais vu le jour sans Abir Tadmouri-Sellier, dont je suis infiniment reconnaissante pour sa confiance et sa bienveillance. Bien qu'à temps partiel auprès des équipes Pierre Fabre, je garde des souvenirs émus des moments avec notre équipe RWE de choc, et en particulier avec les reines de la stat Claire Castagné, Olivia Dialla et Florence Carrère. Michal, many thanks for all the support you provided, both computationally and mentally speaking. J'adresse également un message particulier à Guillaume Desachy, que je remercie pour sa confiance, son enthousiasme et sa bonne humeur constante.

Les jeunes chercheurs et chercheuses de l'équipe HeKA ont égayé mon quotidien, et je tiens à les remercier chaleureusement pour tous les agréables moments que nous avons partagés. Je souhaite adresser des remerciements particuliers à Alice, Emma, Enora, Judith, Linus, Lucas et Stylianos. Je saisie également cette occasion pour exprimer ma gratitude à toutes les personnes qui ont pris part avec enthousiasme aux séminaires juniors HeKA. Un grand merci pour leur indulgence à l'égard de mes fonctions d'organisatrice et de *chairwoman*.

"La recherche, c'est avant tout une aventure collective." Ce mantra a véritablement guidé l'ensemble de cette thèse, où chaque pas en avant a été marqué par un échange fructueux.

Ces quelques années de recherche ont ainsi été ponctuées par des collaborations stimulantes, qui m'ont permis non seulement d'en apprendre beaucoup sur des pathologies spécifiques, mais aussi de m'évader un peu du cadre théorique de mon activité. En particulier, je remercie Line Farah pour sa magnifique introduction au monde des dispositifs médicaux, accompagnée de Nicolas Martelli. Mes remerciements s'adressent également à Dr Marie-Caroline Dieu-Nosjean, à Dr Olivier Espitia, à Dr Bernard Denis, au duo Dr Raphaël Degrave et Pr Jean-Benoît Arlet, à Pr Jean-Luc Diehl, à Dr Nicolas Bréchot, à Dr Juliette Didier, à Dr Coumba Sow, à Dr Stéphanie Baron et à Dr Frédérique Schortgen pour m'avoir fait découvrir les domaines passionnants de l'immunologie, de la médecine interne, de la gastro-entérologie et de la réanimation. Enfin, je souhaite exprimer mon immense gratitude à Olivier Bouaziz, dont les outils en biostatistique sont indispensables à toutes les personnes travaillant dans ce domaine, et dont l'aide et la rigueur ont été d'une valeur inestimable pour ce projet de thèse.

Le déroulement de ma thèse n'aurait pas été aussi agréable sans la bonne humeur et la bienveillance de l'unité de recherche clinique de l'Hôpital Européen Georges Pompidou. Je suis particulièrement et profondément reconnaissante envers Anaïs Charles-Nelson, Dr Anne-Isabelle Tropeano, et Armelle Arnoux pour leur soutien indéfectible dès les premiers instants. Leur encouragement a été d'une importance capitale tout au long de ce parcours. J'espère très sincèrement que nous aurons l'opportunité de travailler à nouveau ensemble dans le futur. Merci pour tout.

Le temps consacré à la thèse est précieux, mais celui qui reste une fois les machines éteintes l'est encore davantage. Je tiens à exprimer ma profonde gratitude à mes ami·es. Je suis entourée de personnes extraordinaires, et je n'oublierai jamais leur amour, leur soutien inconditionnel et les innombrables moments de rire partagés ensemble. Merci du fond du cœur pour votre amitié et votre amour.

Je tiens également à exprimer ma plus profonde reconnaissance envers ma famille. Leur présence constante et leurs encouragements quotidiens ont été une source inépuisable de réconfort et de motivation. Je suis profondément reconnaissante de la vie de les avoir à mes côtés, maintenant et pour toujours.

Enfin, un immense merci à mon compagnon de route de rendre la vie si belle et si facile à aimer.

Valorisation scientifique

En lien avec la thèse

Articles de journaux

- **Murris, J.**, Charles-Nelson, A., Tadmouri Sellier, A., Lavenu, A., & Katsahian, S. (2023). Towards filling the gaps around recurrent events in high dimensional framework : a systematic literature review and application. *Biostatistics & Epidemiology*, 7(1), e2283650.
- Farah, L.* **Murris, J.***, Borget, I., Guilloux, A., Martelli, N., & Katsahian, S. (2023). Assessment of Performance, Interpretability, and Explainability in Artificial IntelligenceBased Health Technologies : What Healthcare Stakeholders Need to Know. *Mayo Clinic Proceedings : Digital Health*, 1(2), 120-138.
- **Murris, J.**, Bouaziz, O., Jakubczak, M., Katsahian, S., & Lavenu, A. (2024). Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event. *Soumis à BMC Medical Research Methodology*.
- **Murris, J.**, Desachy, G., Lavenu, A., & Katsahian, S. (2024). Random survival forest for right-censored data with recurrent events : The RecForest R package. *En préparation*.
- **Murris, J.**, Tzedakis, S., Desachy, G., Lavenu, A., & Katsahian, S. (2024). Predicting hospital readmission after digestive cancer surgery with survival analysis and machine learning. *En préparation*.

Articles de conférence

- **Murris, J.**, Alquier, M., Cassant, E., Simoneau, M., Bhan, M., & Katsahian, S. (2024). Bridging Interpretability and Survival Endpoints in Health Technology Assessment. *Soumis à Artificial Intelligence, Ethics, and Society 2024*.

Communications orales

- **Murris, J.**, Lavenu, A., & Katsahian, S., Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event, *Communication invitée à l'Unité de Recherche Clinique, Hôpital Européen Georges Pompidou (AP-HP)*, 2024.
- **Murris, J.**, Lavenu, A., & Katsahian, S., Random survival forests for analysing survival data with recurrent events, *Survival Analysis for Junior Researchers, Ulm, Allemagne*, 2023.

-
- **Murris, J.**, Lavenu, A., & Katsahian, S., Random survival forests for analysing survival data with recurrent events, *44th Annual Conference of the International Society for Clinical Biostatistics, Milan, Italie, 2023*.
 - **Murris, J.**, Lavenu, A., & Katsahian, S., Decision trees for analyzing survival data with recurrent events, *54e Journées de Statistique, Bruxelles, Belgique, 2023*.
 - **Murris, J.**, Lavenu, A., & Katsahian, S., Introducing recurrent events in high-dimensional frameworks, *Journées de Biostatistique, Rennes, France, 2022*.

Posters

- **Murris J.**, Katsahian, S., & Lavenu, A. (oratrice), Predicting Hospital Readmission after Cancer Surgery with Survival Analysis and Machine Learning, *1st International Conference on Artificial Intelligence in Healthcare 2024*.
- **Murris J.**, Lavenu, A., & Katsahian, S., RecForest : Forêts aléatoires de survie pour l'analyse des événements récurrents en R. *10e Rencontres R, Vannes, France, 2024 (Prix du meilleur poster)*.
- **Murris, J.**, Lavenu, A., & Katsahian, S., Improving survival risk prediction with random survival forests for recurrent events in biological systems, *31st Annual Intelligent Systems For Molecular Biology and 22nd Annual European Conference on Computational Biology, Lyon, France, 2023*.
- Farah, L.* & **Murris, J.***, Key notions in health technology assessment of AI-based medical devices : what healthcare stakeholders need to know, *9th Statistics & Biopharmacy Conference, Paris, France, 2022*.
- **Murris, J.**, Lavenu, A., & Katsahian, S., Towards filling the gap around recurrent events in high dimensional framework : literature review and early comparison, *9th Statistics & Biopharmacy Conference, Paris, France, 2022*.

Collaborations

Articles de journaux

- Deyrat, J., Fuks, D., **Murris, J.**, Lanoy, E., Nassar, A., Dhote, A., ... & Tzedakis, S. (2024). Evolution of laparoscopic liver surgery in France over the last decade. *HPB, 26, S69*.
- Poisson, J., El-Sissy, C., Serret, A., Smith, N., Lebraud, M., ... **Murris, J.**, ... & Granier, C. (2024). Increased levels of GM-CSF and CXCL10 and low CD8+ Memory Stem T Cell count are markers of immune ageing and severe COVID-19 in older people. *Immunity and ageing, 21(1), 1-14*.
- Guénégou-Arnoux, A., **Murris, J.**, Bechet, S., Jung, C., Auchabie, J., Dupeyrat, J., ... & Schortgen, F. (2024). Protocol for fever control using external cooling in mechanically ventilated patients with septic shock : SEPSISCOOL II randomised controlled trial. *BMJ open, 14(1), e069430*.
- Degrave, R., **Murris, J.**, Charles-Nelson, A., Hermine, O., Porcher, R., Ravaud, P., ... & Arlet, J. B. (2024). Risk factors for thromboembolic events in patients hospitalized for COVID-19 pneumonia in a general ward and requiring treatment with oxygen. *Postgraduate Medical Journal, 100(1180), 120-126*.

-
- Espitia, O., Raimbeau, A., Planquette, B., Katsahian, S., Sanchez, O., Espinasse, B., ... & **Murris, J.** (2023). A systematic review and meta-analysis of incidence of post-thrombotic syndrome, recurrent thromboembolism, and bleeding after upper extremity vein thrombosis. *Journal of Vascular Surgery : Venous and Lymphatic Disorders*, 12(1), S2213-333X.
 - Devi-Marulkar, P., Fastenackels, S., Karapentantz, P., Goc, J., Germain, C., ... **Murris, J.**, ... & Dieu-Nosjean, M. C. (2022). Regulatory T cells infiltrate the tumor-induced tertiary lymphoid structures and are associated with poor clinical outcome in NSCLC. *Communications Biology*, 5(1), 1416.
 - Denis, B., Gendre, I., Tuzin, N., **Murris, J.**, Guignard, A., Perrin, P., & Rahmi, G. (2022). Adenoma detection rate is enough to assess endoscopist performance : a population-based observational study of FIT-positive colonoscopies. *Endoscopy International Open*, 10(09), E1208-E1217.
 - Gogas, H., Dummer, R., Ascierto, P. A., Arance, A., Mandalà, M., ... **Murris, J.**, ... & Flaherty, K. T. (2021). Quality of life in patients with BRAF-mutant melanoma receiving the combination encorafenib plus binimatinib : Results from a multicentre, open-label, randomised, phase III study (COLUMBUS). *European Journal of Cancer*, 152, 116-128.

Articles de conférence

- **Murris, J.**, Amadei, T., Kirscher, T., Klein, A., Tropeano, A. I., & Katsahian, S. (2024). A novel methodological framework for the analysis of health trajectories and survival outcomes in heart failure patients. *Learning from Time Series for Health, Workshop at ICLR 2024*.

Communications orales

- Ferrier, H. (oratrice), **Murris, J.**, Simavonian, A., Thébault, J.L., Ibanez, G., & Seroussi, B., Connaissances des patients d'Île-de-France sur le rôle du médecin traitant pour organiser le parcours de soins coordonnés, *15e Congrès Médecine Générale France, Paris, France*, 2022.
- Lavenu, A. (oratrice), **Murris, J.**, Mareau, A., Rouzé, T., Fromont, M., Gares, V., & Katsahian, S., Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation, *53e Journées de Statistique, Lyon, France*, 2022.

Posters

- **Murris, J.**, Dialla, O., Zkik, A., & Tadmouri, A. (2023). The Curse of Data Maturity in Observational Studies : Practical Advice from Protocol Development to Interpretation of Results. *Value in Health*, 26(12), S424-S425.
- Wasan, H. (orateur), Lonardi, S., Desai, J., Folprecht, G., Gallois, C., Polo Marques, E., Khan, S., **Murris, J.**, & Taieb, J. (2023). Novel data visualization and unsupervised machine learning techniques to support optimal management of toxicity profiles of

-
- encorafenib plus cetuximab in patients with BRAFV600E-mutant metastatic colorectal cancer. *Annals of Oncology* (34), S148-S149.
- **Murris, J.**, Zkik, A., Dialla, O., Khan, S., & Tadmouri, A. (2022). Data Visualization in Real-World Studies to Aid Understanding and Interpretation. *Value in Health*, 25(12), S350-S350.
 - Cribier, B., Doutre, M.S., Taieb, C. (orateur), **Murris, J.**, Saint Aromant, M., Richard, M.A., & Petit, A. (2021). Prevalence of visible skin diseases : an international study of 13,138 individuals. *Annals of Dermatology and Venereology*, 1(8), A295-A296.

Autres activités scientifiques

- Journées de Biostatistique, Paris (2024) – membre du comité d’organisation
- HeKA junior seminar, PariSanté Campus (une fois par mois, 2024) – organisatrice et *chair*
- Open Journal Club dans le cadre de la certification Science Ouverte, Université Paris Cité (une fois par mois, 2023) – organisatrice
- Séminaire de l’équipe HeKA, Rouen (2023) – co-organisatrice

Table des matières

Résumé	i
Remerciements	vii
Valorisation scientifique	ix
Liste des figures	xviii
Liste des tableaux	xix
Liste des abréviations	xxi
Introduction	1
I Survie et événements récurrents	13
I.1 Fondements de l'analyse de survie	15
I.1.1 Définitions des dates en analyse de survie	15
I.1.2 Censure et types de données de survie	16
I.1.3 Le temps de survie	17
I.1.4 Fonction de survie et fonction de risque	18
I.2 Méthodes pour l'analyse de survie classique	18
I.2.1 Estimations non paramétriques	19
I.2.2 Modélisation semi-paramétrique et paramétrique	20
I.3 Méthodes adaptées aux événements récurrents	22
I.3.1 Au-delà du premier événement	22
I.3.2 Composantes des événements récurrents	23
I.3.3 Modélisation des données avec événements récurrents	25
I.3.4 En présence d'un événement terminal	30
I.4 Logiciels et outils statistiques	33
I.5 Étude de cas	34
I.5.1 Design de l'étude	34
I.5.2 Résultats	34
I.6 Discussion	37
I.6.1 Avantages et inconvénients des modèles pour les événements récurrents	37
I.6.2 Implications cliniques	38
I.6.3 Perspectives futures et axes de recherche	39

II Apprentissage pour données censurées	41
II.1 Concepts fondamentaux de l'apprentissage	45
II.1.1 Les différents types d'apprentissage	45
II.1.2 Sur- et sous-apprentissage	46
II.1.3 Évaluation des modèles	48
II.2 Apprentissage pour les données censurées	50
II.2.1 Régressions pénalisées	51
II.2.2 Méthodes d'apprentissage automatique	52
II.2.3 Métriques de performance	55
II.3 Apprentissage avec événements récurrents	59
II.4 Discussion	83
III Forêts aléatoires de survie pour l'analyse des événements récurrents	87
III.1 Arbres et forêts de survie	89
III.1.1 Des arbres de décision aux arbres de survie	89
III.1.2 Spécificités des forêts aléatoires	90
III.1.3 Actualités des forêts aléatoires de survie	92
III.2 Développement des forêts aléatoires de survie pour les événements récurrents	93
III.3 Application de RecForest aux données du PMSI	124
III.3.1 Contexte	124
III.3.2 Analyse des readmissions postopératoires auprès des patients atteints de cancer digestif	124
III.4 Discussion	130
III.4.1 Une taxonomie pour traiter les événements récurrents en survie .	130
III.4.2 Axes de développement	130
IV Interprétabilité, santé et survie : vers une utilisation plus transparente des algorithmes d'IA	133
IV.1 Importance des critères d'interprétabilité et d'explicabilité pour les dispositifs médicaux	136
IV.1.1 Contexte des dispositifs médicaux	136
IV.1.2 Évaluation de performance, d'interprétabilité et d'explicabilité pour les dispositifs médicaux	137
IV.2 Interprétabilité et survie	157
IV.2.1 Contexte	157
IV.2.2 Étude de cas	157
IV.3 Une méthode d'interprétabilité <i>model-specific</i> pour la survie	169
IV.3.1 Introduction de TreeShap	169
IV.3.2 Extensions possibles pour la survie	172
IV.4 Discussion	173
Conclusion	175
Bibliographie	179

Annexe A Chapitre 1	197
A.1 Compléments à l'étude de cas	197
A.1.1 Structure des données pour l'ajustement des modèles	197
A.1.2 Programmation en R	198
Annexe B Chapitre 3	201
B.1 Compléments à l'illustration de RecForest	201
B.1.1 Codes utilisés	201
B.1.2 Description des patients	215
Annexe C Chapitre 4	223
C.1 Programmation de TreeSHAP pour l'exemple	223

Liste des figures

1	Recherche dans Pubmed des articles avec des événements récurrents en oncologie ces 5 dernières années (juin 2024)	2
2	Critères de jugement des articles mentionnant les événements récurrents en oncologie ces 5 dernières années	3
3	Divergence de populations entre patients traités et patients inclus dans les essais cliniques (Source : Cerreta et al. [2012])	6
4	Schéma d'un historique médical	8
I.1	Concept de survie <i>classique</i> et événements récurrents	15
I.2	Processus de censure. Le point orange désigne la survenue d'un événement; DO = date d'origine ; DP = date de point	17
I.3	Processus d'événements en survie <i>classique</i>	19
I.4	Processus d'événements récurrents	22
I.5	Processus d'événements récurrents avec événement terminal	31
I.6	Design de l'étude sur les réadmissions à l'hôpital après une première chirurgie pour cancer digestif. BL = <i>baseline</i> ; DO = Date d'origine ; DP = Date de point.	34
I.7	Nombre moyen estimé des réadmissions à l'hôpital par type de chirurgie pour les patients atteints de cancer digestif. MCF = <i>mean cumulative function</i>	35
II.1	Solution de problème par algorithmes	44
II.2	Vue d'ensemble des types d'apprentissage	46
II.3	Les concepts de sur- et sous-apprentissage	47
II.4	Ensembles d'entraînement, de validation et de test	49
II.5	Ordre d'événements réel et prédit pour deux individus	56
II.6	Pondération par la distribution de la censure dans le Brier score	57
II.7	Erreur moyenne absolue	58
III.1	Construction d'un arbre de survie de RecForest - Nœud initial	94
III.2	Construction d'un arbre de survie de RecForest	94
III.3	Nombre cumulé de réadmissions postopératoires par type d'intervention chirurgicale	127
III.4	Importance des variables sur 5 permutations	129
III.5	Taxonomie de l'analyse des événements récurrents	131
IV.1	Concepts menant à la confiance (Source : Ali et al. [2023])	135
IV.2	Arbre de décision simple pour l'illustration de TreeSHAP ($x = X[0]$ et $y = X[1]$)	170

IV.3	Représentation d'une explication SHAP pour une instance i	171
IV.4	Schéma de l'algorithme SurvSHAP (Source : Krzyński et al. [2023])	172
IV.5	Évolution des arbres de décision (Source : Enjoy Algorithms)	177

Liste des tableaux

I.1	Hazard ratios et intervalles de confiance à 95% de cinq modèles pour traiter les événements récurrents, ajustés sur le type de chirurgie	35
I.2	Récapitulatif des modèles présentés	37
II.1	Régressions pénalisées du modèle de Cox	51
II.2	Apprentissage automatique : récapitulatif des méthodes	85
III.1	Erreurs OOB moyennes pour l'optimisation de <i>mtry</i> avec 5 répétitions sur l'ensemble d'entraînement	128
III.2	Performances de RecForest sur l'ensemble de test	128
A.1	Les dix premières observations des données de l'étude de cas	198
A.2	Structure des données pour le modèle PWP avec trois strates	199
A.3	Structure des données pour le modèle WLW avec trois strates	200
B.1	Codes de diagnostic, CCAM et CIM-10	202
B.2	Caractéristiques à l'inclusion des patients ayant bénéficié d'une chirurgie digestive majeure entre 2020 et 2022 en France (Source : PMSI)	216
B.3	Nombre de réadmissions post-opératoires dans les 6 mois des patients ayant bénéficié d'une chirurgie digestive majeure entre 2020 et 2022 en France (Source : PMSI)	220
B.4	Actes performés des réadmissions post-opératoires dans les 6 mois des patients ayant bénéficié d'une chirurgie digestive majeure entre 2020 et 2022 en France (Source : PMSI)	221

Liste des abréviations

- AG** Andersen-Gill. 26, 27, 33, 35–38, 198
- AI-MDs** *Artificial intelligence-based medical devices.* 135–137
- AUC** *Area Under the Curve.* 56
- CCAM** Classification Commune des Actes Médicaux. 34, 125
- CIM-10** Classification Internationale des Maladies, 10e édition. 34, 125, 126, 128
- C-index** Indice de concordance. 55, 56, 83, 95, 96, 127
- CPH** *Cox proportional hazard.* 50, 55, 83
- CRP** Protéine C-réactive. 42, 43
- DDN** Date de dernières nouvelles. 16
- DO** Date d'origine. 15
- DP** Date de point. 16
- EHR** *Electronic Health Records.* 6
- EN** *Elastic-Net.* 51, 52
- HAS** Haute autorité de santé. 136, 137
- HR** *Hazard ratios.* 21, 36
- IC** Intervalle de confiance. 36
- IMSE** *Integrated mean squared error.* 96, 126–128
- IScore** *Integrated score.* 96, 127, 128
- LASSO** *Least Absolute Shrinkage and Selection Operator.* 51, 52
- MAE** *Mean absolute error.* 58, 83, 84
- MCF** *Mean cumulative function,* fonction moyenne cumulative. 30, 33
- mCRC** Cancer colorectal métastatique. 124
- ML** *Machine learning,* apprentissage automatique. 42, 43, 45, 50, 83, 130, 157, 169, 172–174
- MSE** *Mean squared error.* 95, 96
- OOB** *Out-of-bag.* 90, 91, 93, 126, 127
- PMSI** Programme de Médicalisation des Systèmes d'Information. 34, 124, 125

-
- PWP** Prentice, Williams, et Petersen. xix, 27, 29, 37, 38, 197–199
- PWP-GT** Prentice, Willimas, et Petersen avec temps par intervalle. 27, 35, 36, 197, 198
- PWP-TT** Prentice, Willimas, et Petersen avec temps total. 27, 35, 36, 197, 198
- RCTs** *Random clinical trials.* 5, 8
- RSF** *Random Survival Forest.* 54, 55, 83, 90, 92, 93
- RWD** *Real-world data.* 7, 8
- SNDS** Système National des Données de Santé. 6–8, 12
- SSVM** *Survival support vector machines.* 53, 83
- SVM** *Support vector machines.* 52
- VImp** Variable importance. 91
- VTE** Evénements thromboemboliques veineux. 2, 3
- WLW** Wei-Lin-Wessfeld. xix, 29, 33, 35–38, 199, 200

Introduction

Lors de l'édition 2024 du plus grand congrès annuel en cancérologie (*American Society of Clinical Oncology*, ASCO), [Tateishi et al. \[2024\]](#) ont présenté les défis associés à la pré-diction des récidives postopératoires chez les patients atteints de cancer du poumon non à petites cellules. Cette étude se base sur un algorithme d'apprentissage automatique pour la pré-diction de récidive dans les cinq ans suivant une chirurgie, en intégrant des données de séquençage complet de l'exome et des caractéristiques d'imagerie médicale. L'une des dé-couvertes clés était la corrélation entre les mutations des gènes TP53 et RBM10 et le risque élevé de récidive.

Ces découvertes ont été rendues possibles alors que seule la première récidive a été prise en compte, négligeant le fait que certains patients puissent récidiver deux, voire trois fois [[Uramoto and Tanaka, 2014](#)]. Comment les résultats évoluent-ils en prenant en compte les récidives subséquentes ? Et quel est l'apport de l'apprentissage automatique par rapport aux modèles statistiques usuels pour traiter de tels événements ? C'est ce que propose d'explorer ce travail de thèse.

Dans cette introduction, nous ferons d'abord un état des lieux des événements récur-rents en oncologie. Nous verrons ensuite que les données de vie réelle, au-delà des essais cliniques, constituent un cadre idéal pour le développement de nouveaux modèles robustes et pertinents, offrant une meilleure compréhension de l'évolution de la maladie et des ré-pONSES aux traitements. Enfin, nous discuterons des apports de l'apprentissage automatique pour l'analyse des données médicales, en montrant comment ces techniques peuvent améliorer la précision des prédictions et la gestion clinique des patients.

Événements récurrents en oncologie

En oncologie, la rechute de cancer ou encore la progression de la tumeur sont souvent utilisées pour mesurer l'effet d'un traitement ou d'une intervention, selon les recomman-dations RECIST¹ [[Eisenhauer et al., 2009](#)]. Les méthodes classiques visent à modéliser le temps d'apparition de la première rechute, ou de la première progression, comme l'estima-teur de [Kaplan and Meier \[1958\]](#) ou le modèle de [Cox \[1972b\]](#). Des critères de jugement

1. Afin de simplifier et d'uniformiser les critères d'évaluation des essais cliniques, les organismes euro-peen, américain et canadien de recherche sur le cancer ont défini, en 2000, les critères RECIST (*Response Eva-luation Criteria in Solid Tumors*) uniformes permettant de comparer les résultats.

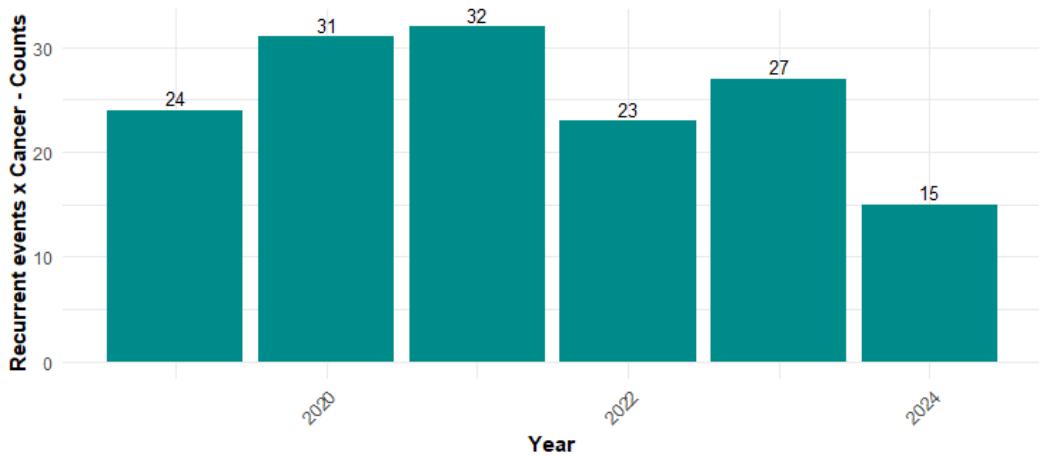


FIGURE 1 – Recherche dans Pubmed des articles avec des événements récurrents en oncologie ces 5 dernières années (juin 2024)

fréquemment utilisés sont par exemple la survie globale (*overall survival*), la survie sans progression (*progression-free survival*) ou encore la survie sans rechute (*relapse-free survival*). Par conséquent, à part lorsque l'intérêt est porté sur le décès, tous les événements qui se produisent après le premier événement sont ignorés dans l'analyse. Or, un événement clinique d'intérêt peut se produire plusieurs fois. L'analyse basée sur le temps jusqu'au premier événement ne peut pas être utilisée pour examiner l'effet des facteurs de risque sur le nombre de récidives au fil du temps [Dancourt et al., 2004, Pandeya et al., 2005]. Il existe alors de nombreuses méthodes d'analyse des événements récurrents, soit conditionnelles, comme les modèles d'[Andersen and Gill \[1982\]](#) ou de [Prentice et al. \[1981\]](#), soit marginales, comme les modèles de [Wei et al. \[1989\]](#).

Les méthodes basées sur les événements récurrents permettent de mieux caractériser le fardeau de la maladie pour un patient, améliorant ainsi la précision statistique et l'interprétation clinique des mesures de l'effet du traitement, selon Claggett et al. [2018]. Par exemple, Rogers et al. [2014] ont réanalysé des données d'essais cliniques et ont montré que les estimations de l'effet du traitement diffèrent entre les analyses du temps jusqu'au premier événement et les analyses des événements récurrents.

Usage actuel des méthodes d'événements récurrents en oncologie dans la littérature médicale

Pourtant, en oncologie, les exemples de modélisation avec événements récurrents demeurent très peu nombreux. Une recherche sur PubMed ne révèle que 112 articles mentionnant une analyse d'événements récurrents en oncologie au cours des cinq dernières années (Figure 1). Seuls 47 de ces 112 articles ont en fait un critère de jugement désignant une récurrence. Les récurrences les plus fréquentes sont les mutations génétiques répétées dans le temps (11 sur 47), la récurrence du cancer (11/47) et la récurrence des événements thromboemboliques veineux (VTE) (10/47) (Figure 2).

INTRODUCTION

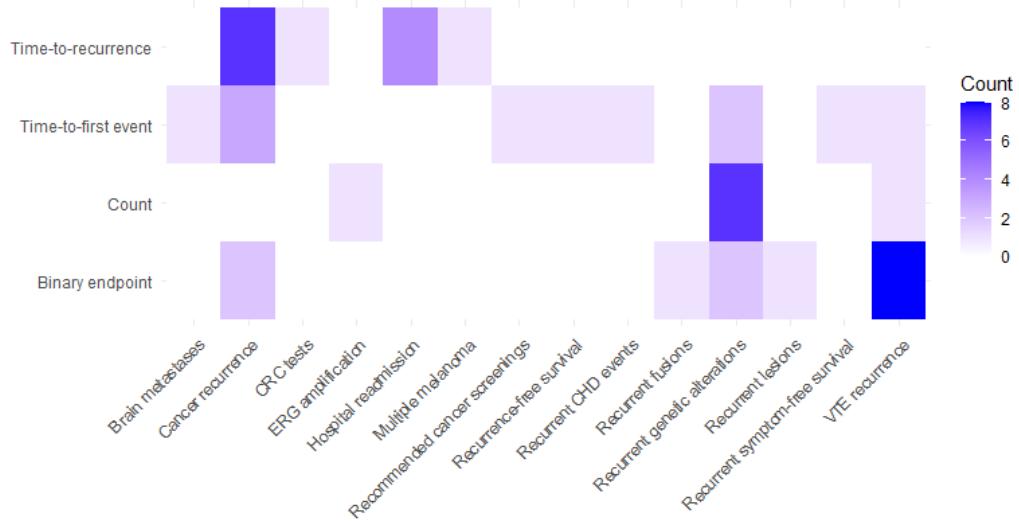


FIGURE 2 – Critères de jugement des articles mentionnant les événements récurrents en oncologie ces 5 dernières années

Cependant, seulement 13 articles traitent effectivement des récurrences avec des méthodes appropriées, incluant des approches conditionnelles (6/13), marginales (4/13) et des modèles joints (3/13). Les récurrences de cancer, en particulier le cancer du sein, sont correctement modélisées dans certaines études, notamment celles de [Osmani et al. \[2019, 2021\]](#), [van Maaren et al. \[2022\]](#). Les réadmissions à l'hôpital sont étudiées dans le contexte du fardeau du cancer de l'enfance par [Smith et al. \[2020\]](#), [Abdalla et al. \[2023, 2024\]](#). En revanche, les études sur les données génétiques semblent moins enclines à utiliser des approches par événements récurrents, comme [DeLeon et al. \[2020\]](#), [Torres et al. \[2022\]](#). En général, seul le nombre de répétitions des mutations est donné. Il est probable que des analyses plus poussées, prenant en compte l'ensemble des mutations et leur temporalité pourraient fournir des perspectives supplémentaires. De même, les récurrences des VTE sont souvent traitées comme des critères binaires, tels que [Girardi et al. \[2023\]](#), [Barca-Hernando et al. \[2023\]](#). Des analyses plus sophistiquées intégrant bien la récurrence des événements pourraient améliorer notre compréhension de ces récurrences et leur gestion clinique.

Par ailleurs, parmi les 112 articles identifiés, 30 sont des nouvelles propositions méthodologiques pour traiter les événements récurrents. Cela montre l'intérêt grandissant pour ces approches en oncologie. Ces nouvelles méthodes ne se limitent pas à la modélisation, mais incluent également des aspects pratiques tels que le calcul de la taille de l'échantillon nécessaire pour les études sur les événements récurrents [[Dinart et al., 2024](#)].

Ces observations soulignent le potentiel sous-exploité des méthodes basées sur les événements récurrents en oncologie. L'acclimatation des chercheurs à ces méthodes pourrait renforcer la robustesse des analyses et, en fin de compte, améliorer la prise en charge des patients.

Complexité des événements récurrents

La complexité d'acquisition et de traitement des données d'événements récurrents peut être la raison de la faible utilisation de méthodes spécifiques. Tout d'abord, la nature hétérogène des cancers et des traitements reçus par les patients crée de la variabilité dans les modèles de récurrence d'événements. En premier lieu, les événements récurrents sont de même nature par définition, mais il semble pertinent de se demander si tous les praticiens s'accordent sur cette approche [Thenmozhi et al., 2019]. Par exemple, une réhospitalisation pour une complication spécifique peut-elle être jugée de même nature qu'une autre hospitalisation pour une raison différente ? Toutes les récurrences de cancer se valent-elles ?

Ensuite, l'hétérogénéité au sein de la population apparaît dans le cas où certains sujets peuvent être plus enclins que d'autres à expérimenter les événements d'intérêt [Andersen and Gill, 1982]. Cette variabilité peut être due à des facteurs non mesurés ou inconnus, ou encore à des périodes de suivi différentes, ce qui peut influencer la probabilité qu'un sujet soit considéré à risque.

Par ailleurs, les événements pour un même individu ne sont pas indépendants [Wei et al., 1989]. La survenue d'un événement peut impacter la survenue des événements suivants, entraînant une corrélation entre les temps d'événements. En d'autres termes, la survenue d'un événement peut modifier le risque de survenue d'événements futurs. Par conséquent, ne pas tenir compte de cette dépendance entre événements peut entraîner une sous-estimation de l'incertitude des paramètres de régression, réduisant ainsi la largeur des intervalles de confiance [Pickles and Crouchley, 1995]. Si les corrélations entre les événements récurrents sont ignorées, l'hypothèse nulle est généralement rejetée, car le modèle de Cox [1972b] n'incorpore pas de corrélation entre les sujets.

Enfin, la censure des données pose des défis supplémentaires, car elle se produit lorsque certains événements ne sont pas observés pendant la période de suivi, compliquant ainsi l'estimation des risques. En particulier, les patients perdus de vue sont considérés comme censurés, puisqu'il n'est pas possible de connaître leur état de santé au-delà de la dernière observation. De même, lorsque des patients ont un événement au-delà de la date de fin de l'étude, cet événement ne peut pas être observé. Ainsi, l'application de modèles statistiques traditionnels peut être inadéquate [Duchateau et al., 2003, Rogers et al., 2014].

Selon Thenmozhi et al. [2019], le choix du modèle approprié pour l'analyse des données sur les événements récurrents dépend de nombreux facteurs, tels que le nombre d'événements, la relation entre les événements subséquents, la corrélation entre les sujets et les différentes covariables, ainsi que la taille de l'échantillon.

Après avoir exploré le sujet des événements récurrents en oncologie, les sources de données de vie réelle méritent une attention particulière en raison de leur rôle important dans l'analyse et l'amélioration des traitements.

Les données de vie réelle au service de la santé publique

Intérêt pour les données de vie réelle

L'étude des données médicales ne peut pas se limiter aux essais cliniques. Les essais contrôlés randomisés (*randomized clinical trials*, RCTs) sont largement reconnus comme étant la méthode de référence pour évaluer l'efficacité et la sécurité des interventions médicales [Hariton and Locascio, 2018]. Leur conception rigoureuse, basée sur la randomisation des patients entre un groupe de traitement et un groupe contrôle, permet de minimiser les biais et de garantir la comparabilité des groupes étudiés [Bothwell et al., 2016]. Cependant, les RCTs présentent des limites majeures pour le développement des soins :

Limites qualitatives

- **Représentativité restreinte** : Les critères d'inclusion stricts des RCTs excluent souvent des groupes de patients importants, tels que les personnes âgées, les patients avec des comorbidités graves (comme l'insuffisance rénale ou hépatique), les femmes enceintes, et les enfants (exemple Figure 3 issue de Cerreta et al. [2012]). En conséquence, les résultats des RCTs peuvent ne pas être généralisables à la population clinique réelle, qui est plus hétérogène [Kim et al., 2015].
- **Conditions optimales non représentatives des soins réels** : Les RCTs se déroulent dans des conditions contrôlées et optimales qui diffèrent souvent de la pratique clinique quotidienne [Kumar et al., 2020]. Par exemple, la stricte adhésion au protocole de traitement, la surveillance et le soutien supplémentaire ne reflètent pas toujours les conditions rencontrées par les patients en dehors du cadre de l'essai.

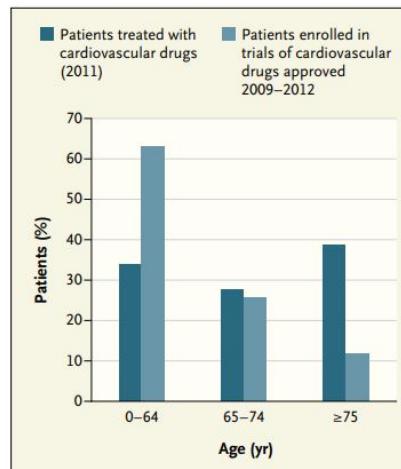
Limites quantitatives

- **Durée limitée du suivi** : La plupart des RCTs sont menés sur une période relativement courte en raison des contraintes de temps et de coûts. Cela signifie que les effets à long terme des traitements, ainsi que les événements indésirables tardifs, peuvent ne pas être détectés ou suffisamment étudiés [Bothwell et al., 2016].
- **Taille des échantillons** : Les RCTs impliquent souvent des échantillons de taille modeste, surtout dans le contexte des maladies rares où le recrutement de participants est particulièrement difficile [Feeley et al., 2009]. Des échantillons plus petits augmentent le risque de ne pas détecter des phénomènes rares, mais cliniquement significatifs, dans un cadre de distribution à queue épaisse (*fat tail distribution*), en raison d'une puissance statistique insuffisante.

Les sources de données de santé sont de plus en plus nombreuses. Deux autres collections de données de recherche sont couramment utilisées en complément aux RCTs. Les études de cohorte impliquent le suivi d'un groupe désigné d'individus sur une période donnée [Porta et al., 2014]. Les registres couvrent de manière exhaustive et collectent systématiquement des données sur une population clinique bien définie.

Cependant, la collecte de données via les cohortes et les registres est coûteuse et peut être limitée à des échantillons restreints, ce qui peut entraver la généralisation des résultats

FIGURE 3 – Divergence de populations entre patients traités et patients inclus dans les essais cliniques (Source : Cerreta et al. [2012])



[Sheikh et al., 2015]. L’expansion des technologies numériques et des systèmes d’information en santé a conduit à une accessibilité (plus ou moins) accrue de sources de données disponibles pour la recherche [Braa et al., 2004, Lupton, 2016, Bossen et al., 2019]. En particulier, les données issues des dossiers de santé électroniques (*electronic health records*, EHR) sont hébergées dans des entrepôts et contiennent des informations détaillées sur les antécédents médicaux des patients, les diagnostics, les traitements et les résultats cliniques. **Ces entrepôts de données de santé centralisent et homogénéisent de façon sécurisée les informations médicales.**

Les entrepôts de données hospitalières, comme l’entrepôt de l’AP-HP regroupant 38 centres ou la base MIMIC (*Medical Information Mart for Intensive Care* du Beth Israel Deaconess Medical Center, USA, accessible depuis Johnson et al. [2023]), contiennent les données collectées auprès des établissements de santé tels que les hôpitaux, les cliniques ou les centres médicaux. D’autres entrepôts regroupent des données médico-administratives [Cadarette and Wong, 2015]. Les bases de données de remboursement des assurances maladie fournissent des informations sur les services médicaux facturés et remboursés, y compris les médicaments prescrits et les procédures effectuées. **En France, la Caisse nationale d’assurance maladie collecte toutes les demandes de remboursement des activités hospitalières et des soins de ville dans une base de données unique.** Le Système National des Données de Santé (SNDS) constitue par ailleurs une avancée considérable pour analyser et améliorer la santé de la population.² D’autres sources de données sont les données de séquençage génomique, les données d’imagerie médicale, ou encore les données issues des dispositifs médicaux connectés [Beaulieu-Jones et al., 2020].

Les données de vie réelle sont idéales pour l’analyse des événements récurrents, car elles reflètent la diversité et la complexité des conditions cliniques des patients, offrant ainsi un aperçu plus complet et représentatif des occurrences répétées d’événements dans des contextes variés et non contrôlés. De plus, lorsque ces données sont collectées en routine, le risque d’omission d’événements importants est minime [Sherman et al., 2016].

2. Site du Health Data Hub.

Toutes ces données constituent les données issues de la vie réelle, que l'on appelle *real-world data* (RWD), et ne sont pas principalement collectées à des fins de recherche [HAS, 2021].

Les promesses des données réelles

La taille et la complexité des EHR augmentent à mesure que les patients interagissent avec les systèmes de santé. Dans le même temps, les investissements publics et privés stimulent le développement de nouvelles méthodes capables de traiter des types de données de plus en plus diversifiés [Bryan and Li, 2024].

De nombreux sponsors (entreprises ou organisations qui mènent un essai clinique) évaluent le potentiel des données de vie réelle dans les essais cliniques [Sherman et al., 2016]. Cette démarche vise à faciliter l'élaboration de designs adaptatifs, à réduire les coûts et la durée des essais cliniques ainsi qu'à accélérer l'obtention des autorisations de mise sur le marché. Par exemple, l'utilisation de bras de contrôle synthétiques dans les essais de phase I, II et III peut accélérer les délais en exploitant les données existantes au lieu de recruter un groupe de comparaison distinct [Makady et al., 2017b]. Une telle conception du design de l'étude peut éviter d'avoir à randomiser les participants pour leur administrer un placebo, rendant ainsi les essais plus attrayants pour les patients, raccourcissant les délais et permettant l'étude simultanée de plusieurs options dans le cadre d'essais de plateforme (*platform trials*) [Burki, 2023].

Les études observationnelles visent à développer une meilleure compréhension de l'expérience vécue par les patients vis-à-vis de leur maladie et de leur traitement. Cela comporte de même une meilleure appréhension de l'impact sur leur vie quotidienne et leur bien-être, au-delà des mesures cliniques courantes [Makady et al., 2017a]. Ces études aident également les organismes payeurs (organisations ou entités qui paient pour les services de soins qu'un prestataire de soins de santé a administrés, comme l'Assurance Maladie en France) à comprendre l'histoire naturelle et le fardeau des maladies, en particulier des maladies rares [Drummond et al., 2015].

Les données du SNDS en France offrent une mine d'informations précieuses pour la recherche clinique et épidémiologique, notamment en oncologie. Le SNDS regroupe des données de plusieurs sources, incluant les bases de données de l'Assurance Maladie, les hôpitaux et les données de mortalité. Ces informations couvrent un large éventail de paramètres tels que les diagnostics, les traitements, les hospitalisations et les remboursements, permettant ainsi une analyse approfondie des parcours de soins des patients. Par exemple, Martin et al. [2024] ont étudié l'impact des réformes d'accès précoce sur l'innovation en oncologie en France à partir des données du SNDS pour examiner les approbations, le nombre de patients et les coûts associés aux nouveaux traitements. D'autres travaux de Prost et al. [2024] ont exploré l'influence du volume de chirurgies du cancer de l'ovaire sur la survie globale et sans progression, démontrant que les patientes opérées dans des centres à haut volume avaient de meilleurs résultats. Enfin, Albigès et al. [2024] ont analysé les schémas de traitement et l'efficacité réelle chez les patients atteints de car-

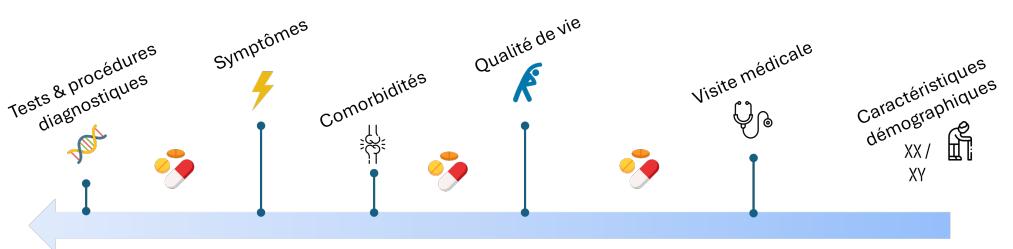


FIGURE 4 – Schéma d'un historique médical

cinome rénal avancé, utilisant les données du SNDS pour mieux comprendre les pratiques cliniques et les résultats thérapeutiques. Ces exemples illustrent comment les données du SNDS peuvent être exploitées pour améliorer la prise en charge des patients et guider les politiques de santé en oncologie.

Néanmoins, il reste plusieurs freins à la mise en œuvre des RWD dans le cadre de la recherche [Makady et al., 2017a]. Ces données sont généralement plus volumineuses et moins structurées que les données recueillies dans le cadre des RCTs. La collecte, le stockage, l'accès et l'analyse en toute sécurité de ces vastes volumes de données peuvent donc s'avérer difficiles. Parmi les problèmes rencontrés, citons la qualité des données, la diversité des sources de données, les questions de confidentialité et d'éthique, ainsi que les incertitudes législatives entourant l'utilisation des RWD.

Les données de vie réelles jouent un rôle de plus en plus important dans la recherche clinique, permettant une meilleure compréhension des maladies et de leur prise en charge [Corrigan-Curay et al., 2018]. Considérer la fiabilité et la pertinence de ces données collectées en routine est néanmoins essentiel afin de répondre efficacement aux exigences tant cliniques que réglementaires.

Les RWD varient largement en termes de format et de qualité. Les données longitudinales et multidimensionnelles posent des défis supplémentaires en termes d'analyse et de modélisation. De nouveaux algorithmes ont été développés pour ces volumes massifs de données comprenant de riches informations (Figure 4).

Apprentissage automatique et recherche médicale

Alors que les algorithmes d'apprentissage sont de plus en plus intégrés à la prise de décision, **les praticiens et les décideurs politiques gagneraient à mieux appréhender leur utilisation et impact sur la santé** [Wong et al., 2018, Deo and Nallamothu, 2016]. Les décisions médicales basées sur les algorithmes d'apprentissage sont plus rapides, plus précises, moins coûteuses et moins biaisées que des jugements cliniques ou examens de cas seuls [Esteva et al., 2019].

Impact de l'apprentissage automatique pour l'analyse de survie

Presque tout type de clinicien, du médecin spécialiste à l'auxiliaire médical, est sujet à une utilisation accrue de l'IA à l'avenir, selon Topol [2019]. L'apprentissage automatique a en effet un impact profond et multidimensionnel sur la recherche médicale, avec des applications allant de la découverte de nouveaux traitements à l'amélioration de diagnostics [Wang et al., 2018]. Ces dernières années, l'apprentissage automatique a démontré sa capacité à améliorer le diagnostic, la détection, la prédition et le pronostic du cancer, par rapport aux méthodes statistiques traditionnelles [Kourou et al., 2015, Munir et al., 2019, Ngiam and Khor, 2019, Zhu et al., 2020, Cuocolo et al., 2020].

Une étude particulièrement probante est celle de Moncada-Torres et al. [2021] sur les patientes atteintes de cancer du sein. Les facteurs associés à la bonne prédition du décès étaient similaires entre l'approche statistique traditionnelle, notamment le modèle de régression de Cox [1972b], et des algorithmes d'apprentissage pour la survie, qui eux rapportaient une meilleure performance. Cette recherche a montré que les modèles d'apprentissage automatique, tels que les forêts aléatoires de survie de Ishwaran et al. [2008], peuvent non seulement surpasser le modèle de Cox en termes de précision prédictive, mais aussi fournir des informations précieuses à propos des facteurs influençant la survie des patientes. Cela souligne le potentiel des algorithmes d'apprentissage automatique dans le domaine de l'oncologie, notamment en termes d'amélioration de la précision des prédictions.

Une autre étude menée par van Zutphen et al. [2021] utilise une extension d'arbres de décision pour la survie, appelée forêts aléatoires de survie, pour analyser l'impact des comportements de vie sur la récidive et la survie globale des patients atteints de cancer colorectal. En particulier, cette recherche met en évidence des associations non linéaires entre la consommation de certains produits et la survie globale des patients. Les comportements identifiés comprennent plusieurs facteurs connus, tels que ceux inclus dans les recommandations de prévention du cancer, mais aussi d'autres comportements liés au mode de vie n'étant pas aussi bien établis. L'utilisation des forêts aléatoires de survie permet de capturer des interactions complexes et des relations non linéaires entre les variables comportementales et les résultats de survie. Ainsi cela rend possible une modélisation plus appropriée que les modèles linéaires traditionnels.

Ainsi, l'importance de l'apprentissage automatique pour traiter les problèmes de survie en oncologie est désormais bien établie. De nombreux exemples illustrent les succès de ces approches. Des revues de la littérature de Wang et al. [2019] et Tizi and Berrado [2023] présentent des perspectives détaillées et des analyses approfondies de l'application des techniques d'apprentissage automatique en oncologie.

Cependant, ces algorithmes d'apprentissage pour la survie semblent se concentrer uniquement sur l'apparition du premier événement, sans prendre en compte la récurrence. Aucun algorithme ne semble avoir été spécifiquement conçu pour gérer les événements récurrents avec censure.

Enjeux des algorithmes d'apprentissage

Lorsqu'il s'agit de développer des algorithmes d'apprentissage, et en particulier à des fins de recherche médicale, plusieurs défis doivent être pris en compte pour assurer leur efficacité et leur précision.

Biais d'échantillonnage Les algorithmes d'apprentissage sont entraînés sur des données d'entraînement pour apprendre à généraliser, et sont évalués sur des jeux de données de test. Le choix des données d'entraînement est crucial. Le biais d'échantillonnage survient lorsque les données utilisées pour entraîner les modèles d'apprentissage automatique proviennent de sources spécifiques ou de sous-populations particulières, ne reflétant pas la diversité de la population générale [Zadrozny, 2004]. Le biais d'échantillonnage est une considération capitale pour la conception d'une étude clinique, car il peut affecter l'appliquabilité des résultats de l'étude à une population réelle de patients [Yu and Eng, 2020].

Les disparités régionales et institutionnelles ont été bien identifiées comme facteurs de biais [Patel et al., 2020]. Plus généralement, il est crucial de s'assurer que les données d'entraînement proviennent de sources variées et soient représentatives de la diversité de la population cible. Les validations externes des algorithmes, utilisant un échantillon indépendant, constituent un moyen efficace pour identifier de tels biais [Marco, 2014].

Supposons qu'une étude visant à prédire le risque de rechutes de patientes atteintes de cancer du sein utilise des données provenant principalement d'hôpitaux situés dans des zones urbaines avec un accès à des soins de santé avancés. Ces hôpitaux peuvent avoir des protocoles de traitement spécifiques, des ressources abondantes et une population de patients qui peut être plus jeune et plus aisée financièrement par rapport à la population générale. Cette source de données peut introduire plusieurs biais. Premièrement, les protocoles de traitement et les ressources disponibles dans les hôpitaux urbains peuvent être plus avancés que ceux disponibles dans les zones rurales ou dans des hôpitaux moins bien financés. Les patientes traitées dans ces hôpitaux peuvent avoir des taux de survie et de réadmission différents par rapport à celles traitées ailleurs. Deuxièmement, la population admise dans les hôpitaux urbains peut ne pas être représentative de la population générale. Si l'échantillon inclut principalement des patientes plus jeunes et plus riches, le modèle peut ne pas bien prendre en compte les risques des patientes plus âgées ou de celles issues de milieux socio-économiques défavorisés. Enfin, les patientes provenant des zones urbaines peuvent avoir un meilleur accès aux soins de suivi et aux services de santé, influençant ainsi les taux de réadmission et de survie. Cela peut fausser les prédictions du modèle lorsqu'il est appliqué à des populations avec un accès limité aux soins.

Exemple fictif

Interprétabilité et transparence Les algorithmes d'apprentissage, surtout les plus complexes, peuvent être décrits comme des "boîtes noires", des algorithmes d'inférence "oracles" rendant des verdicts sans justification [Watson et al., 2019]. Cette problématique est devenue préoccupante, en particulier avec l'adoption du règlement général sur la protection des données (RGPD) en Union européenne, qui confère aux citoyens un "droit à l'explica-

tion".³ Les décisions prises basées sur des algorithmes doivent justifier ces décisions auprès des personnes concernées. Ainsi, si une prédition indique que Y est vrai, une explication indique *pourquoi* Y est vrai [Markus et al., 2021]. La méconnaissance des patients et des médecins de la manière dont les prédictions sont effectuées est un obstacle fréquent à l'adoption de l'apprentissage automatique en santé [Lipton, 2017]. Le caractère automatique des prédictions effectuées par les algorithmes rend essentielle la génération d'explication intelligible pour établir de la confiance entre le patient, le praticien et la machine. Une décision est-elle possible si l'on ne comprend pas pourquoi un modèle effectue une prédition ? La compréhension des algorithmes est donc devenue un enjeu majeur [Doshi-Velez and Kim, 2017].

Éthique et responsabilité Caruana et al. [2015] ont développé un algorithme prédisant les risques de décès des patients hospitalisés pour pneumonie, classant systématiquement les asthmatiques comme étant à faible risque. Cette corrélation était trompeuse (*spurious correlation*), car les patients asthmatiques atteints de pneumonie étaient souvent envoyés directement à l'unité de soins intensifs, où ils recevaient un traitement intensif améliorant leur pronostic de manière significative. Cet exemple soulève la question de la responsabilité. Qui est responsable en cas d'erreur de diagnostic computationnel ? Les cliniciens ? Les statisticiens ? Le débat sur la responsabilité des erreurs d'algorithme n'est pas clos [Johnson and Verdicio, 2019].

Ces enjeux soulignent la complexité de la mise en œuvre de l'apprentissage automatique et l'importance de considérer non seulement les aspects techniques, mais aussi les implications éthiques, sociales et réglementaires [Mittelstadt et al., 2016].

Objectifs de la thèse

L'objectif de cette thèse est double.

Premièrement, ce travail de recherche vise à améliorer la compréhension des événements récurrents afin de proposer une nouvelle méthode basée sur des principes d'apprentissage automatique. Nous avons souhaité illustrer cette nouvelle méthode sur des données observationnelles, tout en répondant à un problème clinique concret pour améliorer la prise en charge des patients atteints de cancer.

Deuxièmement, ce projet de recherche explore les conditions nécessaires à l'adoption de ce type d'algorithme d'apprentissage, en mettant l'accent sur la nécessité de gagner la confiance des utilisateurs. La seconde partie de la thèse se concentre donc sur les défis d'interprétabilité des algorithmes d'apprentissage dans un contexte large, avec une attention particulière sur les modèles de survie.

3. Article en ligne de la CNIL, mai 2016, Le règlement général sur la protection des données - RGPD.

Organisation du manuscrit

Le manuscrit est structuré en quatre chapitres principaux, suivis d'une synthèse dans la section Conclusion. Chaque chapitre est conçu pour être lu de manière indépendante et se termine par un résumé des principaux apports.

Le Chapitre I offre un état de l'art des méthodologies utilisées pour traiter les données censurées. Il présente les principes fondamentaux et les approches permettant de modéliser le temps jusqu'au premier événement. En outre, ce chapitre explore le spectre des méthodes d'analyse de survie appliquées aux événements récurrents, discutant des avantages et des inconvénients de chacune à travers une étude de cas. Il inclut également des informations sur les logiciels et les codes nécessaires pour effectuer ces analyses.

Le Chapitre II se concentre sur les algorithmes d'apprentissage automatique, en particulier leur extension aux analyses de survie et aux métriques spécifiques associées. Ce chapitre inclut un état de l'art des méthodes d'apprentissage pour les événements récurrents, présenté dans l'article intitulé "*Towards filling the gaps around recurrent events in high dimensional framework : a systematic literature review and application*", publié dans le journal *Biostatistics and Epidemiology* en janvier 2023.

Le Chapitre III introduit une nouvelle méthode développée sur la base des fondements de l'analyse de survie pour événements récurrents et des principes d'apprentissage automatique. Cette méthode, nommée RecForest, utilise des forêts aléatoires. L'article intitulé "*Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event*" est soumis au journal *BMC Medical Research Methodology*. Ce chapitre illustre également l'application de RecForest dans le cadre d'une étude observationnelle utilisant une partie des données du SNDS, visant à comprendre les réadmissions postopératoires chez les patients ayant subi une chirurgie pour cancer.

Le Chapitre IV se penche sur l'utilisation des algorithmes d'apprentissage automatique, d'abord de manière générale puis avec un focus sur les analyses de survie. Il s'appuie sur l'article "*Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence-Based Health Technologies : What Healthcare Stakeholders Need to Know*", publié dans le journal *Mayo Clinic Proceedings : Digital Health* en juin 2023. Ce chapitre présente également les méthodes d'interprétabilité appliquées aux analyses de survie, illustrées dans un article soumis à la conférence *Artificial Intelligence, Ethics, and Society 2024* et intitulé "*Bridging Interpretability and Survival Endpoints in Health Technology Assessment*".

La **Conclusion** fournit une synthèse des contributions de la thèse et explore les pistes de développement futur, mettant en lumière les implications pratiques et les perspectives de recherche à venir.

Chapitre I

Survie et événements récurrents

We are often reminded of the story of the gentleman on his 100th birthday who proclaimed that he was looking forward to many more years ahead because "I read the obituaries every day, and you almost never see someone over 100 listed there".

Terry Therneau, Multi-state models and competing risks (juin 2024)

Sommaire

I.1	Fondements de l'analyse de survie	15
I.1.1	Définitions des dates en analyse de survie	15
I.1.2	Censure et types de données de survie	16
I.1.3	Le temps de survie	17
I.1.4	Fonction de survie et fonction de risque	18
I.2	Méthodes pour l'analyse de survie classique	18
I.2.1	Estimations non paramétriques	19
I.2.1.1	Estimation de la fonction de survie	19
I.2.1.2	Estimation de la fonction de risque cumulé	19
I.2.2	Modélisation semi-paramétrique et paramétrique	20
I.2.2.1	Modèle de Cox à risques proportionnels	20
I.2.2.2	Modèles paramétriques	22
I.3	Méthodes adaptées aux événements récurrents	22
I.3.1	Au-delà du premier événement	22
I.3.2	Composantes des événements récurrents	23
I.3.2.1	Échelles temporelles	23

I.3.2.2	Ensemble des sujets à risque	23
I.3.2.3	Processus de comptage	23
I.3.2.4	Processus de Poisson	25
I.3.3	Modélisation des données avec événements récurrents	25
I.3.3.1	Les modèles conditionnels	26
I.3.3.2	Les modèles à fragilité	28
I.3.3.3	Les modèles marginaux	28
I.3.4	En présence d'un événement terminal	30
I.3.4.1	Approche marginale de correction pour les risques compétitifs	31
I.3.4.2	Autres approches	32
I.4	Logiciels et outils statistiques	33
I.5	Étude de cas	34
I.5.1	Design de l'étude	34
I.5.2	Résultats	34
I.6	Discussion	37
I.6.1	Avantages et inconvénients des modèles pour les événements récurrents	37
I.6.2	Implications cliniques	38
I.6.3	Perspectives futures et axes de recherche	39

Introduction

L'analyse des données censurées, ou *time-to-event*, est l'étude du temps écoulé jusqu'à l'occurrence d'un ou plusieurs événements d'intérêt [Clark et al., 2003]. Ce type d'analyse est particulièrement approprié lorsque le temps entre l'exposition (par exemple, le diagnostic ou le début d'un traitement) et l'événement est cliniquement pertinent. Traditionnellement, l'analyse de survie dite *classique* se concentre sur le temps jusqu'au premier événement (Figure I.1) [Altman, 1990].

Cependant, dans de nombreuses situations, le même type d'événements peut se produire de manière répétée chez un même individu. Les événements récurrents, tels que les rechutes multiples ou encore les complications postopératoires (après une chirurgie du cancer), fournissent des informations supplémentaires qui peuvent être d'importance majeure pour une compréhension plus complète des phénomènes étudiés [Lawless and Nadeau, 1995].

L'objectif de ce chapitre est de fournir un état des lieux des méthodes d'analyse de survie classiques ainsi que des méthodes spécifiques aux événements récurrents. Le chapitre est organisé comme suit : après cette introduction, nous présenterons les fondamentaux de l'analyse de survie (Section I.1), suivis par une discussion détaillée des méthodes classiques (Section I.2). Nous introduirons ensuite les événements récurrents et explorerons les différentes approches méthodologiques adaptées à leur analyse (Section I.3), avec les outils

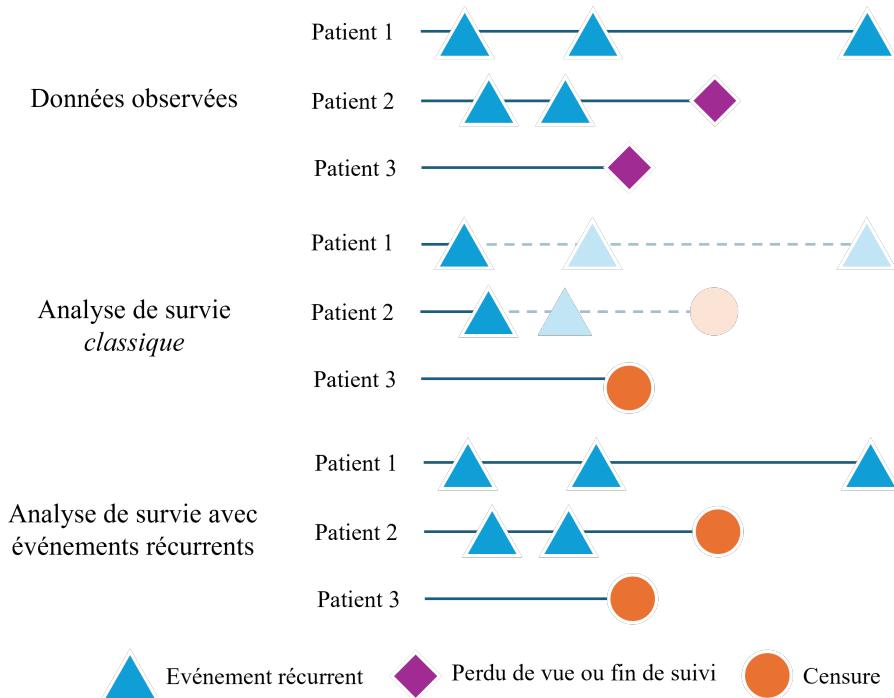


FIGURE I.1 – Concept de survie *classique* et événements récurrents

disponibles associés (Section I.4). Ensuite, nous proposons une étude de cas pour appliquer les méthodes (Section I.5). Enfin, nous conclurons avec une synthèse des méthodes et des perspectives futures dans le domaine (Section I.6).

I.1 Fondements de l'analyse de survie

L'objectif principal de l'analyse de survie est de modéliser le temps jusqu'à l'apparition d'un événement tout en identifiant les facteurs qui l'influencent.

I.1.1 Définitions des dates en analyse de survie

Dans le cadre des études de survie, il est question du délai de survenue d'événements. Le délai est calculé à partir des dates enregistrées dans les données. Voici les définitions des dates clés :

- **Date d'origine (DO)** : C'est le point de départ pour le suivi de chaque patient dans l'étude. Cette date peut correspondre à :
 - La date de randomisation dans le cadre d'un essai clinique, au moment auquel le patient est assigné à un groupe de traitement ou de contrôle.
 - La date d'inclusion dans une étude de suivi de cohorte, qui marque l'entrée du patient dans l'étude et le début de son suivi.

- La date de diagnostic, particulièrement pertinente pour assurer une date d'origine d'état de santé jugé similaire entre les patients face au début de leur maladie.
- **Date de dernières nouvelles (DDN)** : Il s'agit de la dernière date à laquelle des informations fiables sur le patient sont disponibles. Cela peut être :
 - La date de la dernière visite effectuée, pour les patients vivants ou pour lesquels l'événement d'intérêt (autre que le décès) n'a pas été observé.
 - La date de décès, si le patient est décédé.
- **Date de point (DP)** : Cette date est déterminée a priori, souvent dans le protocole de l'étude, et sert de repère pour l'analyse des données. Elle correspond à la date de l'analyse intermédiaire ou finale de l'étude. Après cette date, les informations supplémentaires éventuellement collectées ne sont généralement pas prises en compte dans l'analyse principale. La date de point est fixe pour l'ensemble des individus observés.

I.1.2 Censure et types de données de survie

Dans l'analyse de survie, il est fréquent que le temps d'un événement soit non observé pour certains individus. Ce phénomène, connu sous le nom de **censure**, survient lorsque la date exacte de survenue d'un événement est indéterminée. La censure peut se manifester de différentes manières, chacune ayant des implications sur l'analyse statistique et l'interprétation des résultats. Les types de censure comprennent (Figure I.2) :

- **Censure à gauche** : L'événement s'est produit avant la date d'origine, de sorte que le temps exact de l'événement est inconnu, mais il est inférieur à la date d'origine.
- **Censure par intervalle** : L'événement se produit entre deux points d'observation. Par exemple, si un patient est vu à des visites régulières, l'événement est survenu entre deux visites successives.
- **Censure à droite** : L'événement n'a pas été observé pendant la période d'étude et peut se produire après la fin de l'étude. Ce type de censure est le plus commun et peut se présenter sous deux formes :
 - *Exclu vivant* : Le patient est encore en vie ou bien l'événement n'a pas eu lieu à la date de fin de la période d'observation (ou date de point), et donc le temps d'événement est inconnu et supérieur à cette date.
 - *Perdu de vue* : L'information sur le patient est incomplète, et on ne sait pas si l'événement a eu lieu ou non après le dernier point d'observation connu. Les patients perdus de vue peuvent l'être par exemple en raison d'un retrait de consentement à l'étude, ou d'un déménagement.

La date conservée correspond alors soit à la date de dernière nouvelle, soit la date de point suivant celle qui intervient en premier.

Pour la censure à droite, on ne pourra pas observer le temps d'événement si celui-ci survient après la date de censure. Dans un premier temps, nous ferons l'hypothèse de l'indépendance du processus de censure. Cela signifie que les individus censurés au temps t ne doivent pas constituer un échantillon biaisé de ceux qui sont à risque au même temps t .

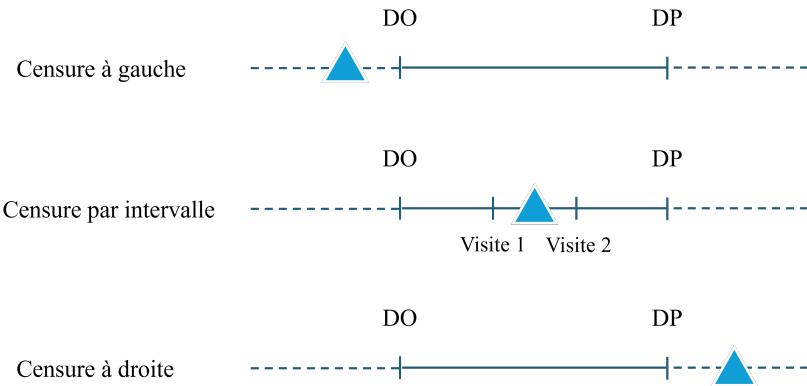


FIGURE I.2 – Processus de censure. Le point orange désigne la survenue d'un événement; DO = date d'origine; DP = date de point

Note : En général, il n'est pas possible de vérifier l'hypothèse de censure indépendante à partir des données disponibles. La censure causée par le fait d'être en vie à la fin de l'étude peut généralement être considérée comme "indépendante" en toute sécurité. Il est tout de même recommandé de toujours suivre les sujets qui sont perdus de vue et de noter les raisons de cette perte de suivi, lorsque cela est possible (par exemple, l'abandon du calendrier de suivi ou l'émigration).

La notion de censure va de pair avec la notion de sujet à risque. Un sujet i est à risque à l'instant t s'il est suivi dans l'étude avec $C_i \geq t$ et qu'il n'a pas encore subi le premier événement avec $T_i \geq t$. Soit $Y_i(t)$ pour $i = 1, \dots, n$ la fonction indicatrice que le sujet i est à risque et sous observation au temps t .

I.1.3 Le temps de survie

Le temps, le délai ou la durée de survie correspond au temps écoulé jusqu'à l'apparition d'un événement spécifique depuis la date d'origine. Soit $C \in [0, \infty)$ le délai de survenue de la censure à droite et T^* le délai de survenue de l'événement d'intérêt deux variables aléatoires, et $\delta \in \{0, 1\}$ une indicatrice du statut de l'événement. Pour un individu i :

- si $T_i^* \leq C_i$, l'événement est observé car il survient avant la censure. Le temps de survenue d'événement est connu et on note $\delta_i = 1$;
- si $T_i^* > C_i$, l'événement n'est pas observé pendant la période d'étude, car s'il survient, c'est après la censure. Le temps de survenue de l'événement n'est pas connu, ou bien l'événement n'a pas eu lieu, et on note $\delta_i = 0$.

Soit $T \perp\!\!\!\perp C$ le temps de survie, T est la variable aléatoire définie par $T = T^* \wedge C$, avec $a \wedge b = \min(a, b)$. La variable aléatoire T est non négative et sa distribution est supposée continue.

I.1.4 Fonction de survie et fonction de risque

Soit T^* le délai de survenue de l'événement d'intérêt. Si T^* est une variable continue et $\forall t \in \mathbb{R}^+$, on peut définir sa fonction de densité $f(t)$ et sa fonction de répartition $F(t)$ de la manière suivante :

$$F(t) = P(T^* \leq t) = \int_0^t f(u)du. \quad (\text{I.1})$$

La **fonction de survie**, notée $S(t)$, est la probabilité de ne pas avoir présenté l'événement d'intérêt avant le temps t . Elle se définit comme suit :

$$S(t) = 1 - F(t) = P(T^* > t), \quad (\text{I.2})$$

où $S(0) = 1$ et $\lim_{t \rightarrow \infty} S(t) = 0$.

La **fonction de risque instantané**, noté $\lambda(t)$ est la probabilité de présenter l'événement d'intérêt dans un petit intervalle de temps juste après t , étant donné que l'événement n'a pas eu lieu jusqu'à ce moment t . À l'inverse de la fonction de survie qui portait sur l'absence de survenue d'événement d'intérêt, la fonction de risque se concentre bien sur la survenue de celui-ci. Elle est définie par :

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T^* < t + \Delta t | T^* \geq t)}{\Delta t}. \quad (\text{I.3})$$

La **fonction de risque cumulé**, noté $\Lambda(t)$ est une quantité intégrale qui représente la somme des risques instantanés dans le temps, et est définie comme :

$$\Lambda(t) = \int_0^t \lambda(u)du. \quad (\text{I.4})$$

Cette fonction donne une mesure globale du risque de l'événement survenant jusqu'au temps t .

Les fonctions $f(t)$, $S(t)$, $\Lambda(t)$ et $\lambda(t)$ sont liées par les relations suivantes :

$$\begin{cases} \lambda(t) = \frac{f(t)}{S(t)}, \\ S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)du\right). \end{cases} \quad (\text{I.5})$$

I.2 Méthodes pour l'analyse de survie classique

Pour la survie *classique*, on se concentre sur la survenue du premier événement. Tel que représenté dans la Figure I.3, l'individu est initialement dans l'état 0, c'est-à-dire qu'il ne présente aucun événement, et peut rester dans cet état. Dès la survenue d'un événement, l'individu passe dans l'état 1.



FIGURE I.3 – Processus d'événements en survie *classique*

I.2.1 Estimations non paramétriques

I.2.1.1 Estimation de la fonction de survie

L'estimation de Kaplan-Meier est une méthode non paramétrique très fréquemment utilisée pour estimer la fonction de survie [Kaplan and Meier, 1958]. Elle produit une fonction de survie en escalier, où chaque palier représente une estimation de la probabilité de survie jusqu'à un certain temps. L'estimation de Kaplan-Meier obtenue est définie à partir des temps ordonnés t_k d'événements :

$$\hat{S}_{KM}(t) = \prod_{k=1}^K \left(1 - \frac{d_k}{n_k}\right), \quad (\text{I.6})$$

où $k : t_k \leq t$ sont les temps où au moins un événement d'intérêt a eu lieu, d_k est le nombre d'événements entre t_k et t_{k-1} , et n_k est le nombre de sujets à risque juste avant t_k . En l'absence de censure, l'estimateur de Kaplan-Meier est équivalent à la fonction de survie empirique.

Le test du log-rank est un outil statistique non paramétrique pour la comparaison des fonctions de survie entre deux ou plusieurs groupes, comme des bras de traitement [Mantel et al., 1966]. L'hypothèse nulle est l'égalité des fonctions de survie pour deux groupes A et B , $H_0 : \hat{S}^A(t) = \hat{S}^B(t)$, et l'hypothèse alternative est $H_1 : \hat{S}^A(t) \neq \hat{S}^B(t)$. Il se base sur la comparaison des temps d'événement observés avec ceux attendus sous l'hypothèse nulle d'égalité des fonctions de survie. La statistique du test du log-rank s'écrit :

$$U_{logrank} = \sum_{j=1}^J \frac{(O_j - E_j)^2}{E_j}, \quad (\text{I.7})$$

où O_j est le nombre d'événements observés et E_j est le nombre d'événements attendus dans le groupe j , et J est le nombre de groupes. La p-valeur pour rejeter l'hypothèse nulle est obtenue en référence à une distribution du χ^2 avec $J - 1$ degrés de liberté [Mantel and Haenszel, 1959]. La p-valeur indique la significativité statistique de la différence entre les fonctions de survie. Les conditions d'application du test du log-rank sont :

- L'indépendance intra- et inter-groupes ;
- Un nombre suffisant d'événements observés au sein de chaque groupe.

I.2.1.2 Estimation de la fonction de risque cumulé

L'estimateur Nelson-Aalen de Lawless and Nadeau [1995] estime le risque cumulé à l'aide d'une fonction croissante en escalier, qui augmente à chaque temps (ordonné) d'évé-

nement. L'estimateur est donné par :

$$\hat{\Lambda}_{NA}(t) = \sum_{k=1}^K \frac{d_k}{n_k}, \quad (\text{I.8})$$

avec $k : t_k \leq t$. L'estimateur de Nelson-Aalen est donc la somme cumulée des taux de risque instantanés estimés à chaque temps d'événement.

Les estimateurs Kaplan-Meier et Nelson-Aalen sont liés de la façon suivante :

$$\hat{S}_{KM}(t) = \prod_{u \leq t} (1 - \Delta \hat{\Lambda}_{NA}(u)), \quad (\text{I.9})$$

où le produit est sur tous les temps d'événements uniques u , $u \leq t$, et $\Delta \hat{\Lambda}_{NA}(u)$ est l'incrément de l'estimateur de Nelson Aalen $\hat{\Lambda}_{NA}$ au temps u .

Les estimations non paramétriques peuvent être utilisées pour identifier des facteurs pronostiques uniques, tels que l'assignation de traitement ou les caractéristiques des patients (l'âge en catégorie, sexe, le stade de la maladie, etc.), mais sans quantification d'un éventuel effet. Cependant, elles ne sont pas conçues pour répondre à des questions sur des données individuelles de patients, car ces estimations ne tiennent pas compte de l'ensemble des caractéristiques des patients.

I.2.2 Modélisation semi-paramétrique et paramétrique

Soit $Z = (Z_1, \dots, Z_n)$ où $Z_i = (Z_{i1}, \dots, Z_{ip})^T$ les p covariables correspondants aux p caractéristiques des n individus, avec $i = 1, \dots, n$.

I.2.2.1 Modèle de Cox à risques proportionnels

Le modèle de Cox à risques proportionnels est un modèle semi-paramétrique et permet d'évaluer l'impact de plusieurs variables explicatives sur le risque de survenue de l'événement d'intérêt [Cox, 1972b]. Le modèle s'écrit

$$\lambda(t|Z) = \lambda_0(t) \cdot \exp(\beta^T Z), \quad (\text{I.10})$$

avec $\exp(\beta^T Z)$ est l'effet multiplicatif des p variables explicatives Z et ne dépend pas du temps, et $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ sont les coefficients à estimer. $\lambda_0(t)$ est non spécifiée et correspond au risque instantané de base lorsque toutes les covariables sont égales à 0 :

$$\lambda(t|Z_{i1} = 0, \dots, Z_{ip} = 0) = \lambda_0(t). \quad (\text{I.11})$$

Le modèle de Cox est dit semi-paramétrique, car il comporte à la fois une partie non paramétrique avec le risque de base, et une partie paramétrique avec la fonction de risque relatif dans l'équation I.10.

Pour les individus i et \tilde{i} et leurs covariables associées Z_i et $Z_{\tilde{i}}$, le ratio des fonctions de risques $\lambda(t|Z_i)$ et $\lambda(t|Z_{\tilde{i}})$ est

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_{\tilde{i}})} = \frac{\lambda_0(t) \cdot \exp(\beta^T Z_i)}{\lambda_0(t) \cdot \exp(\beta^T Z_{\tilde{i}})} = \frac{\exp(\beta^T Z_i)}{\exp(\beta^T Z_{\tilde{i}})}. \quad (\text{I.12})$$

Ce ratio correspond au rapport des risques, le *hazard ratio* (HR). Si le ratio de l'équation I.12 est constant, alors le modèle de Cox (équation I.10) est dit à hasards (ou risques) proportionnels.

Par ailleurs, si toutes les valeurs de Z_i et $Z_{\tilde{i}}$ sont égales à l'exception de la k ème valeur, où $Z_{ik} = Z_{\tilde{ik}} + 1$ et $k \in \{1, \dots, p\}$, alors le ratio de l'équation I.12 est

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_{\tilde{i}})} = \exp(\beta^T(Z_i - Z_{\tilde{i}})) = \exp(\beta_k). \quad (\text{I.13})$$

C'est-à-dire l'effet de l'augmentation d'une unité supplémentaire pour la k ème covariable lorsque toutes les autres ont des valeurs égales (et indépendamment du risque de base).

Estimation des coefficients L'estimation des coefficients β est obtenue par la maximisation de la vraisemblance partielle, qui est définie comme le produit des probabilités conditionnelles de chaque événement observé, compte tenu du risque à ce moment-là :

$$\mathcal{L}_p(\beta) = \prod_{i=1}^n \left(\frac{\exp(\beta^T Z_i)}{\sum_{l \in R^{Cox}(T_i)} \exp(\beta^T Z_l)} \right)^{\delta_i} \quad (\text{I.14})$$

où T_i est le temps d'événement pour l'individu $i = 1, \dots, n$ et δ_i est l'indicateur de censure qui est égal à 1 pour un événement et à 0 pour la censure. L'ensemble des sujets à risque $R^{Cox}(t)$ rassemble les individus exposés à un risque d'événement juste avant le point temporel t , c'est-à-dire tous les individus qui ne sont pas censurés et qui n'ont pas subi d'événement juste avant t . L'ensemble des sujets à risque est donc défini comme suit :

$$R^{Cox}(t) := \{l, l = 1, \dots, n : T_l \geq t\} \quad (\text{I.15})$$

L'estimation des β est ensuite conduite à l'aide de l'algorithme Newton-Raphson, et l'estimateur du risque de base est de type Breslow [Breslow, 1970, Akram and Ann, 2015].

Hypothèses du modèle Le modèle de Cox repose sur les hypothèses suivantes :

- Le risque de base $\lambda_0(t)$ est non-paramétrique ;
- Les effets des covariables sont additifs et linéaires sur l'échelle du log-risque ;
- Les risques sont proportionnels.

Le modèle de Cox est dit à risques proportionnels (*proportional hazards*), ce qui suppose que le rapport des risques entre deux individus reste constant dans le temps, ce qui est connu sous le nom d'hypothèse des risques proportionnels. On a :

$$\frac{\lambda(t|Z_1, \dots, Z_j, \dots, Z_p)}{\lambda(t|Z_1, \dots, 0, \dots, Z_p)} = \exp(\beta_j Z_j), \quad (\text{I.16})$$

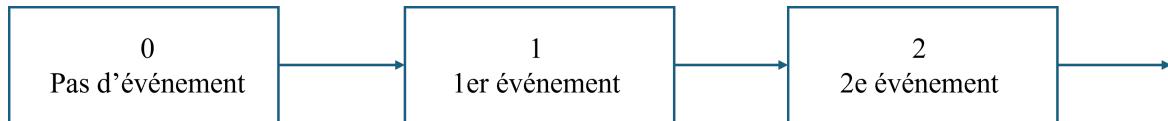


FIGURE I.4 – Processus d'événements récurrents

c'est-à-dire que le taux est constant au cours du temps. Cette hypothèse doit être vérifiée à la fois par inspection visuelle et à l'aide de tests statistiques appropriés, tels que le test basé sur les résidus de Schoenfeld [1982] ou le test de Grambsch and Therneau [1994].

Ces hypothèses sont essentielles et doivent être prises en compte dans toutes les extensions du modèle de Cox.

I.2.2.2 Modèles paramétriques

Les modèles paramétriques offrent une alternative utile lorsque l'on dispose d'informations a priori sur la distribution du temps jusqu'à l'événement ou lorsque l'on souhaite faire des extrapolations.

Le modèle exponentiel suppose une fonction de risque constante dans le temps avec $\lambda(t) = \lambda$ [Friedman, 1982]. Les modèles exponentiels par morceaux ont des risques constants par morceaux avec $\lambda(t) = \lambda_j$ lorsque $s_{j-1} \leq t < s_j$ pour des intervalles prédéfinis, $0 = s_0 < s_1 < \dots < s_J = \infty$. Cela conduit à des taux d'occurrence/exposition spécifiques à chaque intervalle. L'hypothèse de risque constant est jugée restrictive par Sabathé et al. [2020] et semble dans les faits rarement justifiée. Cependant, il est à la base du calcul des taux simples d'occurrence/exposition.

Une extension simple du modèle exponentiel est le modèle de Weibull avec une fonction de risque qui varie avec le temps avec $\lambda(t) = \lambda \alpha t^{\alpha-1}$ [Carroll, 2003]. Ce modèle, bien que rarement utilisé en pratique, est mathématiquement plus flexible, permettant des fonctions de risque croissantes, constantes et décroissantes.

I.3 Méthodes adaptées aux événements récurrents

I.3.1 Au-delà du premier événement

Les processus stochastiques générant des événements du même type et de manière répétée dans le temps sont appelés processus d'événements récurrents (Figure I.4). Les données résultant de ces processus sont appelées données d'événements récurrents (*recurrent event data*). L'analyse des événements récurrents a beaucoup progressé au cours des dernières décennies [Andersen et al., 1993, Prentice and Kalbfleisch, 2003, Cook and Lawless, 2007]. Cette évolution a été principalement motivée par des études biomédicales dans lesquelles les individus sont soumis à des événements répétés.

I.3.2 Composantes des événements récurrents

Nous considérons des sujets indépendants $i = 1, \dots, n$ et, pour le sujet i , $T_{i1} < T_{i2} < \dots \in [0, \infty)$ sont les moments de survenue de l'événement récurrent et T_{ij} est le moment d'entrée dans l'état j dans la Figure I.4.

I.3.2.1 Échelles temporelles

Les méthodes d'événements récurrents peuvent être spécifiées selon deux échelles temporelles différentes : l'échelle temporelle calendaire (ou temps total) et l'échelle temporelle par intervalle [Cook and Lawless, 2007, Kelly and Lim, 2000]. Le temps calendaire correspond au temps mesuré à partir de la date d'origine. Dans le cas du temps par intervalle, le temps est remis à zéro après chaque événement récurrent et l'échelle temporelle se base sur le temps écoulé depuis l'événement précédent.

I.3.2.2 Ensemble des sujets à risque

Les ensembles des sujets à risque à l'instant t regroupent tous les sujets qui sont susceptibles de subir un j ème événement à l'instant t . Trois types d'ensembles de sujets à risque sont possibles : les ensembles sans restriction, les ensembles semi-restreints et les ensembles restreints [Kelly and Lim, 2000]. Ils sont définis tels que :

- L'ensemble des sujets à risque est dit *non restreint* si tous les intervalles à risque contribuent à l'ensemble des sujets à risque pour n'importe quel événement, indépendamment du nombre d'événements précédents.
- L'ensemble de sujets à risque est dit *restreint* si le j ème ensemble de sujets à risque n'inclut que les intervalles à risque du j ème événement des individus qui ont déjà subi $(j - 1)$ événements. Cela signifie que seuls les sujets ayant déjà connu $(j - 1)$ événements sont considérés comme exposés au risque d'un j ème événement.
- Un ensemble des sujets à risque est dit *semi-restreint* si les ensembles des sujets à risque ont des risques spécifiques à l'événement, mais permettent aux individus qui ont moins de $(j - 1)$ événements d'être à risque pour un j ème événement. Les individus sont considérés comme étant à risque pour tous les événements à partir de la date d'origine. Un individu qui a déjà subi un événement, par exemple, est considéré comme exposé simultanément à un deuxième, un troisième, ... événement.

Note : Il n'est pas possible de combiner une échelle temporelle par intervalle et un ensemble de risques semi-restreint.

I.3.2.3 Processus de comptage

Supposons que n processus d'événements récurrents commençant à $t = 0$ soient observés. T_{ij} est défini comme le j ème temps d'événement pour l'individu i , $j = 1, 2, \dots$ et

$i = 1, \dots, n$ avec $T_{ij} \in [0, \infty)$ et $T_{i1} < T_{i2} < T_{i3} < \dots$. Les données d'événements récurrents sont formulées à l'aide de processus de comptage.

Définition I.3.1 (Processus de comptage) *Un processus stochastique continu à droite $N = \{N(t) : 0 \leq t < \infty\}$ est un processus de comptage si $N(t) \in \mathbb{N}$ est le nombre d'événements survenus jusqu'au temps t . Un processus de comptage est de saut de taille 1 aux temps d'événement, est constant entre deux événements et $N(0) = 0$.*

Un processus de comptage pour le sujet i est noté $N_i(t) = \sum_{j=1}^{\infty} \mathbb{1}(T_{ij} \leq t)$ sur $[0, t]$. Le nombre total d'événements au temps t est $N(t) = \sum_{i=1}^n N_i(t)$ et $dN(t) = N(t+dt) - N(t)$ est l'incrément de $N(t)$ sur l'intervalle $[t, t+dt]$ soit le nombre d'événements survenus sur ce même intervalle.

L'historique du processus est généralement un élément important dans l'analyse des événements récurrents et il constitue la base de la différenciation entre les méthodes d'événements récurrents conditionnels et marginaux. Cet historique comprend schématiquement toutes les informations qui ont été générées par le processus de comptage depuis la date d'origine jusqu'à l'instant t . Il contient par exemple les réalisations antérieures du processus de comptage et fournit ainsi des informations sur l'occurrence et la chronologie des événements précédents. On peut plus formellement définir cet historique, ou filtration.

Définition I.3.2 (Historique) *Une filtration, ou un historique, $\{\mathcal{N}(t)\}_{t \geq 0}$ est une famille croissante de tribus telles que $\mathcal{N}(t) := \sigma((N_i(s))_{s \leq t} : i = 1, \dots, n)$ sont générées à partir du processus de comptage $N_i(t)$, pour $i = 1, \dots, n$. En d'autres termes, $\mathcal{N}(t)$ est une tribu pour chaque t et si $s \leq t$, alors $\mathcal{N}(s) \subset \mathcal{N}(t)$. Cela signifie que la quantité d'informations sur le passé augmente au fur et à mesure du temps.*

Alors que la tribu $\mathcal{N}(t)$ représente l'information disponible à l'instant t , la filtration $\{\mathcal{N}(t)\}_{t \geq 0}$ présente l'évolution de l'information au cours du temps. Le processus de comptage $N_i(t)$ pour $i = 1, \dots, n$ est adapté à l'historique $\{\mathcal{N}(t)\}_{t \geq 0}$, ce qui signifie qu'à l'instant t les réalisations de $N_i(s)$ sont connues pour tout $s \leq t$, c'est-à-dire que $N_i(s) \in \mathcal{N}(t)$ pour tout $s \leq t$.

Dans le cas d'une censure à droite, les événements récurrents des processus N_i ne sont pas tous observés. Soit $C_i \in [0, \infty)$ le temps de censure pour l'individu i , $i = 1, \dots, n$. Les n individus étudiés sont alors observés sur l'intervalle de temps $[0, C_i]$, où $t = 0$ correspond au début du processus d'événement récurrent. Le nombre total de sujets à risque au temps t est $Y(t) = \sum_{i=1}^n Y_i(t)$, avec $Y_i(t) = \mathbb{1}(C_i \geq t) \in \{0, 1\}$ pour $i = 1, \dots, n$ une fonction indicatrice indiquant si l'individu i est sous observation et exposé à un risque d'événement juste avant le temps t .

Par ailleurs, pour des périodes de suivi spécifiques à l'individu i sur $[0, C_i]$, on considère $Z_i(t) = \{Z_i(u) : 0 \leq u \leq t\}$ le processus de covariables qui contient p covariables dépendantes ou indépendantes du temps jusqu'au temps t , avec $Z_i(t) = (Z_{i1}(t), \dots, Z_{ip}(t)) \in \mathbb{R}^p$ et $t \leq C_i$.

I.3.2.4 Processus de Poisson

Pour les modèles de Poisson, le processus de comptage sous-jacent est un processus de Poisson avec $\mathbb{E}[N(t)] = \text{Var}(N(t))$.

Définition I.3.3 (Processus de Poisson) *Un processus de comptage $N(t) = \{N(t) : 0 \leq t < \infty\}$ est un processus de Poisson non homogène de paramètre $\alpha(t)$, si*

- $N(0) = 0$;
- Étant donné $T_1 < T_2 < T_3 < T_4$, $N(T_1, T_2)$ est indépendant de $N(T_3, T_4)$, où $N(T_1, T_2) := N(T_2) - N(T_1)$ est le nombre d'événements dans $(T_1, T_2]$, respectivement. Autrement dit, si (T_1, T_2) et (T_3, T_4) sont des intervalles ininterrompus, alors $N(T_1, T_2)$ et $N(T_3, T_4)$ sont indépendants.
- Pour $0 \leq T_1 < T_2$, $N(T_1, T_2)$ suit une loi de Poisson de moyenne $\mu_{\mathcal{P}}(T_1, T_2) = \mu_{\mathcal{P}}(T_1) - \mu_{\mathcal{P}}(T_2) = \int_{T_1}^{T_2} \alpha(u)du$, avec $\mu_{\mathcal{P}}(T_1) = \int_0^{T_1} \alpha(u)du$ et $\mu_{\mathcal{P}}(T_2) = \int_0^{T_2} \alpha(u)du$. On a alors

$$N(T_1, T_2) \sim \text{Poisson}\left(\int_{T_1}^{T_2} \alpha(u)du\right) = \text{Poisson}\left(\mu_{\mathcal{P}}(T_1, T_2)\right), \quad (\text{I.17})$$

et pour $j \in \mathbb{N}$: $\mathbb{P}(N(t) = j) = \frac{\mu_{\mathcal{P}}(t)^j}{j!} \cdot \exp(-\mu_{\mathcal{P}}(t))$ avec $N(t) = N(0, t)$.

Un processus de Poisson $\{N(t) : 0 \leq t < \infty\}$ est dit homogène si les intervalles de temps entre les événements successifs sont des variables aléatoires indépendantes et identiquement distribuées, avec $\alpha(t) \equiv \alpha$ et $\alpha > 0$ ne dépend pas de t . Cela signifie que la loi de $\{N(t) - N(s)\}_{s \leq t}$ ne dépend ni de t ni de s , mais seulement de la longueur de l'intervalle $]s, t[$.

I.3.3 Modélisation des données avec événements récurrents

Selon Therneau et al. [2000], la prise en compte des événements récurrents devrait améliorer la qualité de l'estimation des effets au moment de la modélisation.

La corrélation intra-individuelle des événements récurrents pour un même individu peut être considérée suivant différentes approches : conditionnelle, marginale et fragilité. Les méthodes conditionnelles supposent que la dépendance entre les événements récurrents est entièrement expliquée par des covariables variant dans le temps. Cela signifie que l'incrément temporel entre les événements est conditionnellement non corrélé aux covariables observées. A l'inverse, les modèles marginaux supposent l'indépendance entre les événements récurrents d'un même individu. Enfin, l'approche par fragilité introduit un effet aléatoire ou un terme de fragilité dans le modèle des événements récurrents qui induit une dépendance entre les temps des événements récurrents.

I.3.3.1 Les modèles conditionnels

La plupart des modèles conditionnels sont basés sur l'intensité, correspondant à la probabilité instantanée de présenter un événement dans un petit intervalle de temps, conditionnellement aux événements passés, soit l'historique du patient.

Soit $\mathcal{N}(t)$ une tribu pour chaque temps t , si $s \leq t$ alors $\mathcal{N}(s) \subseteq \mathcal{N}(t)$. La quantité d'historique augmente bien avec le temps. L'intensité des modèles conditionnels s'écrit alors :

$$r(t) = \lim_{dt > 0} \frac{\mathbb{P}(dN(t) = 1 | \mathcal{N}(t-))}{dt}, \quad (\text{I.18})$$

avec $\mathcal{N}(t-) = \sigma((N_i(s))_{s < t} : i = 1, \dots, n)$ l'historique juste avant t . La principale caractéristique de ces modèles réside dans la dépendance entre les événements répétés capturée par des covariables dépendantes du temps [Andersen and Gill, 1982, Aalen et al., 2004]. Cela implique que les incrément de temps entre les événements sont conditionnellement non corrélés, compte tenu des covariables.

Le modèle d'Andersen-Gill Le modèle de comptage de processus d'Andersen-Gill (AG) est le plus utilisé et généralise le modèle de Cox [Andersen and Gill, 1982]. Le modèle s'écrit pour le sujet i :

$$r_i(t) = Y_i(t) \cdot r_0(t) \cdot \exp(\beta^T Z_i(t)). \quad (\text{I.19})$$

La fonction d'intensité de base $r_0(t)$ est commune pour tous les événements et $Y_i(t)$ indique si le sujet est à risque. En présence du processus des sujets à risque ainsi que des covariables, l'historique se réécrit $\mathcal{N}(t) = \sigma((N_i(s), Y_i(s), Z_i(s))_{s \leq t} : i = 1, \dots, n)$ et prend en considération tant les censures passées que l'évolution ou non des valeurs des covariables.

Dans l'équation I.19, le risque instantané de subir un événement au temps t depuis l'entrée dans l'étude est supposé le même. Cette hypothèse forte implique que les événements récurrents sont supposés indépendants. Si cette hypothèse est respectée, le risque peut être estimé en utilisant les temps de chaque événement observé. Ainsi, un patient apporte plus d'information en fonction du nombre d'événements observés individuellement.

Le modèle AG vise alors à estimer la même quantité que le modèle de Cox de l'équation I.10. Cependant, l'estimation est basée sur plus d'information dans la mesure où un individu qui a subi un événement reste à risque pour d'autres événements subséquents. La vraisemblance partielle de l'équation (I.14) est ainsi basée sur un plus grand nombre d'événements et sur un ensemble à risque tel que

$$R^{AG}(t) := \{l, l = 1, \dots, n : \exists j \in \{1, \dots, n_l\}, T_{lj} \geq t\}, \quad (\text{I.20})$$

avec n_i le nombre de temps distincts d'événement, soit le nombre d'événements, pour l'individu i . La vraisemblance partielle s'écrit alors

$$\mathcal{L}_p(\beta) = \prod_{i=1}^n \prod_{j=1}^{n_i} \left(\frac{\exp(\beta^T Z_i(T_{ij}))}{\sum_{l,j \in R^{AG}(T_{ij})} Y_{lj}(T_{lj}) \exp(\beta^T Z_i(T_{lj}))} \right)^{\delta_{ij}}. \quad (\text{I.21})$$

C'est un modèle approprié lorsque les corrélations entre les événements pour chaque individu sont induites par des covariables mesurées. Le modèle AG est particulièrement adapté lorsque l'objectif est d'évaluer l'effet global d'un traitement sur l'intensité des événements récurrents, avec une dépendance entre les événements due aux variables dépendantes du temps.

Le modèle de Prentice, Williams et Peterson Le modèle de Prentice, Williams et Peterson (PWP) est un modèle AG dit stratifié [Prentice et al., 1981]. Le risque de base $r_0(t)$ de l'équation (I.19) n'est plus supposé le même entre les événements, ainsi l'intensité r_{ij} du j ème événement pour l'individu i dépend de l'historique. Le modèle PWP peut être considéré de deux façons, suivant la manière de prendre en compte le temps : (i) le modèle PWP-TT utilise le temps écoulé entre la date d'origine et chaque événement récurrent, ou (ii) le modèle PWP-GT utilise le temps écoulé entre deux événements récurrents. Le modèle s'écrit :

$$r_{ij}(t) = \begin{cases} Y_i(t) \cdot r_{0j}(t) \cdot \exp(\beta_j^T Z_i(t)) & \text{pour le modèle PWP-TT,} \\ Y_i(t) \cdot r_{0j}(t - t_{j-1}) \cdot \exp(\beta_j^T Z_i(t)) & \text{pour le modèle PWP-GT.} \end{cases} \quad (\text{I.22})$$

Pour chaque événement récurrent $j = 1, \dots, n_i$, n_i est le nombre total d'événements vécus par l'individu i sur $[0, C_i]$ tel que $N_i(C_i) := n_i$. De cette façon, une fonction d'intensité est modélisée avec une fonction d'intensité de base spécifique r_{0j} . L'intensité pour un événement récurrent peut changer suite à un événement précédent. Un individu n'est considéré à risque du j ème événement que s'il a connu un $(j-1)$ ème événement antérieur. De cette façon, dans ces deux approches temporelles, l'intensité au temps t pour la j ème récurrence est conditionnelle à tous les événements précédents. La vraisemblance du modèle s'écrit :

$$\mathcal{L}_p(\beta) = \prod_{i=1}^n \prod_{j=1}^{n_i} \left(\frac{\exp(\beta^T Z_i(T_{ij}))}{\sum_{l,j \in R^{PWP}(T_{lj})} Y_{lj}(T_{lj}) \exp(\beta^T Z_i(T_{lj}))} \right)^{\delta_{ij}}. \quad (\text{I.23})$$

Les ensembles de sujets à risque sont définis séparément pour chaque strate, tels que :

$$R_j^{PWP}(t) := \begin{cases} l, l = 1, \dots, n : T_{l(j-1)} < t \leq T_{lj} & \text{pour le modèle PWP-TT,} \\ l, l = 1, \dots, n : (T_{lj} - T_{l(j-1)}) \geq t & \text{pour le modèle PWP-GT.} \end{cases} \quad (\text{I.24})$$

où à nouveau T_{ij} sont les temps d'événement distincts pour l'individu i , et pour l'événement j ème survenant $j = 1, \dots, n_i$.

Lorsque le nombre d'événements récurrents, et donc de strates, augmente dans le modèle PWP, le nombre de sujets à risque par strate devient relativement faible. Cela peut entraîner des estimations instables, c'est pourquoi il est habituellement nécessaire de limiter les données à un nombre spécifique d'événements récurrents [Kelly and Lim, 2000]. Pour cette raison, le modèle PWP est utile dans les situations où le nombre de récidives par sujet est plutôt faible, ou bien lorsque le risque d'événements est différent suivant les récurrences [Amorim and Cai, 2015].

I.3.3.2 Les modèles à fragilité

Les modèles basés sur l'intensité reposent sur l'expression explicite de la dépendance entre les événements répétés au moyen de covariables dépendantes du temps et/ou d'une stratification dépendante du temps. Des effets aléatoires peuvent également être incorporés dans l'équation I.10 pour induire une dépendance à l'égard de l'historique des événements précédents. En effet, la corrélation intra-référence peut être due à de l'hétérogénéité non mesurée. Les modèles de fragilité permettent d'introduire une covariable aléatoire dans le modèle qui induit une dépendance entre les temps d'événements récurrents. Les modèles à fragilité sont également basés sur l'intensité et on écrit :

$$r(t|U) = \lim_{dt>0} \frac{\mathbb{P}(dN(t) = 1 | \mathcal{N}(t-), U)}{dt}, \quad (\text{I.25})$$

avec U l'effet aléatoire non observé.

Le modèle de fragilité le plus couramment utilisé est un modèle de fragilité partagée avec des effets aléatoires u_i de distribution Gamma généralement, et $u_i \sim \Gamma(1, \theta)$. Le modèle est alors exprimé comme suit :

$$r_i(t|u_i) = r_0(t) \cdot u_i \cdot \exp(\beta^T Z_i), \quad (\text{I.26})$$

où u_i est le terme de fragilité qui capture l'hétérogénéité non observée et la dépendance entre les événements récurrents au sein des observations d'un même patient i .

Définition I.3.4 (Distribution Gamma) Soit U une variable aléatoire absolument continue. U suit une distribution Gamma avec un paramètre d'échelle ψ^{-1} et un paramètre de forme ψ^{-1} , soit $U \sim \Gamma(\psi^{-1}, \psi^{-1})$, si :

- Sa fonction de densité de probabilité est $g_U(u) = \frac{\exp(-\psi^{-1})u^{\psi^{-1}-1}}{\psi^{\psi^{-1}}\Gamma(\psi^{-1})}$ pour $u > 0$;
- La moyenne et la variance de U sont respectivement $\mathbb{E}[U] = 1$ et $\text{Var}(U) = \psi$.

La vraisemblance du modèle s'écrit :

$$\begin{aligned} \mathcal{L}_p(\beta) = \prod_{i=1}^n \int_{-\infty}^{+\infty} & \left(\left[\prod_{j=1}^{n_i} (r_0(t) \cdot u_i \cdot \exp(\beta^T Z_i(t)))^{\delta_i(T_{ij})} \right] \right. \\ & \left. \cdot \exp\left(-\int_0^{T_{ij}} Y_i(t) \cdot r_0(t) \cdot u_i \cdot \exp(\beta^T Z_i(t)) dt\right) g(u_i) du_i \right). \end{aligned} \quad (\text{I.27})$$

L'estimation des coefficients varie pour chaque événement lorsque le terme de fragilité est significatif. Ces modèles sont indiqués lorsque la susceptibilité hétérogène au risque d'événements récurrents peut être capturée par un effet aléatoire.

I.3.3.3 Les modèles marginaux

La volonté de modéliser la distribution de l'ensemble du processus des événements récurrents entraîne le risque d'une mauvaise spécification du modèle. Un modèle alternatif

est le modèle marginal, qui peut être interprété en termes de nombre moyen d'événements lorsqu'il n'y a pas de covariables dépendantes du temps.

Ici, l'attention est restreinte à certains paramètres marginaux. Par exemple, la moyenne marginale $\mu(t) = \mathbb{E}[N_i(t)]$ où $\{N_i(t) : 0 \leq t < \infty\}$ est le processus de comptage des événements récurrents [Lawless and Nadeau, 1995, Lin et al., 2000]. Des modèles pour les distributions marginales des temps d'événement ont également été étudiés [Wei et al., 1989]. Dès lors que le processus de comptage n'est pas basé sur l'intensité $r(t)$, on parle de taux marginal.

La **fonction de taux** est la probabilité instantanée marginale de survenue d'un événement au temps t et ne dépend plus des événements passés. La fonction de taux s'écrit :

$$\rho(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(dN(t) = 1)}{dt}. \quad (\text{I.28})$$

Il existe deux types de modèles marginaux, des extensions du modèle de Cox telles que les modèles de Lee et al. [1992] et de Wei et al. [1989], et les équations d'estimation généralisées (GEE) [Twisk et al., 2005]. Seul le modèle de Wei et al. [1989] est présenté ici, car le modèle de Lee et al. [1992] est plus approprié pour les données en grappes (*cluster*) et le modèle GEE ne prend pas en compte le temps de survenue des événements [Charles-Nelson et al., 2019].

Le modèle de Wei, Lee et Weissfeld (WLW) est également stratifié tel que décrit pour le modèle PWP avec échelle temporelle calendaire et utilise un risque de base spécifique à chaque événement [Wei et al., 1989]. Le modèle s'écrit :

$$\rho_i(t) = \rho_{0k}(t) \exp(\beta^T Z_i). \quad (\text{I.29})$$

Contrairement au modèle PWP (équation I.22), un individu est à risque pour chaque événement récurrent tant qu'il est observé. Par conséquent, un individu est à risque d'un événement subséquent même s'il n'y a pas eu d'événement antérieur. Ainsi, la structure de dépendance entre les événements observés pour un individu n'est pas spécifiée. L'ensemble des individus à risque est défini par strate comme suit :

$$R_j^{WLW}(t) := \{l, l = 1, \dots, n : \exists j \in \{1, \dots, n_l\}, T_{lj} \geq t\}. \quad (\text{I.30})$$

Toutefois, contrairement à la définition ci-dessus, si le nombre d'événements observés pour un individu n_i est inférieur au nombre maximal d'événements comptés n_{max} , des temps d'événements "artificiels" $T_{ij} := T_{in_i}, j > n_i$ sont définis avec un indicateur d'événement $\delta_{ij} = 0$ pour ces cas.

La vraisemblance partielle pour ce modèle est :

$$\mathcal{L}_p(\beta) = \prod_{i=1}^n \prod_{j=1}^{n_i} \left(\frac{\exp(\beta^T Z_i(T_{ij}))}{\sum_{l,j \in R^{WLW}(T_{lj})} Y_{lj}(T_{lj}) \exp(\beta^T Z_i(T_{lj}))} \right)^{\delta_{ij}}. \quad (\text{I.31})$$

De la même façon que le modèle stratifié PWP, plus le nombre de strates prises en considération est élevé, moins il reste d'individus par strate, ce qui se traduit par une estimation

de l'effet spécifique à la strate de faible précision. Par conséquent, lors de l'application du modèle de Wei-Lin-Weissfeld, le nombre d'événements par patient à prendre en compte dans l'analyse doit être limité en fonction du nombre total de patients.

La **fonction moyenne cumulée** (*mean cumulative function*, MCF) est le nombre marginal attendu d'événements dans $[0, t]$ et s'écrit :

$$\mu(t) = \mathbb{E}[N(t)] = \int_0^t \rho(u)du. \quad (\text{I.32})$$

À partir de l'équation I.32, on a $d\mu(t) = \rho(t)dt$ et $\mathbb{E}[dN(t)] = d\mu(t)$ pour $\mu(t)$ continue. Ainsi on a $\mathbb{E}[dN(t) - d\mu(t)] = 0$. Pour les n individus, on peut alors écrire $\sum_{i=1}^n Y_i(t)(dN_i(t) - d\mu(t)) = 0$, ce qui mène à l'estimateur non-paramétrique de Cook and Lawless [1997] suivant :

$$\hat{\mu}(t) = \int_0^t d\hat{\mu}(u)du = \int_0^t \frac{\sum_{i=1}^n Y_i(t)dN_i(t)}{\sum_{i=1}^n Y_i(t)}. \quad (\text{I.33})$$

L'estimateur $\hat{\mu}(t)$ est similaire à celui de Nelson-Aalen de l'équation (I.8) lorsque l'on étudie le temps jusqu'à la survenue du premier événement. La MCF de l'équation I.33 correspond alors au nombre estimé d'événements subis par l'individu i à l'instant t .

Le test non-paramétrique associé pour comparer deux MCF est le pseudo-score test. Soit $\{N_{ki}(t) : 0 \leq t < \infty\}$ le processus de comptage de l'individu i dans le groupe de traitement k . $Y_{ki}(t)$ est le processus à risque indiquant si l'individu i est dans le groupe de traitement k et à risque à $t-$, avec $k = 1$ et $k = 2$ désignant le groupe de traitement. Le processus à risque agrégé dans le groupe k est défini comme $Y_k = \sum_{i=1}^{n_k} Y_{ki}(t)$. Les fonctions de moyenne et de taux dans le groupe k sont par $E[dN_{ki}(t)] = \rho_k(t)dt$ et $E(N_{ki}(t)) = \mu_k(t)$ pour $i = 1, \dots, n_k$. La statistique de test est :

$$W(\tau) = \int_0^\tau \frac{Y_1(u)Y_2(u)}{Y_1(u) + Y_2(u)}(d\hat{\mu}_1(u) - d\hat{\mu}_2(u)), \quad (\text{I.34})$$

avec $\tau > 0$ est la durée maximale de suivi [Cook and Lawless, 2007]. La p-valeur est obtenue en référence à une distribution du χ^2 avec 1 degré de liberté pour deux groupes. La p-valeur obtenue indique la significativité statistique de la différence entre les MCF.

Les modèles marginaux considèrent tous les événements récurrents d'un même sujet comme un processus de comptage unique et ne nécessitent pas de covariables dépendantes du temps pour refléter l'historique, ce modèle est donc plus souple et plus parcimonieux que le modèle AG.

I.3.4 En présence d'un événement terminal

Lorsque les informations sur le passé sont "incomplètes" et que le processus d'événements récurrents ne peut être entièrement spécifié, les approches marginales sont préférées [Andersen et al., 2019].

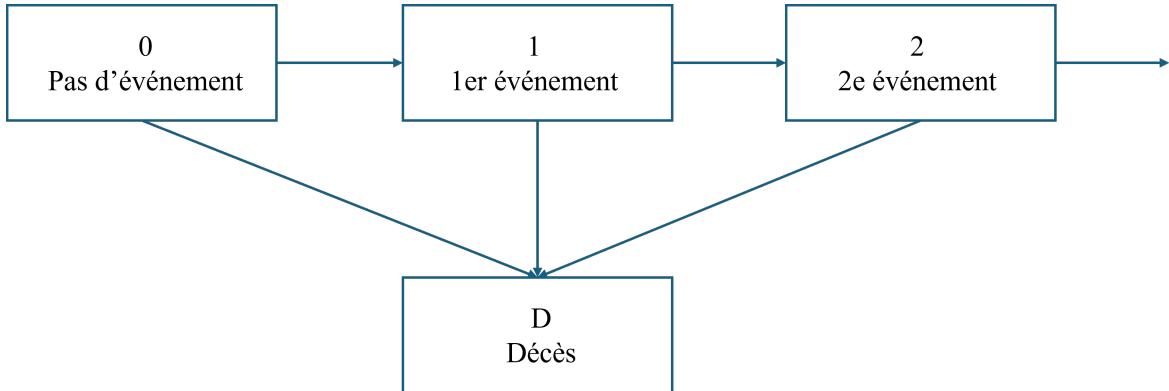


FIGURE I.5 – Processus d'événements récurrents avec événement terminal

Dans de nombreuses applications cliniques des méthodes pour les événements récurrents, des risques compétitifs peuvent être observés sous forme d'événements terminaux, dont la survenue empêche la réalisation d'autres événements [Andersen et al., 2012, Austin et al., 2016]. Ainsi, le décès d'un patient a pour conséquence qu'aucune nouvelle admission à l'hôpital, aucun nouveau cancer, etc., ne pourrait survenir. L'hypothèse de la censure non-informative n'est plus respectée car le processus de décès (jusqu'alors inclus dans le processus de censure) n'est plus indépendant du processus d'événements récurrents [Tai et al., 2002].

Pour pallier cela, les risques compétitifs ont été introduits, et l'événement décès entre en compétition avec les événements récurrents d'intérêt (Figure I.5).

I.3.4.1 Approche marginale de correction pour les risques compétitifs

En présence de décès, l'estimateur Nelson-Aalen défini en équation I.33 est surestimé car les événements récurrents d'intérêt ne peuvent se avoir lieu que tant que le patient est encore en vie. En confondant décès et censure, cela équivaudrait à avoir une population immortelle avec des événements qui peuvent survenir à tout instant.

Nous avons donc besoin d'un estimateur pour $\mu(t)$ qui tienne compte de la mortalité. Soit $\rho_i(t) = \mathbb{E}[dN_i(t)|T_i \geq t]/dt$ la fonction de taux marginal sachant le sujet i en vie (T_i est le temps de survie globale pour le sujet i). La fonction cumulative correspondante $R(t) = \int_0^t \rho(u)du$ (en supposant que tous les sujets ont le même ρ défini en équation I.28) peut être estimée par l'estimateur de Nelson-Aalen. La moyenne marginale est :

$$\mu(t) = \mathbb{E}[N(t)] = \int_0^t S(u)dR(u), \quad (\text{I.35})$$

où $S(\cdot)$ est la fonction de survie à estimer par Kaplan-Meier (vu en Section I.2.1). L'estimateur suggéré par Cook and Lawless [1997] et étudié par Ghosh and Lin [2000] s'écrit :

$$\hat{\mu}(t) = \int_0^t \hat{S}(u-)d\hat{R}(u) = \int_0^t \hat{S}(u-) \frac{\sum_i Y_i(u)dN_i(u)}{\sum_i Y_i(u)}. \quad (\text{I.36})$$

Concernant la modélisation, [Ghosh and Lin \[2002\]](#) ont proposé un modèle de régression pour prendre en compte les covariables indépendantes du temps avec :

$$\mu(t|Z) = \mu_0(t) \cdot \exp(\beta^T Z), \quad (\text{I.37})$$

avec $\mu_0(t)$ non spécifiée. Pour des covariables dépendantes du temps, $d\mu(t | Z) = \mathbb{E}[dN(t) | Z(s)]$ avec $s \geq 0$ et l'équation I.37 se généralise de la façon suivante :

$$d\mu(t|Z) = d\mu_0(t) \cdot \exp(\beta^T Z(t)). \quad (\text{I.38})$$

L'équation I.38 mène à

$$\mu(t|Z) = \int_0^t \exp(\beta^T Z(s)) d\mu_0(s), \quad (\text{I.39})$$

qui se réduit alors à l'équation I.37 lorsque les covariables sont indépendantes du temps.

Dans le cas le plus courant, les temps de censure ne sont pas connus pour tous les individus. Un modèle pour approximer $G(t) = \mathbb{P}(C > t|Z)$ (éventuellement sans Z) est alors nécessaire. Soit $\hat{G}(t)$ l'estimateur pour la distribution de censure, par Kaplan-Meier (Section I.2.1) si C est indépendant de Z , ou un estimateur basé sur le modèle de Cox (Section I.2.2) dans le cas contraire. [Ghosh and Lin \[2002\]](#) ont ensuite introduit les poids

$$w_i(t) = \frac{I(C_i \geq D_i \wedge t) \cdot G(t)}{G(T_i \wedge t)}, \quad (\text{I.40})$$

avec D_i le temps de décès. [Ghosh and Lin \[2002\]](#) ont ensuite montré que $\mathbb{E}[w_i(t)] = G(t)$, et l'estimation des poids de l'équation I.40 est donnée par

$$\hat{w}_i(t) = \frac{I(C_i \geq D_i \wedge t) \cdot \hat{G}(t)}{\hat{G}(T_i \wedge t)}. \quad (\text{I.41})$$

L'équation d'estimation pour β est

$$U(\beta) = \sum_{i=1}^n \int_0^{\inf} (Z_i(t) - \bar{Z}^G(t)) \cdot \hat{w}_i(t) dN_i(t), \quad (\text{I.42})$$

avec

$$\bar{Z}^G(t) = \frac{\sum_{l=1}^n w_l(t) \cdot Z_l(t) \cdot \exp(\beta Z_l(t))}{\sum_{l=1}^n w_l(t) \cdot \exp(\beta Z_l(t))}. \quad (\text{I.43})$$

Enfin, les variances et la fonction de moyenne de base peuvent être estimées, cette dernière par un estimateur de type Breslow (comme en Section I.2.2) :

$$\hat{\mu}_0(t) = \sum_{i=1}^n \int_0^t \frac{\hat{w}_i(u) dN_i(u)}{\sum_{l=1}^n \hat{w}_l(t) \cdot \exp(\beta Z_l(t))}. \quad (\text{I.44})$$

I.3.4.2 Autres approches

Modèles multi-états Ces modèles permettent de représenter de manière détaillée les transitions entre différents états de santé, y compris les événements terminaux [[Andersen and Keiding, 2002](#), [Therneau et al., 2020](#)]. Ils offrent une représentation complète du processus de santé-décès et permettent d'analyser les effets des covariables sur les transitions entre les états.

Modèles joints Les modèles joints permettent de modéliser simultanément la survie et les événements récurrents, en tenant compte de la dépendance potentielle entre ces deux processus [Huang and Liu, 2007, Mazroui et al., 2012, Che and Angus, 2016, Afonso et al., 2024]. Ils sont particulièrement utiles pour étudier l'impact des événements récurrents sur la survie et inversement.

I.4 Logiciels et outils statistiques

L'analyse de survie pour les événements récurrents s'effectue principalement sur R. Cela est possible également en SAS et Stata mais cela n'est pas détaillé ici. Sont présentés ci-après les principaux package R inscrits au CRAN, ainsi que les fonctions associées.

L'estimation non paramétrique de la fonction moyenne cumulée au sens de Nelson-Aalen est disponible dans les packages `reda` et `reReg` à l'aide de la fonction éponyme `mcf` et `Recur`, respectivement [Wang et al., 2022, Chiou et al., 2023]. Dans les deux packages, les représentations graphiques avec ou sans intervalles de confiance ainsi que le pseudo-score test pour la comparaison de deux MCF sont également disponibles. En présence d'un événement terminal, la fonction `recurrent.marginal.mean` du package `timereg` permet de construire l'estimateur non-paramétrique de Ghosh-Lin en intégrant la fonction de survie (de l'événement terminal, soit du décès) [Scheike and Zhang, 2011]. Alternativement, cette fonction de survie peut être estimée par modèle de Cox avec la fonction `recurrent.marginal.coxmean` de ce même package.

Pour la modélisation des événements récurrents basée sur l'intensité, la fonction `coxph` du package `survival` est utilisée avec plusieurs options (voir exemples de code en Annexe A.1) [Therneau, 2024]. La fonction `cph` du package `rms` permet de construire un modèle AG [Harrell, 2023]. Concernant les modèles à fragilité, le package `frailtypack` permet une modélisation avec des variables dépendantes du temps ainsi que la présence d'un événement terminal [Rondeau et al., 2012]. De même, le package `reda` dispose de la fonction `ratereg` pour la modélisation avec fragilité selon une distribution gamma.

Pour la modélisation marginale des événements récurrents, la fonction `coxph` du package `survival` peut également être utilisée pour les modèles de type WLW. En présence d'un événement terminal, la fonction `recreg` du package `mets` permet la modélisation à l'aide d'un modèle Ghosh-Lin [Scheike et al., 2014].

Pour la simulation de données de survie avec événements récurrents, les fonctions `simEvent` et `simEventData` du package `reda` permettent de simuler suivant des processus de Poisson (homogènes ou non) suivant plusieurs options : facteurs longitudinaux, fragilités, présence d'un événement terminal, modèles paramétriques, etc. La fonction `simGSC` du package `reReg` est également une possibilité, ou encore la fonction `simRecurrent` du package `mets`. Le package `simrec` permet par ailleurs de simuler le temps de recrutement pour chaque individu ainsi que des données pour des analyses intermédiaires [Jahn-Eimermacher et al., 2015].

I.5 Étude de cas

Nous proposons ici d'appliquer à un cas concret les modèles de survie spécifiques aux événements récurrents présentés en section I.3.

I.5.1 Design de l'étude

Cette étude rétrospective utilise les données du Programme de Médicalisation des Systèmes d'Information (PMSI), qui référence l'ensemble des hospitalisations en France. L'idée est d'évaluer le risque de réadmission à l'hôpital chez les patients ayant bénéficié d'une première chirurgie pour cancer digestif entre janvier 2020 et décembre 2022. Le suivi de chaque patient a duré six mois après leur intervention chirurgicale. Le design de cette étude de cas est représenté en Figure I.6.

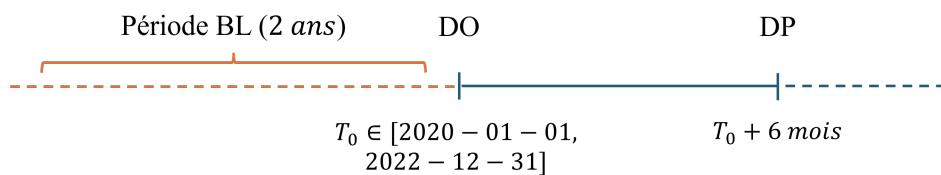


FIGURE I.6 – Design de l'étude sur les réadmissions à l'hôpital après une première chirurgie pour cancer digestif. BL = *baseline*; DO = Date d'origine; DP = Date de point.

Les patients inclus ont été censurés soit à la fin de la période de suivi, soit lorsqu'ils étaient perdus de vue. Les données comprenaient les comorbidités à l'inclusion (*baseline*), recueillies dans les 2 ans avant la première chirurgie à partir des diagnostics codés selon la Classification Internationale des Maladies (CIM-10). L'information sur les actes effectués à chaque réadmission était récupérée à partir de la Classification Commune des Actes Médicaux (CCAM). L'objectif principal de cette étude est de comprendre le risque de réadmission à l'hôpital, au global et par type de chirurgie.

I.5.2 Résultats

Au total, 255 732 patients ont bénéficié d'une chirurgie pour cancer digestif entre janvier 2020 et décembre 2022. Parmi eux, 205 666 (80,4%) patients ont eu au moins une réadmission au cours des six mois suivant leur chirurgie initiale. Le nombre médian (Q1 - Q3) était 1,0 (1,0 - 4,0) réadmissions. En premier lieu, notons que 121 854 (47,6%) des patients ont eu au moins deux réadmissions. L'analyse sur la première réadmission reviendrait à exclure 37,9% des événements observés. La Figure I.7 présente le nombre estimé des réadmissions postopératoires par type de chirurgie par estimateur de Nelson-Aalen.

Pour l'étape de modélisation, nous avons tronqué le jeu de données après le quatrième événement, en raison du faible nombre d'événements dans les strates ultérieures et pour

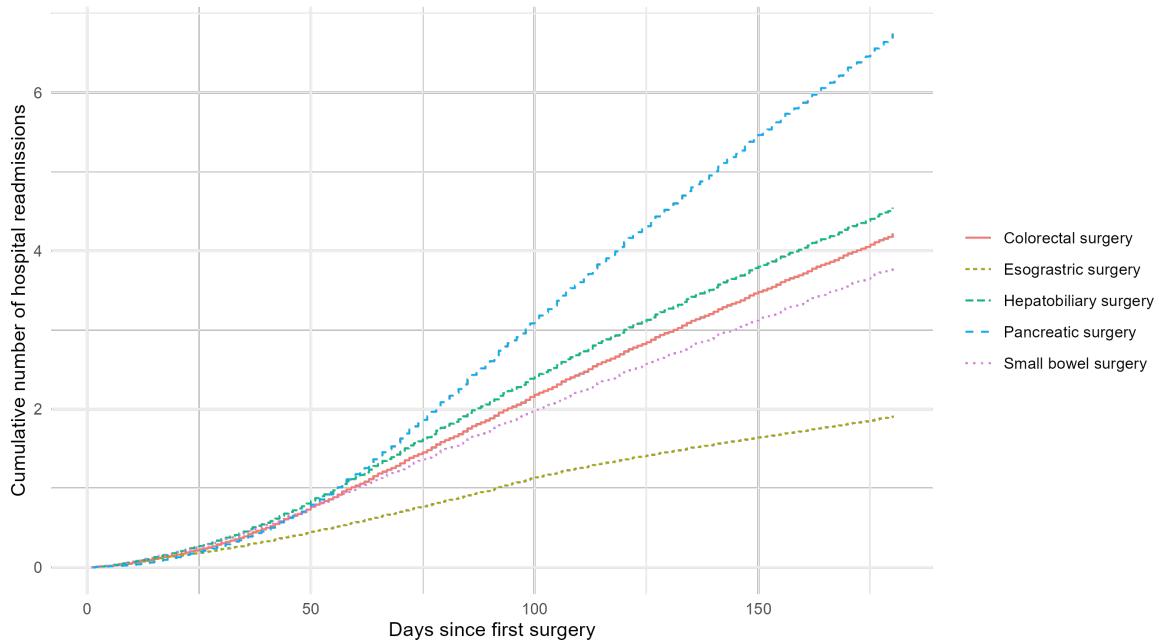


FIGURE I.7 – Nombre moyen estimé des réadmissions à l'hôpital par type de chirurgie pour les patients atteints de cancer digestif. MCF = *mean cumulative function*.

permettre la comparaison avec les modèles stratifiés. La structure des données et le code associé à chaque modèle sont décrits en Annexe A.1. À partir de la littérature scientifique, nous avons également inclus dans les modèles trois facteurs qui semblent associés au risque de réadmission postopératoire. Ces trois facteurs sont : le sexe, l'âge, et le statut diabétique.

Les résultats des hazard ratios sont présentés dans le Tableau I.1. Nous avons limité le nombre d'événements par patient à 4 réadmissions pour l'analyse.

Interprétation Les résultats des modèles s'interprètent de façon différente. Nous prenons l'exemple du statut diabétique du patient. On interprète alors :

- Pour le modèle AG, la probabilité d'avoir une réadmission depuis la chirurgie est

TABLE I.1 – Hazard ratios et intervalles de confiance à 95% de cinq modèles pour traiter les événements récurrents, ajustés sur le type de chirurgie

Modèle	Sexe : M	Âge	Diabète : Oui
AG	1.03 (1.00, 1.05)	1.00 (1.00, 1.00)	2.34 (2.13, 2.56)
PWP-TT	1.00 (0.99, 1.01)	1.00 (1.00, 1.00)	1.04 (1.01, 1.08)
PWP-GT	0.98 (0.97, 0.99)	0.99 (0.99, 0.99)	1.07 (1.03, 1.10)
Fragilité	1.02 (1.00, 1.04)	1.00 (1.00, 1.00)	2.37 (2.16, 2.59)
WLW	0.99 (0.98, 1.00)	0.99 (0.99, 1.00)	1.24 (1.19, 1.29)

AG = Andersen-Gill ; PWP-GT = Prentice, Willimas, et Petersen avec temps par intervalle ; PWP-TT = Prentice, Williams, et Petersen avec temps total ; WLW = Wei, Lin, et Wessfeld.

plus élevée chez les patients diabétiques par rapport aux non diabétiques avec un HR (IC95%) = 2,34 (2,13 - 2,56). Cela signifie que le délai de réadmission depuis la chirurgie est au global plus court pour ces patients par rapport aux patients non diabétiques.

- Pour le modèle PWP-TT, la probabilité de faire la k ème réadmission depuis la chirurgie chez les patients ayant fait la $(k - 1)$ ème réadmission est plus élevée chez les patients diabétiques avec un HR (IC95%) = 1,04 (1,01 - 1,08). Le délai de la k ème réadmission est plus court chez les patients diabétiques que chez les patients non diabétiques ayant fait la $(k - 1)$ ème réadmission.
- Pour le modèle PWP-GT, la probabilité de faire la k ème réadmission depuis la $(k - 1)$ ème réadmission chez les patients est plus élevée chez les patients diabétiques avec un HR (IC95%) = 1,07 (1,03 - 1,10). Le délai entre la k ème et la $(k - 1)$ ème réadmission est plus court chez les patients diabétiques que chez les patients non diabétiques.
- Pour le modèle à fragilité s’interprète de la même façon que le modèle AG, conditionnellement au terme de fragilité.
- Pour le modèle WLW, la probabilité de faire la k ème réadmission depuis la chirurgie pour tous les patients (ayant fait ou non la $(k - 1)$ ème réadmission) est plus élevée pour les patients diabétiques par rapport aux non diabétiques, avec un HR (IC95%) = 1,24 (1,19 - 1,29).

Dans les données disponibles, 4,02% des patients décèdent à l’hôpital dans les 6 mois après la chirurgie. En raison du faible nombre de décès, il semble que les estimations soient similaires en prenant en compte ou non l’événement terminal.

Discussion des résultats Les HR obtenus sont différents d’un modèle à l’autre en raison de la corrélation intra-individuelle et de la contribution à la log-vraisemblance. Les modèles AG et à fragilité rapportent une forte association entre le diabète et les réadmissions après chirurgie digestive. Les modèles PWP-TT et PWP-GT montrent une association beaucoup plus faible, indiquant que l’effet du diabète peut être atténué lorsque l’on prend en compte une stratification sur les récurrences des événements. Enfin, le modèle WLW indique une association intermédiaire, en tenant compte de la possibilité de différents risques pour chaque événement récurrent depuis la chirurgie.

TABLE I.2 – Récapitulatif des modèles présentés

	AG	PWP	Fragilité	WLW
Ensemble de sujets à risque				
Non restreint	X		X	
Restreint		X		
Semi-restreint				X
Échelle de temps				
Temps total	X	X	X	
Temps par intervalle		X		
Temps calendaire				X
Fonction				
Intensité	X	X	X	
Taux marginal				X
Interprétation				
Effet global	X		X	
Effet par événement		X		X

AG = Andersen-Gill; PWP = Prentice, Willimas, et Petersen; WLW = Wei, Lin, et Wessfeld.

I.6 Discussion

Le Tableau I.2 présente un récapitulatif des modèles présentés dans ce chapitre. Dans cette section, nous abordons les points forts et limites de chacune des approches (Section I.6.1), leurs implications cliniques (Section I.6.2) et proposons des axes de recherches actuels sur l’analyse des événements récurrents dans un cadre de survie (Section I.6.3).

I.6.1 Avantages et inconvénients des modèles pour les événements récurrents

L’avantage principal du modèle AG est sa capacité à analyser toutes les hospitalisations pour tous les individus et de mesurer l’effet global d’un facteur donné. Aussi, ce modèle permet de prendre en compte des covariables dépendantes du temps et des intervalles de risque discontinu. En revanche, les événements au sein des observations d’un même sujet sont indépendants conditionnellement aux covariables et le modèle peut être mal spécifié en l’absence de covariable temporelle [Ozga et al., 2018]. Enfin, l’hypothèse de proportionnalité des risques est trop forte en pratique pour ce modèle car elle impose un hazard ratio constant au cours du temps et commun d’un événement à l’autre pour l’ensemble des individus [Charles-Nelson et al., 2019].

Les modèles PWP sont recommandés s’il est raisonnable de supposer que la survenue du premier événement modifie la probabilité d’une récurrence. Ces modèles permettent également d’estimer les effets de chaque événement séparément. L’une des limites principales

de ces modèles est que les ensembles de risques pour les événements ultérieurs deviennent assez petits, ce qui rend les estimations instables [Prentice et al., 1981]. Par conséquent, les données doivent généralement être tronquées.

Les modèles de fragilité sont indiqués lorsqu'un effet aléatoire spécifique au sujet peut expliquer l'hétérogénéité non mesurée qui ne peut être expliquée par les seules covariables. De cette façon, l'interprétation des estimations est spécifique à l'individu. Cela induit également de spécifier l'effet aléatoire qui est paramétrique. Lorsque les effets aléatoires sont importants, un nombre plus faible d'événements semble être approprié [Amorim and Cai, 2015].

Enfin, les modèles marginaux sont particulièrement adaptés lorsque l'intérêt est de modéliser des nombres ou des taux moyens d'événements. Néanmoins, les modèles stratifiés WLW posent la même limite que les modèles conditionnels PWP sur la nécessité de définir en amont un nombre maximum d'événements par patient.

I.6.2 Implications cliniques

Nous avons vu les avantages et inconvénients des modèles pour prendre conscience de leur cadre d'utilisation. En réalité, ces modèles permettent de répondre à des questions cliniques différentes :

- Quel est l'effet du traitement sur le risque global (au sens de probabilité) de survenue d'événement ?
 - Les effets globaux sont mesurés par des modèles non stratifiés, comme le modèle AG ou à fragilité.
- Quel est l'effet du traitement à partir de la troisième récurrence ?
 - Les modèles stratifiés sont alors appropriés car ils apportent des estimations par récurrence.
- Quel est l'effet du traitement sur le nombre d'événements sur la période d'étude ?
 - Les modèles marginaux permettent d'estimer des nombres moyens ou des taux d'événements.

Ainsi, la quantité mesurée de l'effet est essentielle pour guider la modélisation des événements récurrents. Ce point est en fait très lié au concept d'estimand, qui désigne précisément la quantité à estimer pour un critère de jugement donné pour la mesure d'un effet traitement, récemment intégré dans les recommandations internationales d'études cliniques (ICH E9) [ICHE9, 2019].

Plus ces questions cliniques sont finement posées, plus les réponses apportées sont adéquates. En particulier, les implications cliniques sont conséquentes. Dans un premier lieu, l'analyse des événements récurrents mène à une meilleure compréhension des pathologies [Therneau et al., 2000]. Par exemple, dans le cas des chirurgies de cancer, l'analyse des complications postopératoires peut aider à comprendre les facteurs de risque et l'efficacité des interventions [Fico et al., 2023, Bona et al., 2024]. Deuxièmement, les analyses de survie

avec événements récurrents sont de plus en plus fréquentes en oncologie, comme [Galaznik et al. \[2022\]](#) avec l'analyse des différentes lignes de traitement pour les patients atteints de myélome multiple. Par ailleurs, les critères de jugement de type *survie sans progression* ne semblent plus adaptés pour les études sur les patients atteints de cancers récidivants [[Booth et al., 2023](#)]. L'analyse des multiples récidives peut aider à planifier les ressources nécessaires pour la gestion des récidives et les soins associés. Enfin, à des fins médico-économiques, les réadmissions à l'hôpital ont un impact significatif sur les coûts de santé, par exemple dans le cas du cancer colorectal métastatique [[Wick et al., 2011](#)]. Les analyses de survie qui prennent en compte ces événements récurrents peuvent être utilisées pour évaluer l'efficacité coût-efficacité de différentes options thérapeutiques.

I.6.3 Perspectives futures et axes de recherche

[Andersen and Pohar Perme \[2010\]](#) ont démontré l'efficacité des pseudo-observations pour l'analyse des données de survie. [Andersen et al. \[2019\]](#) puis [Erdmann et al. \[2023\]](#) ont récemment étendu aux analyses avec événements récurrents, en présence ou non d'un événement terminal. Pour les études non randomisées, les estimateurs et tests associés peuvent être biaisés en raison de la présence de facteurs de confusion. Pour palier cela, il est souvent question d'estimer l'effet causal moyen (*average causal effect*, ACE) dans l'ensemble de la population, c'est-à-dire la différence entre le résultat moyen dans la population cible si tous les sujets étaient dans le groupe de traitement et le résultat moyen si tous les sujets étaient dans le groupe de contrôle [[Rosenbaum and Rubin, 1983](#)]. En présence d'événements récurrents, [Gao and Zheng \[2016\]](#) ont proposé un effet causal moyen des compilateurs (*complier average causal effect*, CACE) qui est la différence entre les nombres moyens de récurrences dans les groupes de traitement et de contrôle au sein des compilateurs. [Su et al. \[2022\]](#) a récemment proposé des estimateurs construits à partir des pseudo-observations. Néanmoins, ces estimateurs sont basés sur l'hypothèse que l'attribution au traitement est indépendante des autres facteurs, mesurés ou non. Des travaux futurs sur le non-respect de cette hypothèse pour les études observationnelles sont à alors à envisager. Enfin, [Schmidli et al. \[2023\]](#), [Wei et al. \[2023\]](#), [Bühler et al. \[2023\]](#) ont récemment démontré l'importance des estimands dans les études cliniques pour traiter les questions autour des événements récurrents. En revanche, l'hypothèse selon laquelle le traitement et les événements récurrents et compétitifs sont indépendants n'est à ce jour pas encore levée.

Messages-clés de cet état de l'art

Ce chapitre a exposé les principes fondamentaux de l'**analyse de survie**, qualifiée de *classique* lorsqu'elle se focalise sur la survenue du premier événement, puis adaptée aux spécificités des données quand les patients expérimentent plusieurs événements. Ces derniers sont dits *récurrents* et requièrent une attention soutenue pour identifier les dépendances tant interindividuelles qu'intraindividuelles lors de la modélisation. Nous avons exposé deux approches majeures pour gérer les événements récurrents en survie : les **approches conditionnelles** qui se concentrent sur la modélisation de la fonction d'intensité des événements, en intégrant l'historique des événements passés (modèles d'Andersen-Gill, de Prentice, Williams et Peterson), et les **approches marginales** qui se focalisent sur la modélisation de la fonction de taux des événements, sans considérer les événements antérieurs (modèles de Lin, Wei et Ying, et de Wei, Lin et Weissfeld). De plus, nous avons discuté des cas en présence d'un **événement terminal** (notamment le décès) et de l'introduction de risques compétitifs, une situation fréquente dans les études en oncologie. En effet, le décès des patients avant la survenue d'autres événements peut altérer les estimations. Aussi, nous avons présenté une étude clinique portant sur les réadmissions hospitalières postopératoires en chirurgie digestive. Cette première application a permis de souligner les **similarités et divergences** entre les modèles. En conclusion, nous avons mis en évidence l'importance de la sélection du modèle **en adéquation avec la question clinique posée**. Chaque modèle requiert alors une interprétation spécifique des résultats obtenus. Il est donc essentiel de définir précisément ces points en **collaboration étroite avec les cliniciens impliqués**. Ce chapitre n'aborde pas l'utilisation de l'apprentissage automatique, qui a pourtant gagné du terrain dans l'analyse de survie. C'est le sujet que nous aborderons dans le prochain chapitre.

Chapitre II

Apprentissage pour données censurées

"Prediction is very difficult, especially if it's about the future!"

Niels Bohr (1922)

Sommaire

II.1 Concepts fondamentaux de l'apprentissage	45
II.1.1 Les différents types d'apprentissage	45
II.1.2 Sur- et sous-apprentissage	46
II.1.3 Évaluation des modèles	48
II.1.3.1 Optimisation des hyperparamètres et sélection de modèles	48
II.1.3.2 Le théorème <i>No free lunch</i>	50
II.2 Apprentissage pour les données censurées	50
II.2.1 Régressions pénalisées	51
II.2.2 Méthodes d'apprentissage automatique	52
II.2.2.1 Machines à vecteurs de support	52
II.2.2.2 Arbres de survie	53
II.2.2.3 Méthodes d'ensemble	53
II.2.2.4 Boosting	54
II.2.2.5 Apprentissage profond	55
II.2.3 Métriques de performance	55
II.2.3.1 Indice de concordance	55
II.2.3.2 Brier score intégré	57
II.2.3.3 Erreur absolue moyenne	58
II.3 Apprentissage avec événements récurrents	59
II.4 Discussion	83

Introduction

Les statistiques et l'apprentissage automatique (*machine learning*, ML) sont des disciplines distinctes bien que complémentaires. Les statistiques conventionnelles se concentrent sur la modélisation des relations entre variables explicatives et variable à expliquer et reposent sur des hypothèses plus ou moins fortes sur la distribution des données. Les algorithmes de ML mettent plutôt l'accent sur la capacité à bien prédire une valeur cible. Breiman [2001a] a encouragé l'intégration des deux approches afin de profiter de leurs avantages respectifs afin de traiter conjointement de manière optimale analyse de donnée et prédiction. Le ML peut être défini comme suit :

"[Machine learning is] the field of study that gives computers the ability to learn without being explicitly programmed." – Arthur Samuel (1959)

Une définition alternative et un peu plus récente donne plus de détails :

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." – Tom Mitchell (1997)

Par exemple, dans le domaine de l'oncologie, un programme d'apprentissage automatique peut être conçu pour estimer la probabilité de complication post-opératoire après une chirurgie de cancer, basé sur des données médicales historiques. Étant donné des exemples de patients ayant connu (ou non) une complication post-opératoire, le système apprend à identifier les individus à haut risque. Les exemples que le système utilise pour apprendre sont appelés des instances d'entraînement (ou échantillons). Dans ce cas, la tâche T est de prédire la complication post-opératoire pour de nouveaux patients, l'expérience E correspond aux données d'entraînement, et la mesure de performance P est à définir (par exemple, la proportion des cas de récidive correctement classés).

Exemple fictif

Pourquoi avoir recours aux algorithmes d'apprentissage ?

Dans un premier temps, nous considérons intuitivement une approche hypothético-déductive d'un algorithme programmé pour détecter les complications postopératoires après une chirurgie de cancer, illustré en Figure II.1.¹ Les étapes sont :

1. **L'analyse des données** incluant les dossiers médicaux des patients ayant bénéficié d'une chirurgie pour cancer colorectal pourrait permettre de détecter que les patients qui développent des complications postopératoires ont souvent des niveaux élevés de CRP (Protéine C-réactive) et des antécédents de diabète.
2. **Un algorithme est développé**, par exemple, pour analyser les résultats de tests sanguins préopératoires pour détecter des niveaux de CRP supérieurs à une certaine valeur seuil, ou pour identifier les patients ayant un historique de diabète. Le programme

1. "[Une] méthode hypothético-déductive est une méthode scientifique qui consiste à formuler une hypothèse afin d'en déduire des conséquences observables futures (prévision), mais également passées (réintroduction), permettant d'en déterminer la validité." Source : [Wikipédia](#).

pourrait utiliser une règle simple : si un patient présente au moins deux des caractéristiques identifiées (CRP élevée, diabète, chirurgie prolongée, etc.), il est signalé comme étant à haut risque de complications postopératoires.

3. Après avoir **testé le programme sur un nouvel ensemble de patients**, l'algorithme manque certains patients à risque et inclut certains patients à faible risque malgré un bon niveau global de prédiction. Il est alors nécessaire retourner à la première étape pour identifier d'autres caractéristiques potentielles ou affiner les algorithmes existants, en ajoutant par exemple d'autres facteurs comme le score ASA (*American Society of Anesthesiologists*) ou la durée de l'anesthésie.

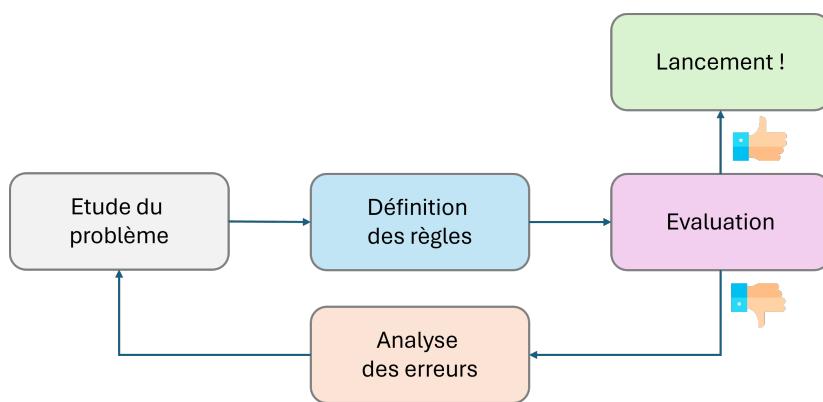
Le problème à résoudre n'étant pas simple à traiter, celui-ci pourrait bien devenir une longue liste de règles complexes. En revanche, un système de prédiction basé sur des techniques de ML apprend automatiquement quelles caractéristiques sont utiles afin de prédirer les complications postopératoires en détectant des schémas inhabituellement fréquents dans les exemples de complications par rapport aux exemples sans complications (Figure II.1(b)). Le programme est plus court, plus facile à maintenir et souvent plus précis. Par ailleurs, si les facteurs de risque de complications postopératoires évoluent avec le temps, ou si de nouveaux protocoles chirurgicaux se mettent en place, un programme basé sur une approche traditionnelle devra constamment être mis à jour pour refléter de tels changements. Un système de prédiction basé sur une approche d'apprentissage pourra automatiquement intégrer ces nouvelles informations et produire de nouvelles prédictions adéquatement (Figure II.1(c)).²

Si une valeur cible peut être déterminée à l'aide de règles simples, de calculs ou d'étapes prédéterminées programmables sans recours à l'apprentissage basé sur des données, l'utilisation de l'apprentissage devient superflue. En revanche, l'apprentissage automatique excelle là où les approches traditionnelles échouent : quand la tâche est trop complexe ou bien lorsqu'aucun algorithme connu n'existe.

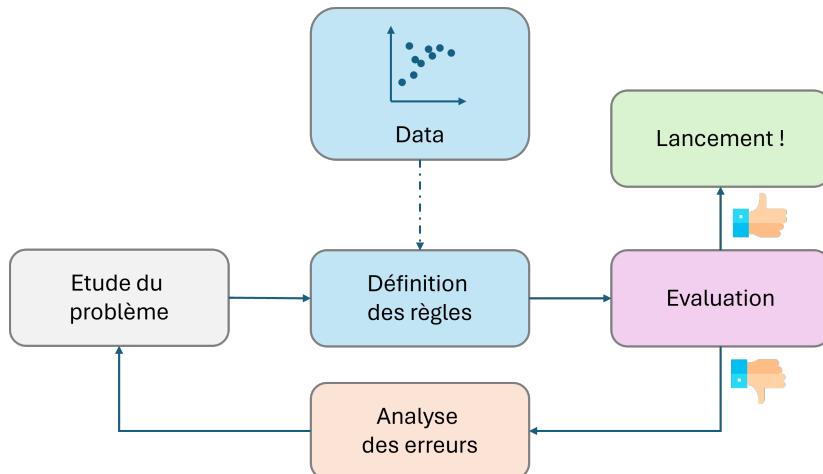
Par exemple, considérons l'analyse de textes issus des dossiers médicaux : un programme simple pourrait être conçu pour distinguer les mentions de "cancer du sein" et "cancer du poumon". Il est possible de remarquer que le terme "poumon" est souvent associé à des mots comme "bronche" ou "alvéole", tandis que le terme "sein" est souvent accompagné de mots comme "mammographie" ou "mastectomie". Un algorithme pourrait alors être développé manuellement pour mesurer la fréquence de ces mots clefs afin de distinguer les deux types de cancer. Cependant, cette solution n'est pas adaptée à un cas de figure où des milliers de termes médicaux sont extraits des dossiers de millions de patients provenant d'environnements variés dans lesquels les langages médicaux et les terminologies diffèrent [Wu et al., 2020]. La meilleure solution à date est de développer un algorithme apprenant par lui-même, en se basant sur un grand volume d'exemples de textes annotés pour chaque catégorie.

Aussi, les jeux de données cliniques sont de plus en plus volumineux, notamment en raison de l'apport des bases médico-administratives (on parle de données en grande échelle ou grande dimension). Les algorithmes de ML peuvent traiter et analyser ces grandes quantités de données pour découvrir des relations entre facteurs qui seraient difficilement identifiées.

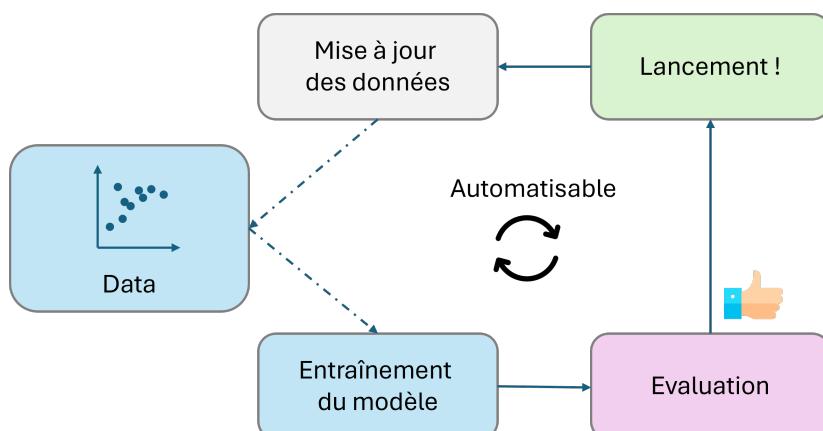
2. Les Figures II.1(a) à II.1(c) sont inspirées du Chapitre 1 de Géron [2022].



(a) Approche hypothético-déductive de solution de problème



(b) Approche par apprentissage de solution de problème



(c) Adaptation automatique pour prendre en compte les évolutions

FIGURE II.1 – Solution de problème par algorithmes

fiables autrement, comme la recherche de biomarqueurs ou de signatures génétiques [Bertsimas and Wiberg, 2020]. Les approches ML peuvent également contribuer à la médecine personnalisée, en profilant par exemple certains sous-groupes de patients, et en adaptant leur traitement [Kourou et al., 2015].

On peut dès lors envisager des algorithmes pour des problèmes de survie *classique* et chercher à prédire un risque de survie de la première complication post-opératoire. Par extension, on peut également imaginer aller au-delà de la première complication post-opératoire et chercher à prédire un risque de survie des multiples complications postopératoires.

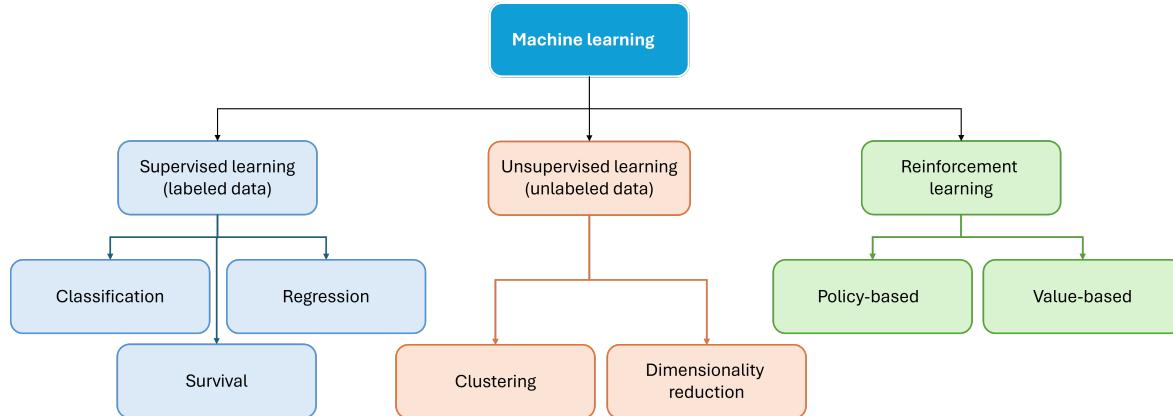
Dans ce chapitre, nous allons voir comment l'analyse de survie a profité de l'avènement du ML. Dans un premier temps, nous rappelons les concepts fondamentaux de l'apprentissage (Section II.1). Ensuite, nous examinons l'extension des algorithmes de ML à l'analyse de survie *classique* (Section II.2) ainsi qu'aux événements récurrents (Section II.3). Enfin, nous discutons des avantages et inconvénients des principes présentés (Section II.4).

II.1 Concepts fondamentaux de l'apprentissage

II.1.1 Les différents types d'apprentissage

Nous proposons de considérer un algorithme comme une série de règles définies pour atteindre une fonction objective notée f . Un modèle de ML \hat{f} associe aux données d'entrée des prédictions [Mitchell, 1997]. La prédiction $\hat{y} = \hat{f}(X)$ est ce que le modèle prévoit être la valeur cible sur la base des caractéristiques X . Hastie et al. [2009] définit plusieurs types d'apprentissage (Figure II.2), couramment rencontrés en recherche médicale :

- L'**apprentissage supervisé** consiste à apprendre à partir d'un jeu de données étiquetées (ou *labellisées*), où chaque exemple d'entraînement est associé à une réponse correcte (*label*). Le *label* peut par exemple correspondre à la présence d'une tumeur, la durée de survie ou la réponse à un traitement. Ce type d'algorithme est particulièrement utile pour l'aide au diagnostic des cancers [Yaqoob et al., 2023]. Les algorithmes d'apprentissage supervisé les plus communs sont : les régressions logistiques ou linéaires, les arbres de décisions, les approches ensemblistes, les machines à vecteur support, et les réseaux de neurones.
- L'**apprentissage non supervisé** permet de découvrir des motifs (*patterns*) dans les données [Li, 2017]. Contrairement à l'apprentissage supervisé, les données utilisées n'ont pas de labels. Cette approche est souvent utilisée dans le cadre d'analyses plus exploratoires. Par exemple, Xu et al. [2023] ont récemment permis de répartir les patients avec un cancer avancé dans des groupes distincts de gravité des symptômes, et les patients présentant une gravité plus élevée avant le traitement étaient plus susceptibles d'être hospitalisés et de décéder. Pour des problèmes de profilage (*clustering*), les algorithmes principaux sont *k-means*, DBSCAN de Learning [2006] et l'analyse hiérarchique des grappes de Bridges Jr [1966]. Pour la réduction de dimension, les algorithmes les plus connus sont l'analyse en composantes principales de Wold et al.



[1987], le t-SNE de Van der Maaten and Hinton [2008] et UMAP de McInnes et al. [2018].

- L'**apprentissage semi-supervisé** utilise les données étiquetées pour former initialement le modèle, puis exploite la structure des données non étiquetées pour affiner le processus d'apprentissage [Learning, 2006]. La plupart des algorithmes d'apprentissage semi-supervisé sont des combinaisons d'algorithmes non supervisés et supervisés.
- L'**apprentissage par renforcement** (*reinforcement learning*) est utilisé pour la prise de décision séquentielle lorsqu'une stratégie doit être apprise à partir de données issues d'interactions entre un agent et son environnement. Les applications comprennent principalement la détermination des schémas thérapeutiques optimaux avec plusieurs lignes de traitement pour les patients atteints d'un cancer [Padmanabhan et al., 2017, Yu et al., 2021].

Les sorties des algorithmes d'apprentissage sont appelées prédictions. Ce terme de "prédiction" peut parfois prêter à confusion [Isariyawongse and Kattan, 2012]. Il peut signifier l'estimation de la probabilité d'événements futurs, tels que la récidive du cancer, ou bien l'évaluation rétrospective d'un événement, tel que déterminer la malignité d'une tumeur déjà diagnostiquée. Dans les deux cas, l'algorithme fournit une estimation visant à éclairer les décisions prises lors d'interventions cliniques. Pour les modèles de survie, la prédiction peut être le temps de survenue d'un événement, un risque de survie ou encore un risque cumulé de survie [Suresh et al., 2022].

II.1.2 Sur- et sous-apprentissage

Le sur-apprentissage (*overfitting*) et le sous-apprentissage (*underfitting*) sont deux concepts clés en apprentissage automatique qui décrivent comment un modèle peut échouer à généraliser correctement à partir des données d'entraînement [Bishop and Nasrabadi, 2006]. L'objectif posé par Hastie et al. [2009] est de trouver le meilleur compromis entre une bonne description des données et une capacité du modèle à généraliser (Figure II.3).

Le **sur-apprentissage** se produit lorsque le modèle s'ajuste trop aux données d'en-

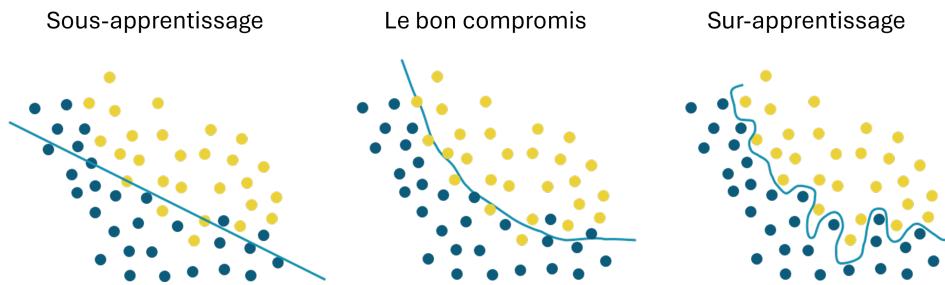


FIGURE II.3 – Les concepts de sur- et sous-apprentissage

traînement, capturant non seulement les tendances sous-jacentes, mais aussi le bruit et les fluctuations spécifiques de ces données. Un modèle ayant sur-appris présente une performance plus élevée sur les données d'entraînement que sur de nouvelles données, car il ne généralise pas bien. Les symptômes du sur-apprentissage incluent :

- Complexité excessive du modèle - Le modèle a trop de paramètres par rapport à la quantité de données, ce qui lui permet de "mémoriser" les données d'entraînement;
- Baisse de la performance sur les données de test - Un écart significatif entre les performances sur les données d'entraînement et les données de test.

Un modèle de prédiction de cancer utilisant un grand nombre de caractéristiques pour prédire la présence de cancer pourrait sur-apprendre si certaines de ces caractéristiques n'étaient que représentatives des patients de l'ensemble d'entraînement.

Le **sous-apprentissage** se produit lorsque le modèle est trop simple pour capturer les tendances sous-jacentes des données. Un modèle en sous-apprentissage a une mauvaise performance aussi bien sur les données d'entraînement que sur les données de test, car il ne peut pas saisir la complexité des données.

Un modèle de régression linéaire utilisé pour prédire les résultats de survie après un traitement contre le cancer pourrait sous-apprendre si les relations entre les caractéristiques (comme l'âge, le type et le stade du cancer, etc.) et la variable à prédire (ou à expliquer) étaient non linéaires et complexes. Un modèle aussi simple manquerait des interactions importantes entre les caractéristiques des patients.

Le **compromis "biais-variance"** est lié aux concepts de sur- et de sous-apprentissage [Hastie et al., 2009]. Le biais fait référence à l'erreur introduite par l'approximation d'un problème réel (qui peut être complexe) par un modèle simplifié. Un biais important peut conduire le modèle à ne pas tenir compte des relations pertinentes entre les variables explicatives et la variable à expliquer (cas de sous-apprentissage). La variance est l'erreur introduite par la sensibilité du modèle aux petites fluctuations de l'ensemble d'apprentissage. Une variance élevée peut inciter le modèle à modéliser le bruit aléatoire des données d'apprentissage plutôt que les résultats escomptés (cas de sur-apprentissage). Plus la complexité du modèle augmente, plus le biais diminue et plus la variance augmente. L'objectif est de trouver un modèle qui soit juste assez complexe pour capturer la structure sous-jacente des données sans sur-apprendre.

II.1.3 Évaluation des modèles

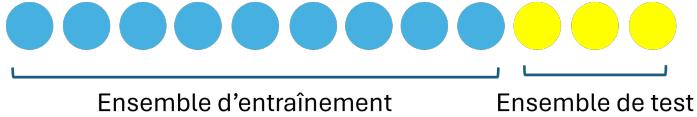
Le choix du **bon modèle de prédiction** est un processus qui nécessite de conjuguer l'analyse des données, la connaissance du contexte médical et la prise en compte réfléchie des différents paramètres. Selon [Box \[1976\]](#), "*all models are wrong, but some are useful*", alors il importe de définir précisément le terme "*useful*". En particulier en recherche clinique, cette utilité est directement liée à la capacité à déterminer si le modèle fournit des prédictions précises et adaptées à un contexte donné [Qi et al. \[2023\]](#). Evaluer la performance revient à évaluer la capacité du modèle à généraliser sur de nouvelles données. Pour ce faire, plusieurs techniques dites de validation existent, pour lesquelles différentes métriques de performance peuvent être calculées.

La méthode de référence consiste à diviser les données disponibles en deux ensembles distincts : l'ensemble d'entraînement (*train set*) et l'ensemble de test (*test set*) (Figure II.4(a)). Le modèle est entraîné à l'aide de l'ensemble d'entraînement, puis testé sur l'ensemble de test. En évaluant le modèle sur l'ensemble de test, on obtient une mesure de performance. En particulier, l'erreur est calculée en comptant le nombre de fois où le modèle prédit une issue différente de la réalité représentée par l'échantillon de test. Cela permet de mesurer la performance du modèle sur des instances qui n'ont jamais été vues auparavant. Une pratique courante est d'utiliser entre 70% et 80% des données pour l'entraînement et de réservé 20% à 30% pour le test [\[Hastie et al., 2009\]](#). Cependant, cette proportion peut varier en fonction de la taille du jeu de données.

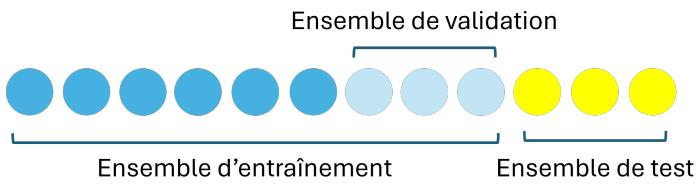
II.1.3.1 Optimisation des hyperparamètres et sélection de modèles

La majorité des modèles ont des paramètres qui leur sont propres et qui peuvent impacter la performance, ce sont les hyperparamètres. Une méthode consiste à entraîner plusieurs modèles avec différentes valeurs de ces paramètres et à sélectionner celui qui minimise l'erreur de généralisation. Cependant, en procédant ainsi, le modèle sélectionné risque d'être seulement adapté à l'ensemble de test, induisant potentiellement une baisse de performance sur de nouvelles données [\[Arlot and Celisse, 2010\]](#).

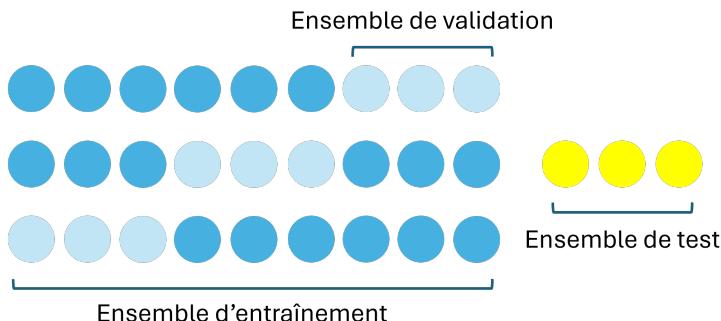
Validation par réserve Une façon de pallier ce risque est de réservé une partie de l'ensemble d'entraînement pour évaluer plusieurs modèles candidats suivant une sélection d'hyperparamètres et sélectionner le meilleur. Cet ensemble réservé est appelé ensemble de validation (*validation set*). Plus précisément, plusieurs modèles sont entraînés avec divers hyperparamètres sur un ensemble d'entraînement réduit (c'est-à-dire l'ensemble d'entraînement complet moins l'ensemble de validation), et le modèle offrant les meilleures performances sur l'ensemble de validation est sélectionné. Ensuite, ce modèle est réentraîné sur l'ensemble d'entraînement complet (y compris l'ensemble de validation), et ses performances finales sont évaluées sur l'ensemble de test pour estimer l'erreur de généralisation (Figure II.4(b)).



(a) Ensembles d'entraînement et de test



(b) Validation par réserve pour l'optimisation des hyperparamètres sur les ensembles d'entraînement et de validation (chaque ligne correspond à un tour de la validation croisée)



(c) Validation croisée pour l'optimisation des hyperparamètres sur les ensembles d'entraînement et de validation

FIGURE II.4 – Ensembles d'entraînement, de validation et de test

Validation croisée Si l'ensemble de validation est trop petit, les évaluations seront moins précises, risquant de sélectionner un modèle sous-optimal. À l'inverse, si l'ensemble de validation est trop grand, l'ensemble d'entraînement restant sera significativement réduit, ce qui peut fausser la comparaison des modèles. Une solution à ce problème est la validation croisée, qui utilise de nombreux petits ensembles de validation. Chaque modèle est évalué une fois par ensemble de validation, après avoir été entraîné sur le reste des données (Figure II.4(c)). En agrégant toutes les évaluations, on obtient une mesure plus précise de la performance du modèle. La validation croisée peut également fonctionner en considérant l'ensemble initial des données.

II.1.3.2 Le théorème *No free lunch*

Un modèle est une version simplifiée des données, conçue pour écarter les détails superflus qui ne sont pas susceptibles de se généraliser à de nouvelles instances [Géron, 2022]. Wolpert and Macready [1997] ont démontré que sans faire aucune hypothèse sur les données, il n'y a aucune raison de préférer un modèle à un autre. C'est ce que l'on appelle le théorème *No Free Lunch*. Pour certains jeux de données, le meilleur modèle est un modèle linéaire, tandis que pour d'autres, c'est un réseau de neurones. Aucun modèle n'est a priori garanti de mieux fonctionner.

Implications de modélisation Pour déterminer quel modèle est le plus adapté, il est nécessaire de les évaluer tous. Étant donné que cela est pratiquement impossible, il est courant de faire des hypothèses raisonnables sur les données et de n'évaluer qu'un ensemble de modèles plausibles. Par exemple, pour des tâches simples, il peut être judicieux d'évaluer des modèles linéaires avec différents niveaux de régularisation, tandis que pour des problèmes complexes, il peut être plus approprié d'explorer d'autres modèles. En oncologie, le théorème *No Free Lunch* a des implications directes. Par exemple, pour prédire la survie des patients atteints de cancer, un modèle linéaire peut suffire si les relations entre les variables cliniques et la survie sont simples. Cependant, pour des tâches plus complexes, comme la prédiction des réponses aux traitements basées sur des données génomiques, des régressions pénalisées, des forêts aléatoires ou encore des réseaux de neurones profonds peuvent être plus appropriés. Le choix du modèle dépendra donc des caractéristiques spécifiques des données disponibles et des hypothèses que les chercheurs sont capables de formuler, et notamment pour la recherche avancée en cancer [Kalantari et al., 2020].

II.2 Apprentissage pour les données censurées

Le modèle de Cox à hasards proportionnels (*Cox proportional hazard*, CPH) vu au Chapitre I est largement utilisé en recherche clinique pour sa facilité de mise en œuvre, sa rapidité de calcul et, surtout, ses résultats exploitables [Wang et al., 2019]. Néanmoins, ce modèle repose sur des hypothèses restrictives, telles que la proportionnalité des risques, et l'absence de corrélation entre les variables explicatives. Enfin, cette méthode ne permet pas de modéliser correctement des effets non linéaires et/ou d'interaction [Moncada-Torres et al., 2021]. Dans cette section, nous présenterons les approches alliant survie et ML puis nous introduirons les différentes métriques utilisées pour les évaluer.

Tizi and Berrado [2023] ont exploré les différents modèles de ML pour la survie en oncologie. Les modèles les plus fréquents sont les arbres de décision et les forêts aléatoires de Ishwaran et al. [2008], les réseaux de neurones tels que Bhambhani et al. [2021], et les machines à vecteurs de support de Van Belle et al. [2011b] ou Pölsterl et al. [2015]. Par ailleurs, une autre revue de Huang et al. [2023] a montré l'intérêt des algorithmes de ML en survie pour les études en vie réelle toutes pathologies confondues. Nous proposons ici un aperçu synthétique des principales approches statistiques et d'apprentissage automatique utilisées dans l'analyse de la survie.

TABLE II.1 – Régressions pénalisées du modèle de Cox

Modèle de Cox pénalisé	Pénalisation
LASSO-Cox	$\lambda \sum_{j=1}^p \beta_j $
Ridge-Cox	$\frac{\lambda}{2} \sum_{j=1}^p \beta_j^2$
EN-Cox	$\lambda \left(\alpha \sum_{j=1}^p \beta_j + \frac{1}{2}(1 - \alpha) \sum_{j=1}^p \beta_j^2 \right)$

LASSO = *Least Absolute Shrinkage and Selection Operator*; EN = Elastic-Net.

II.2.1 Régressions pénalisées

L'intégration de l'ensemble des variables explicatives au modèle de prédiction peut donner lieu à des phénomènes de sur-apprentissage [Van Houwelingen and Putter, 2011]. Des techniques de pénalisation sont utilisées pour restreindre l'espace des estimations des coefficients, pour sélectionner des variables et/ou tenir compte de leur multicolinéarité [Friedman, 2009] : LASSO (*Least Absolute Shrinkage and Selection Operator*), Ridge ou encore Elastic-net. La famille des fonctions de pénalité normales est $L_\gamma : \mathbb{R}^p \rightarrow \mathbb{R}$, avec $L_\gamma(\beta) = \|\beta\|_\gamma = (\sum_{j=1}^p \|\beta_j\|^\gamma)^{\frac{1}{\gamma}}$ pour $\gamma > 0$. Plus la valeur de γ est petite, plus la solution est parcimonieuse, mais lorsque $0 \leq \gamma < 1$, la pénalité n'est pas convexe, ce qui rend le problème d'optimisation plus difficile à résoudre. Nous présentons ici les modèles de Cox régularisés couramment utilisés, dont les régularisateurs sont résumés dans le Tableau II.1.

Selon le principe introduit dans Tibshirani [1997], la pénalité de la norme L_1 a été incorporée dans la log-vraisemblance partielle de l'équation I.14 pour obtenir l'algorithme LASSO-Cox, permettant ainsi la sélection des variables et l'estimation simultanée des coefficients de régression. Plusieurs extensions de la méthode LASSO-Cox ont également été développées. Le LASSO-Cox adaptatif proposé par Zhang and Lu [2007] repose sur une vraisemblance partielle pénalisée avec des pénalités L_1 pondérées de manière adaptative $\lambda \sum_{j=1}^p \tau_j |\beta_j|$ sur les coefficients de régression, où les poids τ_j sont faibles pour les grands coefficients et importants pour les petits coefficients. Dans le LASSO-Cox fusionné de Tibshirani et al. [2005], les coefficients et leurs différences successives sont pénalisés à l'aide de la norme L_1 . Enfin, dans le LASSO-Cox graphique de Friedman et al. [2008], les graphes parcimonieux sont estimés en utilisant la méthode de descente des coordonnées, en appliquant une pénalité L_1 à la matrice de covariance inverse.

La régression Ridge a été proposée à l'origine par Hoerl and Kennard [1970] et a été utilisée avec succès dans le contexte de la régression Cox par Verweij and Van Houwelingen [1994]. Elle incorpore un régularisateur de norme L_2 pour sélectionner les variables corrélées et réduire leurs valeurs l'une par rapport à l'autre. La méthode de Cox régularisée basée sur les variables de Vinzamuri and Reddy [2013] utilise un régularisateur à valeur non négative $R(\beta) = |\beta|^T M |\beta|$ pour la formulation des moindres carrés modifiée de la régression de Cox, avec $M \in \mathbb{R}^{p \times p}$ est une matrice semi-définie positive. Ridge-Cox est un cas particulier lorsque M est la matrice d'identité.

L'approche Elastic-Net (EN) combine les pénalités L_1 et L_2 , permettant d'effectuer la sélection des variables et de traiter simultanément la corrélation entre les variables [Zou and Hastie, 2005]. La méthode EN-Cox a été proposée par Simon et al. [2011] où le terme

de pénalité EN indiqué dans le Tableau II.1 avec $0 \leq \alpha < 1$ et introduit dans la fonction de log-vraisemblance partielle dans l'équation I.14. Contrairement à LASSO-Cox, EN-Cox peut sélectionner plus de n variables si $n \leq p$.

II.2.2 Méthodes d'apprentissage automatique

Les principes d'apprentissage énoncés en Section II.1 s'appliquent initialement pour répondre à des problèmes de classification ou de régression. Nous présentons dans cette partie comment ces concepts s'appliquent également aux données de survie.

II.2.2.1 Machines à vecteurs de support

Les machines à vecteurs de support (*support vector machines*, SVM) sont une méthode d'apprentissage supervisé performante [Wang et al., 2019]. Une machine à vecteur de support construit un hyperplan ou un ensemble d'hyperplans dans un espace de dimension élevée ou infinie, qui peut être utilisé pour la classification ou la régression [Cortes and Vapnik, 1995]. De façon intuitive, on obtient une bonne séparation en utilisant l'hyperplan qui présente la plus grande distance, appelée marge fonctionnelle, par rapport aux données d'apprentissage les plus proches d'une ou l'autre classe, qui sont les vecteurs de supports. Plus cette marge est élevée, plus l'erreur de généralisation est faible. Les SVM peuvent prendre en compte les données censurées de deux manières différentes :

- En considérant un problème de régression, le modèle apprend à prédire le temps de survie ;
- En considérant un problème de *ranking*, le modèle attribue un rang inférieur aux individus avec des temps de survie plus courts.

Approche de régression à vecteur de support Pour la régression, les SVM consistent à ne considérer que les instances présentant des événements dans la régression à vecteur de support (*support vector regression*). Dans ce cadre, la fonction de perte $f(X_i) = \max(0, |f(X_i) - y_i| - \epsilon)$ est minimisée avec l'ajout d'un régularisateur [Smola and Schölkopf, 1998]. Cependant, cette méthode présente un inconvénient majeur : elle ignore les informations d'ordre incluses dans les instances censurées [Shivaswamy et al., 2007].

Approche de classification par vecteur de support Une alternative est d'utiliser la classification par vecteur de support (*support vector classification*) en appliquant l'approche de classification par contrainte de Har-Peled et al. [2002]. Cette méthode impose alors des contraintes dans la formulation du SVM pour deux instances comparables afin de maintenir l'ordre requis.

Machines à vecteur de support pour données censurées Pour répondre aux limitations des méthodes précédentes, Khan and Zubek [2008] ont proposé la régression à vecteur

de support pour les données censurées (*support vector regression for censored data*). Cette approche tire parti de la régression à vecteur de support standard tout en l'adaptant aux cas censurés grâce à une fonction de perte asymétrique mise à jour. Ainsi, elle prend en compte à la fois les instances non censurées et censurées dans le modèle. Van Belle et al. [2011a] ont introduit une approche basée sur le *support vector regression* qui combine les méthodes de classement et de régression dans le contexte de l'analyse de survie. Les machines à vecteur de support (*Survival Support Vector Machine*, SSVM) pour l'analyse de survie offrent une solution complète en intégrant les deux types d'informations pour améliorer les prédictions. Schématiquement, cette méthode consiste à séparer les individus par un hyperplan, et le score de risque est polarisé pour les individus loin de la frontière.

II.2.2.2 Arbres de survie

La structure d'un arbre de survie, comme celle des arbres de décision de type CART (*Classification and Regression Trees*) est composée de nœuds de décision qui partitionnent les données en groupes homogènes selon la variable la plus significative à chaque étape [Breiman et al., 1984]. Chaque nœud représente une variable explicative et un seuil de décision, tandis que les nœuds terminaux fournissent des estimations de survie pour les groupes ainsi définis.

Les arbres de survie présentent plusieurs avantages. Tout d'abord, ils sont interprétables [Molnar, 2020], c'est-à-dire que leur processus de décision est directement intelligible pour l'humain Miller [2019]. La construction des arbres permet en effet une visualisation claire des impacts des différentes variables [Breiman, 2017]. De plus, ils sont capables de capturer les relations non linéaires entre les variables explicatives et le temps de survie, offrant ainsi une modélisation pouvant être plus adéquate dans le cadre de données complexes [Ishwaran et al., 2010]. Toutefois, un arbre est interprétable tant qu'il n'y a pas trop de nœuds et que l'arbre n'est pas trop profond. De plus, ces arbres trop complexes peuvent de même induire du sur-apprentissage, ce qui peut réduire la capacité de généralisation du modèle [Ishwaran et al., 2008]. Le Chapitre III donne plus de détails à ce propos.

II.2.2.3 Méthodes d'ensemble

Pour surmonter l'instabilité des arbres de décision seuls, les techniques ensemblistes telles que le *bagging* (issu de la combinaison de *bootstrap* et *aggregating*) introduit par Breiman [1996] et les forêts aléatoires de Breiman [2001b] sont couramment utilisées pour construire des modèles plus performants.

Arbres de survie en bagging Dans le cadre des arbres de survie en *bagging* (*bagging survival trees*), la fonction de survie agrégée est obtenue en moyennant les prédictions des arbres de survie individuels [Hothorn et al., 2004]. Cette approche comprend trois étapes principales :

1. Tirer B échantillons bootstrap à partir des données disponibles ;

2. Construire un arbre de survie pour chaque échantillon bootstrap, en s'assurant que chaque nœud terminal contient un nombre d'événements supérieur ou égal à un certain seuil;
3. Calculer la fonction de survie agrégée en faisant la moyenne des prédictions des nœuds feuilles.

Pour chaque nœud terminal, la fonction de survie est estimée à l'aide de l'estimateur de Kaplan-Meier de la Section I.2.1 du Chapitre I, en supposant que tous les individus dans un même nœud partagent la même fonction de survie.

Forêts aléatoires de survie Les forêts aléatoires, introduites par [Breiman \[2001b\]](#) suivent une approche similaire aux *bagging survival trees*. La principale différence est le fait qu'à chaque nœud, seul un sous-ensemble aléatoire de variables est utilisé pour le critère de division, ce qui réduit la corrélation entre les arbres et améliore les performances prédictives. [Ishwaran et al. \[2008\]](#) ont étendu cette méthode aux forêts aléatoires de survie (*Random Survival Forest*, RSF), qui utilisent un ensemble d'arbres de survie pour la prédiction. Le Chapitre III donne plus de détails à propos des forêts aléatoires.

Super Learners Cet algorithme combine plusieurs modèles en estimant la performance de la prédiction de chacun au moyen d'une validation croisée en k folds [[Van der Laan et al., 2007](#)]. La recherche de la combinaison optimale des *weak learners* par la minimisation du risque permet de contrôler le sur-apprentissage du modèle ensembliste final. Les algorithmes candidats peuvent aller d'un modèle de [Cox \[1972a\]](#) à des algorithmes d'apprentissage automatique basés sur des arbres, comme les forêts aléatoires de survie décrites ci-dessus.

II.2.2.4 Boosting

Le boosting est une méthode ensembliste combinant plusieurs modèles faibles (*weak learner*) de manière itérative pour corriger les erreurs des prédictions précédentes et ainsi former un unique modèle global [[Ridgeway, 1999](#)]. Les prédictions sont ainsi agrégées à chaque ajout de modèle améliorant (*boosting*) le modèle global. Par conséquent, le modèle global f est un modèle additif de la forme

$$f(\mathbf{x}) = \sum_{m=1}^M \beta_m g(\mathbf{x}, \theta_m) \quad (\text{II.1})$$

avec $M > 0$ le nombre de *weak learners*, $\beta_m \in \mathbb{R}$ le poids associé à la fonction g du *weak learner* paramétrée par θ , et $\sum_{m=1}^M \beta_m = 1$. Chaque *weak learner* peut différer d'un autre suivant ses paramètres, indiqué par l'indice θ_m .

CoxBoost Proposée par [Binder and Schumacher \[2008\]](#), le modèle CoxBoost est une méthode d'apprentissage statistique qui combine les modèles de Cox pour l'analyse de survie à la technique de *boosting*. CoxBoost ajuste le modèle de Cox par étapes successives, en se

basant sur les résidus des prédictions précédentes pour affiner les estimations des coefficients des variables explicatives. Une régularisation est effectuée afin de prendre en compte une éventuelle multicolinéarité et d'éviter le sur-apprentissage. En prenant en compte un ensemble flexible de variables candidates pour la mise à jour à chaque étape du boosting, l'estimation des coefficients du modèle est effectuée par méthode du gradient basée sur le décalage (*offset*). [Binder and Schumacher \[2008\]](#) introduisent également la possibilité d'inclure dans le modèle final certaines covariables obligatoires qui devraient être explicitement prises en compte.

Boosting des arbres de survie Le *boosting* peut également prendre en compte comme *weak learner* un arbre de survie, tel que proposé par [Hothorn et al. \[2006\]](#). Ce modèle est semblable aux RSF, dans la mesure où il repose sur plusieurs *weak learners* pour produire une prédition globale. Cependant, la méthode d'agrégation diffère. Les RSF ajustent un ensemble d'arbres de survie de façon indépendante, puis calculent la moyenne de leurs prédictions, tandis que le modèle à gradient boosté est construit de manière séquentielle.

II.2.2.5 Apprentissage profond

Nous n'aborderons pas dans cette thèse les méthodes d'apprentissage profond (*deep learning*) qui trouvent également leurs applications en analyse de survie. Comme pour l'apprentissage automatique traditionnel, les approches de *deep learning* s'appuient généralement sur des méthodes statistiques classiques de survie tout en exploitant les avantages des réseaux de neurones. En particulier, la revue de littérature proposée par [Wiegreb et al. \[2024\]](#) a souligné leur intérêt pour le traitement des données non structurées ou de haute dimension telles que des images, du texte ou des données omiques.

II.2.3 Métriques de performance

La performance prédictive d'un modèle de survie est généralement basée sur la distance ou l'accord entre les résultats observés et prédits [[Uno et al., 2011](#), [Bylinskii et al., 2018](#)].

II.2.3.1 Indice de concordance

En analyse de survie, la méthode la plus courante pour évaluer un modèle consiste à considérer le risque relatif d'un événement pour différentes instances plutôt que les temps de survie absous pour chaque instance [[Longato et al., 2020](#), [Zhou et al., 2023](#)]. Cela peut se faire en calculant la probabilité de concordance ou l'indice de concordance (C-index) :

$$\mathbb{C} = \mathbb{P}(\eta_i > \eta_j | T_i < T_j), \quad (\text{II.2})$$

avec η_i le score de risque pour l'individu i . En général, $\eta_i := \eta_i(Z_i) = \exp(\beta^T Z_i)$ pour le CPH. Différentes versions du C-index existent, le pionnier ayant été [Harrell et al. \[1982\]](#),

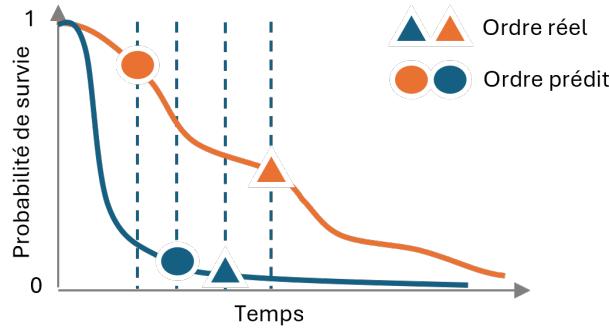


FIGURE II.5 – Ordre d'événements réel et prédit pour deux individus

puis étendues par [Pencina and D'agostino \[2004\]](#), [Uno et al. \[2011\]](#) et [Gerds et al. \[2013\]](#). L'estimateur du C-index de [Harrell et al. \[1982\]](#) s'écrit

$$\hat{C}_H = \frac{\sum_{i \neq j} \mathbb{1}(\hat{\eta}_i > \hat{\eta}_j) \mathbb{1}(T_i < T_j, \delta_i = 1)}{\sum_{i \neq j} \mathbb{1}(T_i < T_j, \delta_i = 1)}, \quad (\text{II.3})$$

avec $\hat{\eta}_i$ le score de risque estimé pour le sujet i , T_i et T_j les temps de survie observés pour les sujets i et j , δ_i un indicateur de censure pour le sujet i (1 si l'événement est observé, 0 si censuré). Ainsi, le numérateur compte le nombre de paires concordantes selon le modèle, c'est-à-dire les paires pour lesquelles le modèle prédit correctement l'ordre des événements (Figure II.5). Le dénominateur normalise en divisant par le nombre total de paires comparables.

Pour évaluer la performance durant une période de suivi spécifique, [Heagerty and Zheng \[2005\]](#) ont défini un C-index pour une période de suivi fixe sur $[0, t^*]$ comme la moyenne pondérée des valeurs de l'AUC (*Area Under the Curve*) à tous les points de temps d'observation possibles. L'AUC dépendant du temps pour un temps de survie t peut être calculée comme suit :

$$\text{AUC}(t) = \mathbb{P}(\eta_i < \eta_j \mid T_i < t, T_j > t). \quad (\text{II.4})$$

L'AUC estimé est

$$\widehat{\text{AUC}}(t) = \frac{1}{\text{num}(t)} \sum_{i:T_i < t} \sum_{j:T_j > t} \mathbb{1}(\hat{\eta}_i < \hat{\eta}_j), \quad (\text{II.5})$$

avec $\text{num}(t)$ le nombre de paires comparables pour le point de temps t , c'est-à-dire les paires (i, j) telles que $T_i < t$ et $T_j > t$. Le C-index sur la période $[0, t^*]$ est la moyenne pondérée de l'AUC dépendante du temps obtenue par l'équation II.5 et est calculé comme suit :

$$\hat{C}_{t^*} = \frac{1}{\text{num}(t^*)} \sum_{i:\delta_i=1} \sum_{j:T_i < T_j} \mathbb{1}(\hat{\eta}_i < \hat{\eta}_j) = \sum_t \widehat{\text{AUC}}(t) \cdot \text{num}(t). \quad (\text{II.6})$$

Ainsi, \hat{C}_{t^*} représente la probabilité que les prédictions soient concordantes avec les résultats observés pour un jeu de données durant la période $[0, t^*]$.

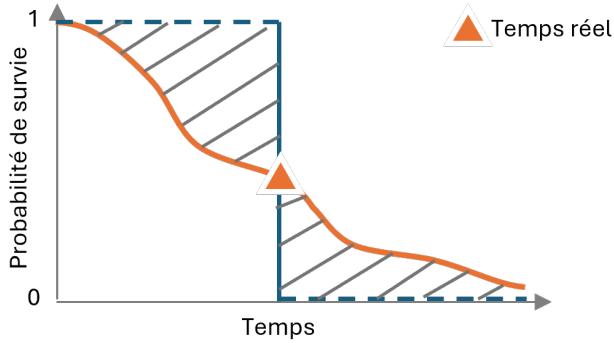


FIGURE II.6 – Pondération par la distribution de la censure dans le Brier score

II.2.3.2 Brier score intégré

Le score de [Brier \[1950\]](#), initialement développé pour prédire l'inexactitude des prévisions météorologiques, permet d'évaluer les modèles de prédiction qui ont des résultats probabilistes. Cela signifie que le résultat doit rester dans la plage $[0, 1]$, et la somme de tous les résultats possibles pour un individu doit être égale à 1. En considérant des prédictions de résultats binaires, le Brier score peut être considéré comme l'erreur quadratique moyenne et s'écrit :

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N [\hat{y}_i - y_i]^2, \quad (\text{II.7})$$

avec $\hat{y}_i \in [0, 1]$ la probabilité prédictée par le modèle d'avoir un événement pour l'individu i et $y_i \in \{0, 1\}$ le résultat réel. Un modèle avec un Brier score plus faible donnera de meilleures prédictions et sera considéré comme bien calibré puisque le Brier score est proche de 0.

Pour l'analyse de survie, le Brier score a été étendu par [Graf et al. \[1999\]](#) pour tenir compte de la censure. Les contributions individuelles sont répondées selon la distribution de la censure. Le Brier score au temps t de l'équation II.7 s'écrit

$$\text{BS}(t) = \frac{1}{\sum_{i=1}^N Y_i(t)} \sum_{i=1}^N \hat{w}_i(t) \left[\hat{S}_i(t) - \mathbb{1}(T_i > t) \right]^2, \quad (\text{II.8})$$

avec $\hat{w}_i(t)$ le poids pour l'individu i estimé à partir de l'estimateur de Kaplan-Meier de la distribution de censure G . On a

$$\hat{w}_i(t) = \begin{cases} \frac{\delta_i}{\hat{G}(t)} & \text{si } T_i \leq t; \\ \frac{1}{\hat{G}(t)} & \text{si } T_i > t. \end{cases} \quad (\text{II.9})$$

Avec cette distribution de poids, les poids pour les instances censurées avant t sont égaux à 0. Cependant, elles contribuent indirectement au calcul du Brier score puisqu'elles sont utilisées pour calculer G . Les poids pour les instances non censurées à t sont supérieurs à 1, de sorte qu'elles contribuent par leur probabilité de survie estimée au calcul du Brier score (Figure II.6).

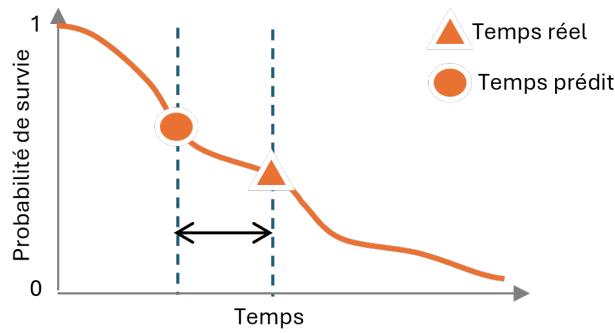


FIGURE II.7 – Erreur moyenne absolue

Le Brier score intégré (*integrated Brier score*, IBS) évalue la performance globale d'un modèle sur une plage de temps. Il est défini comme l'intégrale du Brier score sur la période $[\tau_1, \tau_2]$:

$$\text{IBS} = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \text{BS}(t) dt. \quad (\text{II.10})$$

En général $\tau_1 = 0$ et τ_2 est le temps de suivi maximal observé.

II.2.3.3 Erreur absolue moyenne

Pour les problèmes d'analyse de survie, l'erreur moyenne absolue (*mean absolute error*, MAE) peut être définie comme une moyenne des différences entre les valeurs de temps prédites et les valeurs de temps d'observation réelles (Figure II.7). On écrit :

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N \delta_i |T_i - \hat{T}_i|. \quad (\text{II.11})$$

Il convient de noter que seuls les échantillons pour lesquels l'événement survient sont pris en compte dans cette métrique puisque si $\delta_i = 0$, le terme correspondant deviendra nul.

Une approche naïve consiste à exclure tous les sujets censurés de l'évaluation, puis à utiliser l'équation II.11 pour calculer l'erreur absolue pour chaque patient non censuré et à prendre la moyenne sur les instances non censurées. La distribution (marginale) des sujets censurés et des sujets d'événements peut varier considérablement, ce qui rend cette stratégie susceptible d'être biaisée. En outre, lorsque le taux de censure est élevé, une partie importante des données sera complètement ignorée par la mesure de performance.

[Haider et al. \[2020\]](#) ont proposé d'utiliser la perte de la charnière (*hinge loss*) pour calculer une mesure unilatérale qui ne prend en compte que si le temps prédit est antérieur au temps censuré. Plus récemment, cette mesure utilise la méthode de pondération par la probabilité inverse de censure pour traiter les sujets censurés dans le calcul de l'erreur de prédiction. MAE-margin attribue une valeur "best guess" (temps de marge) à chaque sujet censuré en utilisant l'estimateur non paramétrique de Kaplan-Meier [[Haider et al., 2020](#)]. Ce temps de marge peut être interprété comme une espérance conditionnelle du temps de l'événement étant donné que le temps de l'événement est supérieur au temps de censure.

Plus récemment, [Qi et al., 2023] ont proposé de pondérer par des poids de la censure (*inverse probability censoring weight*) dans l'esprit du Brier score de l'équation II.9, ou encore d'inclure les pseudo-observations de Andersen and Pohar Perme [2010].

II.3 Apprentissage avec événements récurrents

L'article ci-après, intitulé "*Towards filling the gaps around recurrent events in high dimensional framework : a systematic literature review and application*" et publié au journal *Biostatistics and Epidemiology* en janvier 2023, présente un état de l'art des méthodes d'apprentissage existantes pour traiter les événements récurrents. En plus d'une revue systématique de la littérature, nous avons proposé d'appliquer les modèles identifiés sur des données simulées afin de mieux comprendre leurs mécanismes. Les messages clés de ce travail sont :

- En l'absence de recommandations claires, nous avons mis en évidence la prudence des auteurs/investigateurs lorsqu'ils traitent des événements récurrents en grande dimension. Pour cause, très souvent (71/80) la notion de récurrence en survie est évitée, et l'analyse repose soit sur un problème de classification avec récurrence = oui/non, soit sur un problème de survie classique avec le temps jusqu'à la première récurrence ou le temps sans récurrence (*recurrence-free survival*).
- Pour la première fois dans le contexte des événements récurrents, des approches statistiques avec et sans apprentissage ont été confrontées, avec les méthodes "standards", des algorithmes de sélection de variables et un réseau de neurones, sur des données simulées.
- Au moment des analyses, il n'existant pas de critère unique pour évaluer la performance des méthodes qui traitent les événements récurrents et aucune recommandation en regard de la métrique.
- L'analyse des événements récurrents par apprentissage est encore à explorer avec peu d'exemples concrets dans la littérature.



Towards filling the gaps around recurrent events in high dimensional framework: a systematic literature review and application*

Juliette Murris ^{a,b,c}, Anais Charles-Nelson ^{d,e}, Abir Tadmouri Sellier^c,
Audrey Lavenu ^{f,g,h} and Sandrine Katsahian ^{a,b,d,e,i},

^aInserm, Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université, Paris, France; ^bHeKA, Inria, Paris, France; ^cRWE & Data, Pierre Fabre, Boulogne-Billancourt, France; ^dUnité de Recherche Clinique, Hôpital Européen Georges-Pompidou, Assistance Publique – Hôpitaux de Paris (AP-HP), APHP, Centre, Paris, France; ^eCentre d'Investigation Clinique 1418 (CIC1418) Épidémiologie Clinique, Paris, France; ^fFaculté de Médecine, Université de Rennes 1, Rennes, France; ^gInstitut de Recherche Mathématique de Rennes (IRMAR), Rennes, France; ^hCentre de Investigation Clinique (CIC) CIC 1414, Inserm, Université de Rennes 1, Rennes, France; ⁱService d'Informatique Médicale, Biostatistiques et Santé Publique, Hôpital Européen Georges Pompidou, Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, France

ABSTRACT

Individuals may experience repeated events over time. However, there is no consensus about learning approaches to use in a high-dimensional framework for survival data (when the number of variables exceeds the number of individuals, i.e. $p > n$). The aim of this study was to identify learning algorithms for analyzing/predicting recurrent events, and to compare them to standard statistical models on simulated data. A systematic literature review was conducted to provide state-of-the-art methodology. Data were then simulated according to the number of variables, the proportion of active variables, and the number of events. The performance of the models was assessed using Harrell's concordance index, Kim's C-index, and error rate for active variables. Seven publications were identified, of which four were methodological studies, one an application paper and two were reviews. On simulated data, the standard models failed when $p > n$. Penalized Andersen–Gill and frailty models outperformed, whereas RankDeepSurv gave poorer performances. With no current guidelines on a specific approach to use, this study deepens understanding of the mechanisms and limits of investigated methods in this context.

ARTICLE HISTORY

Received 21 July 2022
Accepted 22 February 2023

KEYWORDS

Recurrent events; survival analysis; high-dimensional data; machine learning; simulated data

1. Introduction

Individuals may experience repeated events over time, such as hospitalizations or cancer relapses. In either clinical trials or real-world settings, survival analysis usually focuses on

CONTACT Juliette Murris  juliette.murris@pierre-fabre.com  33 Av. Emile Zola, Boulogne-Billancourt 92100, France

† A.L. and S.K. contributed equally.

* Preprint available at  <https://arxiv.org/abs/2203.15694>.

© 2023 International Biometric Society – Chinese Region

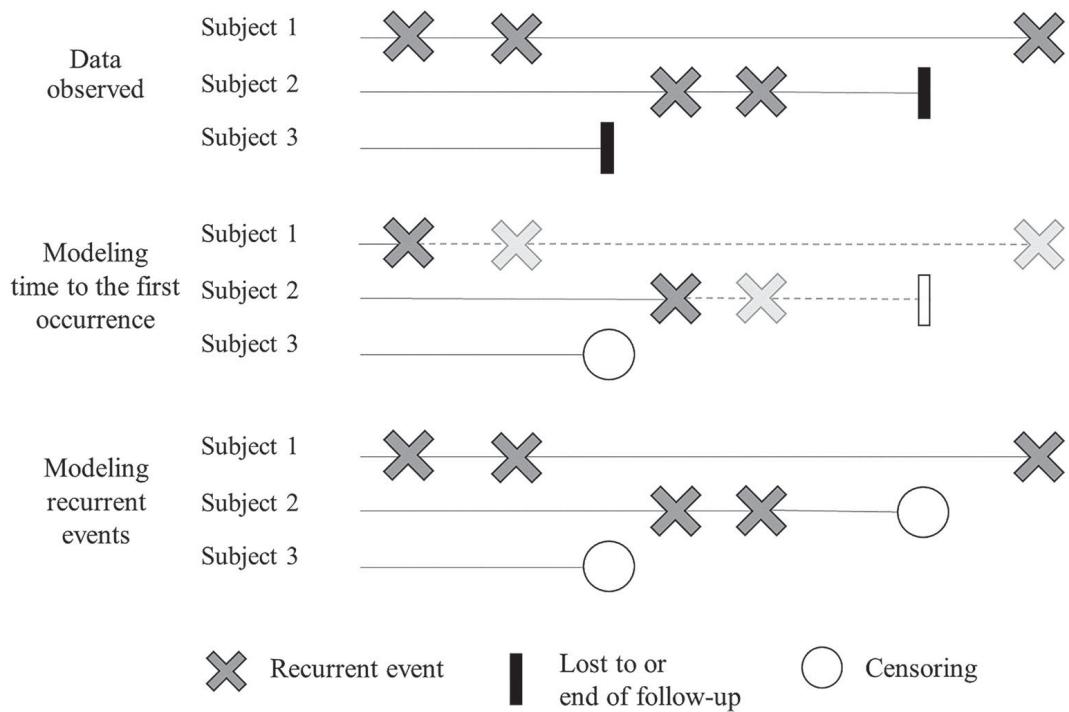


Figure 1. Recurrent event framework.

modeling the time to the first event. However, variables may have a different effect on the first event and on subsequent occurrences. Thus, modeling recurrent events remains a big challenge (Figure 1).

Indeed, two main problems arise when analyzing recurrent events. Firstly, interindividual heterogeneity emerges as some subjects may be more likely than others to experience the event. Secondly, events experienced by an individual are not independent, leading to intraindividual heterogeneity. Various methods have been developed to deal with these two issues and can be classified into marginal and conditional models. Marginal models involve implicitly averaging over the history of previous recurrent events. Conditional models can condition on event past history.

Furthermore, modern technologies enable data to be generated on thousands of variables or observations, as per genomics, medico-administrative databases, disease monitoring by intelligent medical devices, etc. While massive data describes large numbers of observations, high-dimensional data is defined as data with a number of variables of interest p greater than the number of individuals n . In this context, standard statistical models may no longer be applied, as they tend to face convergence problems and non-clinically relevant significance of the variables can arise. Machine learning methods have been developed to handle these problems.

Literature reviews were previously conducted on recurrent events, but none dealt with a high dimensional framework [1–3]. The aim of this article was to review innovative methodology available to analyze and predict high-dimensional recurrent events data. Simulations were performed to study the properties of identified methods compared to standard methods, according to the number of variables at the modeling stage.

Section 2 hereafter describes the methodological setting regarding both the literature review and the statistical approaches for modeling and evaluating the models as well as the data simulation scheme. Then, section 3 provides the findings of the review that enabled the identification of the adequate methods. Next, the application results on simulated data are reported. Section 4 finally relates the discussion and gives a contextual perspective based on related work and theoretical considerations.

2. Materials and method

2.1. Systematic literature review

A systematic literature review (SLR) was performed to identify high-dimensional survival methods for analyzing recurrent events.

2.1.1. Data sources and search strategy

A first search in PubMed, as recommended by Cochrane [4,5], was performed in October 2022 to identify published articles with the following concepts and index terms (MeSh terms) in the title or abstract:

- Survival analysis,
- Recurrent event,
- High-dimensionality,
- Machine Learning.

Appendix 1 details the complete PubMed search strategy.

Secondly, hand searches were also carried out via research engines (Google, Google Scholar, Science direct, Web of Science) and conferences (International Society for Clinical Biostatistics, Association for Computing Machinery, Machine Learning Conference, Journées de Statistique, Medical and Health Informatics) to seek unpublished work such as conference abstracts, papers, and reviews [6]. The hand search strategy included the distinct concepts above that were combined using the following key terms: ‘survival’ or ‘survival analysis’ or ‘time to event,’ ‘recurrence’ ‘recurrent’ or ‘repeated events’ or ‘relapse’ or ‘hospitalization,’ ‘high-dimension’ or ‘machine learning.’

2.1.2. Eligibility criteria

Only published articles in English or French were included. Inclusion criteria were systematic or observational studies that analyzed any recurrent outcome(s), as well as reviews and/or surveys. Exclusion criteria were any Bayesian approach and clinical trial design. The rationale behind this strategy was to ensure consistency with frequentist approaches and real-world applications. Unstructured data such as textual or imaging data were not considered; in our opinion such data are disparate from structured data. Finally, no restrictions regarding the field of healthcare, medical indication or treatment were applied.

2.1.3. Study selection

Two reviewers assessed the eligibility of publications independently and any discrepancies were subsequently discussed. Forward and backward citation tracking was conducted

to avoid missing any relevant literature. Eligible hits were subjected to title and abstracts screening after duplicate removal. The findings of this selection led to the next step in the systematic review process which was the full-text review.

2.1.4. Study characteristics

Study characteristics such as general study setting, location, sample size if applicable, research design, statistical/machine learning approaches, outcomes measured, metrics for evaluation, code availability / reproducibility and application of data (sample description if applicable) were extracted for each included study and summarized. Heterogeneity across studies was not assessed as it was deemed irrelevant to the objective.

2.2. Statistical analysis for application

2.2.1. Notations

Let X_i be a p-dimensional vector of covariates, β the associated regression coefficients, $\lambda_0(t)$ the baseline hazard function, $Y_i(t)$ an indicator of whether subject i is at risk at time t, $\delta_i = 1$ when the subject experienced the event (else 0). Let E_i and C_i be the time to event or censoring, $T_i = E_i \wedge C_i$ for the patient i, with $a \wedge b = \min(a, b)$. $N_i^*(t)$ denoted the number of events over the interval $[0, t]$. Of note, $i = 1, \dots, n$, with n the number of subjects and $\mathbf{X} \in \mathbb{R}^{n*p}$ denoted the covariates matrix for all subjects.

2.2.2. Standard statistical models for modeling recurrent events

Andersen–Gill (AG) [7], Prentice, William and Peterson (PWP) [8], Wei-Lin-Weissfeld (WLW) [9] and the frailty models [10] were developed as extensions of the Cox model [11]. These methodologies commonly use models to handle recurrent event data. Their characteristics are summarized in Table 1. Further details on time scales and how models accounted for subject at risk can be found in Appendix 2. While other statistical approaches exist to model recurrent events, we focused on risk outputs to be able to compare methodologies to one another containing identical metrics. However, this statistical model can handle low-dimensional data only, i.e. when the number of individuals is lower than the number of variables.

2.2.3. Evaluation criteria to measure performance

Few methods are currently available to evaluate a model adjusted for recurrent events. This leads to a lack for model discrimination, i.e. the model cannot differentiate between high- and low-risk individuals who may be subject to the events. We selected the following three criteria to answer the study objectives:

Harrell's Concordance index. Harrell's C-index is a common evaluation criterion in survival analysis [12]. This measure is the proportion of pairs of individuals for which the order of survival times is concordant with the order of the predicted risk. In the presence of censoring, the denominator is the number of pairs of individuals with an event. The C-index is estimated as follows

$$\hat{\mathbb{C}} = \frac{\sum_{i \neq j} I\{\eta_i < \eta_j\} \times I\{T_i > T_j\} \times \delta_j}{\sum_{i \neq j} I\{T_i > T_j\} \times \delta_j} \quad (1)$$

CHAPITRE II Apprentissage pour données censurées

Table 1. Papers identified from the literature review.

#	Year	Author	Title	Type	Description	Data used for application	Evaluation measure	Code availability / reproducibility
#1	2013	Wu et al.	Lasso penalized semiparametric regression on high-dimensional recurrent event data via coordinate descent	Variable selection	Regularization method with penalization Use of the coordinated descent algorithm, which computes in a bi-directional way (forward and backward) the deviations of the optimization problem and updates the parameter value iteratively	Chronic septic granulomatosis	Number of selected predictor variables and regression coefficients	No
#2	2018	Zhao et al.	Variable selection for recurrent event data with broken adaptive ridge regression	Variable selection	Extension of the broken adaptive ridge method to recurrent events, involves repetition and reweighting of penalized L2 models Simultaneous variable selection and parameter estimation, accounts for clustering effects	Chronic septic granulomatosis	MSE Number of predictor variables selected correctly, and number of predictor variables selected incorrectly	Yes
#3	2019	Wang et al.	Machine Learning for Survival Analysis: A Survey	Literature review	Introduction to survival analysis, overview of classical methods and overview of learning methods Recurrent events are mentioned, but ML methods are not developed	/	/	/
#4	2019	Gupta et al.	CRESA: A Deep Learning Approach to Competing Risks, Recurrent Event Survival Analysis.	Deep neural networks	LSTM neural networks with the introduction of the cumulative incidence curve to take into account competitive and/or recurrent events	MIMIC III Machine failure data	Harrell's C-index MAE	No

(continued)

Table 1. Continued.

#	Year	Author	Title	Type	Description	Data used for application	Evaluation measure	Code availability / reproducibility
#5	2019	Jing et al.	A deep survival analysis method based on ranking	Deep neural networks	Extension of the DeepSurv model (neural networks for competitive events) with the use of ranking in the loss function on the differences between observed and predicted values	Myocardial infarction Breast cancer omic data	Harrell's C-index	Yes
#6	2020	Bull et al.	Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods	Literature review	Summary of existing methodology to provide clinical prediction depending on the nature of input data Both statistical and learning approaches are described, but no ML methods for recurrent events highlighted	/	/	/
#7	2021	Kim et al.	Deep Learning-Based Prediction Model for Breast Cancer Recurrence Using Adjuvant Breast Cancer Cohort in Tertiary Cancer Center Registry	Deep neural networks	Use of Weibull Time To Event Recurrent Neural Network, an extension of recurrent neural network to sequentially estimate time to next event	Breast cancer registry in Korea	Harrell's C-index MAE	Yes

Notes: LSTM, long short-term memory; MAE, mean absolute error. Articles were sorted by publication year. #1 to #3 were identified via hand searches and #4 to #7 via PubMed.

With η_i the risk of occurrence of the event. Of note, when two individuals are censored, we cannot know which of the two has the event first. This pair is not included in the calculation. In the same way, if one of the individuals is censored and its censoring time is lower than the event time of another individual, we cannot know which one has the event first. This pair is also not included in the C-index calculation. If the C-index is equal to 1, it means a perfect prediction, and if the C-index ≤ 0.5 , it implies that the model behaves similarly or worse than random. Models with a C-index close to 1 are preferred. Harrell's C-index was computed at each event.

Kim's C-index. Kim et al. [13] proposed a measure of concordance between observed and predicted event counts over a time interval of shared observations. It is the proportion of pairs of individuals for whom the risk prediction and the number of observed events are concordant:

$$\hat{C}_{rec} = \frac{\sum_{i=1}^n \sum_{j=1}^n I\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\} \times I\{\beta^t X_i > \beta^t X_j\}}{\sum_{i=1}^n \sum_{j=1}^n I\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\}} \quad (2)$$

This extension of the C-index implies:

- Two individuals are comparable up to the minimum time of follow-up;
- A pair contributes to the denominator if the two event counts are not equal.

As per Harrell's C-index in Equation (1), a score close to 1 indicates a better performance of the model. As opposed to Harrell's C-index, Kim's C-index was computed once across all the events.

Error rate for active variables. When simulating the datasets, the active status of each variable is known. Methods report the significant variables with a p -value < 0.05 (except deep neural networks). Significant variables are considered as positive tests for their active status. Some active variables likely have a false negative test (FN), and some passive variables have a false positive test (FP). The error rate (err) is the proportion of misclassified variables after prediction:

$$err = \frac{FP + FN}{p} \quad (3)$$

2.2.4. Simulation scheme

The following assumptions were made:

- Active variables were continuous, and have the same (non-zero) effect;
- The variables do not vary over time;
- Individuals were at risk continuously until end of follow-up;
- Censoring is not informative.

The generation of the covariate matrix, $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma(\rho))$. $\mu = (\mu_1 \dots \mu_p) = (a \dots a)$ and $\Sigma(\rho)$ was the covariance matrix with an autoregressive correlation structure and $\rho \in (0, 1)$. The coefficients $\beta = (\beta_1 \dots \beta_p) = (b, \dots, b, 0, \dots, 0)$ were associated with the p covariates. m coefficients were equal to a constant $b \in \mathbb{R}$ (the value of the active coefficients) and $p - m$ coefficients were equal to zero. The sparse rate was described by $\frac{m}{p}$.

The baseline hazard function followed a Weibull distribution with scale $\alpha > 0$ and shape $\gamma > 0$, and $\lambda_0(t) = \alpha\gamma t^{\gamma-1}$. The cumulated baseline hazard function could be expressed as $\Lambda_0(t) = \int_0^t \lambda_0(s)ds = \alpha t^\gamma$. Hence the cumulative hazard function could be expressed as $\Lambda(t) = \Lambda_0(t) \exp(\beta^t X_i)$. Conditional baseline hazard function was then defined as $\tilde{\Lambda}_t(u) := \tilde{\Lambda}^i(u|T_{i-1} = t) = \Lambda(u + t) - \Lambda(t)$. A frailty term z_i i.i.d. was incorporated to account for heterogeneity.

To maintain censoring rates, censored individuals were randomly drawn (censoring is not informative), as per Jahn-Eimermacher et al. [14]. The algorithm of Jahn-Eimermacher et al. [15] was applied to simulate event times k for each subject i :

$$t_{i,1} = \Lambda^{-1}(t)(-\log(\varepsilon_1)) \quad (4)$$

and $t_{i,k+1} = t_{i,k} + \tilde{\Lambda}_{i,t_k}^{-1}(-\log(\varepsilon_{k+1}))$ with $\varepsilon_k \sim U[0, 1]$.

Train-test split was employed with a 70–30% distribution. Datasets were generated with:

- $N = 100$ subjects (low sample size)
- Censoring rate of 20%
- $\rho = 0.7$
- $b = 0.15$
- $\alpha = 1$ and $\gamma = 2$
- $z \sim \text{Gamma}(0.25)$

Scenarios include variations of the number of covariates $p = 25, 50, 100, 150$, and 200 and the sparse rate = 0%, 25%, and 50%. For each one of the 15 scenarios, 100 datasets were generated to account for variability.

3. Results

3.1. Systematic literature review

The search strategy is summarized in Figure 2. Extraction led to the identification of 192 hits through electronic research on the PubMed database. Forty-one studies proceeded to the full-text review step, while the other 151 remaining papers were excluded for further consideration.

Overall, after confirming the outcome of interest dealt with recurrence, the primary reason for exclusion was the non-consideration of recurrent events as time-to-event for each occurrence. Recurrence was considered as a classifier (19/192), as a recurrence-free survival outcome (26/192), or as a time-to-first event (34/192). In this way, the challenge of recurrent data was avoided. The probability of the event was estimated without considering all available information (subsequent event occurrences were omitted in such cases), and were hence biased. This may be the illustration of authors' caution when dealing with recurrent events in high (framework?) dimensions, as no published guidelines or recommendations are available as far as we know. In the field of medicine, the most frequent disease application was cancer (77/192) probably due to the high level of sustained interest in this condition. Among cancers, however, no type stood out (colon/colorectal cancer 34/192, breast cancer 20/192, lung cancer 15/192, other cancer 8/192). In addition, four

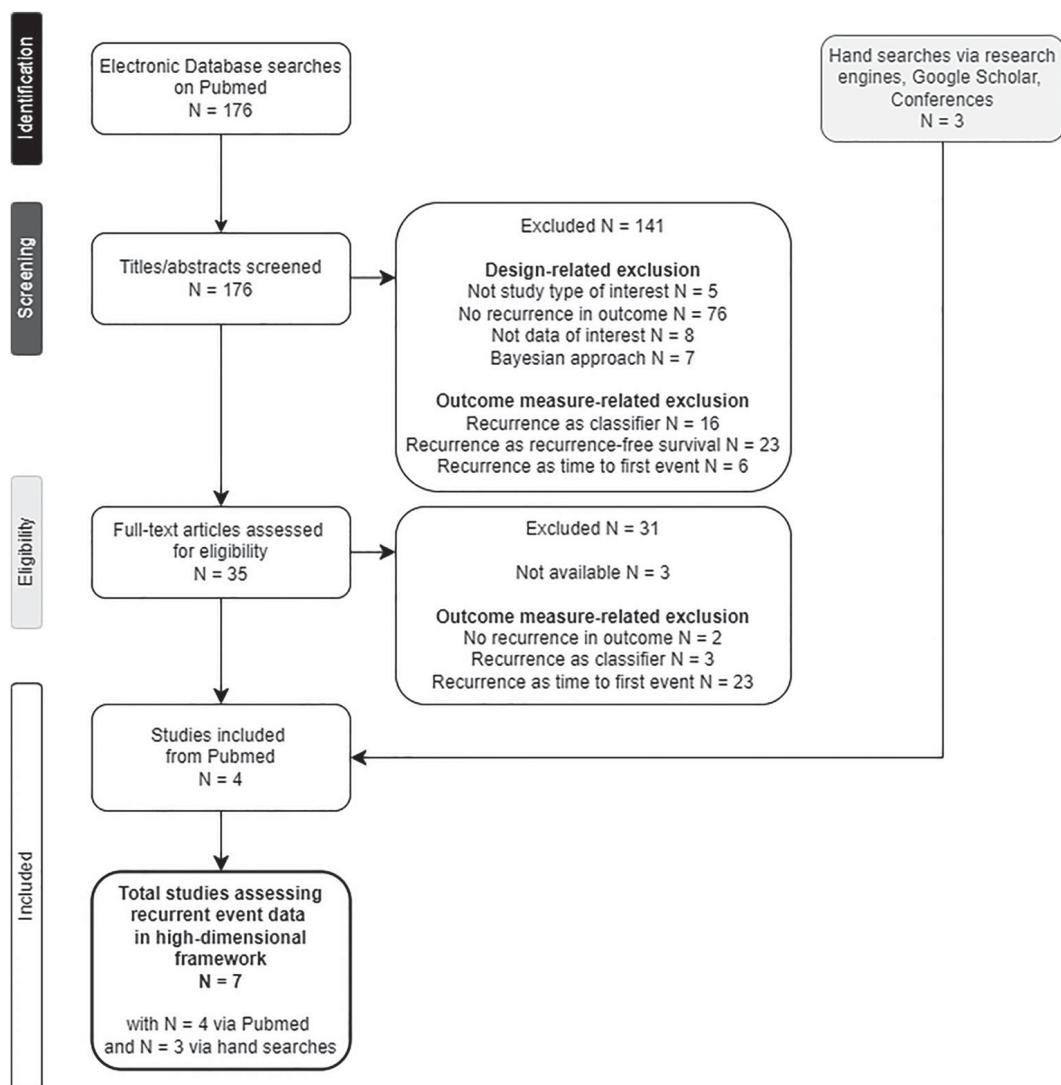


Figure 2. Flowchart of included publications via PubMed.

full-text articles could not be reviewed as they were not available. After title, abstract and full-text thorough review, four publications were included from the electronic database search. Three additional papers from the hand searches were identified. Subsequently, a total of seven relevant publications were selected (Table 2).

The two first publications are literature reviews. Work from Wang et al. [16] and Bull et al. [17] are recent comprehensive surveys in which classical and contemporary methods of survival analysis are reported. They both underline the development over the last decade of more complex approaches to dealing with longitudinal data to predict survival outcomes, e.g. joint models and deep neural networks. Recurrent events may be seen as longitudinal outcomes but were not addressed per se and were only mentioned as specific data structure.

Four methodological articles were also selected presenting a significant variation in methodology. Two articles describing learning algorithms for variable selection strategies

Table 2. Standard statistical models for recurrent events analyses.

Model	Components and specificities
AG	Conditional model, accounts for the counting process as a time scale and unrestricted set for subjects at risk Recurrent events within individuals are independent and share a common baseline hazard function Intensity of the model: $\lambda_i(t) = Y_i(t) \times \lambda_0(t) \times \exp(\beta^t X_i)$
PWP	Conditional model, counting process as time scale and restricted set for subjects at risk Stratified AG, stratum k collects all the kth events of the individuals Hazard function for each event Hazard function: $\lambda_{ik}(t) = Y_i(t) \times \lambda_{0k}(t) \times \exp(\beta_k^t X_i)$
WLW	Marginal model, also stratified, calendar time scale and semi-restricted set for subjects at risk Intra-subject dependence Hazard function: $\lambda_{ik}(t) = Y_i(t) \times \lambda_{0k}(t) \times \exp(\beta_k^t X_i)$
Frailty	Extension of AG model Random term z_i for each individual to account for unobservable or unmeasured characteristics Hazard function: $\lambda_i(t) = Y_i(t) \times \lambda_0(t) \times z_i \times \exp(\beta^t X_i)$

Notes: AG, Andersen–Gill; PWP, Prentice, William and Peterson; WLW, Wei-Lin-Weissfeld.

were identified. Firstly, Wu [18] focused on accelerating coefficient estimation with a co-ordinated descent algorithm and penalizing partial likelihood, followed by Zhao et al. [19] work which provided an extension of Ridge penalization for estimating and selecting variables simultaneously. Finally, the other two selected articles are from Gupta et al. [20] and Jing et al. [21], and developed deep neural networks extensions for the analysis of recurrence.

An additional paper was selected, which aimed at estimating time between two breast cancer recurrences using a Weibull Time To Event Recurrent Neural Network [22]. However, the methodology used was an extension of a recurrent neural network and was not published in any peer-reviewed journal [23].

Findings from the present review highlight the current gap in the literature and vast differences in the context and methods of interest. In particular, not all developed models were based on simulated data, as Jing et al. [21]. Additionally, none of the included publications compared their performance to others. For instance, variable selection approaches were compared to standard statistical model only, while neural networks were compared to other neural networks or random forests. No head-to-head comparison across standard methods, learning algorithms and deep neural networks seem to have been performed.

3.2. Application to simulated data

We proposed testing the identified methods on simulated data. Only two methods had open-sourced code: Variable selection from Zhao et al. [19] and the deep neural network RankDeepSurv from Jing et al. [21].

3.2.1. Methods selected

3.2.1.1. Learning algorithms for variable selection. A common approach to addressing high-dimension challenge is variable selection. Penalizing models helps to reduce the space of parameter coefficients, called shrinkage. Widely used for regression and classification problems, Lasso penalization accepts null coefficients to select variables [24] and Ridge helps to deal with multicollinearity in the data [25]. Both penalization approaches have been extended to Cox models in the standard survival analysis framework [26,27]. The purpose is to solve a constrained optimization problem of the partial log-likelihood of the

Cox model, which is written

$$\mathcal{L}(\beta) = \sum_{i=1}^n \delta_i \beta^t X_i - \sum_{i=1}^n \delta_i \times \log \sum_{j \in \mathcal{R}(\tau_i)} \exp(\beta^t X_j) \quad (5)$$

With $\mathcal{R}(t)$ the set of individuals who are ‘at risk’ for the event at time t . For CoxLasso, regularization is performed using an L_1 norm penalty and $\hat{\beta} = \operatorname{argmax}_{\beta} \mathcal{L}(\beta)$, $\|\beta\|_1 \leq s$ and for CoxRidge an L_2 norm penalty and $\hat{\beta} = \operatorname{argmax}_{\beta} \mathcal{L}(\beta)$, $\|\beta\|_2 \leq s$, with $s \geq 0$. The lower the value of s , the stronger the penalization. Hyperparameters, named penalty coefficients, are used to determine its value, and enable the control of the impact of the penalty.

Zhao et al. [19] proposed an extension of these methods to recurrent events by developing the broken adaptive ridge (BAR) regression. The first iteration consists of a penalized L_2 model

$$\hat{\beta}^{(0)} = \operatorname{argmin}_{\beta} \left(-2 \mathcal{L}_{mod}(\beta) + \xi_n \sum_{j=1}^p \beta_j^2 \right), \xi_n \geq 0 \quad (6)$$

If penalization hyperparameter $\xi_n > 0$, this is a Ridge penalty, and if $\xi_n = 0$ then $\hat{\beta}^{(0)}$ is not penalized. We update for each iteration ω :

$$\hat{\beta}^{(\omega)} = \operatorname{argmin}_{\beta} \left(-2 \mathcal{L}_{mod}(\beta) + \vartheta_n \sum_{j=1}^p \frac{\beta_j^2}{(\hat{\beta}_j^{(\omega-1)})^2} \right), \omega \geq 1 \quad (7)$$

BAR estimates are defined by $\hat{\beta} = \lim_{k \rightarrow \infty} \hat{\beta}^{(\omega)}$. The estimator benefits from the oracle properties of both penalties for model covariate selection and estimation. Cross-validation is recommended to optimize values of hyperparameters ξ_n and ϑ_n . According to Kawaguchi et al. [28], estimates are not sensitive to variations of ξ_n and optimization can be performed only on ϑ_n . In the absence of a consensual single measure on cross-validation under recurrent events, two values for ϑ_n were studied in this paper, thereby covering two separate models. Such penalty was applied to models presented in the previous subsection.

3.2.1.2. Deep neural network. RankDeepSurv is a deep neural network proposed by Jing et al. [21] with fully connected layers (all neurons in one layer are connected to all neurons in another layer). The specificity of the RankDeepSurv neural network lies in the loss function adapted to survival, which results from the sum of two terms: one to constrain the survival model using an extension of the mean square error and the other to evaluate the rank error between observed and predicted values for two individuals. The loss function is written as

$$L_{loss}(\theta) = \alpha_1 L_1(\theta) + \alpha_2 L_2(\theta) + \mu \|\theta\|_2^2 \quad (8)$$

with $\alpha_1, \alpha_2 > 0$ constant values, θ the weights of the network, μ the regularization parameter for L_2 ; $L_1 = \frac{1}{n} \sum_{i=1, I(i)=1}^n (y_{j,pred} - y_{j,obs})^2$, $I(i) = 1$ if i is censored or if the predicted time to event is before the observed time, else 0; $L_2 =$

$\frac{1}{n} \sum_{I(i,j)=1}^n [(y_{j,obs} - y_{i,obs}) - (y_{j,pred} - y_{i,pred})]^2$, $I(i,j) = 1$ if $y_{j,obs} - y_{i,obs} > y_{j,pred} - y_{i,pred}$, else 0. Gradient descent is utilized for solving the minimization of L_{loss} .

3.2.2. Results of the application

Simulated datasets had identical characteristics in terms of number of individuals, structure of covariates, but differed across scenarios in terms of number of covariates and sparse rate. In the variance-covariance matrix, the covariates were highly correlated when they were close, then decreasingly correlated when they were further apart. Appendix Figure A2 captured this relationship across covariates with five datasets, regardless of the number of covariates. Figure 3 provides a visual representation of the history of nine individuals and their events over the follow-up period (and helps to understand Figure 1).

3.2.2.1. Impact of the number of covariates on the average C-index. Average C-indices were computed across all 15 scenarios (Figure 4). As expected, the standard models failed as soon as $p > n$. Whereas the C-indices were also expected to be around 0.5 when the sparse rate was zero, they increased as the sparse rate increased. The best performance was obtained using the frailty model. Other models showed similar trends, except for the WLW and RankDeepSurv models. The C-indices of these two models remained around the value of 0.5 (and even below) regardless of the scenario. The Kim's C-index was more stable across the different number of covariates and sparse rates, although it tended to decrease as the number of covariates increased with sparse rate = 50%. Small differences across penalty values were noticed as 0.05 penalized models and 0.1 penalized models followed similar trends.

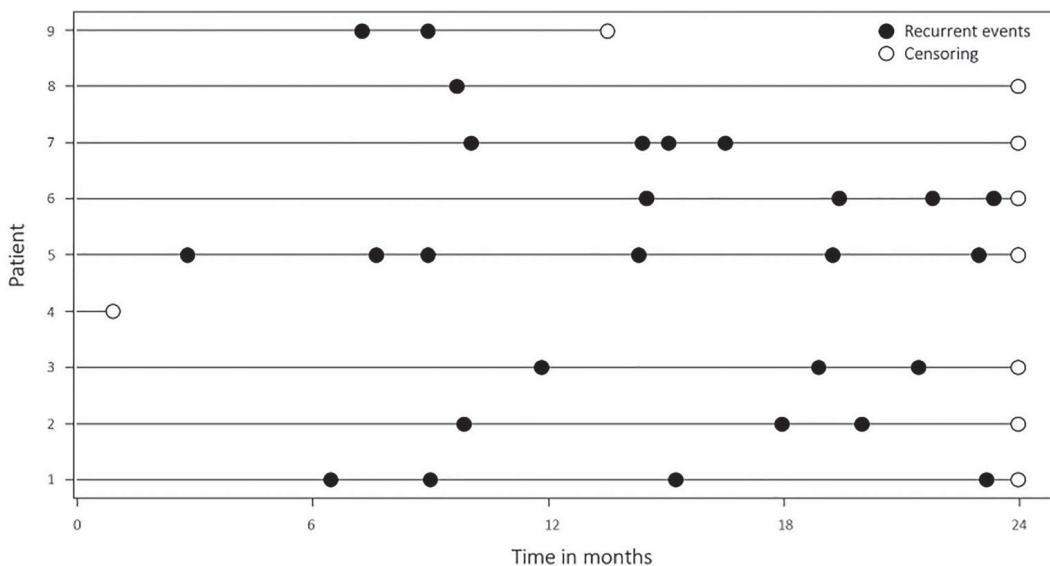


Figure 3. History of the 9 first simulated individuals (for a given training set). Notes: A row referred to a patient with their event history. The time on the x-axis was the follow-up period. Solid circles corresponded to events and empty circles represented censoring.

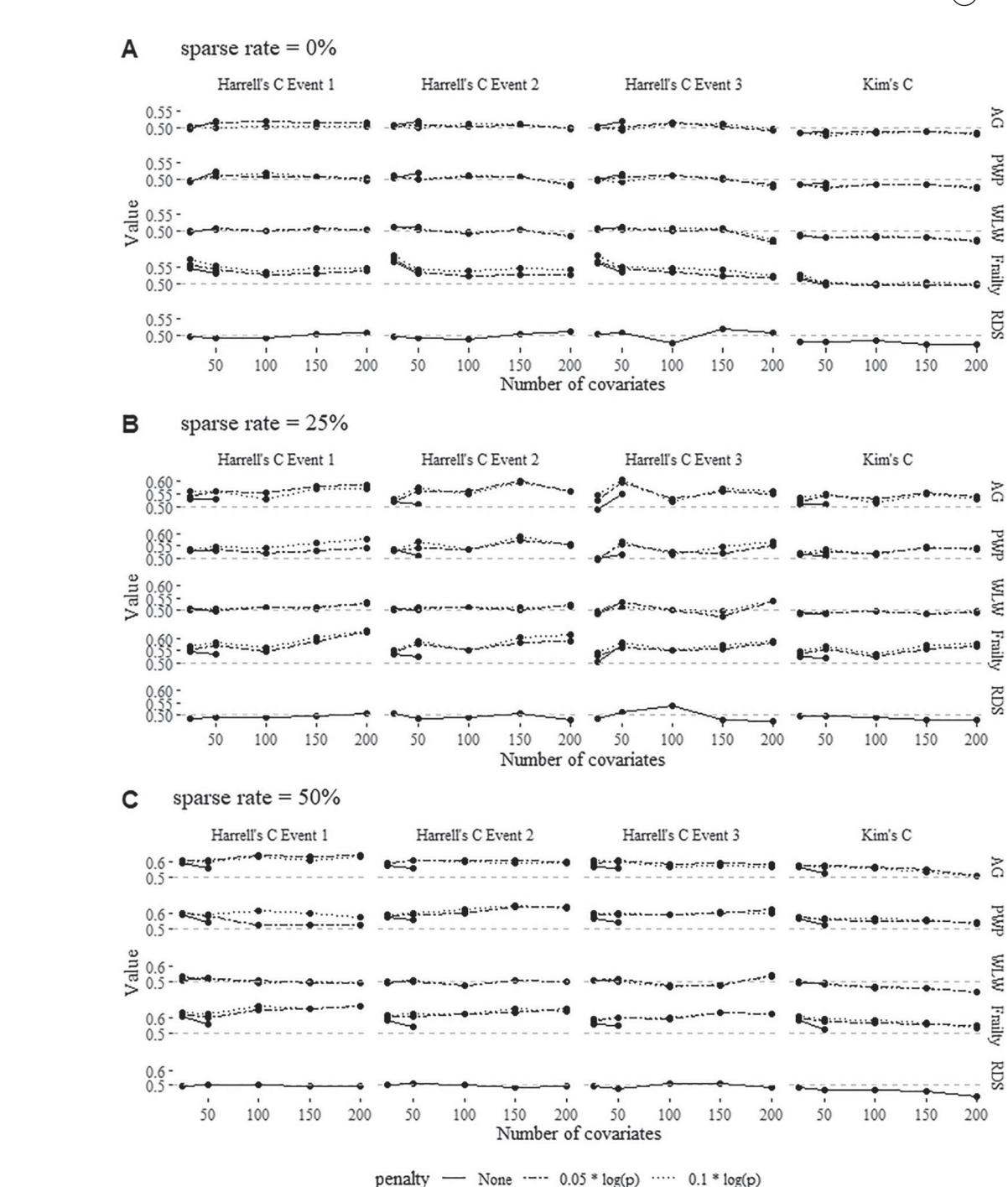


Figure 4. Impact of the number of covariates on average C-indices with sparse rate equal to 0% (A), 25% (B) and 50% (C). Notes: p the number of covariates. For each sparse rate, model and penalty, the average C-indices of the 100 simulated datasets were displayed over the number of covariates. The penalties were equal to 0 (unpenalized), $0.05 \times \log(p)$ and $0.1 \times \log(p)$, respectively. Penalties > 0 were applicable only for standard statistical models, RankDeepSurv deep neural network was hence not penalized. Unpenalized standard statistical models did not converge as soon as $p > n$, performance was therefore not available. AG: Andersen–Gill; PWP: Prentice, RDS: RankDeepSurv, William, and Peterson; WLW: Wei-Lin-Weissfeld.

3.2.2.2. Focus on the variability of C-indices for two extreme scenarios. Two extreme scenarios were thoroughly studied: one with no active variable and only 25 variables (A), and another one in which models overall reported greater performance with a sparse rate of 25% and over 150 variables (B). Similar trends in variability were observed across these two scenarios and for each C-index (Appendix Figure A3). Kim's C-index was the less volatile across models and their penalties, with values ranging between 0.39 and 0.63 and 0.28 and 0.76 for (A) and (B), respectively. Harrell's C-index was increasingly variable in the first event (A: min = 0.26 and max = 0.74; B: min = 0.24 and max = 0.81), second event (A: min = 0.30 and max = 0.75; B: min = 0.17 and max = 0.85), and third event (A: min = 1 and max = 0).

3.2.2.3. Error rate for active variables. Results regarding average error rates are displayed in Appendix Figure A4. For scenarios where the sparse rate was equal to 0%, all models reported average error rates below 0.5, except penalized WLW with error rates around 0.75. Average error rates appeared similar when no penalty was applied. The AG model had the lowest average error rate for each p , with a minimum value of 0.018 for the penalized model at $0.1 \times \log(p)$ and $p = 200$. Average error rate decreased when the penalty increased when $p > n$. For scenarios with a sparse rate equal to 25%, the unpenalized frailty model had the best performance, while the other models provided higher values. Similarly, penalties decreased the average error rate. Penalized AG models reported average error rates lower than 0.3. Finally, when the sparse rate was equal to 50%, almost constant average error rates around 0.5 were observed for each model regardless of p .

4. Discussion

The present systematic literature review enabled the identification of emerging approaches. A total of seven publications were included, highlighting the limitation of available resources in this area. Methods herein identified were based on existing model extensions to the recurrent survival framework, which included variable selection approaches and neural networks. As with the standard models presented, they have both strengths and drawbacks. It is therefore necessary to tailor them to the clinical setting in order to meet the stated goals in a meaningful way.

At the same time, these approaches had not tested against one another. This may lead to erratic behavior and confusion when researchers aim to conduct robust and reliable analyses in this context. The present study thus proposed to evaluate some of the available open-sourced innovative learning algorithms developed to solve the high-dimensional framework when considering recurrent events. The investigation of the beforementioned 15 scenarios on simulated data highlighted specificities of both the methodology and measures used for the evaluation of their performance.

Firstly, unpenalized standard approaches failed as soon as $p > n$ as expected, while penalized approaches helped to improve their performance when $p < n$. This was typically expected as standard statistical models were not designed for $p > n$ cases. AG and PWP models reported equivalent performance, while the frailty model consistently had the best performance. This was due to the construction of the frailty term from the simulation scheme. The WLW model performed in an inferior manner, regardless

of penalization or not. This finding was consistent with results in the literature, suggesting WLW models to be more appropriate with events of different types rather than recurrent events [29,30]. Nevertheless, these models, each with their own specificities, can respond to differing needs, especially related to the research questions [1,3,31]. Secondly, variable selection with penalties did not significantly increase performance, and few variables were even selected when the sparse rate was zero. Since only two values for the hyperparameter were explored, it seemed quite unlikely these would maximize model performance. The deep neural network reported poorer performance; one reason could be that the format of the data was not suitable for the code. In this case, average error rates increased with the sparse rate. When the number of active variables was higher, models tended to select the wrong variables. It appears as if the models have a harder time learning and selecting the true active variables in the advent of a high sparse rate, however they managed to report better C-indices in this situation. This was related to the variance-covariance structure chosen for data simulation. With regards to evaluation metrics, Kim's C-index has shown higher stability and robustness compared to other metrics and stood for a criterion evaluating the entire set of events. Harrell's C, on the other hand, was measured at each event, making it difficult to be interpreted in terms of global performance.

Nevertheless, some limitations should be noted. The literature review presented several drawbacks. Publications whose objective was variable selection without explicit dimension reduction, such as Tong et al. [32] and Chen et al. [33] could not be captured because of the elaborated research strategy. In addition, it is not always simple to assess how the outcome was considered, especially for neural networks that make little mention of the expected structure to process the data. Furthermore, as mentioned above, the lack of hyperparameter optimization for variable selection made BAR approach inconclusive. Lastly, a cross-validation would have highlighted the robustness of the results [34].

Other evaluation measures have been used in the literature, e.g. the mean square error, the mean absolute error, the log-likelihood [18,19,30]. An additional approach to investigating active variables would be to assess the importance of the variables by permutation [35]. When choosing a performance measure beforehand, this consists in permuting k times for the order of the covariates and calculating k times for the performance of the model. We note however that the simulations scheme itself presented several drawbacks. Covariates were not time-dependent and shared the same effect on the outcome, which may seem implausible in real life and made the interpretation of the results difficult to generalize. Also, although the simulation of the data maintained censoring rates, it was not based on a distribution of censoring time, while one should be able to genuinely control [36,37].

5. Conclusion

As far as we know, this is the first study to compare standard methods, variable selection algorithms, and a deep neural network in modeling recurrent events in a high-dimensional framework, and to specifically measure the impact of the number of covariates.

Progress in medical care is leading to the use of embedded artificial intelligence (AI) technologies, evidenced by the booming market for AI medical devices. In this context,

these systems are typically designed to prevent the occurrence of events at the hospital, elderly care home or outpatient setting, for example. Where these events are likely to occur repeatedly, and all available data/knowledge is captured, then thorough, robust and appropriate analysis of recurrent events is crucial [38].

Overall, this work raises many concerns for recurrent event data analysis in high-dimensional settings. In addition, it highlights the current need for developing further approaches in order to assess their performance in a relevant manner.

Acknowledgements

The authors would like to thank the reviewers for their constructive comments that led to improvements of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially funded by a grant from the Association Nationale de la Recherche et de la Technologie, with Pierre Fabre, Convention industrielle de formation par la recherche number 2020/1701.

Data availability statement

All data were simulated in line with simulation scheme detailed in the article.

Notes on contributors

Juliette Murris is a researcher in biostatistics and biomathematics within Inserm and Inria Paris. Besides her PhD project, she is a part-time biostatistician at Pierre Fabre.

Anaïs Charles-Nelson holds a PhD in biostatistics. Her research focused on recurrent and terminal events.

Abir Tadmouri Sellier holds a PhD and an MPH. She is the head of the real-world evidence and data department at Pierre Fabre.

Audrey Lavenu holds a PhD and an authorization in conducting research in applied mathematics. Her research focuses on epidemics, learning and biostatistics.

Sandrine Katsahian is the head the the clinical research unit at the Pompidou hospital. She is also a permanent researcher within Inserm and Inria Paris.

ORCID

Juliette Murris  <http://orcid.org/0000-0002-7017-9865>

Anaïs Charles-Nelson  <http://orcid.org/0001-6437-7059>

Audrey Lavenu  <http://orcid.org/0000-0002-0049-2397>

Sandrine Katsahian  <http://orcid.org/0000-0002-7261-0671>



References

- [1] Rogers JK, Pocock SJ, McMurray JJV, et al. Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-preserved. *Eur J Heart Fail.* **2014**;16(1):33–40. doi:[10.1002/ejhf.29](https://doi.org/10.1002/ejhf.29)
- [2] Twisk J, Smidt N, de Vente W. Applied analysis of recurrent events: a practical overview. *J Epidemiol Community Health.* **2005**;59(8):706–710. doi:[10.1136/jech.2004.030759](https://doi.org/10.1136/jech.2004.030759)
- [3] Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. *Int J Epidemiol.* **2015**;44(1):324–333. doi:[10.1093/ije/dyu222](https://doi.org/10.1093/ije/dyu222)
- [4] Tacconelli E. Systematic reviews: CRD's guidance for undertaking reviews in health care. *Lancet Infect Dis.* **2010**;10(4):226. doi:[10.1016/S1473-3099\(10\)70065-7](https://doi.org/10.1016/S1473-3099(10)70065-7)
- [5] Higgins JPT, Thomas J, Chandler J, et al., editors. *Cochrane handbook for systematic reviews of interventions.* 2nd ed. Chichester (UK): John Wiley & Sons; **2019**. 726 p.
- [6] Guide to the methods of technology appraisal 2013. 2013;94.
- [7] Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat.* **1982**;10(4):1100–1120. doi:[10.1214/aos/1176345976](https://doi.org/10.1214/aos/1176345976)
- [8] Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. *Biometrika.* **1981**;68(2):373–379. doi:[10.1093/biomet/68.2.373](https://doi.org/10.1093/biomet/68.2.373)
- [9] 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. doi:[10.1080/01621459.1989.10478873](https://doi.org/10.1080/01621459.1989.10478873)
- [10] Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography.* **1979**;16(3):439–454. doi:[10.2307/2061224](https://doi.org/10.2307/2061224)
- [11] Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Methodol.* **1972**;34(2):187–202.
- [12] Harrell FE Jr, Calif RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA.* **1982**;247(18):2543–2546. doi:[10.1001/jama.1982.03320430047030](https://doi.org/10.1001/jama.1982.03320430047030)
- [13] Kim S, Schaubel DE, McCullough KP. A C-index for recurrent event data: application to hospitalizations among dialysis patients. *Biometrics.* **2018**;74(2):734–743. doi:[10.1111/biom.12761](https://doi.org/10.1111/biom.12761)
- [14] Jahn-Eimermacher A, Ingel K, Ozga AK, et al. Simulating recurrent event data with hazard functions defined on a total time scale. *BMC Med Res Methodol.* **2015**;15(1):16. doi:[10.1186/s12874-015-0005-2](https://doi.org/10.1186/s12874-015-0005-2)
- [15] Jahn-Eimermacher A. Comparison of the Andersen-Gill model with Poisson and negative binomial regression on recurrent event data. *Comput Stat Data Anal.* **2008**;52(11):4989–4997. doi:[10.1016/j.csda.2008.04.009](https://doi.org/10.1016/j.csda.2008.04.009)
- [16] Wang P, Li Y, Reddy CK. Machine learning for survival analysis: a survey. *ACM Comput Surv.* **2019**;51(6):110:1–110:36. doi:[10.1145/3214306](https://doi.org/10.1145/3214306)
- [17] Bull LM, Lunt M, Martin GP, et al. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. *Diagn Progn Res.* **2020**;4:9. doi:[10.1186/s41512-020-00078-z](https://doi.org/10.1186/s41512-020-00078-z)
- [18] Wu TT. Lasso penalized semiparametric regression on high-dimensional recurrent event data via coordinate descent. *J Stat Comput Simul.* **2013**;83(6):1145–1155. doi:[10.1080/00949655.2011.652114](https://doi.org/10.1080/00949655.2011.652114)
- [19] Zhao H, Sun D, Li G, et al. Variable selection for recurrent event data with broken adaptive ridge regression. *Can J Stat.* **2018**;46(3):416–428. doi:[10.1002/cjs.11459](https://doi.org/10.1002/cjs.11459)
- [20] Gupta G, Sunder V, Prasad R, et al. CRESA: a deep learning approach to competing risks, recurrent event survival analysis. In: Yang Q, Zhou ZH, Gong Z, et al., editors. *Advances in knowledge discovery and data mining.* Cham: Springer International Publishing; **2019**. p. 108–122.
- [21] Jing B, Zhang T, Wang Z, et al. A deep survival analysis method based on ranking. *Artif Intell Med.* **2019**;98:1–9. doi:[10.1016/j.artmed.2019.06.001](https://doi.org/10.1016/j.artmed.2019.06.001)
- [22] Kim JY, Lee YS, Yu J, et al. Deep learning-based prediction model for breast cancer recurrence using adjuvant breast cancer cohort in tertiary cancer center registry. *Front Oncol.* **2021**;11:596364. doi:[10.3389/fonc.2021.596364](https://doi.org/10.3389/fonc.2021.596364)

- [23] Martinsson E. A model for sequential prediction of time-to-event in the case of discrete or continuous censored data, recurrent events or time-varying covariates. 2017.
- [24] Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol.* **1996**;58(1):267–288.
- [25] Hilt DE, Seegrist DW, Northeastern Forest Experiment Station (Radnor, Pa.), United States. Ridge, a computer program for calculating ridge regression estimates. Upper Darby (PA): Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station; **1977**; 10 p.
- [26] Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med.* **1997**;16(4):385–395. doi:[10.1002/\(SICI\)1097-0258\(19970228\)16:4<385::AID-SIM380>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3)
- [27] Simon N, Friedman J, Hastie T, et al. Regularization paths for Cox’s proportional hazards model via coordinate descent. *J Stat Softw [Internet].* **2011** [cited 2022 Feb 16];39(5). Available from: <http://www.jstatsoft.org/v39/i05/>
- [28] Kawaguchi ES, Suchard MA, Liu Z, et al. A surrogate ℓ_0 sparse Cox’s regression with applications to sparse high-dimensional massive sample size time-to-event data. *Stat Med.* **2020**;39(6):675–686. doi:[10.1002/sim.8438](https://doi.org/10.1002/sim.8438)
- [29] Ozga AK, Kieser M, Rauch G. A systematic comparison of recurrent event models for application to composite endpoints. *BMC Med Res Methodol.* **2018**;18(1):2. doi:[10.1186/s12874-017-0462-x](https://doi.org/10.1186/s12874-017-0462-x)
- [30] Ullah S, Gabbett TJ, Finch CF. Statistical modelling for recurrent events: an application to sports injuries. *Br J Sports Med.* **2014**;48(17):1287–1293. doi:[10.1136/bjsports-2011-090803](https://doi.org/10.1136/bjsports-2011-090803)
- [31] 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. *Stat Med.* **2019**;38(18):3476–3502. doi:[10.1002/sim.8168](https://doi.org/10.1002/sim.8168)
- [32] Tong X, Zhu L, Sun J. Variable selection for recurrent event data via nonconcave penalized estimating function. *Lifetime Data Anal.* **2009**;15(2):197–215. doi:[10.1007/s10985-008-9104-2](https://doi.org/10.1007/s10985-008-9104-2)
- [33] Chen X, Wang Q. Variable selection in the additive rate model for recurrent event data. *Comput Stat Data Anal.* **2013**;57(1):491–503. doi:[10.1016/j.csda.2012.06.019](https://doi.org/10.1016/j.csda.2012.06.019)
- [34] Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. *Biom J.* **2018**;60(3):431–449. doi:[10.1002/bimj.201700067](https://doi.org/10.1002/bimj.201700067)
- [35] Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable’s importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res.* **2019**;20(177):1–81.
- [36] Wan F. Simulating survival data with predefined censoring rates for proportional hazards models: simulating censored survival data. *Stat Med.* **2017**;36(5):838–854. doi:[10.1002/sim.7178](https://doi.org/10.1002/sim.7178)
- [37] Pénichoux J, Moreau T, Latouche A. Simulating recurrent events that mimic actual data: a review of the literature with emphasis on event-dependence. *ArXiv150305798 Stat [Internet].* 2015. Available from: <http://arxiv.org/abs/1503.05798>
- [38] Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *NPJ Digit Med.* **2021**;4(1):153. doi:[10.1038/s41746-021-00521-5](https://doi.org/10.1038/s41746-021-00521-5)



Appendices

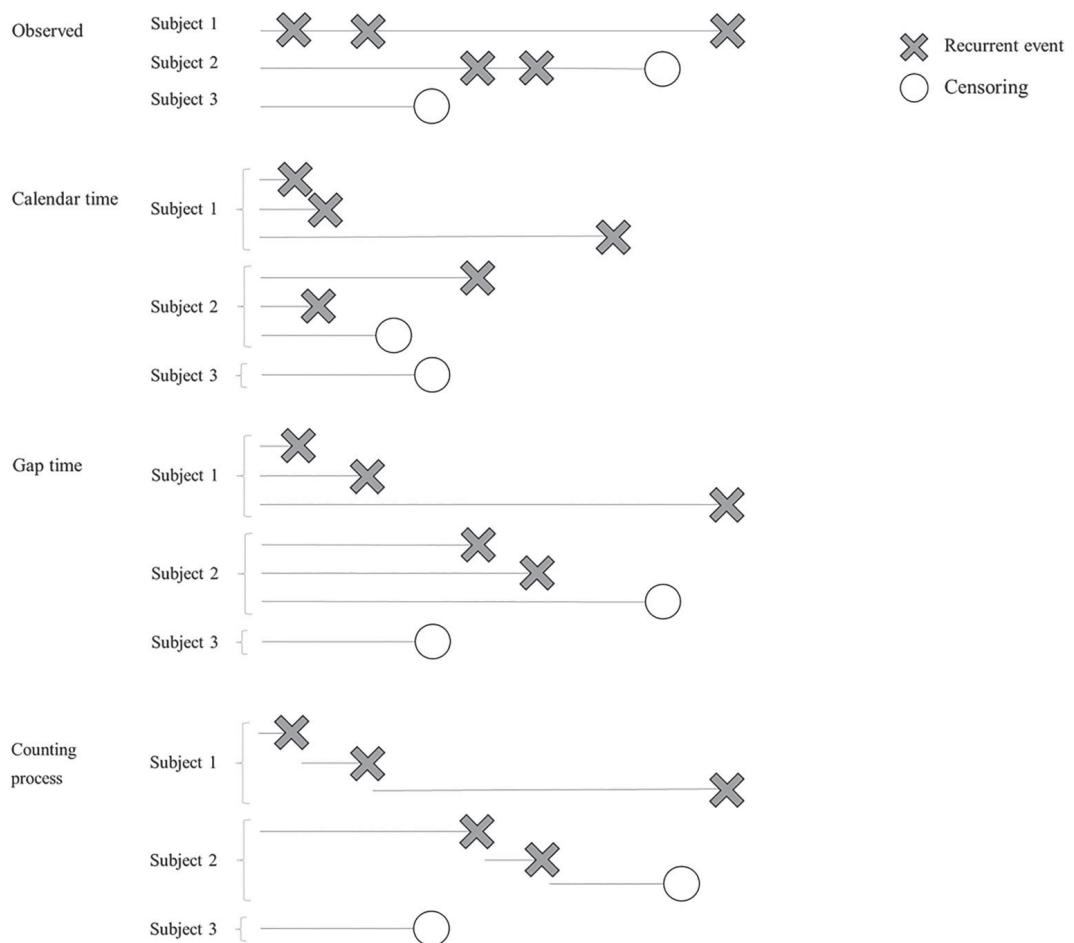


Figure A1. Timescales in recurrent events analysis.

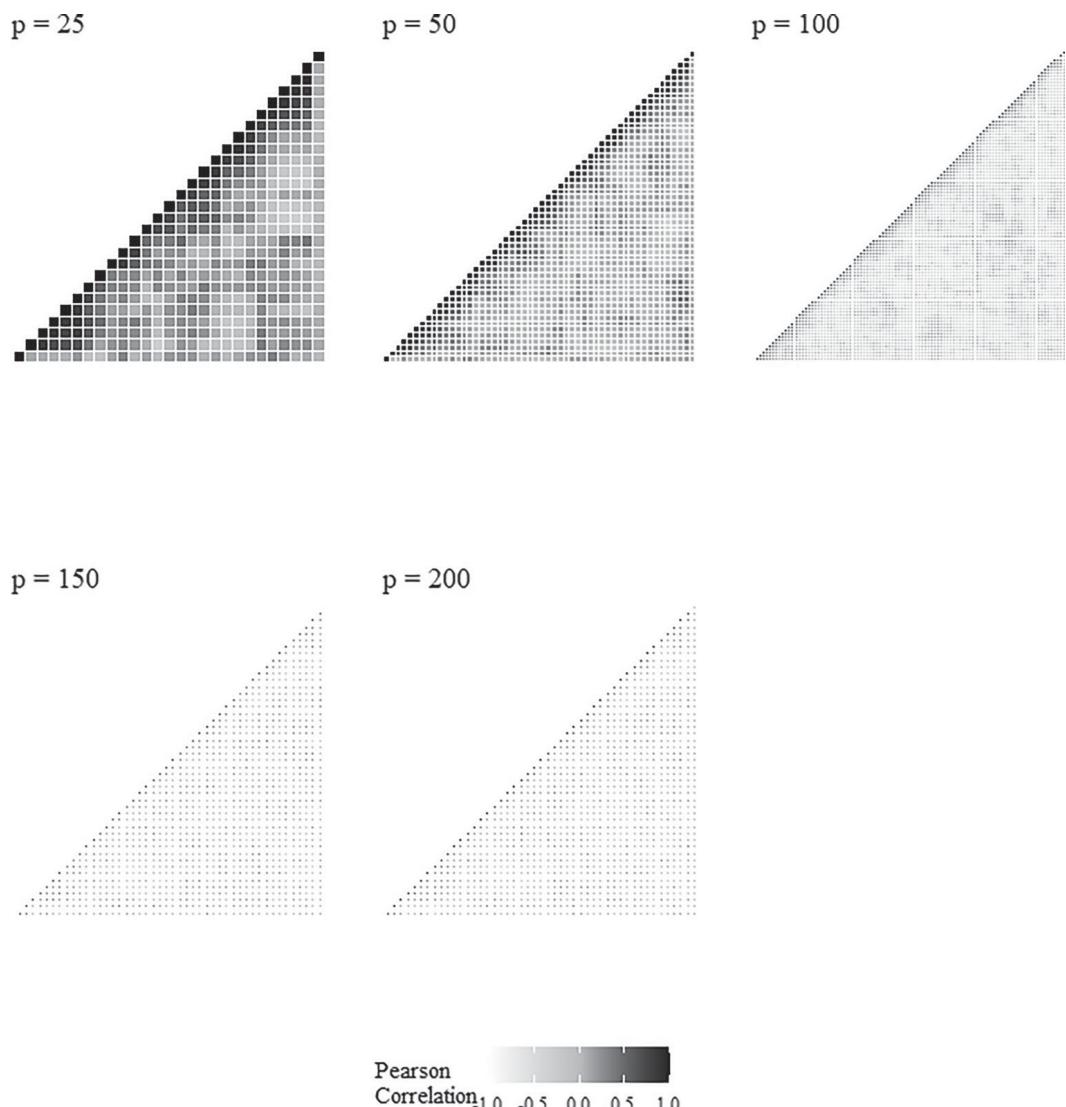


Figure A2. Heatmaps of correlation with variations of the number of variables (25, 50, 100, 150, 200). Notes: Each square provided the Pearson correlation coefficient between the covariate on the x-axis and the one on the y-axis. All coefficients on the diagonal were equal to 1 as it was the correlation coefficient between a covariate and itself.

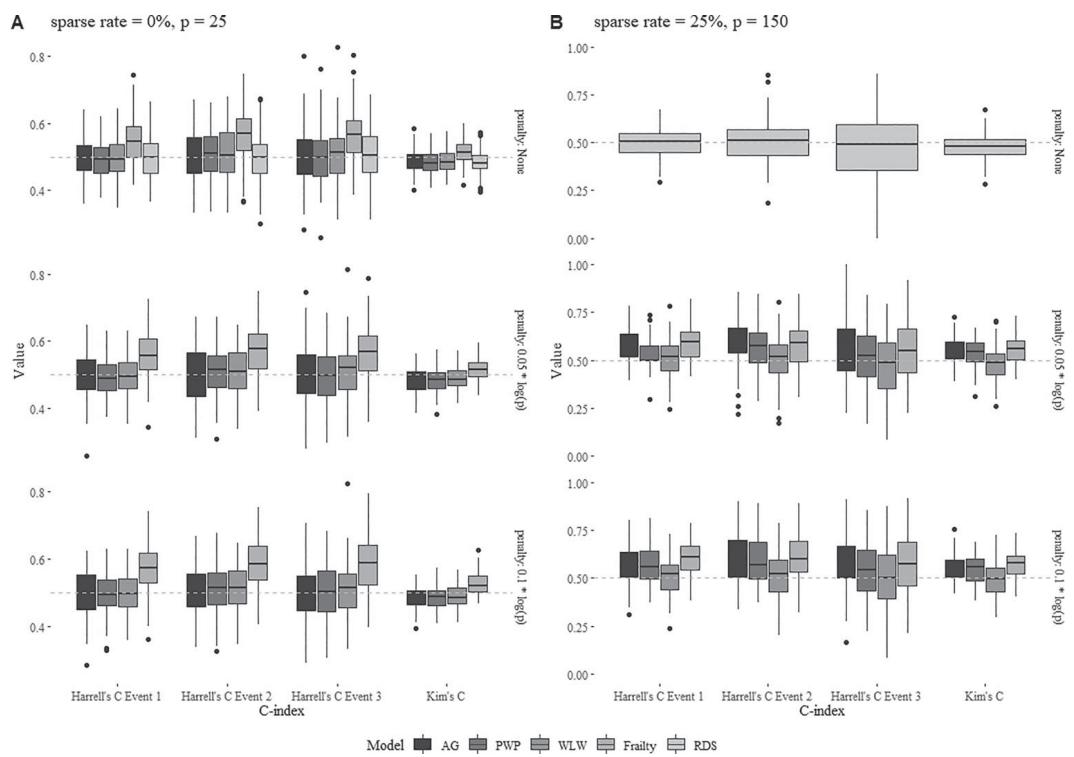


Figure A3. Variability of C-indices for two extreme scenarios: sparse rate = 0%, $p = 25$ (A) and sparse rate = 25%, $p = 150$ (B). Notes: p was the number of covariates. For each model and penalty, the C-indices of the 100 simulated datasets were summarized in a boxplot. The penalties were equal to 0 (unpenalized), $0.05 \times \log(p)$ and $0.1 \times \log(p)$, respectively. Penalties > 0 were applicable only for standard statistical models, RankDeepSurv deep neural network was hence not penalized. Unpenalized standard statistical models did not converge as soon as $p > n$, performance was therefore not available. C-AG: Andersen-Gill; PWP: Prentice, William, and Peterson; RDS: RankDeepSurv, WLW: Wei-Lin-Weissfeld.

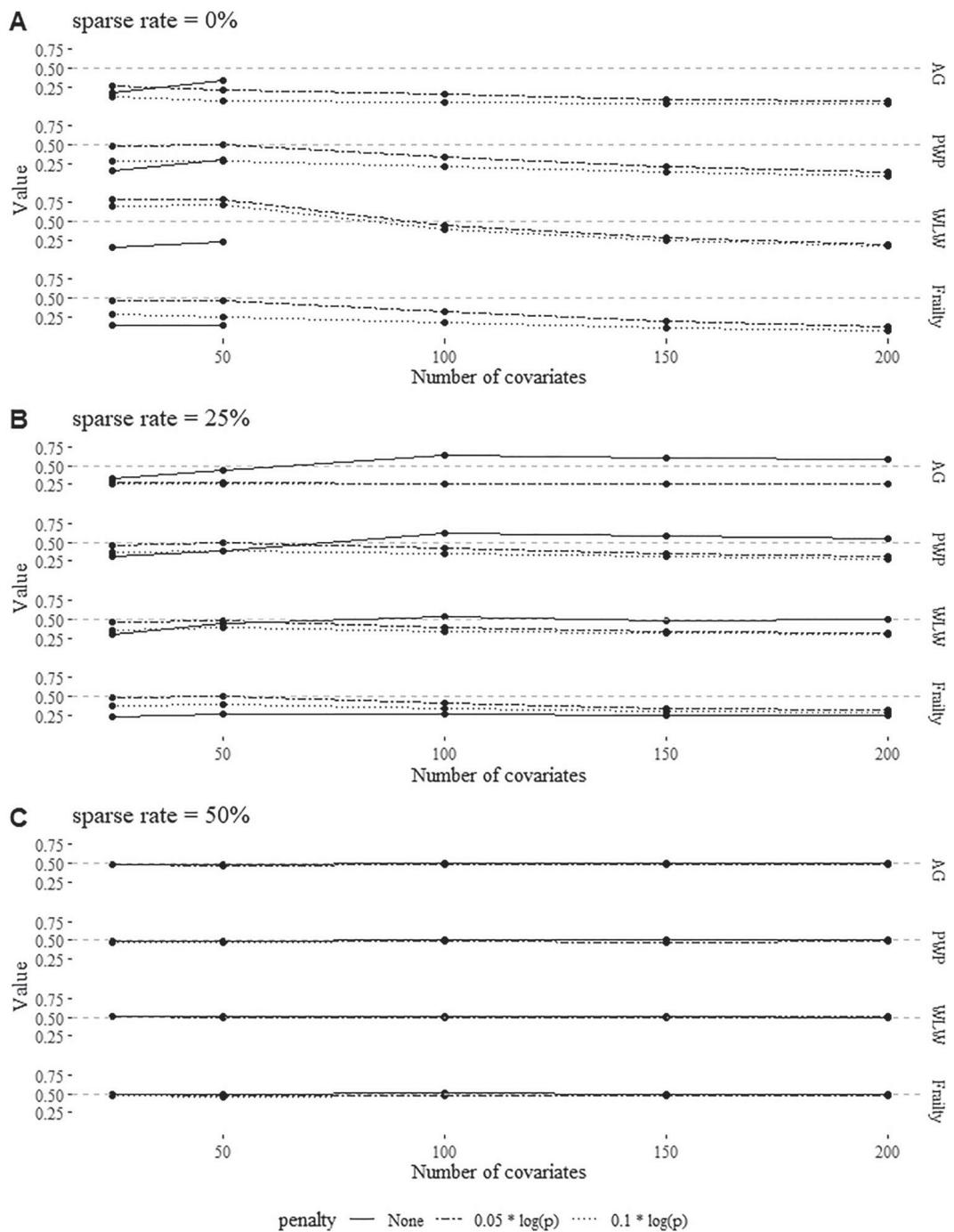


Figure A4. Average error rates with sparse rate equal to 0% (A), 25% (B) and 50% (C). Notes: p was the number of covariates. For each sparse rate, model and penalty, the average error rates of the 100 simulated datasets were displayed over the number of covariates. The penalties were equal to 0 (unpenalized), $0.05 \times \log(p)$ and $0.1 \times \log(p)$, respectively. Unpenalized standard statistical models did not converge as soon as $p > n$, error rate assessment was therefore not available. AG: Andersen–Gill; PWP: Prentice, William and Peterson; RDS: RankDeepSurv; WLW: Wei-Lin-Weissfeld.



Appendix 1. PubMed search strategy

Table A1. Search strategies in PubMed database.

Concept	Research strategy keyword	Research	# Results
Survival analysis	'survival analysis'[MeSH Terms] OR 'survival analysis'[Text Word] OR 'time-to-event'[All Fields]	# 1	338,066
Recurrence	('recurren*'[All Fields] OR ('relapse'[All Fields] OR ('repeated'[All Fields] OR ('multiple'[All Fields] AND ('event'[All Fields])))))	# 2	930,483
High-dimension	'high dimension*'[All Fields]	#3	8435
Machine learning	'machine learning'[MeSH Terms] OR 'machine learning'[Text Word]	#4	69,520
Survival analysis for recurrent events	#1 and #2	#5	75,826
High-dimension or machine learning	#3 or #4	#6	76,554
Total	#5 and #6	#7	192

Appendix 2. Data components for modeling recurrent events when using standard statistical approaches

A2.1. Set of individuals at risk

Standard statistical models described do not encounter for individuals at risk in the same way. This induces prior data management for appropriate application.

The set of individuals at risk for the k th event comprised individuals who were at risk for the event. Different definition existed for the set of individuals at risk, mainly based on baseline hazard function:

- The unrestricted set, in which each subject could be at risk for any event regardless of the number of events presented, at all-time intervals;
- The restricted set contained only the time intervals for the k th event of subjects who had already presented $k - 1$ events;
- The semi-restricted set contained for the k th event the subjects who had $k - 1$ or fewer events.

A2.2. Timescales

Timescales also embody key components to address at the data management stage. Three common timescales are:

- Calendar time, in which the times denotes the time since randomization/beginning of the study until an event occurs;
- Gap time, or waiting scale, resets the time to zero when an event occurs, i.e. it corresponds to the time elapsed since the last event previously observed;
- Counting process is constructed as per calendar time, although it enables late inclusions and/or censoring.

Illustrations for timescales were provided in Figure A1.

II.4 Discussion

Le Tableau II.2 présente un récapitulatif des principales méthodes d'apprentissage. Quel que soit le problème étudié (classification ou survie), ces modèles d'apprentissage tirent parti de leurs avantages inhérents et présentent les mêmes limites par conception.

Choix du modèle Le choix de l'algorithme de ML le plus approprié pour les analyses de survie doit être basé sur la question de recherche ainsi que sur les caractéristiques des données, par exemple la taille de l'échantillon, le nombre de variables disponibles et l'équilibre du critère de jugement [Wang et al., 2019]. Si la taille de la population n'est pas suffisante, le recours à des modèles complexes peut poser un problème de sur-apprentissage. À l'inverse, pour des données de grande dimension et des suspicions de multicolinéarité, les SSVM ou les RSF peuvent être avantageux. Par ailleurs, la simplicité du modèle est toujours à privilégier selon le Rasoir d'Ockham, aussi appelé principe de parcimonie et décrit dans Wang et al. [2016] :

"Les multiples ne doivent pas être utilisés sans nécessité" – Guillaume d'Ockham (XIVe siècle)

ML vs. CPH Les études ne comparent pas systématiquement leurs résultats avec le modèle de référence CPH. Cependant, lorsque cette comparaison est effectuée, les algorithmes d'apprentissage automatique tendent à montrer une meilleure performance. Selon Huang et al. [2023], 94% des études qui ont effectué cette comparaison ont trouvé que les algorithmes de ML surpassaient le modèle CPH, indépendamment des pathologies étudiées. De plus, Moncada-Torres et al. [2021] ont démontré que les méthodes de ML peuvent non seulement surpasser les prédictions de CPH en termes de performance, mais aussi fournir des informations interprétables concernant la survie des patientes atteintes de cancer du sein.

Choix de la métrique L'évaluation de la performance prédictive des modèles de survie est complexe en raison de l'absence de recommandations concernant le meilleur critère à utiliser. Le C-index est utile lorsqu'il est nécessaire de comparer des patients, tels que la priorisation des patients pour les greffes de foie (un patient présentant le risque de décès le plus élevé devrait être traité en premier) [Qi et al., 2023]. Le score de Brier est utilisé pour mesurer à la fois la calibration et la discrimination [DeGroot and Fienberg, 1983]. Pour évaluer les fonctions de survie et de risque prédictives, des mesures ponctuelles et intégrales semblent plus adaptées. Les mesures ponctuelles utilisent une seule valeur de la prévision et réduisent la tâche de survie à un problème de régression, en tenant compte de l'indicateur de censure. Les mesures intégrales comparent les fonctions prédictives et théoriques pour tous les points dans le temps. La MAE semble être une mesure plus intuitive pour évaluer les modèles de prédiction de la survie, car elle mesure la différence attendue entre les temps prédictifs et les temps réels des événements. Cependant, les versions proposées des MAE en survie produisent souvent des valeurs biaisées en cas de censure élevée [Haider et al., 2020].

Aussi, nous notons que MAE peut uniquement être utilisé pour l'évaluation des modèles de survie qui peuvent fournir le temps de l'événement comme valeur cible prédite. Des recommandations sont en attente pour mesurer proprement les prédictions de survie et c'est un champ très actif de recherche [Sylvain et al., 2021, Qi et al., 2023].

Apprentissage pour les événements récurrents Malgré les avancées significatives de l'apprentissage automatique dans l'analyse des données censurées, son application aux événements récurrents reste encore relativement sous-développée. C'est ce que nous allons voir au prochain chapitre.

TABLE II.2 – Apprentissage automatique : récapitulatif des méthodes

Modèle	Concept	Avantage	Inconvénient
Régressions pénalisées (LASSO, Ridge, Elastic-Net)	Modèles linéaires avec ajout de termes dans la fonction de perte pour éviter le sur-apprentissage	<ul style="list-style-type: none"> Prévient l'apprentissage; Réduit la complexité du modèle; Interprétabilité usuelle des coefficients. 	<ul style="list-style-type: none"> Peut conduire à une sous-évaluation si le paramètre de régularisation n'est pas correctement défini; Ne gère pas les données manquantes.
Méthodes d'ensembles	Basées sur l'entraînement de <i>weak learners</i> et agrège les prédictions	<ul style="list-style-type: none"> Capable de traiter des données de grande taille; Gestion automatique des interactions entre les variables explicatives; Résiste aux valeurs aberrantes; 	<ul style="list-style-type: none"> Peut souffrir de temps de computation longs suivant les données et hyperparamètres; Peut souffrir de surapprentissage avec des <i>weak learners</i> de type arbres.
Machines à vecteurs de support	Trouver l'hyperplan dans l'espace des caractéristiques qui maximise la marge entre les classes	<ul style="list-style-type: none"> Gestion des données manquantes. 	<ul style="list-style-type: none"> Efficace dans les espaces de haute dimension; Peu intuitive; Nécessite un choix minutieux des hyperparamètres; Nécessite des méthodes d'interprétabilité plus poussées.

Messages-clés de ce chapitre

Ce chapitre est dédié à l'étude de l'**apprentissage automatique** appliqué à l'analyse de données censurées. Dans un premier temps, les concepts fondamentaux de l'apprentissage automatique ont été abordés, incluant les **différents types d'apprentissage** (supervisé, non supervisé, semi-supervisé, par renforcement), ainsi que les notions de **sur-apprentissage** et **sous-apprentissage**. L'**évaluation des modèles** est également traitée en détail, avec une explication des méthodes de **validation croisée** et de validation par réserve, qui sont essentielles pour **optimiser** les hyperparamètres et **sélectionner les meilleurs modèles**. Ensuite, les régressions **pénalisées**, les méthodes de **machine learning** et les **métriques** de performance spécifiques à l'analyse de survie *classique* (temps jusqu'au premier événement) ont été présentées. La contribution centrale de ce chapitre repose sur l'article intitulé "*Towards filling the gaps around recurrent events in high dimensional framework : a systematic literature review and application*". Cet article présente une revue systématique de la littérature sur les **méthodes d'apprentissage automatique pour traiter les événements récurrents** et leur application sur des données simulées. Il met en lumière le **manque de recommandations** et d'approches adéquates pour traiter les événements récurrents dans le cadre de l'apprentissage automatique. Ainsi, ce chapitre fournit une vue d'ensemble complète des techniques d'apprentissage automatique pour l'analyse de données censurées, avec un focus particulier sur les défis et les lacunes dans le traitement des événements récurrents. Le prochain chapitre répond ainsi à ces enjeux en proposant une nouvelle méthode qui tire parti des méthodes d'ensemble et de l'analyse non- et semi-paramétrique des événements récurrents.

Chapitre III

Forêts aléatoires de survie pour l'analyse des événements récurrents

Priest : If men don't trust each other,
this earth might as well be hell.

Commoner : Right. The world's a
kind of hell.

Priest : No! I don't want to believe
that!

Commoner : No one will hear you, no
matter how loud you shout. Just think.
Which one of these stories do you
believe?

Woodcutter : None makes any sense.

Commoner : Don't worry about it. It
isn't as if men were reasonable.

Akira Kurosawa, Rashōmon (1950)

Sommaire

III.1 Arbres et forêts de survie	89
III.1.1 Des arbres de décision aux arbres de survie	89
III.1.2 Spécificités des forêts aléatoires	90
III.1.2.1 L'échantillon <i>out-of-bag</i>	90
III.1.2.2 Importance des variables	91
III.1.2.3 Biais d'induction des forêts aléatoires	91
III.1.3 Actualités des forêts aléatoires de survie	92
III.1.3.1 Extensions récentes	92
III.1.3.2 Sélection de variables et forêts aléatoires de survie . . .	93

III.2 Développement des forêts aléatoires de survie pour les événements récurrents	93
III.3 Application de RecForest aux données du PMSI	124
III.3.1 Contexte	124
III.3.2 Analyse des réadmissions postopératoires auprès des patients atteints de cancer digestif	124
III.3.2.1 Bref contexte médical	124
III.3.2.2 Objectifs	125
III.3.2.3 Méthodologie	125
III.3.2.4 Résultats	126
III.4 Discussion	130
III.4.1 Une taxonomie pour traiter les événements récurrents en survie	130
III.4.2 Axes de développement	130

Introduction

L'analyse des événements récurrents représente un défi majeur dans le domaine des statistiques et de l'apprentissage automatique, en particulier dans des applications médicales comme l'oncologie [Osmani et al., 2018, Charles-Nelson et al., 2019, Galaznik et al., 2022]. Contrairement aux événements uniques, les événements récurrents impliquent des occurrences multiples d'un événement au fil du temps, nécessitant des méthodes analytiques sophistiquées pour capturer leur complexité (Chapitre I). Par ailleurs, l'apprentissage peut se décliner pour des problèmes de classification et de régression, mais aussi de survie (Chapitre II). Ainsi nous avons mis en évidence le manque de méthodes d'apprentissage développées pour traiter les événements récurrents.

Les forêts aléatoires, introduites par Breiman [2001b], sont une méthode d'ensemble basée sur l'agrégation de multiples arbres de décision pour améliorer la robustesse et la précision des prédictions. Lorsqu'elles sont adaptées à l'analyse de survie *classique* (temps jusqu'au premier événement) par Ishwaran et al. [2008], ces forêts permettent de modéliser efficacement le temps jusqu'à la survenue d'un événement, tout en tenant compte des données censurées [Bohannan et al., 2022, Wang et al., 2023]. Pour les événements récurrents, cette approche doit être modifiée pour capturer l'ensemble des événements observés, offrant une vision plus complète du parcours des patients. L'article de la Section III.2 intitulé "*Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event*" présente une extension de ces forêts aléatoires de survie, nommées RecForest, et a été soumis à *BMC Medical Research Methodology*.

Ainsi, ce chapitre est dédié au développement des forêts aléatoires de survie appliquées aux événements récurrents. Nous commencerons par une présentation des concepts fondamentaux des arbres et forêts de survie, en détaillant leur fonctionnement et leurs avantages par rapport aux méthodes traditionnelles (Section III.1). Ensuite, nous décrirons le développement de l'extension des forêts aléatoires de survie RecForest pour traiter les

événements récurrents. Une illustration de cette méthode sera fournie à travers une application concrète en oncologie en Section III.3. Enfin, nous proposerons une taxonomie à jour des méthodes de survie pour traiter les événements récurrents et discuterons des axes de développement futurs (Section III.4).

III.1 Arbres et forêts de survie

III.1.1 Des arbres de décision aux arbres de survie

Les arbres de survie constituent une forme d'arbres de classification et de régression adaptés pour gérer les données censurées. L'intuition derrière les modèles d'arbres est de partitionner récursivement les données en fonction d'un critère de division particulier, de sorte que les objets similaires entre eux, en fonction de l'événement d'intérêt, soient placés dans le même nœud. La première tentative d'utilisation d'une structure d'arbre pour les données de survie a été réalisée par [Ciampi et al. \[1981\]](#). Cependant, [Gordon and Olshen \[1985\]](#) ont été les premiers à discuter de la création d'arbres de survie.

Critères de division La principale différence entre un arbre de survie et un arbre de décision standard réside dans le choix du critère de division. La méthode des arbres de décision effectue une partition récursive des données en fixant un seuil pour chaque caractéristique, mais elle ne peut ni considérer les interactions entre les caractéristiques ni les informations censurées dans le modèle [[Safavian and Landgrebe, 1991](#)]. Les critères de division pour les arbres sont basés sur des heuristiques qui maximisent l'hétérogénéité entre les sous-ensembles après division ou qui maximisent l'homogénéité à l'intérieur de ces sous-ensembles.

La première catégorie d'approches minimise la fonction de perte en utilisant le critère d'homogénéité au sein des nœuds. [Gordon and Olshen \[1985\]](#) ont mesuré l'homogénéité et les distances de Hellinger entre les fonctions de distribution estimées en utilisant la métrique de Wasserstein. Une fonction de vraisemblance exponentielle a été employée par [Davis and Anderson \[1989\]](#) pour la partition récursive basée sur la somme des résidus du modèle de [Cox \[1972b\]](#). [LeBlanc and Crowley \[1992\]](#) ont mesuré la déviance des nœuds en se basant sur la première étape d'une procédure d'estimation de vraisemblance complète.

Dans la seconde catégorie de critères de division, [Ciampi et al. \[1986\]](#) ont utilisé des statistiques de test du log-rank pour les mesures d'hétérogénéité entre les nœuds. Plus tard, [Ciampi et al. \[1987\]](#) ont proposé une statistique de ratio de vraisemblance pour mesurer la dissimilarité entre deux nœuds. [Segal \[1988\]](#) a introduit une procédure pour mesurer la dissimilarité entre les nœuds basée sur la classe de statistiques à deux échantillons de Tarone-Ware.

L'amélioration principale d'un arbre de survie par rapport à un arbre de décision standard réside dans sa capacité à gérer les données censurées en utilisant la structure de l'arbre. Un autre aspect important de la construction d'un arbre de survie est la sélection de l'arbre

final. Des procédures telles que la sélection arrière ou la sélection avant peuvent être suivies pour choisir l'arbre optimal [Bou-Hamad et al., 2011]. Cependant, un ensemble d'arbres peut éviter le problème de la sélection de l'arbre final et offrir de meilleures performances par rapport à un arbre unique.

Algorithm 1 Forêts aléatoires

```

1: Définir le nombre d'arbres  $B$ 
2: Définir le nombre de caractéristiques à considérer  $mtry$ 
3: Définir le nombre minimal d'échantillons pour diviser un nœud  $minsplit$ 
4: Définir le nombre minimal d'échantillons dans un nœud feuille  $nodesize$ 
5: Initialiser la forêt aléatoire comme un ensemble vide
6: for  $b = 1$  à  $B$  do
7:   Tirer un échantillon bootstrap  $\mathcal{D}_b$  des données d'apprentissage
8:   Initialiser l'arbre  $b$  sans aucune division
9:   while le critère d'arrêt n'est pas atteint do
10:    Sélectionner aléatoirement  $mtry$  caractéristiques
11:    Trouver la meilleure division en maximisant le critère de qualité (par ex. Gini,
        Entropie)
12:    Diviser le nœud en deux sous-nœuds
13:   end while
14:   Ajouter l'arbre  $b$  à la forêt aléatoire
15: end for
16: Fin de l'algorithme

```

Les forêts aléatoires de survie de Ishwaran et al. [2008] sont une extension de l'algorithme des forêts aléatoires traditionnelles (Algorithm 1). Contrairement aux modèles de survie traditionnels qui peuvent supposer une forme spécifique pour la fonction de survie, les forêts aléatoires de survie (*Random Survival Forest*, RSF) ne font pas de telles hypothèses, ce qui les rend flexibles et adaptées à diverses distributions de données. Tout comme les forêts aléatoires, les RSF peuvent traiter un grand nombre de variables prédictives, ce qui est courant dans les données médicales modernes, y compris les données génomiques et protéomiques.

III.1.2 Spécificités des forêts aléatoires

III.1.2.1 L'échantillon *out-of-bag*

Une forêt aléatoire agrège B arbres construits chacun à partir d'un échantillon bootstrap différent des données originales. Les individus qui ne sont pas sélectionnés dans l'échantillon bootstrap pour un arbre sont dits *out-of-bag* (OOB). Ces échantillons OOB permettent d'estimer l'erreur de prédiction du modèle sans avoir besoin d'un ensemble de validation distinct. Pour chaque instance, la moyenne de ses prédictions OOB est calculée pour tous les arbres pour lesquels l'instance est OOB.

III.1.2.2 Importance des variables

Les forêts aléatoires fournissent une mesure non paramétrique de l'importance des variables (*variable importance*, VImp). L'importance d'une variable est déterminée en comparant le taux d'erreur lorsque la variable est perturbée par permutation aléatoire par rapport au taux d'erreur initial [Breiman, 2001b]. La variable Z est perturbée lorsque ses valeurs sont permutées aléatoirement dans un arbre, puis les données ainsi perturbées sont introduites dans l'arbre pour calculer l'erreur du prédicteur résultant. La particularité des forêts aléatoires consiste à utiliser l'estimation OOB plutôt que la validation croisée, souvent coûteuse en termes de calcul. On utilise alors l'erreur OOB pour mesurer l'importance des variables :

$$VImp(Z) = \frac{1}{B} \sum_{b=1}^B (\widehat{err}_{OOB}^Z - err_{OOB}), \quad (\text{III.1})$$

avec err_{OOB} l'erreur OOB associée à l'arbre b , et \widehat{err}_{OOB}^Z est l'erreur OOB associée à l'arbre b après permutation de Z .

Des valeurs positives et élevées de la VImp indiquent une capacité prédictive élevée de la variable, tandis que des valeurs nulles ou négatives suggèrent que la variable pourrait être assimilée à du bruit.

III.1.2.3 Biais d'induction des forêts aléatoires

Le biais d'induction est un concept fondamental en apprentissage automatique, représentant les hypothèses implicites associées au choix d'un modèle particulier pour permettre l'apprentissage [Utgoff, 2012]. Sans ces hypothèses inductives, il serait impossible de prédire quoi que ce soit [Mitchell, 1980].

"[An inductive bias is] any basis for choosing one generalization over another, other than strict consistency with the observed training instances" – Tom Mitchell (1980)

Les biais d'induction ont des conséquences importantes sur la robustesse, l'interprétabilité et l'alignement d'un modèle avec le domaine d'application [Domingos, 2012]. En particulier, les forêts aléatoires héritent de nombreux biais venant des arbres à partir desquels elles sont construites. La sélection aléatoire des variables à chaque nouvelle division (à chaque noeud) a des conséquences non négligeables pour le modèle :

- L'échantillonnage des variables augmente la dépendance du modèle vis-à-vis des variables redondantes. Si une variable bruitée Z_2 est une copie d'une variable importante Z_1 , la sélection aléatoire peut amener l'arbre à se baser sur Z_2 en l'absence de Z_1 , ce qui rend le modèle moins parcimonieux.
- L'échantillonnage affecte l'interprétation de l'importance des variables. En effet, puisque seules quelques variables sont échantillonées à chaque division, l'évaluation de leur importance tend à se répartir plus équitablement entre elles, même si certaines variables sont intrinsèquement plus informatives. C'est pourquoi la plupart du temps la

valeur du nombre de variables échantillonnées est égale à la racine du nombre total de variables (package `randomForest` en R, ou `scikit-learn` et `scikit-survival` en Python).

- L'échantillonnage et la dichotomisation des variables à chaque nœud impliquent qu'une variable dépassant la plage des valeurs observées n'a pas d'impact sur les prédictions.

Grinsztajn et al. [2022] ont étudié les biais d'induction qui rendent les méthodes basées sur les arbres plus puissantes que les réseaux de neurones, et ont constaté que la capacité à traiter la non-linéarité des données tabulaires constituait un point fort de ces méthodes. Les biais d'induction jouent ainsi un rôle essentiel dans la réflexion sur l'apprentissage automatique dans la mesure où ils permettent de généraliser au-delà des données observées.

"If biases and initial knowledge are at the heart of the ability to generalize beyond observed data, then efforts to study machine learning must focus on the combined use of prior knowledge, biases, and observation in guiding the learning process. It would be wise to make the biases and their use in controlling learning just as explicit as past research has made the observations and their use." – Tom Mitchell (1980)

Note : Ce paragraphe a été écrit et inspiré des réflexions de Christopher Molnar dans ses newsletters intitulées "["Mindful Modeler"](#)".

III.1.3 Actualités des forêts aléatoires de survie

III.1.3.1 Extensions récentes

Depuis l'introduction des RSF, plusieurs variantes et améliorations ont été développées pour répondre à des besoins spécifiques dans l'analyse des données de survie. Voici un aperçu des extensions principales, bien que celles-ci soient axées sur le temps jusqu'au premier (ou prochain) événement.

Pour les **risques compétitifs**, Ishwaran et al. [2014] ont proposé des RSF qui adaptent les règles de division en utilisant les tests de Gray. En outre, plusieurs estimateurs d'ensemble, tels que la fonction d'incidence cumulée (*cumulative incidence function*), ont été définis pour ces modèles. Mogensen and Gerd [2013] ont, quant à eux, proposé de remplacer le statut de l'événement censuré par une pseudo-valeur jackknife basée sur l'estimateur marginal d'Aalen-Johansen, comme décrit par Klein and Andersen [2005].

Les **prédictions dynamiques** sont fréquemment utilisées pour intégrer des informations sur les biomarqueurs au cours du temps, permettant de produire des estimations actualisées et plus précises des risques de survenue d'événements [Rizopoulos, 2011]. Pickett et al. [2021] ont introduit les RSF avec une approche *landmark*. Récemment, Devaux et al. [2023a] ont allié prédition dynamique, facteurs longitudinaux et risques compétitifs dans les RSF avec DynForest et son package R associé [Devaux et al., 2023b].

Nous n'avons pas encore abordé les **approches bayésiennes**. Des estimations bayésiennes non paramétriques existent, notamment les BART (*Bayesian additive regression*

trees) de Chipman et al. [2010]. Ces méthodes ont été étendues à la survie et agrégées en forêts par Linero et al. [2022].

III.1.3.2 Sélection de variables et forêts aléatoires de survie

La sélection des variables joue un rôle essentiel dans le traitement des données de (très) haute dimension, notamment avec l'avènement des bases médico-administratives, des données textuelles et d'images, et des données génétiques [Wang and Li, 2017a, Ang et al., 2015]. En plus de son utilisation générale pour la prédiction de la survie, des méthodes basées sur les RSF ont également été développées pour la sélection des variables.

RSF-VH (*variable hunting*), introduit par Ishwaran et al. [2010], est une méthode de régularisation par étapes utilisant la profondeur minimale des sous-arbres pour évaluer l'importance des variables. Cette méthode sélectionne les variables de manière itérative en se basant sur différentes divisions des données.

Aussi, Genuer et al. [2010] ont proposé d'adopter une sélection de variables selon un objectif d'interprétation ou de prédiction. Pour l'interprétation, il convient de construire une série de forêts aléatoires impliquant les k premières variables, pour $k = 1, \dots, m$, et de sélectionner les variables conduisant à l'erreur la plus faible. Pour la prédiction, à partir des variables ordonnées retenues pour l'interprétation, il faut construire une séquence ascendante de forêts aléatoires en utilisant et en testant progressivement les variables. Les variables du dernier modèle sont alors sélectionnées.

Pang et al. [2012] ont proposé un algorithme qui classe les covariables par ordre d'importance, entraîne itérativement des modèles RSF en utilisant les covariables les plus importantes, et détermine l'ensemble optimal en minimisant le taux d'erreur OOB.

Une approche alternative, basée sur une stratégie topologique, a été proposée par Mbo邦ning and Broët [2016]. Cette méthode itère sur la construction d'une forêt de survie de type *bagging* pour générer des p-valeurs pour chaque variable, permettant ainsi une sélection robuste même en présence de nombreuses variables bruitées.

III.2 Développement des forêts aléatoires de survie pour les événements récurrents

Cette section présente RecForest, une extension des forêts aléatoires de survie pour traiter les données de survie avec événements récurrents. C'est au niveau de la règle de division et des estimations des nœuds terminaux que nous avons tiré parti des modèles non- et semi-paramétrique du Chapitre I. Afin de s'adapter au mieux au contexte des études en oncologie, nous avons également introduit la possibilité de prendre en compte les événements terminaux.

La construction des arbres de survie initiaux, c'est-à-dire des premiers nœuds de décision, de RecForest est illustrée en Figure III.1. À partir d'un échantillon de la population,

on peut diviser en deux sous-échantillons à partir de la variable x_1 suivant le seuil a . Cette variable x_1 a été tirée de façon aléatoire parmi les $mtry$ variables. Au seuil a , la variable x_1 maximise la distance entre les deux noeuds enfants créés par $x_1 < a$ et $x_1 > a$ selon le pseudo-score test. Ce test, introduit à la Section I.3 du Chapitre I, est une version étendue du *logrank* pour les événements récurrents. La règle de division, ou *splitting rule*, est basée sur la maximisation de cette statistique de test. En présence d'un événement terminal, un modèle Ghosh-Lin est construit, et la statistique du test de Wald est récupérée pour la maximisation.

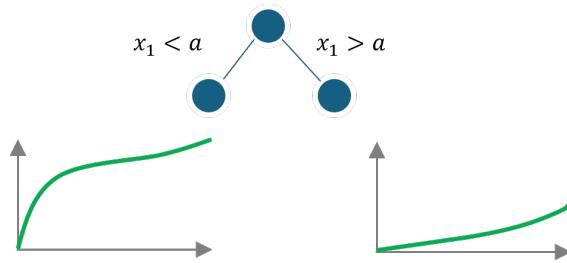


FIGURE III.1 – Construction d'un arbre de survie de RecForest - Nœud initial

De façon itérative, $mtry$ variables sont sélectionnées aléatoirement à chaque nœud jusqu'au respect des règles d'arrêt, qui sont :

- le nœud terminal doit contenir au moins un certain nombre d'individus ;
- le nœud terminal doit contenir au moins un certain nombre d'individus avec au moins un événement.

L'estimation du nombre attendu d'événements récurrents au cours du temps est alors obtenue pour chaque nœud terminal, comme illustré en Figure III.2. Successivement, x_1 , x_2 et x_3 ont maximisé la statistique de test (pseudo-score ou Wald) aux seuils a , b et c , respectivement. Les estimateurs MCF sont ainsi obtenus à chaque nœud terminal pour l'ensemble des individus respectant le chemin de variables et de seuils associés correspondants.

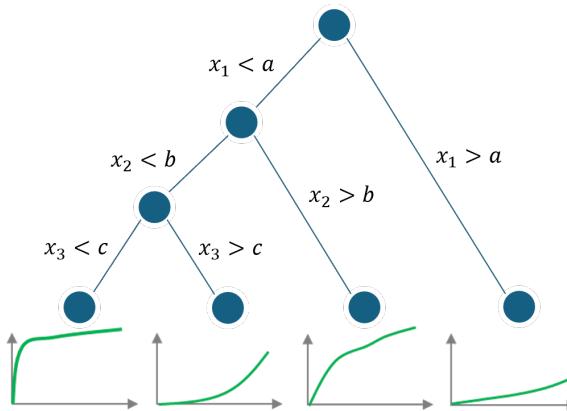


FIGURE III.2 – Construction d'un arbre de survie de RecForest

De façon analogue aux forêts aléatoires, RecForest agrège les estimations de l'ensemble des arbres pour construire l'estimation finale. Chaque individu obtient ainsi le nombre attendu d'événements récurrents au cours du temps.

Comme mis en évidence au Chapitre II, la façon d'évaluer les modèles de survie pour événements récurrents n'est pas claire. Pour évaluer la performance de RecForest, nous avons utilisé deux métriques, spécifiquement étendues aux événements récurrents : un indice de concordance (C-index) et une erreur quadratique moyenne.

Le C-index proposé est une extension du C-index de [Kim et al. \[2018\]](#), qui était basé sur le nombre d'événements récurrents à la fin du suivi. La problématique provient du fait qu'en vie réelle, les individus ont bien souvent des temps de suivi différents. Ainsi, il paraît peu raisonnable de comparer un individu avec 2 événements récurrents après un suivi de 4 mois et un individu avec 2 événements après un suivi de 5 ans. Pour corriger cela, nous avons proposé de prendre en compte un taux d'occurrence d'événements récurrents, soit le nombre d'événements récurrents par unité de temps. Le C-index de [Murris et al. \[2024\]](#) s'écrit alors :

$$\hat{C}_{\text{rec}} = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}(r_i > r_j) \times \mathbb{1}(\hat{r}_i > \hat{r}_j)}{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}(r_i > r_j)}, \quad (\text{III.2})$$

avec $r_i = \frac{N_i(T_i)}{T_i}$ et $\hat{r}_i = \frac{\hat{N}_i(T_i | \mathbf{x}_i)}{T_i}$ les taux d'occurrence des événements observés et prédis, respectivement. Tel que les autres, la valeur dudit indice se situe entre 0 et 1, où 1 indique une concordance parfaite et des valeurs proches de 0,5 suggèrent un modèle proche de l'aléatoire.

[Bouaziz \[2024\]](#) a récemment proposé une mesure pour l'erreur quadratique moyenne (*mean squared error*, MSE) pour les événements récurrents. Nous avons intégré cette métrique, et pour chaque arbre b , on a

$$\widehat{MSE}_b(t, \hat{\mu}_b) = \frac{1}{n} \sum_{i=1}^n \left(\int_0^t \frac{dN_i(u)}{\hat{G}_c(u|\mathbf{x})} - \hat{\mu}_b(t|\mathbf{x}) \right)^2, \quad (\text{III.3})$$

avec $\hat{G}_c(u|\mathbf{x}) = 1 - \hat{G}(u - |\mathbf{x})$ est un estimateur de $G_c(u|\mathbf{x}) = 1 - G(u - |\mathbf{x})$ la fonction de distribution de la censure C compte tenu de \mathbf{x} . En l'absence d'événement terminal, \hat{G} est la fonction de distribution cumulative empirique de la variable censurée. En présence d'un événement terminal, \hat{G} est l'estimateur de Kaplan-Meier de C . Pour agréger sur les B arbres de la forêt, on a

$$\widehat{MSE}(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B \widehat{MSE}_b(t, \hat{\mu}_b). \quad (\text{III.4})$$

Comme le souligne [Bouaziz \[2024\]](#), deux modèles différents peuvent conduire à des valeurs MSE similaires, soulignant la difficulté d'évaluer quel modèle est le meilleur. Un score est donc introduit pour représenter le gain de prédiction par rapport à un estimateur de référence et nous définissons pour chaque arbre b :

$$Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0}) = \widehat{MSE}_b(t, \hat{\mu}_{b,0}) - \widehat{MSE}_b(t, \hat{\mu}_b), \quad (\text{III.5})$$

où $\hat{\mu}_b$ est l'estimateur de l'arbre b et $\hat{\mu}_{b,0}$ l'estimateur de référence. Le Score de la forêt aléatoire s'écrit

$$Score(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0}). \quad (\text{III.6})$$

Un score plus élevé est associé à une meilleure performance.

Le MSE et le Score ci-dessus sont des mesures qui dépendent du temps. Comme démontré dans [Bouaziz \[2024\]](#), le MSE ci-dessus se réduit au score de Brier lorsque les individus ne subissent qu'un seul événement. Dans l'esprit de la version intégrée du score de Brier sur une période $[\tau_1, \tau_2]$, nous intégrons le MSE et le Score des équations III.4 et III.6, respectivement. L'IMSE (*integrated mean squared error*) et le IScore (*integrated score*) s'écrivent :

$$\begin{cases} \widehat{IMSE}(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \widehat{MSE}(t, \hat{M}) dt; \\ IScore(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} Score(t, \hat{M}) dt. \end{cases} \quad (\text{III.7})$$

En général $\tau_1 = 0$ et τ_2 est le temps de suivi maximal observé.

L'article soumis au journal *BMC Medical Research Methodology*, intitulé "*Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event*", présente la méthode RecForest. Les messages clés sont les suivants :

- Une étude de simulation a été menée pour comparer RecForest avec l'estimateur non paramétrique et le modèle Ghosh-Lin.
- La flexibilité de RecForest a été éprouvée dans une variété de contextes simulés, y compris dans le traitement de données manquantes et de grande dimension, attestant de sa robustesse et de sa capacité d'adaptation.
- RecForest a démontré de meilleures performances en termes de C-index et MSE décrits ci-dessus.
- Une démonstration pratique sur des données en libre accès a été effectuée à la fois pour montrer l'impact sur données réelles ainsi que l'optimisation des hyperparamètres de RecForest.

Le développement d'un package R dédié à RecForest est actuellement en cours pour faciliter son adoption et son utilisation dans la communauté scientifique.

Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event

Juliette Murris¹ , Olivier Bouaziz² , Michal Jakubczak³ ,
Sandrine Katsahian⁴ , and Audrey Lavenu⁵  *

¹ HeKA, Inria, Inserm, Centre de recherche des Cordeliers, Université Paris Cité, Paris, France, & R&D, Pierre Fabre, Boulogne-Billancourt, France, e-mail:
juliette.murris-ext@aphp.fr

² Université Paris Cité, CNRS, MAP5, F-75006 Paris, France, e-mail:
olivier.bouaziz@parisdescartes.fr

³ R&D, Pierre Fabre, Boulogne-Billancourt, France, & Ardigen S.A., Kraków, Poland,
e-mail: michal.jakubczak.ext@pierre-fabre.com

⁴ HeKA, Inria, Inserm, Centre de recherche des Cordeliers, Université Paris Cité, Paris, France, Unité de Recherche Clinique, Hôpital Européen Georges-Pompidou, Assistance Publique – Hôpitaux de Paris (AP-HP), Centre, Paris, France, & Centre d'Investigation Clinique 1418 Épidémiologie Clinique, Paris, France, e-mail:
sandrine.katsahian@aphp.fr

⁵ Faculté de Médecine, Université de Rennes, Rennes, France, Institut de Recherche Mathématique de Rennes (IRMAR), Rennes, France, & Centre de Investigation Clinique 1414, Inserm, Université de Rennes, Rennes, France, e-mail:
audrey.lavenu@univ-rennes.fr

Abstract: Random survival forests (RSF) have emerged as valuable tools in medical research. They have shown their utility in modelling complex relationships between predictors and survival outcomes, overcoming linearity or low dimensionality assumptions. Nevertheless, RSF have not been adapted to right-censored data with recurrent events (RE). This work introduces RecForest, an extension of RSF and tailored for RE data, leveraging principles from survival analysis and ensemble learning. RecForest adapts the splitting rule to account for RE, with or without a terminal event, by employing the pseudo-score test or the Wald test derived from the marginal Ghosh-Lin model. The ensemble estimate is constructed by aggregating the expected number of events from each tree. Performance metrics involve a concordance index (C-index) tailored for RE analysis, along with an extension of the mean squared error (MSE). A comprehensive evaluation was conducted on both simulated and open-source data. We compared RecForest against the non-parametric mean cumulative function and the Ghosh-Lin model.

*SK and AL share last authorship.

Across the simulations and application, RecForest consistently outperforms, exhibiting C-index values ranging from 0.64 to 0.80 and lowest MSE metrics. As analysing time-to-recurrence data is critical in medical research, the proposed method represents a valuable addition to the analytical toolbox in this domain.

Keywords and phrases: Random forests, Recurrent events, Survival analyses, Terminal events, High-dimensional data.

1. Introduction

Recurrent events refer to instances where individuals may experience multiple occurrences of the same event over time. In medical research, patients may face recurrent disease relapses, frequent hospitalizations, or repeated surgeries. While traditional survival analyses focus solely on the first occurrence of an event, specific statistical models have been developed to capture the complexity of recurrence in a survival framework. Intensity models rely on instantaneous hazards at each time point and account for dependence amongst event occurrences captured by time-varying covariates ([Andersen and Gill \(1982\)](#); [Prentice, Williams and Peterson \(1981\)](#)). Besides, marginal models centre on the overall distribution of event times and the cumulative event counts ([Wei, Lin and Weissfeld \(1989\)](#); [Cook, Lawless and Lee \(2010\)](#)). For a more in-depth exploration of these models concerning recurrent events, comprehensive discussions can be found in works by [Amorim and Cai \(2015\)](#) and [Ozga, Kieser and Rauch \(2018\)](#).

Time-to-event analyses are systematically challenging due to the presence of censoring, i.e. when the precise timing of an event remains unknown or unobserved. Above methodologies strictly assume the censoring process to be uninformative, hence independent of the underlying event process. Nevertheless, a terminal event may occur in competition, preventing further events of interest from happening. A terminal event is then a specific type of event considered as a termination point for the study period, making the censoring process no longer uninformative. Strategies for handling terminal events include ignoring them, although this approach is acknowledged to be flawed, or accounting for competing risks. Several pertinent statistical models enable to analyse both recurrent events and competing risks ([Charles-Nelson, Katsahian and Schramm \(2019\)](#)).

Navigating medical data introduces numerous challenges, including high-dimensionality, variable selection, and multicollinearity. To address these, survival time-to-first-event approaches have integrated statistical and machine learning techniques. In practice, various algorithms now have their

survival counterparts that are effectively employed to answer medical questions in real-world applications ([Huang et al. \(2023\)](#)). For instance, penalized regression methods, such as LASSO (Least Absolute Shrinkage and Selection Operator), Ridge, and Elastic-Net, have been tailored for Cox models, facilitating variable selection and regularization ([Cox \(1972\)](#); [Tibshirani \(1997\)](#)). Support-vector machines introduced by [Van Belle et al. \(2011\)](#), renowned for their capacity to handle high-dimensional data and non-linearity, have also been extended to survival endpoints. Likewise, random survival forests (RSF) from [Ishwaran et al. \(2008\)](#) embody a powerful ensemble learning technique handling interactions. The RSF algorithm has been extended to model several phenomena, such as competing risks, or longitudinal data ([Ishwaran et al. \(2014\)](#); [Devaux et al. \(2023\)](#)). However, within the survival framework, no machine learning approach has hitherto been extended to recurrent events ([Murris et al. \(2023\)](#)). To address these unmet needs and confront the aforementioned challenges, we introduce the first RSF capable of handling recurrent events, with or without a terminal event. Illustrated in Figure 1, our method entails a 5-step approach that i) discerns the relevance of recurrent and terminal events, ii) grows trees to construct a coherent RSF, iii) thoroughly assesses performance, iv) provides relevant variable importance, and v) enables predictions on new data.

In this paper, we consider n individuals. Let $N^*(t)$ be the number of recurrent events that occur in the time interval $[0, t]$, D the survival time and C the censoring time. The data is made of $(N(\cdot), \Upsilon, \delta)$ where $N(t) = N^*(t \wedge C)$, $\Upsilon = D \wedge C$, $\delta = I(D \leq C)$, where $a \wedge b = \min(a, b)$ and $I(\cdot)$ is the indicator function. For $i = 1, \dots, n$, $(N_i(\cdot), \Upsilon_i, \delta_i)$ are assumed to be independent replicates of $(N(\cdot), \Upsilon, \delta)$. The marginal mean frequency function is $\mu(t) = \mathbb{E}[N(t)]$. An estimator of μ in the absence of a terminal event is the Nelson-Aalen estimator from [Lawless and Nadeau \(1995\)](#), that writes

$$(1) \quad \hat{\mu}(t) = \hat{R}(t) = \int_0^t \frac{dN(u)}{Y(u)}$$

with $N(t) = \sum_i N_i(t)$, and $Y(t) = \sum_i Y_i(t)$ the number of individuals at risk at time t . In presence of a terminal event, we have $\mu(t) = \int_0^t S(u)dR(u)$ where $S(t) = \mathbb{P}(D \geq t)$ and $dR(t) = \mathbb{E}[dN^*(t)|D \geq t]$ ([Cook and Lawless \(1997\)](#); [Ghosh and Lin \(2000\)](#)). The associated estimator writes

$$(2) \quad \hat{\mu}(t) = \int_0^t \hat{S}(u)d\hat{R}(u) = \int_0^t \hat{S}(u) \frac{\sum_i Y_i(u)dN_i(u)}{\sum_i Y_i(u)}$$

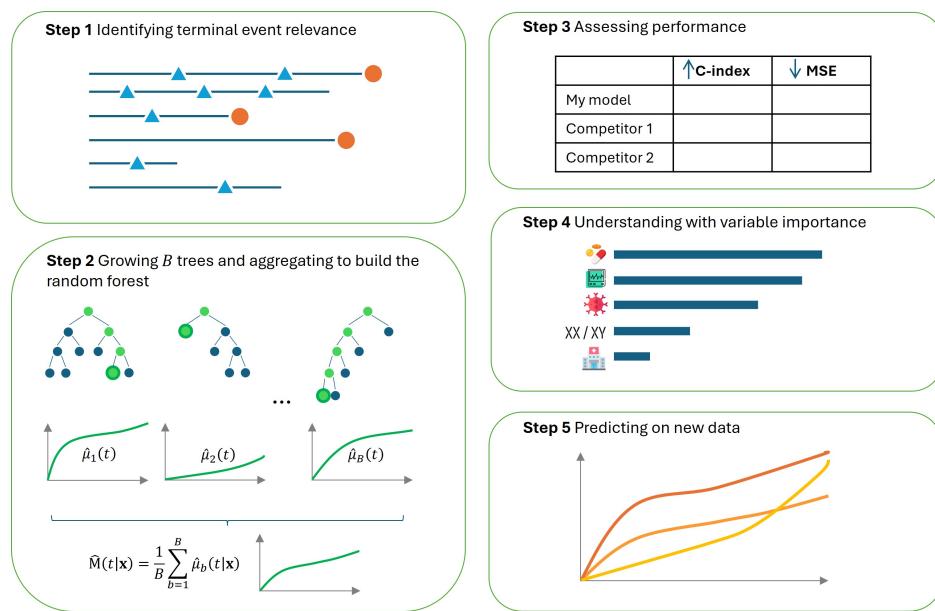


FIG 1. Scheme of the use of RecForest for survival data with recurrent events in presence or absence of a terminal event

And $\hat{S}(t)$ is the Kaplan-Meier estimator of $S(t)$ based on (Υ_i, δ_i) ([Kaplan and Meier \(1958\)](#)).

From the above considerations arises the evaluation of the provided estimations. Within survival framework, a widely common metric is an extension of the area under the ROC curve known as the concordance index (C-index). The principle of the C-index from [Harrell, Lee and Mark \(1996\)](#) and its derivatives as per [Uno et al. \(2011\)](#) is to measure the ability of a model to correctly order pairs of survival times. Recent developments from [Kim, Schaubel and McCullough \(2018\)](#) have expanded the application of the C-index to the recurrent event framework, incorporating the number of subsequent event occurrences. However, the number of events over time is only comparable if individuals have similar follow-up, which is hardly the case in real-world settings. Therefore, we proposed a generalized C-index by introducing event occurrence rate. Additionally, we employ the mean-square error, recently adapted to account for recurrent events by [Bouaziz \(2024\)](#).

Based on the non-parametric estimators and ensemble method principles, the objective of this work is to introduce a new ensemble approach, called RecForest, for the analysis of recurrent events in a survival framework, with or without a terminal event. The overall methodology based on survival decision trees and novel associated evaluation metrics is described in Section 2. Section 3 displays an extended simulation scheme for the comprehension of the proposed methodology. Illustrative examples based on open-source data are used for concrete application in Section 4.

2. Methodology

The proposed Algorithm 1 is an extension of the RSF introduced by [Ishwaran et al. \(2008\)](#). The first step is drawing bootstrap samples to prevent overfitting and capture inherent variability within the original dataset. Then, survival trees are constructed on each bootstrap sample. Unlike the original RSF, our approach accommodates for subsequent events by integrating statistical considerations tailored for recurrent events analysis. As a last step, the algorithm aggregates the results over the constructed recursive survival trees to obtain a comprehensive estimate.

Next subsections describe in further details how survival trees grow for constructing the random forest. Additionally, we provide adequate metrics for the evaluation. Finally, we expound on the computation of variable importance.

Algorithm 1 Overview of RecForest algorithm

Require: Draw $B > 0$ bootstrap samples from the learning data

```

for Each node of survival tree  $b$  do
     $mtry$  predictors are randomly selected with  $mtry \in \mathbb{N}$ ,  $mtry \leq p$ ;
    A greedy algorithm for optimal threshold research is used to maximize the test
    statistic;
    The tree grows until the stopping rule is met based on the minimal number of events
     $minsplit$  and the minimal number of individuals in terminal nodes  $nodesize$ ;
    Estimate  $\hat{\mu}_b$  is computed;
end for
Estimate  $M$  is computed over the  $B$  trees.
```

2.1. Growing trees with recurrent events

2.1.1. Splitting rules

At each node $h \in \mathcal{H}$, the ongoing subsample is split into two daughter nodes denoted $h^{(+)}$ and $h^{(-)}$. The aim of the split is to make the daughter nodes as different as possible with regards to the outcome. The splitting rule requires that each of the $mtry$ randomly drawn variable is dichotomized. For continuous variables, random split points, quartiles, and deciles are considered. Let $\mathbf{x}_h = \{A, B\}$ be the dichotomized vector of a variable inherited from h . With no terminal event, we compare the marginal mean frequency functions $\mu_A(t)$ and $\mu_B(t)$. The null hypothesis is their equality. In absence of a terminal event, we use the two-sample test akin to the logrank test from [Lawless and Nadeau \(1995\)](#). The test statistic writes $U(t) = \int_0^t \frac{Y_A(u)Y_B(u)}{Y_A(u)+Y_B(u)} (d\hat{\mu}_A(u) - d\hat{\mu}_B(u))$. To incorporate the presence of a terminal event, we employ the marginal Gosh-Lin (GL) model from [Ghosh and Lin \(2002\)](#) within the single variable \mathbf{x}_h . Acknowledging there are no further recurrence after the terminal event, the marginal mean up to t associated with \mathbf{x}_h is defined as $\mu_{\mathbf{x}_h}(t) = \mathbb{E}[N^*(t)|\mathbf{x}_h] = \mu_0(t) \times \exp(\beta \mathbf{x}_h)$ with μ_0 left unspecified and β the regression coefficient. To accommodate longitudinal variables, the GL model considers a rate function $d\mu_{\mathbf{x}_h}(t) = d\mu_0(t) \times \exp(\beta \mathbf{x}_h(t))$. The Wald test statistic is then extracted from $\mu_{\mathbf{x}_h}$ and $d\mu_{\mathbf{x}_h}$ to test the null hypothesis of $\beta = 0$.

The variable selected for node h is the one that maximizes the adequate test statistic to generate $h^{(+)}$ and $h^{(-)}$, based on the presence of a terminal event and/or longitudinal variables.

2.1.2. Terminal node estimator

Let b be a bootstrap sample from original data on which a tree is grown and \mathbf{x} a p -dimensional vector of covariates dropped down the tree. The node-specific event count $N_b(t|\mathbf{x})$ is the number of recurrent events before censoring or a terminal event at time t . The associated number of individuals at risk $Y_b(t|\mathbf{x})$ is the number of individuals that were not censored, or that did not encounter a terminal event by time t . We then define a tree-specific estimate as follows

$$(3) \quad \hat{\mu}_b(t|\mathbf{x}) = \hat{R}_b(t|\mathbf{x}) = \int_0^t \frac{N_b(du|\mathbf{x})}{Y_b(du|\mathbf{x})}$$

In case of the presence of a terminal event,

$$(4) \quad \hat{\mu}_b(t|\mathbf{x}) = \int_0^t \hat{S}_b(u|\mathbf{x}) d\hat{R}_b(u|\mathbf{x})$$

Individuals from the same terminal node share similar features inherited from their tree path, along with identical estimates. As per the splitting rule, the terminal node estimator depends on the presence of a terminal event in the original sample.

2.1.3. Pruning trees

A pruning strategy is essential to help find a trade-off to prevent overfitting and improve generalization performance of trees, within a reasonable computational time. Aligned with Devaux et al. (2023), we suggest two stopping rules for each terminal node: (i) a minimal number of events called *minsplit*, and (ii) a minimal number of individuals called *nodelsize*. The validation of either stopping rule designates the current node h as terminal.

2.1.4. Handling missing data

To tackle eventual missing data, we include an adaptive-tree imputation which addresses missing data during the tree-growing stage by selectively drawing from available, non-missing, in-bag data (Ishwaran et al. (2008); Chen and Xu (2023)). At each node h_b from tree b , the method entails imputing random non-missing information specifically from the selected variables. The imputed data is then utilized for making splits within the node h_b . Imputed values are reset to missing as the tree progresses to subsequent nodes.

2.2. From trees to random forests

2.2.1. Ensemble estimates

Once all B trees are grown from the independent bootstrap samples, the ensemble estimate \hat{M} is the aggregation of all B tree-specific estimates. We define \hat{M} as

$$(5) \quad \hat{M}(t|\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_b(t|\mathbf{x})$$

2.2.2. OOB ensemble estimates

By standard bootstrap theory, each bootstrap sample leaves out circa 37% of the data ([Ishwaran et al. \(2008\)](#)). This is the so-called out-of-bag (OOB) sample. OOB data is used to build OOB ensembles. Let $\mathcal{O}_i \subseteq \{1, \dots, B\}$ be the index set of trees where for $b \in \mathcal{O}_i$, $c_{i,b} = 0$, which means the individual i is in the OOB sample. The OOB ensemble estimate \hat{M}^{OOB} of aggregated tree-specific estimates for i which is OOB writes

$$(6) \quad \hat{M}^{OOB}(t|\mathbf{x}_i) = \frac{1}{|\mathcal{O}_i|} \sum_{b \in \mathcal{O}_i} \hat{\mu}_b(t|\mathbf{x}_i)$$

OOB ensemble estimates are typically used for reporting errors.

2.3. Performance

Performance metrics below indicate the ability of the model to predict well from training data to unseen data. In our case, unseen data are either from the OOB sample or external validation data.

2.3.1. Assessing performance with relevant metrics

For the assessment of performance, we introduce an extended version of the C-index and employ the mean-square error (MSE), a derived score, and their integrated versions ([Figure 2](#)).

Concordance index. [Kim, Schaubel and McCullough \(2018\)](#) adapted the C-index to recurrent events and considered the number of events over time across individuals. This metric hence suffers from the potential bias in case of

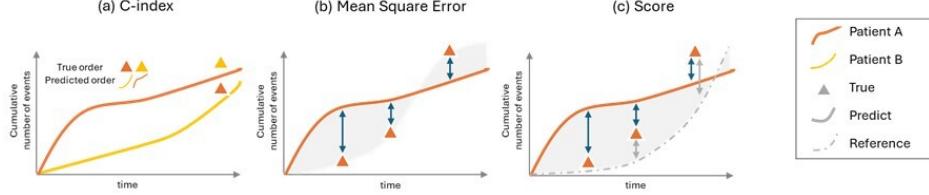


FIG 2. Illustration of the performance metrics with true and predicted cumulative number of events over time

substantial variability in the follow-up times. Individuals with longer follow-up times or a higher number of events might indeed disproportionately influence the C-index calculation. To address this issue, we suggest using occurrence rates by computing event rates per unit time.

The proposed C-index is defined as the proportion of all concordant pairs of individuals where predicted occurrence rates are correctly ordered with respect to observed occurrence rates (as shown in Figure 2a). As occurrence rates can be calculated for all individuals, including censored ones, the proposed C-index is not partial and considers all individuals in the computation. In this work, the C-index then writes

$$(7) \quad \hat{C}_{\text{rec}} = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{r_i > r_j} \times \mathbb{1}_{\hat{r}_i > \hat{r}_j}}{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{r_i > r_j}}$$

with $r_i = \frac{N_i(T_i)}{T_i}$ and $\hat{r}_i = \frac{\hat{\mu}(T_i | \mathbf{x}_i)}{T_i}$ the observed and predicted event occurrence rates, respectively. Like other C-indices, the value of the above C-index falls within the range of 0 to 1, where 1 indicates perfect concordance, and values close to 0.5 suggest randomness in the model.

Mean-squared error and derived score. No MSE measure has been adapted to recurrent events framework until very lately. Bouaziz (2024) filled this gap and suggested a generalization of the Brier score from Graf et al. (1999). For our problematic, for each tree b , we define

$$(8) \quad \widehat{MSE}_b(t, \hat{\mu}_b) = \frac{1}{n} \sum_{i=1}^n \left(\int_0^t \frac{dN_i(u)}{\hat{G}_c(u | \mathbf{x})} - \hat{\mu}_b(t | \mathbf{x}) \right)^2$$

Where $\hat{G}_c(u | \mathbf{x}) = 1 - \hat{G}(u - | \mathbf{x})$ is an estimator of $G_c(u | \mathbf{x}) = 1 - G(u - | \mathbf{x})$ the conditional cumulative distribution function of the censoring variable C

given \mathbf{x} . We assume C and \mathbf{x} to be independent. With no terminal event, \hat{G} is the empirical cumulative distribution function of the censored variable. In the presence of a terminal event, \hat{G} is the Kaplan-Meier estimator of C . As suggested in Figure 2b), the general prediction criterion denoted \widehat{MSE} over our random forest hence writes

$$(9) \quad \widehat{MSE}(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B \widehat{MSE}_b(t, \hat{\mu}_b)$$

As pointed out in [Bouaziz \(2024\)](#), two different models may lead to similar MSE values over time underlining the difficulty in assessing which model is better. A score is thus introduced to represent the prediction gain compared to a reference estimator and we define for each tree b:

$$(10) \quad Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0}) = \widehat{MSE}_b(t, \hat{\mu}_{b,0}) - \widehat{MSE}_b(t, \hat{\mu}_b)$$

Where $\hat{\mu}_b$ is the evaluated estimator and $\hat{\mu}_{b,0}$ the reference estimator over the b samples. In our case, the reference estimator is the tree-specific non-parametric either the Nelson-Aalen or the Ghosh-Lin estimator described above. The ensemble score illustrated in Figure 2c) writes

$$(11) \quad Score(t, \hat{M}) = \frac{1}{B} \sum_{b=1}^B Score_b(t, \hat{\mu}_b, \hat{\mu}_{b,0})$$

A higher score is associated with a better performance.

Integrated counterparts. Above MSE and derived score are time-dependent metrics. While they provide valuable insight of the performance for each time t , there is a need for the estimation of the expectation of single-time MSE and derived score over time (shaded areas in Figure 2). As demonstrated in [Bouaziz \(2024\)](#), above MSE reduces to the Brier score when individuals experience one event at most. In the spirit of the integrated version of the Brier score between two time points τ_1 and τ_2 , we integrate the MSE and the score:

$$(12) \quad \begin{cases} \widehat{IMSE}(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} \widehat{MSE}(t, \hat{M}) dt \\ IScore(\tau_1, \tau_2, \hat{M}) &= \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} Score(t, \hat{M}) dt \end{cases}$$

With $\tau_1 = 0$ and τ_2 the maximum event time on the original sample.

2.3.2. OOB errors

OOB errors are used for tuning hyperparameters and evaluating predictive performances and are computed on OOB samples. They are also particularly useful in the absence of external validation data or when dealing with low-dimensional original samples, where allocating a portion for validation is hardly affordable. OOB predictions are calculated by average predictions from OOB trees, and the error rate is complementary to 1. In this work, we consider the IMSE to assess the OOB error:

$$(13) \quad \text{OOB error} = \widehat{\text{IMSE}}^{OOB}(t, \hat{M}^{OOB})$$

In this way, models exhibiting lower OOB errors are consistently favored. Of note, computing OOB errors is not recommended when the number of trees is low as each one of them may underfit.

2.4. Variable importance

The importance of a variable ($VImp$) is evaluated by permutation, corresponding to the impact of random perturbations in the sample on the OOB error (Breiman (2001)). To quantify the $VImp$ of a covariate, a performance metric, as previously defined, is calculated following the permutation of values associated with this covariate. The $VImp$ is determined as the difference between the original and permuted performance metrics. For covariate j and considering K permutations, $VImp(j)$ writes

$$(14) \quad VImp(j) = \frac{1}{K} \sum_{k=1}^K (\hat{\theta} - \hat{\theta}_k^j)$$

With $\hat{\theta} = \{-\widehat{\text{IMSE}}, \hat{C}\}$ the original performance metric and $\hat{\theta}_k^j$ the permuted performance metric. High relative values of $VImp$ indicate a loss of performance and lower/null values are interpreted as no performance for such covariates.

3. Simulation study

We propose the following simulation settings to illustrate the use of RecForest, inspired by Ishwaran et al. (2014); Bouaziz (2024). Simulation scenarios will cover multiple cases with associated covariates, with or without a terminal event, low- and high-dimensional data, and with or without missing

TABLE 1
Summary of investigated scenarios

	Without a terminal event		With a terminal event	
	Low dimensional $\{n = 250, p = 20\}$	High dimensional $\{n = 250, p = 300\}$	Low dimensional $\{n = 250, p = 20\}$	High dimensional $\{n = 250, p = 300\}$
Complete	x	x	x	x
Missing	x	x	x	x
Random	x	x	x	x

data. 250 learning sets and one external validation set were generated for each scenario with $n = 250$ individuals and p covariates. Table 1 below summarizes investigated scenarios. Next subsections further detail simulation parameters for each case. For each scenario, we grow 100 trees for **RecForest**. We set the following values: the minimal number of events is $minsplit = 5$, the minimal number of individuals is $nodesize = 10$, and the number of random predictors at each node is $mtry = \{1, \sqrt{p}, \log(p)\}$. We compared **RecForest** with non-parametric estimators, as well as a semi-parametric GL model where possible. Performances were measured on the external validation set using the C-index, MSE, the score and their integrated versions.

3.1. Simulation scheme

3.1.1. With and without a terminal event

For $i = 1, \dots, n$, p_0 -dimensional covariate vector, $X_i = (X_{i,1}, \dots, X_{i,p_0})$, $X_{i,1:\lfloor \frac{p_0}{2} \rfloor} \sim \mathcal{B}(0.5)$ Bernoulli variables and $X_{i,\lfloor \frac{p_0}{2} \rfloor+1:p_0} \sim \mathcal{N}(2, 0.5)$ Gaussian variables are simulated. Recurrent events are generated from a non-homogenous Poisson process $\lambda(t|X_i) = \lambda_0(t) \exp(\beta^T X_i)$ with $\lambda_0(t) = \frac{\alpha}{\gamma} (\frac{t}{\gamma})^{\alpha-1}$ a Weibull baseline, $\alpha = 2$ the shape parameter and $\gamma = 0.39$ the scale parameter. The first ten associated coefficients are non-zero, with $\beta = (\beta_1, \dots, \beta_{p_0})^T$ and $\beta_1 = \log(5)$, $\beta_{2:4} = \log(1.3)$, $\beta_{5:p_0} = \log(0.7)$. The true expected number of events is $\mu^*(t|X_i) = \int_0^t \lambda(u|X_i) du = (\frac{t}{\gamma})^\alpha \exp(\beta^T X_i)$. The censoring process is then simulated based on a uniform distribution $\mathcal{U}(0, 3)$. With a terminal event, the recurrent event process and covariates are simulated in the same way as above. The censoring process is simulated based on a uniform distribution $\mathcal{U}(0, 8)$. The terminal event is simulated using a Cox model with shape parameter is 8 and scale parameter is 1.8. The same covariates as the recurrent event process are included with same coefficients β . We set $p_0 = 10$.

3.1.2. Low- and high-dimensional scenarios

To define low- and high-dimensional scenarios, we introduce q independent noise covariates randomly drawn from a standard normal distribution and add them to simulated datasets. We set $q = 10$ for low-dimensional scenarios, and $q = 290$ in high-dimensional scenarios. The total number of covariates for each scenario is $p = p_0 + q$.

Complete. When we analyze scenarios that involve all p generated covariates, we refer to these as ‘complete’ datasets analyses.

Missing. To simulate real-world conditions where datasets may have missing values, we intentionally introduced missing data. Specifically, we randomly set 5% of the covariate X_1 to NA across all individuals. This was done in a completely random manner, ensuring that the missing data does not follow any pattern and is not dependent on any other variables or the values of X_1 itself. The missing data mechanism is completely at random. We refer to such scenarios as ‘Missing’.

Random. We created scenarios where the covariates are generated independently of the recurrent events to simulate a situation where no underlying factors influence the counting process. In such cases, q independent noise covariates are randomly drawn from a standard normal distribution. We created scenarios where the covariates are generated independently of the recurrent events to simulate a situation where no underlying factors influence the counting process. We set $q = 20$ for low-dimensional scenarios, and $q = 300$ in high-dimensional scenarios. The total number of covariates for each scenario is $p = q$, ensuring that all covariates are unassociated with the event data and are purely random. We refer to these scenarios as ‘Random’.

3.2. Results

Performances were assessed in a framework of 250 training sets and one external validation set. The non-parametric estimator uses no covariates, regardless of the dimensionality by construction. For the GL model, no variable selection was performed, meaning all p covariates were included in the model. This limits the analysis for the GL model to low-dimensional scenarios only.

TABLE 2
Means and standard deviations of the C-index without a terminal event

Scenario\Model	Np	GL	RecForest mtry = 1	RecForest mtry = \sqrt{p}	RecForest mtry = $\log(p)$
Low dimensional $\{n = 250, p = 20\}$					
Complete		0.55 (0.12)	0.68 (0.08)	0.71 (0.04)	0.70 (0.05)
Missing	0.55 (0.04)	0.52 (0.10)	0.65 (0.14)	0.69 (0.15)	0.67 (0.15)
Random		0.49 (0.05)	0.56 (0.15)	0.54 (0.15)	0.58 (0.18)
High dimensional $\{n = 250, p = 300\}$					
Complete		/	0.67 (0.21)	0.70 (0.11)	0.70 (0.17)
Missing	0.55 (0.04)	/	0.60 (0.29)	0.64 (0.18)	0.63 (0.24)
Random		/	0.51 (0.31)	0.55 (0.25)	0.56 (0.29)

Np = non-parametric estimator; GL = Gosh-Lin model with no variable selection.
 RecForest was trained with fixed values for *minsplit* = 5 and *nodesize* = 10. 250 learning sets and one external validation set were generated for each scenario. Values closer to 1 indicate higher performance.

3.2.1. Without a terminal event

On average, 62% of the individuals experienced at least one recurrent event, 46% had at least two recurrent events, 26% had at least five recurrent events, and circa four recurrent events per individual. Table 2 below reports performances in terms of C-index values. Overall, performances based on C-index values are greater in scenarios with neither missing data, nor high-dimensionality, both in average and in variability. As expected, scenarios with random inputs lead to randomness with C-index values neighboring 0.50 for each model. The non-parametric estimator provides an average C-index of 0.55. The GL model seems to suffer from not being well-specified, with average C-index values ranging from 0.49 to 0.55 where assessable. RecForest consistently outperforms, irrespective of mtry with values ranging from 0.64 up to 0.71 (random scenarios are not deemed for comparing performance). Besides, it is not impacted by the introduction of massive noisy data, as C-index values remain similar across low- and -high-dimensional scenarios. Table 3 outlines performances in terms of integrated scores. As checked with C-indices, there is no expectations in the interpretation of the random scenarios, hence there are not displayed. The non-parametric estimator is the reference model in the computation of the score. Similar conclusions may be drawn from the different scenarios with the outperformance of RecForest. Yet, higher variability is observed, especially when introducing missing data.

TABLE 3
Means and standard deviations of the integrated score without a terminal event

Scenario\Model	GL	RecForest mtry = 1	RecForest mtry = \sqrt{p}	RecForest mtry = $\log(p)$
Low dimensional $\{n = 250, p = 20\}$				
Complete	50.45 (41.00)	208.10 (102.42)	539.49 (451.96)	161.12 (87.65)
Missing	35.78 (17.91)	325.30 (189.63)	498.75 (415.28)	258.90 (112.14)
High dimensional $\{n = 250, p = 300\}$				
Complete	/	309.47 (134.57)	388.20 (226.95)	355.65 (117.78)
Missing	/	398.70 (229.57)	574.50 (318.75)	475.20 (213.80)

GL = Gosh-Lin model with no variable selection. RecForest was trained with fixed values for *minsplit* = 5 and *nodelsize* = 10. 250 learning sets and one external validation set were generated for each scenario. Higher values indicate higher performance.

3.2.2. With a terminal event

On average, 44% of individuals experienced a terminal event during the observation period. Overall, performance discrepancies in terms of C-index values (Table 4) were observed when dealing with missing data or randomness, with performance being notably lower compared to complete data, as expected. The non-parametric estimator exhibited poor performance (C-index = 0.52 (0.03)). In both low and high-dimensional datasets, RecForest tends to perform better with higher mtry values, always reporting higher C-index values compared to non-parametric estimator and GL model. However, it is notable that in the high-dimensional scenarios, the performance drop is more significant. GL model performs better than non-parametric estimator only with complete data scenarios (C-index = 0.57 (0.07)). Integrated scores for evaluating approaches with a terminal event are displayed in Table 5. We observe similar results than without a terminal event. RecForest consistently yields integrated score values exceeding 300. The decrease in performance compared to the GL model is evident, with IScore values of 110.86 (75.14) and 112.81 (77.64) observed in complete and missing data scenarios, respectively.

In summary of the simulation study, our findings illustrate RecForest superior performance across all examined scenarios. Unlike the comparator GL model, RecForest effectively addresses both missing data and high-dimensionality. Furthermore, in random scenarios, RecForest outputs randomness, implying its reliability when the input lacks discernible patterns.

TABLE 4
Means and standard deviations of the C-index with a terminal event

Scenario\Model	Np	GL	RecForest <i>mtry = 1</i>	RecForest <i>mtry = \sqrt{p}</i>	RecForest <i>mtry = log(p)</i>
Low dimensional $\{n = 250, p = 20\}$					
Complete		0.57 (0.07)	0.79 (0.05)	0.80 (0.04)	0.82 (0.04)
Missing	0.52 (0.03)	0.51 (0.11)	0.73 (0.19)	0.71 (0.13)	0.75 (0.11)
Random		0.45 (0.10)	0.53 (0.16)	0.51 (0.11)	0.50 (0.19)
High dimensional $\{n = 250, p = 300\}$					
Complete		/	0.71 (0.19)	0.69 (0.13)	0.74 (0.10)
Missing	0.52 (0.03)	/	0.64 (0.20)	0.68 (0.13)	0.71 (0.11)
Random		/	0.49 (23.10)	0.48 (12.30)	0.50 (20.09)

Np = non-parametric estimator; GL = Gosh-Lin model with no variable selection.
 RecForest was trained with fixed values for *minsplit* = 5 and *nodesize* = 10. 250 learning sets and one external validation set were generated for each scenario. Values closer to 1 indicate higher performance.

TABLE 5
Means and standard deviations of the integrated score with a terminal event

Scenario\Model	GL	RecForest <i>mtry = 1</i>	RecForest <i>mtry = \sqrt{p}</i>	RecForest <i>mtry = log(p)</i>
Low dimensional $\{n = 250, p = 20\}$				
Complete	110.86 (75.14)	315.76 (119.41)	446.14 (410.88)	410.90 (115.54)
Missing	112.81 (77.64)	368.62 (211.11)	406.20 (275.57)	392.85 (138.23)
High dimensional $\{n = 250, p = 300\}$				
Complete	/	547.89 (229.37)	589.14 (472.33)	628.67 (122.41)
Missing	/	392.34 (39.16)	578.52 (336.71)	512.85 (441.29)

GL = Gosh-Lin model with no variable selection. RecForest was trained with fixed values for *minsplit* = 5 and *nodesize* = 10. 250 learning sets and one external validation set were generated for each scenario. Higher values indicate higher performance.

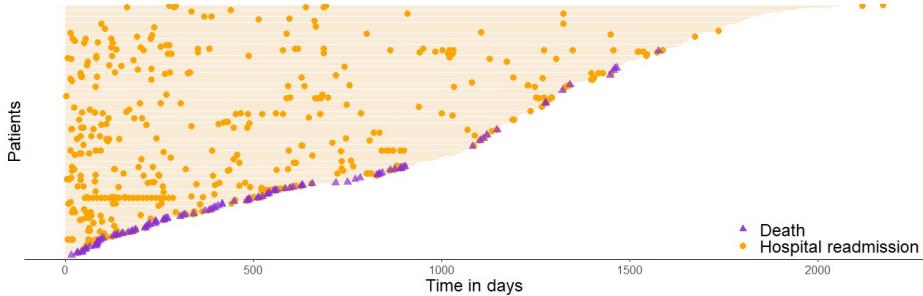


FIG 3. Event plot for readmission data

4. Illustrative example: the readmission data

Readmission dataset from the `frailtypack` R package from Rondeau, Marzroui and Gonzalez (2012) is widely used to demonstrate methodological principles from recurrent events analysis in presence of a terminal event. The data consist of multiple rehospitalizations after surgery in 403 patients diagnosed with colorectal cancer. Available factors are sex (MF), chemotherapy treatment (YesNo), Dukes' tumoral stage (with levels A-B, C, and D), and time-dependent comorbidity Charlson's index (with levels 0, 1-2, and ≥ 3). In average, there were 1.13 (min. – max. = 0 – 22) hospital readmissions per patients, with 199 patients with no admission and a total of 106 deaths (Figure 3).

In absence of an external validation set, performances were assessed with a 10-fold cross-validation procedure. We consider the following models: four multivariate Ghosh-Lin models with arbitrary combinations of factors, and `RecForest`. The reference model is the non-parametric estimator. Hyperparameters from `RecForest` `minsplit`, `nodelsize` and `mtry` were tuned on the total sample and the OOB score was minimized for $\{ntrees = 100, minsplit = 2, nodelsize = 1, mtry = 2\}$ (Figure 6 in the Supplementary).

In our analysis (results in Table 6), the non-parametric estimator registers a C-index = 0.58 (0.05). `RecForest` outperforms with C-index = 0.80 (0.04). All GL models, with one to four covariates for adjustment, maintain relatively consistent C-indices around 0.45 to 0.53. Comparing IMSE and IScore metrics, `RecForest` and the non-parametric estimator are not directly comparable due to construction. Specifically, the non-parametric reference for the integrated score in `RecForest` is constructed for each bootstrap sample. Integrated scores for GL models operate on the overall dataset from the

TABLE 6
Means and standard deviations over the 10-fold cross-validation for readmission dataset

Metric\Model	Np	GL1	GL2	GL3	GL4	RecForest	GL*
C-index ↑	0.58 (0.05)	0.53 (0.08)	0.48 (0.08)	0.48 (0.07)	0.45 (0.05)	0.80 (0.04)	0.60 (0.06)
IMSE ↓	7 883.50 (6 229.47)	7 843.99 (6 106.36)	8 361.16 (6 292.29)	8 229.08 (6 478.35)	9 981.50 (6 064.23)	706.02 (508.96)	7 934.28 (6 606.23)
IScore ↑	ref. ref.	39.41 (230.6)	-477.67 (348.48)	-345.62 (432.6)	-2 098.44 (541.59)	188.22 (89.00)	51.33 (142.63)

Np = non-parametric estimator; GL1 = Gosh-Lin model with sex; GL2 = Gosh-Lin with sex and chemotherapy; GL3 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage; GL4 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage and Charlson's index; GL* = Ghosh-Lin model with best variables from RecForest. Arrows indicate whether higher or lower scores lead to best performances.

ongoing fold. Consequently, IScore values from RecForest do not simply reflect the difference between IMSE values from the non-parametric estimator and RecForest, as opposed to IScores from GL models.

IMSE and IScore for RecForest indicate lower margin of errors. Among GL models, GL1 (Gosh-Lin model with sex) exhibits lower IMSE than the non-parametric estimator, resulting in a higher IScore. GL2 to GL4 (GL2 = Gosh-Lin with sex and chemotherapy; GL3 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage; GL4 = Gosh-Lin model with sex, chemotherapy and Dukes' tumoral stage and Charlson's index) yield negative IScore values, indicating high variability among GL models observed in our simulation study. Besides, high variability is observed across all approaches.

Variable importance for RecForest was based on both the C-index and the opposite of the integrated MSE (Figure 4). Most important variable identified by RecForest was the Charlson comorbidity index. Sex and chemotherapy did not seem to have an impact on the predictive performance. Variable selection enabled to reach better performance for GL* model.

Prediction curves for RecForest as the expected number of recurrent events are displayed in Figure 5. Predictions were generated for two patients, one with the highest Charlson comorbidity score (in orange), and the other with the lowest (in blue). We observe for the patient in orange that the model predicted an expected number of three readmissions as the patient dies after two observed readmissions. For the patient in blue, the model predictions are in line with observed events.

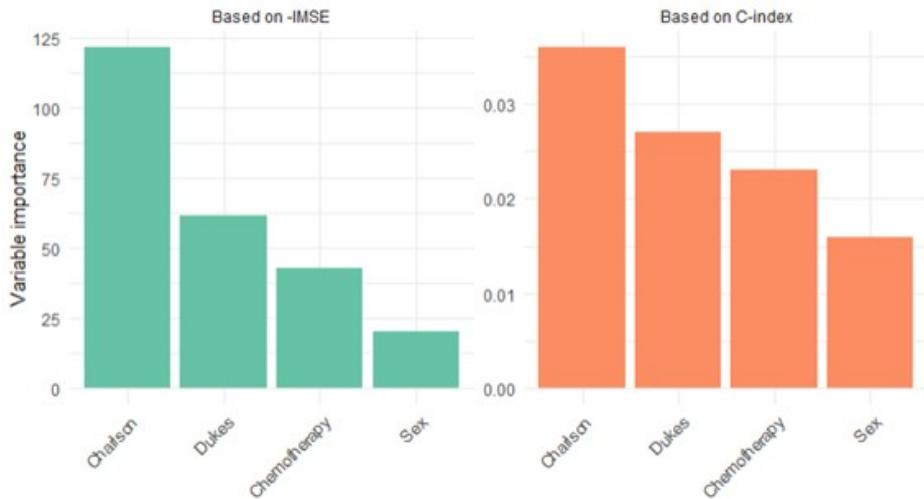


FIG 4. Variable importance of RecForest computed on the C-index and the opposite of the integrated MSE. Charlson refers to Charlson comorbidity index, Dukes refers to tumoral Dukes stage.

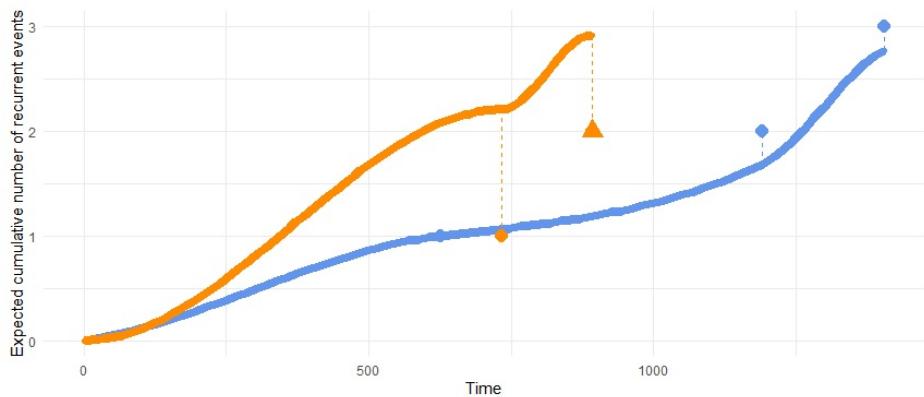


FIG 5. Expected cumulative number of recurrent events with RecForest for two patients, one in orange with the highest Charlson comorbidity score, and the other in blue with the lowest. Data points outside the prediction curves are observed data. Triangle indicates the patient died.

5. Discussion

We developed RecForest by extending the RSF algorithm to handle recurrent events in a survival framework, in potential presence of a terminal event, of longitudinal markers, and of missing data. To do so, the splitting rule at each node was tailor-made for recurrent events analysis and mean cumulative number of events served in terminal node estimators. We characterized the performance both with discrimination and calibration by introducing a generalized C-index for recurrent event analysis and applying an innovative MSE.

In a simulation study, we compared RecForest with two baseline approaches: a non-parametric estimator and a GL model. Scenarios included variations of the presence of a terminal event, the low- or high-dimensionality of the data, and the inclusion of missing values. In instances where missing values or high-dimensional data were present, greater variability was observed across all scenarios for both metrics. In all explored cases, RecForest demonstrated higher performances both in terms of C-index and integrated score values compared with baseline. The impact was quite little when introducing massive noisy data. Besides, RecForest emerged as the sole modeling approach capable of handling high-dimensionality scenarios with no prior variable selection. Furthermore, we presented a practical application using well-known open-source data, showcasing how the fine-tuning of RecForest hyperparameters leads to a more performant model. Again, RecForest exhibited superior predictive performance, achieving a C-index of 0.80 alongside strong calibration metrics (IMSE = 706, IScore = 188). Overall, across both simulated and real-world datasets, RecForest consistently emerged as the most effective modeling approach.

In practical applications, high-dimensional problems involving recurrent events are often sidestepped by transforming the recurrent event survival framework into alternative formats, such as an event count, a time-to-first-event endpoint, or a classification problem. However, each of these transformations may lead to the voluntary omission of valuable information. In response to this, RecForest aims to bridge a recognized gap in handling such applications, ensuring a more comprehensive and nuanced analysis of recurrent event data. Additionally, our algorithm benefits from random forests features, i.e. the ability of handling missing data or multicollinearity, and reducing overfitting thanks to bagging principle.

Additional settings can be explored to integrate a terminal event within the proposed approach. For instance, [Charles-Nelson, Katsahian and Schramm \(2019\)](#) suggested working with inverse probability of survival weighting (IPSW)

to compute coefficient weights in the Ghosh-Lin model, whereas we used inverse probability of censoring weighting. IPSW is typically recommended when modeling the terminal event is also of interest. Another example would be to use frailty models, either joint or additive as per [Rondeau, Marzroui and Gonzalez \(2012\)](#). Besides, natural extensions of random forests, serving as ensemble methods, have been widely used to improve performance through boosting techniques like Gradient Boosting ([Friedman \(2001\)](#)), Extreme Gradient Boosting ([Chen and Guestrin \(2016\)](#)), or LightGBM ([Ke et al. \(2017\)](#)). Since all these methods are grounded in tree-based structures, they offer seamless extensions to the proposed approach, and would hence provide innovative tools for recurrent event analysis.

Our methodology also suffers from several drawbacks. The primary limitation of random forest-like algorithms lies in the computation time, which grows with the number of trees, the dimensionality of the data and the numbers of variables selected at each tree node. Second, we assumed the proportional hazard assumption of the Gosh model, which may not universally hold in real-world settings. Furthermore, variable importance measures provided do not account for potential correlations. To address this limitation, the implementation of grouped variable importance statistics is a promising avenue for further refinement ([Devaux et al. \(2023\)](#); [Gregorutti, Michel and Saint-Pierre \(2015\)](#)). Nevertheless, signs of associations would still be unavailable. Another potential limitation of random forests is their static usage of features. Dynamic predictions could indeed be included as per [Cottin et al. \(2022\)](#) and [Moradian et al. \(2022\)](#).

On the other hand, the issue of interpretability in machine and deep learning, particularly in digital health has garnered significant attention, as pointed out in [Farah et al. \(2023\)](#). Several explainability methods have been proposed such as Local Interpretable Model-agnostic Explanations (LIME, [Ribeiro, Singh and Guestrin \(2016\)](#)), SHapley Additive exPlanations (SHAP, [Lundberg and Lee \(2017\)](#)), and counterfactual explanations ([Guidotti \(2022\)](#); [Bhan et al. \(2023\)](#)). These interpretability techniques have been recently adapted for survival analysis ([Cottin et al. \(2024\)](#); [Kovalev, Utkin and Kasimov \(2020\)](#)). Moreover, random forest-like methods offer a valuable tool for variable selection, especially in addressing high-dimensionality or obtaining hazard ratios that are intrinsically interpretable ([Khan and Shaw \(2016\)](#); [Wang and Li \(2017\)](#)). Approaches such as permutation-based selection, variable hunting, and iterative feature elimination serve as effective means towards this purpose ([Genuer, Poggi and Tuleau-Malot \(2010\)](#); [Ishwaran et al. \(2010\)](#); [Pang et al. \(2012\)](#)).

6. Conclusion

To conclude, we introduced a new algorithm based on survival theory for recurrent events with or without a terminal event and ensemble-based methodology for learning. RecForest is readily accessible to adequately answer further clinical needs.

Funding

Author Murris J reports a grant from the Association Nationale de la Recherche et de la Technologie, with Pierre Fabre, Convention industrielle de formation par la recherche number 2020/1701.

Supplementary Material

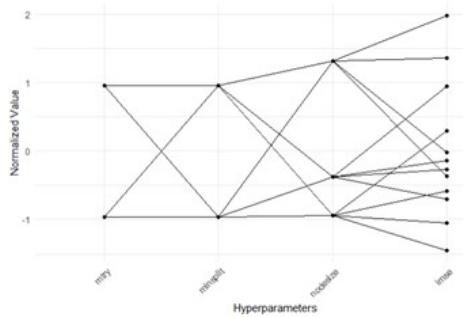


FIG 6. Hyperparameter optimization on readmission data based on out-of-bag scores

Hyperparameter optimization on readmission data based on out-of-bag scores

Hyperparameter optimization was performed using a grid search approach.

References

- AMORIM, L. D. A. F. and CAI, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology* **44** 324–333.
- ANDERSEN, P. K. and GILL, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics* **10** 1100–1120.
- BHAN, M., VITTAUT, J.-N., CHESNEAU, N. and LESOT, M.-J. (2023). TIGTEC: Token Importance Guided TExt Counterfactuals. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part III* 496–512. Springer-Verlag, Berlin, Heidelberg.
- BOUAZIZ, O. (2024). Assessing model prediction performance for the expected cumulative number of recurrent events. *Lifetime Data Analysis* **30** 262–289.
- BREIMAN, L. (2001). Random Forests. *Machine Learning* **45** 5–32.
- CHARLES-NELSON, A., KATSAHIAN, S. and SCHRAMM, C. (2019). 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* sim.8168.

- CHEN, T. and GUESTRIN, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16* 785–794. Association for Computing Machinery, New York, NY, USA.
- CHEN, S. and XU, C. (2023). Handling high-dimensional data with missing values by modern machine learning techniques. *Journal of Applied Statistics* **50** 786–804.
- COOK, R. J. and LAWLESS, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16** 911–924.
- COOK, R. J., LAWLESS, J. F. and LEE, K.-A. (2010). A copula-based mixed Poisson model for bivariate recurrent events under event-dependent censoring. *Statistics in Medicine* **29** 694–707.
- COTTIN, A., PECUCHET, N., ZULIAN, M., GUILLOUX, A. and KATSAGHIAN, S. (2022). IDNetwork: A deep illness-death network based on multi-state event history process for disease prognostication. *Statistics in Medicine* **41** 1573–1598.
- COTTIN, A., ZULIAN, M., PÉCUCHE, N., GUILLOUX, A. and KATSAGHIAN, S. (2024). MS-CPFI: A model-agnostic Counterfactual Perturbation Feature Importance algorithm for interpreting black-box Multi-State models. *Artificial Intelligence in Medicine* **147** 102741.
- COX, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34** 187–202.
- DEVAUX, A., HELMER, C., GENUER, R. and PROUST-LIMA, C. (2023). Random survival forests with multivariate longitudinal endogenous covariates. *Statistical Methods in Medical Research* **32** 2331–2346.
- FARAH, L., MURRIS, J. M., BORGET, I., GUILLOUX, A., MARTELLI, N. M. and KATSAGHIAN, S. I. M. (2023). Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence-Based Health Technologies: What Healthcare Stakeholders Need to Know. *Mayo Clinic Proceedings: Digital Health* **1** 120–138.
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29** 1189–1232.
- GENUER, R., POGGI, J.-M. and TULEAU-MALOT, C. (2010). Variable selection using random forests. *Pattern Recognition Letters* **31** 2225–2236.
- GHOSH, D. and LIN, D. Y. (2000). Nonparametric Analysis of Recurrent Events and Death. *Biometrics* **56** 554–562.
- GHOSH, D. and LIN, D. Y. (2002). Marginal Regression Models for Recurrent and Terminal Events. *Statistica Sinica* **12** 663–688.
- GRAF, E., SCHMOOR, C., SAUERBREI, W. and SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for sur-

- vival data. *Statistics in Medicine* **18** 2529–2545.
- GREGORUTTI, B., MICHEL, B. and SAINT-PIERRE, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis* **90** 15–35.
- GUIDOTTI, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*.
- HARRELL, F. E., LEE, K. L. and MARK, D. B. (1996). MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. *Statistics in Medicine* **15** 361–387.
- HUANG, Y., LI, J., LI, M. and APARASU, R. R. (2023). Application of machine learning in predicting survival outcomes involving real-world data: a scoping review. *BMC medical research methodology* **23** 268.
- ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *The Annals of Applied Statistics* **2** 841–860.
- ISHWARAN, H., KOGALUR, U. B., GORODESKI, E. Z., MINN, A. J. and LAUER, M. S. (2010). High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association* **105** 205–217.
- ISHWARAN, H., GERDS, T. A., KOGALUR, U. B., MOORE, R. D., GANGE, S. J. and LAU, B. M. (2014). Random survival forests for competing risks. *Biostatistics (Oxford, England)* **15** 757–773.
- KAPLAN, E. L. and MEIER, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53** 457–481.
- KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., YE, Q. and LIU, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems* **30**. Curran Associates, Inc.
- KHAN, M. H. R. and SHAW, J. E. H. (2016). Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing* **26** 725–741.
- KIM, S., SCHaubel, D. E. and McCULLOUGH, K. P. (2018). A C-index for recurrent event data: Application to hospitalizations among dialysis patients. *Biometrics* **74** 734–743.
- KOVALEV, M. S., UTKIN, L. V. and KASIMOV, E. M. (2020). SurvLIME: A method for explaining machine learning survival models. *Knowledge-Based Systems* **203** 106164.
- LAWLESS, J. F. and NADEAU, C. (1995). Some Simple Robust Methods for

- the Analysis of Recurrent Events. *Technometrics* **37** 158–168.
- LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17* 4768–4777. Curran Associates Inc., Red Hook, NY, USA.
- MORADIAN, H., YAO, W., LAROCQUE, D., SIMONOFF, J. S. and FRYDMAN, H. (2022). Dynamic estimation with random forests for discrete-time survival data. *Canadian Journal of Statistics* **50** 533–548.
- MURRIS, J., CHARLES-NELSON, A., TADMOURI SELLIER, A., LAVENU, A. and KATSAGIANI, S. (2023). Towards filling the gaps around recurrent events in high dimensional framework: a systematic literature review and application*. *Biostatistics & Epidemiology* **7** e2283650.
- OZGA, A.-K., KIESER, M. and RAUCH, G. (2018). A systematic comparison of recurrent event models for application to composite endpoints. *BMC medical research methodology* **18** 2.
- PANG, H., GEORGE, S. L., HUI, K. and TONG, T. (2012). Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **9** 1422–1431.
- PRENTICE, R. L., WILLIAMS, B. J. and PETERSON, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68** 373–379.
- RIBEIRO, M. T., SINGH, S. and GUESTRIN, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16* 1135–1144. Association for Computing Machinery, New York, NY, USA.
- RONDEAU, V., MARZROUI, Y. and GONZALEZ, J. R. (2012). 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* **47** 1–28.
- TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16** 385–395.
- UNO, H., CAI, T., PENCINA, M. J., D'AGOSTINO, R. B. and WEI, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30** 1105–1117.
- VAN BELLE, V., PELCKMANS, K., VAN HUFFEL, S. and SUYKENS, J. A. K. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in*

Medicine **53** 107–118.

- WANG, H. and LI, G. (2017). A Selective Review on Random Survival Forests for High Dimensional Data. *Quantitative bio-science* **36** 85–96.
- WEI, L. J., LIN, D. Y. and WEISSFELD, L. (1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association* **84** 1065–1073.

III.3 Application de RecForest aux données du PMSI

III.3.1 Contexte

À l'origine du projet de thèse, Abir Tadmouri-Sellier et moi avions envisagé une application finale auprès des patients atteints de cancer colorectal métastatique (mCRC). En 2020, Pierre Fabre a obtenu l'autorisation européenne de mise sur le marché pour une molécule dans le traitement des patients adultes mCRC avec mutation BRAFV600E. Cette annonce a suscité un intérêt accru pour l'étude de cette population en conditions réelles [Martinelli et al., 2022]. C'est dans ce contexte que nous avons débuté une collaboration en juin 2023 avec le Dr S. Tzedakis, chirurgien digestif à l'hôpital Cochin (AP-HP).

Les données utilisées dans cette application de RecForest ont été extraites du Programme de Médicalisation des Systèmes d'Information (PMSI), un système qui recueille et stocke des informations médicales sur l'ensemble des hospitalisations en France. Le PMSI offre un accès à des données démographiques, médicales et administratives détaillées sur un vaste échantillon de patients. L'absence de données de laboratoire constitue l'un des principaux défis des données médico-administratives, rendant impossible l'identification des mutations dans ces bases. Dans l'impossibilité d'identifier les mutations, la population étudiée englobait ainsi plus largement les patients ayant bénéficié d'une chirurgie pour cancer colorectal.

L'objectif principal de l'étude était d'estimer l'incidence des réadmissions hospitalières chez ces patients. Les réadmissions peuvent être liées aux complications postopératoires, ce qui les rend particulièrement pertinentes pour évaluer la qualité des soins et les besoins en suivi postopératoire. L'objectif secondaire visait à évaluer le nombre de réadmissions survenant au cours du temps, notamment dans les 6 mois suivant l'intervention. Cette approche permet de mieux comprendre les risques et les facteurs associés aux réadmissions, facilitant ainsi l'amélioration des stratégies de gestion clinique pour les patients atteints de mCRC. Pour un intérêt de santé publique plus global, nous avons finalement étendu la population aux patients ayant bénéficié d'une chirurgie pour cancer digestif.

Note : Ce travail est encore en cours, notamment pour la vérification de la population et le codage des variables sur cette base de données complexe. Bien que ce projet soit présenté à titre illustratif dans cette thèse, l'intérêt en termes de santé publique est bien réel.

III.3.2 Analyse des réadmissions postopératoires auprès des patients atteints de cancer digestif

III.3.2.1 Bref contexte médical

Les réadmissions à l'hôpital après une chirurgie digestive constituent un enjeu majeur en santé publique, impactant à la fois les patients et les systèmes de santé [El Amrani et al.,

2019]. En effet, les réadmissions postopératoires sont souvent associées à des complications telles que des infections, des obstructions ou des fuites anastomotiques, qui peuvent gravement affecter la récupération des patients et augmenter la morbidité et la mortalité. Selon Merkow et al. [2015], le délai médian de réadmission était de 8 jours (intervalle interquartile, 3-14 jours). Par ailleurs, Symons et al. [2013] ont identifié que les réadmissions fréquentes sont souvent liées à des facteurs comme l'âge avancé, aux comorbidités, et le type d'intervention chirurgicale. Comprendre les réadmissions postopératoires permettrait de développer des stratégies de prévention efficaces, améliorant les résultats cliniques pour les patients, permettant aux chirurgiens d'évaluer la qualité des soins, et réduisant éventuellement les coûts de santé.

III.3.2.2 Objectifs

L'objectif de cette étude est dans un premier temps de décrire les réadmissions postopératoires dans les 6 mois suivant une première chirurgie digestive, puis de développer un modèle prédictif pour évaluer les risques de réadmission.

III.3.2.3 Méthodologie

Source de données Les données ont été extraites de la base de données du Programme de Médicalisation des Systèmes d'Information (PMSI) [Moulis et al., 2015]. La base de données inclut toutes les procédures chirurgicales remboursées réalisées en France, dans tous les hôpitaux, indépendamment de leur affiliation académique ou de leur propriété (publique et privée à but lucratif et privée à but non lucratif). Les données du PMSI sont collectées sous forme de rapports standardisés et comprennent :

- Les données démographiques des patients (âge, sexe, code postal, dates d'entrée et de sortie) ;
- Les diagnostics primaires et associés basés sur la Classification Internationale des Maladies, 10ème édition (CIM-10) ;
- Les procédures thérapeutiques basées sur la Classification Commune des Actes Médicaux (CCAM, 11ème édition), qui est une classification nationale standardisée des procédures médicales.

Un identifiant unique et anonyme est attribué à chaque patient, permettant ainsi d'identifier l'ensemble de ses séjours hospitaliers programmés ou non dans tous les hôpitaux de France. Le consentement du patient n'est pas requis puisque les informations individuelles sont anonymes.

Sélection des patients Nous avons inclus tous les patients adultes qui ont bénéficié d'une chirurgie digestive à partir des codes CCAM (Annexe B.1 Tableau B.1) : chirurgie colorectale, chirurgie de l'intestin grêle, chirurgie hépatobiliaire, chirurgie pancréatique, chirurgie cesogastrique. Nous avons extrait les informations de la base de données du PMSI de janvier 2018 à juin 2023, et inclus les patients ayant bénéficié d'une première chirurgie

digestive de janvier 2020 à décembre 2022. Nous avons exclu les patients dont le codage des procédures était ambigu et les patients présentant une erreur de codage.

Analyse des données En premier lieu, nous avons conduit une analyse descriptive. Les variables qualitatives sont présentées sous forme de nombres et de proportions. Les caractéristiques démographiques et cliniques continues sont présentées sous forme de médiane avec intervalle interquartile. Nous avons décrit la population au global et par type de chirurgie. Les patients pouvaient bénéficier de plusieurs types de chirurgie, ainsi aucun test statistique n'a été fait pour comparaison. Nous avons également décrit les réadmissions à l'hôpital à l'aide de l'estimateur de Nelson-Aalen de [Lawless and Nadeau \[1995\]](#). Les patients inclus ont été censurés soit à la fin de la période de suivi, soit lorsqu'ils étaient perdus de vue. Si le nombre de décès est négligeable (inférieur à 5%), nous considérerons le décès comme une censure non-informative.

Application de RecForest Au total, nous avons utilisé 113 variables prédictives pour RecForest. Parmi elles, on compte 2 variables démographiques (sexe et tranche d'âge), 5 types de chirurgie, 3 indices de comorbidité, 66 codes CIM-10 pour les comorbidités, et 37 types d'actes médicaux réalisés. Le nombre attendu de réadmissions à l'hôpital a été prédit en fonction du temps écoulé depuis la première chirurgie digestive.

Pour évaluer la performance du modèle, nous utilisons comme ensemble de test les patients ayant bénéficié d'une première chirurgie digestive au cours des 12 derniers mois. Cela permet de garantir que le modèle sera performant avec les données futures. L'ensemble d'entraînement de RecForest se compose des patients opérés entre janvier 2020 et décembre 2021, et l'ensemble de test des patients opérés entre janvier et décembre 2022.

Pour l'optimisation des hyperparamètres, nous utilisons l'ensemble d'entraînement en entier plutôt que la validation croisée, conformément à [Breiman \[2001b\]](#) : "*Therefore, using the out-of-bag error estimate removes the need for a set aside test set*". Nous avons fixé $n_{trees} = 50$, $minsplit = 30$ et $nodesize = 20$. L'hyperparamètre $mtry$ a été optimisé avec $mtry \in \{1, 5, 10\}$ sur l'ensemble d'entraînement, avec 5 répétitions pour minimiser l'erreur OOB. L'importance des variables a été calculée avec 5 permutations et est présentée en termes de pourcentage de variation de l'opposé du IMSE.

III.3.2.4 Résultats

Nous avons identifié 255 732 patients adultes ayant bénéficié d'une première chirurgie digestive entre janvier 2020 et décembre 2022. Les caractéristiques des patients sont décrivées en Annexe B.1 Tableau B.2. La première chirurgie était pour 131 260 (51%) patients une chirurgie colorectale, pour 21 663 (8,5%) patients une chirurgie hépatobiliaire, pour 10 078 (3,9%) patients une chirurgie pancréatique, pour 13 280 (5,2%) patients une chirurgie de l'intestin grêle et pour 70 298 (27%) patients une chirurgie œsogastrique. 145 102 (57%) patients étaient des femmes et l'âge médian (Q1 - Q3) était de 62 (47 - 73) ans. 81 495 (32%) patients souffraient d'obésité à l'inclusion. L'indice de comorbidité de Charlson était nul pour 106 201 (42%) patients. La chirurgie œsogastrique concernait plutôt les femmes (73%)

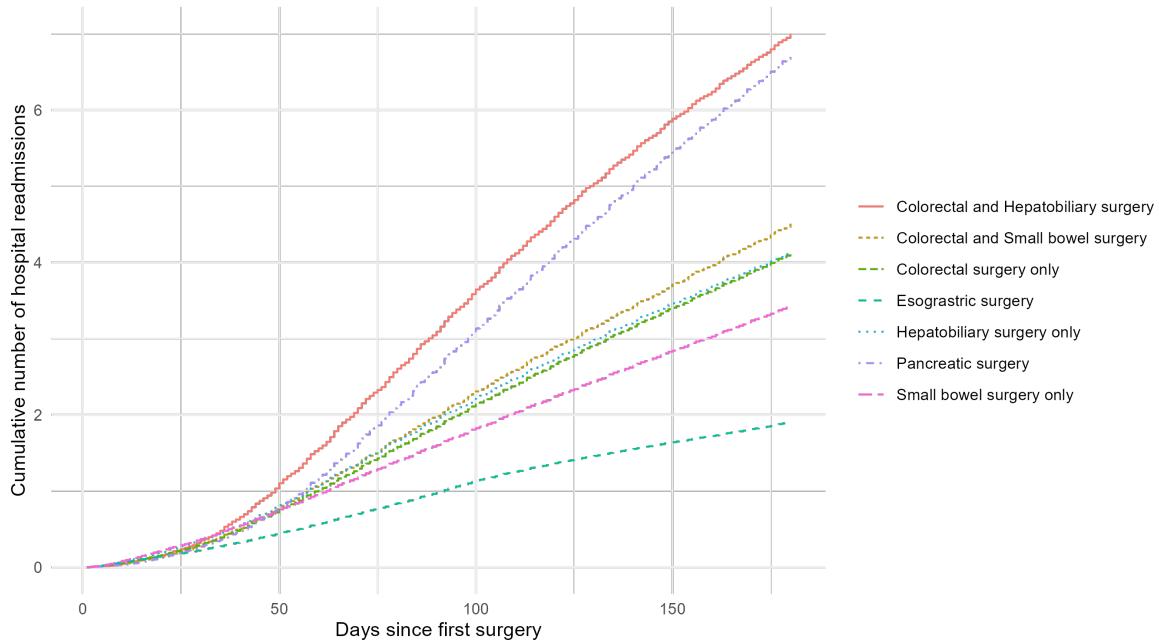


FIGURE III.3 – Nombre cumulé de réadmissions postopératoires par type d'intervention chirurgicale

et des patients plus jeunes (médiane ($Q_1 - Q_3$) = 42 (32 - 54)). L'indice de comorbidité de Charlson était nul pour 393 (3.9%) patients pour la chirurgie pancréatique, et 44 927 (64%) pour la chirurgie œsogastrique. Parmi les patients avec une première chirurgie pour l'intestin grêle, 3 623 (27%) ont également bénéficié d'une première chirurgie colorectale. Au total, 10 230 (4.0%) sont décédés dans les 6 mois qui ont suivi l'intervention chirurgicale.

Au total, 686 918 réadmissions ont été identifiées, et les actes associés sont décrits dans le Tableau B.4. L'acte le plus commun était pour complication rénale, avec 28 791 (4.2%) réadmissions. On remarque que 413 617 (60%) réadmissions ont eu lieu après une première chirurgie colorectale. 205 666 (80.4%) patients ont eu au moins une réadmission postopératoire et 121 854 (47.6%) patients en ont eu au moins deux dans les 6 mois après la première chirurgie. Le nombre médian ($Q_1 - Q_3$) était 1.0 (1.0, 4.0) réadmissions (Tableau B.3). 51 281 (39%) patients ayant bénéficié d'une chirurgie colorectale ont eu au moins trois réadmissions, 9 419 (43%) patients ayant bénéficié pour une chirurgie hépatobiliare et 6 667 (66%) patients ayant bénéficié pour une chirurgie pancréatique. Le nombre cumulé de réadmissions postopératoires par type d'intervention chirurgicale est présenté en Figure III.3.¹

L'ensemble d'entraînement comptait 184 834 patients, soit 72,28%. L'erreur OOB était minimisée pour $mtry = 5$ avec une moyenne de 1 181,12 (Tableau III.1). Le modèle retenu pour RecForest utilise les hyperparamètres suivants : $\{n_{trees} = 50, mtry = 5, \text{minsplit} = 30, \text{nodesize} = 20\}$.

Les performances sur l'ensemble de test sont présentées dans le Tableau III.2 en termes de C-index, d'IMSE et d'IScore. Le C-index, compris entre 0 et 1, mesure la proportion de

1. Cette Figure est différente de la Figure I.7 du Chapitre I Section I.5, car il était d'intérêt de regarder des combinaisons de type d'intervention chirurgicale.

TABLE III.1 – Erreurs OOB moyennes pour l'optimisation de *mtry* avec 5 répétitions sur l'ensemble d'entraînement

<i>mtry</i> = 1	<i>mtry</i> = 5	<i>mtry</i> = 10
3 275,77	1 181,12	1 996,56

Avec $\{n_{trees} = 50, \text{minsplit} = 30, \text{nodesize} = 20\}$.

patients correctement classés en fonction de leurs taux d'événements observés et prédits ($C - \text{index} = 0,72$). L'IMSE est calculé sur l'ensemble de la période, et une valeur plus faible est préférable ($IMSE = 1398,04$). Comme IMSE est une valeur brute, l'IScore compare le gain de prédiction de RecForest par rapport à un estimateur de référence, ici l'estimateur non-paramétrique de Nelson-Aalen. Un IScore positif et le plus élevé possible est souhaitable ($IScore = 409,32$).

TABLE III.2 – Performances de RecForest sur l'ensemble de test

C-index ↑	IMSE ↓	IScore ↑
0,72	1 398,04	409,32

Avec $\{n_{trees} = 50, \text{mtry} = 5, \text{minsplit} = 30, \text{nodesize} = 20\}$.

L'importance des variables ($Vimp$) est présentée dans la Figure III.4 selon leur catégorie (données démographiques, codes CIM-10, actes, indices de comorbidité, types de chirurgie). Trois groupes de variables se distinguent :

- Les variables les plus importantes, avec $\%Vimp \geq 4\%$;
- Les variables moyennement importantes, avec $1\% \leq \%Vimp < 4\%$;
- Les variables les moins importantes, avec $\%Vimp < 1\%$.

Parmi les variables les plus importantes, on retrouve la maladie rénale chronique avancée, le diabète mellitus non compliqué, le statut de receveur d'une greffe (autre que foie), et la cardiopathie chronique. La chirurgie pancréatique est également parmi les variables les plus importantes.

III.3. Application de RecForest aux données du PMSI



FIGURE III.4 – Importance des variables sur 5 permutations

III.4 Discussion

III.4.1 Une taxonomie pour traiter les événements récurrents en survie

Avec l'introduction de RecForest dans le paysage des méthodes disponibles, nous proposons une taxonomie pour traiter les événements récurrents dans un cadre de survie, comme illustré en Figure III.5. Cette taxonomie vise à clarifier les différentes approches et techniques utilisables pour modéliser et analyser les événements récurrents, en tenant compte des spécificités des données et des besoins analytiques.

La première étape de cette taxonomie consiste à définir le type d'analyse souhaité. Cette décision est influencée par plusieurs facteurs :

- **Données disponibles** : Nombre de patients, caractéristiques recueillies, fréquence des événements récurrents, et la durée de suivi ;
- **Objectif de l'étude** : Prédiction d'un risque, identification de facteurs de risque, ou évaluation de l'efficacité d'un traitement ;
- **Acceptation par le clinicien** : Acceptation de nouvelles méthodes statistiques ou ML ou préférences pour des méthodes plus classiques ;
- **Ressources computationnelles** : Capacités de calcul et de temps disponibles pour mener des analyses complexes.

Une fois le type d'analyse déterminé, il est crucial de recentrer l'objectif de l'analyse à travers une approche par estimand. Cette approche permet de préciser quelle quantité on cherche à estimer avec le modèle. Comme discuté en Section I.6 du Chapitre I, les modèles peuvent estimer différentes quantités, telles que :

- Le taux d'incidence cumulée des événements ;
- Le temps moyen entre les événements récurrents ;
- La probabilité de survenue du prochain événement dans un intervalle de temps donné.

Nous avons également apposé les sujets associés, qui nous semblent être des sujets de recherche à part entière, et qui seront discutés en prochaine section. Cette taxonomie se veut offrir une structure claire pour aborder l'analyse des événements récurrents dans un cadre de survie. Elle aide les chercheurs et les cliniciens à choisir les méthodes les plus appropriées en fonction de leurs données, de leurs objectifs et des ressources disponibles, assurant ainsi des analyses robustes et pertinentes.

III.4.2 Axes de développement

Notre première priorité est de rendre RecForest disponible et accessible au travers du développement d'un package R dédié. Cette méthode a attiré beaucoup d'attention lors de congrès, notamment en remportant un prix du meilleur poster aux Rencontres R de Vannes en juin 2024. Ce projet de développement est en cours et devrait être finalisé d'ici la fin de l'année 2024.



FIGURE III.5 – Taxonomie de l'analyse des événements récurrents

Ensuite, nous souhaitons approfondir notre analyse des réadmissions en examinant de manière détaillée les résultats selon le type de chirurgie, toujours en collaboration avec Dr S. Tzedakis et les laboratoires Pierre Fabre. Cela nous permettra de formuler des recommandations cliniques claires et précises pour améliorer les soins aux patients.

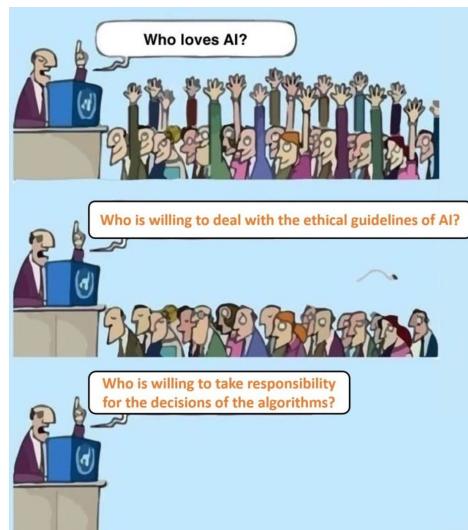
Enfin, les actualités présentées dans la Section III.1.3 concernaient les forêts aléatoires de survie avec une approche axée sur le temps jusqu'au premier événement. Ces développements ouvrent de nombreuses perspectives pour RecForest, notamment en ce qui concerne la sélection de variables. Cela permettrait également de mieux comprendre les facteurs influençant les résultats cliniques, renforçant ainsi la pertinence et l'applicabilité de RecForest dans un contexte de survie complexe.

Messages-clés de ce chapitre

La contribution majeure de ce chapitre est l'algorithme RecForest, qui combine des **approches statistiques non paramétriques et semi-paramétriques** avec le principe des **méthodes d'ensemble**. Nous avons adapté les forêts de survie aléatoires en modifiant les **règles de division** à chaque nœud et les **estimations** des nœuds terminaux afin de prendre en compte les événements récurrents. L'objectif de RecForest est d'estimer le **nombre attendu d'événements récurrents pour chaque individu au fil du temps**. Nous avons démontré la **performance** de RecForest à travers des données simulées et réelles, soulignant sa **flexibilité** dans des contextes variés, notamment en présence d'événements terminaux, de facteurs longitudinaux, de données manquantes et de grande dimension. De plus, nous avons introduit un **nouveau C-index** spécifique aux méthodes de survie avec événements récurrents, qui s'adapte à tous les temps de suivi entre les individus, ce qui en fait une métrique particulièrement pertinente pour les données observationnelles. Nous avons également mis en évidence l'importance d'un MSE spécifique pour la **calibration** des modèles de type RecForest. En collaboration avec des médecins, nous répondons à un **besoin de santé publique** : l'analyse des **réadmissions post-opératoires** chez les patients ayant bénéficié d'une chirurgie pour cancer digestif. Cette thématique de recherche a des **implications multiples**, dans la mesure où elle vise à améliorer la qualité de vie des patients, à soutenir les chirurgiens dans la qualité des soins prodigues, et aide les hôpitaux à mieux gérer leurs éventuelles surcharges. Enfin, nous proposons une **taxonomie** pour traiter les événements récurrents dans un cadre de survie, visant à **structurer la compréhension des événements récurrents et à favoriser des approches harmonisées** dans la recherche clinique et statistique.

Chapitre IV

Interprétabilité, santé et survie : vers une utilisation plus transparente des algorithmes d'IA



Murat Durmus (2022)

Sommaire

IV.1 Importance des critères d'interprétabilité et d'explicabilité pour les dispositifs médicaux	136
IV.1.1 Contexte des dispositifs médicaux	136
IV.1.2 Évaluation de performance, d'interprétabilité et d'explicabilité pour les dispositifs médicaux	137
IV.2 Interprétabilité et survie	157
IV.2.1 Contexte	157

IV.2.2 Étude de cas	157
IV.3 Une méthode d'interprétabilité <i>model-specific</i> pour la survie	169
IV.3.1 Introduction de TreeShap	169
IV.3.2 Extensions possibles pour la survie	172
IV.4 Discussion	173

Note : Dans ce chapitre, nous utiliserons l'acronyme IA pour désigner l'ensemble des algorithmes d'apprentissage automatique, qu'ils relèvent de l'apprentissage profond (*deep learning*) ou d'apprentissage automatique plus traditionnel (*machine learning*).

Introduction

Les modèles d'apprentissage automatique ont révolutionné l'analyse de données, couvrant des problèmes allant de classification, de régression et de survie [Wang et al., 2019]. En particulier, ces avancées permettent une analyse approfondie des données médicales et une aide à la décision dans le cadre clinique. Au Chapitre III, nous avons présenté RecForest, un modèle basé sur les forêts aléatoires, et avons démontré ses performances dans divers scénarios. Une question persiste : à quel point pouvons-nous nous fier à ces prédictions ?

Pour y répondre, nous définissons d'abord les concepts d'interprétabilité et d'explicabilité. De très nombreuses définitions sont proposées dans la littérature visant à rapprocher ou à distinguer ces deux concepts [Graziani et al., 2023]. Nous nous limitons aux définitions de Markus et al. [2021] afin d'être alignés avec les articles présentés dans ce chapitre :

“An AI system is explainable if the task model is intrinsically interpretable or if the non-interpretable task model is complemented with an interpretable and faithful explanation.”

“An explanation is interpretable if the explanation is unambiguous, i.e., it provides a single rationale that is similar for similar instances, and if the explanation is not too complex, i.e., it is presented in a compact form.”

Imaginons un monde où RecForest est déployé pour aider les médecins à diagnostiquer un stade de santé à partir de la prédiction d'événements multiples. Malgré l'importance des variables qui donne une idée plutôt imprécise de la justification des prédictions, l'algorithme n'est pas construit pour fournir de manière inhérente une explication pour justifier sa prédiction [Ribeiro et al., 2016]. Dans le domaine médical, où chaque décision peut avoir des conséquences sur la vie des patients, l'absence d'interprétabilité de tels modèles est plus qu'une simple préoccupation académique [Doshi-Velez and Kim, 2017]. C'est une question de confiance (au sens *fairness* en anglais), de responsabilité, et de sécurité des patients.

Sans une compréhension claire des mécanismes sous-jacents qui guident les décisions du modèle, les praticiens peuvent être réticents à adopter ces technologies [Liao et al., 2020]. Les autorités de santé ont bien compris cet enjeu d'importance majeure dans la régulation

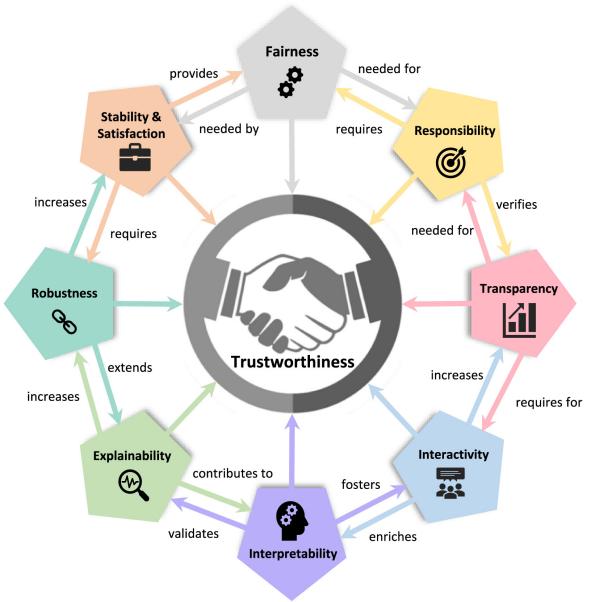


FIGURE IV.1 – Concepts menant à la confiance (Source : Ali et al. [2023])

des algorithmes, pouvant être qualifiés de dispositifs médicaux embarquant de l’IA (*artificial intelligence-based medical device, AI-MDs*) [Gerke et al., 2020]. L’IA Act, récemment voté au Parlement Européen, est une tentative de réponse harmonieuse aux problématiques éthiques basée sur une taxonomie des niveaux de risques [Down and Act, 2021].

Ainsi, le champ de l’IA explicable (*explainable artificial intelligence*) s’est développé afin de répondre au besoin grandissant d’être en mesure d’expliquer les décisions prises par les algorithmes d’IA [Ali et al., 2023]. L’explicabilité peut se définir comme le processus consistant à élucider ou à révéler les mécanismes de prise de décision des modèles. Ainsi, il s’agit de la capacité à comprendre *pourquoi* les modèles d’IA prennent leurs décisions [Markus et al., 2021]. Si l’interprétabilité relève de la compréhension du fonctionnement sous-jacent d’un modèle d’IA, l’explicabilité vise à comprendre *comment* les algorithmes prennent leurs décisions [Markus et al., 2021]. Une méthode interprétable est alors jugée explicable si ses opérations ayant donné lieu à une prédiction sont compréhensibles pour un utilisateur humain. La Figure IV.1 tirée de Ali et al. [2023] représente les caractéristiques clés nécessaires à la confiance envers un modèle d’IA, ainsi que la place qu’y occupent l’explicabilité et l’interprétabilité. De cette façon, l’explicabilité contribue à l’interprétabilité, qui elle-même favorise l’interactivité avec l’algorithme.

La Section IV.1 présente un état des lieux des critères des dispositifs médicaux avec IA dans les recommandations publiées par huit agences d’évaluation des technologies de santé. On y retrouve les notions de performance, d’interprétabilité et d’explicabilité. Dans l’article intitulé “*Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence Based Health Technologies : What Healthcare Stakeholders Need to Know*”, nous avons proposé de préciser ces notions, ainsi que de détailler les outils pour les évaluer.

La Section IV.1 fait brièvement mention de l’usage de l’interprétabilité pour les problématiques de survie. Puisque cette thématique de recherche est également au cœur de mon travail de thèse, nous avons souhaité porter un intérêt particulier à l’intersection des deux.

L'interprétabilité en survie est un sujet encore relativement peu exploré. De cette façon, nous avons proposé un article intitulé "*Bridging Interpretability and Survival Endpoints in Health Technology Assessment*" se focalisant sur des algorithmes de survie communément utilisés afin d'exhiber leur fonctionnement. Nous avons proposé un article à visée pédagogique, en Section IV.2.

L'interprétabilité des modèles relatifs au cadre de survie étendu aux événements récurrents mériterait cependant plus d'attention. C'est pourquoi nous proposons une nouvelle approche allant dans cette direction en Section IV.3.

IV.1 Importance des critères d'interprétabilité et d'explicabilité pour les dispositifs médicaux

IV.1.1 Contexte des dispositifs médicaux

Les AI-MDs désignent le recours aux algorithmes à des fins d'exécution de tâches humaines au moyen de technologies de santé [Farah et al., 2023a]. Leur remboursement est une question délicate et complexe, qui dépend des modalités nationales d'accès au marché [Unsworth et al., 2022]. En Europe, les stratégies d'accès au marché et de remboursement des AI-MDs varient d'un pays à l'autre. L'Allemagne est pionnière, avec une procédure accélérée pour le remboursement des innovations en matière de santé numérique par l'assurance maladie, par le biais du répertoire des applications de santé numérique (*digital health applications*). La France a mis en place son propre processus d'évaluation et de négociation pour les AI-MDs¹ :

- Le programme d'innovation prévoit un financement anticipé des études cliniques des médicaments [Adenot et al., 2020];
- La prise en charge anticipée permet d'accéder aux produits avant leur soumission pendant un an;
- La prise en charge anticipée permet un remboursement anticipé avant autorisation, sur la base de preuves cliniques d'efficacité et de sécurité.

Au moment de l'évaluation de ces preuves, la haute autorité de santé (HAS) a défini 42 critères, classés en quatre catégories. La quatrième catégorie s'articule autour des caractéristiques fonctionnelles de l'algorithme et comprend, outre la performance de l'algorithme, les critères d'explicabilité et d'interprétabilité. L'explicabilité et l'interprétabilité sont en effet au cœur de l'évaluation dans la mesure où elle permet aux professionnels de la santé et aux parties prenantes de mieux comprendre le processus de prise de décision de l'IA [Muehlematter et al., 2021]. Cette compréhension est essentielle pour plusieurs raisons. D'abord, il est important de comprendre le processus décisionnel de l'algorithme pour évaluer la sécurité et l'efficacité du dispositif. Cela peut permettre d'identifier d'éventuels biais ou erreurs dans les algorithmes d'IA qui pourraient conduire à des erreurs de prédiction

1. Article en ligne HAS 2023, Dispositifs médicaux numériques : création à la HAS d'un guichet unique pour une évaluation transversale

IV.1. Importance des critères d’interprétabilité et d’explicabilité pour les dispositifs médicaux

[Dimanov et al., 2020, Slack et al., 2020]. Par ailleurs, les cliniciens et les patients sont plus enclins à faire confiance et à utiliser les AI-MDs s’ils peuvent comprendre comment les décisions sont prises [LaRosa and Danks, 2018]. On peut également noter que l’interprétabilité peut aider les développeurs à affiner les algorithmes d’IA en identifiant les facteurs expliquant de faibles performances, induisant ainsi un processus continu d’amélioration de l’outil technologique.

Selon le rapport HAS 2002, "les professionnels de santé peuvent utiliser des dispositifs médicaux numériques dans le cadre d’un acte médical sans être pleinement éclairés de leurs performances ou de leurs limites ou, *a contrario*, être réticents à leur utilisation pour ces mêmes raisons".²

IV.1.2 Évaluation de performance, d’interprétabilité et d’explicabilité pour les dispositifs médicaux

L’article en premier co-autorat, intitulé "*Assessment of Performance, Interpretability, and Explainability in Artificial IntelligenceBased Health Technologies : What Healthcare Stakeholders Need to Know*", a été publié dans le journal *Mayo Clinic Proceedings : Digital Health* en juin 2023. Ce travail résulte d’une collaboration avec Line Farah, pharmacienne et économiste de la santé à la Délégation ministérielle au numérique en santé. Pour sa thèse de science, L. Farah cherche à construire un nouveau modèle médico-économique pour évaluer les AI-MDs. Notre collaboration a débuté en constatant le manque de clarté dans les critères d’évaluation des autorités de santé. Nous avons donc entrepris de recenser l’ensemble des critères d’évaluation des autorités de santé à travers le monde et de les mettre en relation avec les outils techniques de la statistique et de la science des données. Notre objectif était de fournir une lecture accessible à tous les acteurs (*stakeholders*), afin que chaque personne impliquée dans l’élaboration d’un AI-MD, du développeur à l’utilisateur, puisse comprendre les prérequis essentiels pour l’évaluation des AI-MDs.

Les points clés de notre travail sont les suivants :

- L’état des lieux des autorités de santé dans le monde souligne l’importance de l’évaluation des performances, de l’interprétabilité et de l’explicabilité pour renforcer la confiance des AI-MDs ;
- Pour l’évaluation rigoureuse, nous reprenons point par point les critères de performance, d’explicabilité et d’interprétabilité en fournissant les outils et méthodes existant pour aider à comprendre comment et pourquoi les algorithmes fonctionnent ;
- Le besoin de standardisation dans les processus d’accès au marché pour harmoniser les processus d’évaluation et ainsi faciliter l’accès aux patients.

Je tiens à remercier une nouvelle fois Line Farah, Nicolas Martelli, et Pr Isabelle Borget pour cette collaboration très enrichissante. Pour en savoir plus sur les AI-MDs, je recommande le podcast passionnant de Nicolas Martelli intitulé "*Deux mots sur les dispositifs médicaux*".³

2. Rapport d’élaboration HAS 2022, Intégration des dispositifs médicaux numériques à usage professionnel dans la pratique

3. Disponible [ici](#).

Assessment of Performance, Interpretability, and Explainability in Artificial Intelligence-Based Health Technologies: What Healthcare Stakeholders Need to Know

Line Farah, PharmD; Juliette M. Murris, MSc; Isabelle Borget, PhD, PharmD; Agathe Guilloux, PhD; Nicolas M. Martelli, PhD, PharmD; and Sandrine I.M. Katsahian, MD, PhD

Abstract

This review aimed to specify different concepts that are essential to the development of medical devices (MDs) with artificial intelligence (AI) (AI-based MDs) and shed light on how algorithm performance, interpretability, and explainability are key assets. First, a literature review was performed to determine the key criteria needed for a health technology assessment of AI-based MDs in the existing guidelines. Then, we analyzed the existing assessment methodologies of the different criteria selected after the literature review. The scoping review revealed that health technology assessment agencies have highlighted different criteria, with 3 important ones to reinforce confidence in AI-based MDs: performance, interpretability, and explainability. We give recommendations on how and when to evaluate performance on the basis of the model structure and available data. In addition, should interpretability and explainability be difficult to define mathematically, we describe existing ways to support their evaluation. We also provide a decision support flowchart to identify the anticipated regulatory requirements for the development and assessment of AI-based MDs. The importance of explainability and interpretability techniques in health technology assessment agencies is increasing to hold stakeholders more accountable for the decisions made by AI-based MDs. The identification of 3 main assessment criteria for AI-based MDs according to health technology assessment guidelines led us to propose a set of tools and methods to help understand how and why machine learning algorithms work as well as their predictions.

© 2023 THE AUTHORS. Published by Elsevier Inc on behalf of Mayo Foundation for Medical Education and Research. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>) ■ Mayo Clin Proc Digital Health 2023;1(2):120–138



From the Groupe de Recherche et d'accueil en Droit et Economie de la Santé Department (L.F., I.B., N.M.M.), University Paris-Saclay, Orsay, France; Innovation Center for Medical Devices (L.F.), Délégation à la Recherche Clinique et à l'Innovation, Hôpital Foch, Suresnes, France; Inserm (J.M.M., S.I.M.K.), Centre de Recherche des Cordeliers, Sorbonne

Affiliations continued at the end of this article.

Understanding of algorithms in general and in artificial intelligence (AI) in healthcare has become an essential criterion following the new regulation processes for AI (AI Act), data (General Data Protection Regulation), and medical devices (MDs) (Medical Device Regulation) in Europe. Among these, the AI Act is the first regulation to divide applications of AI into different risk categories: (1) unacceptable risk, (2) high risk, and (3) low or minimal risk.¹

In medicine, AI can be used not only in combination with an MD but also as an MD by itself. In fact, MDs are defined in the

European Medical Device Regulation as “any instrument, apparatus, appliance, software, implant, reagent, material, or other article intended by the manufacturer to be used, alone or in combination, for human beings for specific medical purposes.”² Artificial intelligence-based MDs are health technologies employed to improve human capabilities for several applications, including prediction or identification of diseases, data classification or analysis for disease outbreaks, optimization of medical therapy, or disease diagnosis.² The Food and Drug Administration (FDA) in the United States defines an AI-based MD as

IV.1. Importance des critères d’interprétabilité et d’explicabilité pour les dispositifs médicaux

ASSESSMENT OF AI-BASED MDS

“Software as a Medical Device” when the algorithm is intended to prevent, diagnose, treat, mitigate, or cure diseases.³

An increase in approved AI-based MDs has been recorded, with 222 devices in the United States and 240 devices in Europe between 2015 and 2020.⁴ Methodological frameworks are designed by health technology assessment (HTA) agencies to assess these technologies, and these agencies aim to evaluate them using a standardized method through multiple domains, such as safety, clinical effectiveness, costs and economic evaluation, organizational aspects, patients, and social and legal aspects.⁵ The assessment of AI-based MDs is performed by health technology agencies, such as Haute autorité de santé (HAS) in France, the National Institute for Health and Care Excellence in the United Kingdom, or FDA in the United States. In addition to the usual technical, clinical, and health economics criteria used for MD assessment, the need for specific criteria to assess AI technologies in healthcare has been highlighted.⁶

For instance, in France, HAS has defined 42 criteria, classified into 4 categories, to assess AI-based MDs. The fourth category, on functional characteristics, includes, in addition to algorithm performance, the criteria of explainability and interpretability. In the United States, FDA takes into account either the real-world or the human–AI team performance, the latter of which relates to how interpretable the model outputs are for humans, with an emphasis on the performance of the model. The performance of AI technology is often prioritized; however, an inability to understand the algorithms raises serious concerns in terms of fairness, ethics, and trust, and both interpretability and explainability refer to this capacity to understand algorithms.

From a healthcare perspective, the opacity of some AI models led to a decline in adoption by healthcare professionals. Several authors have highlighted the need for making these AI-based MDs more interpretable; however, the authors have also insisted on the explainability for trustworthy AI.^{7–9} On the contrary, Ghassemi et al¹⁰ advocated the rigorous internal and external validation of AI models owing to the lack of suitable explainability methods. However, these notions seem to be important

ARTICLE HIGHLIGHTS

- The level of confidence in artificial intelligence (AI)-based medical devices relies on transparency (interpretability and explainability of outputs) and ethics (in terms of trustworthiness and regulation).
- To provide interpretability, we identified that metrics and methodologies for “explainable AI” need to be associated with ethical and legal analysis.
- Specific explainability and interpretability assessment by regulators increased and led to stakeholders being increasingly held accountable for the decisions made by AI-based medical devices.
- Acceptable standards for explainability are context-dependent and reliant on the risks in the clinical scenario.
- Raising awareness about these concepts is essential for their widespread adoption and to answer ethical questions.

to develop trustworthy AI using several principles proposed by Hasani et al,¹¹ such as transparency, explainability, technical robustness, or stakeholder involvement. Thus, there is a growing need for appropriate assessment methodologies for explainable and interpretable AI-based MDs.¹²

Therefore, the aim of this study was to specify the different concepts that are essential for the development of AI-based MDs and to shed light on how performance, interpretability, and explainability are key in the development of health technology models.

To meet this objective, we aimed to address these 3 fundamental aspects of the evaluation of all criteria involved in the development and use of such technologies. After presenting AI ecosystems in healthcare with a focus on HTA agencies (section 1: State of the art of the assessment of AI-based MDs by HTA agencies), we will examine how the performance of AI-based MDs is measured (section 2: How to measure the performance of AI-based MDs?) and then provide elements for integrating interpretability and explainability issues into the core of algorithm development (section 3: How can we evaluate interpretability and explainability in AI health technologies?). Finally, we will discuss the major relevance of these notions for all stakeholders and offer a decision-making tool to

CHAPITRE IV Interprétabilité, santé et survie : vers une utilisation plus transparente des algorithmes d'IA

122

TABLE 1. Identification of Key Specific Criteria for Artificial Intelligence-Based Medical Device Assessment in the Reviewed Guidelines on Health Technology Assessment of Artificial Intelligence Technologies^a

Country	Guidelines (date)	Criteria ^b	Description ^b	Reference, year ^c
Finland	Digi-HTA: Health technology assessment framework for digital healthcare services (2019)	AI	Capacity of the staff to understand the operational logic of AI? (interpretability) Transparency of the conclusions and decisions of the AI solution, that is, understanding of medical staff about the origin of the decisions (explainability)	Haverinen et al, 2019⁷⁸
		Technical stability	The testing process and company's process for handling error messages	
		Cost	Costs of using the product for a healthcare customer	
		Effectiveness	The product provides clinical benefits to the end users by improving their behavior related to their own health	
		Clinical safety	Risks, possible side effects, or other undesirable effects associated with using the product; research evidence available related to clinical safety	
		Data security	Information security and data protection requirements	
		Usability and accessibility	The process of the company to continue to evaluate and develop accessibility. Product compatibility with usability guidelines (if applicable)	
		Interoperability	The product interfaces into the website and software, the healthcare services, and electronic patient records	
		Robotics	Safety risks for healthcare personnel or customers and the robot's design to avoid them	
		Purpose	Specify the benefit of the information provided or decisions made by machine learning processes	Haute autorité de santé, 2020⁸¹
France	Liste des produits et prestations remboursables (LPPR) Guide: Dossier submission to the Medical Device and	Data	Describe samples used, input data involved for initial model learning or relearning, and input data involved in decision making	

Continued on next page

IV.1. Importance des critères d'interprétabilité et d'explicabilité pour les dispositifs médicaux

TABLE 1. Continued

Country	Guidelines (date)	Criteria ^b	Description ^a	Reference, year ^c
	Health Technology Evaluation Committee (2020)	Model Functional characteristics	Describe training, validation, and testing before and after MD deployment Performance and qualification, system robustness and resilience, explainability, and interpretability	
Australia	Clinician checklist for assessing suitability of machine learning applications in healthcare (2021)	Purpose Data Performance Interpretability, explainability, and explicability Workflow Patient harm Ethical, legal, and social	Purpose of the algorithm The quality of the data used to train the algorithm: accurate and free of bias, standardized and interoperable, and sufficient quantity of data Algorithm performance Algorithm transferability to new clinical settings Evidence generation related to the algorithm's impact on patient care improvement and outcomes Clinically intelligible outputs of the algorithm: interpretability and explainability Algorithm fitting into and complementing current workflows Avoiding patient harm Ethical, legal, or social concerns raised by the algorithm	Scott et al, ⁶³ 2021
United States, Canada, United Kingdom	Good Machine Learning Practice for Medical Device Development: Guiding Principles	Product life cycle Security practices Clinical study participants and datasets	Understanding of a model's intended integration into clinical workflow (interpretability and explicability) Balance between desired benefits and associated patient risks Safety, effectiveness, and clinically meaningful needs addressed over the lifecycle of the device Good software engineering practices, data quality assurance, data management, and cybersecurity practices Data collection: relevant characteristics of the intended patient population sufficiently represented in a sample of adequate size in the clinical study and training and test	Korean Minis US Food and Drug Administration, Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency, 2021 ⁸²

Continued on next page

CHAPITRE IV Interprétabilité, santé et survie : vers une utilisation plus transparente des algorithmes d'IA

124

Country	Guidelines (date)	Criteria ^b	Description ^c	Reference, year ^d
		Training datasets/test sets	datasets, management of bias, promotion of appropriate and generalizable performance across the intended patient population	
		Selected reference datasets	Training and test datasets were selected and maintained to be appropriately independent of one another	
		Model design and intended use of the device	Accepted, best available methods for developing a reference dataset; accepted reference datasets in model development and testing that promote and demonstrate model robustness and generalizability	
		Performance of the human–AI team	A model design supporting mitigation of known risks, such as overfitting, performance degradation, and security risks. Clinical benefits and risks are well understood, used to derive clinically meaningful performance goals for testing the product can safely achieve its intended use	
		Device performance	Model as a "human in the loop," consideration of human factors and the human interpretability of the model outputs are addressed with emphasis on the performance of the human–AI team	
		Clear and essential information for users	Statistically sound test plans developed and executed to generate clinically relevant device performance information independently of the training dataset	
		Performance and retraining risks	Users are provided with ready access to clear, contextually relevant information that is appropriate for the intended audience (such as healthcare providers or patients), including the product's intended use and indications for use, performance of the model for appropriate subgroups, user interface interpretation (interpretability), and clinical workflow integration of the model.	
			Capability to be monitored in "real-world" use with a focus on maintained or improved safety and performance	

Continued on next page

IV.1. Importance des critères d’interprétabilité et d’explicabilité pour les dispositifs médicaux

TABLE 1. Continued

Country	Guidelines (date)	Criteria ^b	Description ^a	Reference, year ^c
Europe (Greece)	Presenting AI, DL, and ML studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research checklist proposal (2021)	Data Performance Ethical considerations and methodological biases	Outcome imbalances/training and testing/missing data/overfitting Evaluation metrics The confusion table Measuring performance Performance curves and AUC Image segmentation or localization Continuous measurements Multiple measurements Data and privacy Bias and fairness Informed consent and autonomy Safety and interpretability Responsibility and liability	Olczak et al, 2021 ⁷⁹
United States	Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist (2020)	Study design Data and optimization Model performance Model examination/assessment in clinical practice	Clarity of the design Characteristics of the cohorts (training and test sets) and representativity of real-world clinical settings Comparator Origin of the data, data quality, independence between training and test sets, data quantity, targeted population, input data type Primary metric selected to evaluate algorithm performance (eg, AUC, F-score, etc) Performance comparison between baseline and proposed model with the appropriate statistical significance Explainability: clinically intelligible outputs of the algorithm and explainability of the algorithm Algorithm fitting into and/or complementing current clinical workflows Ethical, legal, or social concerns raised by the algorithm	Norgeot et al, 2020 ⁸⁰

Continued on next page

CHAPITRE IV Interprétabilité, santé et survie : vers une utilisation plus transparente des algorithmes d'IA

126

TABLE 1. Continued				
Country	Guidelines (date)	Criteria ^b	Description ^b	Reference, year ^c
South Korea	Guideline on Review and Approval of Artificial Intelligence (AI) and big data-based Medical Devices (For Industry)	Characteristics Performance	Medical device classification criteria Validate the essential requirements and clinical effectiveness Clinical validation (clinical performance and efficacy)	Korean Ministry of Food and Drug Safety, 2020³³

^aAI, artificial intelligence; AUC, area under the curve; DP, deep learning; HTA, health technology assessment; MD, medical device; ML, machine learning.

^bThe lines highlighted in bold correspond to the specific criteria related to artificial intelligence–based medical devices in each guideline.

^cOur review selected 7 articles (out of 64), including guidelines on health technology assessment of artificial intelligence–based medical devices from 8 countries. For each guideline, 3 criteria are highlighted in green: performance, interpretability, and explainability.

IV.1. Importance des critères d’interprétabilité et d’explicabilité pour les dispositifs médicaux

ASSESSMENT OF AI-BASED MDS

facilitate the HTA process (section 4: Discussion: to what extent can the explainability and interpretability of AI be as useful as performance for HTA?).

STATE OF THE ART OF THE ASSESSMENT OF AI-BASED MDS BY HTA AGENCIES

The AI ecosystem involves a large diversity of stakeholders with heterogeneous competencies and knowledge essential to tackle the development, validation, assessment, and deployment of AI-based MDs. In addition to those who usually contribute to the creation of MDs and their assessment, the AI health sector includes new stakeholders specialized in data, information technology, and engineering: AI public research institutes (*Supplemental Figure 1*, available online at <https://www.mcpdigitalhealth.org/>). The AI healthcare area gathers multiple actors from different areas (health, information technology, robotics, the tech industry, and ethics). Therefore, a crucial step in assessing these technologies is to identify the various stakeholders and understand a common taxonomy and the key notions to bridge the gap between them, thereby guaranteeing a common basis for assessments. Assessing the requirements of different international HTA agencies related to the evaluation of AI-based MDs shows that, in addition to the usual HTA criteria, such as performance and safety, the need for interpretability is crucial for clinical diagnosis, prevention, or treatment. The need for explainability is also important to comply with the “right to explanation” provided by the European General Data Protection Regulation.

The European guidelines for trustworthy AI include the principles of explainability and interpretability in addition to fairness and prevention of harm.¹³

Objective and Methods

A literature review, following Preferred Reporting Items for Systematic Reviews and Meta-Analyses recommendations, was performed to highlight the specific key criteria needed for an HTA of AI-based MDs (review protocol provided in Supplemental Material, available online at <https://www.mcpdigitalhealth.org/>).

Results

Of 64 articles, 7 were selected after full-text screening. They included guidelines on HTA of AI-based MDs from 8 countries. For each guideline, the following 3 criteria were highlighted: performance, interpretability, and explainability (*Table 1*). Nevertheless, no methodology has been proposed to measure these criteria.

On the one hand, some HTA agencies only focus on interpretability. On the other hand, other agencies, such as HAS in France, highlight these notions as essential to be defined in the reimbursement dossiers of AI-based MDs that are submitted by companies. Interpretability is an important criterion; assessors ask for the parameters that influence the decision and for the methods used to identify them. For explainability, this agency focuses on understanding the factors that lead to the decision-making process.

Even when there is no legal obligation, it is important for HTA agencies and clinicians to be able to justify their decision-making process to patients.^{14–16} Explainability allows comparisons of algorithms with current recommendations; however, explaining how the predictions are derived can be a time-consuming process and, hence, could be suggested in specific situations. For AI-based MDs with high risks for patient safety (for instance, those that impact morbimortality), explanations are vital. In addition, explanations can be required when an algorithm’s clinical performance has not yet been proven.¹⁷

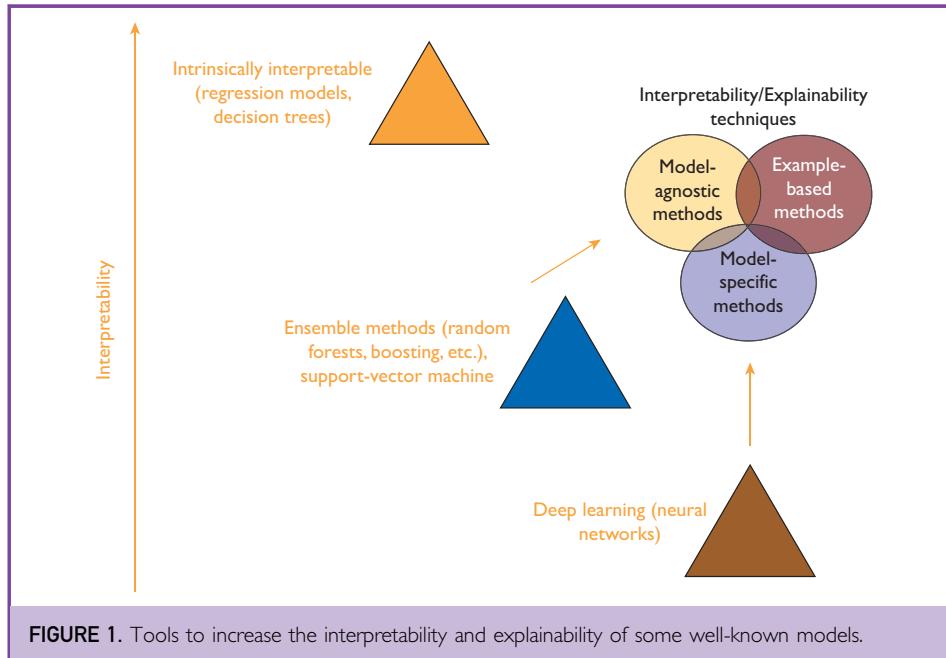
Therefore, the next part of this study focused on the methods and tools used to assess performance, interpretability, and explainability to answer the need in an HTA process.

HOW TO MEASURE THE PERFORMANCE OF AI-BASED MDS?

In this section, we outline which tools are available for measuring the performance of an algorithm and how to use them.

Definition

Performance measurement consists of evaluating the error of the model (hence the reliability) by assessing the difference between predicted and observed data. It is usually



based on a score, an error metric for which lower values indicate better results.

Tools for Measurement

Various metrics exist to evaluate the performance of a model. Each meets different purposes according to the global objective of the modeling strategy (regression or classification).¹⁸ The list of metrics presented in the *Supplemental Table* (available online at <https://www.mcpdigitalhealth.org/>) is not exhaustive because the number of metrics is currently exploding to meet the needs of new applications. Some metrics are based on mean differences between estimated and true values (such as mean square-errors and R^2 -like measures; this is called calibration. Besides, discrimination describes the capacity of algorithm estimates to distinguish between individual observations, which does not imply to know whether the output is true.¹⁹ In any case, all metrics are subject to some limitations that should be outlined in the development of AI-based MDs (for further literature, see the online book from Biecek and Burzykowski²⁰).

Evaluation

The goal is to learn an algorithm that best maps input data to the outcome. The learning

process consists of 3 main components: the space of assumptions, training data, and the loss function. The space of assumptions describes the overall authorized set for the algorithm. The training data include the set of input data and outcome used by the learning algorithm to adjust for the best parameters. The loss function measures the error between true and predicted outcomes. The relationship between the complexity of the space of assumptions, the size of the training data and the generalization error of the learned algorithm defines the bias-variance trade-off, which is both a fundamental concept and a key challenge. The generalization error is the difference between the expected error of the learned function on new data and the training error on the data used to learn the function. We assume in this section that training and test data are independent and identically distributed.

The Bias-Variance Trade-off. It is generally accepted that evaluating the algorithm on the same data it has learned on is a methodological mistake.^{21,22} Overfitting is when a model is able to predict perfectly well on fitted data but not on yet unseen data. When a model overfits, it typically leads to higher prediction

IV.1. Importance des critères d'interprétabilité et d'explicabilité pour les dispositifs médicaux

ASSESSMENT OF AI-BASED MDS

TABLE 2. Models with a High Interpretability Level					
Algorithm	Linear explanation	Monotone relationship	Task	Interpretable coefficient	Examples in healthcare (specialty, pathology, intended use of algorithm)
Linear regression	Yes	Yes	Linear	Linear coefficient	<ul style="list-style-type: none"> ● Endocrinology, diabetes, prediction of severity (Butt et al, 2021)⁸⁴ ● Genetics, prediction of gene expression (Zeng et al, 2017)⁸⁵
Logistic regression	No	Yes	Classification	Odds ratio	<ul style="list-style-type: none"> ● Gastroenterology, fatty liver, prediction of disease in the general population (Bedogni et al, 2006)⁸⁶ ● Radiology, breast cancer, computer-aided diagnostic system (Nemat et al, 2018)⁸⁷
Cox regression model	Yes	Yes	Time to event	Hazard ratios	<ul style="list-style-type: none"> ● Cardiology, heart failure, prediction of mortality (Cheng et al, 2017)⁸⁸ ● Oncology, gastric cancer, prognosis prediction (Wei et al, 2021)⁸⁹
Decision trees	No	Yes	All	Nodes	<ul style="list-style-type: none"> ● Psychiatry, mental disorders, risk prediction (Van Hoffen et al, 2020)⁹⁰ ● Cardiology, malignant ventricular arrhythmia, diagnosis prediction (Mandala et al, 2020)⁹¹

errors because the model is too specific for the data and is barely generalizable. Predictions for individuals already in the database will naturally match with themselves, and therefore, there will be no prediction error. However, should there be small fluctuations in the training data, some error would be introduced by the sensitivity of the algorithm. This is called variance, which is highly dependent on small variations within the training sample. High variance with great capacity in fitting training data leads to overfit, whereas small variance has a small capacity to fit the training data and will underfit.

The opposite problem is bias, when error is introduced by approximating a complex problem using a simpler algorithm. High bias has a small capacity to fit the training data and will underfit, whereas low bias with a great capacity in fitting training data leads to overfit. The bias-variance trade-off is the balance between these 2 sources of

error. A good trade-off point is achieved when the algorithm has low bias and low variance, which corresponds to a good balance between fitting the training data and generalizing to new data.

We have listed some best practices around the bias-variance trade-off and summarized them in [Supplemental Figure 2](#) (available online at <https://www.mcpdigitalhealth.org/>).

Which Data to Use, When, and How. A common practice to avoid overfitting is to evaluate the algorithm on a random sample held outside of the data used to train it.²³ The main idea is that the data on which the predictive model is applied, known as the test data, should be different from the training data. A systematic way to evaluate the aforementioned trade-off is an iterative split called cross-validation, in which the dataset is divided into different subsets and the model's error is measured on each subset.²⁴

TABLE 3. Pros and Cons of Methods Serving Interpretability and Explainability ^a						
	Easy to understand		Computation time	Data type	Limitations	Examples in healthcare (specialty, pathology, intended use of algorithm)
	For engineers ^b	For end users ^c				
Feature importance, SHAP, LIME	Yes	Intermediate, rarely shown	Low	Image, text, or tabular	Feature importance—sensitive to multicollinearity SHAP—sensitive to categorical variables and feature interactions LIME—difficulty in setting distance threshold	<ul style="list-style-type: none"> Cardiology, cardiac surgery—associated acute kidney injury, prediction (Tseng et al, 2020)⁹² Computational neuroscience, brain age prediction (Lombardi et al, 2021)⁹³ Pediatric medicine, organ transplantation, prediction of posttransplant health outcomes (Kilian et al, 2011)⁹⁴
Counterfactual explanations	Yes	Yes	High	Image, text, and mainly tabular	Difficulty in generating feasible and actionable explanations. Causal constraints	<ul style="list-style-type: none"> Neurology, prediction errors in the human brain (Boorman et al, 2011)⁹⁵

^aLIME, Local Interpretable Model-agnostic Explanations; SHAP = Shapley Additive exPlanations.

^bThe engineers include, but are not limited to, developers, data scientists, and statisticians.

^cThe end users include, but are not limited to, healthcare professionals, decision-makers, medical experts, and patients.

Settings of the algorithm are commonly called hyperparameters and drive the inherent complexity in controlling the learning process.²⁵ Hyperparameter optimization hence allows to find the optimal complexity of the algorithm that performs best both on the training and unseen data. The idea is to find the hyperparameter combinations that optimize the cross-validation metric. More inputs on hyperparameter optimization are given in the Supplemental Material (available online at <https://www.mcpdigitalhealth.org/>).

The final evaluation can be performed either on a test set previously held out or on external data.

HOW CAN WE EVALUATE INTERPRETABILITY AND EXPLAINABILITY IN AI HEALTH TECHNOLOGIES?

The following methods are intended to provide an understanding of model prediction and behavior as part of an evaluation dossier. They do not cover how the methods can be used to debug or improve a model. Therefore,

interpretability and explainability are ideals to be achieved, rather than assets.

Definitions

Artificial intelligence raises numerous questions because of its opaque decision-making process. Both interpretability and explainability aim to help understand algorithms and answer user-based questions regarding AI's input, output, and performance (such as why, how, what if, and why not).²⁶

Existing definitions for explainability and interpretability have been previously and widely discussed in the literature, and it seems that there is no clear taxonomy of concepts.^{17,26–28} Even though some authors consider the 2 concepts to be similar, some HTA agencies distinguish between them during the evaluation process (Table 1). Following are the definitions proposed by Markus et al¹⁷ in 2021:

1. “An AI system is explainable if the task model is intrinsically interpretable or if the non-interpretable task model is

IV.1. Importance des critères d'interprétabilité et d'explicabilité pour les dispositifs médicaux

ASSESSMENT OF AI-BASED MDS

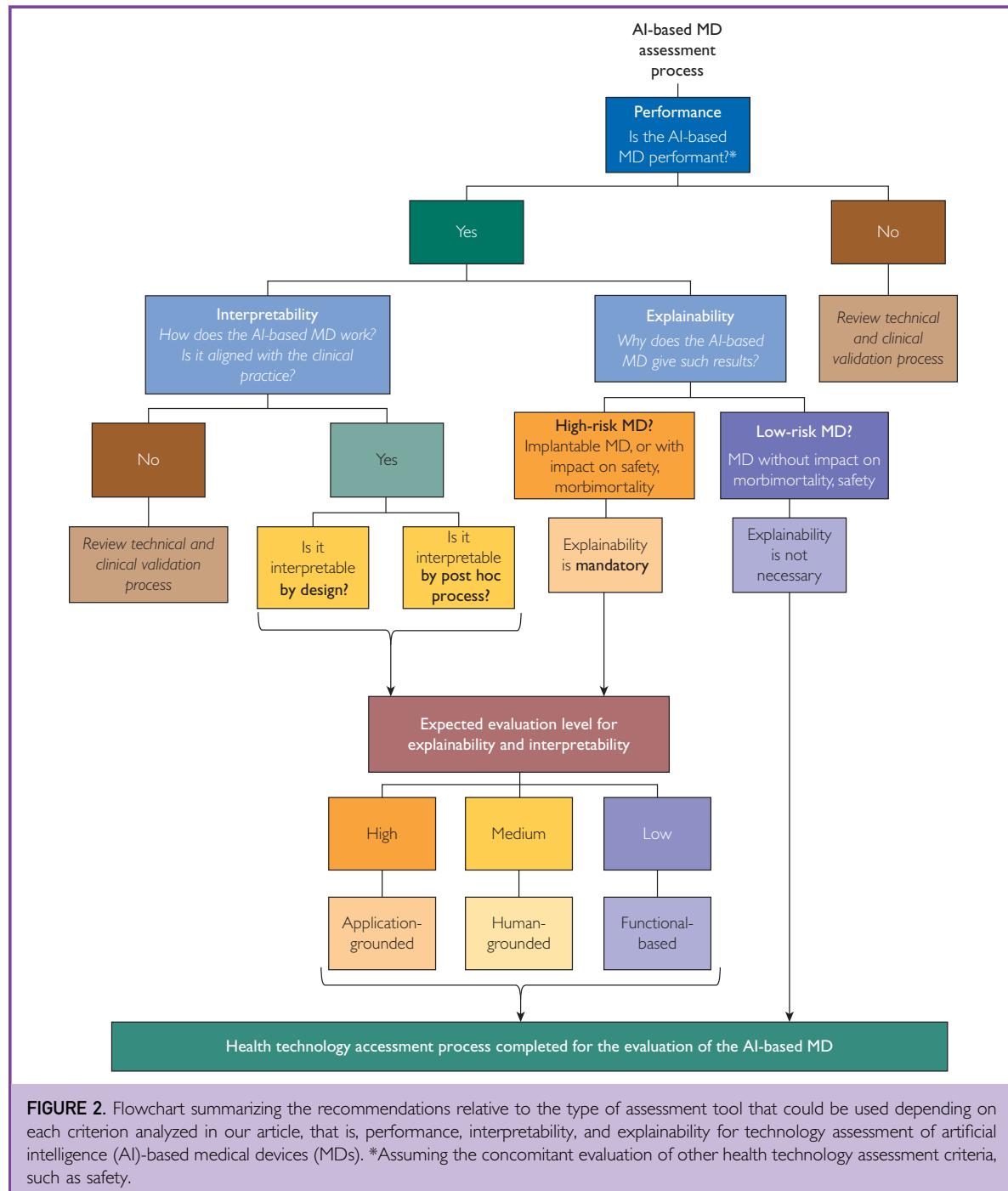


FIGURE 2. Flowchart summarizing the recommendations relative to the type of assessment tool that could be used depending on each criterion analyzed in our article, that is, performance, interpretability, and explainability for technology assessment of artificial intelligence (AI)-based medical devices (MDs). *Assuming the concomitant evaluation of other health technology assessment criteria, such as safety.

- complemented with an interpretable and faithful explanation.”
2. “An explanation is interpretable if the explanation is unambiguous, i.e., it provides a single rationale that is similar for similar instances, and if the explanation is not too complex, i.e., it is presented in a compact form.”

If a task model is interpretable, it is hence very likely to be explainable.

Tools for Measurement

Interpretability is difficult to define mathematically. Although there are many different machine learning (ML) algorithms, not all of them are explainable straightforwardly. However, 3 levels of interpretability (high, medium, and low) have been identified. We extended the figure proposed by Dam et al²⁹ (2018) by including existing tools that can increase the interpretability of the most well-known models while taking into account that black-box algorithms do not necessarily lead to higher performance (Figure 1).³⁰ Figure 1 was not generated from any real data, and the y-axis has no quantification.

Interpretable by Design. Some models are interpretable by design under specific constraints, such as monotonicity, causality, and additivity (Table 2).³¹ They indeed already include internal functioning ready for interpretation, that is, they are intrinsically interpretable. Table 2 also provides relevant examples of application in the healthcare sector. Regression models are tangible equations with interpretable coefficients that can be read as linear coefficients with linear models, odds ratios for logistic regressions, and hazard ratios in Cox models to handle time-to-event data.³² Decision trees include interpretable rules and are greatly adapted to human thinking.^{33,34} Such methods should be the accepted baseline owing to their simple and fast processing and are highly preferable to any extremely complex model.³⁰

Post Hoc Explanations. Ensemble methods (such as random forests or boosting), support-vector machines, or deep neural networks are uninterpretable algorithms. Post hoc explanations can either be global or local.

Global explanations relate to the algorithm’s overall behavior, typically considering the overall importance of the covariates or features, and provide insight into how the algorithm makes predictions on a general basis. Conversely, local explanations refer to explanations at the scale of specific data points, detailing the reasons why the model chose these particular outcomes. Many toolkits and classifications are available in the literature to better describe how such post hoc explanations work.^{35–38}

In this section, we decided to list only the most well-known approaches of post hoc explanation. Because there is a growing need for interpretability to manage the exponential growth of the number of parameters in models, many approaches have been developed recently, and several typologies exist to classify them.^{39–41} Model-specific methods will not be discussed here because they depend highly on the model used for the prediction (such as gradient-based saliency maps, which are typically used for neural networks and imagery and providing each pixel’s importance). The main advantage of model-agnostic methods is that they can be applied in a post hoc manner to any kind of ML model.

Advantages and disadvantages for each method as well as relevant examples in healthcare are provided in Table 3. Overall, any element that can help understand the choices made by the AI algorithm are very welcome (eg, the study by Selvaraju et al⁴²). Methods are yet to be made readily accessible to all stakeholders, from the developer to the end user. Work is currently underway to address this matter.^{43–46}

Evaluation of Performance, Interpretability, and Explainability

According to Ossa et al,⁴⁷ in some cases, fewer explanations are acceptable if the risk-to-benefit ratio is clearly defined. Low-stakes decisions can tolerate less explainable AI as long as the mortality and morbidity risks are limited. In contrast, the diagnosis of a fatal disease requires that the AI algorithm provide doctors and patients with a complete understanding of its decision. The conceptualization of explainability in healthcare seems to be driven by and should focus on the context of clinical implementation. To date, no

IV.1. Importance des critères d’interprétabilité et d’explicabilité pour les dispositifs médicaux

ASSESSMENT OF AI-BASED MDS

consensual approach exists for the evaluation of interpretability. However, Doshi-Velez and Kim⁴⁸ have performed rigorous evaluation of interpretability and explainability, and their findings are outlined in the following sections.

Evaluations Involving Humans. First, application-grounded evaluation ensures that the algorithm performs the task for which it is designed by conducting human experiments. The principle is to involve end users (eg, physicians or radiologists) and show them explanations provided by the algorithm. The second step is to ask them what the machine would do and then present them with the actual output of the machine, working through a real-world example. By giving such tasks, you can quantitatively assess the difference in the performances of the humans and the model. Including both outliers and false assumptions in the algorithm also helps in spotting the expected outcomes. This constitutes a straightforward way of validating the objective, and, hence, the success of the algorithm’s performance. Application-level evaluations are yet to be deployed in healthcare.^{49–51}

Second, human-grounded evaluation is similar to application-grounded evaluation but provides a simpler framework. The people involved are not experts anymore but lay people. Such experiments are typically recommended when objectives are wider than the assessment of interpretability/explainability of an algorithm. They are also cheaper because they do not require the involvement of high-level experts.

Evaluations Not Involving Humans. Functionally grounded evaluation does not involve human intervention. The aim is to formalize the algorithm’s components as an indicator of the quality of the explanation, favoring ease of use and simplicity. For example, a tree with a small depth is preferable to another with a large depth. Easy to formalize, function-based evaluation helps and is a valuable addition to human-based strategies.

Numerous measures to evaluate interpretability and explainability are emerging in the literature, including stability, simplicity, and faithfulness.^{41,52–56} Further guidance is also available elsewhere in the literature.^{17,39}

Notably, the authors agree on the impossibility of fulfilling all properties for “good” explanations. However, human-based experiments are highly recommended whenever possible.

DISCUSSION: TO WHAT EXTENT CAN THE EXPLAINABILITY AND INTERPRETABILITY OF AI BE AS USEFUL AS PERFORMANCE FOR HTA?

To sum up the HTA process of AI-based MDs, we established a flowchart that maps our recommendations toward the type of assessment tool that could be used depending on each criterion analyzed in the present article (Figure 2). We assumed that concomitant evaluation of other HTA criteria, such as safety, would be undertaken at the same time. Performance and interpretability should be evaluated for each category of AI-based MDs, whereas explainability might not be mandatory for low-risk AI-based MDs (in contrast to high-risk MDs), that is, devices with no impact on morbimortality or safety.

Complex Trade-off Between Performance and Interpretability and Explainability

The predictive performance of AI systems is a key issue. However, the importance of explainability depends on the specific AI and its intended use. If explainability is not important and if a black-box model could be acceptable, the model with the best predictive performance is more interesting because explanations can be expensive. When a model has a high level of explainability, the selection of explainable AI methods could be considered.^{8,9,57,58} It is difficult to satisfy all properties of explainability. Holzinger et al⁵⁹ suggested a brief overview of 17 explainable AI methods, including Local Interpretable Model-agnostic Explanations, Anchors, Graph Local Interpretable Model Explanations, Shapley Flow, Textual Explanations of Visual Models, Integrated Gradients, Causal Models, or Meaningful Perturbations. For instance, Arras et al⁶⁰ proposed to adapt the Layer-wise Relevance Propagation technique used for explaining the predictions of feed-forward networks to the Long Short Term Memory architecture used for sequential data modeling in healthcare. Thus, as the developer of an AI system, it is important to establish the relative importance of explainability compared with

predictive performance and what is desired by end users of the AI system.

Performance, Interpretability, and Explainability: Key Requirements for a Trustworthy AI

At an international level, healthcare professionals seem to have difficulties trusting AI-based MDs. A study by Oh et al⁶¹ highlighted that only 5.9% of Korean doctors reported having good familiarity with AI. Among 999 Japanese physicians interviewed, only 44.7% expressed an intention to use AI-driven medicine.⁶² Another study showed that companies require more data, funding, and regulatory certainty, and clinicians and patients insist on trustworthy AI-based MDs.⁶³

There are several issues that can decrease physicians' trust in AI in their clinical practice, such as the low number of randomized clinical trials assessing the performance of AI-based MDs, the lack of transparency within these technologies, the risk of inequity introduced by AI biases, and insufficient regulatory clarity.¹² The need for trustworthy AI exponentially increased in the healthcare ecosystem with the several considerations in medical imaging, as Hasani et al¹¹ highlighted with a proposition of 14 core principles to promote trustworthy AI-based MDs in medical imaging, such as transparency, explainability, technical robustness, or stakeholders involvement. Holzinger et al⁶⁴ insisted on bridging the gap between research and practical applications in the context of future trustworthy medical AI with human-centered AI design methods.

According to Ossa et al,⁴⁷ explainability needs to be sufficient but not exhaustive for doctors and patients. The acceptable standards for explainability are context-dependent and rely on the risks of the clinical scenario, and factors that form part of AI's explainability include usefulness and uncertainty, risk of bias, responsibility attribution, and the AI's involvement in decision making.

To provide interpretability, methodologies for explainable AI need to be associated with ethical and legal analysis.^{65–69} For instance, Currie et al⁷⁰ confirmed the need of addressing the ethical and legal challenge of AI in nuclear medicine. Naik et al⁷¹ showed that as we rely more on AI for decision making, it

becomes important to ensure that they are made ethically and free from unjust biases to tackle the responsible AI notion with devices that are transparent, explainable, and accountable.

A Regulatory Need Toward Responsible AI

The 3 notions that we covered in this article are also part of the process of creating confidence in AI in healthcare. The level of confidence in an algorithm in fact relies heavily on transparency (interpretability and explainability of outputs) and ethics (in terms of trustworthiness and regulation).⁷²

The work by Liao et al²⁶ led to the identification of diverse motivations based on AI users' needs, such as gaining further insights for decision making, appropriately evaluating algorithm capability, and highlighting the ethical responsibilities of AI products. The lack of explanation for some "black-box" algorithms raises ethical questions, particularly in healthcare.²⁷ Closely related concepts are fairness and ethical AI. Fairness refers to the idea that an algorithm should make predictions that are unbiased and do not discriminate against any group of individuals.⁷³ Ethical AI describes the use and design of an algorithm that are in line with human values and the rights and well-being of individuals.⁶⁵ The relationship between such concepts is that interpretability and explainability can help to strive toward fairness and ethical AI. Providing interpretability and explainability for an algorithm's predictions typically means bringing forward transparency and accountability by detecting (and addressing) potential biases or ethical issues (even though some explanations can hide unfairness, as underlined by Dumanov et al⁷⁴ and Slack et al⁷⁵). In this way, stakeholders can better understand how the algorithm works and can evaluate whether fair and unbiased decisions are made.⁷⁶ The aim of transparency and explainability of AI-based MDs hence contributes to fair and accountable algorithmic decision-making processes.⁷⁷

For these reasons, initiatives are awaited from institutions. For instance, the Confiance.ai program was launched in July 2021 and gathers 13 private and public institutes. Together, they aim to build a trusted AI in

IV.1. Importance des critères d’interprétabilité et d’explicabilité pour les dispositifs médicaux

ASSESSMENT OF AI-BASED MDS

the industry to ensure the reliability, security, and certification of AI-based systems.

CONCLUSION

After the identification of 3 main assessment criteria for AI-based MDs according to HTA guidelines, we provided a set of tools and methods to help understand how and why ML algorithms work as well as their predictions. We also highlighted the increase in the importance of explainability and interpretability techniques for HTA agencies to hold stakeholders more accountable for the decisions made by AI-based MDs given how crucial such understanding is in high-stakes decisions. Finally, we believe that raising awareness of these concepts is essential for their widespread adoption and confidence.

POTENTIAL COMPETING INTERESTS

Author Murris reports a grant from the Association Nationale de la Recherche et de la Technologie, with Pierre Fabre, Convention industrielle de formation par la recherche number 2020/1701.

ACKNOWLEDGMENTS

The authors thank Milan Bhan for his methodological inputs. The authors also thank the reviewers for taking time and effort to review the manuscript. Dispose d'un menu contextuel

Dr Farah and author Murris contributed equally to this article and share the first authorship. Drs Martelli and Katsahian contributed equally to this article and share the last authorship.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at <https://www.mcpdigitalhealth.org/>. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AI**, artificial intelligence; **FDA**, Food and Drug Administration; **HAS**, Haute autorité de santé; **HTA**, health technology assessment; **MD**, medical device; **ML**, machine learning

Affiliations (Continued from the first page of this article): University, University Paris Cité, Paris, France; Inria (J.M.M., A.G., S.I.M.K.), Health data- and model- driven

Knowledge Acquisition, PariSantéCampus; Real World Evidence & Data Department (J.M.M.), Pierre Fabre, Boulogne-Billancourt, France; Department of Biostatistics and Epidemiology (I.B), Gustave Roussy, University Paris-Saclay, Villejuif, France; Oncostat U1018 (I.B), Inserm, University Paris-Saclay, Équipe Labelisée Ligue Contre le Cancer, Villejuif, France; Hôpital Européen Georges Pompidou, (N.M.M.), Pharmacy Department, Paris, France; Inserm (S.I.M.K.), Centre d'Investigation Clinique 1418 (CIC1418) Épidémiologie Clinique, Paris, France; and Hôpital Européen Georges Pompidou (S.I.M.K.), Department of Bioinformatics, Biostatistics and Public Health, Assistance Publique des Hôpitaux de Paris, Paris, France.

Grant support: This work was funded by a grant from the Association Nationale de la Recherche et de la Technologie, with Pierre Fabre, Convention industrielle de formation par la recherche number 2020/1701 (J.M.M.).

Correspondence: Address to Line Farah, PharmD, Groupe de Recherche et d'accueil en Droit et Economie de la Santé Department, University Paris-Saclay, Orsay, France (l.farah@hopital-foch.com).

ORCID

Line Farah:  <https://orcid.org/0000-0002-4021-6776>; Juliette M. Murris:  <https://orcid.org/0000-0002-7017-9865>; Isabelle Borget:  <https://orcid.org/0000-0002-6295-6361>; Agathe Guilloux:  <https://orcid.org/0000-0003-0473-1970>; Nicolas M. Martelli:  <https://orcid.org/0000-0001-5959-231X>; Sandrine I.M. Katsahian:  <https://orcid.org/0000-0002-7261-0671>

REFERENCES

1. The Artificial Intelligence Act. The Artificial Intelligence Act. <https://artificialintelligenceact.eu/>. Accessed September 23, 2022.
2. Règlement (UE) 2017/745 du Parlement européen et du Conseil du 5 Avril 2017 Relatif aux Dispositifs Médicaux, Modifiant la Directive 2001/83/CE, le Règlement (CE) N° 178/2002 et le Règlement (CE) N° 1223/2009 et Abrogeant les Directives du Conseil 90/385/CEE et 93/42/CEE (Texte présentant de l'intérêt pour l'EEE). OJ L vol. 117 (2017), <https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32017R0745>. Accessed September 25, 2022.
3. Harvey HB, Gowda V. How the FDA regulates AI. *Acad Radiol*. 2020;27(1):58-61. <https://doi.org/10.1016/j.acra.2019.09.017>.
4. Muehlematter UJ, Dianore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health*. 2021;3(3):e195-e203. [https://doi.org/10.1016/S2589-7500\(20\)30292-2](https://doi.org/10.1016/S2589-7500(20)30292-2).
5. Lampe K, Mäkelä M, Garrido MV, et al. The HTA Core Model: a novel method for producing and reporting health technology assessments. *Int J Technol Assess Health Care*. 2009;25(suppl 2): 9-20. <https://doi.org/10.1017/S0266462309990638>.
6. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPj Digit Med*. 2020;3:53. <https://doi.org/10.1038/s41746-020-0262-2>.
7. Dey S, Chakraborty P, Kwon BC, et al. Human-centered explainability for life sciences, healthcare, and medical

- informatics. *Patterns (N Y)*. 2022;3(5):100493. <https://doi.org/10.1016/j.patter.2022.100493>.
8. Saraswat D, Bhattacharya P, Verma A, et al. Explainable AI for Healthcare 5.0: opportunities and challenges. *IEEE Access*. 2022; 10:84486-84517. <https://doi.org/10.1109/ACCESS.2022.3197671>.
 9. Khodabandehloo E, Riboni D, Alimohammadi A. HealthXAI: collaborative and explainable AI for supporting early diagnosis of cognitive decline. *Future Gener Comput Syst*. 2021;116:168-189. <https://doi.org/10.1016/j.future.2020.10.030>.
 10. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(1):e745-e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
 11. Hasani N, Morris MA, Rhamin A, et al. Trustworthy artificial intelligence in medical imaging. *PET Clin*. 2022;17(1):1-12. <https://doi.org/10.1016/j.petcl.2021.09.007>.
 12. Vollmer S, Mateen B, Bohner G, et al. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. *Preprint. Posted online December*. 2018;21. arXiv 1812.10404. <https://doi.org/10.48550/arXiv.1812.10404>.
 13. European Commission, Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI, Publications Office. 2019. <https://data.europa.eu/doi/10.2759/346720>. Accessed August 25, 2022.
 14. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learn Syst*. 2021;32(11):4793-4813. <https://doi.org/10.1109/TNNLS.2020.3027314>.
 15. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep*. 2019; 49(1):15-21. <https://doi.org/10.1002/hast.973>.
 16. Lötsch J, Krügel D, Ullsch A. Explainable artificial intelligence (XAI) in biomedicine: making AI decisions trustworthy for physicians and patients. *BioMedInformatics*. 2022;2(1):1-17. <https://doi.org/10.3390/biomedinformatics2010001>.
 17. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655. <https://doi.org/10.1016/j.jbi.2020.103655>.
 18. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc*. 2022;29(9):1525-1534. <https://doi.org/10.1093/jamia/ocac093>.
 19. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996; 15(4):361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
 20. Biecek P, Burzykowski T. *Exploratory Model Analysis*. 1st ed. Chapman & Hall/CRC; 2021.
 21. Dietterich T. Overfitting and undercomputing in machine learning. *ACM Comput Surv*. 1995;27(3):326-327. <https://doi.org/10.1145/212094.212114>.
 22. Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci*. 2004;44(1):1-12. <https://doi.org/10.1021/ci0342472>.
 23. Ying X. An overview of overfitting and its solutions. *J Phys Conf Ser*. 2019;1168(2):022022. <https://doi.org/10.1088/1742-6596/1168/2/022022>.
 24. Refaeizadeh P, Tang L, Liu H. Cross-validation. In: Liu L, Özsu MT, eds. *Encyclopedia of Database Systems*. Springer; 2009:532-538.
 25. Probst P, Boulesteix AL, Bischl B. Tunability: importance of hyperparameters of machine learning algorithms. *J Mach Learn Res*. 2019;20:1934-1965.
 26. Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery; 2019:1-13. <https://doi.org/10.1145/3290605.3300809>.
 27. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2019;51(5):1-42. <https://doi.org/10.1145/3236009>.
 28. Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38. <https://doi.org/10.1016/j.artint.2018.07.007>.
 29. Dam HK, Tran T, Ghose A. Explainable software analytics. In: *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*. Association for Computing Machinery; 2018:53-56. <https://doi.org/10.1145/3183399.3183424>.
 30. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. <https://doi.org/10.1038/s42256-019-0048-x>.
 31. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2013:623-631. <https://doi.org/10.1145/2487575.2487579>.
 32. Cox DR. Regression models and Life-Tables. *J R Stat Soc Series B Stat Methodol*. 1972;34(2):187-202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
 33. Fürnkranz J, Kliegr T, Paulheim H. On cognitive preferences and the plausibility of rule-based models. *Mach Learn*. 2020;109(4): 853-898. <https://doi.org/10.1007/s10994-019-05856-5>.
 34. Müller W, Wiederhold E. Applying decision tree methodology for rules extraction under cognitive constraints. *Eur J Oper Res*. 2002;136(2):282-289. [https://doi.org/10.1016/S0377-2217\(01\)00115-1](https://doi.org/10.1016/S0377-2217(01)00115-1).
 35. Arya V, Bellamy RK, Chen P, et al. One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. *Preprint. Posted online September*. 2019. arXiv 1909.03012. <https://doi.org/10.48550/arXiv.1909.03012>.
 36. Danilevsky M, Dhanorkar S, Li Y, Popa L, Qian K, Xu A. Explainability for natural language processing. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2021:4033-4034. <https://doi.org/10.1145/347548.3470808>.
 37. Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. Benchmarking and survey of explanation methods for black box models. *Astrophysics Data System*. <https://ui.adsabs.harvard.edu/abs/2021arXiv210213076B>. Accessed September 2, 2022.
 38. Varshney KR. Interpretability and explainability. In: *Trustworthy Machine Learning*. Independently published 2022. chap 12.
 39. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub; 2020.
 40. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy (Basel)*. 2020;23(1):E18. <https://doi.org/10.3390/e230018>.
 41. Leiter C, Lertvittayakumjorn P, Fomicheva M, et al. Towards explainable evaluation metrics for natural language generation. *Preprint. Posted online March 21, 2022*. arXiv 2203.11131. <https://doi.org/10.48550/arXiv.2203.11131>.
 42. Selvaraju RR, Das Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Preprint. Posted online October 7 2016*. arXiv 1610.02391. <https://doi.org/10.1007/s11263-019-01228-7>.
 43. Hohman F, Head A, Caruana R, DeLine R, Drucker SM. Gamut: a design probe to understand how data scientists understand machine learning models. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery; 2019:1-13. <https://doi.org/10.1145/3290605.3300809>.

IV.1. Importance des critères d’interprétabilité et d’explicabilité pour les dispositifs médicaux

ASSESSMENT OF AI-BASED MDS

44. Lage I, Chen E, He J, et al. An evaluation of the human-interpretability of explanation. Preprint. Posted online January. 2019;31. arXiv 1902.00006. <https://doi.org/10.48550/arXiv.1902.00006>.
45. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J Exp Psychol Gen.* 2015;144(1):114-126. <https://doi.org/10.1037/xge0000033>.
46. Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2020:648-657. <https://doi.org/10.1145/3351095.3375624>.
47. Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. *Digit Health.* 2022;8:2055207621074488. <https://doi.org/10.1177/2055207621074488>.
48. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017. Preprint. Posted online February. 2017. arXiv 1702.08608. <https://doi.org/10.48550/arXiv.1702.08608>.
49. Schmidt P, Biessmann F. Quantifying interpretability and trust in machine learning systems. Preprint. Posted online January. 2019; 20. arXiv 1901.08558. <https://doi.org/10.48550/arXiv.1901.08558>.
50. Liang G, Newell B. Trusting algorithms: performance, explanations, and sticky preferences. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Cognitive Science Society; 2022:708-714.
51. Hase P, Bansal M. Evaluating explainable AI: which algorithmic explanations help users predict model behavior?. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2020:5540-5552. <https://doi.org/10.18653/v1/2020.acl-main.491>.
52. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. Preprint. Posted online November. 2017. arXiv1711.06104. <https://doi.org/10.48550/arXiv.1711.06104>.
53. Yeh CK, Hsieh CY, Suggala A, Inouye DI, Ravikumar PK. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems.* 2019;32.
54. Margot V, Luta G. A new method to compare the interpretability of rule-based algorithms. *AI.* 2021;2(4):621-635. <https://doi.org/10.3390/ai2040037>.
55. Chiaburu T, Biessmann F, Hausser F. Towards ML methods for biodiversity: a novel wild bee dataset and evaluations of XAI methods for ML-assisted rare species annotations. Preprint. Posted online June. 2022. arXiv 2206.07497. <https://doi.org/10.48550/arXiv.2206.07497>.
56. Liao QV, Zhang Y, Luss R, Doshi-Velez F, Dharandhar A. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI. *HCOMP. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 10.* 2022;10(1):147-159. <https://doi.org/10.1609/hcomp.v10i1.21995>.
57. Bhateja V, Satapathy SC, Satori H. Explainable AI for healthcare: from black box to interpretable models. In: Adadi A, Berrada M, eds. *Embedded Systems and Artificial Intelligence*, vol 1076. Springer; 2020.
58. Dave D, Naik H, Singhhal S, Patel P. Explainable AI meets healthcare: a study on heart disease dataset. Preprint. Posted online November. 2020. arXiv 2011.03195. <https://doi.org/10.48550/arXiv.2011.03195>.
59. Holzinger A, Saranti A, Molnar C, Biecek P, Samek W. Explainable AI methods—a brief overview. In: Holzinger A, et al., eds. *xxAI—Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, Vienna, Austria. Revised and Extended Papers*. Springer International Publishing; 2022:13-38. https://doi.org/10.1007/978-3-031-04083-2_2.
60. Aras L, Arjona-Medina J, Widrich M, et al. Explaining and interpreting LSTMs. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, eds. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing; 2019:211-238. https://doi.org/10.1007/978-3-030-28954-6_11.
61. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. *J Med Internet Res.* 2019;21(3):e12422. <https://doi.org/10.2196/12422>.
62. Tamori H, Yamashina H, Mukai M, Morii Y, Suzuki T, Ogasawara K. Acceptance of the use of artificial intelligence in medicine among Japan’s doctors and the public: a questionnaire survey. *JMIR Hum Factors.* 2022;9(1):e24680. <https://doi.org/10.2196/24680>.
63. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform.* 2021; 28(1):e100450. <https://doi.org/10.1136/bmjhci-2021-100450>.
64. Holzinger A, Dehmer M, Emmert-Streib F, et al. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Inf Fusion.* 2022;79:263-278. <https://doi.org/10.1016/j.inffus.2021.10.007>.
65. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intell Based Med.* 2020; 1-2:100001. <https://doi.org/10.1016/j.ibmed.2020.100001>.
66. Jobson D, Mar V, Freckleton I. Legal and ethical considerations of artificial intelligence in skin cancer diagnosis. *Australas J Dermatol.* 2022;63(1):e1-e5. <https://doi.org/10.1111/ajd.13690>.
67. Wilhelm D, Hartwig R, McLennan S, et al. [Ethical, legal and social implications in the use of artificial intelligence-based technologies in surgery: principles, implementation and importance for the user]. Article in German. *Chirurg.* 2022; 93(3):223-233. <https://doi.org/10.1007/s00104-022-01574-2>.
68. Lang M, Bernier A, Knoppers BM. Artificial intelligence in cardiovascular imaging: “unexplainable” legal and ethical challenges? *Can J Cardiol.* 2022;38(2):225-233. <https://doi.org/10.1016/j.cjca.2021.10.009>.
69. O’Sullivan S, Nevejans N, Allen C, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int J Med Robot.* 2019;15(1):e1968. <https://doi.org/10.1002/rcs.1968>.
70. Currie G, Hawk KE. Ethical and legal challenges of artificial intelligence in nuclear medicine. *Semin Nucl Med.* 2021;51(2):120-125. <https://doi.org/10.1053/j.semnuclmed.2020.08.001>.
71. Naik N, Hameed BMZ, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg.* 2022;9:862322. <https://doi.org/10.3389/fsurg.2022.862322>.
72. Ferrario A, Loi M. How explainability contributes to trust in AI. In: *ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery; 2022:1457-1466. <https://doi.org/10.1145/3531146.3533202>.
73. Castelnovo A, Grupi R, Greco G, Regoli D, Penco IG, Cosentini AC. A clarification of the nuances in the fairness metrics landscape. *Sci Rep.* 2022;12(1):4209. <https://doi.org/10.1038/s41598-022-07939-i>.
74. Dimanov B, Bhatt U, Jamnik M, Weller A. You shouldn’t trust me: learning models which conceal unfairness from multiple explanation methods. https://mlg.eng.cam.ac.uk/adrian/ECAI20-You_Should%20%99_Trust_Me.pdf. Accessed September 27, 2022.
75. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM; 2020:180-186. <https://doi.org/10.1145/3375627.3375830>.
76. Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst.* 2021;11(3-4):1-45. <https://doi.org/10.1145/3387166>.

77. Bhatt U, Andrus M, Weller A, Xiang A. Machine learning explainability for external stakeholders. *Preprint. Posted online July, 2020;10:05408. arXiv 2007.05408.* <https://doi.org/10.48550/arXiv.2007.05408>.
78. Haverinen J, Keränen N, Falkenbach P, Majala A, Kolehmainen T, Reponen J. Digi-HTA: Health technology assessment framework for digital healthcare services. *FirjeHeW.* 2019;11(4):326-341.
79. Olczak J, Pavlopoulos J, Prijis J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta Orthop.* 2021;92(5):513-525.
80. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the M-CLAIM checklist. *Nat Med.* 2020;26(9):1320-1324.
81. Haute Autorité de santé. Dossier submission to the Medical Device and Health Technology Evaluation Committee. 2020. https://www.has-sante.fr/upload/docs/application/pdf/2020-10/guide_dm_vf_english_publi.pdf. Accessed August 19, 2022.
82. US. FDA, Health Canada and MHRA. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>. Accessed August 13, 2022.
83. Korean Ministry of Food and Drug Safety. Guideline on Review and Approval of Artificial Intelligence (AI) and big data-based Medical Devices (For Industry). 2020. https://www.mfds.go.kr/eng/brdm_40/view.do?seq=72623&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=1. Accessed August 16, 2022.
84. Butt UM, Letchmunan S, Ali M, et al. Machine learning based diabetes classification and prediction for healthcare applications. *J Healthc Eng.* 2021;2021:9930985.
85. Zeng P, Zhou X, Huang S. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genomics.* 2017;18:368.
86. Bedogni G, Bellentani S, Miglioli L, et al. The fatty liver index: a simple and accurate predictor of hepatic steatosis in the general population. *BMC Gastroenterol.* 2006;6:33.
87. Nemat H, Fehri H, Ahmadinejad N, Frangi AF, Gooya A. Classification of breast lesions in ultrasonography using sparse logistic regression and morphology-based texture features. *Med Phys.* 2018. <https://doi.org/10.1002/mp.13082>.
88. Cheng Y-L, Sung SH, Cheng HM, et al. Prognostic nutritional index and the risk of mortality in patients with acute heart failure. *J Am Heart Assoc.* 2017;6:e004876.
89. Wei J, Zeng Y, Gao X, Liu T. A novel ferroptosis-related lncRNA signature for prognosis prediction in gastric cancer. *BMC Cancer.* 2021;21:1221.
90. van Hoffen MFA, Norder G, Twisk JWR, Roelen CAM. External validation of a prediction model and decision tree for sickness absence due to mental disorders. *Int Arch Occup Environ Health.* 2020;93:1007-1012.
91. Mandala S, Cai Di T, Sunar MS. ECG-based prediction algorithm for imminent malignant ventricular arrhythmias using decision tree. *PLoS One.* 2020;15:e0231635.
92. Tseng P-Y, Chen YT, Wang CH, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care.* 2020;24:478.
93. Lombardi A, Diacono D, Amoroso N, et al. Explainable deep learning for personalized age prediction with brain morphology. *Front Neurosci.* 2021;15:674055.
94. Killian MO, Payrovaziri SN, Gupta D, Desai D, He Z. Machine learning-based prediction of health outcomes in pediatric organ transplantation recipients. *JAMIA Open.* 2021;4:ooab008.
95. Boorman ED, Behrens TE, Rushworth MF. Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biol.* 2011;9:e1001093.

IV.2 Interprétabilité et survie

IV.2.1 Contexte

Les outils d'IA pour l'analyse de survie sont devenus essentiels pour la médecine personnalisée, et ils permettent d'éclairer les décisions thérapeutiques et d'améliorer les soins aux patients [Park et al., 2020, Quazi, 2022]. Alors que de nombreuses méthodes d'interprétabilité ont été proposées pour les tâches classiques de classification et de régression [Molnar, 2020], leur extension aux modèles de survie n'est pas aussi avancée [Langbein et al., 2024]. Cette contradiction implique un besoin important d'interprétabilité pour les modèles de survie afin d'atteindre le même niveau de transparence que dans d'autres domaines du ML. Par conséquent, les méthodes d'interprétabilité des modèles de survie ne sont pas seulement une nécessité technique, mais aussi un impératif réglementaire et clinique croissant pour répondre aux exigences des autorités de santé et maintenir l'intégrité des processus décisionnels en matière de soins de santé.

IV.2.2 Étude de cas

Nous proposons dans cette section une illustration de trois modèles de ML de survie, avec trois méthodes d'interprétabilité sur des données en accès ouvert. Rappelons que nous nous fixons dans le cadre de survie du temps jusqu'au premier événement. L'article qui suit, intitulé "*Bridging Interpretability and Survival Endpoints in Health Technology Assessment*" a été soumis à la conférence *Artificial Intelligence, Ethics and Society 2024*.

Ce travail a été réalisé dans le cadre d'un projet collaboratif avec les étudiants de l'ENSAE, Mattéo Alquier, Émile Cassant et Marama Simoneau, que je tiens à remercier pour leur engagement. Je souhaite également exprimer ma gratitude au Pr Nicolas Chopin pour son enthousiasme et son soutien.

Bridging Interpretability and Survival Endpoints in Health Technology Assessment

Anonymous submission

Abstract

The requirements of interpretability for Health Technology Assessment (HTA) is critical with the emergence of AI-based medical devices, where the need for transparency is paramount. In oncology, clinicians and researchers typically focus on survival endpoints, which are pivotal in evaluating treatment efficacy and informing patient care strategies. However, there are few examples of interpretability methods tailored specifically for survival analysis, leaving a gap in the ability to fully understand and trust the predictions made by AI models in this domain. We bridge this gap by offering a detailed illustration in oncology of three survival machine learning models, and applying three interpretability methods to each. Our work assesses the functionality of these interpretability techniques, providing a set of recommendations and best practices for their use. Additionally, we propose a two-step process for optimizing survival ML models, enhancing their predictive accuracy while maintaining interpretability. Lastly, we identify opportunities for the development of new interpretability methods within the survival analysis framework. By addressing the current limitations and exploring future directions, our research contributes to the facilitation of the integration of AI into HTA and supporting the advancement of transparent AI-based medical devices.

Introduction

Artificial intelligence (AI) and machine learning (ML) technologies offer transformative potential in healthcare by analyzing large volumes of data, and reaching high level of performance in a wide range of tasks. These technologies are applied in various medical device (MD) areas, such as image processing, disease detection, diagnosis, prognosis, risk assessment, pattern identification in physiology and disease, personalized diagnostics, and monitoring treatment responses (Secinaro et al. 2021). The use of AI algorithms in MDs is rapidly expanding, with 151 newly authorized AI-based MDs between August 1, 2023, and March 31, 2024 by the US HTA body Food and Drug Administration (FDA) (US. FDA 2024a).

However, AI-based MDs pose regulatory challenges due to the sensitivity and complexity of clinical data (US. FDA 2024b). Regulatory agencies and health technology assessment (HTA) bodies aim to address these stakes with different kinds of efforts, such as: developing robust methods for training AI algorithms under small data regimes, devis-

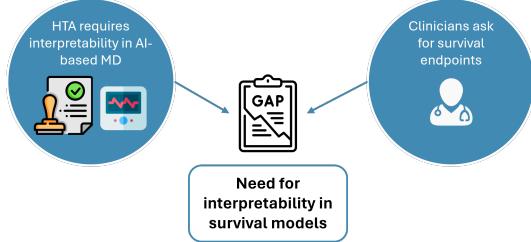


Figure 1: Integrating interpretability with survival endpoints enhances the transparency and utility of health technology assessments for AI-based medical devices.

ing strategies to reduce bias, establishing performance metrics and reference standards, and evaluating safety and effectiveness (Regier et al. 2022; Farah et al. 2023). Post-market surveillance is also crucial to monitor the performance of deployed AI-based MDs (US. FDA 2023). However, the assessment of AI-based MDs is not limited to performance evaluation. Typically, the interpretability of these algorithms is paramount in healthcare to answer the need for transparency and understandability coming from clinicians and regulators (Vollmer et al. 2018). To this end, interpretability facilitates the integration of AI into clinical practice, where the rationale behind AI-driven recommendations must be clear to healthcare providers and patients (LaRosa and Danks 2018).

On the other hand, healthcare studies often concentrate on specific clinical outcomes, such as mortality. The critical clinical need is to determine the occurrence and timing of these events. Survival endpoints are constructed to capture this information, typically comprising a binary indicator related to the occurrence of an event and a time component measuring the duration until the event occurs. This necessitates a specialized form of statistical analysis known as survival analysis (Jenkins 2005). Nevertheless, mortality is not the only event of interest to clinicians. Depending on the disease or condition being studied, survival endpoints are carefully chosen and defined to align with clinical significance. For instance, an asthma patient's survival endpoint might be the occurrence of a flare-up, while for certain cancers with metastatic progression, it could be the time

to disease progression or response to therapy (Emmerson and Brown 2021). Survival analysis benefited from the advent of AI and ML as well. Traditional ML algorithms were indeed adapted to handle survival data, leading to the development of survival-specific counterparts (Wang, Li, and Reddy 2019). These AI-driven survival analysis tools became central in personalized medicine, and inform treatment decisions and improve patient care (Park et al. 2020; Quazi 2022).

While a wide range of interpretability methods has been proposed for conventional classification and regression tasks (Molnar 2020), their extension to survival models is not as advanced (Langbein et al. 2024). This discrepancy implies a strong need for more interpretability for survival models to reach the same level of transparency as compared to other common areas of ML. Consequently, interpretability methods for survival models is not solely a technical necessity, but also a growing regulatory and clinical imperative to meet the sharp requirements of HTA bodies and to maintain the integrity of healthcare decision-making processes (Figure 1).

To tackle the aforementioned lack of interpretability for survival endpoints in the HTA evaluation of AI-based MDs, this paper proposes the following main contributions:

- We offer a detailed illustration of three survival ML models, with three methods of interpretability on open-source oncology data;
- We draw up guidance and assessment of how each interpretability method works, and provide recommendations and best practices;
- We identify avenues for the further development of new interpretability methods in survival framework to complement existing research in the field.

In the "Background and Related Work" section, we provide an overview of the necessity for interpretability within HTA, elucidating the current landscape of state-of-the-art interpretability methods. Additionally, we lay the groundwork by presenting the essential principles of survival analysis, setting the stage for the subsequent application of these concepts. The "Methodology" section offers a detailed explanation of the survival models and interpretability methods that we have applied to an open-source dataset in oncology. We also outline our protocol for the impact of hyperparameters, which serves as a rigorous framework for evaluating the interpretability of our models. In the "Results" section, we conduct a thorough investigation into the effects of the applied survival models and interpretability methods, examining their performance and the insights they yield. Finally, the "Discussion" section reflects on the strengths and limitations of our approach. We critically assess the implications of our results and suggest directions for future research that could further advance the field of interpretability for survival endpoints in HTA.

Background and related work

In this section, we recall some basic principles of interpretability in healthcare and survival models.

Background

HTA is a multidisciplinary process which refers to the systematic evaluation of properties, effects, and/or impacts of health technology (Organization et al. 2011). The European Union Medical Device Regulation (EU MDR) categorizes medical devices into one of four classes: Class I, Class IIa, Class IIb, and Class III medical devices. The MDR medical device classification is based on the device's potential risk of harm to users (Group et al. 2021).

Interpretability in healthcare: a fundamental need Understanding and elucidating the reasoning process behind ML models outcomes is crucial in healthcare (Nazar et al. 2021; Bharati, Mondal, and Podder 2023). Interpretability (or explainability) has several connections with bias detection, transparency, user acceptance, and trust in technologies (Srinivasu et al. 2022) (Figure 2).

Clinicians are increasingly sensitive to the integration of AI tools into their daily routines, and they need to trust the decisions made by algorithms to incorporate them into their practice (LaRosa and Danks 2018). If the reasoning behind these decisions is opaque, it can lead to skepticism and reluctance to adopt the technology. Besides, LaRosa and Danks (2018) emphasize the need for doctors to be educated about AI systems, and for patients to give educated consent before AI is used in their care. Transparent models with high interpretability thus foster trust and increase the likelihood of adoption (Asan, Bayrak, and Choudhury 2020). On the other hand, regulatory oversight is provided by HTA bodies, which carry a broader responsibility towards all healthcare stakeholders. Developing HTA standards in interpretability for AI-based MDs could also streamline market and patient access across countries (Singh 2022).

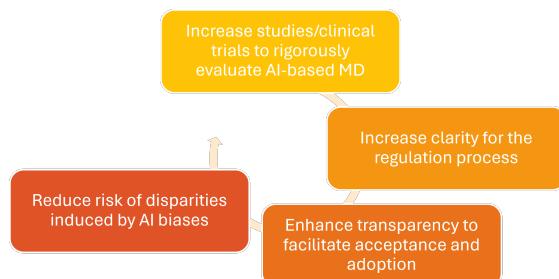


Figure 2: Proposed cycle for trust in AI-based medical devices

Moreover, clinical research for discovering new treatments can greatly benefit from models that provide valuable insights about potential underlying patterns and complex relationships in patient data (Jiménez-Luna, Grisoni, and Schneider 2020). Clinicians can learn from these insights to enhance their understanding of diseases, treatments, and patient outcomes, leading to improved healthcare practices.

Given the aforementioned importance of interpretability in the HTA of AI-based MDs, recent efforts have aimed to consolidate the requirements set forth by various HTA bod-

ies worldwide (Farah et al. 2023). For instance, the FDA and the National Institute for Health and Care Excellence (NICE, UK) emphasize the importance of "understanding of a model's intended integration into clinical workflow (interpretability and explicability)" (US. FDA, Health Canada and MHRA 2021). Similarly, the Haute Autorité de santé (HAS, France) mandates considerations such as "performance and qualification, system robustness and resilience, explainability, and interpretability" (Haute Autorité de Santé 2020). However, despite the recognized importance of these criteria, a standardized methodology for their measurement is still missing.

Interpretability techniques: a review We distinguish two types of interpretability methods: *gradient-based* and *permutation-based* approaches. *Perturbation-based* methods, such as Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) and SHAPley Additive exPlanations (SHAP) (Lundberg and Lee 2017), compute feature importance by perturbing feature values and observing the resultant variations in model output. *Gradient-based* approaches, in contrast, derive feature importance directly from the model's gradients. Techniques such as Integrated Gradients (Sundararajan, Taly, and Yan 2017) and DeepLIFT (Deep Learning Important FeaTures) (Shrikumar, Greenside, and Kundaje 2017) backpropagate the prediction error to the input features, thereby quantifying each feature's contribution to the final prediction.

Both *perturbation-based* and *gradient-based* methods have their strengths and are often chosen based on the model type and the specific interpretability requirements of the task at hand. While perturbation-based methods are generally *model-agnostic* and can be applied to any machine learning model, gradient-based methods are *model-specific* and are typically more suited to neural networks.

Survival analysis: rationale, concepts and definitions Regulatory agencies, such as the FDA and the European Medicines Agency (EMA), often require survival endpoints as primary efficacy measures in clinical trials, especially for therapies intended to treat life-threatening diseases (Pavlovic et al. 2014). Such endpoints are also very commonly used both in real-world studies:

- Overall survival measures the time from the start of the study (.e.g treatment initiation) to death from any cause. This endpoint is commonly used in clinical trials as it informs on the prolongation of life.
- Event-free survival measures the time from the start of treatment until the occurrence of a predefined event, such as disease progression, or relapse. It is particularly relevant in conditions where disease progression is a critical concern, and treatments aim to stabilize the disease.
- Composite endpoints with a survival event combine multiple outcomes into a single measure, accounting for different events that could occur, such as disease progression or death. They are useful in scenarios where multiple significant clinical outcomes are expected and are equally important for patient prognosis.

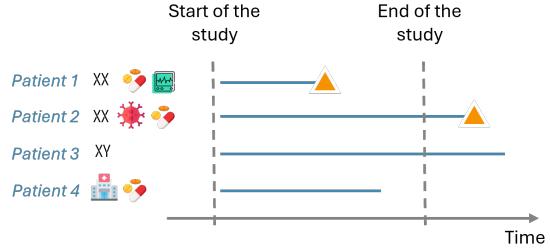


Figure 3: Patient data collected for survival analysis includes their status (alive or deceased), the time to the event (e.g., death), and other relevant factors like demographics and comorbidities.

Survival analysis refers to the analysis of these time-to-event data. In many cases, not all patients will experience the event of interest within the study period. Censoring is defined in cases when the event of interest has not occurred for some patients by the end of the study or before they are lost to follow-up. To understand well the principles of censoring, patient 1 from Figure 3 is not censored because the event happened in the study window. However, patients 2, 3 and 4 are censored because either they were alive at the end of the study (patients 2 and 3), either they did not have the event of interest and the information was not available by the end of the study (patient 4).

For this reason, survival analysis requires specific tools to handle censored data for time-to-event endpoints for the robust estimations of the associated survival probabilities. Key notions include:

- The survival time is the time-to-event information and represents the time from a specific starting point (like diagnosis or treatment initiation) until the occurrence of the event of interest (or censoring time).
- The survival function is defined by the probability that the event of interest has not occurred by time t . Mathematically, the survival function $S(t)$ writes

$$S(t) = P(T > t) \quad (1)$$

where T is the survival time.

- The hazard function is the probability that the event occurs at time t , given no censoring up to time t , divided by the width of a very small time interval. Mathematically, the hazard function $\lambda(t)$ writes

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (2)$$

Alongside the non-parametric Kaplan-Meier (KM) methodology (Kaplan and Meier 1958), the Cox proportional hazard (CPH) model (Cox 1972) is the gold standard for the estimation of the survival function. The CPH can be considered as an extension of classical regression model with a log-function. This semi-parametric approach fit a log-linear function of the hazard function $\lambda(t)$ and writes

$$\lambda(t) = \lambda_0(t) \cdot \exp(\beta X) \quad (3)$$

with $\lambda_0(t)$ the baseline hazard function, β the coefficients associated with covariates X . Outputs from the CPH model include the survival probabilities over time for a new patient with a specific set of covariates, the risk scores for each patient as the linear predictor, and the relative risk of two patients by taking the ratio of their hazard functions.

Two main assumptions arise from this modelisation: (i) the proportionality of hazards, and (ii) the log-linearity for continuous predictors. The proportional hazard assumption (PHA) states that the hazard rate for any individual is proportional to the hazard rate of any other individual at any given time, and this proportionality remains constant over time. The PHA also simplifies the interpretation of the results as it implies that the effect of a predictor variable on the hazard rate is the same over time. The PHA thus allows for fair comparisons between groups or levels of predictor variables over time.

To overcome such assumptions and to capture further complexity of data, machine and deep learning has extended the field of survival analysis, namely with random survival forests (RSF), survival support vector machine (SSVM), and survival neural networks (Wang, Li, and Reddy 2019). Further details on these approaches are given in the methodology section.

Related work

While we observe a growing imperative for transparency in AI-based MDs, survival analysis remains the gold standard for evaluating effective therapies, notably in oncology. However, the literature presents limited examples of interpretability in survival models.

Very recently, Langbein et al. (2024) presented a comprehensive review of existing interpretability methods for survival analysis (Langbein et al. 2024). In this review, the presented *model-agnostic* interpretability methods are survival extensions. Permutation importance is a global interpretation method for assessing the importance of variables in a survival model (Breiman 2001). The underlying principle suggests that altering the values of an important feature is likely to have an impact on the survival model performance. Unlike the original LIME, SurvLIME leverages the CPH model to better capture the time-dependent risk (Kovalev, Utkin, and Kasimov 2020). Similarly, SHAP has been recently expanded with SurvSHAP and accommodates models that produce functional outputs, such as survival functions (Alabdallah et al. 2022). Survival counterfactual explanations extend by adapting the model's input features and prediction (Kovalev et al. 2021). Cottin et al. (2024) also proposed a counterfactual perturbation feature importance for survival multi-state processes.

In practice, *model-agnostic* approaches are not very often used to address clinical purpose. SurvSHAP has only been applied in two cases: heart failure survival explanation from a survival Gradient boosting model (Moreno-Sanchez 2023), and conversion risk interpretability in patients with Alzheimer's disease with a random survival forest (Sarica et al. 2023). Besides, Moncada-Torres et al. (2021) demonstrates that SurvSHAP enables better understand higher performances from Extreme Gradient Boosting in predicting

breast cancer survival, compared to the CPH model. On the other hand, SurvLIME was not applied to any data in oncology.

Another way to compute explanations is to apply *model-specific* interpretability methods. These approaches provide explanations directly based on internal model parameters. For instance, the combination of the CPH model with neural networks provides insights into feature importance, as demonstrated by Xu and Guo (2023). Additionally, feature selection processes embedded within generalized additive models have been explored for their interpretability, as shown in the work of Van Ness and Udell (2024). Besides, in gene discovery, models focus on pathway-level predictions and leverage gradient-weighted class activation mapping (Wang et al. 2024) or compute gene occurrence probabilities (Hou et al. 2023) to provide interpretable insights. However, these *gradient-based* approaches have been specifically developed to address distinct clinical concerns and may be limited to those particular applications. Therefore, caution should be exercised when considering the use of these methods in different clinical settings.

As pointed out in Langbein et al. (2024), there is a notable gap in the provision of concrete examples and practical applications of interpretability within survival analysis frameworks, which is essential to address specific and real-world needs effectively.

Methodology

In this section, we outline the methodological reflexion we used to create an open-source pipeline program in Python. First we list survival ML models, and interpretability methods that we use in our analysis.

Common ML algorithms and survival counterparts

Common ML algorithm have been extended to adapt for survival data. These methods are popular in survival analysis due to their ability to handle complex and high-dimensional data, accommodate censored observations, and provide accurate predictions of time-to-event outcomes (Wang, Li, and Reddy 2019).

Random survival forests RSF works by constructing multiple decision trees on various subsets of the data and using them to estimate the survival function for each instance (Ishwaran et al. 2008). At each node, the splitting rule is based on KM curves. The best split for a given node is made when maximizing the difference between the two KM curves based on Logrank statistic. Terminal node estimations are in turn made with KM methodology. The final prediction is made by averaging the survival functions across all trees. Additional outputs from RSF are associated risk scores.

Survival eXtreme Gradient Boosting For survival analysis, XGBoost can be adapted to handle censored data by using appropriate loss functions (Friedman 2001, 2002). We use loss functions specific to survival analysis, which take censoring into account, such as Cox's partial true likelihood loss function. It builds an ensemble of survival trees in a sequential manner, where each tree corrects the errors of the

previous ones. When adapted for survival analysis, Survival GBoost (SGB) can output either a risk score, or a survival function.

Survival Support Vector Machines Similar to traditional SVM, SSVM aims to find a hyperplane that separates the data (Van Belle et al. 2008). However, the objective function is modified to account for censored observations and to maximize the margin between individuals who have experienced the event and those who have not, while also minimizing the risk of misclassification. The model tries to order the individuals by their risk scores in such a way that corresponds to the actual order of events. SSVM models typically output a risk score for each instance.

Performance evaluation

The area under the curve (AUC), concordance index time-dependent (c-index), and integrated Brier score (IBS) are commonly performance metrics in survival models.

Time-dependent and cumulative dynamic AUC The cumulative dynamic AUC accounts for the fact that a patient's health status is not static but changes over time. This time-dependent AUC considers 'cumulative cases'—individuals who have experienced an event by a certain time—and 'dynamic controls'—those who have not experienced the event by that time. The cumulative dynamic AUC at a specific time point reflects the model's ability to differentiate between subjects who experience an event by that time and those who do not.

Concordance Index The concordance index, referred to as the C-index, is a generalization of the ROC-AUC for censored data, providing a reliable ranking of survival times based on individual risk scores (Harrell et al. 1982, 1984). This metric is defined as the proportion of concordant pairs over the total number of comparable pairs. Concordant pairs refer to pairs with higher predicted risk scores with longer survival time, and comparable pairs are individual with same outcomes. Higher values, closer to 1, indicate better predictive performance.

Integrated Brier Score The Brier score represents an adaptation of the mean squared error for right-censored data, commonly employed to evaluate a model's calibration (Graf et al. 1999). The integrated Brier score averages the Brier score over a range of time points, providing an overall summary of the model's performance across the entire time span of interest.

Interpretability methods extended to survival model outputs

Our work focused on *model-agnostic* methods, requiring only a survival prediction function. These methods can therefore be applied to all kinds of models.

Permutation feature importance The performance metric used for evaluation might be the concordance index or (integrated) Brier score for survival models.

SurvSHAP SurvSHAP computes time-dependent feature attributions, which measure how each feature contributes to the survival model's prediction at different time points (Krzysztof et al. 2023). SurvSHAP can provide both local explanations for individual predictions and global explanations by aggregating the Shapley values across multiple instances.

SurvLIME To solve the optimization problem, SurvLIME seeks to find the optimal coefficients that minimize the weighted average distance between the predicted cumulative hazard functions of the black-box model and those estimated by the Cox proportional hazards surrogate model for the perturbed samples (Kovalev, Utkin, and Kasimov 2020). The local importance values for features are computed based on the coefficients obtained from the surrogate model, reflecting their influence on the prediction for the specific instance.

Impact of hyperparameters

We focus on the optimization of hyperparameters as a means to reduce model complexity without compromising performance. A simpler model inherently offers better interpretability, which is crucial for understanding the underlying mechanisms of prediction. For example, an optimized random forest with a smaller number of trees is preferred, as it allows for a detailed examination of each individual tree. This can shed light on how specific decisions are made within the model, making the interpretability of complex models more accessible and transparent.

The impact of hyperparameters has been demonstrated in survival framework (Vabalas et al. 2019). Hyperparameter optimization is an essential part of finding the best setup when training a model, thereby improving its performance metrics. Our optimization process combines two methods. We first use the set of optimal hyperparameters given by a Bayesian search performed with the Optuna algorithm (Akiba et al. 2019). This results in a more relevant and smaller grid, which we use as input for a grid search by cross-validation (LeCun et al. 1998). In the same spirit of permutation feature importance, we used values from hyperparameters and evaluated their importance on the performance metric.

Data source

The readmission dataset serves as a pivotal resource for illustrating methodological concepts in the survival analysis of recurrent events (González et al. 2005). This dataset encapsulates the medical history of 403 patients who underwent surgery for colorectal cancer, tracking multiple instances of rehospitalization and death. It includes pertinent variables such as the administration of chemotherapy, patient sex, the Dukes' tumoral stage, and the time-varying Charlson comorbidity index, which provides a measure of concurrent health issues. The objective is to understand which factors influence the survival status amongst this patients.

Results

This section presents the findings in answer to study objectives. First, we evaluate the survival models performances, followed by an in-depth analysis of their interpretability. Subsequently, we examine the influence of hyperparameters on each model's performance.

Performances

Data were splitted in training and test set. The training set was used to train and fine-tune the survival models, and the test set enabled to compute performance metrics. The training-test split was repeated five times to capture variability.

Table 1 reports average performances in terms of concordance index, time dependent AUC and integrated Brier score for all three models. RSF and SGB provide similar performances, with slightly better results for RSF. As SSVM does not output any survival functions for each patient, the IBS cannot be computed for such model. Therefore, if the interest of an SSVM model may less be in the calibration than the predictive performance.

	C-index ↑	Cum. dyn. AUC ↑	IBS ↓
RSF	0.892	0.924	0.112
SGB	0.886	0.923	0.113
SSVM	0.873	0.909	/

Table 1: Average performance metrics in terms of concordance index, cumulative dynamic AUC and integrated Brier score for Random Survival Forest, Survival GBoost and Survival Support Machine.

Interpretability in survival predictions

Permutation feature importance PFI was computed using the C-index as the metric provides a measure of how much each feature contributes to the predictive accuracy of a survival model (Figure 4). As one would do with outputs from any regression model, it is convenient to conduct a thorough analysis of the PFI.

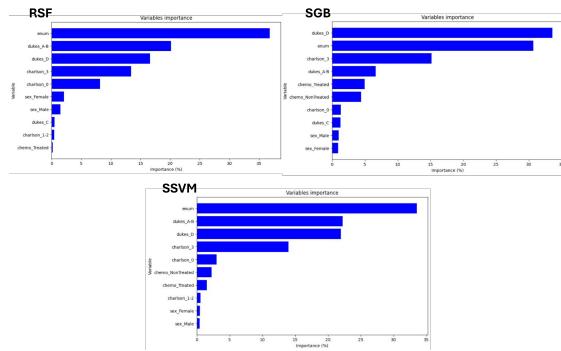


Figure 4: Permutation feature importance based on the C-index

Across all three models, 'enum' refers to the number of hospital readmissions and consistently emerges as a highly influential feature (#1 in RSF and SSVM, #2 in SGB). This underscores the importance of the number of hospitalizations in predicting survival outcomes. The Dukes' stage, particularly modality D , is a strong predictor in the RSF and SSVM models, while modalities $A - B$ are more influential in the SGB model. The Charlson comorbidity index, especially modality 3, is an impactful predictor across all models. Higher Charlson score define higher number of comorbidities. A higher score with a higher impact on mortality is then expected. Sex and chemotherapy treatment do not seem impactful overall.

Generally, PFI values are expected to be non-negative, as they quantify the increase in prediction error — or conversely, the decrease in model performance—when the values of a feature are randomly shuffled. However, the occurrence of negative PFI values can suggest potential overfitting within the model. Overfitting implies that the model has learned the training data too closely, including noise and outliers. Consequently, when a feature is permuted, introducing noise, it may counterintuitively enhance the model's performance on the validation set by mitigating the overfitting effect. Encountering negative PFI values warrants a thorough investigation into their underlying causes. It necessitates a careful examination of the model's structure, the integrity and relevance of the data, and the suitability of the chosen evaluation metric to confirm the accuracy of the PFI computation and the model's appropriateness for the data at hand.

PFI is subject to several limitations. A primary concern is its inability to accurately capture the importance of features that are highly correlated with each other. When one feature is permuted, the model may still rely on the correlated features to make predictions, leading to an underestimation of the permuted feature's importance. Additionally, permuting features can create unrealistic data points that are outside the joint distribution of the training data. Utilizing standardized features may mitigate this issue to some extent. It is also crucial to note that PFI does not offer causal explanations; it merely reflects the extent to which the model's predictions depend on each feature, without asserting any causal influence of the feature on the outcome.

Using SHAP Before reading the results on the dataset, we shall recall how to interpret SHAP values. We suggest to start looking at the magnitude, which indicates the strength of the impact a feature has on the model's prediction. Larger values (positive or negative) mean the feature has a bigger impact. Then, the sign of the SHAP value indicates the direction of the impact: positive SHAP values indicate that the presence of the feature pushes the model's prediction higher; and negative SHAP values indicate that the presence of the feature pushes the model's prediction lower. Besides, SHAP values are calculated with respect to a baseline, which is typically the average prediction of the model over the training set. Each SHAP value indicates how much a feature's value for a given observation changes the prediction from the baseline. Finally, the sum of all SHAP values for a given prediction, plus the baseline value, equals the actual prediction for

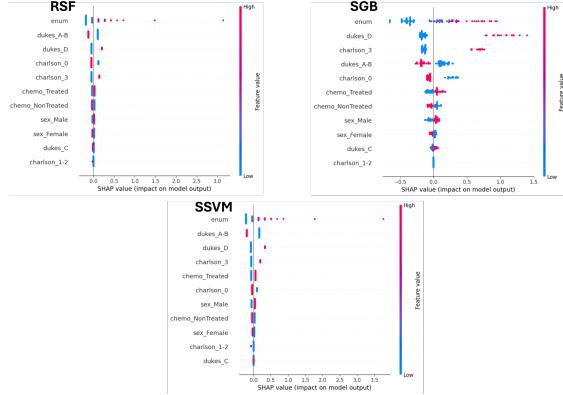


Figure 5: Interpretability using SHAP

that instance.

Figure 5 is the distribution of the impact of each feature across all data points. Overall, RSF and SSVM do not report high magnitudes across the different features. The ranking is however consistent with PFI values investigated above. We observe similar signs for each feature across the three models. For example, higher magnitudes and positive values are reported for the number of hospital readmissions. This means that not only this feature has a considerable importance, but it evolves in the same way as the model output, i.e. mortality.

A more nuanced approach to interpreting SHAP explanations involves examining response curves, which depict the evolution of the average contribution as a function of changes in the variable's value. An illustrative example of this can be seen in Figure 6, where the relationship between the variable's influence and the model's output is clearly visualized.

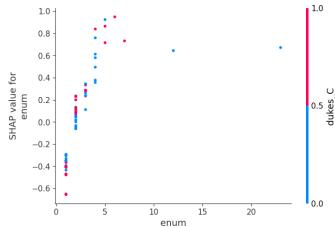


Figure 6: SHAP values between the number of hospital readmission and Duke stage

A significant benefit of SHAP values over PFI is their ability to detect and quantify feature interactions, providing a deeper understanding of intricate relationships within the data. Nevertheless, when dealing with high-dimensional datasets, the interpretation of SHAP values can become daunting. In such cases, employing dimensionality reduction techniques may be necessary to distill the insights from SHAP values into a more comprehensible form.

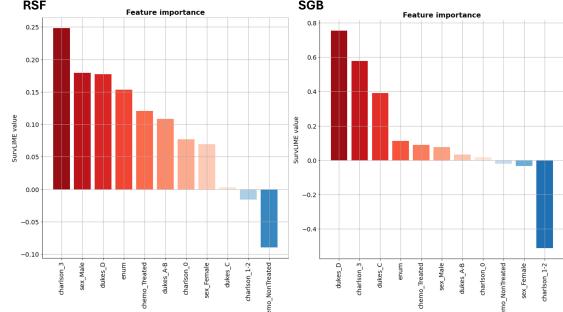


Figure 7: Interpretability using SurvLIME

Using SurvLIME In SurvLIME, each feature's influence on the prediction is quantified through coefficients that the method provides (Figure 7). The absolute value of the coefficients indicates the strength of the impact. Larger values mean the feature has a more significant effect on the model's prediction.

Besides, SurvLIME libraries offer many accessible visualization tools and provide visual aids can help in interpreting the results. For example, survival curves from the ML model representing the original complex model's predictions and the SurvLIME curve showing the approximated predictions based on the local explanation provided by the Cox model can help in an intuitive reasoning. Comparing these two curves allows for an assessment of the SurvLIME model's accuracy in capturing the black-box model's behavior. A close fit between the curves suggests a reliable explanation, whereas a significant divergence might indicate that the SurvLIME model is not fully accounting for the ML model's predictive nuances.

However, that SurvLIME's strength lies in providing explanations for individual predictions rather than a global understanding of the model's overall behavior.

Importance of hyperparameters

In exploring the impact of hyperparameter across RSF, SGB and SSVM models, distinct patterns emerge that highlight the varying influence of specific hyperparameters on each model's performance (Figure 8).

For the RSF model, the number of trees (*n_estimators*) and the maximum depth of the trees (*max_depth*) are the most influential hyperparameters. These parameters significantly affect the model's ability to capture complex patterns and prevent overfitting. In contrast, the SGB model places the greatest importance on the learning rate, and then the number of trees. The learning rate, in particular, is critical as it determines the step size at each iteration while moving toward a minimum of a loss function, directly impacting the model's convergence and generalization. Finally, the SSVM model relies on a different set of influential hyperparameters. The gamma parameter, which defines the influence of a single training example, and the maximal number of iterations are the most impactful. These scores suggest that

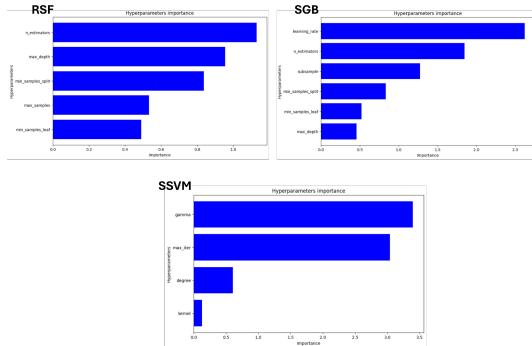


Figure 8: Impact of hyperparameters in terms of importance

the convergence criteria and the complexity of the model are paramount in determining the SSVM's performance.

Towards a robust pipeline for survival ML models evaluation in performance and interpretability

From model design and interpretation of provided results, we designed an open-source user-friendly pipeline designed in Figure 9. The associated code will be available in Github upon acceptance.

The first step consists in the choice and fine-tuning of an appropriate survival model, ensuring it is adapted to the specific characteristics of the data. This involves an adequate balance of choosing the right model, from traditional CPH model to advanced ML techniques. Additionally, the pipeline integrates how to carefully optimize hyperparameters based on the two-step approach, to maximize model performance. Once the model is finely calibrated, the subsequent step entails a rigorous evaluation of its predictive performance, employing relevant metrics such as the C-index or the IBS to ensure accuracy and reliability. The third step is dedicated to interpretability. We have demonstrated that this is a crucial aspect that provides deep insights into the model's predictive mechanisms and guarantees transparency in the model's decision-making process. Finally, the fourth step involves a detailed analysis of the impact of hyperparameters, as they enable to lead to less complex models.

Discussion

Physicians can be seen as noisy decision systems and introduce variability into the clinical decision-making process. Integrating AI-based copilots alongside them aims to mitigate this noise, enhancing decision-making accuracy and improving patient care outcomes. This study contributes to the wider adoption of AI-driven decision-making in medical settings by offering a comprehensive overview of tailored interpretability techniques specifically designed for survival endpoints.

Our primary recommendations for employing interpretability techniques in survival models are the following. For permutation feature importance, we recommended to carefully analyze the model's structure and data integrity,

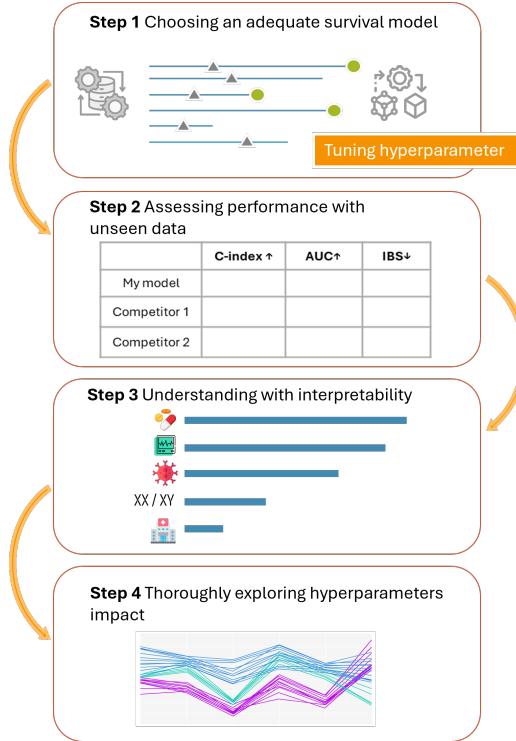


Figure 9: Proposed pipeline for survival ML model performance and interpretability

and to be aware of its limitations, such as the inherent challenge of interpreting features with high collinearity. SurvSHAP is advised for computing time-dependent feature attributions, with attention to the magnitude and sign of SHAP values for understanding feature impact, while also considering the baseline prediction. Lastly, SurvLIME is a good option for its ability to quantify the impact of features on individual predictions using coefficients, emphasizing its utility in providing instance-specific explanations rather than broad model insights, complemented by its user-friendly visualization tools for enhanced interpretability.

We have explored the benefits and limitations of *model-agnostic* interpretability methods. While these methods are versatile and widely applicable, there is a growing need for model-specific techniques that are optimized for survival analysis. For instance, TreeShap is a potent interpretability tool designed for ensemble models that use decision trees as weak learners (Campbell et al. 2022). Adapting TreeShap to accommodate the aspects of survival trees would involve minor modifications, yet it could significantly enhance the interpretability of these models.

It is important to note that the interpretability methods discussed should not be conflated with tools for causal inference. Their primary purpose is to aid in understanding, debugging, and refining model predictions rather than

establishing causal relationships. However, interpretability methods have been developed in classification tasks to provide causal explanations, such as Structural Causal Models, Causal Bayesian Networks, or counterfactual reasoning (Moraffah et al. 2020). The introduction of analogous causal interpretability approaches in survival analysis is highly anticipated, as they would offer valuable insights into the causal mechanisms underlying time-to-event data.

Above considerations regarding interpretability extend beyond the HTA process, as suggested by the authors. Interpretability may indeed also serve as an indicator of openness. Open data initiatives and collaborative research networks foster transparency, collaboration, and innovation within the biotech and medtech sectors. However, the prominence of intellectual property protection and regulatory compliance poses significant barriers to full openness. Striking a balance between openness and protection of intellectual property rights is essential for fostering innovation while ensuring responsible research and development in biotech and medtech (Kshirsagar et al. 2021). Addressing these challenges requires ongoing dialogue and collaboration among stakeholders to promote transparency, facilitate knowledge sharing, and advance scientific discovery in medical research (for a compelling example, refer to Hyland et al. (2020)). This is also in line with ongoing collaboration between the FDA's Centers for Biologics Evaluation and Research, Center for Drug Evaluation and Research, Center for Devices and Radiological Health, and Office of Combination Products to ensure the safe and ethical development and use of AI in medical products (US. FDA 2024b).

Conclusion

In conclusion, while our emphasis has been on evaluating AI-based MD, it is important to acknowledge that all components discussed are applicable to a broader spectrum of applications. These insights remain relevant whenever non-intrinsically interpretable survival models are utilized, extending the reach of our findings beyond the specific context of HTA. By highlighting the generalizability of our approach, we aim to facilitate the adoption and implementation of interpretability methods across diverse domains, fostering transparency, trust, and informed decision-making.

Ethical statement

In this study, we have adhered to ethical standards concerning the use of open-source data and software. The data utilized in our research is sourced from open repositories that ensure the privacy and anonymity of individuals. The code developed for our analysis is written in Python and will be made available as open-source upon the acceptance.

The proposed pipeline is designed with the intent to foster communication between physicians and data scientists at a specific hospital, which, for the purpose of maintaining anonymity during the review process, is not disclosed in this manuscript.

Additionnally, we acknowledge that the interpretability methods employed in our study are computationally greedy.

References

- Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.
- Alabdallah, A.; Pashami, S.; Rögnvaldsson, T.; and Ohlsson, M. 2022. SurvSHAP: a proxy-based algorithm for explaining survival models with SHAP. In *2022 IEEE 9th international conference on data science and advanced analytics (DSAA)*, 1–10. IEEE.
- Asan, O.; Bayrak, A. E.; and Choudhury, A. 2020. Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet research*, 22(6): e15154.
- Bharati, S.; Mondal, M. R. H.; and Podder, P. 2023. A review on explainable artificial intelligence for healthcare: why, how, and when? *IEEE Transactions on Artificial Intelligence*.
- Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.
- Campbell, T. W.; Roder, H.; Georgantas III, R. W.; and Roder, J. 2022. Exact Shapley values for local and model-true explanations of decision tree ensembles. *Machine Learning with Applications*, 9: 100345.
- Cottin, A.; Zulian, M.; Péchuet, N.; Guilloux, A.; and Katsahian, S. 2024. MS-CPFI: A model-agnostic Counterfactual Perturbation Feature Importance algorithm for interpreting black-box Multi-State models. *Artificial Intelligence in Medicine*, 147: 102741.
- Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2): 187–220.
- Emmerson, J.; and Brown, J. 2021. Understanding survival analysis in clinical trials. *Clinical Oncology*, 33(1): 12–14.
- Farah, L.; Murris, J. M.; Borget, I.; Guilloux, A.; Martelli, N. M.; and Katsahian, S. I. 2023. Assessment of performance, interpretability, and explainability in artificial intelligence-based health technologies: what healthcare stakeholders need to know. *Mayo Clinic Proceedings: Digital Health*, 1(2): 120–138.
- Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.
- Friedman, J. H. 2002. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4): 367–378.
- González, J. R.; Fernandez, E.; Moreno, V.; Ribes, J.; Peris, M.; Navarro, M.; Cambray, M.; and Borràs, J. M. 2005. Sex differences in hospital readmission among colorectal cancer patients. *Journal of Epidemiology & Community Health*, 59(6): 506–511.
- Graf, E.; Schmoor, C.; Sauerbrei, W.; and Schumacher, M. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17–18): 2529–2545.
- Group, M. D. C.; et al. 2021. Guidance on classification of medical devices. *MDCG*, 24: 2021–10.

- Harrell, F. E.; Califf, R. M.; Pryor, D. B.; Lee, K. L.; and Rosati, R. A. 1982. Evaluating the yield of medical tests. *Jama*, 247(18): 2543–2546.
- Harrell, F. E.; Lee, K. L.; Califf, R. M.; Pryor, D. B.; and Rosati, R. A. 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2): 143–152.
- Haute Autorité de Santé. 2020. Dossier Submission to the Medical Device and Health Technology Evaluation Committee. https://www.has-sante.fr/upload/docs/application/pdf/2020-10/guide_dm_vf_english_publi.pdf.
- Hou, Z.; Leng, J.; Yu, J.; Xia, Z.; and Wu, L.-Y. 2023. PathExpSurv: pathway expansion for explainable survival analysis and disease gene discovery. *BMC bioinformatics*, 24(1): 434.
- Hyland, S. L.; Faltys, M.; Hüser, M.; Lyu, X.; Gumbusch, T.; Esteban, C.; Bock, C.; Horn, M.; Moor, M.; Rieck, B.; et al. 2020. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3): 364–373.
- Ishwaran, H.; Kogalur, U. B.; Blackstone, E. H.; and Lauer, M. S. 2008. Random survival forests.
- Jenkins, S. P. 2005. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42: 54–56.
- Jiménez-Luna, J.; Grisoni, F.; and Schneider, G. 2020. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10): 573–584.
- Kaplan, E. L.; and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282): 457–481.
- Kovalev, M.; Utkin, L.; Coolen, F.; and Konstantinov, A. 2021. Counterfactual explanation of machine learning survival models. *Informatica*, 32(4): 817–847.
- Kovalev, M. S.; Utkin, L. V.; and Kasimov, E. M. 2020. SurvLIME: A method for explaining machine learning survival models. *Knowledge-Based Systems*, 203: 106164.
- Krzyżysiński, M.; Spytek, M.; Baniecki, H.; and Biecek, P. 2023. SurvSHAP (t): time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262: 110234.
- Kshirsagar, M.; Robinson, C.; Yang, S.; Gholami, S.; Klyuzhin, I.; Mukherjee, S.; Nasir, M.; Ortiz, A.; Oviedo, F.; Tanner, D.; et al. 2021. Becoming good at ai for good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 664–673.
- Langbein, S. H.; Krzyżysiński, M.; Spytek, M.; Baniecki, H.; Biecek, P.; and Wright, M. N. 2024. Interpretable Machine Learning for Survival Analysis. arXiv:2403.10250.
- LaRosa, E.; and Danks, D. 2018. Impacts on trust of healthcare AI. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 210–215.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Molnar, C. 2020. *Interpretable machine learning*. Lulu.com.
- Moncada-Torres, A.; van Maaren, M. C.; Hendriks, M. P.; Siesling, S.; and Geleijnse, G. 2021. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific reports*, 11(1): 6968.
- Moraffah, R.; Karami, M.; Guo, R.; Raglin, A.; and Liu, H. 2020. Causal interpretability for machine learning—problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1): 18–33.
- Moreno-Sánchez, P. A. 2023. Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *Frontiers in Cardiovascular Medicine*, 10: 1219586.
- Nazar, M.; Alam, M. M.; Yafi, E.; and Su’ud, M. M. 2021. A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. *IEEE Access*, 9: 153316–153348.
- Organization, W. H.; et al. 2011. Health technology assessment of medical devices.
- Park, C.-W.; Seo, S. W.; Kang, N.; Ko, B.; Choi, B. W.; Park, C. M.; Chang, D. K.; Kim, H.; Kim, H.; Lee, H.; et al. 2020. Artificial intelligence in health care: Current applications and issues. *Journal of Korean medical science*, 35(42).
- Pavlovic, M.; Teljeur, C.; Wieseler, B.; Klemp, M.; Cleemput, I.; and Neyt, M. 2014. ENDPOINTS FOR RELATIVE EFFECTIVENESS ASSESSMENT (REA) OF PHARMACEUTICALS. *International Journal of Technology Assessment in Health Care*, 30(5): 508–513.
- Quazi, S. 2022. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8): 120.
- Regier, D. A.; Pollard, S.; McPhail, M.; Bubela, T.; Hanna, T. P.; Ho, C.; Lim, H. J.; Chan, K.; Peacock, S. J.; and Weymann, D. 2022. A perspective on life-cycle health technology assessment and real-world evidence for precision oncology in Canada. *NPJ Precision Oncology*, 6(1): 76.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. ” Why should i trust you? ” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.
- Sarica, A.; Aracri, F.; Bianco, M. G.; Arcuri, F.; Quattrone, A.; Quattrone, A.; and Initiative, A. D. N. 2023. Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer’s disease. *Brain Informatics*, 10(1): 31.
- Secinaro, S.; Calandra, D.; Secinaro, A.; Muthurangu, V.; and Biancone, P. 2021. The role of artificial intelligence in healthcare: a structured literature review. *BMC medical informatics and decision making*, 21: 1–23.

- Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, 3145–3153. PMLR.
- Singh, H. 2022. Fair, Robust, and Data-Efficient Machine Learning in Healthcare. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 914–914.
- Srinivasu, P. N.; Sandhya, N.; Jhaveri, R. H.; and Raut, R. 2022. From blackbox to explainable AI in healthcare: existing tools and case studies. *Mobile Information Systems*, 2022: 1–20.
- Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.
- US. FDA. 2023. Artificial Intelligence Program: Research on AI/ML-Based Medical Devices. <https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/artificial-intelligence-program-research-aiml-based-medical-devices>.
- US. FDA. 2024a. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>.
- US. FDA. 2024b. Artificial Intelligence and Machine Learning in Software as a Medical Device. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>.
- US. FDA, Health Canada and MHRA. 2021. Good Machine Learning Practice for Medical Device Development: Guiding Principles. <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>.
- Vabalas, A.; Gowen, E.; Poliakoff, E.; and Casson, A. J. 2019. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11): e0224365.
- Van Belle, V.; Pelckmans, K.; Suykens, J. A.; and Van Huffel, S. 2008. Survival SVM: a practical scalable algorithm. In *ESANN*, 89–94.
- Van Ness, M.; and Udell, M. 2024. Interpretable Prediction and Feature Selection for Survival Analysis. *arXiv preprint arXiv:2404.14689*.
- Vollmer, S.; Mateen, B. A.; Bohner, G.; Király, F. J.; Ghani, R.; Jonsson, P.; Cumbers, S.; Jonas, A.; McAllister, K. S.; Myles, P.; et al. 2018. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. *arXiv preprint arXiv:1812.10404*.
- Wang, P.; Li, Y.; and Reddy, C. K. 2019. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6): 1–36.
- Wang, S.; Liu, Y.; Zhang, H.; and Liu, Z. 2024. SurvConvMixer: robust and interpretable cancer survival prediction based on ConvMixer using pathway-level gene expression images. *BMC bioinformatics*, 25(1): 133.
- Xu, L.; and Guo, C. 2023. CoxNAM: An interpretable deep survival analysis model. *Expert Systems with Applications*, 227: 120218.

IV.3 Une méthode d'interprétabilité *model-specific* pour la survie

Note : Cette sous-section propose une piste de développement de recherche autour de l'interprétabilité pour l'analyse de survie. Par conséquent, les algorithmes mentionnés ne sont pas détaillés, leur principe est simplement illustré.

IV.3.1 Introduction de TreeShap

Des méthodes d'interprétabilité propres aux modèles (*model-specific*) de ML basés sur des arbres de décision ont été proposées. En particulier, TreeSHAP, introduit par [Lundberg et al. \[2018\]](#), est une méthode d'interprétabilité spécifiquement développée pour estimer les valeurs de Shapley de modèles de ML basés sur des arbres. Les valeurs de [Shapley et al. \[1953\]](#) mesurent la contribution de chaque covariable à la différence entre la prédiction du modèle pour un échantillon donné et la prédiction moyenne du modèle. Une mesure d'interprétabilité globale peut être dérivée de l'agrégation des valeurs de Shapley sur plusieurs instances [\[Lundberg and Lee, 2017\]](#).

Dans TreeSHAP, les estimations des valeurs de Shapley pour une instance i sont obtenues à l'aide d'un algorithme de perturbation des covariables dépendant du chemin (*path-dependent feature perturbation algorithm*, détaillé de manière extensive dans [Lundberg et al. \[2018\]](#)). Le principe consiste à calculer les éléments suivants pour chaque feuille et chaque élément i sur le chemin menant à cette feuille :

- La proportion des sous-ensembles S à la feuille qui contiennent i et la proportion des sous-ensembles S qui ne contiennent pas i ;
- Pour chaque cardinalité, la proportion des ensembles de cette cardinalité contenus dans la feuille.

En tenant compte des chemins allant de la racine aux feuilles et des prédictions des nœuds, les valeurs de Shapley sont calculées comme suit :

$$\phi_i = \sum_{j=1}^L \sum_{P \in S_j} \frac{w(|P|, j)}{M_j \binom{|P|}{M_j}} (p_o^{i,j} - p_z^{i,j}) v_j, \quad (\text{IV.1})$$

où S_j est l'ensemble des sous-ensembles d'éléments présents à la feuille j , L est le nombre de feuilles, M_j est la longueur du chemin et $w(|P|, j)$ est la proportion de tous les sous-ensembles de cardinalité P à la feuille j , $p_o^{i,j}$ et $p_z^{i,j}$ représentent les fractions des sous-ensembles qui contiennent ou ne contiennent pas l'élément i respectivement, et v_j est la valeur de la feuille avec l'indice j , c'est-à-dire la sortie du modèle. Ainsi, pour chaque covariable, TreeShap calcule la contribution à la prédiction le long des différents chemins de l'arbre, pondérée par la probabilité de suivre chaque chemin.

Nous considérons un jeu de données fictif de 10 échantillons avec trois variables indépendantes numériques x, y, z et une variable cible *value*, par exemple la taille d'une tumeur.

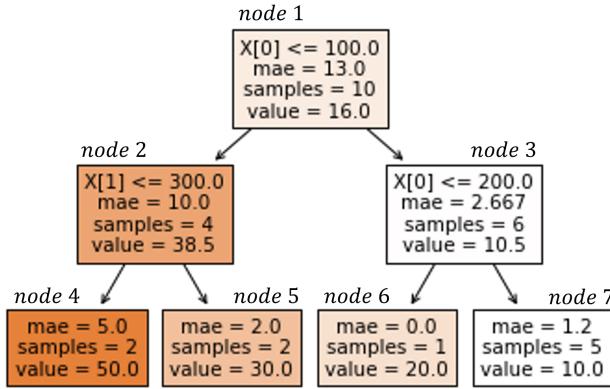


FIGURE IV.2 – Arbre de décision simple pour l'illustration de TreeSHAP ($x = X[0]$ et $y = X[1]$)

Pour l'illustration de TreeSHAP, nous considérons un arbre de régression simple comme présenté dans la Figure IV.2. Le programme Python est disponible en Annexe C.

Nous obtenons les attributions SHAP en commençant par un modèle vide sans aucune variable, puis en calculant la contribution marginale moyenne lorsque chaque variable est ajoutée à ce modèle dans une séquence. La moyenne est calculée sur l'ensemble des séquences possibles. Prenons une instance i avec

$$Z_i = \{x = 150, y = 75, z = 200\}. \quad (\text{IV.2})$$

La prédiction pour cette instance est $value = 20$. Ici, nous avons 3 variables indépendantes, nous devons donc considérer $3! = 6$ séquences. Nous allons calculer les contributions marginales pour chaque séquence. SHAP suppose que la prédiction du modèle avec n'importe quel sous-ensemble S de variables indépendantes est la valeur attendue de la prédiction compte tenu du sous-ensemble. La prédiction pour le modèle nul ϕ^0 (également appelée valeur de base) est prédiction moyenne pour l'ensemble d'apprentissage, ici égale à $(50 \cdot 2 + 30 \cdot 2 + 20 \cdot 1 + 10 \cdot 5)/10 = 23$.

Nous commençons par la séquence $x > y > z$:

1. En premier lieu, la variable x est ajoutée au modèle nul. Pour l'instance i sélectionnée, nous pouvons calculer la prédiction exacte avec cette seule information, car seule la variable x est utilisée dans les noeuds ($node 1$ et $node 3$) menant au noeud feuille $node 6$. La prédiction du modèle avec la seule variable x est donc de 20. Par conséquent, la contribution marginale de x dans cette séquence, $\phi_x^1 = 20 - 23 = -3$.
2. Ensuite nous ajoutons la variable y au modèle ci-dessus (à l'étape 1). Étant donné que l'ajout de y ne modifie pas la prédiction pour l'instance i sélectionnée, la contribution marginale de y dans cette séquence, $\phi_y^1 = 20 - 20 = 0$.
3. De même, la contribution marginale pour z dans cette séquence, $\phi_z^1 = 0$.

Nous considérons ensuite la séquence $y > z > x$:

1. Tout d'abord, la variable y est ajoutée au modèle nul. Le premier noeud $node 1$ utilise x comme variable de séparation, puisque x n'est pas encore disponible, nous



FIGURE IV.3 – Représentation d'une explication SHAP pour une instance i

calculons la prédiction comme $(4/10) \cdot (\text{prédiction du nœud enfant gauche } node 2) + (6/10) \cdot (\text{prédiction du nœud enfant droit } node 3)$; 100, 60 et 40 étant le nombre d'échantillons d'apprentissage tombant dans les nœuds $node 1$, $node 2$ et $node 3$ respectivement. La prédiction du $node 2$ est 50, et la prédiction du $node 3$ est $(1/6) \cdot 20 + (5/6) \cdot 10 = 70/6$. On a alors $\phi_y^2 = 27 - 23 = 4$.

2. Ensuite, nous ajoutons la variable z au modèle ci-dessus. Étant donné que z n'est pas utilisé comme variable de division dans les nœuds internes de l'arbre, l'ajout de cette caractéristique ne modifie en rien la prédiction. Ainsi, la contribution marginale de z dans cette séquence, $\phi_z^2 = 0$.
3. Enfin, nous ajoutons la variable x au modèle, ce qui donne une prédiction de 20. Par conséquent, la contribution marginale de x dans cette séquence est $\phi_x^2 = 20 - 27 = -7$.

De la même façon, nous calculons la contribution marginale de chaque valeur des variables pour les séquences restantes :

- Séquence $x > z > y : \phi_x^3 = -3, \phi_y^3 = 0, \phi_z^3 = 0,$
- Séquence $z > x > y : \phi_x^4 = -3, \phi_y^4 = 0, \phi_z^4 = 0,$
- Séquence $z > y > x : \phi_x^5 = -7, \phi_y^5 = 4, \phi_z^5 = 0,$
- Séquence $y > x > z : \phi_x^6 = -7, \phi_y^6 = 4, \phi_z^6 = 0.$

Ainsi, les attributions SHAP pour l'instance i sont :

$$\begin{cases} \phi_x = \frac{\sum_{k=1}^6 \phi_x^k}{6} = -5, \\ \phi_y = \frac{\sum_{k=1}^6 \phi_y^k}{6} = 2, \\ \phi_z = \frac{\sum_{k=1}^6 \phi_z^k}{6} = 0. \end{cases} \quad (\text{IV.3})$$

L'explication de la prédiction pour l'instance i est égale à $\phi^0 + \phi_x + \phi_y + \phi_z = 23 + (-5) + 2 + 0 = 20$. La valeur de base de la prédiction en l'absence de toute information sur les variables indépendantes est 23. Le fait de connaître $x = 150$ a diminué la prédiction de 5 et le fait de connaître $y = 75$ a augmenté la prédiction de 2, ce qui donne une prédiction finale de 20. La connaissance de $z = 300$ n'a eu aucun impact sur la prédiction du modèle. SHAP fournit une représentation visuelle de cette explication (Figure IV.3). La couleur bleue indique que la valeur $x = 150$ a diminué la prédiction et la couleur rouge indique que la valeur $y = 75$ a augmenté la prédiction.

La complexité de l'algorithme ci-dessus est de l'ordre de $O(LB2^M)$, où B est le nombre d'arbres dans le modèle, L est le nombre maximum de feuilles dans chaque arbre et M est le nombre de variables. Dans l'article TreeSHAP, Lundberg et al. [2018] proposent une version modifiée de cet algorithme qui tient compte du nombre de sous-ensembles S qui

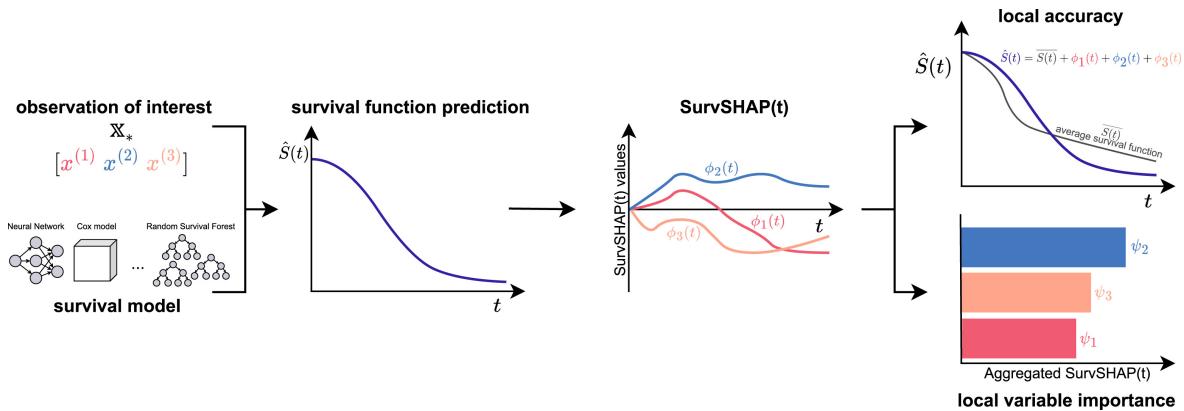


FIGURE IV.4 – Schéma de l'algorithme SurvSHAP (Source : Krzyżysiński et al. [2023])

entrent dans chaque noeud de l'arbre. L'algorithme modifié a une complexité de calcul de $O(LBD^2)$ où D est la profondeur maximale de l'arbre.

IV.3.2 Extensions possibles pour la survie

En intégrant des explications calculées et évaluées relativement à la fonction de survie, on peut adapter TreeSHAP aux temps dépendants.

L'algorithme SurvSHAP de Krzyżysiński et al. [2023] génère des explications à chaque temps pour les modèles de ML de survie (Figure IV.4). $SurvSHAP(t)$ permet d'expliquer les prédictions d'un modèle de survie en fonction du temps. Les valeurs de $SurvSHAP(t)$ s'ajoutent à la fonction de survie prédictive par le modèle et, agrégées dans le temps, peuvent être traitées comme une mesure locale de l'importance des variables.

La contribution d'une observation donnée Z_i au temps t est une espérance conditionnelle et s'écrit

$$e^D(Z_i, t) = \mathbb{E}[\hat{S}(t|Z)|Z = Z_i^D], \quad (\text{IV.4})$$

avec $S(t|Z)$ la fonction de survie, et le conditionnement s'applique à toutes les variables de l'ensemble D .

Les fonctions $\phi_{t_0}(Z_i, j), \dots, \phi_{t_{\max}}(Z_i, j)$ pour toutes les covariables $j \in \{1, \dots, p\}$ mesurent les attributions (i.e. une valeur d'importance) en fonction du temps pour expliquer la prédiction du modèle. L'estimation de $SurvSHAP(t)$ est effectuée via l'algorithme d'échantillonnage classique des valeurs de Shapley, ce qui permet de définir la contribution de la covariable j à l'observation Z_i comme suit

$$\phi_t(Z_i, j) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} (e(Z_i, t)^{\text{before}(\pi, j) \cup j} - e(Z_i, t)^{\text{before}(\pi, j)}), \quad (\text{IV.5})$$

avec Π l'ensemble des permutations des covariables $\{1, \dots, p\}$, $\text{before}(\pi, j)$ est le sous-ensemble des covariables juste avant j dans l'ordre de $\pi \in \Pi$. Pour faciliter la comparaison entre différents modèles et points dans le temps, cette valeur peut être normalisée pour

obtenir des valeurs sur une échelle comparable, et on a

$$\phi_t^*(Z_i, j) = \frac{\phi_t(Z_i, j)}{\sum_{l=1}^p |\phi_t(Z_i, l)|}. \quad (\text{IV.6})$$

Extensions possibles pour la survie avec événements récurrents Les explications fournies par *SurvSHAP*(t) sont basées sur la fonction de survie. En considérant dans l'algorithme une fonction moyenne cumulée au sens de [Cook and Lawless \[1997\]](#) (*mean cumulative function*, équation I.33, Chapitre I), cela pourrait mener à prendre en compte les événements récurrents pour les individus :

$$e_\mu^D(Z_i, t) = \mathbb{E}[\hat{\mu}(t|Z)|Z = Z_i^D]. \quad (\text{IV.7})$$

Les valeurs de *SurvSHAP*(t) s'ajouteraient alors à la fonction moyenne cumulée prédite par le modèle et, agrégées dans le temps, pourraient être traitées comme un classement local de l'importance des variables.

Ainsi, la combinaison de TreeSHAP et de SurvSHAP aboutirait à une méthode d'interprétabilité spécifique aux modèles de ML de survie basés sur les arbres.

IV.4 Discussion

Les cliniciens sont de plus en plus sensibilisés à l'intégration des outils d'IA dans leurs activités quotidiennes, et il est essentiel qu'ils puissent faire confiance aux décisions prises par ces algorithmes pour les intégrer efficacement dans leur pratique [[LaRosa and Danks, 2018](#)]. Si le raisonnement derrière ces décisions reste opaque, cela peut engendrer scepticisme et réticence à adopter la technologie [[Liao et al., 2020](#)]. Par ailleurs, [LaRosa and Danks \[2018\]](#) soulignent la nécessité de former les médecins aux fonctionnements et à l'utilisation des systèmes d'IA, tout en donnant la possibilité aux patients de donner leur consentement de manière éclairée avant que l'IA ne soit utilisée dans leur prise en charge. Les modèles capables de donner des garanties quant à leur transparence et leur interprétabilité favorisent ainsi la confiance et augmentent la probabilité d'adoption de ces technologies [[Asan et al., 2020](#)].

La recherche clinique pour la découverte de nouveaux traitements peut également grandement bénéficier des modèles capables de fournir des explications à propos des mécanismes sous-jacents guidant leur prise de décision, éclairant ainsi à propos des relations complexes propres aux données des patients [[Jiménez-Luna et al., 2020](#)]. Les cliniciens peuvent alors exploiter ces nouvelles connaissances pour mieux comprendre les interactions entre maladie, traitement, et réaction du patient, améliorant ainsi sa prise en charge.

Ce chapitre avait pour objectif de mettre en perspective l'importance des méthodes d'interprétabilité permettant d'améliorer la confiance d'un utilisateur envers un algorithme. De ce fait, la revue de littérature des critères d'évaluation des différentes agences de réglementation de technologie de santé dans le monde a souligné que l'interprétabilité faisait l'objet

CHAPITRE IV Interprétabilité, santé et survie : vers une utilisation plus transparente des algorithmes d'IA

d'un intérêt croissant (Section IV.1). Toutefois, cette analyse doit être épaulée par une application concrète (Section IV.2). Ainsi, il est essentiel de démontrer l'intérêt de ces méthodes à travers des exemples d'implémentations illustrant la manière dont on pouvait en tirer des explications. Le code associé a été rendu public pour faciliter la reproductibilité. Enfin, pour faire écho à la méthode développée et présentée au Chapitre III, nous avons proposé une piste de développement d'une nouvelle méthode d'interprétabilité de forêts aléatoires de survie, en présence ou non d'événements récurrents. Cette nouvelle approche vise à améliorer la transparence des modèles de ML de survie, en donnant la possibilité aux cliniciens de générer des explications adaptées.

En conclusion, l'intégration des méthodes d'interprétabilité dans les algorithmes médicaux est indispensable pour renforcer la confiance des utilisateurs, améliorer la prise en charge des patients et promouvoir l'adoption des technologies d'IA dans le domaine médical. Ce chapitre a mis en lumière l'importance de ces méthodes et a proposé des solutions concrètes.

Messages-clés de ce chapitre

Ce chapitre aborde un sujet d'une importance grandissante, particulièrement dans le domaine de la santé : l'**interprétabilité** et l'**explicabilité** des algorithmes d'apprentissage automatique. La contribution principale de ce chapitre réside dans l'**état des lieux détaillé** des critères d'interprétabilité et d'explicabilité pertinents pour les **autorités de santé**. Nous avons apporté de nombreuses **clarifications** essentielles pour **améliorer la compréhension** et l'application de ces critères dans les pratiques de santé publique. Pour ancrer cette perspective dans le contexte spécifique de la survie, nous illustrons les méthodes d'interprétabilité appliquées aux algorithmes d'apprentissage couramment utilisés dans ce domaine, tels que Random Survival Forests, Survival Support Vector Machines et CoxBoost. Bien que ces méthodes soient **agnostiques aux modèles**, nous soulignons également l'existence de **méthodes d'interprétabilité spécifiquement développées pour les arbres de décisions**. En particulier, nous mettons en lumière une piste de développement sur l'utilisation de TreeShap pour l'**interprétabilité des modèles de survie, applicable non seulement aux forêts aléatoires de survie**, mais aussi à leurs extensions, y compris celles impliquant des événements récurrents. Cette approche permettrait une interprétation plus fine et adaptée aux caractéristiques spécifiques des modèles de survie basés sur des arbres. Ce chapitre apporte une contribution significative en fournissant une **vue d'ensemble des outils disponibles** pour l'interprétabilité et l'explicabilité, et en proposant des **méthodes avancées et adaptées aux besoins des autorités de santé** pour une **meilleure prise de décision**.

Conclusion

"Modeling is not only about math and methods, but it's about the mindset through which you see the world."

Leo Breiman (2001)

Synthèse des travaux

Ce projet de thèse avait d'abord pour objectif d'améliorer la compréhension des événements récurrents afin de proposer une nouvelle méthode basée sur des principes d'apprentissage automatique. À travers un état de l'art détaillé des méthodes existantes alliant méthodologie pour événement récurrent et apprentissage, nous avons pu mettre en évidence le manque d'approches pour répondre aux enjeux à la fois des données réelles, mais aussi des attentes en termes de recherche médicale [Murris et al., 2023]. Nous avons ainsi développé et validé RecForest, inspirée des mécanismes des forêts aléatoires, qui présente une approche innovante pour la prédiction d'événements récurrents [Murris et al., 2024]. RecForest a été appliquée dans le cadre des réhospitalisations postopératoires pour les patients atteints de cancer digestif.

Le second objectif était d'explorer les conditions nécessaires à l'adoption de ce type d'algorithme d'apprentissage, en mettant l'accent sur la nécessité de gagner la confiance des utilisateurs. En étudiant méthodiquement les critères d'évaluation des différentes autorités de santé à l'échelle internationale, nous avons mis en avant l'importance de l'interprétabilité et de l'explicabilité [Farah et al., 2023b]. Pour ancrer pleinement cette problématique dans le projet de recherche de la thèse, nous avons illustré les méthodes les plus courantes dans le contexte de la survie, démontrant leur pertinence et leur applicabilité dans des situations réelles.

Ainsi, notre recherche contribue à la fois à l'avancement des techniques de prédiction des événements récurrents et à la compréhension des défis associés à l'implémentation de ces méthodes dans des contextes médicaux, ouvrant la voie à des applications plus efficaces et mieux acceptées dans le domaine de la santé.

Perspectives

En cours de développement

Ce travail de recherche a encore de beaux jours devant lui, en creusant en particulier les pistes suivantes qui seront prochainement finalisées.

Validation de la cohorte RecForest a été appliquée pour les réhospitalisations post-opératoires des patients atteints de cancer digestif. Cependant, nous n'avons pas encore validé la cohorte avec tous les critères d'inclusion/exclusion et toutes les covariables. Actuellement, ce travail sert d'illustration pour cette thèse, mais une validation complète est prévue très prochainement en collaboration avec le Dr S. Tzedakis. Cette validation comprendra une analyse rigoureuse des données cliniques afin de s'assurer que tous les facteurs pertinents sont pris en compte, garantissant ainsi la robustesse et la fiabilité de notre méthode dans un contexte réel pour répondre à la problématique posée.

Développement d'un package R Le code open source est essentiel pour l'application future de RecForest. Nous travaillons actuellement à l'implémentation du package dans R, afin de permettre une utilisation plus large et de faciliter l'intégration de notre méthode dans divers projets de recherche et applications médicales. Cette implémentation vise à offrir une documentation complète et des exemples pratiques pour garantir une adoption aisée par la communauté scientifique et médicale [Wilson et al., 2017]. Aussi, nous prévoyons de développer des tutoriels pour aider les utilisateurs à se familiariser avec RecForest, à optimiser son utilisation et à adapter l'algorithme à leurs besoins spécifiques.

Vers de nouvelles collaborations Nous envisageons aussi d'autres collaborations avec des équipes de recherche pour tester RecForest dans divers contextes cliniques. Ces collaborations pourraient offrir des perspectives supplémentaires pour affiner l'algorithme et adapter ses fonctionnalités à des besoins spécifiques, tout en renforçant la crédibilité et l'adoption de RecForest dans le domaine médical.

Futures pistes de recherche

Au cours de ce manuscrit de thèse, nous avons évoqué plusieurs pistes de recherche qui sont reprises ici.

Vers de nouveaux modèles d'ensemble RecForest est une méthode d'ensemble basée sur les forêts aléatoires, offrant ainsi une approche robuste pour la prédiction et l'analyse des données. Toutefois, plusieurs autres méthodes d'ensemble pourraient bénéficier de l'intégration des arbres *weak learners* de RecForest. Cette intégration pourrait potentiellement enrichir ces méthodes en améliorant leur performance et leur adaptabilité.

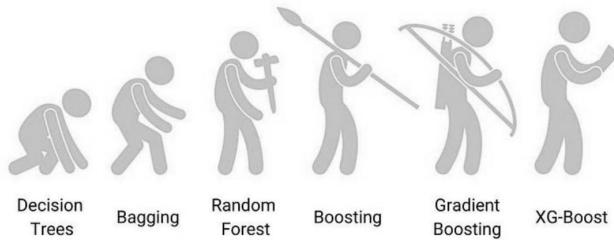


FIGURE IV.5 – Évolution des arbres de décision (Source : [Enjoy Algorithms](#))

à divers types de données et problèmes cliniques (Figure IV.5). Voici quelques-unes de ces méthodes :

- Les Gradient Boosting Machines (GBM) de [Friedman \[2001\]](#) sont actuellement utilisés pour des problèmes de régression et de classification, mais aussi de survie (avec une approche sur le temps jusqu’au premier événement) avec [Hothorn and Zeileis \[2021\]](#) ou encore [Bai et al. \[2022\]](#). Les GBM sont construits à partir d’une séquence d’arbres, où chaque nouvel arbre vise à corriger les erreurs des arbres précédents. En intégrant les arbres de RecForest dans un cadre de GBM, nous pourrions bénéficier de leur robustesse tout en optimisant les prédictions à chaque itération.
- LightGBM, proposé par [Ke et al. \[2017\]](#), est une version optimisée du gradient boosting qui utilise des histogrammes pour une gestion efficace de la mémoire et une accélération des calculs. En utilisant les arbres de RecForest, nous pourrions combiner la rapidité et l’efficacité de LightGBM avec la capacité de RecForest à capturer des interactions complexes dans les données.
- CatBoost de [Prokhorenkova et al. \[2018\]](#) est une méthode de boosting par gradient qui excelle dans le traitement des variables catégorielles grâce à ses représentations sous forme d’*embedding*. L’intégration des arbres de RecForest dans CatBoost pourrait améliorer la performance globale en profitant des capacités uniques de gestion des variables catégorielles tout en bénéficiant de la robustesse des arbres issus de RecForest.

En explorant ces intégrations, nous pourrions non seulement tirer parti des forces spécifiques de chaque méthode d’ensemble mais aussi améliorer la robustesse et la précision des prédictions fournies par RecForest [[Grinsztajn et al., 2022](#)]. Cette approche hybride permettrait d’exploiter les avantages combinés des différentes techniques d’apprentissage, créant ainsi des modèles plus performants et mieux adaptés aux exigences variées des applications cliniques et médicales.

Interprétabilité pour la survie Nous avons vu que l’interprétabilité des algorithmes d’analyse de survie était cruciale pour comprendre *comment* et *pourquoi* certaines variables influencent le temps jusqu’à un événement d’intérêt, comme une rechute ou une réhospitalisation. Dans le domaine médical, où les décisions doivent souvent être basées sur des analyses complexes et où les conséquences peuvent être graves, il est essentiel de pouvoir fournir des explications claires aux praticiens [[Markus et al., 2021](#)]. Les modèles traditionnels comme le modèle de régression de Cox [[Cox, 1972b](#)] sont intrinsèquement interprétables grâce aux coefficients estimés, qui indiquent l’effet direct des covariables sur le risque de

survenue d'un événement. Cependant, pour les algorithmes d'apprentissage automatique plus complexes, tels que les forêts aléatoires de Ishwaran et al. [2008] ou les méthodes de *boosting*, il est nécessaire de recourir à des techniques supplémentaires pour décomposer les contributions des variables et explorer les interactions complexes entre elles.

Les forêts aléatoires de survie fournissent la plupart du temps des mesures d'importance des variables, comme dans Ishwaran et al. [2008], Wang and Li [2017b], Naseije et al. [2017], Ishwaran and Lu [2019], Devaux et al. [2023a], qui fournissent une vue d'ensemble sur la contribution relative de chaque variable au modèle. Bien que ces mesures soient utiles, elles ne permettent pas toujours de saisir les interactions complexes ou les effets spécifiques des valeurs des covariables sur les prédictions de survie.

C'est ici que des outils avancés comme TreeShap de Lundberg et al. [2018] peuvent jouer un rôle important. TreeShap est une méthode dérivée des valeurs de SHAP (SHapley Additive exPlanations), qui attribue une valeur précise à chaque variable en termes d'impact sur les prédictions d'un modèle d'arbre de décision. Cette approche permet de décomposer les prédictions complexes en contributions individuelles des variables, offrant ainsi une vue détaillée de leur influence sur les prévisions de survie. En utilisant TreeShap, il est possible de visualiser comment chaque covariable affecte les prédictions de survie, fournissant ainsi des explications plus transparentes et compréhensibles.

Le Chapitre IV présente les outils nécessaires pour le développement d'une extension de TreeShap spécialement adaptée à l'analyse de survie. Cette approche constitue une piste sérieuse pour améliorer l'interprétabilité des algorithmes d'apprentissage pour la survie basés sur des arbres, en offrant des explications détaillées des prédictions issues de ces modèles complexes. La mise en œuvre de cette approche pourrait donc représenter une avancée significative pour rendre les modèles de survie plus accessibles et compréhensibles pour les praticiens, facilitant ainsi leur utilisation dans des contextes cliniques réels.

Conclusion générale

Cette thèse a mis en lumière la valeur ajoutée d'une approche intégrée qui combine les principes de l'apprentissage automatique et les méthodes statistiques traditionnelles. Comme le souligne Breiman [2001a], l'alliance de ces disciplines permet de tirer parti de leurs forces respectives, offrant des modèles à la fois robustes et flexibles. En combinant la rigueur théorique des statistiques avec les capacités de traitement avancées des méthodes d'apprentissage automatique, nous avons développé des outils qui non seulement améliorent la précision des prédictions tout en garantissant un certain niveau d'interprétabilité.

En conclusion, cette recherche offre des horizons prometteurs pour améliorer les prédictions de survie et la gestion des événements récurrents dans un contexte clinique. Elle ouvre la voie à de futures investigations visant à affiner les méthodes développées, à explorer leur application à d'autres contextes médicaux et à intégrer de nouvelles techniques pour optimiser l'impact de ces analyses sur la pratique clinique et la recherche en oncologie.

Bibliographie

- O. O. Aalen, J. Fosen, H. Weedon-Fekjær, Ø. Borgan, and E. Husebye. Dynamic analysis of multivariate failure time data. *Biometrics*, 60(3) :764–773, 2004.
- T. Abdalla, T. Walwyn, D. White, C. S. Choong, M. Bulsara, D. B. Preen, and J. L. Ohan. Hospitalizations and cost of inpatient care for physical diseases in survivors of childhood cancer in western australia : A longitudinal matched cohort study. *Cancer Epidemiology, Biomarkers & Prevention*, 32(9) :1249–1259, 2023.
- T. Abdalla, D. B. Preen, J. D. Pole, T. Walwyn, M. Bulsara, A. Ives, C. S. Choong, and J. L. Ohan. Psychiatric disorders in childhood cancer survivors : A retrospective matched cohort study of inpatient hospitalisations and community-based mental health services utilisation in western australia. *Australian & New Zealand Journal of Psychiatry*, 58(6) : 515–527, 2024.
- I. Adenot, D. Camus, A.-A. É. de Fleurian, D. Tassy, S. Bourguignon, N. Chabin, P.-Y. Chambrin, D. Costagliola, L. Huot, A.-S. Joly, et al. Early patient access to health technologies : Is innovation needed for early management ? *Therapies*, 75(1) :71–83, 2020.
- P. M. Afonso, D. Rizopoulos, A. K. Palipana, E. Gecili, C. Brokamp, J. P. Clancy, R. D. Szczesniak, and E.-R. Andrinopoulou. A joint model for (un) bounded longitudinal markers, competing risks, and recurrent events using patient registry data. *arXiv preprint arXiv* :2405.16492, 2024.
- S. Akram and Q. U. Ann. Newton raphson method. *International Journal of Scientific & Engineering Research*, 6(7) :1748–1752, 2015.
- L. Albigès, C. Bellera, S. Branchoux, M. Arnaud, A. Gouverneur, S. Néré, A.-F. Gaudin, I. Durand-Zaleski, and S. Négrier. Real-world treatment patterns and effectiveness of patients with advanced renal cell carcinoma : A nationwide observational study. *Clinical Genitourinary Cancer*, 22(2) :295–304, 2024.
- S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable artificial intelligence (xai) : What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99 :101805, 2023.
- D. G. Altman. *Practical statistics for medical research*. Chapman and Hall/CRC, 1990.
- L. D. A. F. Amorim and J. Cai. Modelling recurrent events : a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1) :324–333, 2015. ISSN 1464-3685. doi : 10.1093/ije/dyu222.
- P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes : A Large Sample Study. *The Annals of Statistics*, 10(4) :1100–1120, 1982. ISSN 0090-5364. URL <https://www.jstor.org/stable/2240714>.

-
- P. K. Andersen and N. Keiding. Multi-state models for event history analysis. *Statistical methods in medical research*, 11(2) :91–115, 2002.
- P. K. Andersen and M. Pohar Perme. Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1) :71–99, 2010.
- P. K. Andersen, Ø. Borgan, R. D. Gill, N. Keiding, P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. Nonparametric estimation. *Statistical models based on counting processes*, pages 176–331, 1993.
- P. K. Andersen, R. B. Geskus, T. de Witte, and H. Putter. Competing risks in epidemiology : possibilities and pitfalls. *International journal of epidemiology*, 41(3) :861–870, 2012.
- P. K. Andersen, J. Angst, and H. Ravn. Modeling marginal features in studies of recurrent events in the presence of a terminal event. *Lifetime Data Analysis*, 25(4) :681–695, Oct. 2019. ISSN 1572-9249. doi : 10.1007/s10985-019-09462-4.
- J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed. Supervised, unsupervised, and semi-supervised feature selection : a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5) :971–989, 2015.
- S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4 :40–79, 2010.
- O. Asan, A. E. Bayrak, and A. Choudhury. Artificial intelligence and human trust in healthcare : focus on clinicians. *Journal of medical Internet research*, 22(6) :e15154, 2020.
- P. C. Austin, D. S. Lee, and J. P. Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6) :601–609, 2016.
- M. Bai, Y. Zheng, and Y. Shen. Gradient boosting survival tree with applications in credit scoring. *Journal of the Operational Research Society*, 73(1) :39–55, 2022.
- M. Barca-Hernando, S. Lopez-Ruz, S. Marin-Romero, V. Garcia-Garcia, T. Elias-Hernandez, R. Otero-Candela, M. Carrier, and L. Jara-Palomares. Risk of recurrent cancer-associated thrombosis after discontinuation of anticoagulant therapy. *Research and Practice in Thrombosis and Haemostasis*, 7(2) :100115, 2023.
- B. K. Beaulieu-Jones, S. G. Finlayson, W. Yuan, R. B. Altman, I. S. Kohane, V. Prasad, and K.-H. Yu. Examining the use of real-world evidence in the regulatory process. *Clinical Pharmacology & Therapeutics*, 107(4) :843–852, 2020.
- D. Bertsimas and H. Wiberg. Machine learning in oncology : methods, applications, and challenges. *JCO clinical cancer informatics*, 4, 2020.
- H. P. Bhambhvani, A. Zamora, E. Shkolyar, K. Prado, D. R. Greenberg, A. M. Kasman, J. Liao, S. Shah, S. Srinivas, E. C. Skinner, et al. Development of robust artificial neural networks for prediction of 5-year survival in bladder cancer. In *Urologic oncology : seminars and original investigations*, volume 39, pages 193–e7. Elsevier, 2021.
- H. Binder and M. Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC bioinformatics*, 9 :1–10, 2008.
- C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- Z. S. Bohannan, F. Coffman, and A. Mitrofanova. Random survival forest model identifies novel biomarkers of event-free survival in high-risk pediatric acute lymphoblastic leukemia. *Computational and structural biotechnology journal*, 20 :583–597, 2022.

- D. Bona, M. Manara, G. Bonitta, G. Guerrazzi, J. Guraj, F. Lombardo, A. Biondi, M. Cavalli, P. G. Bruni, G. Campanelli, et al. Long-term impact of severe postoperative complications after esophagectomy for cancer : Individual patient data meta-analysis. *Cancers*, 16(8) :1468, 2024.
- C. M. Booth, E. A. Eisenhauer, B. Gyawali, and I. F. Tannock. Progression-free survival should not be used as a primary end point for registration of anticancer drugs. *Journal of Clinical Oncology*, 41(32) :4968–4972, 2023.
- C. Bossen, K. H. Pine, F. Cabitza, G. Ellingsen, and E. M. Piras. Data work in healthcare : An introduction, 2019.
- L. E. Bothwell, J. A. Greene, S. H. Podolsky, D. S. Jones, et al. Assessing the gold standard—lessons from the history of rcts. *N engl j med*, 374(22) :2175–2181, 2016.
- I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5 :44, 2011.
- O. Bouaziz. Assessing model prediction performance for the expected cumulative number of recurrent events. *Lifetime Data Analysis*, 30(1) :262–289, Jan. 2024. ISSN 1572-9249. doi : 10.1007/s10985-023-09610-x.
- G. E. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356) : 791–799, 1976.
- J. Braa, E. Monteiro, and S. Sahay. Networks of action : sustainable health information systems across developing countries. *MIS quarterly*, pages 337–362, 2004.
- L. Breiman. Bagging predictors. *Machine learning*, 24 :123–140, 1996.
- L. Breiman. Statistical modeling : The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3) :199–231, 2001a.
- L. Breiman. Random Forests. *Machine Learning*, 45(1) :5–32, Oct. 2001b. ISSN 1573-0565. doi : 10.1023/A:1010933404324. URL <https://doi.org/10.1023/A:1010933404324>.
- L. Breiman. *Classification and regression trees*. Routledge, 2017.
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. Cart. *Classification and regression trees*, 1984.
- N. Breslow. A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, 57(3) :579–594, 1970.
- C. C. Bridges Jr. Hierarchical cluster analysis. *Psychological reports*, 18(3) :851–854, 1966.
- G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1) :1–3, 1950.
- J. Bryan and D. Li. Comments on contemporary uses of machine learning for electronic health records. *NC Med J*, 85(4) :263–265, 2024.
- A. Böhler, R. J. Cook, and J. F. Lawless. Multistate models as a framework for estimand specification in clinical trials of complex processes. *Statistics in Medicine*, 42(9) :1368–1397, 2023.
- T. Burki. Platform trials : the future of medical research ? *The Lancet Respiratory Medicine*, 11(3) :232–233, 2023.

-
- Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3) :740–757, 2018.
- S. M. Cadarette and L. Wong. An introduction to health care administrative data. *The Canadian journal of hospital pharmacy*, 68(3) :232, 2015.
- K. J. Carroll. On the use and utility of the weibull model in the analysis of survival data. *Controlled clinical trials*, 24(6) :682–701, 2003.
- R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare : Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.
- F. Cerreta, H.-G. Eichler, and G. Rasi. Drug policy for an aging population—the european medicines agency’s geriatric medicines strategy. *New England Journal of Medicine*, 367 (21) :1972–1974, 2012.
- A. Charles-Nelson, S. Katsahian, and C. Schramm. 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*, page sim.8168, 2019. ISSN 0277-6715, 1097-0258. doi : 10.1002/sim.8168. URL <https://onlinelibrary.wiley.com/doi/10.1002/sim.8168>.
- X. Che and J. Angus. A new joint model of recurrent event data with the additive hazards model for the terminal event time. *Metrika*, 79 :763–787, 2016.
- S. H. Chiou, G. Xu, J. Yan, and C.-Y. Huang. Regression modeling for recurrent events possibly with an informative terminal event using R package reReg. *Journal of Statistical Software*, 105(5) :1–34, 2023. doi : 10.18637/jss.v105.i05.
- H. A. Chipman, E. I. George, and R. E. McCulloch. Bart : Bayesian additive regression trees. 2010.
- A. Ciampi, R. Bush, M. Gospodarowicz, and J. Till. An approach to classifying prognostic factors related to survival experience for non-hodgkin’s lymphoma patients : Based on a series of 982 patients : 1967–1975. *Cancer*, 47(3) :621–627, 1981.
- A. Ciampi, J. Thiffault, J.-P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition : a comparison of three methods of analysis for survival data with covariates. *Computational statistics & data analysis*, 4 (3) :185–204, 1986.
- A. Ciampi, C.-H. Chang, S. Hogg, and S. McKinney. Recursive partition : A versatile method for exploratory-data analysis in biostatistics. *Biostatistics : Advances in Statistical Sciences Festschrift in Honor of Professor VM Joshi’s 70th Birthday Volume V*, pages 23–50, 1987.
- B. Claggett, S. Pocock, L. Wei, M. A. Pfeffer, J. J. McMurray, and S. D. Solomon. Comparison of time-to-first event and recurrent-event methods in randomized clinical trials. *Circulation*, 138(6) :570–577, 2018.
- T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. Survival analysis part i : basic concepts and first analyses. *British journal of cancer*, 89(2) :232–238, 2003.

- R. J. Cook and J. F. Lawless. Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine*, 16(8) :911–924, Apr. 1997. ISSN 0277-6715. doi : 10.1002/(sici)1097-0258(19970430)16:8<911::aid-sim544>3.0.co;2-i.
- R. J. Cook and J. F. Lawless. Models and frameworks for analysis of recurrent events. *The Statistical Analysis of Recurrent Events*, pages 27–58, 2007.
- J. Corrigan-Curay, L. Sacks, and J. Woodcock. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *Jama*, 320(9) :867–868, 2018.
- C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20 :273–297, 1995.
- D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–220, 1972a.
- D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202, 1972b. ISSN 2517-6161. doi : 10.1111/j.2517-6161.1972.tb00899.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x>.
- R. Cuocolo, M. Caruso, T. Perillo, L. Ugga, and M. Petretta. Machine learning in oncology : a clinical appraisal. *Cancer letters*, 481 :55–62, 2020.
- V. Dancourt, C. Quantin, M. Abrahamowicz, C. Binquet, A. Alioum, and J. Faivre. Modeling recurrence in colorectal cancer. *Journal of clinical epidemiology*, 57(3) :243–251, 2004.
- R. B. Davis and J. R. Anderson. Exponential survival trees. *Statistics in medicine*, 8(8) :947–961, 1989.
- M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society : Series D (The Statistician)*, 32(1-2) :12–22, 1983.
- T. T. DeLeon, D. R. Almquist, B. R. Kipp, B. T. Langlais, A. Mangold, J. L. Winters, H. E. Kosiorek, R. W. Joseph, R. S. Dronca, M. S. Block, et al. Assessment of clinical outcomes with immune checkpoint inhibitor therapy in melanoma patients with cdkn2a and tp53 pathogenic mutations. *PLoS One*, 15(3) :e0230306, 2020.
- R. C. Deo and B. K. Nallamothu. Learning about machine learning : the promise and pitfalls of big data and the electronic health record, 2016.
- A. Devaux, C. Helmer, R. Genuer, and C. Proust-Lima. Random survival forests with multivariate longitudinal endogenous covariates. *Statistical Methods in Medical Research*, 32(12) :2331–2346, Dec. 2023a. ISSN 0962-2802. doi : 10.1177/09622802231206477. URL <https://doi.org/10.1177/09622802231206477>.
- A. Devaux, C. Proust-Lima, and R. Genuer. Random forests for time-fixed and time-dependent predictors : The dynforest r package. *arXiv preprint arXiv:2302.02670*, 2023b.
- B. Dimanov, U. Bhatt, M. Jamnik, and A. Weller. You shouldn't trust me : Learning models which conceal unfairness from multiple explanation methods. In *ECAI 2020*, pages 2473–2480. IOS Press, 2020.
- D. Dinart, C. Bellera, and V. Rondeau. Sample size estimation for recurrent event data using multifrailty and multilevel survival models. *Journal of Biopharmaceutical Statistics*, pages 1–16, 2024.
- P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10) :78–87, 2012.

-
- F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv :1702.08608*, 2017.
- L. Down and I. Act. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021.
- M. F. Drummond, M. J. Sculpher, K. Claxton, G. L. Stoddart, and G. W. Torrance. *Methods for the economic evaluation of health care programmes*. Oxford university press, 2015.
- L. Duchateau, P. Janssen, I. Kezic, and C. Fortpied. Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society Series C : Applied Statistics*, 52(3) :355–363, 2003.
- E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, et al. New response evaluation criteria in solid tumours : revised recist guideline (version 1.1). *European journal of cancer*, 45(2) :228–247, 2009.
- M. El Amrani, X. Lenne, G. Clement, J.-R. Delpero, D. Theis, F.-R. Pruvot, A. Bruandet, and S. Truant. Specificity of procedure volume and its association with postoperative mortality in digestive cancer surgery : a nationwide study of 225,752 patients. *Annals of Surgery*, 270(5) :775–782, 2019.
- A. Erdmann, J. Beyersmann, and E. Bluhmki. Comparison of nonparametric estimators of the expected number of recurrent events. *Pharmaceutical Statistics*, 2023.
- A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1) :24–29, 2019.
- L. Farah, J. Davaze-Schneider, T. Martin, P. Nguyen, I. Borget, and N. Martelli. Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment ? a systematic review. *Artificial intelligence in medicine*, 140 :102547, 2023a.
- L. Farah, J. M. Murris, I. Borget, A. Guilloux, N. M. Martelli, and S. I. Katsahian. Assessment of performance, interpretability, and explainability in artificial intelligence-based health technologies : what healthcare stakeholders need to know. *Mayo Clinic Proceedings : Digital Health*, 1(2) :120–138, 2023b.
- N. Feeley, S. Cossette, J. Côté, M. Héon, R. Stremler, G. Martorella, and M. Purden. The importance of piloting an rct intervention. *Canadian Journal of Nursing Research Archive*, pages 84–99, 2009.
- V. Fico, G. Altieri, M. Di Grezia, V. Bianchi, M. M. Chiarello, G. Pepe, G. Tropeano, and G. Brisinda. Surgical complications of oncological treatments : A narrative review. *World Journal of Gastrointestinal Surgery*, 15(6) :1056, 2023.
- J. Friedman. The elements of statistical learning : Data mining, inference, and prediction. (*No Title*), 2009.
- J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441, 2008.
- J. H. Friedman. Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, 29(5) :1189–1232, Oct. 2001. ISSN 0090-5364, 2168-8966. doi : 10.1214/aos/1013203451. URL <https://projecteuclid.org>.

- <https://doi.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full>.
- M. Friedman. Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1) :101–113, 1982.
- A. Galaznik, W. Dudley, and N. Coombs. Recurring event survival analyses : A methodological approach to model recurring event data in cancer patients receiving multiple lines of therapy., 2022.
- X. Gao and M. Zheng. Causal inference for recurrent events data with all-or-none compliance. *Communications in Statistics-Theory and Methods*, 45(24) :7306–7325, 2016.
- R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern recognition letters*, 31(14) :2225–2236, 2010.
- T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13) :2173–2184, 2013.
- S. Gerke, B. Babic, T. Evgeniou, and I. G. Cohen. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *npj Digital Medicine*, 3(1) :1–4, Apr. 2020. ISSN 2398-6352. doi : 10.1038/s41746-020-0262-2. URL <https://www.nature.com/articles/s41746-020-0262-2>.
- A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.
- D. Ghosh and D. Y. Lin. Nonparametric Analysis of Recurrent Events and Death. *Biometrics*, 56(2) :554–562, 2000. ISSN 0006-341X. URL <https://www.jstor.org/stable/2677000>.
- D. Ghosh and D. Y. Lin. Marginal Regression Models for Recurrent and Terminal Events. *Statistica Sinica*, 12(3) :663–688, 2002. ISSN 1017-0405. URL <https://www.jstor.org/stable/24306989>.
- L. Girardi, T.-F. Wang, W. Ageno, and M. Carrier. Updates in the incidence, pathogenesis, and management of cancer and venous thromboembolism. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 43(6) :824–831, 2023.
- L. Gordon and R. A. Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69 (10) :1065–1069, 1985.
- E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18) :2529–2545, Sept. 1999. ISSN 0277-6715. doi : 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.
- P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3) :515–526, 1994.
- M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano, et al. A global taxonomy of interpretable ai : unifying the terminology for the technical and social sciences. *Artificial intelligence review*, 56 (4) :3473–3504, 2023.

-
- L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35 :507–520, 2022.
- H. Haider, B. Hoehn, S. Davis, and R. Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85) :1–63, 2020.
- S. Har-Peled, D. Roth, and D. Zimak. Constraint classification : A new approach to multiclass classification. In *Algorithmic Learning Theory : 13th International Conference, ALT 2002 Lübeck, Germany, November 24–26, 2002 Proceedings* 13, pages 365–379. Springer, 2002.
- E. Hariton and J. J. Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG : an international journal of obstetrics and gynaecology*, 125(13) :1716, 2018.
- F. E. Harrell. *reda : Regression Modeling Strategies*, 2023. URL <https://github.com/harrelfe/rms>. R package version 6.8-0.
- F. E. Harrell, Jr, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18) :2543–2546, 1982. ISSN 0098-7484. doi : 10.1001/jama.1982.03320430047030. URL <https://doi.org/10.1001/jama.1982.03320430047030>.
- HAS. Real-world studies for the assessment of medicinal products and medical devices. *Haute Autorité de Santé (HAS) : Saint-Denis, France*, page 50, 2021.
- T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*, volume 2. Springer, 2009.
- P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1) :92–105, 2005.
- A. E. Hoerl and R. W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.
- T. Hothorn and A. Zeileis. Predictive Distribution Modeling Using Transformation Forests. *Journal of Computational and Graphical Statistics*, 30(4) :1181–1196, Oct. 2021. ISSN 1061-8600. doi : 10.1080/10618600.2021.1872581. URL <https://doi.org/10.1080/10618600.2021.1872581>.
- T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1) :77–91, 2004.
- T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3) :355–373, 2006.
- X. Huang and L. Liu. A joint frailty model for survival and gap times between recurrent events. *Biometrics*, 63(2) :389–397, 2007.
- Y. Huang, J. Li, M. Li, and R. R. Aparasu. Application of machine learning in predicting survival outcomes involving real-world data : a scoping review. *BMC medical research methodology*, 23(1) :268, Nov. 2023. ISSN 1471-2288. doi : 10.1186/s12874-023-02078-1.
- H. G. ICHE9. Estimands and sensitivity analysis in clinical trials. 2019.
- B. K. Isariyawongse and M. W. Kattan. Prediction tools in surgical oncology. *Surgical Oncology Clinics of North America*, 21(3) :439–447, 2012.

- H. Ishwaran and M. Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine*, 38(4) : 558–582, 2019.
- H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3) :841–860, Sept. 2008. ISSN 1932-6157, 1941-7330. doi : 10.1214/08-AOAS169. URL <https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Random-survival-forests/10.1214/08-AOAS169.full>.
- H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer. High-Dimensional Variable Selection for Survival Data. *Journal of the American Statistical Association*, 105(489) :205–217, Mar. 2010. ISSN 0162-1459. doi : 10.1198/jasa.2009.tm08622. URL <https://doi.org/10.1198/jasa.2009.tm08622>.
- H. Ishwaran, T. A. Gerds, U. B. Kogalur, R. D. Moore, S. J. Gange, and B. M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4) :757–773, 2014.
- A. Jahn-Eimermacher, K. Ingel, A.-K. Ozga, S. Preussler, and H. Binder. Simulating recurrent event data with hazard functions defined on a total time scale. *BMC medical research methodology*, 15 :1–9, 2015.
- J. Jiménez-Luna, F. Grisoni, and G. Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10) :573–584, 2020.
- A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1) :1, 2023.
- D. G. Johnson and M. Verdicchio. Ai, agency and responsibility : the vw fraud case and beyond. *Ai & Society*, 34 :639–647, 2019.
- J. Kalantari, H. Nelson, and N. Chia. The unreasonable effectiveness of inverse reinforcement learning in advancing cancer research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 437–445, 2020.
- E. L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282) :457–481, June 1958. ISSN 0162-1459. doi : 10.1080/01621459.1958.10501452. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452>.
- G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm : A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.
- P. J. Kelly and L. L.-Y. Lim. Survival analysis for recurrent event data : an application to childhood infectious diseases. *Statistics in medicine*, 19(1) :13–33, 2000.
- F. M. Khan and V. B. Zubek. Support vector regression for censored data (svrc) : a novel tool for survival analysis. In *2008 Eighth IEEE International Conference on Data Mining*, pages 863–868. IEEE, 2008.
- E. S. Kim, D. Bernstein, S. G. Hilsenbeck, C. H. Chung, A. P. Dicker, J. L. Ersek, S. Stein, F. R. Khuri, E. Burgess, K. Hunt, et al. Modernizing eligibility criteria for molecularly driven trials. *Journal of Clinical Oncology*, 33(25) :2815–2820, 2015.

-
- S. Kim, D. E. Schaubel, and K. P. McCullough. A C-index for recurrent event data : Application to hospitalizations among dialysis patients. *Biometrics*, 74(2) :734–743, 2018. ISSN 1541-0420. doi : 10.1111/biom.12761.
- J. P. Klein and P. K. Andersen. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, 61(1) :223–229, 2005.
- K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13 :8–17, 2015.
- M. Krzyziński, M. Spytek, H. Baniecki, and P. Biecek. Survshap (t) : time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262 :110234, 2023.
- A. Kumar, Z. D. Guss, P. T. Courtney, V. Nalawade, P. Sheridan, R. R. Sarkar, M. P. Banegas, B. S. Rose, R. Xu, and J. D. Murphy. Evaluation of the use of cancer registry data for comparative effectiveness research. *JAMA Network Open*, 3(7) :e2011985–e2011985, 2020.
- S. H. Langbein, M. Krzyziński, M. Spytek, H. Baniecki, P. Biecek, and M. N. Wright. Interpretable machine learning for survival analysis, 2024.
- E. LaRosa and D. Danks. Impacts on trust of healthcare ai. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 210–215, 2018.
- J. F. Lawless and C. Nadeau. Some Simple Robust Methods for the Analysis of Recurrent Events. *Technometrics*, 37(2) :158–168, 1995. ISSN 0040-1706. doi : 10.2307/1269617. URL <https://www.jstor.org/stable/1269617>.
- S.-S. Learning. Semi-supervised learning. *CSZ2006. html*, 5 :2, 2006.
- M. LeBlanc and J. Crowley. Relative risk trees for censored survival data. *Biometrics*, pages 411–425, 1992.
- E. W. Lee, L. Wei, D. A. Amato, and S. Leurgans. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival analysis : state of the art*, pages 237–247, 1992.
- Y. Li. Deep reinforcement learning : An overview. *arXiv preprint arXiv :1701.07274*, 2017.
- Q. V. Liao, D. Gruen, and S. Miller. Questioning the AI : Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, pages 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. doi : 10.1145/3313831.3376590. URL <https://doi.org/10.1145/3313831.3376590>.
- D. Y. Lin, L.-J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 62(4) :711–730, 2000.
- A. R. Linero, P. Basak, Y. Li, and D. Sinha. Bayesian survival tree ensembles with submodel shrinkage. *Bayesian Analysis*, 17(3) :997–1020, 2022.
- Z. C. Lipton. The doctor just won’t accept that! *arXiv preprint arXiv :1711.08037*, 2017.
- E. Longato, M. Vettoretti, and B. Di Camillo. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of biomedical informatics*, 108 :103496, 2020.

- S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv :1802.03888*, 2018.
- D. Lupton. Towards critical digital health studies : Reflections on two decades of research in health and the way forward. *Health* ;, 20(1) :49–61, 2016.
- A. Makady, A. de Boer, H. Hillege, O. Klungel, W. Goettsch, et al. What is real-world data ? a review of definitions based on literature and stakeholder interviews. *Value in health*, 20(7) :858–865, 2017a.
- A. Makady, R. Ten Ham, A. de Boer, H. Hillege, O. Klungel, W. Goettsch, et al. Policies for use of real-world data in health technology assessment (hta) : a comparative study of six hta agencies. *Value in Health*, 20(4) :520–532, 2017b.
- N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4) :719–748, 1959.
- N. Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3) :163–170, 1966.
- S. Marco. The need for external validation in machine olfaction : emphasis on health-related applications. *Analytical and bioanalytical chemistry*, 406 :3941–3956, 2014.
- A. F. Markus, J. A. Kors, and P. R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care : A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113(C), Jan. 2021. ISSN 1532-0464. doi : 10.1016/j.jbi.2020.103655. URL <https://doi.org/10.1016/j.jbi.2020.103655>.
- T. Martin, C. Rioufol, B. Favier, N. Martelli, I. Madelaine, C. Chouaid, and I. Borget. Impact of early access reform on oncology innovation in france : Approvals, patients, and costs. *BioDrugs*, 38(3) :465–475, 2024.
- E. Martinelli, C. Cremolini, T. Mazard, J. Vidal, I. Virchow, D. Tougeron, P.-J. Cuyle, B. Chibaudel, S. Kim, I. Ghanem, et al. Real-world first-line treatment of patients with brafv600e-mutant metastatic colorectal cancer : the capstan crc study. *ESMO open*, 7(6) :100603, 2022.
- Y. Mazroui, S. Mathoulin-Pelissier, P. Soubeyran, and V. Rondeau. General joint frailty model for recurrent event data with a dependent terminal event : application to follicular lymphoma data. *Statistics in medicine*, 31(11-12) :1162–1176, 2012.
- C. Mbogning and P. Broët. Bagging survival tree procedure for variable selection and prediction in the presence of nonsusceptible patients. *BMC bioinformatics*, 17 :1–21, 2016.
- L. McInnes, J. Healy, and J. Melville. Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*, 2018.
- R. P. Merkow, M. H. Ju, J. W. Chung, B. L. Hall, M. E. Cohen, M. V. Williams, T. C. Tsai, C. Y. Ko, and K. Y. Bilmoria. Underlying reasons associated with hospital readmission following surgery in the united states. *Jama*, 313(5) :483–495, 2015.
- T. Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, 267 :1–38, Feb. 2019. ISSN 0004-3702. doi : 10.1016/j.artint.2018.07.007. URL <https://www.sciencedirect.com/science/article/pii/S0004370218305988>.

-
- T. M. Mitchell. The need for biases in learning generalizations. 1980.
- T. M. Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.
- B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms : Mapping the debate. *Big Data & Society*, 3(2) :2053951716679679, 2016.
- U. B. Mogensen and T. A. Gerdts. A random forest approach for competing risks based on pseudo-values. *Statistics in medicine*, 32(18) :3102–3114, 2013.
- C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.
- A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse. Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Scientific reports*, 11(1) :6968, 2021.
- G. Moulis, M. Lapeyre-Mestre, A. Palmaro, G. Pugnet, J.-L. Montastruc, and L. Sailler. French health insurance databases : what interest for medical research ? *La Revue de médecine interne*, 36(6) :411–417, 2015.
- U. J. Muehlematter, P. Daniore, and K. N. Vokinger. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20) : a comparative analysis. *The Lancet Digital Health*, 3(3) :e195–e203, Mar. 2021. ISSN 2589-7500. doi : 10.1016/S2589-7500(20)30292-2. URL [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(20\)30292-2/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30292-2/fulltext).
- K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi. Cancer diagnosis using deep learning : a bibliographic review. *Cancers*, 11(9) :1235, 2019.
- J. Murris, A. Charles-Nelson, A. Tadmouri Sellier, A. Lavenu, and S. Katsahian. Towards filling the gaps around recurrent events in high dimensional framework : a systematic literature review and application*. *Biostatistics & Epidemiology*, 7(1) :e2283650, Jan. 2023. ISSN 2470-9360. doi : 10.1080/24709360.2023.2283650. URL <https://doi.org/10.1080/24709360.2023.2283650>.
- J. Murris, O. Bouaziz, M. Jakubczak, S. Katsahian, and A. Lavenu. Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event. 2024.
- J. B. Nasejje, H. Mwambi, K. Dheda, and M. Lesosky. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC medical research methodology*, 17 :1–17, 2017.
- K. Y. Ngiam and W. Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5) :e262–e273, 2019.
- F. Osmani, E. Hajizadeh, and A. A. Rasekhi. Association between multiple recurrent events with multivariate modeling : a retrospective cohort study. *Journal of research in health sciences*, 18(4) :e00433, 2018.
- F. Osmani, E. Hajizadeh, A. Rasekhi, and M. E. Akbari. Prognostic factors associated with locoronal relapses, metastatic relapses, and death among women with breast cancer. population-based cohort study. *The Breast*, 48 :82–88, 2019.
- F. Osmani, E. Hajizadeh, and M. E. Akbari. Prognostic factors associated with curing in patients with breast cancer : A joint frailty model. *International Journal of Preventive Medicine*, 12(1) :9, 2021.

- A.-K. Ozga, M. Kieser, and G. Rauch. A systematic comparison of recurrent event models for application to composite endpoints. *BMC medical research methodology*, 18(1) :2, 2018. ISSN 1471-2288. doi : 10.1186/s12874-017-0462-x.
- R. Padmanabhan, N. Meskin, and W. M. Haddad. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Mathematical biosciences*, 293 :11–20, 2017.
- N. Pandeya, D. M. Purdie, A. Green, and G. Williams. Repeated occurrence of basal cell carcinoma of the skin and multifailure survival analysis : follow-up data from the nambour skin cancer prevention trial. *American journal of epidemiology*, 161(8) :748–754, 2005.
- H. Pang, S. L. George, K. Hui, and T. Tong. Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(5) :1422–1431, 2012.
- C.-W. Park, S. W. Seo, N. Kang, B. Ko, B. W. Choi, C. M. Park, D. K. Chang, H. Kim, H. Kim, H. Lee, et al. Artificial intelligence in health care : Current applications and issues. *Journal of Korean medical science*, 35(42), 2020.
- M. I. Patel, A. M. Lopez, W. Blackstock, K. Reeder-Hayes, E. A. Moushey, J. Phillips, and W. Tap. Cancer disparities and health equity : a policy statement from the american society of clinical oncology. *Journal of Clinical Oncology*, 38(29) :3439–3448, 2020.
- M. J. Pencina and R. B. D'agostino. Overall c as a measure of discrimination in survival analysis : model specific population value and confidence interval estimation. *Statistics in medicine*, 23(13) :2109–2123, 2004.
- K. L. Pickett, K. Suresh, K. R. Campbell, S. Davis, and E. Juarez-Colunga. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC medical research methodology*, 21 :1–14, 2021.
- A. Pickles and R. Crouchley. A comparison of frailty models for multivariate survival data. *Statistics in Medicine*, 14(13) :1447–1461, 1995.
- S. Pölsterl, N. Navab, and A. Katouzian. Fast training of support vector machines for survival analysis. In *Machine Learning and Knowledge Discovery in Databases : European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II* 15, pages 243–259. Springer, 2015.
- M. S. Porta, S. Greenland, M. Hernán, I. dos Santos Silva, and J. M. Last. *A dictionary of epidemiology*. Oxford University Press, USA, 2014.
- R. L. Prentice and J. D. Kalbfleisch. Aspects of the analysis of multivariate failure time data. *SORT. 2003, Vol. 27, Núm. 1 [January-June]*, 2003.
- R. L. Prentice, B. J. Williams, and A. V. Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2) :373–379, 1981. ISSN 0006-3444, 1464-3510. doi : 10.1093/biomet/68.2.373. URL <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/68.2.373>.
- L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost : unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.
- P. Prost, M. Duraes, V. Georgescu, L. Rebel, G. Mercier, and G. Rathat. Impact of ovarian cancer surgery volume on overall and progression-free survival : A population-based

-
- retrospective national french study. *Annals of Surgical Oncology*, 31(5) :3269–3279, 2024.
- S.-a. Qi, N. Kumar, M. Farrokh, W. Sun, L.-H. Kuan, R. Ranganath, R. Henao, and R. Greiner. An effective meaningful way to evaluate survival models. *arXiv preprint arXiv* :2306.01196, 2023.
- S. Quazi. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8) :120, 2022.
- M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" : Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi : 10.1145/2939672.2939778. URL <https://doi.org/10.1145/2939672.2939778>.
- G. Ridgeway. The state of boosting. *Computing science and statistics*, pages 172–181, 1999.
- D. Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3) :819–829, 2011.
- J. K. Rogers, S. J. Pocock, J. J. V. McMurray, C. B. Granger, E. L. Michelson, J. Östergren, M. A. Pfeffer, S. D. Solomon, K. Swedberg, and S. Yusuf. Analysing recurrent hospitalizations in heart failure : a review of statistical methodology, with application to CHARM-Preserved. *European Journal of Heart Failure*, 16(1) :33–40, 2014. ISSN 1879-0844. doi : 10.1002/ejhf.29.
- V. Rondeau, Y. Mazroui, and J. R. Gonzalez. 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*, 47(4) :1–28, 2012. URL <https://www.jstatsoft.org/v47/i04/>.
- P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1) :41–55, 1983.
- C. Sabathe, P. K. Andersen, C. Helmer, T. A. Gerdts, H. Jacqmin-Gadda, and P. Joly. Regression analysis in an illness-death model with interval-censored data : A pseudo-value approach. *Statistical methods in medical research*, 29(3) :752–764, 2020.
- S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3) :660–674, 1991.
- T. H. Scheike and M.-J. Zhang. Analyzing competing risk data using the R timereg package. *Journal of Statistical Software*, 38(2) :1–15, 2011. URL <https://www.jstatsoft.org/v38/i02/>.
- T. H. Scheike, K. K. Holst, and J. B. Hjelmborg. Estimating heritability for cause specific mortality based on twin studies. *Lifetime Data Analysis*, 20(2) :210–233, 2014. doi : 10.1007/s10985-013-9244-x.
- H. Schmidli, J. H. Roger, and M. Akacha. Estimands for recurrent event endpoints in the presence of a terminal event. *Statistics in Biopharmaceutical Research*, 15(2) :238–248, 2023.
- D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1) :239–241, 1982.

- M. R. Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.
- L. S. Shapley et al. A value for n-person games. 1953.
- A. Sheikh, H. S. Sood, and D. W. Bates. Leveraging health information technology to achieve the “triple aim” of healthcare reform. *Journal of the American Medical Informatics Association*, 22(4) :849–856, 2015.
- R. E. Sherman, S. A. Anderson, G. J. Dal Pan, G. W. Gray, T. Gross, N. L. Hunter, L. LaVange, D. Marinac-Dabic, P. W. Marks, M. A. Robb, et al. Real-world evidence—what is it and what can it tell us. *N Engl J Med*, 375(23) :2293–2297, 2016.
- P. K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 655–660. IEEE, 2007.
- N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5) :1, 2011.
- D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP : Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, New York NY USA, Feb. 2020. ACM. ISBN 978-1-4503-7110-0. doi : 10.1145/3375627.3375830. URL <https://dl.acm.org/doi/10.1145/3375627.3375830>.
- L. Smith, A. W. Glaser, D. C. Greenwood, and R. G. Feltbower. Cumulative burden of subsequent neoplasms, cardiovascular and respiratory morbidity in young people surviving cancer. *Cancer Epidemiology*, 66 :101711, 2020.
- A. J. Smola and B. Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.
- C.-L. Su, R. W. Platt, and J.-F. Plante. Causal inference for recurrent event data using pseudo-observations. *Biostatistics*, 23(1) :189–206, 2022.
- K. Suresh, C. Severn, and D. Ghosh. Survival prediction models : an introduction to discrete-time modeling. *BMC medical research methodology*, 22(1) :207, 2022.
- T. Sylvain, M. Luck, J. P. Cohen, H. Cardinal, A. Lodi, and Y. Bengio. Exploring the wasserstein metric for survival analysis. 2021.
- N. Symons, K. Moorthy, A. Almoudaris, A. Bottle, P. Aylin, C. Vincent, and O. Faiz. Mortality in high-risk emergency general surgical admissions. *Journal of British Surgery*, 100(10) :1318–1325, 2013.
- B.-C. Tai, I. R White, V. Gebski, and D. Machin. On the issue of ‘multiple’first failures in competing risks analysis. *Statistics in medicine*, 21(15) :2243–2255, 2002.
- A. Tateishi, H. Horinouchi, K. Takasawa, N. Kouno, T. Mizuno, Y. Okubo, Y. Yoshida, S.-i. Watanabe, M. Miyake, M. Kusumoto, et al. Prediction of recurrence in stage i egfr mutation-positive nsclc : Combination of ct appearance and selected co-occurring gene alterations by machine learning., 2024.
- M. Thenmozhi, V. Jeyaseelan, L. Jeyaseelan, R. Isaac, and R. Vedantam. Survival analysis in longitudinal studies for recurrent events : applications and challenges. *Clinical Epidemiology and Global Health*, 7(2) :253–260, 2019.
- T. Therneau, C. Crowson, and E. Atkinson. Multi-state models and competing risks. *CRAN-R* (<https://cran.r-project.org/web/packages/survival/vignettes/compete.pdf>), 2020.

-
- T. M. Therneau. *A Package for Survival Analysis in R*, 2024. URL <https://CRAN.R-project.org/package=survival>. R package version 3.7-0.
- T. M. Therneau, P. M. Grambsch, T. M. Therneau, and P. M. Grambsch. *The cox model*. Springer, 2000.
- R. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4) :385–395, 1997. ISSN 0277-6715. doi : 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.
- R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 67(1) :91–108, 2005.
- W. Tizi and A. Berrado. Machine learning for survival analysis in cancer research : A comparative study. *Scientific African*, 21 :e01880, 2023.
- E. J. Topol. High-performance medicine : the convergence of human and artificial intelligence. *Nature medicine*, 25(1) :44–56, 2019.
- A. N. B. Torres, R. C. Melchers, L. Van Grieken, J. J. Out-Luiting, H. Mei, C. Agaser, T. B. Kuipers, K. D. Quint, R. Willemze, M. H. Vermeer, et al. Whole-genome profiling of primary cutaneous anaplastic large cell lymphoma. *Haematologica*, 107(7) :1619, 2022.
- J. W. Twisk, N. Smidt, and W. de Vente. Applied analysis of recurrent events : a practical overview. *Journal of Epidemiology & Community Health*, 59(8) :706–710, 2005.
- H. Uno, T. Cai, M. J. Pencina, R. B. D’Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10) :1105–1117, 2011. ISSN 1097-0258. doi : 10.1002/sim.4154. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4154>.
- H. Unsworth, V. Wolfram, B. Dillon, M. Salmon, F. Greaves, X. Liu, T. MacDonald, A. K. Denniston, V. Sounderajah, H. Ashrafian, et al. Building an evidence standards framework for artificial intelligence-enabled digital health technologies. *The Lancet Digital Health*, 4(4) :e216–e217, 2022.
- H. Uramoto and F. Tanaka. Recurrence after surgery in patients with nsclc. *Translational lung cancer research*, 3(4) :242, 2014.
- P. E. Utgoff. *Machine learning of inductive bias*, volume 15. Springer Science & Business Media, 2012.
- V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. Suykens. Support vector methods for survival analysis : a comparison between ranking and regression approaches. *Artificial intelligence in medicine*, 53(2) :107–118, 2011a.
- V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. K. Suykens. Support vector methods for survival analysis : a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2) :107–118, Oct. 2011b. ISSN 1873-2860. doi : 10.1016/j.artmed.2011.06.006.
- M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.
- L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

- H. Van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.
- M. C. van Maaren, B. Rachet, G. S. Sonke, A. Mauguen, V. Rondeau, S. Siesling, and A. Belot. Socioeconomic status and its relation with breast cancer recurrence and survival in young women in the netherlands. *Cancer epidemiology*, 77 :102118, 2022.
- M. van Zutphen, F. J. van Duijnhoven, E. Wesselink, R. W. Schrauwen, E. A. Kouwenhoven, H. K. van Halteren, J. H. de Wilt, R. M. Winkels, D. E. Kok, and H. C. Boshuizen. Identification of lifestyle behaviors associated with recurrence and survival in colorectal cancer patients using random survival forests. *Cancers*, 13(10) :2442, 2021.
- P. J. Verweij and H. C. Van Houwelingen. Penalized likelihood in cox regression. *Statistics in medicine*, 13(23-24) :2427–2436, 1994.
- B. Vinzamuri and C. K. Reddy. Cox regression with correlation based regularization for electronic health records. In *2013 IEEE 13th International Conference on Data Mining*, pages 757–766. IEEE, 2013.
- H. Wang and G. Li. A selective review on random survival forests for high dimensional data. *Quantitative bio-science*, 36(2) :85, 2017a.
- H. Wang and G. Li. A Selective Review on Random Survival Forests for High Dimensional Data. *Quantitative bio-science*, 36(2) :85–96, 2017b. ISSN 2288-1344. doi : 10.22283/qbs.2017.36.2.85. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6364686/>.
- H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng. Machine learning basics. *Deep learning*, pages 98–164, 2016.
- H. Wang, X. Chen, and G. Li. Survival Forests with R-Squared Splitting Rules. *Journal of Computational Biology*, 25(4) :388–395, Apr. 2018. ISSN 1066-5277. doi : 10.1089/cmb.2017.0107. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5905875/>.
- P. Wang, Y. Li, and C. K. Reddy. Machine Learning for Survival Analysis : A Survey. *ACM Computing Surveys*, 51(6) :110 :1–110 :36, 2019. ISSN 0360-0300. doi : 10.1145/3214306. URL <https://doi.org/10.1145/3214306>.
- W. Wang, H. Fu, and J. Yan. *reda : Recurrent Event Data Analysis*, 2022. URL <https://github.com/wenjie2wang/reda>. R package version 0.5.4.
- Y. Wang, Y. Deng, Y. Tan, M. Zhou, Y. Jiang, and B. Liu. A comparison of random survival forest and cox regression for prediction of mortality in patients with hemorrhagic stroke. *BMC Medical Informatics and Decision Making*, 23(1) :215, 2023.
- D. S. Watson, J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes, and L. Floridi. Clinical applications of machine learning algorithms : beyond the black box. *Bmj*, 364, 2019.
- J. Wei, T. Mütze, A. Jahn-Eimermacher, and J. Roger. Properties of two while-alive estimands for recurrent events and their potential estimators. *Statistics in Biopharmaceutical Research*, 15(2) :257–267, 2023.
- L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, 84(408) :1065–1073, 1989. ISSN 0162-1459. doi : 10.2307/2290084. URL <https://www.jstor.org/stable/2290084>.

-
- E. C. Wick, A. D. Shore, K. Hirose, A. M. Ibrahim, S. L. Gearhart, J. Efron, J. P. Weiner, and M. A. Makary. Readmission rates and cost following colorectal surgery. *Diseases of the Colon & Rectum*, 54(12) :1475–1479, 2011.
- S. Wiegreb, P. Kopper, R. Sonabend, B. Bischl, and A. Bender. Deep learning for survival analysis : a review. *Artificial Intelligence Review*, 57(3) :65, 2024.
- G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal. Good enough practices in scientific computing. *PLoS computational biology*, 13(6) :e1005510, 2017.
- S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3) :37–52, 1987.
- D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1) :67–82, 1997.
- J. Wong, M. Murray Horwitz, L. Zhou, and S. Toh. Using machine learning to identify health outcomes from electronic health record data. *Current epidemiology reports*, 5 :331–342, 2018.
- S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, et al. Deep learning in clinical natural language processing : a methodical review. *Journal of the American Medical Informatics Association*, 27(3) :457–470, 2020.
- H. Xu, M. Mohamed, M. Flannery, L. Peppone, E. Ramsdale, K. P. Loh, M. Wells, L. Jamieson, V. G. Vogel, B. A. Hall, et al. An unsupervised machine learning approach to evaluating the association of symptom clusters with adverse outcomes among older adults with advanced cancer : a secondary analysis of a randomized clinical trial. *JAMA network open*, 6(3) :e234198–e234198, 2023.
- A. Yaqoob, R. Musheer Aziz, and N. K. verma. Applications and techniques of machine learning in cancer classification : A systematic review. *Human-Centric Intelligent Systems*, 3(4) :588–615, 2023.
- A. C. Yu and J. Eng. One algorithm may not fit all : how selection bias affects machine learning performance. *Radiographics*, 40(7) :1932–1937, 2020.
- C. Yu, J. Liu, S. Nemati, and G. Yin. Reinforcement learning in healthcare : A survey. *ACM Computing Surveys (CSUR)*, 55(1) :1–36, 2021.
- B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114, 2004.
- H. H. Zhang and W. Lu. Adaptive lasso for cox’s proportional hazards model. *Biometrika*, 94(3) :691–703, 2007.
- H. Zhou, H. Wang, S. Wang, and Y. Zou. Survmetrics : An r package for predictive evaluation metrics in survival analysis. *R J*, 14(4) :252–263, 2023.
- W. Zhu, L. Xie, J. Han, and X. Guo. The application of deep learning in cancer prognosis prediction. *Cancers*, 12(3) :603, 2020.
- H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 67(2) :301–320, 2005.

Annexe A

Chapitre 1

A.1 Compléments à l'étude de cas

A.1.1 Structure des données pour l'ajustement des modèles

Selon le processus de comptage utilisé pour la saisie des données, chaque patient est représenté par plusieurs lignes de données en fonction du nombre d'événements qu'il a subis, le temps étant organisé en intervalles représentant (date d'origine, premier événement], (premier événement, deuxième événement],(k ième événement, date de dernière nouvelle]. La différence majeure pour l'adaptation des modèles consiste en la mise en forme de données adéquates. Pour mettre en évidence les caractéristiques importantes de la structure des données, nous présentons quelques informations sur les celles que nous avons utilisés dans nos applications.

Les 10 premières observations de l'application sur la chirurgie des patients atteints de cancer digestif sont présentées dans le Tableau A.1. id est l'identifiant du patient, $t.start = 0$ pour le premier événement et prend ensuite la valeur du dernier temps d'événement. $t.stop$ est le temps de l'événement lorsque $event = 1$, ou de censure pour $event = 0$, ou de décès pour $death = 1$. $(t.start, t.stop)$ définit l'intervalle à risque. $enum$ est la strate correspondante.

La même structure de données peut être utilisée pour ajuster l'AG, PWP et à fragilité. L'information supplémentaire requise pour l'ajustement des modèles PWP-TT et PWP-GT est l'indicateur d'événement, qui sera utilisé pour la stratification. Les strates correspondront au numéro de l'événement.

TABLE A.1 – Les dix premières observations des données de l'étude de cas

<i>id</i>	<i>t.start</i>	<i>t.stop</i>	<i>time (= t.stop - t.start)</i>	<i>event</i>	<i>death</i>	<i>enum</i>
1	0	10	10	1	0	1
1	10	175	165	1	0	2
1	175	182	7	0	0	3
2	0	55	55	1	0	1
2	55	83	28	0	1	2
3	0	29	29	0	1	1
4	0	1	1	1	0	1
4	1	8	7	1	0	2
4	8	86	78	0	0	3
5	0	142	142	1	0	1

A.1.2 Programmation en R

La librairie `survival` de R permet de modéliser des données d'événements récurrents avec toutes les approches discutées.

Modèle AG. La structure de données requise pour le modèle Andersen-Gill est indiquée dans le Tableau A.1. Le code R permettant d'appliquer le modèle AG au jeu de données *dt* est le suivant

```
1  coxph(Surv(t.start, t.stop, event) ~ group + cluster(id), data =
2    dt)
```

Modèle PWP. La structure de données requise pour le modèle Prentice, Williams et Peterson est indiquée dans le Tableau A.2. Le code R permettant d'appliquer le modèle PWP-TT au jeu de données *dt* est le suivant

```
1  coxph(Surv(t.start, t.stop, event) ~ group + cluster(id) + strata(
2    enum), data = dt)
```

Le code R permettant d'appliquer le modèle PWP-GT au jeu de données *dt* est le suivant

```
1  coxph(Surv(t.start, time, event) ~ group + cluster(id) + strata(
2    enum), data = dt)
```

Modèle à fragilité. La structure de données requise pour le modèle à fragilité est indiquée dans le Tableau A.1. Le code R permettant d'appliquer le modèle à fragilité au jeu de données *dt* est le suivant

```
1  coxph(Surv(t.start, t.stop, event) ~ group + frailty(id), data =
2    dt)
```

TABLE A.2 – Structure des données pour le modèle PWP avec trois strates

<i>id</i>	<i>t.start</i>	<i>t.stop</i>	<i>time (= t.stop - t.start)</i>	<i>event</i>	<i>death</i>	<i>enum</i>
1	0	10	10	1	0	1
1	10	175	165	1	0	2
1	175	182	7	0	0	3
2	0	55	55	1	0	1
2	55	83	28	0	1	2
2	55	83	28	0	1	3
3	0	29	29	0	1	1
3	0	29	29	0	1	2
3	0	29	29	0	1	3
4	0	1	1	1	0	1
4	1	8	7	1	0	2
4	8	86	78	0	0	3
5	0	142	142	1	0	1
5	0	142	142	1	0	2
5	0	142	142	1	0	3

Modèle WLW. La structure de données requise pour le modèle WLW est indiquée dans le Tableau A.3. Le code R permettant d’appliquer le modèle WLW au jeu de données *dt* est le suivant

```

1  coxph(Surv(t.start, t.stop, event) ~ group + cluster(id) + strata
2  (enum), data = dt)

```

TABLE A.3 – Structure des données opur le modèle WLW avec trois strates

<i>id</i>	<i>t.start</i>	<i>t.stop</i>	<i>event</i>	<i>death</i>	<i>enum</i>
1	0	10	1	0	1
1	0	175	1	0	2
1	0	182	0	0	3
2	0	55	1	0	1
2	0	83	0	1	2
2	0	83	0	1	3
3	0	29	0	1	1
3	0	29	0	1	2
3	0	29	0	1	3
4	0	1	1	0	1
4	0	8	1	0	2
4	0	86	0	0	3
5	0	142	1	0	1
5	0	142	1	0	2
5	0	142	1	0	3

Annexe B

Chapitre 3

B.1 Compléments à l'illustration de RecForest

B.1.1 Codes utilisés

TABLE B.1 – Codes de diagnostic, CCAM et CIM-10

Diagnosis	ICD-10/medical procedure codes
Bile Duct Surgery	HMFA010, HMFA009, HMLA001, HMLC001, HMFA001, HMFA002, HMFA005, HMFA006, HMFC003, HMFC005, HMFA003, HMFA008, HMFA004, HMFC002, HMCA005, HMCA006, HMCA007, HMCA008, HMFA010, HMFA009, HMLA001, HMLC001, HMFA001, HMFA002, HMFA005, HMFA006, HMFC003, HMFC005, HMFA003, HMFA008, HMFA004, HMFC002, HMCA005, HMCA006, HMCA007, HMCA008, HMFA010, HMFA009, HMLA001, HMLC001, HMFA001, HMFA002, HMFA005, HMFA006, HMFC003, HMFC005, HMFA003, HMFA008, HMFA004, HMFC002, HMCA005, HMCA006, HMCA007, HMCA008
Colectomy	HHFA002, HHFA004, HHFA005, HHFA006, HHFA008, HHFA009, HHFA010, HHFA014, HHFA017, HHFA018, HHFA021, HHFA022, HHFA023, HHFA024, HHFA026, HHFA028, HHFA029, HHFA030, HHFA031, HHFC040, HHFC296, HHFA002, HHFA004, HHFA005, HHFA006, HHFA008, HHFA009, HHFA010, HHFA014, HHFA017, HHFA018, HHFA021, HHFA022, HHFA023, HHFA024, HHFA026, HHFA028, HHFA029, HHFA030, HHFA031, HHFC040, HHFC296, HHFA002, HHFA004, HHFA005, HHFA006, HHFA008, HHFA009, HHFA010, HHFA014, HHFA017, HHFA018, HHFA021, HHFA022, HHFA023, HHFA024, HHFA026, HHFA028, HHFA029, HHFA030, HHFA031, HHFC040, HHFC296
Esophagectomy	HEFA, HEFA, HEFA
Complication Transfusion	FELF003, FELF004, FELF001, FELF003, FELF004, FELF006, FELF008, FELF011, FELF003, FELF004, FELF001, FELF003, FELF004, FELF006, FELF008, FELF011, FELF003, FELF004, FELF001, FELF003, FELF004, FELF006, FELF008, FELF011

Complication Cardiac	B376, I200, I200+0, I201, I208, I209, I2100, I2100, I2108, I2110, I21100, I2120, I21200, I2128, I2130, I21300, I2140, I21400, I2190, I21900, I2198, I22000, I2288, I2290, I236, I238, I240, I248, I249, I313, I330, I339, I38, I400, I409, I410, I460, I461, I469, I470, I471, I472, I479, I480, I481, I483, I484, I490, I500, I501, I509, I513, DDAF001, DDAF003, DDAF004, DDAF006, DDAF007, DDAF008, DDAF010, DDQH006, DDQH009, DDQH01, DCJB001, DCJA001, DERD001, DERF001, DERF002, DERF003, DERF004, DERP005
Complication Respiratory Assistance	GLLD008, GLLD002, GLLD003, GLLD004, GLLD006, GLLD008, GLLD002, GLLD003, GLLD004, GLLD006

Peritoneal Surgery	HPBA001, HPFA003, HPFA004, HPBA001, HPFA003, HPFA004, HPFA001, HPFA003, HPFA004, HPFA001, HPFA003, HPFA004, HPFA001, HPFA003, HPFA004
Rectal Resection	HJFA, HJFA, HJFA
Small Bowel Resection	HGFA001, HGFA007, HGFC014, HGFC016, HGFC021, HGFA001, HGFA001, HGFA007, HGFA007, HGFC014, HGFC016, HGFC021
Surgical Shunt	EHCA003, EHCA006, EHCA009, EHCA007, EHCA004, EHCA002, EHCA005, EHCA010, EHCA001, HEPA005, HEPA004, HEPA007, EHCA003, EHCA006, EHCA009, EHCA007, EHCA004, EHCA002, EHCA005, EHCA010, EHCA001, HEPA005, HEPA007, EHCA003, EHCA006, EHCA009, EHCA007, EHCA004, EHCA002, EHCA005, EHCA010, EHCA001, HEPA005, HEPA004, HEPA007
Radiologic Drainage Abdomen	HLHJ004, ZZJH005, ZZJH003, ZZJH006, ZZJH004, ZZJH005, ZZJH006, ZZJJ007, ZZJJ008, ZZJJ010, ZZJJ013, ZZJJ003, ZZJJH008, ZZJJH007, ZZJH002, ZZJH001, HPJB001, HPHB003, HMCH001, HLHH002, ZZJH004, ZZJH005, ZZJH003, ZZJH006, ZZJJ004, ZZJJ005, ZZJJH006, ZZJJ007, ZZJJ008, ZZJJ010, ZZJJ013, ZZJJ003, ZZJJH008, ZZJJH007, ZZJH002, ZZJH001, HPJB001, HPHB003, HMCH001, HLHH002, HLHJ004, ZZJH005, ZZJH003, ZZJH006, ZZJJ004, ZZJJ005, ZZJJH006, ZZJJ007, ZZJJ008, ZZJJ010, ZZJJ013, ZZJJ003, ZZJJH008, ZZJJH007, ZZJH002, ZZJH001, HPJB001, HPHB003, HMCH001, HLHH002, ZZHH001, ZZHH004, ZZHH006, ZZHH008, ZZHH009, ZZHH010, ZZHH011, ZZHH012, ZZHB002, ZZHH001, ZZHH009, ZZHH010, ZZHH013, ZZHH017, ZZHH019, ZZHH013, ZZHH017, ZZHH022, ZZHH001, ZZHH004, ZZHH006, ZZHH008, ZZHH009, ZZHH010, ZZHH011, ZZHH012, ZZHH010, ZZHH011, ZZHH012, ZZHB002, ZZHH001, ZZHH017, ZZHH019, ZZHH009, ZZHH010, ZZHH013, ZZHH017, ZZHH019, ZZHH022

Radiologic Arteriography Intervention	EDQH008, EDQH003, EDQH001, EDQH005, EDQH006, EDQH007, DGQH001, DDQH012, DDQH013, DDQH014, DDQH015, DGQH002, DGQH003, DGQH004, EDPF004, EDNF003, EDJF002, EDSF016, EDSF015, EDSF014, EDSF009, EDSF012, EDSF008, EDSF001, EDSF005, EDSF006, EDPF003, EDLF005, EDLF007, EDLF004, EDLF006, EDLF008, EDQH008, EDQH003, EDQH001, EDQH005, EDQH006, EDQH007, DGQH001, DDQH012, DDQH013, DDQH014, DDQH015, DGQH002, DGQH003, DGQH004, EDPF004, EDNF003, EDJF002, EDSF016, EDSF015, EDSF014, EDSF009, EDSF012, EDSF008, EDSF001, EDSF005, EDSF006, EDPF003, EDLF005, EDLF007, EDLF004, EDLF006, EDLF008, JVJF002, JVJF003, JVJF005, JVJF006, JVJF007, JVJF005, JVJF002
Renal Dialysis	Thoracic Drainage

Reintervention	ZCQA001, ZCQC001, ZCQC002, QZJA011, ZCJA004, ZCJA002, ZCJA005, ZCJC001, GGJA002, GGJA001, GGJC001, GGJC002, HGCA008, HGCA008, HGLA001, HGCA004, HGCA001, HGCA005, HGCC003, HFCA004, HGFA003, HGFA004, HGCA007, HGCC003, HFCC001, HFCA003, HGFA005, DGCA001, DGS005, DHCA001, EDCA001, EDCA002, EDC0015, EDFA002, EDFA005, EDFA009, EDFA010, EDKA003, EDPA002, EDSA003, EDPF002, EDSA001, EHCA008, EHFA001, ELFA001, FFFA001, FFFA002, FFFC001, FFJA001, FFSA001, HECA004, HECA002, HFCA002, HFFA002, EDSA001, LMSA002, HPPC003, HPPA002, HNJC001, GGJA003, GGJC002, GGJC001, HHCC001, HHMA002, HJCA001, HMCA009, LLMA008, LMMA004, LMMA010, LMMC015, LMSA002, ZCJD001, HNJA001, HNFA012, HNCA006, HLJC001, HLJA001, HHCC007, HHCA003, HHCA002, HHCA001, HGPA005, HGPA004, HGPA003, HGPA002, HGPA001, HGFA013, HGFA005, HGFA007, HGFC021, HGFA003, HGCC026, HGCC003, HGCA008, HGCA002, HFP001, ZCQA001, ZCQC001,
----------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

ZCQC002, QZJA011, ZCJA004, ZCJA002, ZCJA005, ZCJC001,
GGJA002, GGJA001, GGJC001, GGJC002, HGCC002, HGCC003, HGCA008,
HGLA001, HGCA004, HGCA001, HGCA005, HGCC003, HFCA004,
HGFA003, HGFA004, HGCA007, HGCC003, HFCC001, HFCA003,
HGFA005, DGCA001, DGS005, DHCA001, EDCA001, EDCA002,
EDCC015, EDFA002, EDFA005, EDFA009, EDFA010, EDKA003,
EDPA002, EDSA003, EDPF002, EDSA001, EHCA008, EHFA001,
ELFA001, FFFA001, FFFA002, FFFC001, FFJA001, FFSA001, HECA004,
HFCA002, HFFA002, HFFA003, HFFA005, LMSA002, HPPC003,
HPPA002, HNJ001, GGJA003, GGJC002, GGJC001, HHCC001,
HHMA002, HJCA001, HMCA009, LLMA008, LMMA004, LMMA010,
LMMC015, LMSA002, ZCJD001, HNJA001, HNFA012, HNCA006,
HLJC001, HLJA001, HHCC007, HHCA003, HHCA002, HHCA001,
HGPA005, HGPA004, HGPA003, HGPA002, HGPA001, HGFA013,
HGFA005, HGFA007, HGFC021, HGFA003, HGCC026, HGCC003,
HGCA008, HGCA002, HFPA001, ZCQA001, ZCQC001, ZCQC002,
QZJA011, ZCJA004, ZCJA002, ZCJA005, ZCJC001, GGJA002, GGJA001,
GGJC001, GGJC002, HGCC026, HGCA008, HGLA001, HGCA004,
HGCA001, HGCA005, HGCC003, HFCA004, HGFA003, HGFA004,
HGCA007, HGCC003, HFCC001, HFCA003, HGFA005, DGCA001,
DGSA005, DHCA001, EDCA001, EDCA002, EDCC015, EDFA002,
EDFA005, EDFA009, EDFA010, EDKA003, EDPA002, EDSA003,
EDPF002, EDSA001, EHCA008, EHFA001, ELFA001, FFFA001,
FFFA002, FFFC001, FFJA001, FFSA001, HECA004, HFCA002, HFFA002,
HFFA003, HFFA005, LMSA002, HPPC003, HPPA002, HNJ001,
GGJA003, GGJC002, GGJC001, HHCC001, HHMA002, HJCA001,
HMCA009, LLMA008, LMMA004, LMMA010, LMHC015, LMSA002,
ZCJD001, HNJA001, HNFA012, HNCA006, HLJC001, HLJA001,
HHCC007, HHCA003, HHCA002, HHCA001, HGPA005, HGPA004,
HGPA003, HGPA002, HGPA001, HGFA013, HGFA005, HGFA007,
HGFC021, HGFA003, HGCC026, HGCC003, HGCA008, HGCA002,
HFPA001

Vascular Resection	ELFA001, EHFA001, EHCA008, EHCA010, DHFA004, DHFA005, DHFA007, DHFA001, DHFA006, ELFA001, EHFA001, EHFA001, EHCA008, EHCA010, DHFA004, DHFA005, DHFA007, DHFA001, DHFA006, EHFA001, EHFA001, EHCA008, EHCA010, DHFA004, DHFA005, DHFA007, DHFA001, DHFA006
Radiologic Portal Vein Embolisation	EHSF001, EHSF001, EHSF001, EHSF001
Aids	B20, B21, B22, B24, B20, B21, B22, B24
Alcohol Use Disorders	E244, E511, F101, F102, F103, F104, F105, F106, F107, F108, F109, F1020, F1021, F1022, F1023, G312, G621, G721, I426, K292, K852, K860, O354, Z502, Z714, Z721, E244, E511, F101, F102, F103, F104, F105, F106, F107, F108, F109, F1020, F1021, F1022, F1023, G312, G621, G721, I426, K292, K852, K860, O354, Z502, Z714, Z721
Alcoholic Liver Disease	K70, K70
Anemia	D500, D63, D62, D500, D63, D62
Anticoagulant Treatment	Z921, Y442, Z921, Y442
Ascitis	R18, HPHB003, HPJB001, R18, HPHB003, HPJB001
Autoimmune Hepatitis	K754, K754
Cardiac Valve	Z95, Z95
Cardiac Arrhythmia	I48, I48
Cerebral Infarction	I63, I63
Cerebrovascular Disease	G45, G46, H340, I6, G45, G46, H340, I6
Chemotherapy for Cancer	Z511, Z512, Z926, Z511, Z512, Z926

Chronic Kidney Disease Advanced	I120, I131, N00, N01, N03, N05, N183, N184, N185, N19, N250, Z490, Z491, Z492, Z940, Z992, HGPC005, HPGAA001, HPJP001, HPKA002, HPKB001, HPKC014, HPLA005, HPLB004, HPLC035, HPPA004, HPPP002, JVRP007, JVRP008, JAEA003, JVJB001, JVRP004, JVJF007, JVQF001, JVQF007, JVQP002, JVQP009, YYYY007, I120, I131, N00, N01, N03, N05, N183, N184, N185, N19, N250, Z490, Z491, Z492, Z940, Z992, HGPC005, HPGAA001, HPJP001, HPKA002, HPKB001, HPKC014, HPLA005, HPLB004, HPLC035, HPPA004, HPPP002, JVRP007, JVRP008, JAEA003, JVJB001, JVRP004, JVJF007, JVQF001, JVQF007, JVQP002, JVQP009, YYYY007
Chronic Kidney Disease	E102, E112, E122, E132, E142, I151, JAHB001, JAHC001, JAHA001, JAHH006, JAHH007, N02, N04, N06, N07, N08, N18, N25, N083, E102, E112, E122, E132, E142, I151, JAHB001, JAHH002, JAHC001, JAHA001, JAHH007, N02, N04, N06, N07, N08, N18, N19, N25, N083
Chronic Obstructive Pulmonary Disease	I278, I279, J40, J41, J42, J43, J44, J45, J46, J47, J60, J61, J62, J63, J64, J65, J66, J67, J684, J701, J703, I278, I279, J40, J41, J42, J43, J44, J45, J46, J47, J60, J61, J62, J63, J64, J65, J66, J67, J684, J701, J703
Cirrhosis	I859, I864, I982, I9829, K703, K717, K743, K744, K745, K746, K766, I859, I864, I982, I9829, K703, K717, K743, K744, K745, K746, K766
Congestive Heart Failure	I099, I110, I130, I132, I255, I420, I425, I426, I427, I428, I429, I43, I50, P290, I099, I110, I130, I132, I255, I420, I425, I426, I427, I428, I429, I43, I50, P290
Chronic Cardiopathy	I25, I25
Connective Tissue Disorder	M05, M06, M315, M32, M33, M34, M351, M353, M360, M05, M06, M315, M32, M33, M34, M351, M353, M360
Crohn Disease	K50, K50
Dementia	F00, F01, F02, F03, F051, G30, G311, F00, F01, F02, F03, F051, G30, G311

Diabetes Mellitus Complicated	E102, E103, E104, E105, E107, E112, E113, E114, E115, E117, E122, E123, E124, E125, E127, E132, E133, E134, E135, E137, E142, E143, E144, E145, E147, H360, N083, H280, G632, E102, E103, E104, E105, E107, E112, E113, E114, E115, E117, E122, E123, E124, E125, E127, E132, E133, E134, E135, E137, E142, E143, E144, E145, E147, H360, N083, H280, G632
Diabetes Mellitus Uncomplicated	E100, E101, E106, E108, E109, E110, E111, E116, E118, E119, E120, E121, E126, E128, E129, E130, E131, E136, E138, E139, E140, E141, E146, E148, E149, E100, E101, E106, E108, E109, E110, E111, E116, E118, E119, E120, E121, E126, E128, E129, E130, E131, E136, E138, E139, E140, E141, E146, E148, E149
Complication Thromboembolic	I260, I269, I740, I741, I742, I743, I744, I745, I748, I800, I801, I802, I803, I808, I809, I81, I821, I822, I823, I828, I829, K751, I260, I269, I740, I741, I742, I743, I744, I745, I748, I800, I801, I802, I803, I808, I809, I81, I821, I822, I823, I828, I829, K751
Gastro Esophageal Varices not Bleeding	I859, I864, I982, I9829, I859, I864, I982, I9829
Hemiplegia	G041, G114, G801, G802, G81, G82, G830, G831, G832, G833, G834, G839, G041, G114, G801, G802, G81, G82, G830, G831, G832, G833, G834, G839
Hepatic Encephalopathy	K704, K711, K74, K704, K711, K74
Hepatitis Viral	B181, B162, B169, B182, B160, B161, B180, B170, B181, B162, B169, B182, B160, B161, B180, B170
Hepatopulmonary Syndrome	I280, I280
Hepatorenal Syndrome	K767, K767
Hydrothorax	J948, J948
Hyperlipidemia	E780, E781, E782, E784, E785, E780, E781, E782, E784, E785
Hypertension	I10, I11, I12, I13, R030, I10, I11, I12, I13, R030
Hypertensive Heart Failure with or without Renal Failure	I130, I132, I130, I132

Hypertriglyceridemia	E871, E871
Jaundice	R17, R17
Jaundice Obstructive	K831, K831
Leukemia	C91, C92, C93, C94, C95, C96, C91, C92, C93, C94, C95, C96
Lithiasis Gallbladder	K802
Liver Disease Mild	K713, K714, K715, K73, K740, K741, K742, K743, K760, K764, K765
Liver Disease Moderate	K703, I850, I859, I864, I982, J948, K703, K704, K711, K717, K721, K729,
Severe	K744, K745, K746, K65, K66, K767, R17, R18, EHBD001, EHNE002, HESE001, HESE002, HPHB003, HPJB001, EHCA003, EHCA006, EHCA009, EHCA007, EHCA004, EHCA002, EHCA005, EHCA010, EHCA001, HEP A005, HEP A004, HEP A007
Liver Transplantation	Z944, HLEA001, HGEA002, HLEA002, HGEA004
Lymphoma	C81, C82, C83, C84, C85, C86, C87, C89, C90
Malnutrition	E40, E41, E42, E43, E44, R630, R634
Metastasis Liver	C787
Myelodysplastic Syndromes	D46
Myocardial Infarction	I21, I22, I252, I255
Non Melanoma Neoplasm of Skin	C44
Non Viral Non Metabolic Chronic Liver Disease	E831, E84, FEJF003, FEJF006, FEJF008, E830, NI63, I820, K743, K754, Q44, T864, Z944
Obesity	E66
Paraplegia Hemiplegia	G041, G114, G801, G802, G82, G830, G831, G832, G833, G834, G839
Peptic Ulcer Disease	K25, K26, K27, K28
Peripheral Vascular Disease	I70, I71, I731, I738, I739, I771, I790, I792, K551, K558, K559, R02, Z958, Z959
Phlebitis and Thrombophlebitis	I80
Portal Hypertension	K766

Portal Vein Thrombosis	I81
Primary Liver Cancer	C220, C221, C229
Pulmonary Embolism	I26
Renal Transplantation	T861, Z940, JAEA003
Smoking	F17, Z716, Z720, T652
Solid Tumor Localised	C0, C10, C11, C12, C13, C14, C15, C16, C17, C20, C21, C22, C23, C24, C25, C26, C3, C40, C41, C43, C45, C46, C47, C48, C49, C5, C6, C70, C71, C72, C73, C74, C75, C76
Solid Tumor Metastatic	C77, C78, C79, C80
Sleep Apnea Syndrome	G473
Tips	EHCF002, EHAFO04, EHPPF001, EHNF001
Transplant Recipient without Liver	Z940, Z941, Z942, Z943, Z9481, Z9482, T861, T862, T863, T8680, T8681, T8682, Z94801, Z94802, Z94803, Z94804, Z94809, T860, JAEA003
Ulcerative Colitis	K51

B.1.2 Description des patients

TABLE B.2 – Caractéristiques à l'inclusion des patients ayant bénéficié d'une chirurgie digestive majeure entre 2020 et 2022 en France (Source : PMSI)

Characteristic	Whole Sample		Colorectal Surgery		Hepatobiliary Surgery		Pancreatic Surgery		Small Bowel Surgery		Esogastric Surgery	
	N = 255,732	N = 131,260	N = 21,663	N = 10,078	N = 13,280	N = 70,298 ¹						
Demographics												
Sexe												
1	110,630 (43%)	68,042 (52%)	11,497 (53%)	5,432 (54%)	6,332 (48%)	19,204 (27%)						
2	145,102 (57%)	63,218 (48%)	10,166 (47%)	4,646 (46%)	6,948 (52%)	51,094 (73%)						
Age	62 (46, 73)	68 (57, 77)	66 (56, 74)	68 (60, 74)	65 (50, 75)	42 (32, 54)						
Clinical characteristics												
Aids	454 (0.2%)	225 (0.2%)	59 (0.3%)	28 (0.3%)	29 (0.2%)	111 (0.2%)						
Alcohol use disorders	7,264 (2.8%)	4,243 (3.2%)	1,202 (5.5%)	575 (5.7%)	445 (3.4%)	941 (1.3%)						
Alcoholic liver disease	2,386 (0.9%)	1,015 (0.8%)	1,032 (4.8%)	77 (0.8%)	104 (0.8%)	171 (0.2%)						
Anemia	34,122 (13%)	22,115 (17%)	2,989 (14%)	2,392 (24%)	2,386 (18%)	4,112 (5.8%)						
Anticoagulant treatment	12,555 (4.9%)	8,590 (6.5%)	1,160 (5.4%)	610 (6.1%)	831 (6.3%)	1,163 (1.7%)						
Ascitis	10,813 (4.2%)	5,348 (4.1%)	1,538 (7.1%)	933 (9.3%)	734 (5.5%)	642 (0.9%)						
Autoimmune hepatitis	82 (<0.1%)	34 (<0.1%)	32 (0.1%)	4 (<0.1%)	5 (<0.1%)	8 (<0.1%)						
Cardiac valve	14,220 (5.6%)	9,560 (7.3%)	1,516 (7.0%)	780 (7.7%)	931 (7.0%)	1,478 (2.1%)						
Cardiac arrhythmia	19,289 (7.5%)	13,404 (10%)	1,784 (8.2%)	859 (8.5%)	1,260 (9.5%)	2,031 (2.9%)						
Cerebral infarction	1,508 (0.6%)	1,075 (0.8%)	134 (0.6%)	53 (0.5%)	96 (0.7%)	112 (0.2%)						
Cerebrovascular disease	5,853 (2.3%)	4,071 (3.1%)	549 (2.5%)	238 (2.4%)	369 (2.8%)	558 (0.8%)						
Chemotherapy for cancer	42,881 (17%)	22,366 (17%)	6,338 (29%)	2,751 (27%)	2,416 (18%)	7,573 (11%)						
Chronic kidney disease advanced	7,320 (2.9%)	5,063 (3.9%)	668 (3.1%)	308 (3.1%)	618 (4.7%)	638 (0.9%)						
Chronic kidney disease	10,531 (4.1%)	7,156 (5.5%)	1,018 (4.7%)	456 (4.5%)	822 (6.2%)	1,030 (1.5%)						
Chronic obstructive pulmonary disease	17,041 (6.7%)	9,592 (7.3%)	1,501 (6.9%)	845 (8.4%)	928 (7.0%)	4,073 (5.8%)						
Cirrhosis	4,355 (1.7%)	1,604 (1.2%)	2,022 (9.3%)	244 (2.4%)	186 (1.4%)	368 (0.5%)						
Congestive heart failure	13,505 (5.3%)	9,383 (7.1%)	1,247 (5.8%)	659 (6.5%)	892 (6.7%)	1,376 (2.0%)						
Chronic cardiopathy	12,969 (5.1%)	8,600 (6.6%)	1,369 (6.3%)	697 (6.9%)	802 (6.0%)	1,534 (2.2%)						
Connective tissue disorder	1,979 (0.8%)	1,195 (0.9%)	152 (0.7%)	101 (1.0%)	127 (1.0%)	331 (0.5%)						

		A_{NN}^S	A_{NE}^S	A_{EX}^S	A_{ES}^S
Crohn disease	5,240 (2.0%)	4,469 (3.4%)	78 (0.4%)	27 (0.3%)	839 (6.3%)
Dementia	2,435 (1.0%)	1,981 (1.5%)	137 (0.6%)	47 (0.5%)	147 (1.1%)
Diabetes mellitus complicated	4,890 (1.9%)	2,822 (2.1%)	586 (2.7%)	362 (3.6%)	283 (2.1%)
Diabetes mellitus uncomplicated	33,069 (13%)	17,233 (13%)	3,840 (18%)	2,934 (29%)	1,668 (13%)
Complication thromboembolic	13,997 (5.5%)	8,350 (6.4%)	1,670 (7.7%)	1,281 (13%)	991 (7.5%)
Gastro esophageal varices not bleeding	823 (0.3%)	348 (0.3%)	325 (1.5%)	30 (0.3%)	53 (0.4%)
Hemiplegia	3,385 (1.3%)	2,382 (1.8%)	283 (1.3%)	95 (0.9%)	235 (1.8%)
Hepatic encephalopathy	3,003 (1.2%)	840 (0.6%)	1,710 (7.9%)	111 (1.1%)	89 (0.7%)
Hepatitis viral	926 (0.4%)	246 (0.2%)	484 (2.2%)	47 (0.5%)	32 (0.2%)
Hepatopulmonary syndrome	8 (<0.1%)	5 (<0.1%)	2 (<0.1%)	1 (<0.1%)	0 (0%)
Hepatorenal syndrome	89 (<0.1%)	36 (<0.1%)	42 (0.2%)	3 (<0.1%)	8 (<0.1%)
Hydrothorax	143 (<0.1%)	60 (<0.1%)	24 (0.1%)	11 (0.1%)	5 (<0.1%)
Hyperlipidemia	20,841 (8.1%)	11,971 (9.1%)	2,053 (9.5%)	1,223 (12%)	1,115 (8.4%)
Hypertension	78,522 (31%)	46,163 (35%)	7,791 (36%)	4,125 (41%)	4,326 (33%)
Hypertensive heart failure with or without renal failure	92 (<0.1%)	75 (<0.1%)	6 (<0.1%)	1 (<0.1%)	6 (<0.1%)
Hypertriglyceridemia	19,179 (7.5%)	12,317 (9.4%)	2,110 (9.7%)	1,528 (15%)	1,339 (10%)
Jaundice	3,398 (1.3%)	338 (0.3%)	1,089 (5.0%)	1,939 (19%)	62 (0.5%)
Jaundice obstructive	6,184 (2.4%)	904 (0.7%)	2,066 (9.5%)	3,041 (30%)	152 (1.1%)
Leukemia	752 (0.3%)	507 (0.4%)	87 (0.4%)	35 (0.3%)	62 (0.5%)
Lithiasis gallbladder	3,857 (1.5%)	1,240 (0.9%)	1,176 (5.4%)	369 (3.7%)	181 (1.4%)
Liver disease mild	8,253 (3.2%)	834 (0.6%)	1,458 (6.7%)	200 (2.0%)	130 (1.0%)
Liver disease moderate to severe	43,940 (17%)	25,819 (20%)	5,476 (25%)	4,130 (41%)	3,714 (28%)
Liver transplantation	435 (0.2%)	105 (<0.1%)	275 (1.3%)	11 (0.1%)	21 (0.2%)
Lymphoma	1,951 (0.8%)	1,166 (0.9%)	174 (0.8%)	97 (1.0%)	323 (2.4%)
Malnutrition	46,955 (18%)	29,770 (23%)	4,265 (20%)	4,955 (49%)	3,148 (24%)
Metastasis liver	14,965 (5.9%)	7,727 (5.9%)	7,788 (36%)	424 (4.2%)	549 (4.1%)
Myelodysplastic syndromes	386 (0.2%)	294 (0.2%)	28 (0.1%)	5 (<0.1%)	39 (0.3%)
Myocardial infarction	9,139 (3.6%)	6,125 (4.7%)	958 (4.4%)	597 (4.5%)	1,036 (1.5%)
Non melanoma neoplasm of skin	1,192 (0.5%)	836 (0.6%)	113 (0.5%)	45 (0.4%)	78 (0.6%)
Non viral non metabolic chronic liver disease	1,286 (0.5%)	430 (0.3%)	592 (2.7%)	69 (0.7%)	145 (0.2%)

Obesity	81,495 (32%)	15,298 (12%)	3,063 (14%)	2,324 (18%)	58,128 (83%)
Paraplegia hemiplegia	3,385 (1.3%)	2,382 (1.8%)	283 (1.3%)	95 (0.9%)	327 (0.5%)
Peptic ulcer disease	5,021 (2.0%)	1,750 (1.3%)	383 (1.8%)	361 (3.6%)	2,390 (3.4%)
Peripheral vascular disease	11,031 (4.3%)	7,518 (5.7%)	1,059 (4.9%)	633 (6.3%)	1,034 (1.5%)
Phlebitis and thrombophlebitis	6,397 (2.5%)	4,203 (3.2%)	515 (2.4%)	362 (3.6%)	454 (3.4%)
Portal hypertension	1,239 (0.5%)	484 (0.4%)	464 (2.1%)	146 (1.4%)	72 (0.5%)
Portal vein thrombosis	1,081 (0.4%)	304 (0.2%)	461 (2.1%)	240 (2.4%)	79 (0.6%)
Primary liver cancer	5,367 (2.1%)	352 (0.3%)	4,550 (21%)	461 (4.6%)	61 (0.5%)
Pulmonary embolism	4,270 (1.7%)	2,504 (1.9%)	474 (2.2%)	318 (3.2%)	284 (2.1%)
Renal transplantation	587 (0.2%)	350 (0.3%)	61 (0.3%)	51 (0.5%)	69 (0.5%)
Smoking	16,060 (6.3%)	8,583 (6.5%)	1,654 (7.6%)	1,019 (10%)	950 (7.2%)
Solid tumor localised	68,657 (27%)	30,286 (23%)	10,167 (47%)	8,944 (89%)	3,753 (28%)
Solid tumor metastatic	46,942 (18%)	26,612 (20%)	9,006 (42%)	3,216 (32%)	2,470 (19%)
Sleep apnea syndrome	28,990 (11%)	6,023 (4.6%)	1,181 (5.5%)	617 (6.1%)	844 (6.4%)
Tips	100 (<0.1%)	49 (<0.1%)	19 (<0.1%)	7 (<0.1%)	8 (<0.1%)
Transplant recipient without liver	763 (0.3%)	459 (0.3%)	77 (0.4%)	59 (0.6%)	90 (0.7%)
Ulcerative colitis	2,184 (0.9%)	1,863 (1.4%)	77 (0.4%)	25 (0.2%)	124 (0.9%)
Indices for comorbidities					
Charlson comorbidity index					
0	106,201 (42%)	47,680 (36%)	4,413 (20%)	393 (3.9%)	4,338 (33%)
1-2	65,945 (26%)	32,422 (25%)	5,668 (26%)	3,109 (31%)	3,353 (25%)
3-4	37,354 (15%)	23,045 (18%)	3,486 (16%)	2,342 (23%)	2,996 (23%)
>=5	46,155 (18%)	28,113 (21%)	8,096 (37%)	4,234 (42%)	2,593 (20%)
Bannay comorbidity index					
0	54,635 (21%)	10,941 (8.3%)	1,532 (7.1%)	137 (1.4%)	1,883 (14%)
1-2	49,949 (20%)	24,568 (19%)	2,647 (12%)	631 (6.3%)	2,106 (16%)
3-4	63,048 (25%)	41,046 (31%)	5,213 (24%)	2,463 (24%)	3,908 (29%)
>=5	88,023 (34%)	54,705 (42%)	12,271 (57%)	6,847 (68%)	5,383 (41%)
Quan comorbidity index					
0	119,033 (47%)	54,718 (42%)	4,907 (23%)	531 (5.3%)	4,997 (38%)
1-2	62,819 (25%)	30,413 (23%)	6,250 (29%)	4,316 (43%)	3,180 (24%)
3-4	28,917 (11%)		18,421 (14%)	2,644 (12%)	2,502 (19%)

	>=5	44,886 (18%)	27,708 (21%)	7,862 (36%)	4,014 (40%)	2,601 (20%)	2,634 (34%)
Surgery type							ANNEXE
Colorectal surgery	131,260 (51%)	131,260 (100%)		2,343 (11%)	587 (5.8%)	3,623 (27%)	759 (14%)
Hepatobiliary surgery	21,663 (8.5%)	2,343 (1.8%)		21,663 (100%)	481 (4.8%)	181 (1.4%)	541 (0.8%)
Pancreatic surgery	10,078 (3.9%)	587 (0.4%)		481 (2.2%)	10,078 (100%)	106 (0.8%)	521 (0.7%)
Small bowel surgery	13,280 (5.2%)	3,623 (2.8%)		181 (0.8%)	106 (1.1%)	13,280 (100%)	292 (0.4%)
Esogastric surgery	70,298 (27%)	759 (0.6%)		541 (2.5%)	521 (5.2%)	292 (2.2%)	70,298 (100%)
Death postoperative within 6 mos	10,230 (4.0%)	6,802 (5.2%)		1,067 (4.9%)	642 (6.4%)	753 (5.7%)	887 (13%)

n (%) ; Median (IQR)

TABLE B.3 – Nombre de réadmissions post-opératoires dans les 6 mois des patients ayant bénéficié d'une chirurgie digestive majeure entre 2020 et 2022 en France (Source : PMSI)

Characteristic	Whole Sample	Colorectal Surgery	Hepatobiliary Surgery	Pancreatic Surgery	Small Bowel Surgery	Esogastric Surgery
	N = 255,732	N = 131,260	N = 21,663	N = 10,078	N = 13,280	N = 70,298 ¹
Total readmissions	1.0 (0.0, 3.0)	1.0 (0.0, 4.0)	1.0 (0.0, 6.0)	5.0 (1.0, 9.0)	1.0 (0.0, 3.0)	0.00 (0.00, 1.00)
Readmission category						
0	105,285 (41%)	43,399 (33%)	7,090 (33%)	1,419 (14%)	5,277 (40%)	42,889 (61%)
1	53,281 (21%)	30,028 (23%)	4,190 (19%)	1,573 (16%)	3,029 (23%)	13,770 (20%)
2	22,712 (8.9%)	13,444 (10%)	1,915 (8.8%)	796 (7.9%)	1,325 (10.0%)	4,993 (7.1%)
>=3	74,454 (29%)	44,389 (34%)	8,468 (39%)	6,290 (62%)	3,649 (27%)	8,646 (12%)

n (%); Median (IQR)

ANNEXE 8

TABLE B.4 – Actes performés des readmissions post-opératoires dans les 6 mois des patients ayant bénéficié d'une chirurgie digestive majeure entre 2020 et 2022 en France (Source : PMSI)

Act	Whole Sample		Colorectal Surgery		Hepatobiliary Surgery		Pancreatic Surgery		Small Bowel Surgery		Esogastric Surgery	
	N = 686,918	N = 413,617	N = 73,650	N = 56,211	N = 34,978	N = 76,828						
Bile duct surgery	119 (<0.1%)	19 (<0.1%)	71 (<0.1%)	28 (<0.1%)	3 (<0.1%)	6 (<0.1%)						
Bisegmentectomy	178 (<0.1%)	142 (<0.1%)	41 (<0.1%)	0 (0%)	6 (<0.1%)	3 (<0.1%)						
Colectomy	2,328 (0.3%)	1,696 (0.4%)	355 (0.5%)	43 (<0.1%)	113 (0.3%)	71 (<0.1%)						
Complication abdominal wall abscess	1,213 (0.2%)	764 (0.2%)	87 (0.1%)	38 (<0.1%)	106 (0.3%)	152 (0.2%)						
Complication cardiac	2,260 (0.3%)	1,496 (0.4%)	258 (0.4%)	110 (0.2%)	148 (0.4%)	206 (0.3%)						
Complication renal	28,791 (4.2%)	19,086 (4.6%)	2,582 (3.5%)	618 (1.1%)	2,827 (8.1%)	3,489 (4.5%)						
Complication respiratory assistance	3,680 (0.5%)	2,082 (0.5%)	430 (0.6%)	305 (0.5%)	227 (0.6%)	724 (0.9%)						
Complication transfusion	13,189 (1.9%)	7,574 (1.8%)	1,491 (2.0%)	1,101 (2.0%)	1,144 (1.5%)							
Gastrectomy	389 (<0.1%)	57 (<0.1%)	48 (<0.1%)	26 (<0.1%)	18 (<0.1%)	230 (0.3%)						
Hepatectomy central	1,445 (0.2%)	1,105 (0.3%)	457 (0.6%)	9 (<0.1%)	47 (0.1%)	20 (<0.1%)						
Hepatectomy laparotomy	24 (<0.1%)	22 (<0.1%)	3 (<0.1%)	0 (0%)	1 (<0.1%)	0 (0%)						
Hepatectomy laparoscopy	1,092 (0.2%)	826 (0.2%)	373 (0.5%)	8 (<0.1%)	36 (0.1%)	8 (<0.1%)						
Hepatectomy left	366 (<0.1%)	289 (<0.1%)	88 (0.1%)	1 (<0.1%)	11 (<0.1%)	12 (<0.1%)						
Hepatectomy major	71 (<0.1%)	52 (<0.1%)	23 (<0.1%)	1 (<0.1%)	1 (<0.1%)	0 (0%)						
Hepatectomy minor	507 (<0.1%)	356 (<0.1%)	259 (0.4%)	4 (<0.1%)	17 (<0.1%)	4 (<0.1%)						
Hepatectomy right	534 (<0.1%)	440 (0.1%)	115 (0.2%)	4 (<0.1%)	13 (<0.1%)	8 (<0.1%)						
Hepatectomy wedge	364 (<0.1%)	254 (<0.1%)	204 (0.3%)	3 (<0.1%)	13 (<0.1%)	4 (<0.1%)						
Lobectomy left	680 (<0.1%)	547 (0.1%)	154 (0.2%)	2 (<0.1%)	25 (<0.1%)	10 (<0.1%)						
Lobectomy right	87 (<0.1%)	81 (<0.1%)	9 (<0.1%)	3 (<0.1%)	0 (0%)	1 (<0.1%)						
Oesophagectomy	51 (<0.1%)	31 (<0.1%)	31 (<0.1%)	0 (0%)	2 (<0.1%)	0 (0%)						
Pancreatectomy	45 (<0.1%)	9 (<0.1%)	12 (<0.1%)	0 (0%)	1 (<0.1%)	20 (<0.1%)						
Peritoneal surgery	130 (<0.1%)	19 (<0.1%)	62 (<0.1%)	34 (<0.1%)	4 (<0.1%)	9 (<0.1%)						
Radiologic arteriography intervention	825 (0.1%)	349 (<0.1%)	57 (<0.1%)	11 (<0.1%)	40 (0.1%)	39 (<0.1%)						
	1,490 (0.2%)	736 (0.2%)	358 (0.5%)	286 (0.5%)	86 (0.2%)	140 (0.2%)						

Radiologic biopsy	1,217 (0.2%)	802 (0.2%)	164 (0.2%)	70 (0.1%)	81 (0.2%)
Radiologic drainage	4,676 (0.7%)	1,929 (0.5%)	1,240 (1.7%)	583 (1.0%)	216 (0.6%)
abdomen					516 (0.7%)
Radiologic portal vein embolisation	286 (<0.1%)	171 (<0.1%)	185 (0.3%)	5 (<0.1%)	6 (<0.1%)
Rectal resection	1,066 (0.2%)	647 (0.2%)	360 (0.5%)	1 (<0.1%)	19 (<0.1%)
Renal dialysis	1,335 (0.2%)	848 (0.2%)	159 (0.2%)	122 (0.2%)	88 (0.3%)
Reintervention	15,264 (2.2%)	10,414 (2.5%)	1,203 (1.6%)	486 (0.9%)	945 (2.7%)
Segmentectomy	242 (<0.1%)	199 (<0.1%)	57 (<0.1%)	0 (0%)	5 (<0.1%)
Segmentectomy 1	23 (<0.1%)	18 (<0.1%)	8 (<0.1%)	0 (0%)	1 (<0.1%)
Small bowel resection	2,506 (0.4%)	2,262 (0.5%)	73 (<0.1%)	9 (<0.1%)	54 (<0.1%)
Surgical shunt	5 (<0.1%)	2 (<0.1%)	1 (<0.1%)	1 (<0.1%)	1 (<0.1%)
Thoracic drainage	1,639 (0.2%)	695 (0.2%)	324 (0.4%)	145 (0.3%)	77 (0.2%)
Trisegmentectomy	51 (<0.1%)	42 (<0.1%)	10 (<0.1%)	1 (<0.1%)	3 (<0.1%)
Vascular resection	49 (<0.1%)	19 (<0.1%)	25 (<0.1%)	7 (<0.1%)	1 (<0.1%)

Annexe C

Chapitre 4

C.1 Programmation de TreeSHAP pour l'exemple

✓ Tree SHAP- Example

```
!pip install shap

Requirement already satisfied: shap in c:\users\murrisj\anaconda3\lib\site-packages (0.40.0)
Requirement already satisfied: numba in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (0.54.1)
Requirement already satisfied: scikit-learn in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (1.0.2)
Requirement already satisfied: tqdm>4.25.0 in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (4.62.3)
Requirement already satisfied: pandas in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (1.4.2)
Requirement already satisfied: slicer==0.0.7 in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (0.0.7)
Requirement already satisfied: numpy in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (1.20.3)
Requirement already satisfied: packaging>20.9 in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (21.0)
Requirement already satisfied: cloudpickle in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (2.0.0)
Requirement already satisfied: scipy in c:\users\murrisj\anaconda3\lib\site-packages (from shap) (1.7.1)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\murrisj\anaconda3\lib\site-packages (from packaging>20.9->shap) (3.0.4)
Requirement already satisfied: colorama in c:\users\murrisj\anaconda3\lib\site-packages (from numba->shap) (0.4.4)
Requirement already satisfied: setuptools in c:\users\murrisj\anaconda3\lib\site-packages (from numba->shap) (58.0.4)
Requirement already satisfied: llvmlite<0.38,>=0.37.0rc1 in c:\users\murrisj\anaconda3\lib\site-packages (from numba->shap) (0.37.0)
Requirement already satisfied: pytz>=2020.1 in c:\users\murrisj\anaconda3\lib\site-packages (from pandas->shap) (2021.3)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\murrisj\anaconda3\lib\site-packages (from pandas->shap) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\murrisj\anaconda3\lib\site-packages (from python-dateutil>=2.8.1->pandas->shap)
Requirement already satisfied: joblib>=0.11 in c:\users\murrisj\anaconda3\lib\site-packages (from scikit-learn->shap) (1.1.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\murrisj\anaconda3\lib\site-packages (from scikit-learn->shap) (2.2.0)

import numpy as np
import pandas as pd
import shap
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor, plot_tree
%matplotlib inline

np.random.seed(100)
X_train = pd.DataFrame({'x':[206]*5 + [194] + [6]*4,
                       'y': list(np.random.randint(100, 400, 6)) + [299, 299, 301, 301],
                       'z': list(np.random.randint(100, 400, 10))})
X_train

      x    y    z
0  206  108  114
1  206  380  390
2  206  179  340
3  206  153  380
4  206  166  243
5  194  326  328
6     6  299  158
7     6  299  237
8     6  301  193
9     6  201  196

# target_train = pd.Series([10]*5 + [20] + [50]*2 + [30]*2)
target_train = pd.Series([8, 9, 10, 11, 12] + [20] + [45, 55] + [28, 32])
target_train.name = 't'
target_train

0    8
1    9
2   10
3   11
4   12
5   20
6   45
7   55
8   28
9   32
Name: t, dtype: int64

tree_model = DecisionTreeRegressor(criterion='mae', max_depth=2, random_state = 100)

tree_model.fit(X=X_train, y=target_train)
```

https://colab.research.google.com/drive/1594h8Vfst0PzT_m6TIC9MywQdDO0lVtB#scrollTo=mGkZlTkrcHP4&printMode=true

1/2

ANNEXES

05/08/2024 16:07 Tree_SHAP_Hypothetical_Example.ipynb - Colab

```
Criterion 'mae' was deprecated in v1.0 and will be removed in version 1.2. Use `criterion='absolute_error'` which is equivalent.
DecisionTreeRegressor(criterion='mae', max_depth=2, random_state=100)

plot_tree(tree_model, filled=True)

[Text(0.5, 0.8333333333333334, 'X[0] <= 100.0\nmae = 13.0\nsamples = 10\nvalue = 16.0'),
 Text(0.25, 0.5, 'X[1] <= 300.0\nmae = 10.0\nsamples = 4\nvalue = 38.5'),
 Text(0.125, 0.1666666666666666, 'mae = 5.0\nsamples = 2\nvalue = 50.0'),
 Text(0.375, 0.1666666666666666, 'mae = 2.0\nsamples = 2\nvalue = 30.0'),
 Text(0.75, 0.5, 'X[0] <= 200.0\nmae = 2.667\nsamples = 6\nvalue = 10.5'),
 Text(0.625, 0.1666666666666666, 'mae = 0.0\nsamples = 1\nvalue = 20.0'),
 Text(0.875, 0.1666666666666666, 'mae = 1.2\nsamples = 5\nvalue = 10.0')]



shap.initjs()  
shap.force_plot(explainer.expected_value, shap_values[0,:], X_test.iloc[0,:])


```

