Modélisation flexible pénalisée d'évènements récurrents. Développement d'un modèle d'intensité marginale avec application aux effets indésirables associés à l'immunothérapie anti-cancéreuse

Elsa Coz

Université Claude Bernard Lyon 1

THESE de DOCTORAT DE L'UNIVERSITE CLAUDE BERNARD LYON 1

Ecole Doctorale ED 341
Ecosystèmes Evolution Modélisation Microbiologie (E2M2)

Discipline: Biostatistiques, Santé Publique

Soutenue publiquement le 04/11/2024, par :

Elsa Coz

Modélisation flexible pénalisée d'évènements récurrents. Développement d'un modèle d'intensité marginale avec application aux effets indésirables associés à l'immunothérapie anti-cancéreuse

---

Devant le jury composé de :

RONDEAU, Virginie	DR INSERM, Université de Bordeaux	Rapporteure
JOLY, Pierre	PU, Université de Bordeaux	Rapporteur
CIUPERCA, Gabriela	PU, Université de Lyon	Examinatrice
BOUAZIZ, Olivier	PU, Université de Lille	Examinateur
MAUCORT-BOULCH, Delphine	PU-PH, Université de Lyon	Directrice de thèse
HACID, Mohand-Saïd	PU, Université de Lyon	Co-directeur de thèse
FAUVERNIER, Mathieu	MCU, Université de Lyon	Invité

Résumé

Les mécanismes associés aux effets indésirables sous immunothérapie et l'identification de profils à risque chez les patients font à ce jour l'objet de nombreuses études. Cependant, les données de toxicités associées à un traitement sont souvent d'une grande complexité et posent de nombreux défis aussi bien en termes de recueil que d'un point de vue méthodologique. L'objectif de cette thèse est d'explorer des approches statistiques permettant de comparer des profils de toxicité associés à un traitement. Du fait de la multi-dimensionnalité de ces données (e.g. temporalité, récurrence, sévérité), évaluer la toxicité d'un traitement peut se faire à travers de nombreux indicateurs donnant soit une mesure absolue du risque de toxicité étudié (e.g. probabilités, taux), soit une mesure relative (e.g. odds ratios, hazard ratios). Une première revue des indicateurs liés aux modèles de régression proposés dans la littérature a mis en lumière des aspects méthodologiques fréquemment négligés dans l'évaluation du risque, en particulier en ce qui concerne la représentation du risque absolu au cours du temps. En réponse à cette observation, deux modèles ont ensuite été explorés au cours de ces travaux.

Le premier modèle est un modèle flexible pénalisé du taux de survenue d'un unique évènement de toxicité (e.g. première apparition, grade maximum). L'intérêt de ce modèle réside dans sa capacité à décrire de manière flexible la dynamique du taux au cours du temps en utilisant des splines, tout en intégrant les effets non linéaires et non proportionnels des covariables. L'utilisation de splines s'accompagne toutefois de problématiques de sur-ajustement potentiel. Introduire une pénalisation de la vraisemblance, avec un objectif de lissage, permet de réduire ce phénomène. Le modèle de taux pénalisé a été exploré par simulation en tenant compte des spécificités du contexte des données de toxicité : échantillons de taille modeste, risques compétitifs, non-proportionnalité, dynamique différente de celle rencontrée pour la survenue du décès. L'application du modèle sur des données observationnelles de patients traités par immunothérapie dans le cadre du projet Européen Qualitop a également été réalisée.

Le second modèle offre une description de la dynamique de survenue des événements basée sur des splines, tout en prenant en compte la possibilité d'une récurrence de ces événements. L'indicateur utilisé est l'intensité marginale (IM), qui présente l'avantage de décrire un processus récurrent au sein d'une population de manière synthétique. Deux approches ont été proposées dans cette thèse pour modéliser l'indicateur avec des splines. La première, dite directe, reprend le cadre du modèle d'IM semi-paramétrique basé sur des équations d'estimation, qui ne nécessite pas d'hypothèses strictes sur la corrélation des événements chez un même sujet. La seconde, dite indirecte, estime l'IM à partir d'un modèle d'intensité basé sur le processus de Poisson avec effet aléatoire. Les performances des deux approches sont comparées par une étude de simulation originale, basée sur des processus de type multi-états. Les deux approches ont ensuite été étendues au cadre pénalisé, avec objectif de lissage, afin de limiter le sur-ajustement.

Mots clés : Effets indésirables, Survie, Evènements récurrents, Splines pénalisées, Taux, Intensité marginale

Abstract

The mechanisms associated with adverse events under immunotherapy and the identification of risk profiles in patients are currently the subject of numerous studies. However, toxicity data associated with treatment are often highly complex and present many challenges both in terms of data collection and in terms of methodological approaches. The objective of this thesis is to explore statistical methods for comparing toxicity profiles associated with treatment. Due to the multidimensionality of these data (e.g., temporality, recurrence, severity), evaluating treatment toxicity can involve numerous absolute indicators (e.g., probabilities, hazards) and relative indicators (e.g., odds ratios, hazard ratios). A preliminary review of the indicators associated with regression models proposed in the literature highlighted methodological aspects that are frequently overlooked in risk assessment, particularly regarding the representation of absolute risk over time. In response to this observation, two models were explored during this research.

The first model is a penalized flexible hazard model focusing on the occurrence of a single event (e.g. first event, maximum grade). The value of this model lies in its ability to flexibly describe the dynamics of the hazard according to time using splines while also accounting for non-linear and non-proportional effects of covariates. However, the use of splines may lead to overfitting issues. Adding a penalty to the likelihood, with a smoothing objective, helps mitigate this issue. The so-called penalized hazard model was explored through simulation, taking into account the specificities of adverse events data : modest sample sizes, competing risks, non-proportionality, and dynamics different from those encountered with death as the event of interest. The model was also applied to observational data from patients treated with immunotherapy within the European Qualitop project.

The second model also describes the dynamics of event occurrence using splines while accounting for the possibility of event recurrence. The indicator used is the marginal intensity (MI), which provides a synthetic description of a recurrent process within a population. Two approaches were proposed in this thesis for modelling the indicator with splines. The first, called direct, follows the framework of the semi-parametric MI model based on estimating equations, which does not require strict assumptions about the correlation of events within the same subject. The second, called indirect, estimates the MI from a Poisson process intensity-based model with a random effect. The performances of these two approaches are compared through an original simulation study based on multi-state processes. Both approaches were then extended to the penalized smoothing framework in order to limit overfitting.

Key words : Adverse events, Survival, Recurrent events, Penalized splines, Hazard, Marginal intensity

Remerciements

Je souhaite adresser mes remerciements à ma directrice de thèse, Delphine Maucort-Boulch, ainsi qu'à mon co-directeur Mohand-Saïd Hacid pour m'avoir offert cette opportunité, pour leurs conseils et leurs grandes qualités humaines. Je remercie, en particulier, Delphine pour son écoute, sa confiance et sa patience.

Je remercie également Mathieu Fauvernier, mon co-encadrant, pour son enthousiasme et son optimisme sans fin, ses grandes qualités scientifiques et pédagogiques (et ses blagues !). Un grand merci pour les passionnantes sessions de *brain-storming* devant un tableau plus ou moins blanc.

Mais ces grandes discussions n'auraient pas eu la même saveur sans mes co-équipiers de bureau. Merci à Malek pour sa légendaire bonne humeur, sa gentillesse et son écoute. Merci à Nicolas pour son humour corrosif et pour être un excellent *science buddy*[1] en toute occasion.

Je remercie également chaleureusement toute l'équipe du service de biostatistiques des HCL pour leur accueil et leur bienveillance. En particulier, merci à Malek, Mad-Hélénie et Sylvain pour leur aide et leur soutien dans l'aventure Qualitop. Merci à Paola, Véronique et surtout Stéphanie pour toute leur aide pour nous simplifier la vie sur l'aspect administratif. Merci également à Zoé et Laurent pour leurs remarques toujours très constructives et leur intérêt pour mes (nombreuses) questions. Merci à mes compagnons d'aventure de doctorats : Charles-Hervé, Alexandre, Levi, Maxime, Yue et Nicolas.

Je remercie Hadrien Charvat pour ses précieux apports scientifiques, sa bienveillance et ses incroyables talents d'optimisation.

Le projet Qualitop fut également l'occasion de belles rencontres et je remercie toute l'équipe pour ces échanges et en particulier le groupe des lyonnais : Marta, Stéphane, Juba, Aurore, Alexandra, Léa, pour leur bonne humeur lors des déplacements à travers l'Europe.

Je remercie Virginie Rondeau, Pierre Joly, Gabriela Ciuperca et Olivier Bouaziz pour avoir accepté d'être membres de mon jury.

Je remercie tous mes proches, ma famille et amis pour leur soutien, en particulier mes parents et ma soeur. Merci à mon Papa pour la relecture de ce manuscrit.

Merci Alexandre, pour ta confiance, ta patience et ton écoute (en particulier, les longs monologues sur les modèles) pendant toute cette aventure.

1. Yanai, I., Lercher, M.J. It takes two to think. Nat Biotechnol 42, 18–19 (2024). https ://doi.org/10.1038/s41587-023-02074-2

Table des matières

Liste des abréviations

AG Andersen-Gill

EI Effets indésirables

GAM *Generalized Additive Models*

ICI Immune checkpoint inhibitors

IM Intensité marginale

ir-AE Effets indésirables immuno-médiés (*Immune related adverse events)*

MAP Maximum *a posteriori*

MFT Modèle flexible du taux

PWP Prentice-Williams-Petersen

REML *Restricted Maximum Likelihood*

RMSE *Rooted Mean Squared Error*

Table des figures

Partie I Modéliser des données de toxicité

Partie II Modéliser les effets indésirables avec un modèle de taux flexible

Partie III Les modèles flexibles sur l'intensité marginale dans un contexte d'évènements récurrents

Annexes

Liste des tableaux

Contexte et objectifs de la thèse

Contexte

L'immunothérapie anti-cancéreuse a révolutionné la prise en charge de nombreux cancers au cours des dernières années. Les inhibiteurs du contrôle immunitaire (*Immune Checkpoint Inhibitors*) (ICI) constituent une famille de traitements agissant en bloquant certains récepteurs du système immunitaire comme le *cytotoxic T-lymphocyte-associated antigen 4* (CTLA-4), le *programmed death cell protein 1* (PD-1) ou le PD-ligand 1 (PD-L1). En bloquant ces récepteurs, les ICI provoquent une réponse accrue de la part des lymphocytes T visant à éliminer les cellules tumorales. De plus en plus de molécules ont obtenu des autorisations de mise sur le marché : anti-PD1 (e.g. nivolumab, pembrolizumab, cemiplimab), anti-CTLA4 (ipilimumab), anti-PDL1 (e.g. atezolizumab, durvalumab). Les ICI sont utilisés pour le traitement de nombreux cancers comme le mélanome avancé (Carlino et al., 2021), le cancer du poumon (Tang et al., 2022), la liste des nouvelles applications s'allongeant chaque année (Yin et al., 2023). Par ailleurs, ils sont aussi utilisés comme adjuvant (en complément d'une chirurgie). Les ICI peuvent être prescrits en monothérapie mais aussi en combinaison avec d'autres ICI (e.g. nivolumab + ipilimumab), de la chimiothérapie ou des thérapies ciblées.

En stimulant la réponse immunitaire des lymphocytes T pour combattre le cancer, des réactions auto-immune peuvent alors se produire. Ces réactions sont communément appelées effets indésirables immuno-médiés (*immune-related adverse events* ir-AEs) et peuvent affecter n'importe quel organe. La fréquence d'ir-AEs est importante bien que cela dépende de l'ICI utilisé. Par exemple, les troubles endocriniens, e.g. hyperthyroïdie, hypothyroïdie, hypophysite, sont très fréquents sous anti-PD1 (Barroso-Sousa et al., 2018). La sévérité des ir-AEs est variable, allant de symptômes bénins à des évènements fatals (e.g. pneumonites, myocardites). Afin d'améliorer la prise en charge de ces toxicités, l'identification de facteurs prédictifs des évènements indésirables est un champ de recherche actif.

La révolution dans la prise en charge de cancers auparavant associés à un mauvais pronostic (e.g. le mélanome métastatique) a conduit de plus en plus d'essais cliniques à considérer la qualité de vie comme résultat d'intérêt, en complément de la survie. Certains essais cliniques ont rapporté : une meilleure qualité de vie entre le début du traitement et la 12^{eme} semaine chez les patients sous pembrolizumab que sous chimiothérapie (Schadendorf et al., 2016), une plus faible propension à la détérioration de la qualité de vie sous nivolumab que sous chimiothérapie (Long et al., 2016), une différence de qualité de vie non cliniquement significative en cours de traitement comparé à l'initiation pour des thérapies adjuvantes (Bottomley et al., 2021). L'association entre ir-AEs et qualité de vie (Malkhasyan et al., 2017) a été peu étudiée et la plupart des études sont basées sur des patients sélectionnés pour les essais cliniques.

Cette thèse s'inscrit dans le cadre du projet européen Qualitop (*QUality of Life after cancer ImmunoTherapy*, dont l'objectif général était de créer une cohorte multi-nationale de patients traités en routine par immunothérapie anti-cancéreuse. Plusieurs sujets de recherche ont été

abordés avec les données collectées dans cette cohorte : recherche de marqueurs prédictifs des ir-AEs, des déterminants de la qualité de vie de ces patients et notamment son lien avec les ir-AEs. L'un des objectifs était également d'intégrer les résultats sur une plateforme digitale.

Objectifs

L'objectif principal de cette thèse était de proposer une modélisation du risque de toxicité, pour le décrire et identifier des profils à risque, chez des patients traités par immunothérapie anti-cancéreuse.

Une première partie exploratoire a donc été réalisée afin d'identifier (i) les caractéristiques propres aux données d'effets indésirables collectées à partir des dossiers patients, (ii) les problématiques statistiques associées, (iii) les modèles de régression statistiques proposés dans la littérature permettant de gérer ces données.

Constatant que la survenue des évènements au cours du temps est rarement décrite dans la littérature (e.g. simple présentation des temps médians d'occurrence), la seconde partie de la thèse a donc porté sur le modèle flexible du taux. Ces derniers ont été explorés dans le contexte des effets indésirables, notamment dans un cadre de risques compétitifs.

La récurrence des évènements indésirables étant souvent négligée et la modélisation limitée au premier évènement, la troisième partie de la thèse propose une modélisation flexible de l'intensité marginale (souvent appelée *rate* dans la littérature) au cours du temps.

Partie I

Collecter et modéliser des données de toxicité

1 La cohorte Qualitop

1.1 Description générale

La cohorte prospective Qualitop est une cohorte multi-centrique et multi-nationale (Vinke et al., 2023) de patients sous immunothérapie anti-cancéreuse. Deux principales catégories de traitements peuvent être identifiées dans cette cohorte : les *immune checkpoint inhibitors* (ICI) et les CAR-T cells [1]. Les patients peuvent être inclus dans la cohorte s'ils sont âgés de plus de 18 ans à la date de signature du consentement éclairé et durant la date de décision pour l'immunothérapie jusqu'au début du second cycle de traitement.

La cohorte implique les populations suivants :

- **Les Hospices Civils de Lyon (France)**
  La cohorte française inclus des patients traités par ICI (tous types de cancers) et par CAR-T cells (lymphomes).
- **University Medical Center Groningen (Pays-Bas)**
  Les patients inclus sont atteints d'un cancer du poumon traité par ICI (cohorte OncoLifeS).
- **La cohorte nationale de patients traités par CAR-T (Pays-Bas)**
  Les patients sont traités pour un lymphome par CAR-T cells.
- **Instituto Português de Oncologia, Lisboa (Portugal)**
  Les patients sont traités pour un lymphome par ICI ou par CAR-T cells.
- **Hospital Clinic de Barcelona (IDIBAPS) (Espagne)**
  Cette cohorte se compose de patients atteints de mélanomes traités par ICI dans le département de dermatologie.

1. Les traitements par CAR-T cells (*Chimeric Antigenic Receptor T*) constituent un autre type d'immunothérapie. Ce sont des lymphocytes T modifiés génétiquement afin qu'ils puissent reconnaître et éliminer les cellules cancéreuses. Le profil de toxicité de ces traitements est très différent de celui des ICI, incluant des syndromes de neurotoxicité et de relargage cytokinique.

1.2 La cohorte des mélanomes

Parmi les axes de recherche identifiés dans le cadre du projet, il a semblé pertinent de faire un focus sur la cohorte des mélanomes. Cette cohorte est plutôt homogène en termes de traitements, les patients étant principalement traités en monothérapie par anti-PD1 (nivolumab ou pembrolizumab) pour de la première ligne ou en adjuvant. Dans la suite de cette thèse, les modèles ont été appliqués sur les données issues de cette cohorte.

Les patients mélanomes sont inclus en France et en Espagne. En France, les patients traités par ICI ont été inclus entre le 15 juillet 2019 et le 16 février 2024 et en Espagne, entre le 8 août 2021 et le 11 mars 2024. A la fin des inclusions, la cohorte totale comptait 423 patients atteints de mélanomes. Dans la suite, nous en conservons 279, traités en ligne 1 ou adjuvant et en monothérapie d'anti-PD1, comme indiqué dans le flowchart en Figure 1.1.

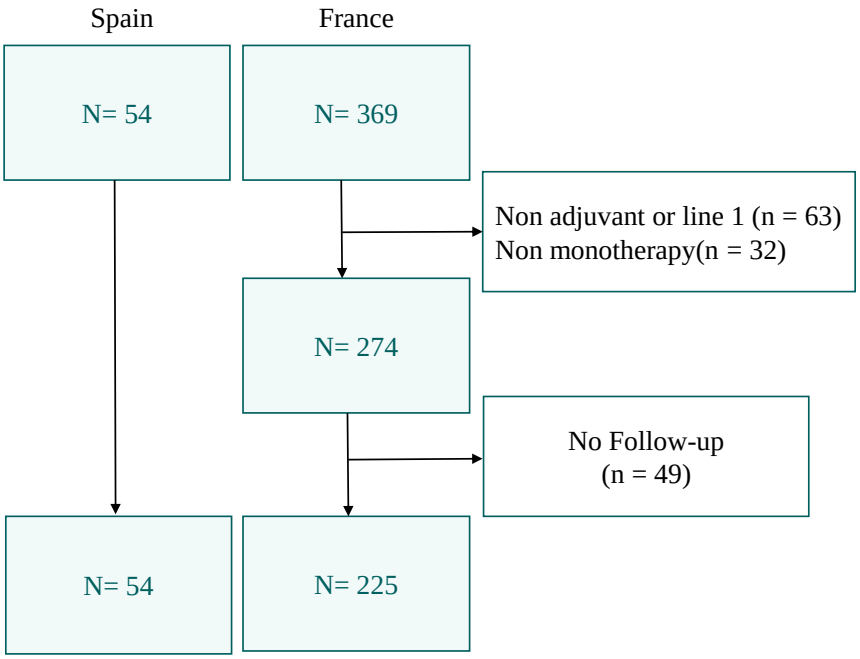

FIGURE 1.1 – Flowchart de la cohorte des mélanomes

Les caractéristiques des patients au début du traitement sont détaillées en Table 1.1. La principale différence entre les populations des deux pays est la proportion de patients avec des métastases distantes. En Espagne, les patients avec un mélanome non résécable sont traités dans un autre département que celui qui réalise les recrutements et donc non inclus dans l'étude. L'âge médian des patients de la cohorte est de 66 ans et la plupart des patients ont reçu du nivolumab (74%).

	France, N = 225	**Espagne**, N = 54	**Total**, N = 279
Sexe			
Homme	127 (56%)	25 (46%)	152 (54 %)
Femme	98 (44%)	29 (54%)	127 (46%)
Age en début d'immunothérapie			
	68 (53 - 76)	63 (53 - 74)	66 (53 - 75)
Métastases distantes			
Oui	66 (30%)	1 (2%)	67 (24 %)
Non	155 (70%)	52 (98%)	207 (76%)
Manquant	4	1	5
Historique de maladie auto-immune			
Oui	25 (11%)	4 (7%)	29 (10 %)
Non	200 (89%)	50 (93%)	250 (90%)
Type d'anti-PD1			
Nivolumab	170 (76%)	36 (67%)	206 (74 %)
Pembrolizumab	55 (24%)	18 (33%)	73 (26%)
Ratio neutrophiles sur lymphocytes (NLR)			
	2.4 (1.7 - 3.4)	2.4 (1.7 - 3.1)	2.4 (1.7 - 3.3)
Manquant	1	0	1

TABLE 1.1 – Caractéristiques des patients mélanomes à l'inclusion. *Les variables continues sont décrites par la médiane et l'intervalle interquartile.*

2 Données

2.1 Mode de collecte de données

Les informations cliniques des patients ont été collectées manuellement à partir du dossier patient de façon rétrospective. En France, le recueil des données a été effectué sous *Easily*, une plateforme de dossiers patients informatisés développée par les Hospices Civils de Lyon. Dans les autres centres, les données ont été collectées dans REDCap (*Research Electronic Data Capture*). Il s'agit d'une plate-forme en ligne conçue pour recueillir des données de recherche.

Les données cliniques du patient ont été recueillies depuis le dossier patient jusqu'à un maximum de 24 mois (sauf interruption de l'immunothérapie et début d'une autre ligne de traitement). Au-delà, seul le statut vital et la date de dernière injection ont été mis à jour à intervalles réguliers. Pour les patients sous immunothérapie, il est possible d'identifier plusieurs motifs de visites, conduisant à une fréquence de recueil d'information plus ou moins importante selon les patients :

— *Visites pour traitement*
Elles ont lieu à chaque cycle d'immunothérapie, soit toutes les 2 à 6 semaines selon la dose et le type de traitement. Une analyse biologique est réalisée à cette occasion avant l'injection du traitement.

— *Visites pour toxicité*
Elle peut survenir entre deux cycles en cas de toxicité et peut conduire à une réduction de la dose lors du cycle suivant voire à un arrêt temporaire ou définitif du traitement. Une hospitalisation est parfois nécessaire. Des analyses biologiques complémentaires à celles des visites de routines peuvent être réalisées.

— *Visite d'évaluation*
A cette occasion, une évaluation de la réponse au traitement est réalisée afin de décider de la poursuite de ce dernier. Une évaluation est, en général, réalisée au bout des 6 premiers mois mais elle peut également survenir plus tôt (par exemple au bout de 3 mois ou au bout d'un certain nombre de cycles). Elle peut coïncider avec une visite pour traitement.

2.2 Variables collectées

Plusieurs types de variables ont été collectées dans la base de données Qualitop à partir des dossiers patients :

- *Variables démographiques* : date de naissance, sexe ;
- *Historique des co-morbidités* : ces dernières peuvent être recueillies sous forme d'un texte libre ou classées selon la classification des maladies CIM-10.
- *Caractéristiques du/des traitements* : en cours et passés (doses, fréquences, molécules) ;
- *Caractéristiques du cancer en début d'immunothérapie* : type histologique, présence et localisation des métastases ;
- *Traitements concomitants* : ces traitements sont recueillis sous forme de texte libre ;
- *Résultats d'examens biologiques* ;
- *Résultats d'examens cliniques* : poids, statut ECOG (*Eastern cooperative oncology group*), pression artérielle, etc. ;
- *Effets indésirables* : La section 2.3 est dédiée à la description de la collecte et de la structure de ces derniers.

2.3 Les évènements indésirables en données observationnelles

Les informations associées aux évènements indésirables ont été collectées depuis le dossier patient, incluant la date de début, la date de fin, la nécessité d'une hospitalisation. Une imputabilité au traitement est parfois proposée. Cependant, cette information peut être partielle, manquante ou manquer d'objectivité (Lineberry et al., 2016).

Classiquement, les évènement ont été classés et gradés selon la *Common Terminology Criteria for Adverse Events* (CTCAE) Version 5.0 de l'Institut National des Cancers. La sévérité y est décrite par un grade allant de 1 (symptômes légers ou cliniques) à 5 (décès). La CTCAE comporte des éléments objectifs de type analyses de laboratoires (e.g. augmentation des lipases), des évènements cliniques mesurables/observables (e.g. rash maculopapuleux) et des évènements symptomatiques subjectifs (e.g. nausées, douleurs) (Basch et al., 2014).

Plusieurs facteurs peuvent influencer le niveau d'exhaustivité du dossier patient en termes d'effets indésirables. En premier lieu, l'appréciation du clinicien joue un rôle central au moment de l'évaluation lors d'une visite. En particulier, il a été observé une tendance à sous-estimer les évènements symptomatiques, à la fois en fréquence mais aussi du point de vue de la sévérité, par rapport aux patients (Atkinson et al., 2016). C'est pourquoi il est de plus en plus commun d'inclure des observations rapportées par le patient (*Patient Reported Outcome* (PRO)) dans les essais pour compléter les données collectées par les cliniciens. En second lieu, l'engagement du patient pour sa santé, son âge et sa littéracie en santé (Whittaker et al., 2017) ainsi que la communication entre l'équipe médicale et le patient sont d'autres facteurs qui influencent l'identification des effets indésirables. Enfin, la fréquence des visites est un facteur majeur de cette collecte, les évènements ponctuels entre deux visites pouvant être complètement omis.

Dans les essais cliniques, il est fortement recommandé de spécifier les évènements indésirables d'intérêt dans le protocole de l'essai (Lineberry et al., 2016). En effet, pour certains

évènements d'intérêt, le CRF (*case-report form*) devrait être spécialement conçu pour capturer certains évènements d'intérêt, qui pourrait nécessiter des investigations complémentaires non nécessairement réalisées en routine pour en établir le diagnostic (Crowe et al., 2009). Avec des données collectées à partir du dossier patient, on comprend que ce problème est omniprésent.

Quelques cas particuliers à l'immunothérapie

Bien que la CTCAE soit en constante évolution pour répondre aux besoins spécifiques des traitements les plus récents, elle n'est pas toujours optimale pour les décrire. Certains évènements spécifiques à l'immunothérapie, dits immuno-médiés peuvent être reportés de façon hétérogène entre les études (Xie et al., 2021). D'une part, ils n'ont pas de terme dédié et d'autre part établir le caractère immuno-médié de façon objective n'est pas toujours évident.

Deux ir-AEs fréquents sous immunothérapie sont les diarrhées et les colites immuno-médiées. La CTCAE ne prévoit pas de distinction entre les diarrhées immuno-médiée et infectieuses. De plus, bien qu'il existe deux termes dans la CTCAE pour grader la diarrhée et la colite, les critères présentent un fort chevauchement ce qui fait qu'en pratique les termes peuvent être interchangés (Abu-Sbeih et al., 2020).

Dans de nombreux essais cliniques évaluant l'immunothérapie, la toxicité hépatique a été gradée en se basant sur des anomalies des indicateurs biochimiques hépatiques sériques : ALAT, ASAT, GGT, ALP, biliubine en utilisant la catégorisation proposée par la CTCAE. Le grade d'hépatotoxicité est alors le plus élevé parmi ces anormalités (Suzman et al., 2018). Notons cependant qu'une augmentation de grade 3-4 des ALAT, ASAT n'est pas une preuve directe d'une perte de la fonction hépatique. Souvent asymptomatique et généralement sans évolution notable évaluée par imagerie (Suzman et al., 2018), l'évaluation de la toxicité hépatique associée à l'immunothérapie à partir de la CTCAE est donc difficile à caractériser.

La thyroïdite est également un ir-AE fréquent. Cette dernière n'a pas de terme dédié dans la CTCAE. Sous immunothérapie, elle se caractérise souvent par une phase (brève et souvent asymptomatique) d'hyperthyroïdie suivie par de l'hyperthyroïdie. Ces deux évènements existent dans la CTCAE, sont collectés en routine mais caractérisent donc un même évènement sous-jacent. La phase d'hyperthyroïdie étant très courte, elle peut être non diagnostiquée, créant de l'hétérogénéité entre les études. On comprend donc que caractériser le début de la thyroïdite à partir d'un dossier patient peut être sujet à une importante variabilité et poser des difficultés d'inférence en cas de modélisation.

3 Article : Vue d'ensemble des modèles de régression pour l'analyse des effets indésirables

Ces dernières années, plusieurs revues de littérature ont souligné que l'analyse des effets indésirables dans les essais cliniques n'est pas optimale. En effet, le contexte entourant l'analyse des événements indésirables implique souvent un nombre très importants d'événements, un manque de puissance pour trouver des associations, mais aussi un manque de formation spécifique concernant ces données complexes. Dans les essais contrôlés randomisés ou dans les études observationnelles, la comparaison de la survenue d'événements indésirables en fonction d'une covariable d'intérêt (par exemple, le traitement) est une question récurrente dans l'analyse des données relatives à la sécurité des médicaments. Ce travail propose un tour d'horizon des modèles de régression existants pour comparer les événements indésirables et pour discuter du choix du modèle, en fonction des caractéristiques des événements indésirables d'intérêt. De nombreuses dimensions peuvent être pertinentes pour comparer les événements indésirables entre les patients (par exemple, le moment, la récurrence et la gravité). De nouveaux modèles ont été proposés récemment pour en tenir compte dans la modélisation. Pour les traitements chroniques, l'apparition d'événements intercurrents au cours du suivi du patient nécessite généralement l'adaptation de l'approche de modélisation (au moins en ce qui concerne leur interprétation). En outre, l'analyse basée sur des modèles de régression ne doit pas se limiter à l'estimation des effets relatifs. En effet, les risques absolus issus du modèle doivent être présentés systématiquement pour en faciliter l'interprétation, valider le modèle et favoriser la comparaison des études.

Drug Safety (2024) 47:205–216
https://doi.org/10.1007/s40264-023-01380-7

REVIEW ARTICLE

An Overview of Regression Models for Adverse Events Analysis

Elsa Coz[1,2,3,4] · Mathieu Fauvernier[1,2,3,4] · Delphine Maucort-Boulch[1,2,3,4]

Accepted: 2 November 2023 / Published online: 25 November 2023

Abstract

Over the last few years, several review articles described the adverse events analysis as sub-optimal in clinical trials. Indeed, the context surrounding adverse events analyses often imply an overwhelming number of events, a lack of power to find associations, but also a lack of specific training regarding those complex data. In randomized controlled trials or in observational studies, comparing the occurrence of adverse events according to a covariable of interest (e.g., treatment) is a recurrent question in the analysis of drug safety data, and adjusting other important factors is often relevant. This article is an overview of the existing regression models that may be considered to compare adverse events and to discuss model choice regarding the characteristics of the adverse events of interest. Many dimensions may be relevant to compare the adverse events between patients, (e.g., timing, recurrence, and severity). Recent efforts have been made to cover all of them. For chronic treatments, the occurrence of intercurrent events during the patient follow-up usually needs the modeling approach to be adapted (at least with regard to their interpretation). Moreover, analysis based on regression models should not be limited to the estimation of relative effects. Indeed, absolute risks stemming from the model should be presented systematically to help the interpretation, to validate the model, and to encourage comparison of studies.

Key Points

For the comparison of adverse events between patients, regression models adjusting important covariables may be considered both in observational and randomized controlled trials.

Time-to-event models have been advocated in adverse events analysis. However, models like logistic regression (with rare event corrections) may be considered for rare events.

If possible, the absolute risk from the regression model should be presented systematically because it may help validation and interpretation, particularly in the competing risk settings, and comparison between studies with different follow-up times.

Due to the multiple facets of adverse events data, several risk measures may be relevant to accurately evaluate and compare patients' toxicity profiles.

✉ Mathieu Fauvernier
mathieu.fauvernier@chu-lyon.fr

1 Université de Lyon, 69000 Lyon, France

2 Université Lyon 1, 69100 Villeurbanne, France

3 Hospices Civils de Lyon, Pôle Santé Publique, Service de Biostatistique et Bioinformatique, 69003 Lyon, France

4 CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, 69100 Villeurbanne, France

1 Introduction

Over the past few years, there have been several studies undertaken to depict the situation with regard to the collection, reporting and analysis of drug safety data in randomized controlled trials (RCTs). In a review, the practice regarding the collection, reporting and the analysis of adverse events (AEs) were described as inconsistent and sub-optimal [1]. The CONSORT harm extensions [2, 3] provided guidelines to cover the reporting of harms but have not been sufficiently adopted [1]. With regard to the statistical part of AE analysis, guidelines hardly tackle the question, but focus mainly on collection and reporting [4]. A scoping review of the statistical methods identified several studies designed specifically for AE data, but those methods have rarely been applied [5]. A recent survey conducted by statisticians from academia and industry to understand the current practice [6] identified several barriers that limit the application of those methods. Beyond the barrier of the design and characteristics of the RCT (e.g., the overwhelming number of events and the underpowered sample size for harm outcomes), many participants indicated a lack of guidelines, of awareness of appropriate methods, and of training on the subject.

For treatment comparison, even for randomized designs, it may be interesting to consider multivariate regression models to adjust on important covariables to improve precision and reduce sample size [7–9]. However, they are barely used in practice [10]. In pharmacoepidemiology, i.e., the study of the risks and therapeutic effects of drugs in real-life populations [11], regression modeling is even more relevant, as those studies rely on observational data with no control for confounders, so adjusted analysis may be necessary to reduce biases. However, due to the multidimensional nature of AEs (e.g., timing, multiplicity, severity) modeling is challenging [12]. For chronic treatments or in oncology, the occurrence of intercurrent events (e.g., death, treatment interruption, treatment discontinuation) is unavoidable even in controlled designs and must be considered during the choice of the model and of the risk measures [13].

Over the last few years, several authors advocated the use of competing risk methods to avoid bias related to the death in time-to-event outcomes of AE analysis [14,]. Nonparametric estimation of risk measure was discussed in the context of pharmacoepidemiology, but without considering competing events, and more recently in the context of RCTs [15]. Models for neglected dimensions were proposed like recurrence [16] or severity [17]. To our knowledge, there is no specific overview of which regression model to choose depending on the research question and the data structure. The purpose of this article is to (1) provide an overview of regression models available to compare adverse events between patients, (2) to discuss and give some clues regarding the choice of the model according to the characteristic of the event, the clinical framework, and the objective.

2 Models

2.1 Towards the Model

The definition of the event(s) and its (their) specificities highly affect the modeling choices. Figure 1 gives a general overview of the characteristics that may considered when identifying a suitable model. The AE might occur early

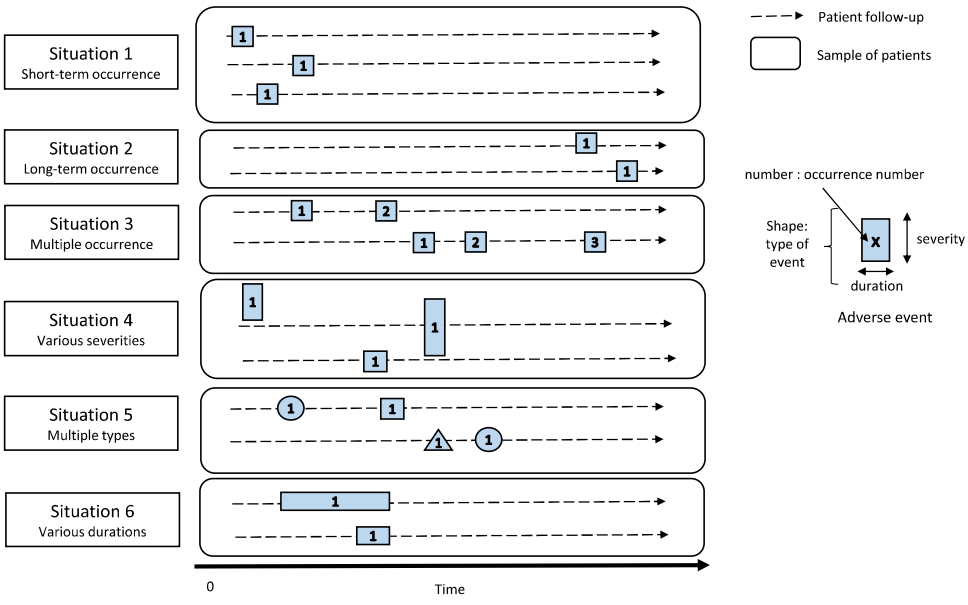

Fig. 1 Event characteristics driving the modeling choices

or late (situation 1 or 2) after treatment initiation, it might occur more than once (situation 3) or have various severities (situation 4). Furthermore, different AE types may be modelled together to catch a common effect (situation 5) (e.g., different AEs from a same body system [18]). Finally, when available, AE duration may be of interest because long durations may highly affect patients' quality of life (situation 6).

Additional to the event characteristics, the therapeutic pathway of the patients may influence the characteristics of the model. In a chronic disease such as cancer, a patient's follow-up may last several months or years so that during the treatment, many incidents may happen, referred as intercurrent events in the *Addendum on estimands and sensitivity analysis in clinical trials* of the ICH Guidelines [13, 19]. We will mention some of them that usually require the model to be adapted. Figure 2 depicts common successions of incidents as they may be seen during a patient's follow-up.

Patient A and B belong to the classical setting of survival analysis, the event of interest (e.g., the AE) occurring at different times and potentially censored (patient B). A first type of intercurrent event is death (Patient C), which obviously prevents the AEs from occurring. When interested in comparing the occurrences at a given time for all patients, the model usually needs to account for a terminating event. Targeting the "direct" effect of the covariable on the occurrence of AEs (i.e., not mediated by the competing event) is much more complicated, as it requires the hypothetical scenario in which the intercurrent event would not occur. This is not achievable, unlike sometimes making unrealistic/untestable assumptions [20, 21] (e.g., independence between the competing events and the events of interest). We will not address those methods in the article. Patient D discontinues the treatment, which usually reduces or even removes the risk of AEs. Hence, this phenomenon counts as an informative censoring event. Contrary to death, the outcome may not be defined because AE collection may be reduced in frequency or even discontinued as in patient D bis (e.g., if the patient begins another treatment) [22]. Treatment interruption for a given period (Patient E), co-medication (Patient F), or dosage change (Patient G) modify the risk of exposure to AEs and raise important methodological issues, not to mention that collecting information about co-medication from the patients is very difficult and when present may have thousands of modalities.

2.2 Single Event Models

The usual practice in AE analysis is to focus on one occurrence of an event (e.g., the first, or the most serious). This section details regression models for a single event outcome. For a formal definition of the outcome and associate risk measure in each case, the reader may refer to Table 1.

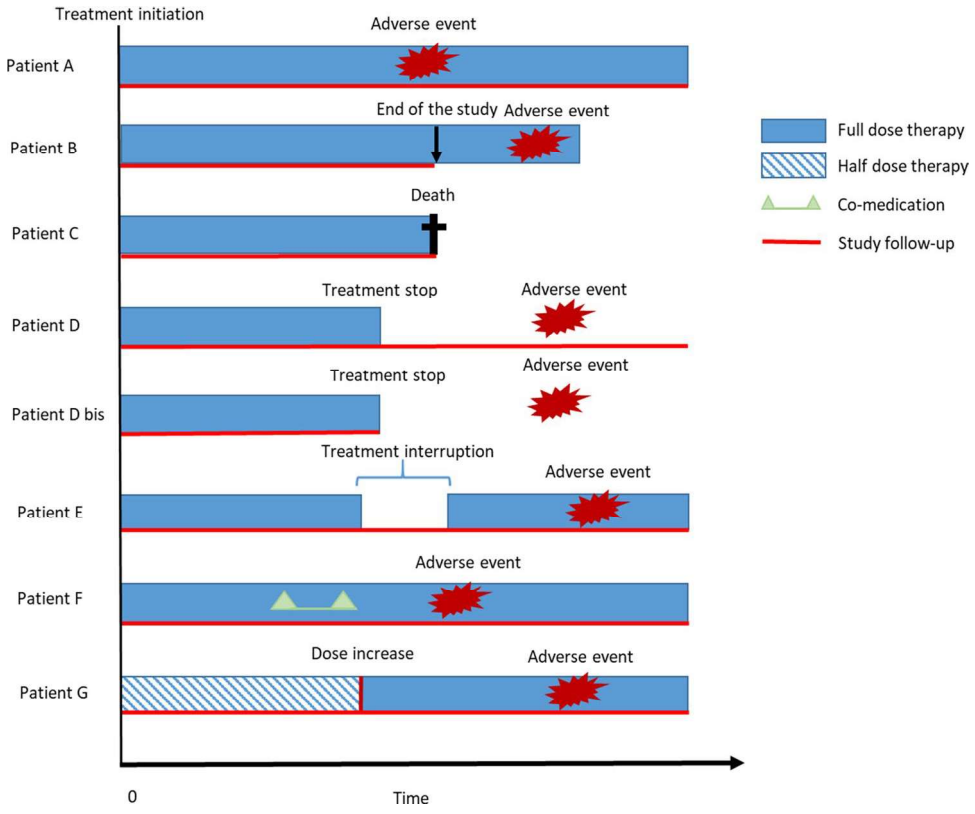

Fig. 2 Typical patients' therapeutic pathways

Table 1 Summary of the models and risk measures of the single-event models section

Single-event models						
Observation		Model	Absolute risk measures			Associated relative risk measure
			Indicator	Definition	In case of a terminating event occurring at time $\boldsymbol{D}$	
Event occurrence indicator	$\boldsymbol{Y}$	Logistic (with/without rare event correction)	Probability	$P(Y = 1)$	Unchanged	Risk ratio[a] OR Probability difference
Censoring indicator, time to event	$\boldsymbol{C}, \boldsymbol{T}$	Time-to-event models (e.g., Cox model)	Hazard $\lambda(t)$	$\lim_{\mathrm{d}t \to 0} \frac{P(t<T\leq t+\mathrm{d}t \mid T>t)}{\mathrm{d}t}$	Use cause-specific hazards: $\lim_{dt \to 0} \frac{P(t<T\leq t+\mathrm{d}t \mid T\geq t, D\geq t)}{\mathrm{d}t}$	HR
			Cumulative incidence function $F(t)$	$P(T < t)$ Directly related to the hazard	- Use cause-specific hazard models of each cause of failure to compute the quantity - Fine-Gray model to directly link the covariables to $F(t)$ [b]	Probability difference
One measure from an ordered severity scale	$\boldsymbol{G}$ where $G = 0$ means no event	Ordinal logistic regression	Probabilities of severity G	$P(G = g)$	Unchanged	OR CR
Ordered severity scale measures evaluated at several times	$\boldsymbol{G_1}, \boldsymbol{G_2}, \ldots, \boldsymbol{G_K}$	Mixed model	Mean grade	$\mathbb{E}(G(t))$	Joint modeling with a time-to-event model on the terminal event may be considered to avoid bias on the covariable effects	Mean grade difference
		Longitudinal ordinal logistic regression	Probabilities of severity G at time t	$P(G(t) = g)$		OR CR
Time of occurrence, censoring indicator and ordinal severity scale	$\boldsymbol{T}, \boldsymbol{C}, \mathbf{G}$	Berridge and Whitehead	Cumulative incidence function of each severity category	$P(T < t, G = g)$		CR
			Hazard $\lambda(t)$	Cf. time-to-event models		HR

CR continuation ratio, *HR* hazard ratio, *OR* odds ratio

[a]AKA relative risk, risk ratio

[b]The Fine-Gray model is actually based on sub-distributional hazards defined as: $1 - F(t) = exp(-\int_0^t \widetilde{h}(s)\mathrm{d}s)$

2.2.1 Logistic Regression

Logistic regression is the most obvious way of relating covariables with a single event. In AE analyses, it may not always be appropriate. Regarding the characteristics we described above, we identified the following situations to be appropriate for this model.

- *Early occurrence* Estimates stemming from the logistic regression (e.g., probabilities or odds ratios [OR]) may be highly biased in presence of censoring (loss to follow-up, administrative censoring) so early AEs may be less sensitive to this issue (situation 1 in Fig. 1) [23]. Moreover, for those events, the times of occurrence may be quite homogenous between patients so that modeling their timing may have limited interest.
- *Rare events* When considering rare events, the logistic regression may be a modeling option. As the event probabilities may be underestimated in those setting [24], it can be combined with useful type of penalization (e.g., Firth's correction) to reduce biases [25]. Variants of Firth's corrections have also been developed to improve the estimates of ORs [26]; they may be particularly useful for analyses of rare but serious AEs [27, 28] with case-control designs [22].
- *Case-control designs* The ORs directly stemming from the logistic regression are known to be interesting for risk evaluation in the case-control study designs, in which the baseline probability is not available.

Odds ratios are often interpreted as probability ratios but this should be done with caution when the probability of an event occurrence is greater than 0.1 [29]. Indeed, in that case, ORs tend to differ from probability ratios and may therefore lead to overestimation of the association between the event and the risk factor. Estimating the risk ratio or risk difference with non-rare outcomes may avoid misleading interpretations, although they require more complex methods (e.g., binomial model) [30].

Handling terminal intercurrent events In case of competing events, all those risk measures stemming from the logistic regression provide the total effect of the covariables on the AE occurrence, meaning a combination of both the direct effect on the AE occurrence and the effect mediated by the competing events [31]. That is why indicators based on probabilities are sometimes criticized for not providing information regarding potential differences in follow-up durations between patients [32]. A classic example is the comparison between treatments with the same hazards of AEs in case one of them increases survival or progression-free survival. Considering the probability of event, the conclusion is that the risk of toxicity is larger in the group with increased survival; hence, the explanation is that patients who die rapidly do not have time to experience AEs in the other treatment arm [19].

2.2.2 Time-to-Event Models

To deal with censoring and to describe the occurrence of a single type of AE over time, time-to-event models have been advocated to improve the analysis of safety data [4, 19, 33]. The proportional hazards model (or Cox model) is the most commonly used regression time-to-event model. The following situations (non-exhaustive list) are well suited for time-to-event models:

- *Long-term toxicity analysis with censoring* As long-term cohort analyses are usually affected by censoring (e.g., administrative), so time-to-event models are the only ones to guarantee unbiased estimates of the probability of events in that situation.
- *Most non-recurrent AEs* If the event of interest is not recurrent (e.g., serious events that may lead to treatment discontinuation), building a time-to-event model is usually valuable because in addition to the treatment effect, it provides an interesting description of the risk of toxicity over time (e.g., cumulative hazard or survival in the Cox model). The proportionality assumption may be questionable and systematically checked, particularly when comparing treatments with very different toxicity profiles (e.g., immunotherapy vs chemotherapy). If the proportionality is not valid, flexible models allowing for time-varying effects may be useful [34].
- *Special interest in time-varying covariables such as drug exposure* Another valuable aspect of time-to-event models is their ability to consider important time-dependent covariables, such as exposure. In 2019, Danieli and Abrahamowicz [34] tackled the modeling of drug exposure and treatment interruptions in a time-to-event model to be used in observational studies. They considered a weighted cumulative exposure to deal with both the time elapsed since the last exposure and the cumulative dose. Flexible estimates described the way past exposures affect the hazard of a specific event. This approach is interesting in case of late AEs, potentially triggered by drug accumulation (Fig. 1 situation 2).

Hazard ratios (see Table 1) measure the association between the event and the covariables, although we generally cannot use them to establish causality [35], even with a correctly Cox specified model and a proper randomization at baseline. Indeed, by conditioning on survival, the risk set may be modified over time if the covariable of interest has an actual treatment effect or in presence of an unmeasured covariable [36]. Cumulative probabilities difference or ratio may be considered instead for causal inference. In

both cases, reporting absolute risk measures (e.g., hazards or cumulative hazards, cumulative probabilities) is highly advised [37].

Handling terminal intercurrent events As previously mentioned, competing risks are often encountered in AE analysis, particularly in oncology [14] complicating time-to-event analyses and their interpretation [38]. Two main approaches may be used for that type of analysis. The first is to model cause-specific hazards based on the times of occurrence for a single cause of failure (see Table 1). Cox model can be used to perform a regression on the cause-specific hazards. A distinct model for each cause of failure is needed to compute the cause-specific cumulative incidence function. The second approach relies on sub-distributional hazards whose famous associated regression model is the Fine-Gray model [39]. Despite its unclear biological interpretation [40], the strength of the model is to directly link the variables to the cumulative incidence function in a single model [41]. In both cases, covariable effects measures (e.g., difference or ratio) derived from the cumulative probabilities provide the same total effect we previously described for logistic regression [31]. Dealing with the direct effect of the treatment cannot definitely be solved using the cause-specific or the sub-distributional hazards, which have generally no causal interpretation, even in randomized designs [31]. Causation in the competing risks setting is constituting an active fields of statistical research and other estimands may be considered (e.g., survivor average causal effect [SACE]) [21].

2.2.3 Models Comparing Severities

The severity of AEs is usually quoted using an ordered categorical or a numeric scale. The well-known National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) uses a grade from 1 (mild) to 5 (death). Then, we may want to compare the impact of a covariable on the severity of an event (situation 4 of Fig. 1). For example, one may be interested in the level of toxicity of a treatment regarding various dose schemes [42] or by the risk factors of severity [43]. A natural option is to conduct an ordinal logistic regression (e.g., proportional odds model or continuation ratio model) [44]. The idea of those models is to associate the covariables with the probabilities of the levels of severities. However, as with the classical logistic regression, the ordinal logistic regression does not account for the timing of the AEs and it may be highly biased in the presence of censoring limiting the situations of application of those models. Hence, other modeling options may be considered according to the aim of the study.

– *The interest is in the evolution of the severity of the adverse events in patients over time* In this case, repeated measurements of the level of severity are collected and longitudinal models may be considered. A first option is the ordinal logistic regression with a random effect on the patient's level. For example, Augustin et al [45] built a longitudinal proportional odds model on an oral mucositis score, with the cumulative dose and the mouth sites as covariables, to improve the planning of radiation therapy. The second option is to consider the grade (e.g., CTCAE grades) as a repeated measure over time using a linear mixed model [46–48]. The latter approach captures the complete toxicity trajectory of the patients, including the burden of low-grade AEs. It may provide a comprehensive, visual description of the toxicity profile despite the uncertain assumption of grade normal distribution.

Handling terminal intercurrent events When the longitudinal outcome and the occurrence of a terminal event (i.e., informative dropout) are correlated, the linear mixed model will lead to biased estimates. To reduce the bias, a joint model may be considered [49], i.e., a mixed model for the longitudinal data and a survival model for the time to death. Shared parameters like random effects link the two outcomes.

– *The interest is in comparing the cumulative probabilities of occurrence for each level of severity over time (potentially the maximum grade per patient)* To handle this issue, Berridge and Whitehead [47] built a two-component model to estimate the occurrence probabilities of AEs according to their severities. One component is a proportional hazard model, used to estimate the all-grade probability of events over time in an unbiased manner, and the second component is an ordinal logistic model which ventilates the probability over the levels of severity.

2.3 Recurrent Events Models

Harm studies tend to focus on severe life-threatening AEs (e.g., CTCAE grade 3–4), which are recommended to be systematically collected [4, 50]. As those events may lead to treatment discontinuation, single-event models seem rather suited to them. However, it became increasingly common to include patient reported outcomes (PRO) (e.g., PRO-CTCAE [51]) in clinical trials to complete AE collection because physicians tend to under-report symptoms in terms of frequency and/or severity, compared to patients themselves [52]. Those events may be mild but recurrent (situation 3 in Fig. 1) and may highly affect the quality of life. However, the statistical methods used to deal with these recurrent events have been repeatedly reported as inappropriate (e.g., restricting the analysis to the first or the worst-grade event) in clinical trials [1, 12, 16]. In this section, we discuss the use of some common recurrent events methods

in the context of AEs. The reader may refer to Table 2, for a summary of the models and formal definitions of the outcomes and risk measures.

Previously, in the time-to-event model section, the hazard was defined as the instantaneous probability by unit of time of experiencing an AE in patients who had not experienced an AE previously. Here, the counting process theory is used to extend the notion of hazard to recurrent events. The intensity of AEs is defined as the instantaneous probability by unit of time of experiencing an AE given the history of the process (e.g., the timing of the previous events) [53]. Additionally, 'rate' will stand for the instantaneous probability of experiencing an AE. We identified the two following clinical questions that may guide the choice of the recurrent event model.

- *Interest in finding associations between the covariables and the overall occurrence of AEs over time*: The most natural modeling approach is the Poisson and the Negative Binomial model [10]. However, the Andersen-Gill (AG) model [54] has been found to perform better in complex situations and should be preferred (unless the available data are aggregated counts) [55]. The latter is marginal (e.g., based on quantities such as the rate or the cumulative mean) semi-parametric regression on the rate (extension of the Cox model). The AG model may be easily applied with standard statistical software handling the Cox model, by simply rearranging the dataset [53, 56]. A robust estimator of the variance is usually needed to account for correlation between the events of a same subject [57]. As marginal quantities, covariables effect on the rate or on the cumulative mean number of events do have a causal interpretation (no selection of the population over time). Hence the rate is particularly interesting in randomized designs to compare treatment effects [58]. The cumulative mean number of events, that are easier to understand, may be obtained from the rate function.

  Handling terminal intercurrent events Terminal events may be managed similarly to the single event hazards by considering the modeling of the rate of AEs at a given time, conditioned on the survival from the terminal event at the same time (see Table 2) [53]. Due to this conditioning, the rate-based covariable effect estimates no longer have a causal interpretation [58]. Furthermore, the rate is no longer directly related to the cumulative mean function (it may also depend on the terminating event rate).

- *Interest is to compare the patient's individual risks of AEs given their number of previous events* The Prentice-Williams-Petersen (PWP) [59] model may be considered in that situation. It is another semi-parametric proportional model based on intensities and the history of the process being merely summarized by the number of previous events. Hence, it can be seen as a multi-state model. Two formulations of the model are possible depending on the knowledge of the process. The first formulation is defined by the time of the events. The covariable effect on the occurrence of AEs is evaluated over the entire observation period and may be allowed to vary according to the number of previous events. In particular, the event occurrence does not depend on the delay since the previous event. The second formulation relies on inter-times or gap times, i.e., the delay between two subsequent events. In that formulation, the occurrence of an event does not depend on the delay since the beginning of the follow-up (e.g., the beginning of the treatment). Like the AG model, the coefficient estimates may be obtained from a standard statistical software handling the Cox model, by rearranging the dataset and stratifying the number of previous events [53, 60]. As for the hazard, the covariable effects in PWP models do not have a causal interpretation (selection of the population) despite randomization at baseline [61]. Deriving the cumulative mean number of AEs from intensity models can be very complex, as well as can its interpretation [53].

  Handling terminal intercurrent events The terminal event (e.g., death) may be related to the recurrent process (e.g., serious events). In that case, the joint modeling of the hazard of the terminating event and of the intensity of the process may be necessary (e.g., joint frailty model) [53, 62].

  Comparing severities in recurrent event models The model of Berridge and Whitehead discussed in the single-event model section was extended to deal with recurrent events by replacing the proportional hazard model by a recurrent model such as PWP [17].

2.4 Multi-type Event Models

As discussed in section 2, patients may experience multiple types of AEs (situation 5 in Fig. 1). It is always possible to model them independently but this would result in a loss of power. Modeling AEs jointly seems more attractive. The Wei Lin and Weissfeld (WLW) model [63] is a Cox model extension that is able to handle several types of events (by stratifying the type). It provides a common measure of the variable effect across all types of AEs considered. This comes with an important gain in power but at the cost of the strong hypothesis that covariable's effects are the same whatever the type of AEs.

Comparing severities in recurrent event models An extension of the Berridge and Whitehead model was proposed [58] in which the two components are respectively replaced

Table 2 Summary of the models and risk measures of the recurrent event models and multi-type events sections

Observation		Model	Absolute risk measures			Associated relative risk measure
			Indicator	Definition	In case of a terminating event occurring at time $\boldsymbol{D}$	
Recurrent events models						
Times to events, censoring time	$\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_k, \boldsymbol{C}$	Andersen-Gill + robust estimator of variance	Rate $\rho(t)$	$\frac{\mathrm{d}\mu(t)}{\mathrm{d}t} = \lim_{\mathrm{d}t \to 0} \frac{P(\mathrm{d}N(t)=1)}{\mathrm{d}t}$ [a]	Rather consider: $\lim_{\mathrm{d}t \to 0} \frac{P(dN(t)=1 \mid D \geq t)}{\mathrm{d}t}$ [b]	Rate ratio
			Cumulative mean function $\mu(t)$	$\mathbb{E}(N(t))$[a] Directly linked to the rate	An additional cause-specific hazard model of the terminating event is required to derive $\mu(t)$	Mean difference
		Prentice-Williams-Petersen (PWP)	Intensities $\lambda_k(t)$[b]	$\lim_{\mathrm{d}t \to 0} \frac{P\left(t < T_k \leq t + \mathrm{d}t \mid N(t) = k-1\right)}{\mathrm{d}t}$	The joint modeling of the hazard and intensity may be needed	Intensity ratio
Inter-events times, censoring time	$\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_k, \boldsymbol{C}$	PWP gap times formulation (PWP gap)	Intensities $\lambda_k(t)$[b]	$\lim_{dt \to 0} \frac{P\left(t < U_k \leq t + \mathrm{d}t \mid N(t) = k-1\right)}{\mathrm{d}t}$		
Times to events, ordered severity scale measures	$\left(\boldsymbol{T}_1, \boldsymbol{G}_1\right), \left(\boldsymbol{T}_2, \boldsymbol{G}_2\right), \ldots \left(\boldsymbol{T}_k, \boldsymbol{G}_k\right), \boldsymbol{C}$	Gebski et al: Ordinal logistic regression + PWP or PWP gap	Cumulative probabilities over time of severity levels	$P\left(T_k < t, G_k = g\right)$		CR
			Intensity/hazard	Cf. related recurrent event component		
Multiple-events models						
Times of each individual type of event, censoring indicator	$\boldsymbol{T}_1, \boldsymbol{T}_2 \ldots, \boldsymbol{T}_m, \boldsymbol{C}$	Wei-Lin-Weissfeld (WLW)	Hazards	$\lim_{\mathrm{d}t \to 0} \frac{P\left(t < T_k \leq t + \mathrm{d}t \mid T_k > t\right)}{\mathrm{d}t}$		HR

CR continuation ratio, *HR* hazard ratio, *OR* odds ratio

[a]We denote the counting process $N(t)$, associated with the recurrent times and the history of the process $H(t)$

[b]Actually, PWP is a special case of the intensity function, the history of the process being summarized by $N(t)$, whose general definition may be written: $\lim_{\mathrm{d}t \to 0} \frac{P(\mathrm{d}N(t)=1 \mid H(t))}{\mathrm{d}t}$

by a multinomial logistic regression and a recurrent event model (PWP or WLW). One difficulty here is to carefully define mutually exclusive categories of AEs.

3 Discussion

In this article, we discussed various regression models for AE outcomes. We considered various dimensions of those events, including severity. Models dealing with severity are more complex (two-component models) and most have been proposed recently. More practical examples to facilitate their interpretation, as well as the implementation of software, would be useful so that they can be used routinely. In most cases, models should be adapted in the presence of a terminal event like death (at least with regard to their interpretation).

Although we extensively searched the literature to illustrate the methods of this overview, this article does not claim to be representative of the actual practice nor to be an exhaustive list of the methods used in that context. We first identified methodological articles dealing with the statistical issues in AE analysis in Pubmed and Google Scholar using keywords: "toxicities", "adverse events" or "drug safety" with "statistical analysis". We did not exclude any period of publication. From those articles, we built a list of regression models. To enrich the discussion, we managed to identify articles that apply the models in practice with drug safety data by searching the name of the model with the keywords "toxicities", "adverse events" or "drug safety" in PubMed and Google Scholar as previously. Hence, this article provides some methodological tools that may suit common situations and answer some clinical questions. Most of the references we provided dealt with the comparison of AEs between treatment arms in RCTs but their usage may be extended to observational studies, like the Qualitop project [64], which motivated this article, and various covariables of interest.

Often, risks measures are used to "map the AE data to a single value" [19] for the purpose of safety evaluation. However, unlike efficacy, AE comparisons may not rely on a single value due to the complex dimensionality of those data. For example, providing both absolute and relative risk measures is commonly advised [37]. For non-parametric estimation, incidence rate is often advised compared to the overall probabilities to account for studies with various follow-up durations (e.g., due to different durations of two treatment arms) [19]. However, by considering the overall cumulative incidence function over time instead of an overall probability, quantities are more comparable. Graphical representations of the absolute risk stemming from the regression model should be done systematically as it may help to validate the adequacy of the model (e.g., comparison with non-parametric estimates) and to interpret the model, particularly in the presence of competing risks. Moreover, it should facilitate further meta-analyses, mixing studies with various follow-up durations.

All throughout the article, we considered the outcome (AEs) of the models to be clearly defined. However, the number of AEs collected may be huge. For example, the CTCAE has narrowed the keyword field used to describe AEs but its version 4.0 still includes more than 1000 terms. Hence, the analyses and comparisons have to then focus on a small number of events of interest whereas the criteria for their selection are often unclear, ill documented or based on arbitrary rules (e.g., frequencies $\geq 5\%$). Some authors considered grouping AEs according to the body systems [65] but assigning types of AEs into body systems is not always as easy as it may seem and the grouping choices may highly influence the conclusions of the study [18]. Selecting the AEs according to their attributability to the treatment is a more difficult task. In their 2016 recommendations, Lineberry et al did not insist on this kind of selection because of its inherent subjectivity and limited value in clinical trials [4].

One common limitation of all the models we discussed is the reliability of data collection. If mild AEs may be of interest regarding the quality of life of the patients, they are often under-reported by clinicians [52]. Therefore, using Patient Reported Outcomes (PRO) may be more relevant in that situation. Furthermore, we discussed models using severity that may be difficult to collect reliably over the whole follow-up, particularly in observational studies. For severe events, the patient is most likely to come for a consultation or may be hospitalized, so the collection of the AE is close to the time of occurrence. Otherwise, AE collection is usually performed when patient meets clinician for a follow-up visit (e.g., once a month). The time of occurrence is therefore not precisely collected, which leads to interval censoring and modeling issues. Moreover, some AEs may arise and be resolved in between visits (e.g., transient hyperthyroidism during immunotherapy), so they may not be collected (truncation). Hence, comparing the occurrence of such AEs in patients with various visit frequencies may be misleading.

4 Conclusion

Comparing the adverse events between groups of patients is a recurrent occurrence with drug safety data. Regression models adjusting on important covariables may be considered in both observational and RCTs. Time-to-event models are advocated for AE analysis; however, the interpretation of those models are complicated because of competition with death or treatment discontinuation. Hence, the absolute risk from the regression model should be presented

systematically because it may help validation and interpretation, particularly in the competing risk settings, and comparison between studies with different follow-up times. Flexible time-to-event models dealing with baseline risks (unlike semi-parametric models) as well as non-linear and time-dependent covariate effects have proven to be useful and should be explored further in this context. Rare events are a recurrent issue in drug safety data and few models may suit rare outcomes. Hence, the logistic regression (with rare event corrections) may be a useful option. Recent articles proposed models accounting for severity; however, their interpretation may be difficult and real-life application should be performed to extend their use.

Acknowledgments We thank Jean Iwaz for his careful lecture and corrections. We also thanks the three anonymous reviewers whose comments significantly improve the manuscript.

Declarations

Funding This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement N° 875171. The findings and conclusions in this report are those of the authors and do not necessarily represent the view of the consortium.

Conflict of Interest The authors have no relevant financial or non-financial interests to disclose.

Code Availability Not applicable.

Ethics Approval Not applicable as the study does not involve human participants.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Availability of Data And Materials Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Author Contributions EC completed the literature review, manuscript figures, data summary, and manuscript revisions with input from MF and DMB. All authors were involved in critical revision and approval of the final manuscript drafts.

References

1. Phillips R, Hazell L, Sauzet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open. 2019;9(2): e024537. https://doi.org/10.1136/bmjopen-2018-024537.
2. Ioannidis JPA, Evans SJW, Gøtzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med. 2004;141(10):781–8. https://doi.org/10.7326/0003-4819-141-10-200411160-00009.
3. Junqueira DR, Zorzela L, Golder S, et al. CONSORT Harms 2022 statement, explanation, and elaboration: updated guideline for the reporting of harms in randomised trials. BMJ. 2023;381: e073725. https://doi.org/10.1136/bmj-2022-073725.
4. Lineberry N, Berlin JA, Mansi B, et al. Recommendations to improve adverse event reporting in clinical trial publications: a joint pharmaceutical industry/journal editor perspective. BMJ. 2016;355: i5078. https://doi.org/10.1136/bmj.i5078.
5. Phillips R, Sauzet O, Cornelius V. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. BMC Med Res Methodol. 2020;20(1):288. https://doi.org/10.1186/s12874-020-01167-9.
6. Phillips R, Cornelius V. Understanding current practice, identifying barriers and exploring priorities for adverse event analysis in randomised controlled trials: an online, cross-sectional survey of statisticians from academia and industry. BMJ Open. 2020;10(6): e036875. https://doi.org/10.1136/bmjopen-2020-036875.
7. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Stat Med. 2008;27(23):4658–77. https://doi.org/10.1002/sim.3113.
8. FDA C for DE and. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. Published May 25, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products. Accessed September 28, 2023.
9. Ishii J, Ohshimo S, Shime N. Potential Confounders for Applying a Novel Sepsis Care Quality Improvement Program. Crit Care Med. 2020;48(2):e161–2. https://doi.org/10.1097/CCM.0000000000004069.
10. Patson N, Mukaka M, Otwombe KN, et al. Systematic review of statistical methods for safety data in malaria chemoprevention in pregnancy trials. Malar J. 2020;19(1):119. https://doi.org/10.1186/s12936-020-03190-z.
11. Quartey G, Wang J, Kim J. A review of risk measures in pharmacoepidemiology with tips for statisticians in the pharmaceutical industry. Pharm Stat. 2011;10(6):548–53. https://doi.org/10.1002/pst.521.
12. Cabarrou B, Gomez-Roca C, Viala M, et al. Modernizing adverse events analysis in oncology clinical trials using alternative approaches: rationale and design of the MOTIVATE trial. Invest New Drugs. 2020;38(6):1879–87. https://doi.org/10.1007/s10637-020-00938-x.
13. FDA. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Published 2019. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf. Accessed 9 Oct 2023.
14. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. Pharm Stat. 2016;15(4):297–305. https://doi.org/10.1002/pst.1739.
15. Stegherr R, Schmoor C, Lübbert M, Friede T, Beyersmann J. Estimating and comparing adverse event probabilities in the presence

of varying follow-up times and competing events. Pharm Stat. 2021. https://doi.org/10.1002/pst.2130.
16. Hengelbrock J, Gillhaus J, Kloss S, Leverkus F. Safety data from randomized controlled trials: applying models for recurrent events. Pharm Stat. 2016;15(4):315–23. https://doi.org/10.1002/pst.1757.
17. Gebski V, Byth K, Asher R, Marschner I. Recurrent time-to-event models with ordinal outcomes. Pharm Stat. 2021;20(1):77–92. https://doi.org/10.1002/pst.2057.
18. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. Biometrics. 2004;60(2):418–26. https://doi.org/10.1111/j.0006-341X.2004.00186.x.
19. Unkel S, Amiri M, Benda N, et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. Pharm Stat. 2019;18(2):166–83. https://doi.org/10.1002/pst.1915.
20. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. Proc Natl Acad Sci U S A. 1975;72(1):20–2.
21. Stensrud MJ, Young JG, Didelez V, Robins JM, Hernán MA. Separable effects for causal inference in the presence of competing events. J Am Stat Assoc. 2022;117(537):175–83. https://doi.org/10.1080/01621459.2020.1765783.
22. Singh S, Loke YK. Drug safety assessment in clinical trials: methodological challenges and opportunities. Trials. 2012;13(1):138. https://doi.org/10.1186/1745-6215-13-138.
23. Crowe BJ, Xia HA, Berlin JA, et al. Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team. Clin Trials. 2009;6(5):430–40. https://doi.org/10.1177/1740774509344101.
24. King G, Zeng L. Logistic regression in rare events data. Polit Anal. 2001;9:137–63.
25. Firth D. Bias Reduction of Maximum Likelihood Estimates. Biometrika. 1993;80(1):27–38. https://doi.org/10.2307/2336755.
26. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions? Stat Med. 2017;36(14):2302–17. https://doi.org/10.1002/sim.7273.
27. Southworth H, O'Connell M. Data mining and statistically guided clinical review of adverse event data in clinical trials. J Biopharm Stat. 2009;19(5):803–17. https://doi.org/10.1080/10543400903105232.
28. Scalorbi F, Argiroffi G, Baccini M, et al. Application of FLIC model to predict adverse events onset in neuroendocrine tumors treated with PRRT. Sci Rep. 2021;11(1):19490. https://doi.org/10.1038/s41598-021-99048-8.
29. Katz KA. The (Relative) risks of using odds ratios. Arch Dermatol. 2006;142(6):761–4. https://doi.org/10.1001/archderm.142.6.761.
30. Holmberg M, Andersen L. Estimating risk ratios and risk differences: alternatives to odds ratios. JAMA. 2020;324(11):1098–9. https://doi.org/10.1001/jama.2020.12698.
31. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. Stat Med. 2020;39(8):1199–236. https://doi.org/10.1002/sim.8471.
32. Siddiqui O. Statistical methods to analyze adverse events data of randomized clinical trials. J Biopharm Stat. 2009;19(5):889–99. https://doi.org/10.1080/10543400903105463.
33. O'Neill RT. Statistical analyses of adverse event data from clinical trials: special emphasis on serious events. Drug Inf J. 1987;21(1):9–20. https://doi.org/10.1177/009286158702100104.
34. Danieli C, Abrahamowicz M. Competing risks modeling of cumulative effects of time-varying drug exposures. Stat Methods Med Res. 2019;28(1):248–62. https://doi.org/10.1177/0962280217720947.
35. Hernán MA. The hazards of hazard ratios. Epidemiol Camb Mass. 2010;21(1):13–5. https://doi.org/10.1097/EDE.0b013e3181c1ea43.
36. Martinussen T. Causality and the Cox regression model. Annu Rev Stat Its Appl. 2022;9(1):249–59. https://doi.org/10.1146/annurev-statistics-040320-114441.
37. Zavala S, Stout JE. Understanding and communicating risk: assessing both relative and absolute risk is absolutely necessary. JID Innov Skin Sci Mol Popul Health. 2022;2(2): 100097. https://doi.org/10.1016/j.xjidi.2022.100097.
38. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26(11):2389–430. https://doi.org/10.1002/sim.2712.
39. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94(446):496–509. https://doi.org/10.1080/01621459.1999.10474144.
40. Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. Stat Med. 2012;31(11–12):1074–88. https://doi.org/10.1002/sim.4385.
41. Ozenne B, Sørensen A, Lyngholm ST, Torp-Pedersen C, Alexander GT. riskRegression: predicting the risk of an event using cox regression models. R J. 2017;9(2):440. https://doi.org/10.32614/RJ-2017-062.
42. Doussau A, Thiébaut R, Paoletti X. Dose-finding design using mixed-effect proportional odds model for longitudinal graded toxicity data in phase I oncology clinical trials. Stat Med. 2013;32(30):5430–47. https://doi.org/10.1002/sim.5960.
43. Kulothungan V, Subbiah M, Ramakrishnan R, Raman R. Identifying associated risk factors for severity of diabetic retinopathy from ordinal logistic regression models. Biostat Epidemiol. 2018;2(1):34–46. https://doi.org/10.1080/24709360.2017.1406040.
44. Harrell JFE. Regression modeling strategies. 2nd ed. Springer International Publishing AG; 2015. (**2015 édition**).
45. Augustin NH, Kim SW, Uhlig A, Hanser C, Henke M, Schumacher M. A flexible multivariate random effects proportional odds model with application to adverse effects during radiation therapy. Biom J Biom Z. 2017;59(6):1339–51. https://doi.org/10.1002/bimj.201600142.
46. Thanarajasingam G, Atherton PJ, Novotny PJ, Loprinzi CL, Sloan JA, Grothey A. Longitudinal adverse event assessment in oncology clinical trials: the Toxicity over Time (ToxT) analysis of Alliance trials NCCTG N9741 and 979254. Lancet Oncol. 2016;17(5):663–70. https://doi.org/10.1016/S1470-2045(16)00038-3.
47. Thanarajasingam G, Leonard JP, Witzig TE, et al. Longitudinal Toxicity over Time (ToxT) analysis to evaluate tolerability: a case study of lenalidomide in the CALGB 50401 (Alliance) trial. Lancet Haematol. 2020;7(6):e490–7. https://doi.org/10.1016/S2352-3026(20)30067-3.
48. Wong ML, Gao J, Thanarajasingam G, et al. Expanding beyond maximum grade: chemotherapy toxicity over time by age and performance status in advanced non-small cell lung cancer in CALGB 9730 (Alliance A151729). Oncologist. 2021;26(3):e435–44. https://doi.org/10.1002/onco.13527.
49. Rizopoulos D. Joint models for longitudinal and time-to-event data: with applications in R. CRC Press; 2012.
50. EMA. ICH E19 Guideline - Optimization of Safety Data Collection. Published online 2019.
51. Basch E, Reeve BB, Mitchell SA, et al. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events

(PRO-CTCAE). J Natl Cancer Inst. 2014;106(9):dju244. https://doi.org/10.1093/jnci/dju244.
52. Atkinson TM, Ryan SJ, Bennett AV, et al. The association between clinician-based common terminology criteria for adverse events (CTCAE) and patient-reported outcomes (PRO): a systematic review. Support Care Cancer. 2016;24(8):3669–76. https://doi.org/10.1007/s00520-016-3297-9.
53. Cook RJ, Lawless JF. The statistical analysis of recurrent events. Springer, New York, 2007. p. 82–9, 218–24, 171–77.
54. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat. 1982;10(4):1100–20. https://doi.org/10.1214/aos/1176345976.
55. 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):4989–97. https://doi.org/10.1016/j.csda.2008.04.009.
56. 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. https://doi.org/10.1186/s12874-017-0462-x.
57. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. J R Stat Soc Ser B Stat Methodol. 2000;62(4):711–30. https://doi.org/10.1111/1467-9868.00259.
58. Janvin M, Young JG, Ryalen PC, Stensrud MJ. Causal inference with recurrent and competing events. Lifetime Data Anal. 2023. https://doi.org/10.1007/s10985-023-09594-8.
59. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. Biometrika. 1981;68(2):373–9. https://doi.org/10.1093/biomet/68.2.373.
60. Amorim LD, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44(1):324–33. https://doi.org/10.1093/ije/dyu222.
61. Zhong Y, Cook RJ. The effect of omitted covariates in marginal and partially conditional recurrent event analyses. Lifetime Data Anal. 2019;25(2):280–300. https://doi.org/10.1007/s10985-018-9430-y.
62. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. Biostatistics. 2007;8(4):708–21. https://doi.org/10.1093/biostatistics/kxl043.
63. 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):1065–73.
64. Vinke PC, Combalia M, de Bock GH, et al. Monitoring multidimensional aspects of quality of life after cancer immunotherapy: protocol for the international multicentre, observational QUALITOP cohort study. BMJ Open. 2023;13(4): e069090. https://doi.org/10.1136/bmjopen-2022-069090.
65. Güttner A, Kübler J, Pigeot I. Multivariate time-to-event analysis of multiple adverse events of drugs in integrated analyses. Stat Med. 2007;26(7):1518–31. https://doi.org/10.1002/sim.2637.

4 | Discussion & Conclusion

A travers cette revue de littérature, nous avons identifié de nombreux modèles de régression proposés pour évaluer et comparer des risques de toxicité. De nombreuses dimensions des données de toxicité y sont prises en compte, y compris la temporalité, la sévérité, la récurrence. Dans de nombreuses situation (e.g. les risques compétitifs), l'utilisation de plusieurs indicateurs semble nécessaire pour permettre une description correcte du risque.

Nous avons également vu que les modèles de régression classiquement utilisés dans le cadre de l'analyse des effets indésirables sont principalement semi-paramétriques. Ils ne proposent donc pas de décrire la dynamique de survenue des évènements au cours du temps, qui pourrait être une source d'information complémentaire très utile dans les analyses de données de toxicité. La complexité ajoutée par la présence quasiment systématique de risques compétitifs rend la représentation de la dynamique du taux au cours du temps d'autant plus pertinente. De plus, l'hypothèse de proportionnalité souvent utilisée dans les modèles semi-paramétriques est discutable lorsque l'on compare la toxicité entre des traitements très différents (par exemple, chimiothérapie versus immunothérapie).

Afin de rendre compte de la complexité de la dynamique du taux, i.e. son évolution au cours du temps, l'utilisation des splines présente de nombreux avantages (Danieli and Abrahamowicz, 2019). Plusieurs modèles flexibles du taux basés sur les splines ont été présentés ces dernière années (Remontet et al., 2019), afin de rendre compte de la dynamique du taux mais également des effets potentiellement non-linéaires et non proportionnels des covariables. Toutefois, dans un contexte d'inférence classique basé sur la vraisemblance, l'utilisation de modèles flexibles s'accompagne d'un risque accru de sur-ajustement. Afin de limiter cet écueil, des modèles pénalisés du taux ont été proposés. La partie suivante s'intéresse donc à l'utilisation des ces modèles pénalisés dans le cadre de données de toxicité.

Partie II

Modéliser les effets indésirables avec un modèle de taux flexible

1 Notions théoriques

1.1 Le modèle flexible sur le logarithme du taux

1.1.1 Notations

On s'intéresse au temps jusqu'à la survenue d'un unique évènement en relation avec des variables. Notons T ce temps d'évènement. Comme brièvement évoqué dans la première partie de ce manuscrit, les indicateurs les plus utilisés pour ce type de données sont la fonction de survie et le taux. On définit la survie par :

$$S(t) = P(T > t) \tag{1.1}$$

La survie est donc une fonction décroissante, donnant la proportion d'individus n'ayant pas encore présenté l'évènement à un instant t.

Une autre quantité d'intérêt dans ce contexte est le taux :

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} \tag{1.2}$$

Le taux décrit la probabilité instantanée d'évènement par unité de temps pour un individu, sachant que cet individu ne l'a pas encore présenté. Il est donc simplement défini sur $\mathbb{R}_+$ sans contrainte de monotonie. La survie et le taux sont reliés de la façon suivante :

$$S(t) = \exp\left(- \int_0^t h(s) ds \right) \tag{1.3}$$

Enfin, la densité de T est donnée par la relation :

$$f(t) = h(t) S(t) \tag{1.4}$$

1.1.2 Modèle du logarithme du taux

Le modèle le plus connu est le modèle de Cox à taux proportionnels, défini de la façon suivante, sur un échantillon de n individus :

$$\{\log(h(t, \mathbf{x}_i))\}_{i=1,\dots,n} = \log(h_0(t)) + \mathbf{X}\boldsymbol{\beta} \tag{1.5}$$

h_0 étant une fonction de base laissée non spécifiée, $\mathbf{X}$ la matrice de design des variables des individus de l'échantillon et $\boldsymbol{\beta}$ un vecteur de paramètres à estimer. L'exponentielle des paramètres du modèle s'interprète simplement comme des taux relatifs. L'une des limites de ce modèle, déjà discutée, est l'absence d'estimation du risque absolu (taux de base $h_0(t)$). Cela

peut conduire à des interprétations discutables, notamment dans des contextes plus complexes, tels que les risques compétitifs, que nous aborderons par la suite. Le taux de base peut être modélisé de façon paramétrique à l'aide d'une loi comme celle de Weibull ou de Gompertz, par exemple, mais cela laisse peu de flexibilité à sa dynamique, potentiellement très complexe. Une alternative est donc de modéliser le taux de base par une spline. Au-delà de la modélisation du taux de base, une spline peut également être utilisée pour ajuster des effets non-linéaires de covariables et des interactions. En particulier, on pourra considérer des interactions avec le temps, c'est-à-dire, des effets non-proportionnels des covariables.

Spline

Une fonction spline est une fonction polynômiale par morceaux, que l'on peut donc écrire comme une combinaison linéaire :

$$f(x, \boldsymbol{\beta}) = \sum_{i=0}^{M+D+1} \beta_i B_i(x)$$

- D est le degré des polynômes (i.e. linéaire, quadratique, cubique...),
- M correspond au nombre de noeuds de la spline,
- $B_i(x)$ sont des fonctions de base de l'espace des polynômes,
- β_i sont les coefficients associés à la spline.

Plus le nombre de noeuds de la spline est élevé plus on pourra obtenir une forme complexe mais cela implique une augmentation du risque de sur-ajustement. Le nombre de degrés de liberté de la spline est égal à M + d + 1.

Pour bien comprendre, on peut par exemple définir la base des puissances tronquées dont la construction est facile à appréhender. On considère $x_1, ..., x_M$, des noeuds intérieurs d'un intervalle $[a, b]$. La spline de degré D peut s'écrire :

$$f(x, \beta) = \sum_{d=0}^{D} \beta_d x^d + \sum_{m=1}^{M} \beta_{D+m} I(x \leq x_d)$$

où, I est la fonction indicatrice.

En pratique, des bases de splines plus efficaces lui sont préférées comme la base B-splines ou des splines cubiques de régression (Wood, 2017) que nous utiliserons tout au long de cette thèse. Une spline cubique ou une B-spline peuvent être erratiques aux extrémités des intervalles (définis par les noeuds extérieurs). Il est alors possible d'imposer la nullité des dérivées d'ordre 2 et plus aux noeuds extérieurs. Les splines prennent alors le nom de splines restreintes ou naturelles (Perperoglou et al., 2019).

Il est possible d'écrire le modèle (1.5) de façon plus générale :

$$log(h(t, \mathbf{x})) = \sum_{l=1}^{L} f_l(t, \mathbf{x}) \tag{1.6}$$

où, t est le temps de suivi, $\mathbf{x}$ est un vecteur de covariables et $f_l(t, \mathbf{x})$ sont des fonctions des

covariables et/ou du temps (e.g. linéaire, spline, produit tensoriel).

Pour chaque fonction f_l, on définit une base de fonctions $b_{lk}(t, \mathbf{x}), k \in \{1, .., K_l\}$ permettant d'écrire f_l (ou du moins une approximation) comme une combinaison linéaire de ces fonctions, paramétrée par le vecteur $\boldsymbol{\beta}_l$:

$$f_l(t, \mathbf{x}) = \sum_{k=1}^{K_L} \beta_{lk} b_{kl}(t, \mathbf{x})$$

On peut alors écrire le modèle avec les notations matricielles suivantes, qui seront utilisées tout au long de cette thèse :

$$\{log(h(t, \mathbf{x}_i))\}_{1 \leq i \leq n} = \mathbf{X}(t)\boldsymbol{\beta} \tag{1.7}$$

où, $\mathbf{X}(t)$ est la matrice de design et $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ ... \\ \boldsymbol{\beta_L} \end{pmatrix}$. Notons que pour un individu i, la fonction $log(h(t, \mathbf{x}_i))$ est définie pour tout t, pas simplement en t_i, le temps d'observation pour l'individu i. On note $X_i(t)$ la ligne de la matrice $\mathbf{X}(t)$ correspondant à l'individu i.

Par exemple, si l'on souhaite étudier l'association entre l'âge et le taux, on pourra considérer les modèles suivants :

— Modèle 1 : Effet non linéaire proportionnel de l'âge

$$log(h(t, age)) = f(t) + g(age)$$

où, f et g sont des splines

— Modèle 2 : Effet non linéaire non proportionnel de l'âge

$$log(h(t, age)) = f(t, age)$$

où, f est un produit tensoriel permettant de modéliser conjointement la non-linéarité et la non-proportionnalité.

1.1.3 Vraisemblance du modèle

Classiquement, les analyses de survie s'accompagnent d'un phénomène de censure qui empêche d'observer la réalisation du temps d'évènement. Les observations sont alors les réalisations d'un couple de variables aléatoires (T, D). D vaut 1 si l'évènement est observé et 0 sinon. T est le dernier temps observé pour l'individu (correspondant au temps survenant en premier entre le temps d'évènement et le temps de censure). Dans la suite, nous supposons que la survenue de la censure est indépendante de la survenue du temps d'évènement (censure non-informative) sachant les covariables du modèle. Ce type de censure peut être, par exemple, de type administratif (fin de l'étude) ou une perte de vue pour une raison non liée à l'évènement d'intérêt.

Nous observons la réalisation de n couples (T, D) indépendants que l'on notera $(t_1, \delta_1), ..., (t_n, \delta_n)$. On note respectivement g et G, le taux et la fonction de répartition des temps de censure. La probabilité d'observer le couple (t_i, δ_i) peut donc s'écrire comme suit :

— si l'évènement est observé :

$$p(t_i, \delta_i) = S(t_i)h(t_i)(1 - G(t_i))$$

— si l'évènement est censuré :

$$p(t_i, \delta_i) = S(t_i)g(t_i)(1 - G(t_i))$$

La log-vraisemblance du modèle peut donc s'écrire :

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} log\{S(t_i, \boldsymbol{\beta})\} + log\{1 - G(t_i)\} + \delta_i log(h(t_i, \boldsymbol{\beta})) + (1 - \delta_i)log(g(t_i))$$

comme G et g ne dépendent pas de $\boldsymbol{\beta}$, on obtient :

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i log(h(t_i, \boldsymbol{\beta})) - \int_0^{t_i} h(s, \boldsymbol{\beta})ds \tag{1.8}$$

1.1.4 Estimation des paramètres du modèle

Approximation numérique

La maximisation de la vraisemblance demande donc d'évaluer l'intégrale du taux, appelée taux cumulé. Une première solution est d'utiliser une méthode numérique, comme une quadrature de Gauss-Legendre, pour se ramener à une somme de termes :

$$\mathcal{L}_i(\boldsymbol{\beta}) = log(h(t_i, \boldsymbol{x_i}, \boldsymbol{\beta}))\delta_i - \sum_{q=1}^{Q} w_q^{GL} h(t_q^{GL}, x_i, \boldsymbol{\beta})$$

où, w_m^{GL} et t_m^{GL} sont respectivement les poids et les noeuds de la quadrature.

Les dérivées premières et secondes par rapport aux paramètres $\boldsymbol{\beta}$ sont alors calculables, ce qui permet d'utiliser des algorithmes de recherche de maximum sur une fonction, du type Newton-Raphson (Fauvernier et al., 2019). Cette approche est, par exemple, employée dans les packages R `mexhaz`, `survPen`.

Le modèle exponentiel par morceaux

Une autre approche pour maximiser (1.8) est de supposer que le taux est constant par morceaux, c'est-à-dire d'utiliser un modèle exponentiel par morceaux (Michael Friedman, 1982). L'idée de cette approche est de découper le temps de survenue d'un évènement t_i en petits intervalles sur lesquels le temps est supposé suivre une loi exponentielle. Notons les points de ce découpage pour l'individu i, $(t_{[i0]}, t_{[i1]}, ..., t_{[ik_i]})$. On note $h_{ij} = exp(X_{ij}\boldsymbol{\beta})$ la valeur du taux pour l'individu i sur son intervalle $[t_{[ij-1]}, t_{[ij]}]$ et X_{ij} est la ligne de la matrice de design de l'individu i pour chaque intervalle $[t_{[ij-1]}, t_{[ij]}]$. On peut alors écrire la contribution d'un sujet à la log-vraisemblance du modèle :

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{\beta}) &= \sum_{j=1}^{k_i} I(j = k_i) \log(h_{ij}) I(\delta_i = 1) - (t_{[j+1]} - t_{[j]})h_{ij} \\
&= \sum_{j=1}^{k_i} I(j = k_i) \log(h_{ij}) I(\delta_i = 1) - \exp\left(\log(h_{ij}) + \log(t_{[j+1]} - t_{[j]}) \right)
\end{aligned}$$

On reconnaît dans cette expression la log-vraisemblance d'un modèle de Poisson de k_i individus indépendants et d'offset $\Delta = \log(t_{[j+1]} - t_{[j]})$. Maximiser la vraisemblance du modèle exponentiel par morceaux revient donc à maximiser celle d'un modèle de Poisson sur un jeu de données augmenté (*data-splitting*), constitué de duplicata des individus associés à chaque période d'observation.

L'intérêt de cette approche est qu'elle permet d'ajuster la régression en utilisant les outils conçus pour le modèle de Poisson (et ses extensions mixtes ou pénalisés) et notamment du cadre des GAM (*Generalized Additive Models*). Le choix du pas de découpage des données par intervalles doit être suffisamment fin pour garantir la validité de l'hypothèse de constance des taux à l'intérieur de chaque intervalle, mais pas trop petit afin d'éviter des temps de calcul excessivement longs. Le package R `pammtools` constitue une aide à la constitution du jeu de données augmenté.

1.2 Modèles de taux avec fragilité

Dans le chapitre précédent, nous avons présenté le modèle flexible du taux, dans un cadre où les observations sont indépendantes conditionnellement aux covariables. Cependant, il est souvent impossible de tenir compte de toutes les variables influentes dans le modèle, laissant une hétérogénéité non observée. Ainsi, des sujets issus d'une même zone géographique, par exemple, peuvent partager certaines caractéristiques non observées (Charvat et al., 2016). Dans ce cas, les temps de survie peuvent être corrélés, mettant à mal l'inférence statistique issue du modèle. Le modèle mixte permet de prendre en compte cette corrélation. En effet, si l'on spécifie que les individus issus d'un même groupe (*cluster*) partagent un effet aléatoire, on obtient un modèle avec fragilité partagée (*shared fraily models*) permettant d'expliquer la corrélation intra-groupe. Ce type de modèle peut également être utile lorsque l'on observe plusieurs temps d'évènements par individus (évènements récurrents) (Balan and Putter, 2020; Rondeau et al., 2012).

Si un modèle de taux avec fragilité peut être très pertinent dans un contexte d'analyse d'évènements indésirables, par exemple pour des données multicentriques, l'objectif de ce chapitre est également d'introduire quelques notions qui nous seront utiles lorsque nous aborderons la pénalisation et la modélisation des évènements récurrents (partie 3), comme la notion de *vraisemblance marginale*. Nous commençons par présenter le cadre des modèles linéaires mixtes, avant d'aborder les spécificités associées au modèle du taux.

1.2.1 Régression linéaire à effets mixtes

Présentation du modèle

On considère une variable aléatoire réponse Y_{ij} pour la $j^{\text{ième}}$ unité (j = 1,2,...,m) d'un groupe i (i = 1,2,...,n)(par exemple, un patient j au sein d'un hôpital i). On suppose qu'il existe une corrélation entre les observations d'un même groupe. L'hypothèse d'indépendance des observations du modèle linéaire n'est donc plus valide. On peut introduire un effet aléatoire b_i non observé, commun à tous les sujets du groupe i, tel que :

$$Y_{ij} = \mu + b_i + \epsilon_{ij} \tag{1.9}$$

où, μ est la moyenne de l'ensemble des observations, les b_i et les ϵ_{ij} sont iid de lois respectives $\mathcal{N}(0, \sigma_a^2)$ et $\mathcal{N}(0, \sigma^2)$. De plus, les b_i et les ϵ_{ij} sont indépendants.

b_i est un intercept aléatoire, qui caractérise la différence de moyennes des individus du groupe i par rapport à la moyenne globale. Ainsi, la moyenne dans le groupe i est égale à $\mu + b_i$. Conditionnellement à b_i, les réponses Y_{ij} sont supposées indépendantes. Les ϵ_{ij} sont les erreurs du modèle. Contrairement à un effet fixe, b_i n'est pas un paramètre du modèle. Nous cherchons seulement à estimer μ, σ et σ_a.

En pratique, il est possible d'ajouter plusieurs effets aléatoires dans un modèle. Pour les besoins de cette thèse et pour simplifier les notations, nous ne considérerons que le cas de l'intercept aléatoire. Dans la suite, nous utiliserons une écriture sous forme matricielle. On suppose que la moyenne μ est une combinaison des valeurs des covariables du modèle et d'un vecteur de paramètres $\boldsymbol{\beta}$ à estimer. On peut écrire le modèle :

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \\ \mathbf{b} &\sim \mathcal{N}(0, V_\theta) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(0, W_\theta) \end{aligned} \tag{1.10}$$

$\mathbf{b}$ est le vecteur des effets aléatoires $(n \times 1)$, $\mathbf{Z}$ est la matrice $(nm \times n)$ donnant la structure des effets aléatoires et $\boldsymbol{\epsilon}$ est le vecteur des résidus du modèle $(nm \times 1)$. V_θ et W_θ sont respectivement les matrices de variance-covariance des effets aléatoires et des erreurs. On note $\boldsymbol{\theta}$ le vecteur des paramètres de V_θ et W_θ.

Pour un nombre fixe de trois individus par groupe, la matrice $\mathbf{Z}$ a la structure suivante :

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 1 & 0 & 0 & 0 & ... & 0 \\ 1 & 0 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & 0 & ... & 0 \\ ... & & & & & \\ 0 & 0 & 0 & 0 & ... & 1 \\ 0 & 0 & 0 & 0 & ... & 1 \\ 0 & 0 & 0 & 0 & ... & 1 \end{pmatrix}$$

On peut remarquer que le modèle (1.10) peut également s'écrire de la façon suivante :

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \Lambda_\theta) \tag{1.11}$$

où, Λ_θ est la matrice de variance-covariance des observations $(nm \times nm)$, qui dépend d'un vecteur θ de paramètres. On peut l'écrire : $\Lambda_\theta = \mathbf{Z}V_\theta\mathbf{Z}^T + W_\theta$. Dans le cas du modèle linéaire gaussien avec intercept aléatoire gaussien, $\Lambda_\theta = \sigma_a^2\mathbf{B}_{nm} + \sigma^2 I_{nm}$, où I_{nm} et σ^2 sont respectivement la matrice identité $(nm \times nm)$ et le paramètre caractérisant la variance résiduelle du modèle, et $\mathbf{B}_{nm} = \mathbf{Z}\mathbf{Z}^T$ $(nm \times nm)$.

Dans notre exemple précédent, $\mathbf{B}_{nm}$ prend la forme suivante :

$$\mathbf{B}_{nm} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & ... & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & ... & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & ... & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & ... & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & ... & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & ... & 0 \\ ... & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & ... & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & ... & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & ... & 1 \end{pmatrix}$$

Estimation des paramètres $\boldsymbol{\beta}$ du modèle

Pour estimer les paramètres du modèle linéaire mixte, on va chercher à maximiser la log-vraisemblance du modèle, soit $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \log(f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}))$. Il est possible d'estimer les paramètres du modèle linéaire mixte de plusieurs manières plus ou moins efficaces (Foulley et al., 2002). Nous en détaillons trois dans cette section :

(i) Par maximisation de la vraisemblance de la loi normale multivariée
En utilisant l'écriture (1.11) du modèle, on peut écrire la vraisemblance du modèle :

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}^n |\boldsymbol{\Lambda}_\theta|^{n/2}} \exp\left(- \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\boldsymbol{\Lambda}_\theta^{-1}}}{2} \right) \tag{1.12}$$

où l'on notera $\|\mathbf{X}\|^2_{\boldsymbol{\Lambda}_\theta^{-1}} = \mathbf{X}^T \boldsymbol{\Lambda}_\theta^{-1} \mathbf{X}$.

La maximisation de cette vraisemblance par rapport au couple de paramètres $(\boldsymbol{\beta}, \boldsymbol{\theta})$ n'est pas faisable de façon analytique et nécessiterait un algorithme de type Newton-Raphson. Cependant, dans le cas linéaire gaussien, pour $\boldsymbol{\theta}$ fixé :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Lambda}_\theta^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_\theta^{-1} \mathbf{y} \tag{1.13}$$

On peut réinjecter la formule de l'estimateur dans $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta})$ pour obtenir la vraisemblance profilée, évitant de rechercher le maximum de vraisemblance de façon itérative. On remarque cependant que cette méthode nécessite une inversion de la matrice $\boldsymbol{\Lambda}_\theta$ de taille $(nm \times nm)$, ce qui est donc coûteux d'un point de vue calculatoire.

(ii) En utilisant la vraisemblance marginale
Une seconde manière de procéder est de voir la vraisemblance $f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$ comme une vraisemblance dite marginale, dans laquelle on aura éliminé le vecteur $\mathbf{b}$, en intégrant $f(y, \mathbf{b}|\boldsymbol{\beta}, \boldsymbol{\theta})$

par rapport à $\mathbf{b}$:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta}) &= \int_{\mathbf{b}} f(\mathbf{y},\mathbf{b}|\boldsymbol{\beta},\boldsymbol{\theta})d\mathbf{b} \\ &= \int_{\mathbf{b}} f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{b})f(\mathbf{b})d\mathbf{b} \end{aligned} \tag{1.14}$$

En inférence Bayésienne, la vraisemblance marginale est utile pour éliminer un paramètre de nuisance de la vraisemblance. En utilisant un développement de Taylor (exact dans ce cas) en $\hat{\mathbf{b}}$, le maximum de $f(\mathbf{y},\mathbf{b}|\boldsymbol{\beta},\boldsymbol{\theta})$, on a :

$$f(\mathbf{y},\mathbf{b}|\boldsymbol{\beta},\boldsymbol{\theta}) = \exp\left(\log(f(\mathbf{y},\hat{\mathbf{b}}|\boldsymbol{\beta},\boldsymbol{\theta})) + \frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^T \frac{\partial^2 log(f(\mathbf{y},\hat{\mathbf{b}}|\boldsymbol{\beta},\boldsymbol{\theta}))}{\partial \mathbf{b}^T \partial \mathbf{b}}(\mathbf{b}-\hat{\mathbf{b}})\right)$$

comme $f(\mathbf{y},\hat{\mathbf{b}}|\boldsymbol{\beta},\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta},\hat{\mathbf{b}})f(\hat{\mathbf{b}})$, on peut remplacer les deux densités dans l'expression précédente :

$$\begin{aligned} f(\mathbf{y},\mathbf{b}|\boldsymbol{\beta},\boldsymbol{\theta}) = \exp\Bigg(&-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|W_\theta|) - \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{b}}\right\|^2_{W_\theta^{-1}} - \frac{m}{2}\log(2\pi) \\ &- \frac{1}{2}log(|V_\theta|) - \frac{1}{2}\hat{\mathbf{b}}^T V_\theta^{-1}\hat{\mathbf{b}}\Bigg) \exp\left(\frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^T[-V_\theta^{-1} - \mathbf{Z}W_\theta^{-1}\mathbf{Z}^T](\mathbf{b}-\hat{\mathbf{b}})\right) \end{aligned}$$

Ainsi,

$$\begin{aligned} \int_{\mathbf{b}} f(\mathbf{y},\mathbf{b}|\boldsymbol{\beta},\boldsymbol{\theta})d\mathbf{b} = \exp\Bigg(&-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|W_\theta|) - \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{b}}\right\|^2_{W_\theta^{-1}} - \frac{m}{2}\log(2\pi) \\ &- \frac{1}{2}log(|V_\theta|) - \frac{1}{2}\hat{\mathbf{b}}^T V_\theta^{-1}\hat{\mathbf{b}}\Bigg) \int_{\mathbf{b}} \exp\left(\frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^T[-V_\theta^{-1} - \mathbf{Z}W_\theta^{-1}\mathbf{Z}^T](\mathbf{b}-\hat{\mathbf{b}})\right)d\mathbf{b} \end{aligned}$$

Comme l'intégrale de la densité d'une loi normale multivariée vaut 1 :

$$\int_{\mathbf{b}} \exp\left(-\frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^T[\mathbf{Z}W_\theta^{-1}\mathbf{Z}^T + V_\theta^{-1}](\mathbf{b}-\hat{\mathbf{b}})\right)d\mathbf{b} = (2\pi)^{m/2}|\mathbf{Z}W_\theta^{-1}\mathbf{Z}^T + V_\theta^{-1}|^{-1/2}$$

En passant ensuite au logarithme :

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta}) = &-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|W_\theta|) - \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{b}}\right\|^2_{W_\theta^{-1}} - \frac{1}{2}\log(|V_\theta|) \\ &- \frac{1}{2}\hat{\mathbf{b}}^T V_\theta^{-1}\hat{\mathbf{b}} - \frac{1}{2}\log(|\mathbf{Z}W_\theta^{-1}\mathbf{Z}^T + V_\theta^{-1}|) \end{aligned}$$

Comme les termes de variance ne dépendent pas de $\boldsymbol{\beta}$ mais qu'en revanche, $\hat{\mathbf{b}}$ dépend de $\boldsymbol{\beta}$, on va donc chercher à maximiser la fonction suivante par rapport à $\mathbf{b}$ et $\boldsymbol{\beta}$:

$$-\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\hat{\mathbf{b}}\right\|^2_{W_\theta^{-1}} - \frac{1}{2}\hat{\mathbf{b}}^T V_\theta^{-1}\hat{\mathbf{b}} \tag{1.15}$$

Les $\hat{\mathbf{b}}$ ne sont pas des paramètres du modèles, on les appelle les BLUPs (Henderson, 1973)(Accronyme anglais pour le Meilleur prédicteur linéaire sans biais ou *Best Linear Unbiased Predictors*).

(iii) En utilisant le maximum a posteriori

Posons le modèle linéaire mixte gaussien dans un cadre Bayésien. On considère une loi *a priori* uniforme sur les paramètres $\boldsymbol{\beta}$. La loi *a posteriori* pour $\boldsymbol{\beta}$ et $\mathbf{b}$ est donnée par :

$$\begin{aligned} f(\boldsymbol{\beta}, \mathbf{b}|\mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{b}) f(\boldsymbol{\beta}) f(\mathbf{b})}{f(\mathbf{y})} \\ &\propto f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{b}) f(\mathbf{b}) \end{aligned} \tag{1.16}$$

en passant au logarithme, on obtient :

$$\propto -\frac{1}{2} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b} \right\|^2_{W_\theta^{-1}} - \frac{1}{2} \mathbf{b}^T V_\theta^{-1} \mathbf{b} \tag{1.17}$$

dont le maximum $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}})$ est appelé le maximum *a posteriori* (MAP).

On retrouve l'expression (1.15), les estimations du maximum de vraisemblance et le MAP coïncident donc dans le cas du modèle linéaire mixte gaussien.

Par ailleurs, la loi *a posteriori* pour $\boldsymbol{\beta}$ et $\mathbf{b}$ est la suivante (voir Wood 2017 page 80) :

$$\begin{pmatrix} \mathbf{b}|\mathbf{y} \\ \boldsymbol{\beta}|\mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}, \begin{pmatrix} \mathbf{Z}^T W_\theta^{-1} \mathbf{Z} + V_\theta^{-1} & \mathbf{Z}^T W_\theta^{-1} \mathbf{X} \\ \mathbf{X}^T W_\theta^{-1} \mathbf{Z} & \mathbf{X}^T W_\theta^{-1} \mathbf{X} \end{pmatrix} \right)^{-1} \tag{1.18}$$

Estimation du paramètre de variance $\boldsymbol{\theta}$

L'estimateur du paramètre de variance $\boldsymbol{\theta}$ du modèle par maximum de vraisemblance sont asymptotiquement sans biais mais dans un échantillon fini, le biais peut être non négligeable. Comme l'estimation des $\boldsymbol{\beta}$ interfère avec celle de $\boldsymbol{\theta}$, on aura à nouveau recours à une vraisemblance marginale ou REML (*Restricted Maximum Likelihood*) afin d'éliminer $\boldsymbol{\beta}$:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta}) &= \int_{\boldsymbol{\beta}} f(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta} \\ &= \int_{\boldsymbol{\beta}} f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta} \end{aligned} \tag{1.19}$$

où, $f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta})$ est la vraisemblance du modèle linéaire mixte et $f(\boldsymbol{\beta}|\boldsymbol{\theta})$ est la loi *a priori* des $\boldsymbol{\beta}$ sachant $\boldsymbol{\theta}$.

En considérant une loi *a priori* uniforme pour $f(\boldsymbol{\beta}|\boldsymbol{\theta})$, et en notant $\hat{\boldsymbol{\beta}}$, le maximum de vraisemblance du modèle $(f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}))$, on réalise le développement de Taylor suivant autour de $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\beta}|\boldsymbol{\theta}) &= f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= \exp\left(\log(f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \boldsymbol{\theta})) + \frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \frac{\partial^2 \log(f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}))}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right) \end{aligned} \tag{1.20}$$

Notons que ce développement est exact car les dérivées d'ordre supérieur à 2 sont toutes nulles dans le cas linéaire gaussien. En réinjectant dans (1.19) et en remplaçant $\frac{\partial^2 \log(f(\mathbf{y}|\hat{\boldsymbol{\beta}},\boldsymbol{\theta}))}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}$ par son expression :

$$f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}|\hat{\boldsymbol{\beta}},\boldsymbol{\theta}) \int_{\boldsymbol{\beta}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \mathbf{X}^T \Lambda_\theta^{-1} \mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \right) d\boldsymbol{\beta}$$

On reconnaît la densité d'une loi normale multivariée, dont l'intégrale fait 1 :

$$(2\pi)^{-p/2} |X^T \Lambda_\theta^{-1} X|^{1/2} \int_{\beta} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \mathbf{X}^T \Lambda_\theta^{-1} \mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \right) d\boldsymbol{\beta} = 1$$

Ce qui permet d'écrire la log-vraisemblance marginale ou log-REML de $\boldsymbol{\theta}$:

$$\begin{aligned} l_r(\boldsymbol{\theta}) &= \mathcal{L}(\mathbf{y}|\hat{\boldsymbol{\beta}},\boldsymbol{\theta}) + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{X}^T\Lambda_\theta^{-1}\mathbf{X}|) \\ &= \frac{p-n}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{X}^T\Lambda_\theta^{-1}\mathbf{X}|) - \frac{1}{2}\log(|\Lambda_\theta|) - \frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2_{\boldsymbol{\Lambda}_\theta^{-1}} \end{aligned} \tag{1.21}$$

1.2.2 Le modèle du taux avec terme de fragilité

Présentation du modèle

Le modèle qui nous intéresse est celui du logarithme du taux avec fragilité. Nous observons le couple $(\mathbf{t}, \boldsymbol{\delta}) = \{(t_{ij}, \delta_{ij})\}_{i=1,..,n, j=1,..,m_i}$ pour chaque sujet j du groupe i.

Pour tout t, on définit le modèle du taux avec fragilité par :

$$h(t, \mathbf{x}_{ij}, b_i) = h(t, \mathbf{x}_{ij}) b_i$$

b_i étant l'effet aléatoire au niveau du groupe, dont la distribution est caractérisée par le paramètre $\boldsymbol{\theta}$. Classiquement, cette distribution pourra être Log-Normale ou Gamma.

Comme précédemment, on définit le modèle de façon flexible :

$$\log(h(t, \mathbf{x}_{ij}, b_i)) = \mathbf{X}_{ij}(t)\boldsymbol{\beta} + \log(b_i) \tag{1.22}$$

où, $\mathbf{X}_{ij}(t)$ est la ligne de la matrice de design $\mathbf{X}(t)$ correspondant au sujet j du groupe i.

Vraisemblance marginale du modèle

En reprenant les notations de la section précédente pour $\mathbf{y} = (\mathbf{t}, \boldsymbol{\delta})$, on constate qu'il n'est plus possible d'écrire directement $f(\mathbf{t}, \boldsymbol{\delta}|\boldsymbol{\beta}, \boldsymbol{\theta})$. Nous disposons en revanche de $f(\mathbf{t}, \boldsymbol{\delta}|\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta})$ et de $f(\mathbf{b}|\boldsymbol{\theta})$. On pourra donc se servir de l'écriture marginale de la vraisemblance par rapport aux effets aléatoires pour estimer les paramètres du modèle :

$$\begin{aligned} f(\mathbf{t}, \boldsymbol{\delta}|\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int_{\mathbf{b}} f(\mathbf{t}, \boldsymbol{\delta}, \mathbf{b}|\boldsymbol{\beta}, \boldsymbol{\theta}) d\mathbf{b} \\ &= \prod_{i=1}^{n} \int_{b_i} f(\mathbf{t}_i, \boldsymbol{\delta}_i, b_i|\boldsymbol{\theta}, \boldsymbol{\beta}) db_i \\ &= \int_{b_i} \prod_{i=1}^{n} f(\mathbf{t}_i, \boldsymbol{\delta}_i|\boldsymbol{\beta}, b_i, \boldsymbol{\theta}) f(b_i|\boldsymbol{\theta}) db_i \\ &= \int_{b_i} \prod_{i=1}^{n} \left[\prod_{j=1}^{m_i} f(t_{ij}, \delta_{ij}|\boldsymbol{\beta}, b_i, \boldsymbol{\theta}) \right] f(b_i|\boldsymbol{\theta}) db_i \end{aligned} \tag{1.23}$$

En utilisant l'écriture de la vraisemblance du modèle de taux (1.8), on obtient :

$$f(\mathbf{t}, \boldsymbol{\delta}|\boldsymbol{\beta}, \boldsymbol{\theta}) \propto \int_b \prod_{i=1}^{n} \left[\prod_{j=1}^{m_i} \{h(t_{ij}) b\}^{\delta_{ij}} \exp\left(- \int_0^{t_{ij}} h(s) b ds \right) \right] f(b|\boldsymbol{\theta}) db \tag{1.24}$$

Calcul du maximum de vraisemblance

Contrairement au cas linéaire, l'intégrale sur $\mathbf{b}$ de la vraisemblance (1.24) ne s'écrit pas de façon exacte et il faut alors recourir à une approximation. Décrivons brièvement quelques méthodes utilisées pour réaliser cela.

(i) Approximation de l'intégrale par une quadrature : La quadrature de Gauss-Hermite
La quadrature de Gauss-Hermite permet d'approximer des intégrales de la forme :

$$\int_{-\infty}^{+\infty} f(x) \exp(-x^2) dx \tag{1.25}$$

par une somme définie de la façon suivante :

$$\sum_{q}^{Q} w_q f(x_q)$$

où, x_q sont les noeuds de la quadrature, égaux aux racines du polynôme d'Hermite d'ordre Q [1] et $w_q = \frac{2^{Q+1}Q!\sqrt{\pi}}{[H'_Q(x_q)]^2}$ sont les poids associés. On peut étendre (1.25) pour n'importe quelle densité gaussienne $\phi(x; \mu, \sigma^2)$:

$$\int_{-\infty}^{+\infty} f(x)\phi(x; \mu, \sigma^2)dx \tag{1.26}$$

en prenant $w_q^* = \frac{w_q}{\sqrt{\pi}}$ et $x_q^* = \mu + 2^{1/2}\sigma x_q$, ce qui permet donc d'approximer l'intégrale (1.23).

En pratique, cette quadrature demande beaucoup de points pour atteindre une bonne précision car cette méthode n'est pas spécifique à la fonction f. Les noeuds peuvent atteindre de très grandes valeurs qui peuvent poser des problèmes numériques. On lui préfère donc sa version adaptative.

(ii) Approximation de l'intégrale par une quadrature : La quadrature adaptative de Gauss-Hermite
Cette quadrature permet une généralisation de l'approximation au cadre suivant :

$$\int_{-\infty}^{+\infty} \exp(p\ l(x))dx$$

où, p est un scalaire et l une fonction unimodale qui ne dépend pas de p.

(Liu and Pierce, 1994) ont proposé une redéfinition des poids et des noeuds dans ce cadre. L'intégrale de Gauss-Hermite adaptative peut donc se définir comme suit, pour un nombre de points de quadrature Q :

$$\int_x \exp(p\ l(x))dx \ \approx\ 2^{1/2}\hat{s}\ p^{-1/2} \sum_{q=1}^{Q} w_q^* l(x_q^*) \tag{1.27}$$

où,

$$\hat{s}^2 = -\left\{ \frac{\partial^2 l(\hat{\mu})}{\partial x^2} \right\}^{-1}$$
$$x_q^* = \hat{\mu} + 2^{1/2}\hat{s}\ p^{-1/2}\ x_q$$
$$w_q^* = w_q \exp(x_q^2)$$

et μ est le mode de l.

Il existe une version multivariée de la quadrature (Jin and Andersson, 2020), permettant d'intégrer par rapport à un vecteur :

$$\int_{\mathbf{x}} \exp(p\, l(\mathbf{x}))d\mathbf{x} \ \approx\ 2^{m/2} \left| \frac{\partial^2 l(\hat{\boldsymbol{\mu}})}{\partial \mathbf{x}^T \partial \mathbf{x}} \right|^{-1/2} \sum_{j_1,..j_m=1}^{Q} \left\{ \prod_{i=1}^{m} w_{j_l} exp(z_{j_l}) \right\} \exp(p\, l(2^{1/2} \frac{\partial^2 l(\hat{\boldsymbol{\mu}})}{\partial \mathbf{x}^T \partial \mathbf{x}}^{-1/2} \mathbf{z}_{j_1,\ldots,j_m} + \hat{\boldsymbol{\mu}}))$$

1. Le polynôme d'Hermite d'ordre Q est défini par : $H_Q(x) = (-1)^Q e^{x^2/2} \frac{d^Q e^{-x^2/2}}{dx^Q}$

où, $\mathbf{z}_{j_1,...,j_z} = (z_{j_1}, ...z_{j_K})^T$ pour est un vecteur $(m \times 1)$

Appliquons cela à notre modèle (1.22) en prenant un seul noeud sur la quadrature. Notre objectif est d'approximer $\int_{\mathbf{b}} f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{b}) f(\mathbf{b}|\boldsymbol{\theta}) d\mathbf{b}$. L'intégrale se fait par rapport au vecteur aléatoire $\mathbf{b}$ mais comme on fait l'hypothèse que les individus sont indépendants, il est possible de se ramener à un produit d'intégrales univariées au niveau de chaque individu. Ainsi, pour l'individu i, on va approximer $\int_{b_i} f(\mathbf{y}_i|\boldsymbol{\beta}, b_i, \boldsymbol{\theta}) f(b_i|\boldsymbol{\theta}) db_i$. On prend donc $p = 1$ et $l = \log(f(\mathbf{y}_i|\boldsymbol{\beta}, b_i, \boldsymbol{\theta}) f(b_i|\boldsymbol{\theta}))$.

On note $\hat{b}_i$, le mode de $\log(f(\mathbf{y}_i, b_i|\boldsymbol{\beta}, \boldsymbol{\theta}))$. Pour Q=1, le polynôme d'Hermite est l'identité donc la racine $z_1 = 0$ et le poids associé est $w_1 = \sqrt{\pi}$. De plus,

$$-\left\{ \frac{\partial^2 log(f(\mathbf{y}_i, \hat{b}_i|\boldsymbol{\beta}, \boldsymbol{\theta}))}{\partial b_i^2} \right\}^{-1} = (h(\hat{b}_i) + \sigma_a^{-2})^{-1}$$

où, $h_\theta(\hat{bi}) = -\frac{d^2 \log(f(\mathbf{y}_i|\boldsymbol{\beta}, \hat{b}_i, \boldsymbol{\theta}))}{db_i^2}$

Au final, la formule (1.27) devient :

$$\int_{b_i} f(\mathbf{y}_i|\boldsymbol{\beta}, b_i, \boldsymbol{\theta}) f(b_i|\boldsymbol{\theta}) db_i \approx \sqrt{2\pi}(h_\theta(\hat{b}_i) + \sigma_a^{-2})^{-1/2} f(\mathbf{y}_i, \hat{b}_i|\boldsymbol{\beta}, \boldsymbol{\theta})$$

De la même manière, en utilisant la forme multivariée sur la vraisemblance totale, on a :

$$\int_b f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta}, \boldsymbol{\theta}) d\mathbf{b} \approx f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta}, \boldsymbol{\theta})(2\pi)^{m/2} |\mathbf{H}_\theta(\hat{\mathbf{b}}) + V_\theta^{-1}|^{-1/2} \tag{1.28}$$

en notant, $\hat{\mathbf{b}}$ le mode de $f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta}, \boldsymbol{\theta})$ et $H_\theta(\hat{\mathbf{b}}) = -\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\beta}, \hat{\mathbf{b}}, \boldsymbol{\theta})}{\partial \mathbf{b}^T \mathbf{b}}$. Notons que contrairement au modèle linéaire, le terme H peut dépendre de $\boldsymbol{\beta}$.

Nous verrons que ce résultat est identique à celui obtenu par l'approximation de Laplace (d'ordre 1) à la section suivante.

(iii) Approximation de Laplace (d'ordre 1)
Une autre manière de procéder est de considérer l'approximation de Laplace pour l'intégrale. On définit $\hat{\mathbf{b}}$ maximisant $f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\theta}) f(\mathbf{y}, \mathbf{b}|\boldsymbol{\theta})$, c'est-à-dire le maximum *a posteriori* (MAP) de la vraisemblance de $\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{b}$ avec une distribution *a priori* gaussienne. On fait un développement de Taylor à l'ordre 2 :

$$f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta}, \boldsymbol{\theta}) \approx \exp\left(\log(f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta}, \boldsymbol{\theta})) + \frac{1}{2} (\mathbf{b} - \hat{\mathbf{b}})^T \frac{\partial^2 \log(f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta}, \boldsymbol{\theta}))}{\partial \mathbf{b}^T \partial \mathbf{b}} (\mathbf{b} - \hat{\mathbf{b}}) \right)$$

Contrairement au cas du modèle linéaire gaussien, nous faisons une approximation car les dérivées d'ordre supérieur à 2 ne sont plus nulles.

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) \approx f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta}, \boldsymbol{\theta}) \int_b \exp\left(-\frac{1}{2} (\mathbf{b} - \hat{\mathbf{b}})^T (H_\theta(\hat{\mathbf{b}}) + V_\theta^{-1})(\mathbf{b} - \hat{\mathbf{b}}) \right) d\mathbf{b}$$

On reconnaît à nouveau la densité d'une loi normale multivariée. En reprenant les notations de la section précédente :

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\theta}) \approx f(\mathbf{y}, \hat{\mathbf{b}}|\boldsymbol{\beta}, \boldsymbol{\theta})(2\pi)^{m/2} |H_\theta(\hat{\mathbf{b}}) + \mathbf{V}_\theta^{-1}|^{-1/2}$$

L'expression obtenue est connue sous le nom d'approximation de Laplace. On reconnaît l'expression (1.28).

En développant un peu plus cette expression et en passant au logarithme, la log-vraisemblance à maximiser peut donc s'écrire :

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\beta}, \hat{\mathbf{b}}, \boldsymbol{\theta}) - \frac{1}{2}\hat{\mathbf{b}}^T V_\theta^{-1}\hat{\mathbf{b}} - \frac{1}{2}\log(|V_\theta|) - \frac{1}{2}\log(|H_\theta(\hat{\mathbf{b}}) + V_\theta^{-1}|) \tag{1.29}$$

Maximum *a posteriori*
Si l'on reprend le cadre Bayésien défini en section 1.2.1 sur le modèle linéaire mixte gaussien, le vecteur des MAP de $\boldsymbol{\beta}, \mathbf{b}$ s'obtient en recherchant pour $\boldsymbol{\theta}$ fixé, le maximum de :

$$\mathcal{L}(\boldsymbol{\beta}, \hat{\mathbf{b}}, \boldsymbol{\theta}) - \frac{1}{2}\hat{\mathbf{b}}^T V_\theta^{-1}\hat{\mathbf{b}} \tag{1.30}$$

Comme le terme $\frac{1}{2}\log(|H_\theta(\hat{\mathbf{b}}) + V_\theta^{-1}|)$ dépend de $\boldsymbol{\beta}$ dans (1.29), on voit donc que le maximum de vraisemblance et le MAP ne coïncident plus exactement. Approximer le MLE par le MAP est cependant une technique populaire pour les modèles semi-paramétriques (*penalized likelihood method*) (Balan and Putter, 2020; Ripatti and Palmgren, 2000) et notamment implémenté dans le package R `coxme`. Nous verrons son lien avec la pénalisation dans le chapitre suivant.

1.3 Pénalisation

1.3.1 Définition générale

Le nombre et l'emplacement des noeuds des splines introduites dans la section 1.1.1 peuvent avoir une forte influence sur le modèle. Trop de noeuds conduit à du surajustement, pas assez à une mauvaise approximation de la fonction à modéliser. Pour aider au choix du modèle, il est possible d'introduire un terme de pénalisation dans la log-vraisemblance du modèle :

$$\mathcal{L}_p(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - pen(\boldsymbol{\beta}, \boldsymbol{\kappa}) \tag{1.31}$$

où, $\boldsymbol{\kappa}$ est un vecteur de paramètres permettant de contrôler le niveau de pénalisation du critère et *pen* est la fonction introduisant la pénalisation sur le vecteur de paramètres $\boldsymbol{\beta}$.

Les fonctions de pénalisation peuvent avoir diverses propriétés de régularité permettant d'avoir des effets différents (Antoniadis et al., 2011) : lissage, sélection de variables, dans le contexte de la grande dimension (nombre de colonnes p $\gg$ nombre de lignes n), ajustement en présence de covariables corrélées (*shrinkage*).

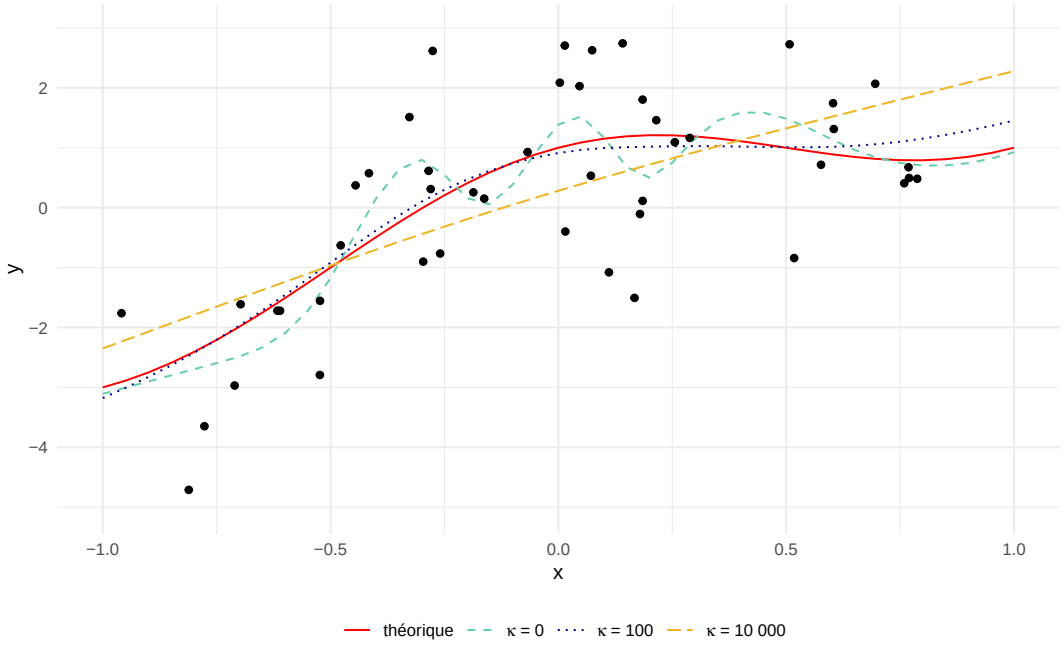

FIGURE 2.1 – Impact du paramètre de lissage κ sur l'estimation des paramètres. *Dans cet exemple, les données* $\boldsymbol{y}$ *suivent une loi normale conditionnellement à la covariable continue* $\boldsymbol{x}$*. La forme fonctionnelle théorique entre* $\boldsymbol{y}$ *et* $\boldsymbol{x}$ *est représentée en rouge. On ajuste un modèle pénalisé avec une spline cubique naturelle à 10 noeuds équirépartis, en faisant varier le paramètre de lissage* κ*.*

Dans notre contexte, nous cherchons à introduire des effets non-linéaires, tout en limitant le sur-ajustement, c'est-à-dire à lisser les prédictions. Pour faire du lissage sur une spline f telle que $f(x) = \sum_{i=1}^{I} \beta_i B_i(x)$, on définit une pénalité sur sa dérivée seconde :

$$pen(\boldsymbol{\beta}, \kappa) = \frac{1}{2}\kappa \int f''(x)^2 dx \tag{1.32}$$

Comme illustré en Figure 2.1, le paramètre κ fait un compromis entre ajustement aux données et lissage. Ainsi, plus κ est grand plus la forme fonctionnelle tend vers la linéarité. On peut écrire $f(x) = \mathbf{B}\boldsymbol{\beta}$ avec $\boldsymbol{\beta}^T = (\beta_1 \ ... \ \beta_I)$ et $\mathbf{B}(x) = (B_1(x) \ ... \ B_I(x))$. En notant $\mathbf{B}''(x) = (\mathbf{B}_1''(x) \ ... \ \mathbf{B}_I''(x))$, on obtient :

$$\begin{aligned}\int f''(x)^2 dx &= \int (\boldsymbol{\beta}\mathbf{B}''(x))^T(\boldsymbol{\beta}\mathbf{B}''(x))dx \\ &= \boldsymbol{\beta}^T\mathbf{S}\boldsymbol{\beta}\end{aligned}$$

où, $\mathbf{S} = \int \mathbf{l}''(x)^T\mathbf{l}''(x)dx$

Pénalisation et splines multidimensionnelles

Lorsque l'on souhaite modéliser des interactions entre deux variables continues, par exemple le temps et l'âge, on s'intéresse à l'estimation d'une surface. Si on note $f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x)$ et $f_y(y) = \sum_{j=1}^{J} \beta_i b_j(y)$ les splines unidimensionnelles (dites marginales), on peut définir la spline multidimensionnelle associée à la surface par un produit tensoriel entre les deux splines marginales :

$$f_{x,y}(x,y) = \sum_{i=1}^{I}\sum_{j=1}^{J} \delta_{ij} a_i(x) b_j(y) \tag{1.33}$$

Classiquement, à chaque dimension sera associé un paramètre de lissage. En pratique, les matrices de pénalisation associées au tensor peuvent être définies à partir des matrices de pénalisation marginales (celles associées à f_x et f_y dans notre exemple.)

Cette définition de la spline multidimensionnelle pénalisée se généralise au-delà du cas bidimensionnel présenté ici.

Dans la suite, nous nous plaçons dans le cadre de régression flexible défini par Wood 2004 pour les GAM, puis de façon plus générale Wood et al. 2016 pour toute vraisemblance régulière. A chaque fonction f_j, on associe une pénalisation quadratique de son vecteur des paramètres $\boldsymbol{\beta}_j$ composée d'un (ou plusieurs en cas de spline multidimensionnelle) paramètre de lissage κ_j à déterminer et d'une matrice de pénalisation connue $\mathbf{S}_j$.

On définit la fonction de log-vraisemblance pénalisée de la façon suivante :

$$\mathcal{L}_p(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{J} \kappa_j \boldsymbol{\beta}_j^T \boldsymbol{S}_j \boldsymbol{\beta}_j \tag{1.34}$$

On écrira $\boldsymbol{S}_j$ sous forme d'une matrice diagonale par blocs contenant la matrice de pénalisation associée aux paramètres de la spline j et des zéros ailleurs. Notons que le vecteur $\boldsymbol{\beta}$ contient également les autres paramètres du modèle qui ne sont pas pénalisés. La matrice de pénalisation est donc nulle pour ces paramètres.

De façon synthétique on écrira :

$$\mathcal{L}_p(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T \mathbf{S}_{\boldsymbol{\kappa}} \boldsymbol{\beta} \tag{1.35}$$

où, $\mathbf{S}_{\boldsymbol{\kappa}} = \begin{bmatrix} 0 & \kappa_1\mathbf{S}_1 & 0 & 0 & 0 \\ 0 & 0 & \kappa_2\mathbf{S}_2 & 0 & 0 \\ 0 & & & \ddots & \\ 0 & 0 & 0 & 0 & \kappa_L\mathbf{S}_L \end{bmatrix}$

1.3.2 Vision Bayésienne de la pénalisation

Nous voyons clairement apparaître dans (1.35) le lien avec la vraisemblance *a posteriori* du modèle mixte (1.30). Le vecteur des paramètres du modèle pénalisé peut être vu comme un vecteur d'effets fixes (paramètres non pénalisés) et d'effets aléatoires (paramètres pénalisés). Estimer le paramètre de lissage revient à estimer le paramètre de variance des effets aléatoires du modèle mixte.

Un certain nombre de concepts empruntés à l'inférence Bayésienne et introduits dans le chapitre précédent sont donc également utilisés pour estimer les paramètres et faire de l'inférence dans le modèle pénalisé.

Vraisemblance marginale pour les paramètres de lissage

Critère LAML

Il est possible d'utiliser la vraisemblance restreinte REML pour estimer les paramètres de lissage (Wahba, 1985). En utilisant une approximation de Laplace pour calculer l'intégrale on peut ainsi construire le critère LAML (*Laplace approximate marginal likelihood*) (Wood et al., 2016). Soit $f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\kappa})$, la vraisemblance du modèle. On considère une loi *a priori* impropre gaussienne sur $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta}|\boldsymbol{\kappa}) = \frac{1}{\sqrt{2\pi}^{p-M_p}} |\mathbf{S}_{\boldsymbol{\kappa}}|_+^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\mathbf{S}_{\boldsymbol{\kappa}}\boldsymbol{\beta} \right)$$

où, p est la dimension de $\boldsymbol{\beta}$, M_p est le nombre de paramètres non pénalisés et $|\mathbf{S}_{\boldsymbol{\kappa}}|_+$ est le produit des valeurs propres positives de la matrice de pénalisation.

On définit le critère LAML par :

$$\mathcal{V}(\boldsymbol{\kappa}) = \log\left(\int_{\boldsymbol{\beta}} f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\kappa}) f(\boldsymbol{\beta}|\boldsymbol{\kappa}) d\boldsymbol{\beta} \right)$$

En notant $\hat{\boldsymbol{\beta}}$ le maximum de $f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\kappa})f(\boldsymbol{\beta}|\boldsymbol{\kappa})$ (et donc *a fortiori* de (1.35)), on procède par un développement de Taylor en $\hat{\boldsymbol{\beta}}$ comme dans le chapitre précédent, ce qui nous conduit à l'expression suivante :

$$\mathcal{V}(\boldsymbol{\kappa}) = \mathcal{L}(\hat{\boldsymbol{\beta}}) + \frac{M_p}{2}\log(2\pi) + \frac{1}{2}\log(|\mathbf{S}_{\boldsymbol{\kappa}}|_+) - \frac{1}{2}\log(|\mathcal{H}|) \tag{1.36}$$

en notant $\mathcal{H}$ l'opposé de la hessienne de la vraisemblance pénalisée (1.35).

LAML versus GCV/LCV

Une façon alternative d'estimer les paramètres $\boldsymbol{\kappa}$ est par cross-validation, e.g. *leave-one-out* cross-validation. En pratique, comme la *cross-validation* est gourmande en ressources de calcul, des formules d'approximation de la cross-validation existent comme la GCV *Generalized Cross-Validation* (Craven and Wahba, 1978).

En survie, on pourra chercher à minimiser le critère LCV (Joly et al., 1998) :

$$LCV(\boldsymbol{\kappa}) = -\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\kappa}) + tr\left\{\mathcal{H}^{-1}H\right\} \tag{1.37}$$

où, $tr(.)$ est la trace de la matrice, $\mathcal{H}$ l'opposé de la hessienne de la vraisemblance pénalisée et H l'opposé de la hessienne de la vraisemblance (non pénalisée).

$tr\left\{\mathcal{H}^{-1}H\right\}$ peut être vue comme une estimation des degrés de liberté effectifs du modèle (Gray, 1992), ce qui fait du critère LCV un équivalent du critère AIC.

Pour de petits échantillons, Wahba 1985 ont établi par simulation que GCV et LAML donnaient des résultats proches. Cependant, lorsque la taille augmente et asymptotiquement, les erreurs quadratiques moyennes sont plus faibles pour GCV que pour LAML. Les deux critères peuvent présenter plusieurs maxima locaux. Le phénomène touche davantage GCV que LAML (Reiss and Todd Ogden, 2009). Dans le contexte particulier des modèles de taux flexibles pénalisés, LAML a donné des résultats plus lisses que LCV par simulation (Fauvernier et al., 2019). Enfin, l'un des avantages de LAML est qu'il permet d'estimer l'incertitude sur les paramètres de lissage.

Estimateur Bayésien empirique de la variance des paramètres

La vraisemblance pénalisée peut être vue comme une loi *a posteriori* du modèle avec une loi *a priori* (impropre) gaussienne sur les paramètres $\boldsymbol{\beta}$. Il est donc possible de construire un intervalle de crédibilité Bayésien pour les paramètres du modèle (Wood et al., 2016) (Supplemtary B.4) :

$$f(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\kappa}) \propto \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{V}_\beta) \tag{1.38}$$

où,

$$\mathbf{V}_\beta = (H + \mathbf{S}_{\boldsymbol{\kappa}})^{-1}, \tag{1.39}$$

avec H l'opposé de la hessienne de la vraisemblance du modèle.

Malgré le biais introduit dans l'estimateur des $\hat{\beta}$ par la pénalisation, l'intervalle de crédibilité bayésien conduit à de bonnes propriétés de couverture car il prend justement en compte ce biais (Nychka, 1988). En particulier, il fournit une bien meilleure couverture que l'estimateur fréquentiste défini par :

$$\mathbf{V}_\beta^F = (H + \mathbf{S}_{\boldsymbol{\kappa}})^{-1} H (H + \mathbf{S}_{\boldsymbol{\kappa}})^{-1} \tag{1.40}$$

1.4 Modèles à risques compétitifs

Nous avons brièvement évoqué le contexte des risques compétitifs dans la première partie de cette thèse. Nous présentons le modèle de taux dans ce contexte de façon plus détaillée.

1.4.1 Cadre de la modélisation

Modélisation "cause-spécifiques"

Comme dans la section 1.1.1, on s'intéresse au temps de survenue d'évènement T chez un sujet. Cependant, dans un cadre compétitif, l'évènement associé E peut être multiple et prendra les valeurs $\{0...J\}$. Par convention, on notera $E = 0$ en cas de censure.

On définit alors le taux cause-spécifique *(cause-specific hazard)* par :

$$h_j(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt, E = j | T \geq t)}{dt} \tag{1.41}$$

Lorsque les évènements sont disjoints, on a donc la relation suivante entre les différents taux cause-spécifiques :

$$h(t) = \sum_{j=1}^{J} h_j(t) \tag{1.42}$$

h est le taux de sortie toutes causes pour lequel on pourra définir la survie toutes causes S comme au chapitre 1.1 :

$$S(t) = \exp\left(- \int_0^t \sum_{j=1}^{J} h_j(u) du \right) \tag{1.43}$$

On définit également la fonction d'incidence cumulée pour la cause de sortie j par :

$$F_j(t) = P(T \leq t,\ E = j) = \int_0^t h_j(u) S(u) du \tag{1.44}$$

On comprend donc que la fonction d'incidence cumulée dépend du taux cause-spécifique h_j mais également des autres taux cause-spécifiques via le terme $S(u)$. Cela pose des difficultés d'interprétation dans ce contexte, puisque l'effet d'une variable sur le taux cause-spécifique n'est pas immédiatement transposable à la fonction d'incidence cause-spécifique[2]. Cette dernière dépend aussi de l'effet de la variable sur les taux des autres causes (Putter et al., 2007). Il est donc souvent recommandé de réaliser les analyses à la fois sur l'échelle de la probabilité d'incidence cumulée et sur l'échelle du taux (Latouche et al., 2013).

La loi de probabilité du couple de variables aléatoires T et E peut s'écrire :

$$f(t, j) = h_j(t) P(T \geq t), \qquad j \in \{0, ..., J\} \tag{1.45}$$

2. En d'autres termes, $F_j(t) \neq \int_t h_j(s) S_j(s) ds$

Modélisation des temps latents d'évènements

Une façon alternative de présenter le cadre compétitif est de s'intéresser à la distribution jointe des variables aléatoires temps d'évènements dits "latents" $T^1, ..., T^J$ (*latent failure time models*), où T^j est le temps écoulé jusqu'à la survenue de l'évènement j. On se place donc dans un cadre hypothétique où chacun de ces temps existe pour chaque individu, mais seul l'un d'eux est observé, censurant les autres. On s'intéresse alors à la fonction de survie jointe de ces temps d'évènements :

$$S(t_1, ..., t_n) = P(T^1 > t_1, ..., T^n > t_n)$$

Il est alors possible de définir des taux sur les temps d'évènements T^j :

$$\tilde{h}_j(t) = \lim_{dt \to 0} \frac{P(t \leq T^j < t + dt | T^j \geq t)}{dt}$$

Sans hypothèse complémentaire sur la dépendance entre ces temps d'évènements, il s'agit d'un problème non identifiable (infinité de solutions) (Tsiatis, 1975). Cette dépendance peut notamment être modélisée en utilisant une copule. Le principal intérêt de cette approche est de répondre à une question de prédiction dans un contexte où l'on retirerait ou réduirait l'évènement compétiteur. Ce cadre ne sera pas approfondi davantage dans cette thèse, et nous nous concentrerons sur la modélisation "cause-spécifique".

Risques compétitifs et censure informative

Dans une étude de survie, on dit que l'on observe de la censure lorsque l'information sur les temps d'évènements n'est pas disponible pour certains participants de l'étude (Leung et al., 1997). Dans un cadre de survie classique, la censure est supposée "non-informative", c'est-à-dire que les individus censurés ont le même profil de risque que ceux encore à risque. Dans un cadre de risques compétitifs, on évoque souvent la présence d'une censure dite "informative" à travers la survenue des évènements compétiteurs.

En réalité, l'évènement concurrent peut être ou non de la censure selon la manière dont on définit le problème et de ce que l'on cherche à mesurer (Young et al., 2020). Dans le cadre cause-spécifique, on observe la variable aléatoire T et le type d'évènement survenu E. L'observation du processus en lui-même n'est pas censurée même si l'intérêt réside dans l'une des causes de sortie en particulier. En revanche, dans le cadre des temps latents, la survenue d'un évènement empêche d'observer les autres temps d'évènement, ce qui en fait de la censure.

1.4.2 Modèles de régression

Il existe deux modèles très connus pour faire de la régression dans un cadre semi-paramétrique : le modèle de Cox et le modèle de Fine-Gray. Le premier est basé sur les taux cause-spécifiques et le second sur les taux dits de sous-répartition (*subdistribution hazards*).

Modèle de taux cause-spécifiques

Modèle

Dans un cadre de risques compétitifs, on définit le modèle semi-paramétrique du taux cause-spécifique par :

$$h_j(t, \mathbf{X}) = h_{0,j}(t) \exp(\mathbf{X}\boldsymbol{\beta}) \tag{1.46}$$

Plus généralement, il est possible d'utiliser un modèle flexible du taux comme présenté en section 1.1, permettant d'obtenir une prédiction de la dynamique du taux au cours du temps.

Vraisemblance (complète) du modèle

En pratique, il est possible d'ajuster ce modèle en considérant les autres causes de sorties au même titre que les évènements de censure (même s'ils n'en sont pas, il s'agit uniquement d'ajuster le modèle). Pour un individu i, nous observons le couple (t_i, e_i). En utilisant la densité (1.45), on peut écrire la vraisemblance du modèle du logarithme du taux :

$$\begin{aligned} f(\mathbf{t}, \mathbf{e}|\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(t_i, e_i|\boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} h_{e_i}(t_i) P(T \geq t_i) \end{aligned}$$

On peut ici définir la quantité suivante :

$$S^{(j)}(t) = \exp\left(-\int_0^t h_j(s)ds\right) \tag{1.47}$$

de telle sorte que :

$$S(t) = \prod_{j=0}^{J} S^{(j)}(t)$$

Si les modèles sur les taux cause-spécifiques ne font pas d'hypothèse de partage de paramètres, on peut donc réécrire la vraisemblance :

$$f(\mathbf{t}, \mathbf{e}|\boldsymbol{\beta}) = \prod_{j=0}^{J} \prod_{i=1}^{n} h_j(t_i|\boldsymbol{\beta}_j)^{\mathbb{1}\{e_i=j\}} S^{(j)}(t_i|\boldsymbol{\beta}_j)$$

où $\boldsymbol{\beta}_j$ sont les vecteurs de paramètres associés au taux cause-spécifique j et $\mathbb{1}\{e_i = j\}$ est une indicatrice prenant la valeur 1 lorsque $e_i = j$ et 0 sinon.

On identifie ainsi, pour chaque terme du produit indicé en j, la vraisemblance d'un modèle qui ferait l'hypothèse de censurer toutes les causes de sortie autres que la cause j. On peut donc, en pratique, estimer les paramètres du modèle en construisant un modèle pour chaque cause de sortie et en censurant les autres comme si elles étaient non informatives. Il s'agit d'une astuce de calcul pour estimer les paramètres, nous ne faisons en réalité pas l'hypothèse que les causes compétitives sont non informatives. Notons que le terme indicé par $j = 0$ est le terme de censure que l'on suppose non informative et peut donc être retiré de la vraisemblance car indépendant des paramètres.

Probabilités d'incidence cumulée

Dans un cadre semi-paramétrique, le calcul des probabilités d'incidence cumulées (et de leurs intervalles de confiance) peut se faire à partir des modèles cause-spécifiques (Ozenne et al., 2017). Cette approche est implémentée dans le package R `riskRegression`.

Si l'on dispose d'une estimation des taux cause-spécifiques, il est possible de recalculer la probabilité d'incidence cumulée en utilisant la relation (1.44). Comme il n'existe pas de

forme analytique de l'intégrale, une estimation ponctuelle de celle-ci peut être réalisée de façon numérique par une quadrature de Gauss-Legendre, par exemple. Pour construire l'intervalle de confiance de la probabilité d'incidence cumulée, il est possible d'utiliser la delta-méthode (Kipourou et al., 2019). [3]

Modèle de taux de sous-répartition de Fine-Gray

Modèle

Le modèle de Fine-Gray (Fine and Gray, 1999) relie directement l'effet de la covariable à la fonction d'incidence cumulée. Pour cela, on définit une quantité $\tilde{h}_j$ appelée taux de sous-répartition (*subdistribution hazard*) telle que :

$$1 - F_j(t) = \exp\left(- \int_0^t \tilde{h}_j(s)ds \right) \tag{1.48}$$

Une régression semi-paramétrique à taux proportionnels est ensuite réalisée pour évaluer l'effet des covariables :

$$\tilde{h}_j(t) = \tilde{h}_{0,j}(t) \exp(\boldsymbol{X}\boldsymbol{\beta}_j) \tag{1.49}$$

Cela signifie qu'il n'est pas nécessaire de réaliser un modèle pour chaque évènement pour avoir la fonction d'incidence cumulée. Il est cependant difficile d'interpréter les coefficients associés aux taux de sous-répartition (Andersen and Keiding, 2012) puisqu'il s'agit du taux instantané de sortie pour la cause j au temps t parmi les patients qui ne sont pas sortis pour la cause j. Ainsi, tous les patients sortis pour une autre cause que j sont considérés comme encore à risque. Par exemple, si l'on considère le décès comme une cause de sortie compétitive, cela signifie que l'on garde les patients décédés dans l'effectif à risque. En outre, le signe des coefficients associés aux covariables renseigne sur le sens de l'association entre la covariable et la fonction d'incidence cumulée mais l'ampleur d'effet est difficile à appréhender.

Probabilités d'incidence cumulée

L'estimation de la probabilité d'incidence cumulée peut être obtenue avec un estimateur de type Breslow pour $H_{j0}(t) = -\int_0^t \tilde{h}_{j0}(s)ds$, de telle sorte que :

$$\hat{F}_j(t) = 1 - \exp[- \hat{H}_{j0}(t) \ \exp(\boldsymbol{X}\boldsymbol{\beta}_j)].$$

Des intervalles de confiance pour la prédiction peuvent être obtenus par une procédure *bootstrap* (Fine and Gray, 1999).

3. Le delta-méthode permet d'estimer la variance d'une fonction f d'une variable aléatoire, ici notre vecteur des paramètres β :

$$\hat{Var}(f(\beta)) = [\nabla f(\beta)]_{|\beta=\hat{\beta}} \ \Sigma_\beta \ {}^t[\nabla f(\beta)]_{|\beta=\hat{\beta}}$$

$[\nabla f(\beta)]_{|\beta=\hat{\beta}} = (\frac{\partial f}{\partial \beta_1}(\hat{\beta}), ..., \frac{\partial f}{\partial \beta_p}(\hat{\beta})$ est le gradient de la fonction f calculé en $\hat{\beta}$ et Σ_β est la matrice de variance-covariance associée au vecteur de coefficients β.

Intervalles de confiance de la probabilité d'incidence cumulée du modèle de taux cause-spécifique

Pour simplifier les notations, on considère un cadre à deux évènements compétiteurs mais la généralisation à plus de deux causes de sortie est immédiate. On indicera les quantités cause-spécifiques (incidence cumulées, taux) de la cause de sortie d'intérêt par j et celles associées à la cause de sortie compétitive j^c. Le gradient de F_j peut donc être calculé comme suit :

— si β_k est un paramètre de h_j, alors :

$$\frac{\partial F_j(t,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} = \int_0^t S_j(u,\boldsymbol{\beta},\mathbf{x}) S_{j^c}(u,\boldsymbol{\beta},\mathbf{x})[\frac{\partial h_j(u,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} - h_j(u,\boldsymbol{\beta},\mathbf{x}) \int_0^u \frac{\partial h_j(v,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} dv] du$$

expression que l'on peut simplifier en utilisant l'expression du taux : $h_j(t,x) = \exp(P[x,t])$, $P[x,t]$ étant un polynôme.

$$\frac{\partial F_j(t,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} = \int_0^t S_j(u,\boldsymbol{\beta},\mathbf{x}) S_{j^c}(u,\boldsymbol{\beta},\mathbf{x}) h_j(u,\boldsymbol{\beta},\mathbf{x})[\frac{\partial P_j(u,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} - \int_0^u h_j(v,\boldsymbol{\beta},\mathbf{x}) \frac{\partial P_j(v,\beta,\mathbf{x})}{\partial \beta_k} dv] du$$

— sinon,

$$\frac{\partial F_j(t,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} = \int_0^t h_j(u,\boldsymbol{\beta},\mathbf{x}) S_j(u,\boldsymbol{\beta},\mathbf{x}) S_{j^c}(u,\boldsymbol{\beta},\mathbf{x})[- \int_0^u \frac{\partial h_{j^c}(v,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} dv] du$$

simplifié en :

$$\frac{\partial F_j(t,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} = \int_0^t h_j(u,\boldsymbol{\beta},\mathbf{x}) S_j(u,\boldsymbol{\beta},\mathbf{x}) S_{j^c}(u,\boldsymbol{\beta},\mathbf{x}) [- \int_0^u h_{j^c}(v,\boldsymbol{\beta},\mathbf{x}) \frac{\partial P_j(v,\boldsymbol{\beta},\mathbf{x})}{\partial \beta_k} dv] du$$

La construction de l'intervalle de confiance peut se faire par approximation normale sur la quantité $log(-log(F_j))$ plutôt que directement sur F_j. Cela permet notamment de s'assurer des bornes entre 0 et 1 pour la probabilité (Pintilie, 2006). En appliquant la delta-méthode sur $log(-log(F_j))$, on obtient :

$$\hat{Var}(log(-log(F_j(t,\boldsymbol{\beta},\mathbf{x})))) = \frac{\hat{Var}(F_j(t,\boldsymbol{\beta},\mathbf{x}))}{(log(F_j(t,\hat{\boldsymbol{\beta}},\mathbf{x}))F_j(t,\hat{\boldsymbol{\beta}},\mathbf{x}))^2}$$

Les bornes de l'intervalle de confiance à 100(1-α)% pour la prédiction F_j sont ensuite obtenues par la formule :

$$F_j^{\exp\left(\pm z_{1-\alpha/2}\sqrt{\hat{Var}(log(-log(F_j(t,\boldsymbol{\beta},\mathbf{x}))))}\right)}$$

2 | Application dans le cadre des effets indésirables

A partir des notions théoriques détaillées dans le chapitre précédent, on se propose d'explorer l'utilisation du modèle de taux flexible dans le cadre de données d'effets indésirables. L'objectif peut être simplement descriptif ou d'associer les temps d'évènements avec des covariables d'intérêt comme par exemple comparer des traitements.

2.1 Stratégie de modélisation

Evènement

Comme on s'intéresse à la survenue d'un unique évènement par patient, on pourra définir le temps d'évènement de plusieurs manières en fonction de la question de recherche : temps d'observation du grade maximum, temps d'observation du premier évènement (tous grades) etc.

Evènements intercurrents

Dans notre contexte, nous pouvons identifier deux principales causes de sortie empêchant l'occurrence de l'évènement indésirable : le décès et l'arrêt du traitement. Notons que contrairement au décès, l'arrêt du traitement n'empêche pas complètement la survenue d'un évènement indésirable. Cependant, il peut être très difficile d'obtenir une collecte des données d'évènements indésirables après arrêt du traitement (Unkel et al., 2019) (changement de traitement, sortie d'étude...). L'estimateur de la probabilité d'incidence cumulée, calculée en l'absence de collecte de données après arrêt du traitement peut s'interpréter de deux manières :

- La probabilité d'incidence cumulée correspondrait à celle obtenue si les données avaient été collectées après le traitement, en supposant que les événements survenus après l'arrêt du traitement sont négligeables ;
- La probabilité d'incidence cumulée d'évènements en cours de traitement (*while on treatment estimand* selon l'addendum R1 aux ICH E9 (FDA, 2019)).

Afin de simplifier la modélisation, nous regrouperons les causes de sortie compétitives pour ne former qu'un seul évènement. Cela évite d'avoir des causes de sortie avec peu d'évènements (Ozenne et al., 2017).

Modèles

Deux modèles flexibles de taux (MFT) cause-spécifiques sont donc envisagés :

$$\{\log(h_{EI}(t,\mathbf{x}(t)))\}_{i=1..n} = \mathbf{X}_{EI}(t)\boldsymbol{\beta}_{EI}$$
$$\{\log(h_{D-AT}(t,\mathbf{x}(t)))\}_{i=1..n} = \mathbf{X}_{D-AT}(t)\boldsymbol{\beta}_{D-AT}$$

où, h_{EI} et h_{D-AT} sont respectivement les taux cause-spécifiques associés aux évènements indésirables et au décès/arrêt du traitement, $\boldsymbol{\beta}_{EI}$ et $\boldsymbol{\beta}_{D-AT}$ sont les vecteurs des paramètres associés aux taux d'évènements indésirables et de décès/arrêt du traitement, $\mathbf{X}_{EI}(t)$ et $\mathbf{X}_{D-AT}(t)$ sont les matrices de design des modèles d'EI et de D-AT respectivement, fonction des covariables $\mathbf{x}$ et du temps. Il est possible de prévoir des paramètres communs entre les différentes causes de sortie (Belot et al., 2010). Dans un contexte de régression sur des évènements indésirables, il y a peu de chances d'avoir des effets communs à la survenue de l'évènement indésirable et à l'arrêt du traitement et au décès. Nous considérons donc des vecteurs de paramètres distincts pour les deux taux cause-spécifiques.

Les noeuds des splines peuvent être placés au niveau des quantiles de temps d'évènements (et des valeurs des covariables s'il y en a). La modélisation est envisagée en présence ou non de pénalisation afin d'en évaluer le bénéfice. Le critère LAML discuté au chapitre précédent est utilisé pour estimer les paramètres de lissage.

Pour compléter les analyses, nous calculons également la probabilité d'incidence cumulée (en tenant compte des risques compétitifs comme détaillé en 1.4.2).

2.2 Etude de simulation

Dans ce chapitre, nous examinons, pour la première fois à notre connaissance, le comportement du modèle flexible dans un contexte de risques compétitif, en utilisant des données simulées basées sur l'observation d'effets indésirables chez des patients en oncologie. En effet, le modèle du taux flexible a été beaucoup étudié dans le cadre de données issues de registres de cancers qui présentent des différences majeures avec les données d'EI, notamment en termes de dynamique et de nombre d'évènements observés (Uhry et al., 2020).

L'objectif de cette simulation est d'évaluer la qualité d'ajustement des taux, des taux relatifs et des probabilités d'incidence cumulées (obtenus en tenant compte de la compétition). En particulier, l'intérêt est porté sur l'impact du nombre de noeuds et de la pénalisation sur la qualité d'ajustement.

2.2.1 Simuler des temps d'évènements dans un cadre de risques compétitifs

Pour simuler des temps d'évènements selon n'importe quelle loi dont on connaît la fonction de répartition F, on procède par la méthode d'inversion (Bender et al., 2005). Le principe de la méthode est de simuler une variable aléatoire selon une loi uniforme $\mathcal{U}[0,1]$ et de calculer $F^{-1}(U)$ qui suivra la distribution caractérisée par la fonction de répartition F.

Dans le cadre du modèle de survie, si T suit la loi définie par F, alors $F(T) \sim \mathcal{U}[0,1]$ mais également $S(T) \sim \mathcal{U}[0,1]$. Comme $S(t) = \exp(-\int_0^t h(s)ds)$, le temps de survie peut être

exprimé par :

$$T \sim \Lambda^{-1}(-\ln(U)),$$

en notant Λ le taux cumulé.

Une première approche consiste à simuler les temps d'évènements de chacune des causes et à considérer le minimum des deux temps. Le problème de cette approche est qu'il faut spécifier la structure de dépendance entre les temps d'évènements, difficilement appréhendable en pratique. Ce type de simulation est plus adapté lorsque l'on souhaite étudier des modèles à temps latents.

Une seconde approche, proposée par (Beyersmann et al., 2009), consiste à utiliser les taux cause-spécifiques. Les étapes de simulation sont les suivantes :

- Spécifier les taux d'évènements compétiteurs h_1 et h_2.
- Simuler les temps T à partir du taux de sortie toutes causes $h = h_1 + h_2$.
- Pour chaque temps T, on considère un tirage aléatoire de type Bernoulli de probabilité $h_1(T)/(h_1(T)+h_2(T))$ déterminant si la sortie est provoquée par la cause 1 (elle est donc de cause 2 dans le cas contraire).
- Il est éventuellement possible de simuler une variable C indépendante des temps simulés précédents pour appliquer une censure non-informative.

2.2.2 Scénarios de simulation

Design

Comme on s'intéresse au contexte spécifique des évènements indésirables, on souhaite simuler des jeux de données inspirés de situations réelles issues de patients suivis en oncologie. Dans cette étude, nous nous intéressons à une régression en présence d'une variable binaire x. Les jeux de données seront construits avec le même nombre de patients pour chaque modalité de x. Plusieurs tailles d'échantillons ont été considérées : 200, 400 et 1000 sujets. Pour chaque taille d'échantillon et chaque scénario, 500 échantillons ont été simulés.

Scénarios

Les taux cause-spécifiques théoriques en Figure 2.2 ont été obtenus à partir de données réelles en ajustant le modèle avec une spline non pénalisée pour le scénario 1 et des polynômes fractionnaires (Sauerbrei et al., 2007) pour le scénario 2. La question n'étant pas l'aspect médical des données ayant servi à inspirer les simulations, nous ne nous attardons pas sur leur description. Les modèles théoriques ont ensuite été utilisés pour simuler les données comme expliqué en 2.2.1. Dans le scénario 1, l'effet de x a été spécifié en proportionnel, contrairement au scénario 2.

La forme analytique des probabilités d'incidence cumulées n'étant pas disponible, il est possible de les obtenir par intégration numérique (Figure 2.3). La compétition est particulièrement importante sur le scénario 2 dans lequel le taux d'EI pour $x = 0$ est bien supérieur à celui pour $x = 1$, alors que les probabilités d'incidence cumulées sont confondues. La Figure 2.4 présente les ratios de taux théoriques obtenus pour les deux scénarios.

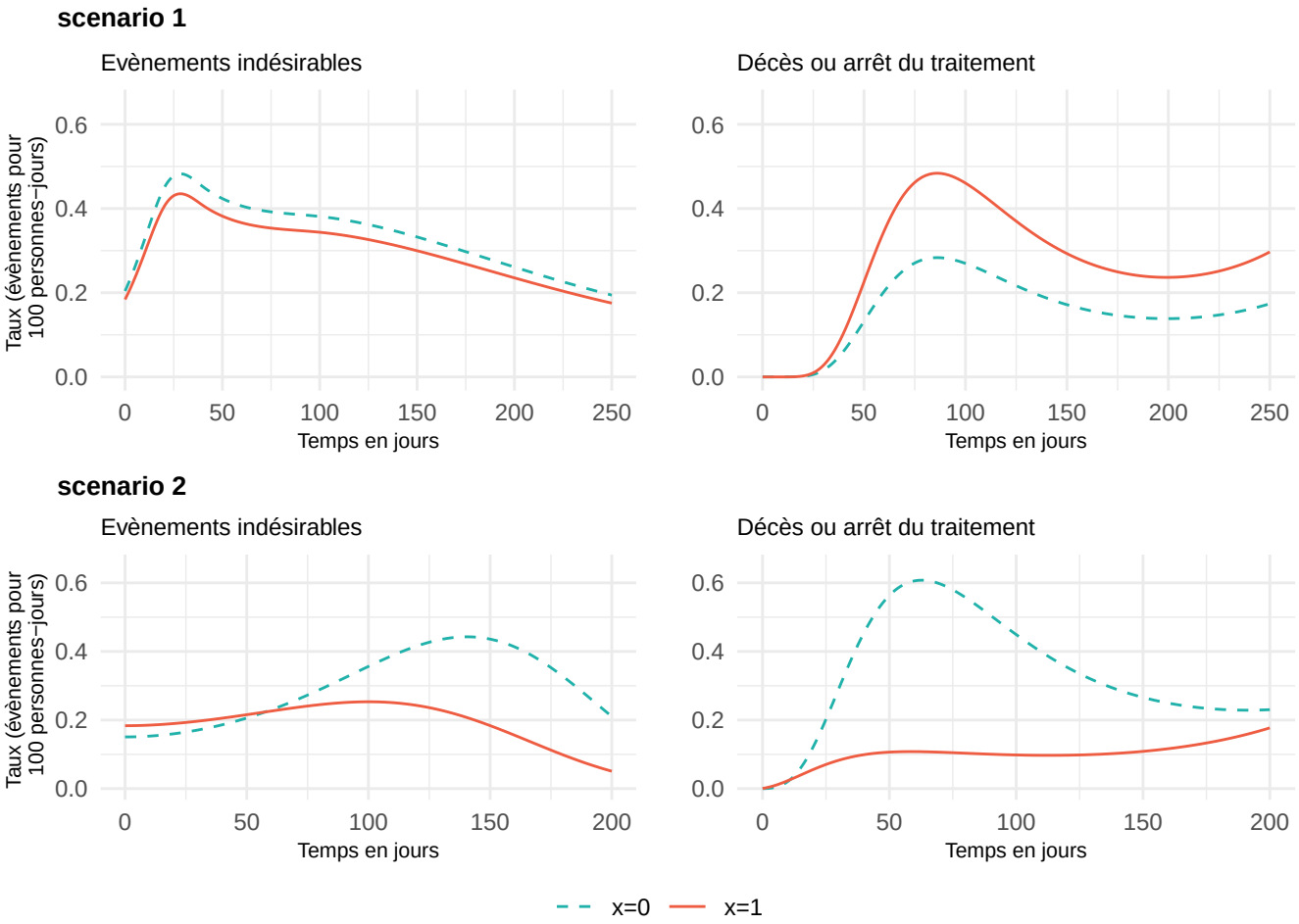

FIGURE 2.2 – Taux cause-spécifiques théoriques des deux scénarios de simulation

2.2.3 Modèles ajustés

Pour explorer le comportement du modèle, on considère plusieurs spécifications du MFT :

— Version proportionnelle

$$\log(h_{AE}(t)) = s_{AE,x=0}(t) + \beta_{AE}\ \mathbb{1}_{x=1}$$
$$\log(h_{D-AT}(t)) = s_{D-AT,x=0}(t) + \beta_{D-AT}\ \mathbb{1}_{x=1}$$

— Version non-proportionnelle

$$\log(h_{AE}(t)) = s_{AE,x=0}(t)\ \mathbb{1}_{x=0} + s_{AE,x=1}(t)\ \mathbb{1}_{x=1}$$
$$\log(h_{D-AT}(t)) = s_{D-AT,x=0}(t)\ \mathbb{1}_{x=0} + s_{D-AT,x=1}(t)\ \mathbb{1}_{x=1}$$

où, les fonctions $s()$ sont des splines.

Pour chaque scénario, on considère :

— (i) un MFT proportionnel pénalisé avec une spline à 10 noeuds ;
— (ii) un MFT non-proportionnel pénalisé avec une spline à 10 noeuds ;
— (iii) un MFT non-proportionnel non pénalisé avec une spline à 10 noeuds ;
— (iv) un MFT non-proportionnel non pénalisé avec une spline à 5 noeuds ;

Nous sommes également intéressés par la comparaison du modèle (i) et (ii) par critère AIC corrigé (AICc) (Wood et al., 2016). Le critère utilise une version corrigée du nombre de degrés

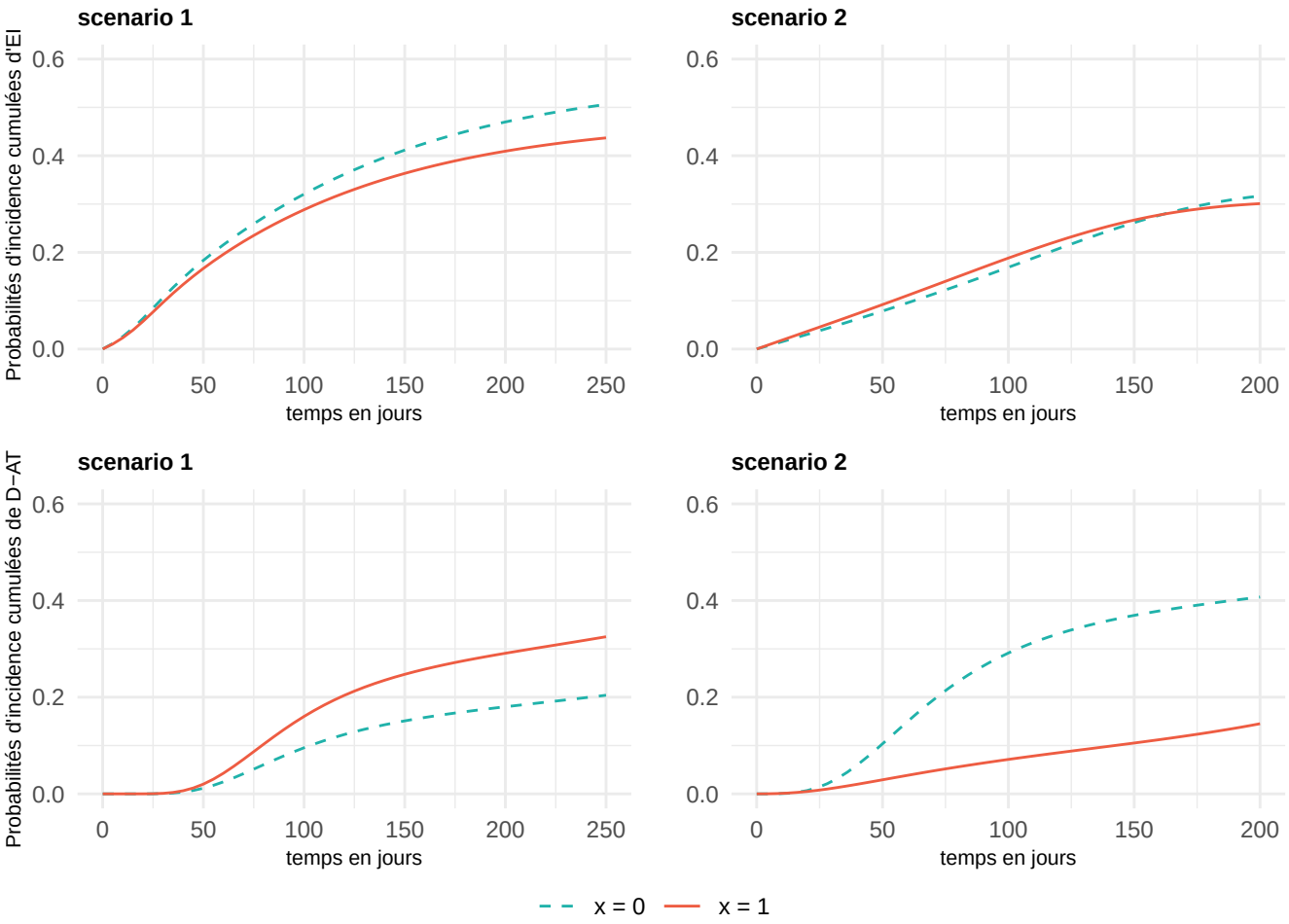

FIGURE 2.3 – Probabilités d'incidence cumulées théoriques d'EI (*haut*) et de D-AT (*bas*)

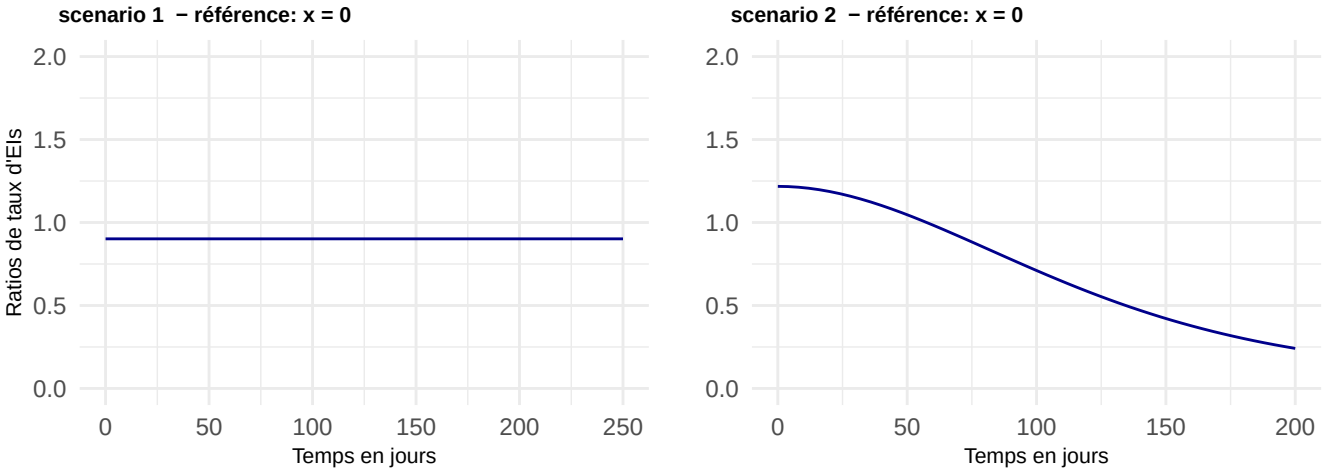

FIGURE 2.4 – Ratios de taux théoriques (entre x = 1 et x = 0)

de libertés, permettant de prendre en compte l'incertitude liée à l'estimation du paramètre de lissage. Le modèle le plus simple est retenu tant que l'écart d'AICc entre les deux modèles est inférieur à 2 (Burnham and Anderson, 2004).

2.2.4 Evaluation des performances des approches

Afin de comparer les différents modèles, nous estimerons les statistiques suivantes sur l'ensemble des échantillons simulés par scénario. On note S le nombre d'échantillons simulés pour chaque scénario, f la quantité d'intérêt théorique et $\hat{f}_s$ la quantité prédite sur l'échantillon s (ici le taux, le taux relatif ou la probabilité d'incidence cumulée).

— La racine de l'erreur quadratique moyenne (*Rooted Mean Squared Error* RMSE) évaluée pour en un temps t donné et pour une valeur de x.

$$RMSE(t,x) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(f(t,x) - \hat{f}_s(t,x)\right)^2} \tag{2.1}$$

— Le biais, évalué pour un temps t donné et un vecteur de covariables $\mathbf{x}$

$$biais(t,x) = \frac{1}{S}\sum_{s=1}^{S}\left(f(t,x) - \hat{f}_s(t,x)\right) \tag{2.2}$$

2.2.5 Résultats

	Scénario 1			Scénario 2		
Taille des échantillons	Méd.	Min.	Max.	Méd.	Min.	Max.
n = 200	95	75	117	62	42	82
n = 400	189	154	217	123	97	154
n = 1000	472	436	518	318	278	358

TABLE 2.1 – Description des échantillons simulés : Médiane, minimum et maximum d'EI simulés

La Table 2.1 présente quelques caractéristiques des échantillons simulés (médiane, minimum, maximum d'EI).

Taux d'évènements indésirables

Scénario 1 : Tous les modèles ont convergé pour le scénario 1, quelle que soit la taille de l'échantillon. On observe un biais important en début de suivi pour le MFT pénalisé (10 noeuds) en Figure 2.7. Le biais diminue très lentement avec la taille de l'échantillon. En regardant les prédictions par échantillon en Figure 2.5, on observe que le modèle pénalisé lisse systématiquement le début de suivi même pour les tailles d'échantillon de 1000 individus, ce qui empêche de décrire la forme en cloche et génère le biais. Sur les échantillons de taille 200, la forme de la courbe est très bruitée pour le modèle non pénalisé à 10 noeuds. En termes de RMSE en Figure 2.8, pour les tailles d'échantillons de 200 et 400, le modèle pénalisé est bien plus performant que les modèles non pénalisés sauf sur les premiers temps. Pour les échantillons de taille 1000, l'erreur du modèle pénalisé sur la partie en cloche de la courbe est toujours élevée en comparaison des modèles non pénalisés.

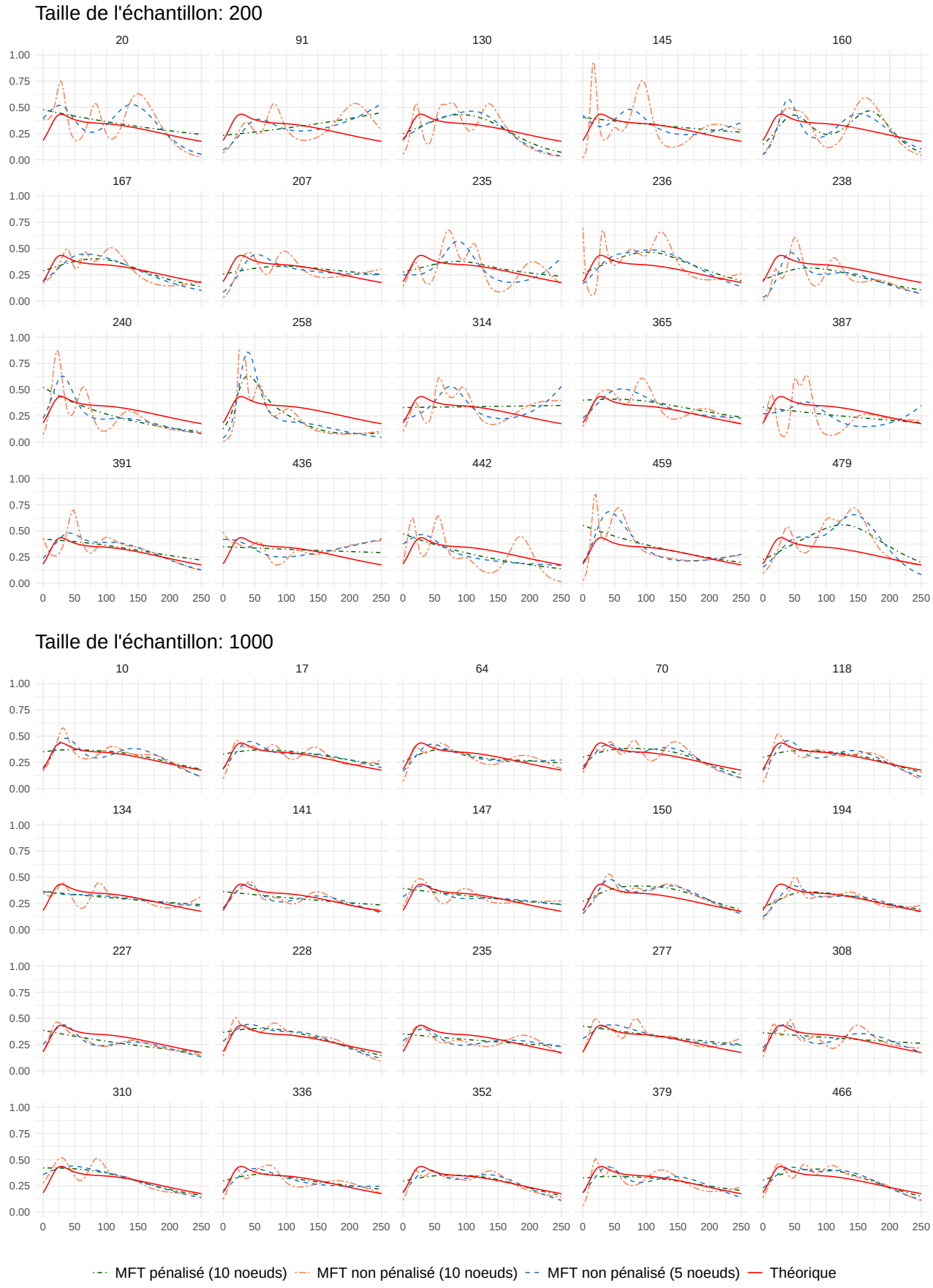

FIGURE 2.5 – Scénario 1 : Ajustement des taux pour $x = 1$ par échantillon. *Les échantillons présentés ont été sélectionnés de façon aléatoire parmi les 500 échantillons simulés. Pour des raisons de lisibilité le scénario à 400 sujets n'est pas présenté.*

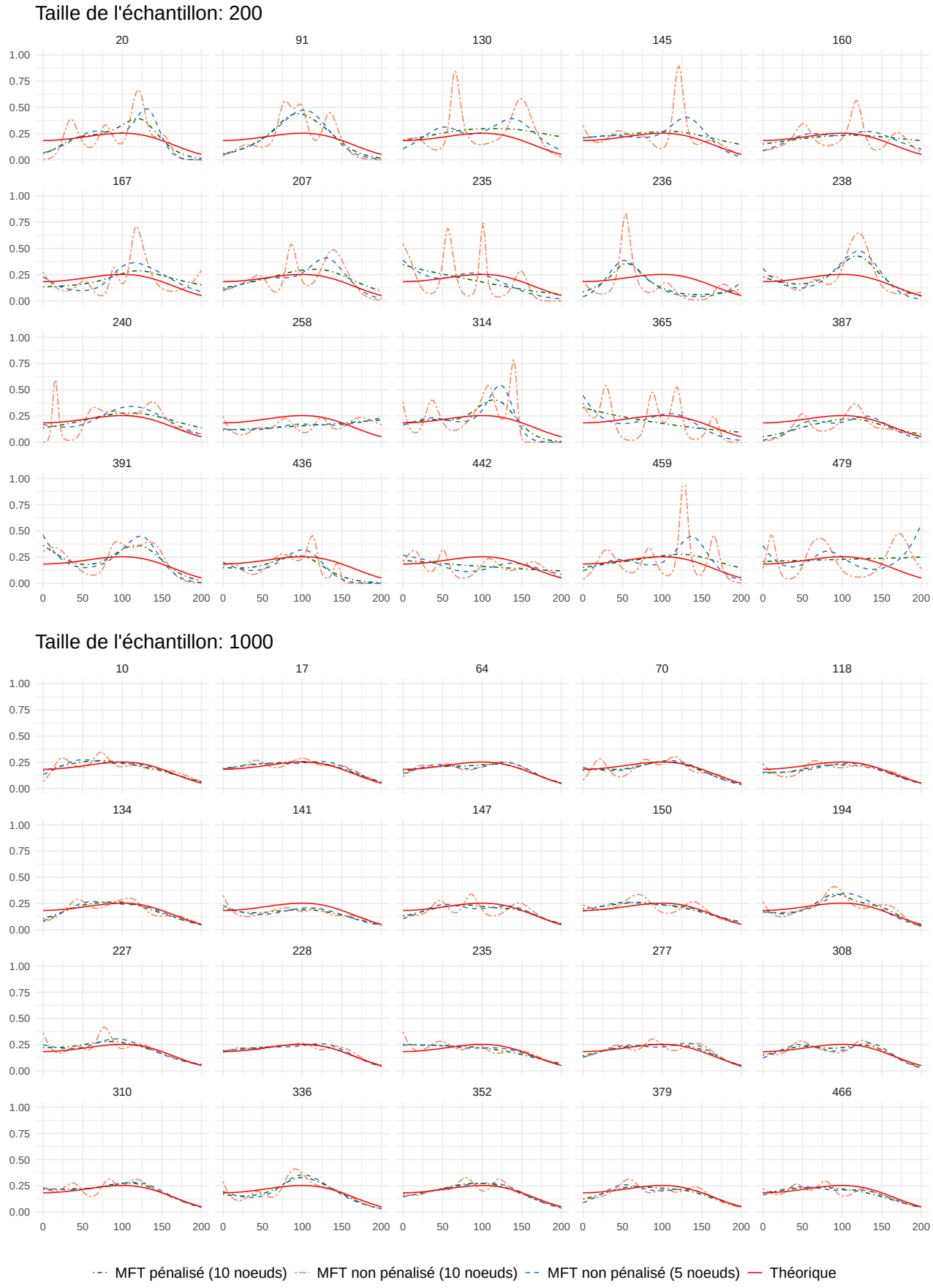

FIGURE 2.6 – Scénario 2 : Ajustement des taux pour $x = 1$ par échantillon. *Les échantillons présentés ont été sélectionnés de façon aléatoire parmi les 500 échantillons simulés*

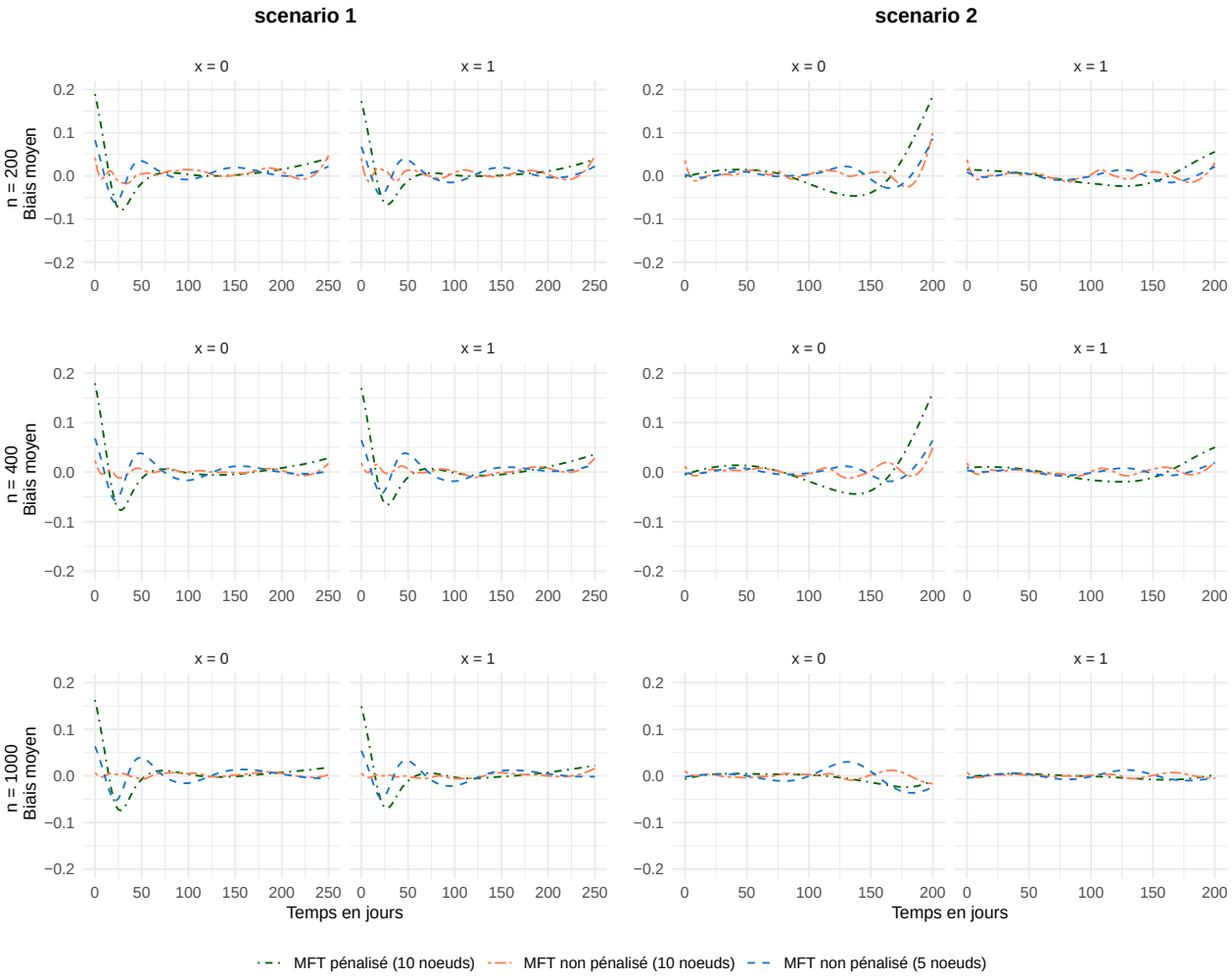

FIGURE 2.7 – Biais sur l'estimation du taux

Scénario 2 : Dans ce scénario, 5 modèles non pénalisés avec la spline à 10 noeuds sur les 500 ajustés ont échoué à converger pour les échantillons de taille 200. Le modèle pénalisé présente un biais sur la fin de la période de temps. La forme en cloche n'est pas systématiquement captée et remplacée par une tendance linéaire croissante. Sur les échantillons de taille 200, le bruit est très important sur le modèle non pénalisé à 10 noeuds, il est donc difficile de dégager une tendance. En termes de RMSE, le modèle pénalisé a toujours une meilleure performance.

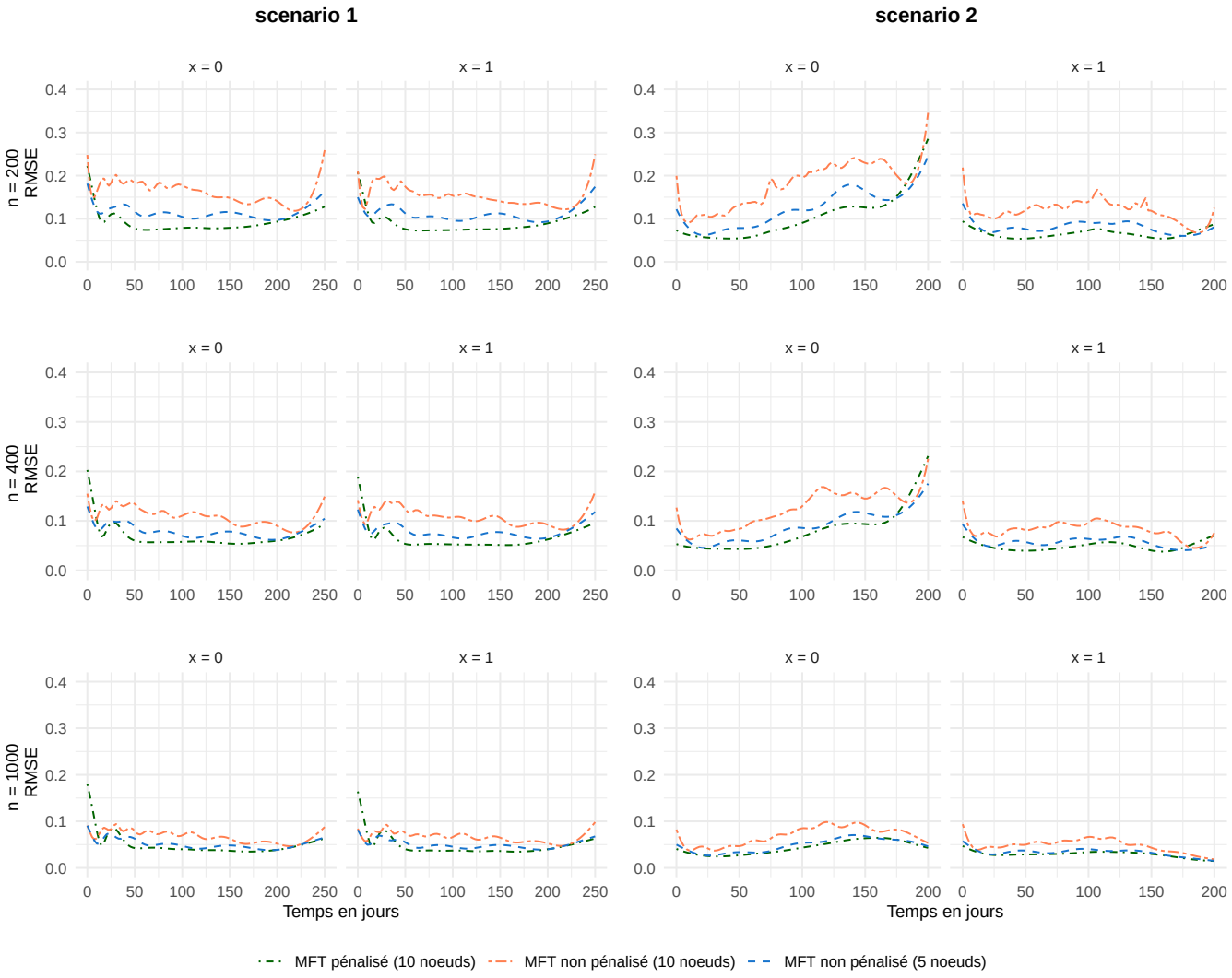

FIGURE 2.8 – RMSE sur l'estimation du taux

Taux relatifs

Dans les deux scénarios, sur les échantillons de taille 200, les taux relatifs estimés par le MFT non pénalisé avec 10 noeuds sont très instables. On observe des biais très importants au début et en fin de la période de temps en Figure 2.9. Le modèle pénalisé est le plus performant en termes de biais, quel que soit le scénario et la taille d'échantillon. Le graphique des RMSE étant très semblable à celui du biais, il n'est pas présenté ici. Sur les prédictions par échantillons en Figures 2.10 et 2.11, la forme des taux relatifs est très variable et parfois très éloignée de la forme théorique pour les petits échantillons (n = 200).

Dans le scénario 1, la comparaison du modèle proportionnel et non proportionnel par critère AICc (avec une différence de 2 pour être en faveur du modèle le plus complexe) conduit à conserver le modèle non proportionnel dans respectivement 9.2%, 10.2% et 9.6% des cas pour les tailles d'échantillons de 200, 400 et 1000. Dans le scénario 2, le modèle non proportionnel est favorisé dans respectivement 39.6%, 57,8% et 91.4% pour les tailles d'échantillons de 200, 400 et 1000.

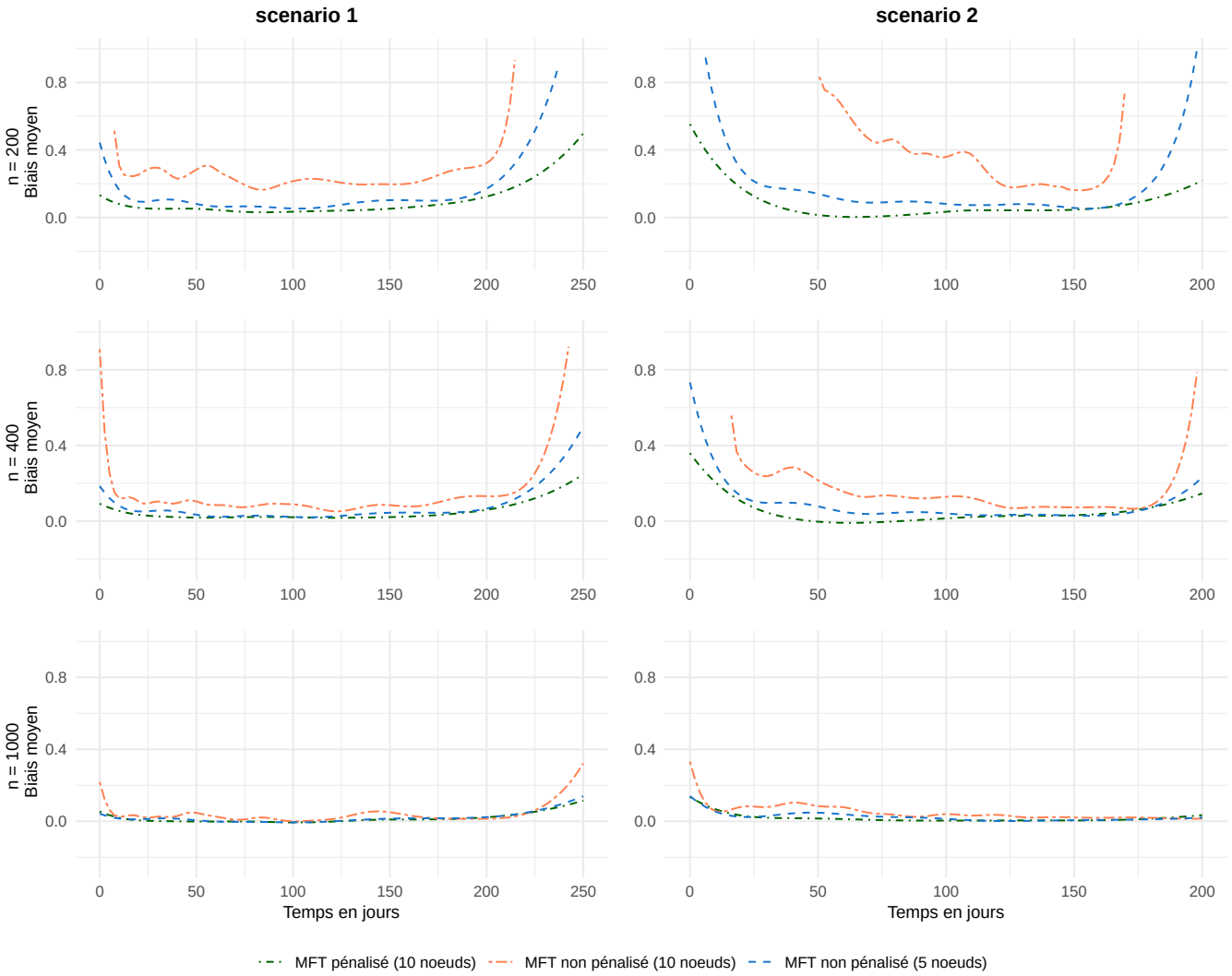

FIGURE 2.9 – Biais sur l'estimation des taux relatifs. *(entre $x = 1$ et $x = 0$)*

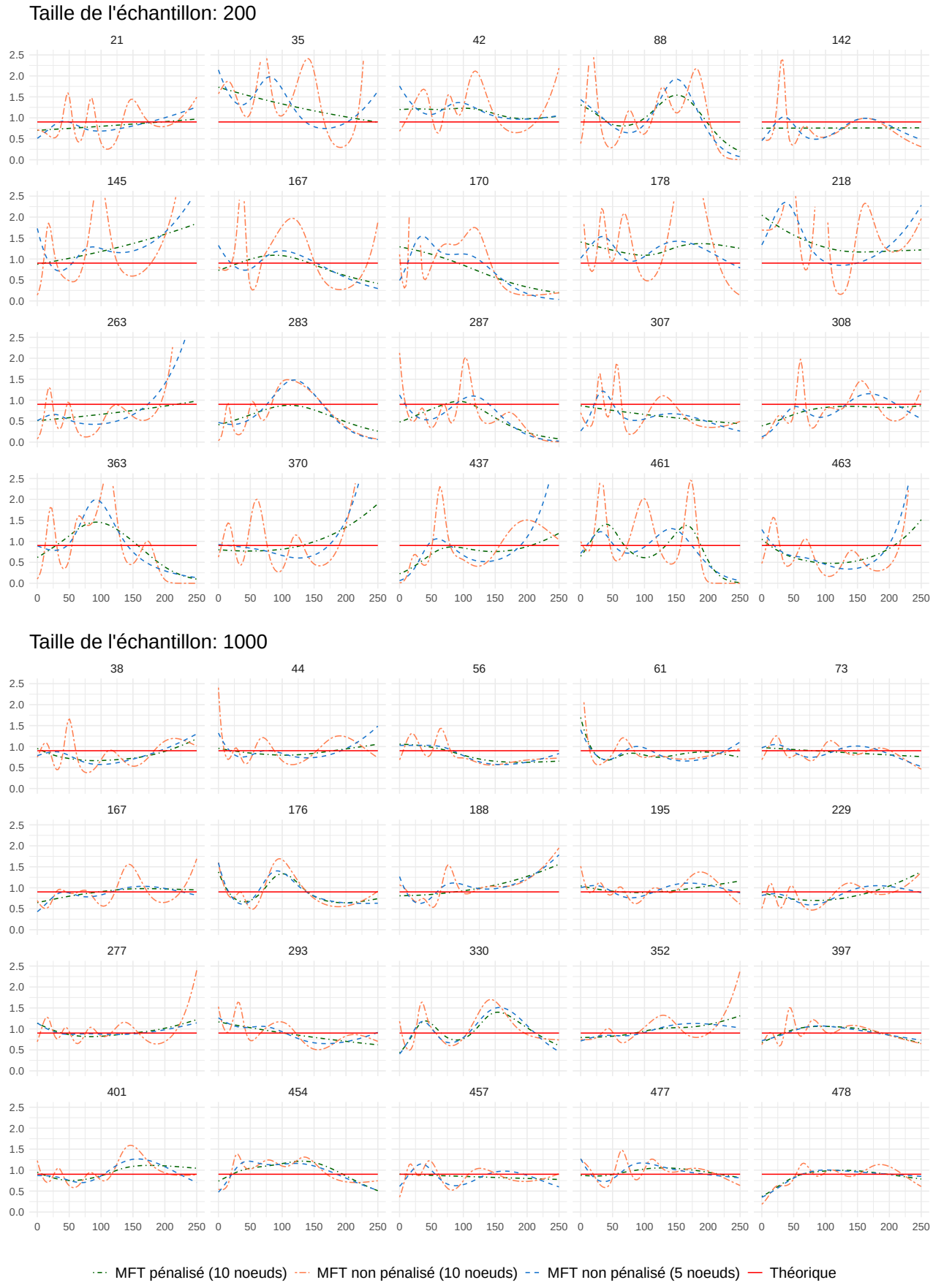

FIGURE 2.10 – Scénario 1 : Ajustement des taux relatifs par échantillon (entre x = 1 et x = 0). *Les échantillons ont été sélectionnés de façon aléatoire parmi les 500 simulés*

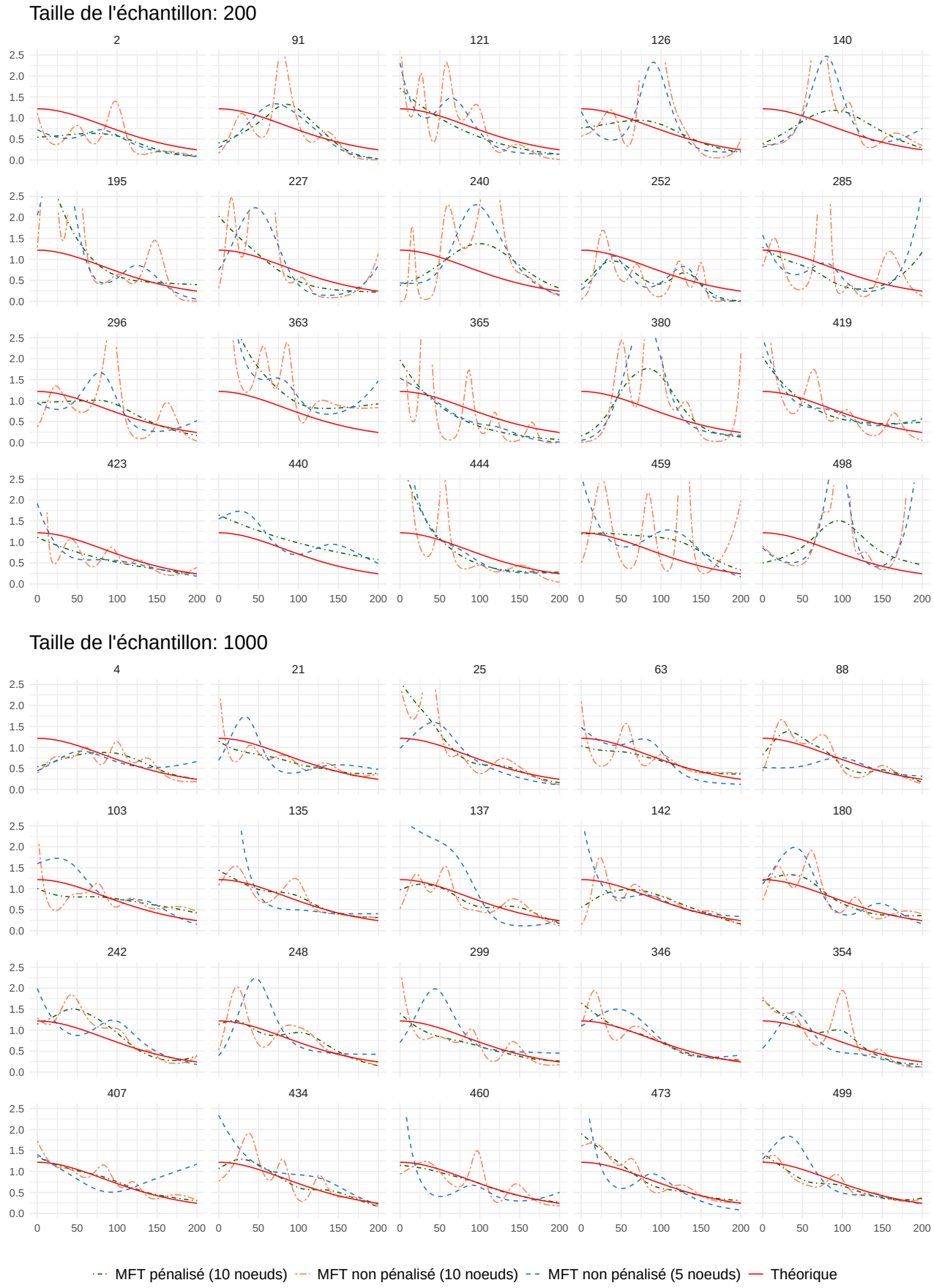

FIGURE 2.11 – Scénario 2 : Ajustement des taux relatifs par échantillon (entre x = 1 et x = 0). *Les échantillons ont été sélectionnés de façon aléatoire parmi les 500 simulés*

Probabilités d'incidence cumulées

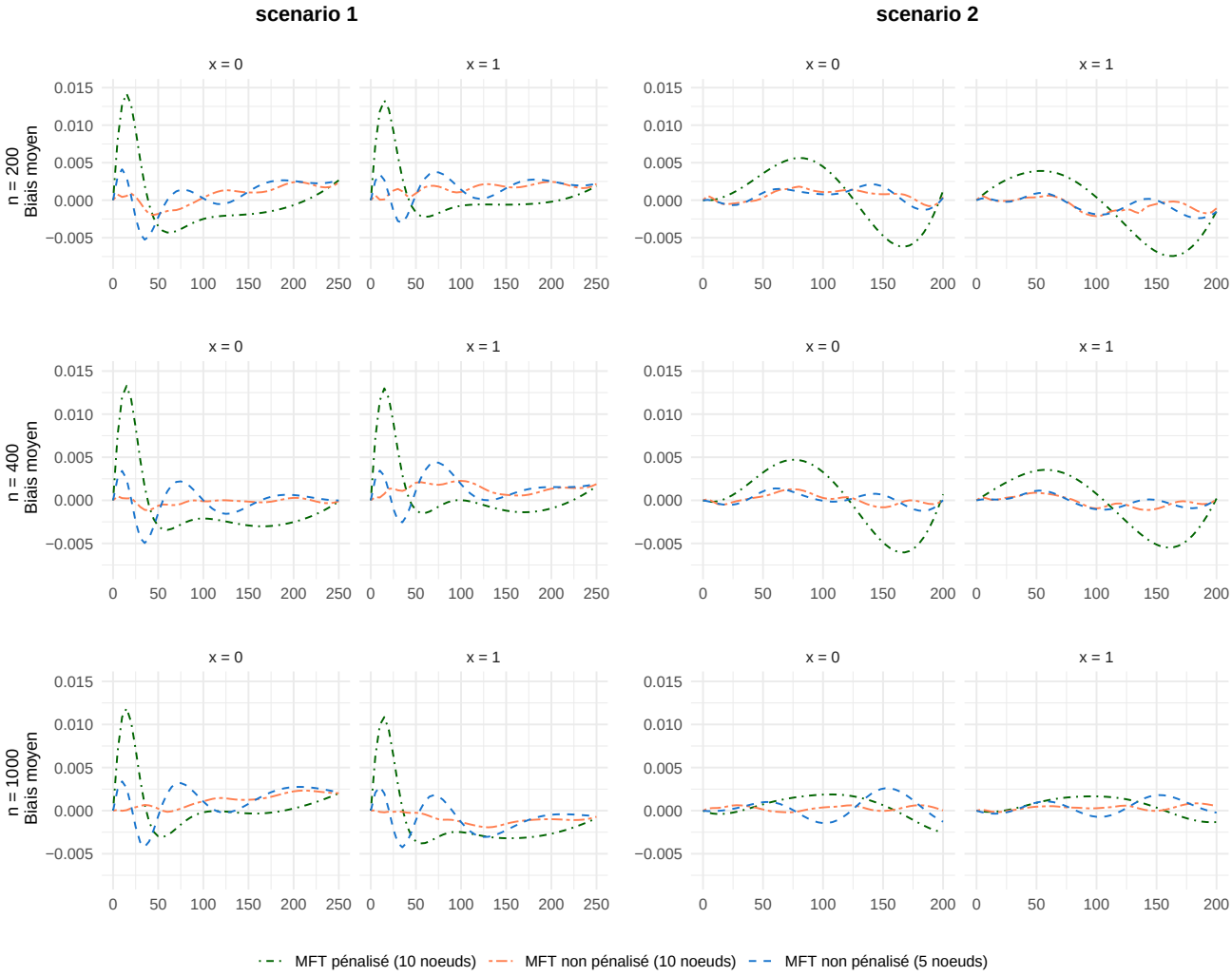

FIGURE 2.12 – Biais sur l'estimation des probabilités d'incidence cumulée d'EI

Scénario 1 : D'une manière générale, le biais observé est faible pour l'ensemble des modèles (1%). Le biais généré par le lissage trop important du modèle pénalisé en début de la période temps se reporte sur les probabilités d'incidence cumulées sur toutes les tailles d'échantillon et en particulier pour une taille d'échantillons de 1000 2.12. La forme de la courbe est cependant beaucoup plus stable que celle du taux et reste proche du théorique, même pour les petits échantillons (Figure 2.13).

Scénario 2 : Le modèle pénalisé est toujours meilleur en termes de RMSE quelle que soit la taille d'échantillons (Figure 2.15).

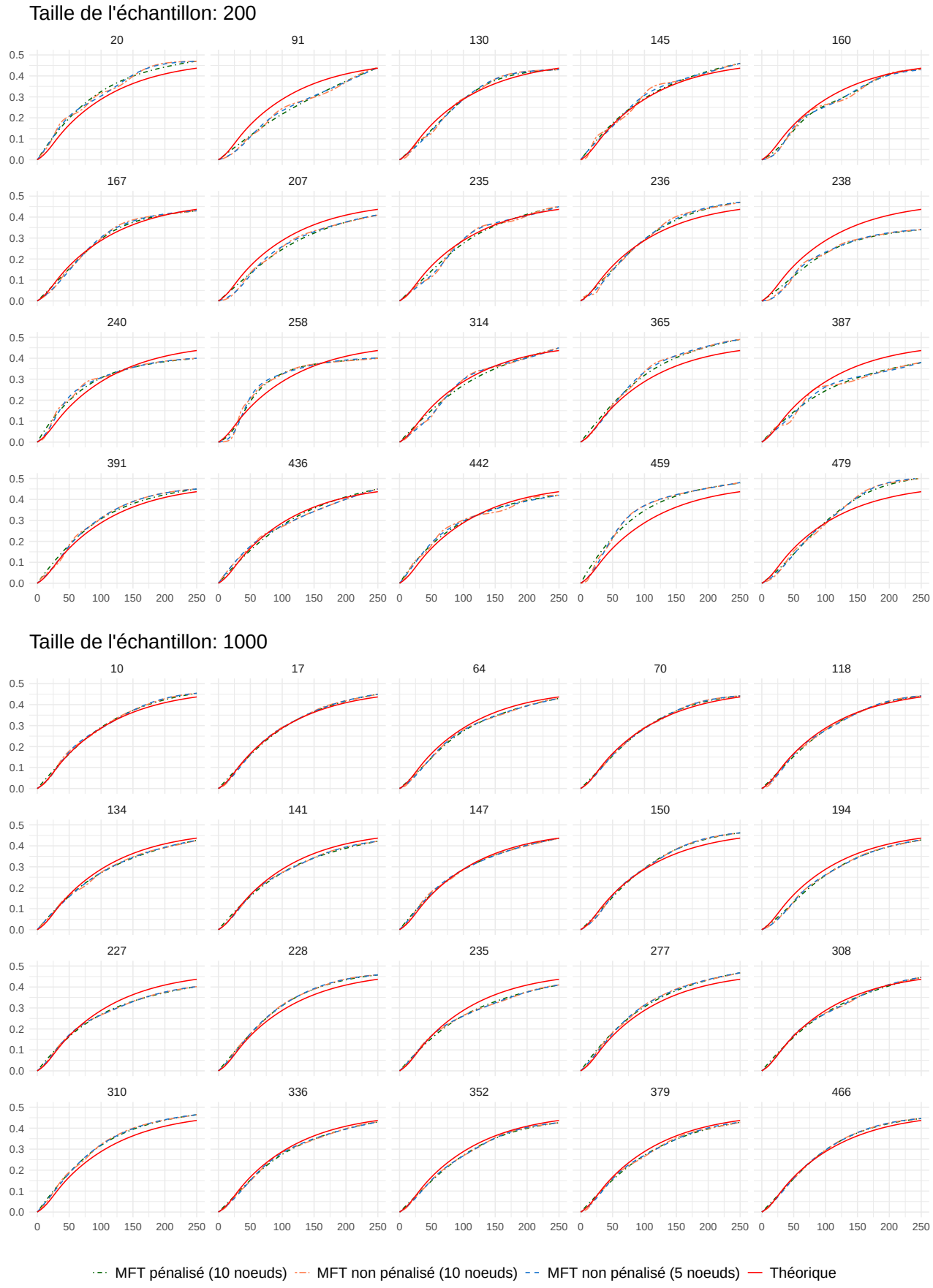

FIGURE 2.13 – Scénario 1 : Ajustement des taux relatifs par échantillon (pour $x = 1$). *Les échantillons ont été sélectionnés de façon aléatoire parmi les 500 simulés*

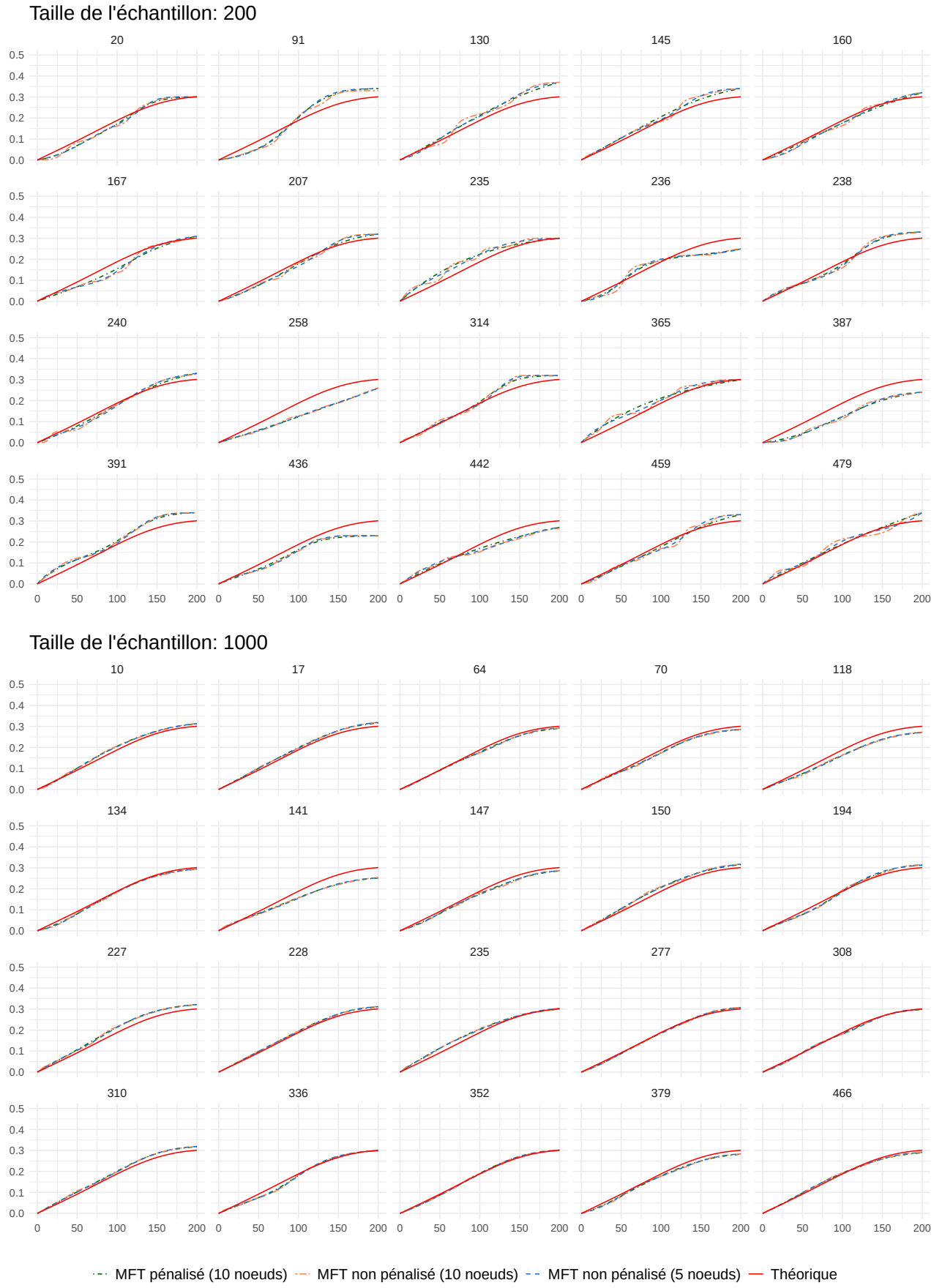

FIGURE 2.14 – Scénario 2 : Ajustement des taux relatifs par échantillon (pour $x = 1$). *Les échantillons ont été sélectionnés de façon aléatoire parmi les 500 simulés*

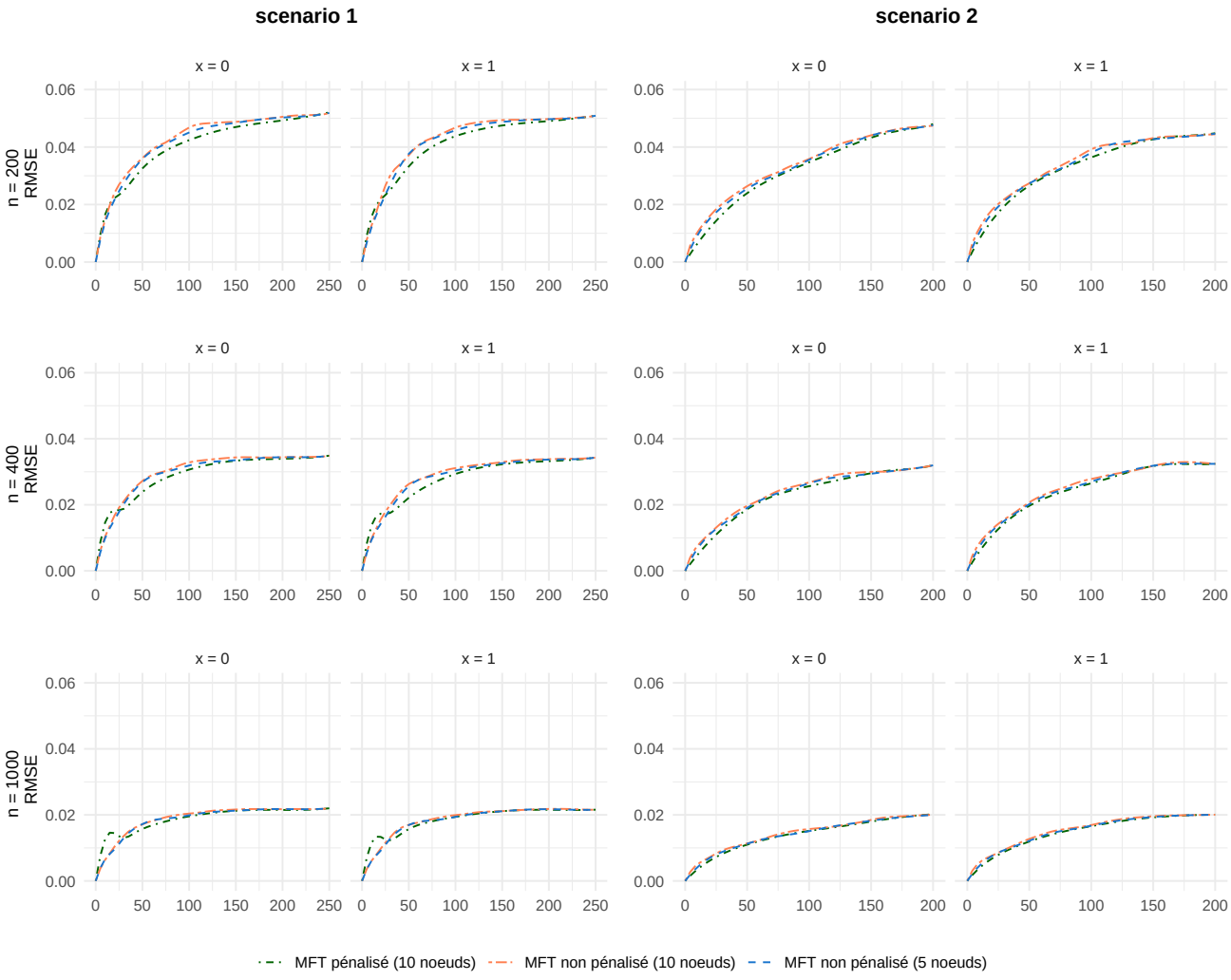

FIGURE 2.15 – Courbes des RMSE d'incidence cumulées d'EI

2.3 Application : Evènements indésirables dans la cohorte des mélanomes

2.3.1 Contexte et objectif

Nous nous intéressons aux EI observés dans la cohorte des mélanomes décrite en partie I section 1.2. Pour commencer, on se propose de décrire le taux des évènements les plus fréquemment associés à l'immunothérapie en utilisant le MFT.

2.3.2 Méthode

Nous considérons le temps au premier évènement tous grades confondus. Un modèle flexible du taux avec le temps comme seule variable est ajusté sur les évènements indésirables les plus fréquents. On considère deux modèles : un MFT pénalisé avec une spline à 10 noeuds pour le temps et un MFT non pénalisé avec une spline à 5 noeuds pour le temps. Les noeuds ont été placés sur les quantiles des temps d'évènements. Cependant, lorsque les valeurs de deux quantiles étaient trop proches ($<= 1$ jour d'écart), l'un des deux noeuds a été retiré pour éviter des problèmes de convergence.

2.3.3 Résultats

La Figure 2.16 présente les taux d'évènements indésirables les plus fréquemment observés (hors fatigue et anomalies biologiques). Dans l'ensemble le modèle pénalisé et le modèle non pénalisé proposent des dynamiques du taux similaires. On observe un écart sur le délai d'apparition des dyspnées. Le modèle non pénalisé propose un léger délai sur leur apparition par rapport au début du traitement, contrairement au modèle pénalisé qui lisse complètement la courbe. Ce cas de figure pourrait être similaire à celui du scénario 1 de l'étude de simulation. Les pentes sont également plus marquées sur la survenue d'évènements à l'initiation du traitement pour les nausées, les diarrhées et les maux de tête. Le modèle pénalisé propose deux modes pour l'hyperthyroïdie, contrairement au modèle non pénalisé. Cela pourrait s'expliquer par un manque de noeuds dans le second. Les deux modes coïncident avec les visites. L'hyperthyroïdie sous immunothérapie est souvent asymptomatique et donc détectée lors des examens biologiques réalisés avant chaque prise de traitement. Dans la plupart des cas, l'hypothyroïdie et l'hyperthyroïdie sont la manifestation d'une même pathologie : une thyroïdite médiée par les cellules T cytotoxiques contre la glande thyroïdienne. L'hyperthyroïdie est donc un évènement transcient, souvent asymptotique et pouvant donc être manqué (Barroso-Sousa et al., 2018). En effet, nous voyons sur les graphiques qu'elle survient sur une fenêtre très courte. La distribution des évènements de type prurit et rash maculopapuleux ne semble pas présenter de mode bien marqué.

La Figure 2.17 présente les taux d'anomalies biologiques recueillies au cours du suivi. Une censure par intervalle est clairement observée sur l'augmentation des CPK par exemples, les modes de la distribution coïncidant avec des dates de visites. Dans ce cas, le modèlé pénalisé et le modèle non pénalisé fournissent des résultats plutôt différents. Le nombre de noeuds pourrait ne pas être suffisant dans le modèle non pénalisé pour capter la dynamique complexe observée du fait de la censure par intervalles.

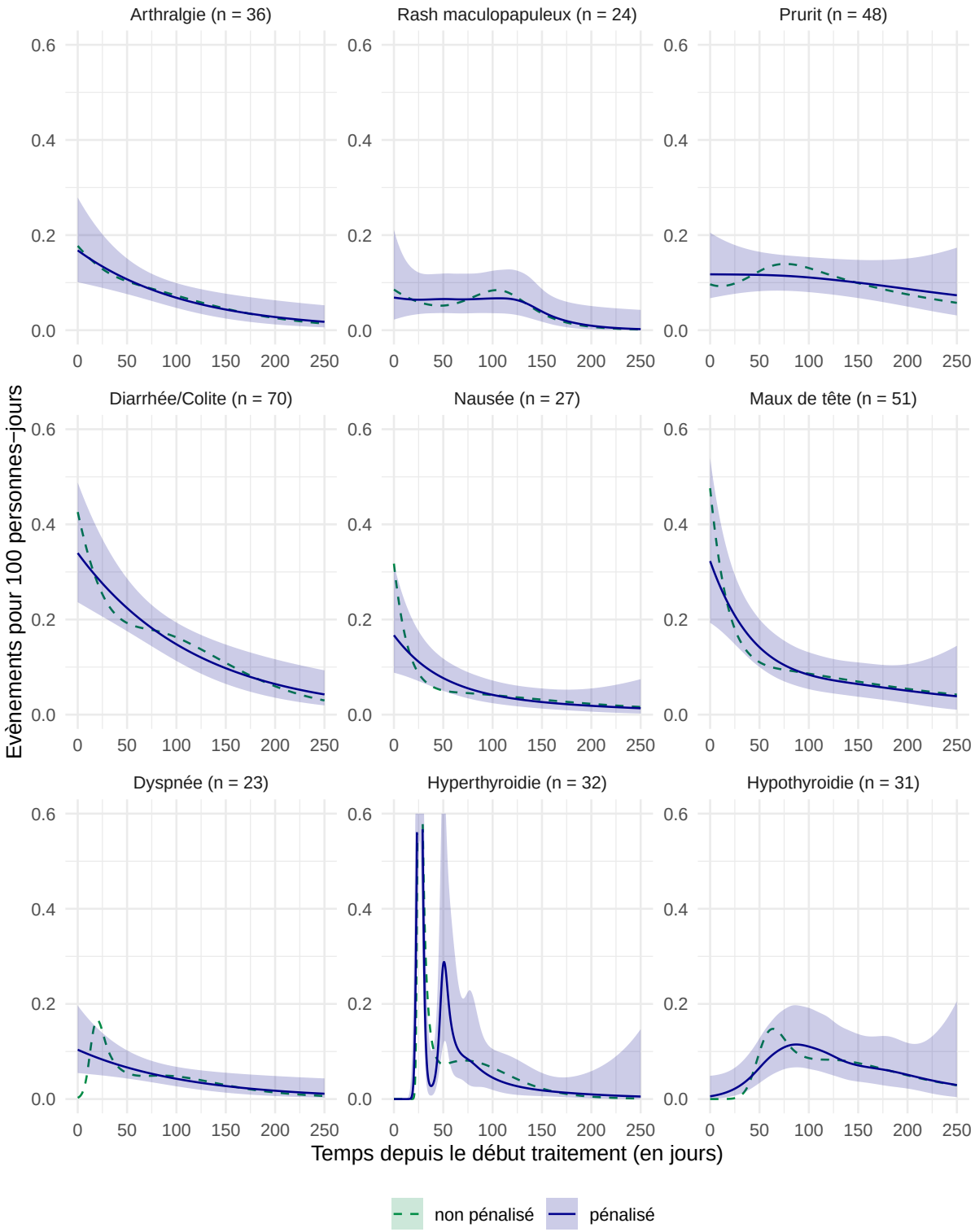

FIGURE 2.16 – Taux d'évènements indésirables

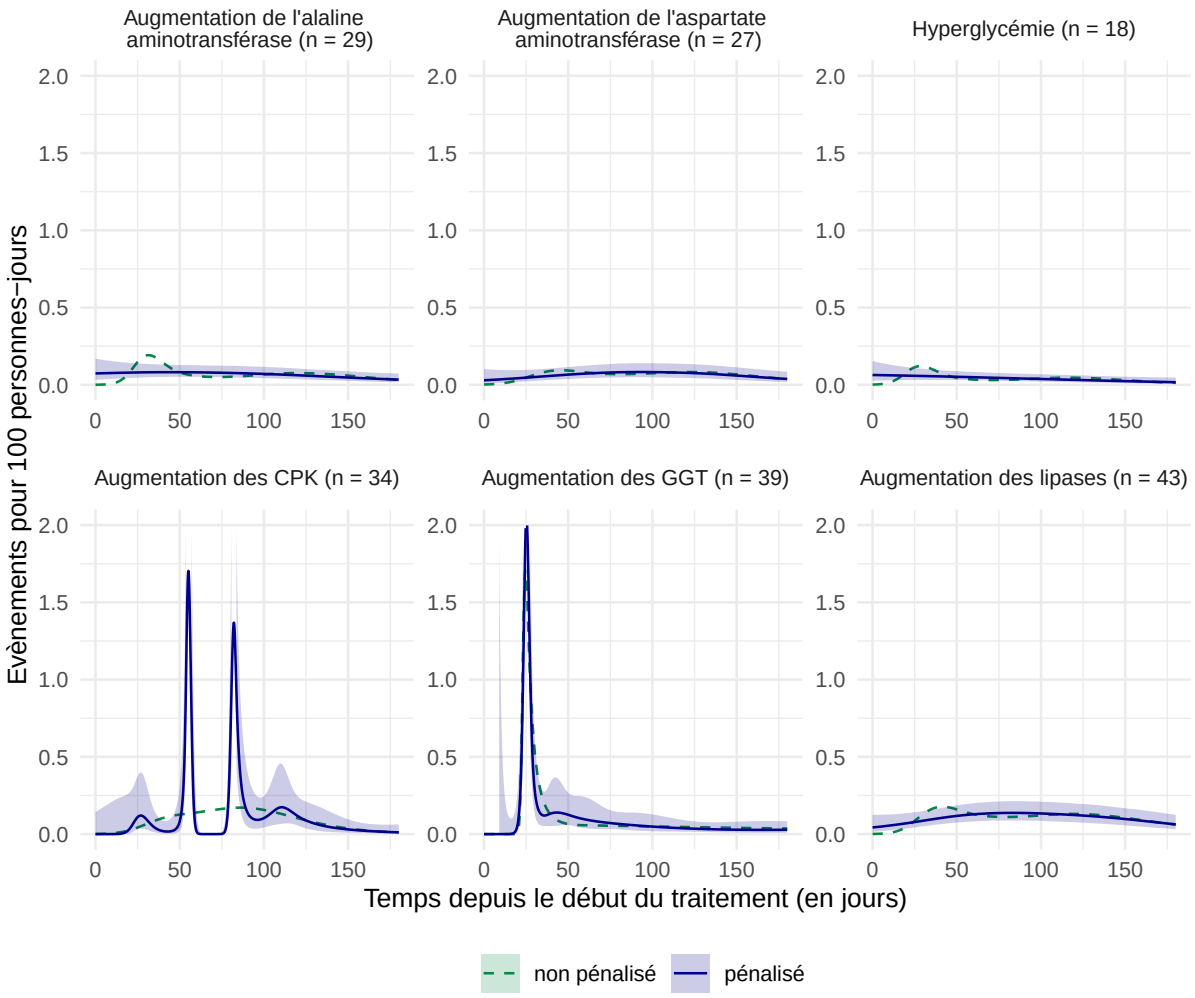

FIGURE 2.17 – Taux d'évènements indésirables : examens biologiques

2.4 Application : Evènements indésirables dermatologiques dans la cohorte des mélanomes

2.4.1 Contexte et objectif

Nous nous intéressons aux évènements dermatologiques survenant dans la cohorte des patients atteints de mélanomes décrite en section 1.2. On souhaite identifier des prédicteurs de ces évènements. Les évènements dermatologiques font partie des évènements indésirables les plus fréquents sous immunothérapie. Parmi les plus fréquents, on compte les rash maculopapuleux, le prurit, le vitiligo, la pemphigoïde bulleuse.

Nous explorons l'association entre les évènements dermatologiques et les covariables suivantes : l'âge au début du traitement, le sexe, le ratio neutrophiles sur lymphocytes, la présence de métastases distantes, la présence d'un historique de maladie auto-immune (MAI) chez le patient et le pays. Le choix des variables repose sur la littérature. En particulier, le ratio neutrophiles sur lymphocytes (NLR) est défini par le compte de neutrophiles et de lymphocytes qui sont des marqueurs de réponse inflammatoire. Un NLR élevé en début de traitement est souvent associé avec un mauvais pronostic (Guthrie et al., 2013), et notamment dans le cas de l'immunothérapie (Bilen et al., 2019; Lalani et al., 2018; Wu et al., 2022). En revanche, son association avec la survenue d'EI n'a pas été systématiquement établie (Fukihara et al., 2019; Owen et al., 2018; Nakanishi et al., 2019; Lee et al., 2021; Eun et al., 2019). Un effet sur les D-AT est anticipé, on adopte donc la stratégie de modélisation proposée en section 2.1.

2.4.2 Méthode

Plusieurs modèles sont envisagés pour construire les modèles de taux pénalisé sur les EI et les D-AT : (i) sans effet non proportionnel, (ii) avec un effet non proportionnel de l'âge, (iii) avec un effet non proportionnel du NLR :

— *Modèle 1 :* $log(h(t)) = \beta_0 + s(t) + s(NLR) + s(age) + \beta_{sexe} I(sexe = "Femme") + \beta_{métastases} I(Métastases = "Oui") + \beta_{MAI} I(MAI = "Oui")$

— *Modèle 2 :* $log(h(t)) = \beta_0 + tensor(t, age) + s(NLR) + \beta_{sexe} I(sexe = "Femme") + \beta_{métastases} I(Métastases = "Oui") + \beta_{MAI} I(MAI = "Oui")$

— *Modèle 3 :* $log(h(t)) = \beta_0 + tensor(t, NLR) + s(age) + \beta_{sexe} I(sexe = "Femme") + \beta_{métastases} I(Métastases = "Oui") + \beta_{MAI} I(MAI = "Oui")$

Les modèles sont ensuite comparés en utilisant un critère AIC corrigé (AICc) (Wood et al., 2016). Les splines sont spécifiées avec 5 noeuds répartis aux quantiles des temps d'évènements et des valeurs de l'échantillon pour l'âge et le NLR.

2.4.3 Résultats

Le modèle 1, sans interaction, est retenu pour les EI (AICc respectifs : 1543.6, 1546.0 et 1545.7) et pour les décès/arrêts de traitement (AICc respectifs : 617.7, 620.3 et 618.3).

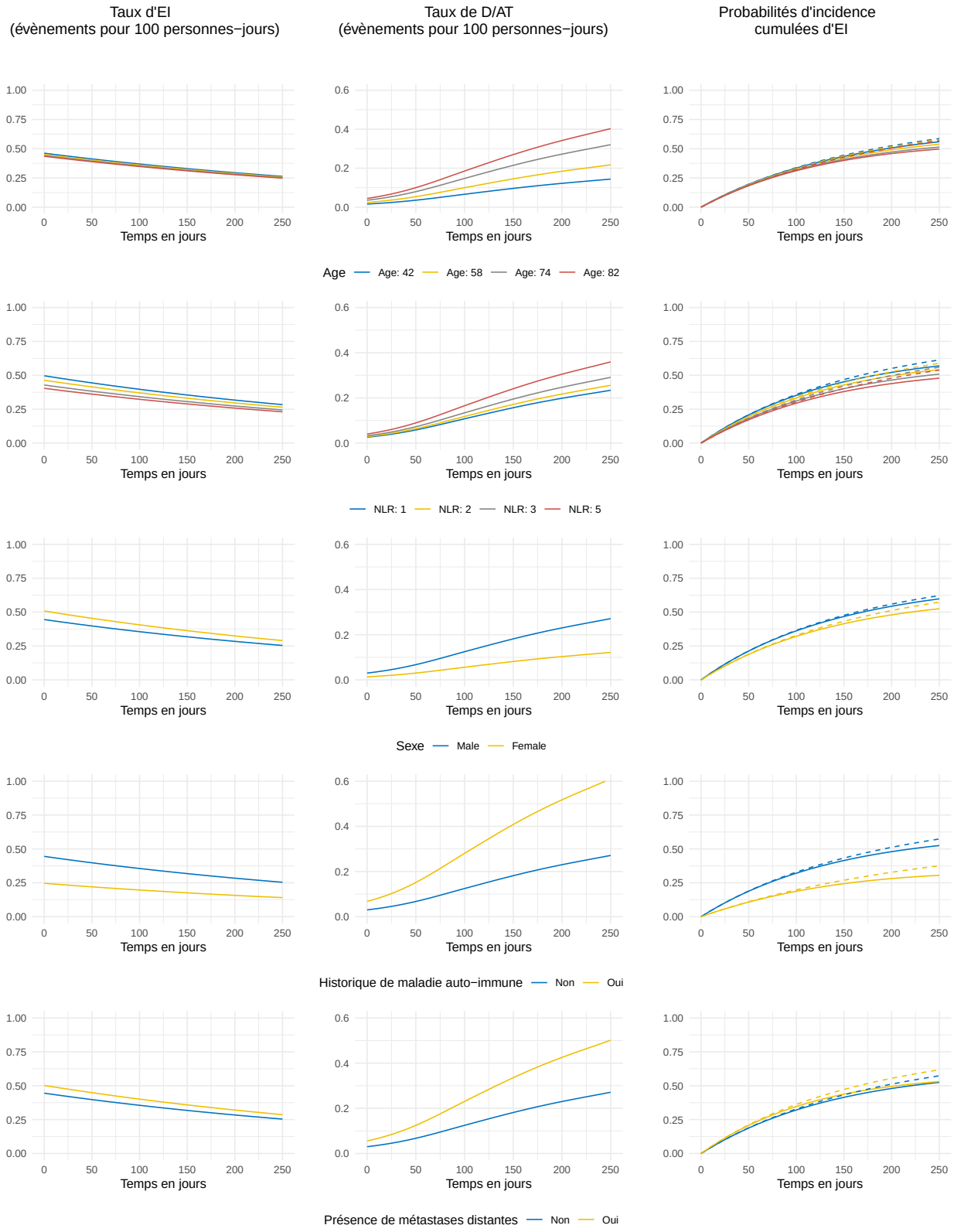

FIGURE 2.18 – Effets des covariables sur (*gauche*) le taux d'évènements indésirables, (*milieu*) le taux de décès ou d'arrêt de traitement, (*droite*) la probabilité cumulée d'évènements indésirables. *A titre indicatif, la probabilité cumulée d'évènements obtenue sans tenir compte de la compétition est tracée en pointillés.*

La Figure 2.19 présente les taux relatifs estimés pour les variables avec un effet proportionnel. Pour les EI, aucun effet significatif n'est observé à part pour le pays. Pour le D/AT, les femmes présentent un taux plus faible que les hommes.

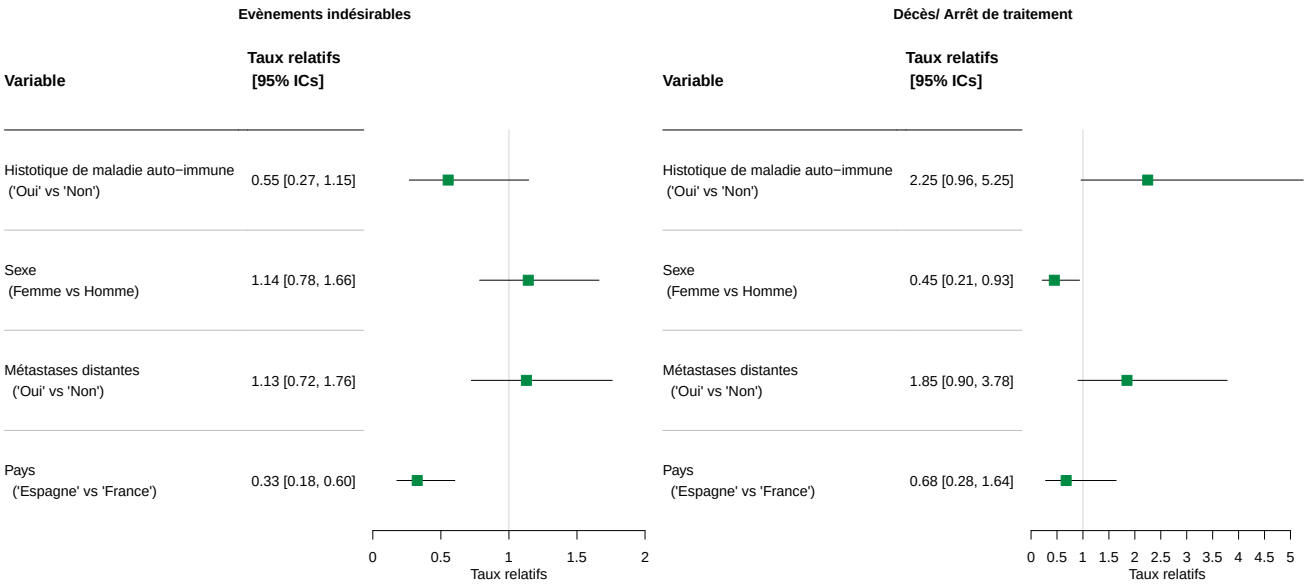

FIGURE 2.19 – Taux relatifs des variables avec effet proportionnel(*gauche*) sur le taux d'EI, (*droite*) sur le taux de D-AT

La Figure 2.18 présente les effets des différentes covariables sur les taux d'EI, les taux de décès et d'arrêt du traitement et la probabilité cumulée d'incidence des EIs. L'âge et le NLR ne semblent pas impacter le taux d'EI. Un taux de D/AT plus élevé est en revanche observé pour les âges et les NLR élevés. La compétition a en revanche un effet modeste sur la probabilité d'incidence cumulée, quelle que soit la variable considérée, car la majorité des évènements surviennent tôt dans le suivi.

3 | Discussion & Conclusion

Au cours de ces travaux, nous avons exploré l'utilisation de modèles de taux flexibles pour décrire des données de toxicité, en menant une étude de simulation qui compare les versions pénalisée et non pénalisée du modèle, ainsi que des cas pratiques sur des données de patients traités par immunothérapie.

Performance du modèle

Dans les simulations, le modèle pénalisé surpasse presque toujours les modèles non pénalisés en termes de RMSE. La prédiction du taux fournie par le modèle révèle une tendance globale, bien que sa précision soit limitée dans les petits échantillons. Nous avons vu que le modèle pénalisé tendait à trop lisser certaines parties de la courbe, notamment les formes en cloches proches des temps initiaux. Le modèle non pénalisé à 10 noeuds donne des résultats trop bruités pour les niveaux d'information explorés dans cette étude. En revanche, le modèle à 5 noeuds permet de décrire le phénomène de cloche masqué par le lissage. Ajuster le modèle à 5 noeuds en complément du modèle pénalisé à 10 noeuds, comme test de sensibilité, peut donc représenter une option intéressante. Le lissage adaptatif pourrait également être exploré, permettant de prendre en compte le manque d'homogénéité du taux d'évènements au cours du temps (Wood, 2017).

La forme des taux relatifs semble très instable et l'estimation est peu informative pour les petits échantillons (n=200). Il doit donc être considéré avec précautions dans ce contexte. L'estimation de la probabilité d'incidence dans le contexte compétitif est très correcte même sur les petits échantillons et ce malgré la variabilité observée sur le taux.

Pertinence dans un contexte d'évènements indésirables

Le modèle propose une représentation de la dynamique d'occurrence des évènements au cours du temps, souvent rarement proposée dans les analyses d'EI. Il permet de prendre en compte une éventuelle censure (liée par exemple à la date de point), qui peut être présente lorsque l'on étudie des évènements à long terme. Pour obtenir une estimation correcte de la dynamique du taux, le nombre d'évènements doit cependant être suffisant, ce qui peut donc être limitant dans les analyses d'évènements indésirables pour décrire par exemple les évènements rares mais graves (e.g. grades 3-4).

En décrivant les EI de la cohorte des mélanomes, nous avons pu constater que les temps d'évènements sont censurés par intervalle. Pour certains évènements, comme l'hyperthyroïdie, l'information peut même être tronquée. Considérer un modèle censuré par intervalle pourrait être une solution (Gentleman and Geyer, 1994). Cependant, dans le contexte des données observationnelles, définir ces intervalles n'est pas toujours simple. La précision du temps d'événement

peut varier en fonction du type d'événement et de sa gravité. Pour les événements identifiés par des analyses biologiques réalisées régulièrement (par exemple, l'alanine aminotransférase et l'aspartate aminotransférase contrôlées à chaque administration du traitement sous immunothérapie), les dates des intervalles sont facilement identifiables. En revanche, ce ne sera pas le cas pour les anomalies biologiques détectées en fonction de la symptomatologie. Certains événements graves, nécessitant une hospitalisation ou une consultation en urgence, auront une date bien définie. En ce qui concerne la date de début d'un symptôme (par exemple, la nausée), elle peut être rapportée de manière plus ou moins précise par le patient lors de la consultation avec le clinicien. Cette diversité de situations complique l'application du modèle censuré par intervalle, surtout si le modèle est utilisé pour du descriptif.

Le modèle ne considère qu'un unique évènement. Il est courant de définir le temps d'évènement par le temps jusqu'au grade maximum ou le temps au premier évènement. Cependant, lorsque l'évènement d'intérêt est récurrent, cela conduit à une perte d'information. Dans ce cas d'autres modèles pourraient être considérés. Ce sera l'objet des développements de la troisième partie de cette thèse.

Partie III

Les modèles flexibles sur l'intensité marginale dans un contexte d'évènements récurrents

Comme discuté dans la première partie, la prise en compte de la récurrence des effets indésirables est un aspect négligé des analyses de toxicité. Dans la partie précédente, nous avons présenté une approche épidémiologique de l'analyse des effets indésirables à savoir, décrire leur occurrence au cours du temps en fonction de certaines caractéristiques des patients. Dans ce cadre, nous étions intéressés par la description du taux de survenue du premier évènement au cours du temps. Ce modèle ne permet donc pas de prendre en compte un caractère éventuellement récurrent du phénomène. Nous avons déjà évoqué quelques modèles de régression pour les évènements récurrents dans la première partie de cette thèse. Ces derniers sont des extensions du modèle de Cox, et ils ne permettent pas de décrire le risque absolu d'évènements au cours du temps. L'objectif de cette partie sera donc de proposer un modèle permettant de décrire le risque absolu d'évènements au cours du temps, ainsi que les effets potentiellement non-linéaires et non proportionnels des covariables.

1 Notions théoriques

1.1 Processus de comptage

1.1.1 Définitions

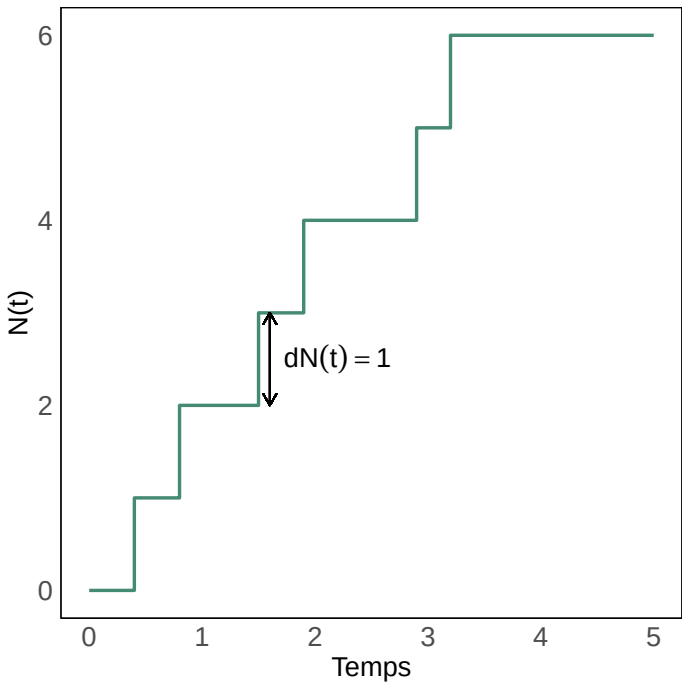

FIGURE 3.1 – Processus de comptage

Une succession d'évènements peut naturellement être décrite par un processus de comptage $N(t)$ (Figure 3.1), c'est-à-dire un processus constant par morceaux, continu à droite, non-

décroissant et de pas égaux à 1, prenant la valeur 0 à l'origine. A un instant donné t, $N(t)$ s'interprète ainsi comme le nombre cumulé d'évènements survenus jusqu'à cet instant. Les temps d'évènements auxquels nous nous intéresserons sont les "sauts" de ce processus. Par exemple, $N(t)$ peut représenter le nombre cumulé d'évènements indésirables observés entre le début du traitement et un instant t. Chaque saut correspond à la survenue d'un nouvel évènement. En notant $T_1, T_2...$ ces temps d'évènements, T_k étant le k^e temps d'évènement, on peut donc écrire :

$$N(t) = \sum_{k=1}^{K} I(T_k \leq t) \tag{1.1}$$

Nous considérons le temps comme une variable continue et supposerons donc que deux évènements ne peuvent pas survenir au même moment. On notera $N(s,t) = N(t) - N(s)$ le nombre d'évènements survenus sur l'intervalle $(s,t]$.

Comme dans le cadre à un unique évènement, un processus de comptage peut être sujet à censure. On définit donc un processus d'observation $Y(t)$ valant 1 lorsque le processus est observé et 0 sinon.

Intensité du processus

Pour caractériser un processus, nous aurons besoin de son historique $H(t)$, c'est-à-dire la filtration naturelle à laquelle $N(t)$ est adaptée. A l'instar du taux lorsque l'on observe un unique évènement, on peut définir l'intensité i.e. la probabilité d'évènement par unité de temps conditionnellement à l'historique du processus :

$$\lambda(t) = \lim_{dt \to 0} \frac{P(dN(t) = 1 | H(t))}{dt} \tag{1.2}$$

où, $dN(t) = N(t + dt) - N(t)$.

L'intensité permet de spécifier complètement le processus et sera notamment indispensable pour définir une vraisemblance. Au cours de la thèse, nous serons amenés à appeler cette quantité intensité conditionnelle pour la distinguer de l'intensité marginale qui sera définie par la suite.

Densité

Pour la modélisation, nous aurons besoin de la loi du processus, notamment pour construire une vraisemblance. Nous observons le processus $N(t)$ d'intensité $\lambda(t|H(t))$ sur un intervalle de temps $[\tau_0, \tau]$. Nous définissons l'évènement : "m évènements, observés aux temps $t_1, t_2, ..., t_m$, sachant $H(\tau_0)$". La densité associée à cet évènement peut alors s'écrire :

$$f(t_1, t_2, ..., t_m | H(\tau_0)) = \prod_{k=1}^{m} \lambda(t_k | H(t_k)). \exp\left(- \int_{\tau_0}^{\tau} \lambda(s | H(s)) ds \right) \tag{1.3}$$

Pour une démonstration de ce résultat, on pourra consulter Cook and Lawless 2007 (p 28-30).

Indicateurs marginaux

Il est également possible de décrire le processus par des quantités dites *marginales*, c'est-à-dire pour lesquelles on ne conditionne pas sur l'historique. On peut ainsi définir la moyenne d'évènements cumulés :

$$\mu(t) = \mathbb{E}(N(t)) \tag{1.4}$$

Une autre quantité, qui sera notre indicateur d'intérêt dans le cadre de cette thèse, est appelé *rate* dans la littérature (Lin et al., 2000; Cook and Lawless, 2007) :

$$\rho(t) = \lim_{dt \to 0} \frac{P(dN(t) = 1)}{dt} \tag{1.5}$$

Il s'agit d'une intensité marginale (IM) et nous en ferons référence sous ce nom dans la suite. L'IM n'a pas d'interprétation individuelle, elle est une moyenne d'intensités pour l'ensemble des trajectoires d'évènements possibles dans la population. En l'absence d'évènement terminal pour le processus (c'est-à-dire des évènement interrompant le processus comme un décès), on a une relation directe entre l'IM et la moyenne cumulée d'évènements :

$$\mu(t) = \int_0^t \rho(s) ds \tag{1.6}$$

1.1.2 Quelques exemples

Avant d'aller plus loin, commençons par illustrer les quantités introduites : intensité, intensité marginale (IM), moyenne cumulée sur deux cas particuliers de processus : le processus de Poisson et le processus de type multi-états.

Le processus de Poisson

Un cas particulier très important de processus de comptage est le processus de Poisson. Il peut être défini par la propriété d'indépendance du nombre d'évènements sur des intervalles disjoints :

$$\text{si } s_2 < s_3 \text{ alors } N(s_1, s_2) \text{ et } N(s_3, s_4) \text{ sont indépendants.}$$

En termes d'intensité, cela se traduit par une indépendance à l'historique du processus, ainsi :

$$\begin{aligned} \lambda(t) &= \lim_{dt \to 0} \frac{P(dN(t) = 1 | H(t))}{dt} \\ &= \lim_{dt \to 0} \frac{P(dN(t) = 1)}{dt} \\ &= \rho(t). \end{aligned} \tag{1.7}$$

L'intensité et l'IM sont égaux dans ce cas particulier. Lorsque ρ est constant, le processus est dit homogène, et il est qualifié de non-homogène dans le cas contraire.

Le lien avec la distribution de Poisson provient de la propriété suivante :

$N(s,t)$ suit une distribution de Poisson de moyenne $\mu(s,t) = \mu(t) - \mu(s)$ pour $0 \leq s \leq t$.

La Figure 3.2 présente un exemple de fonction d'intensité associée à un processus de Poisson, ainsi que la moyenne cumulée d'incidence associée à ce processus. Les fonctions d'IM et d'intensité sont confondues.

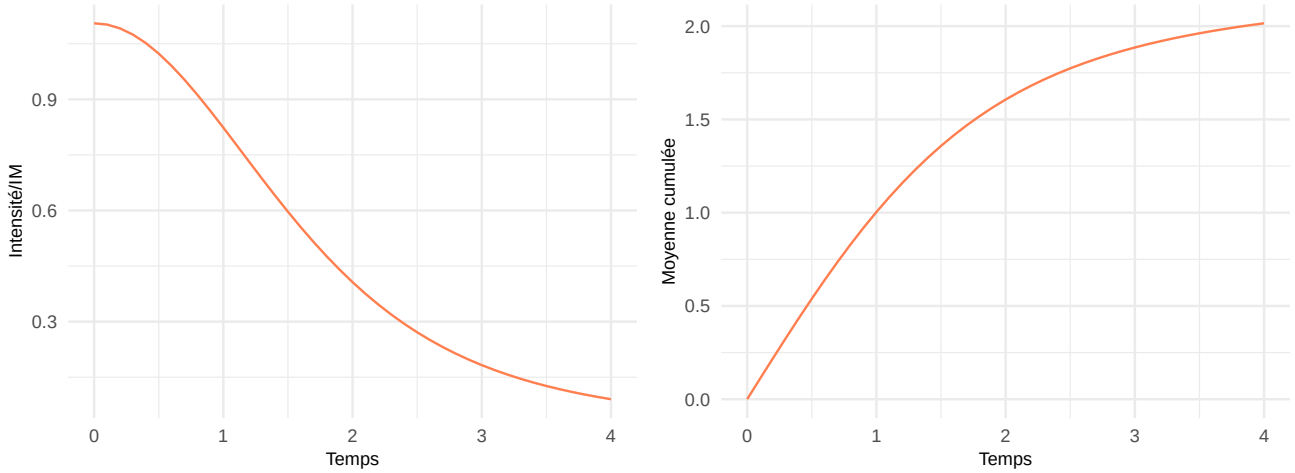

FIGURE 3.2 – Quantités d'intérêt d'un processus de Poisson *(gauche)* Fonction d'intensité et d'IM qui sont confondues pour le processus de Poisson. La fonction d'intensité est identique pour tous les individus. *(droite)* Moyenne cumulée d'évènements au cours du temps.

Processus de type multi-états

La récurrence d'évènements peut être vue comme un modèle multi-états, le sujet restant dans un état jusqu'à survenue d'un nouvel évènement, comme schématisé en Figure 3.3.

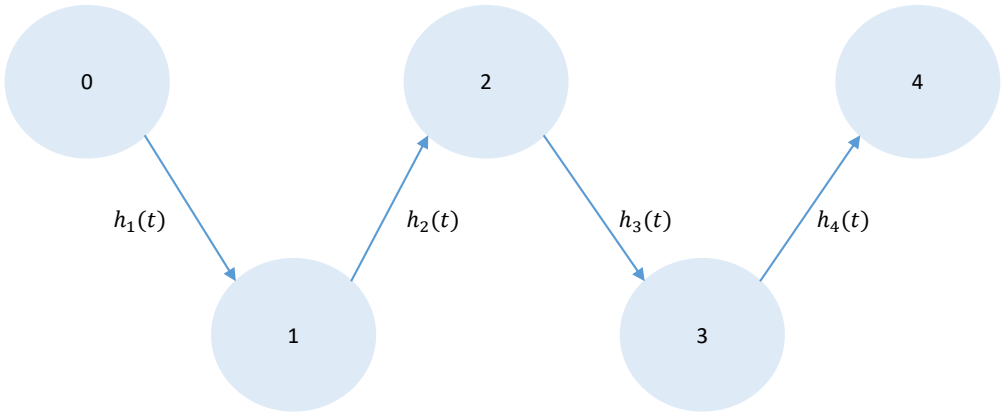

FIGURE 3.3 – Représentation du processus récurrent multi-états

Sous cette formulation, on définit des taux de transition, c'est-à-dire des probabilités de passage d'un état à un autre par unité de temps. Selon la définition de l'origine du temps considérée, il est possible de les écrire de deux manières différentes.

— L'origine est le début du suivi, que nous appellerons processus multi-états type "temps calendaires" :

$$h_k(t) = \lim_{dt \to 0} \frac{P(t \leq T_k < t + dt | T_k \geq t, N(t) = k - 1)}{du}$$

— L'origine est le temps du précédent l'évènement, que nous appellerons processus multi-

états type "inter-temps" :

$$h_k(w) = \lim_{dw \to 0} \frac{P(w \leq W_k < w + dw | W_k \geq w, N(t) = k-1)}{dw}$$

où, $W_k = T_k - T_{k-1}$

Inter-temps

Il est également possible de décrire un processus de comptage, non pas à partir de ses temps d'évènements mais de ses inter-temps (*gap-times*) $W_1, W_2, ..., W_m$ tels que : $T_K = \sum_{k=1}^{K} W_k$. Cette formulation sera très utile quand nous simulerons un processus de comptage.

Lorsque les inter-temps sont indépendants et suivent la même distribution, on dit que c'est un processus de renouvellement (*Renewal process*). On définit alors les taux d'inter-temps :

$$h_k(w) = \lim_{dw \to 0} \frac{P(w \leq W_k < w + dw | W_k \geq w)}{dw}, \tag{1.8}$$

L'origine $w = 0$ correspond au temps de survenue de l'évènement précédent. Lorsque la distribution des inter-temps est exponentielle, le processus est un processus de Poisson homogène (Cook and Lawless, 2007).

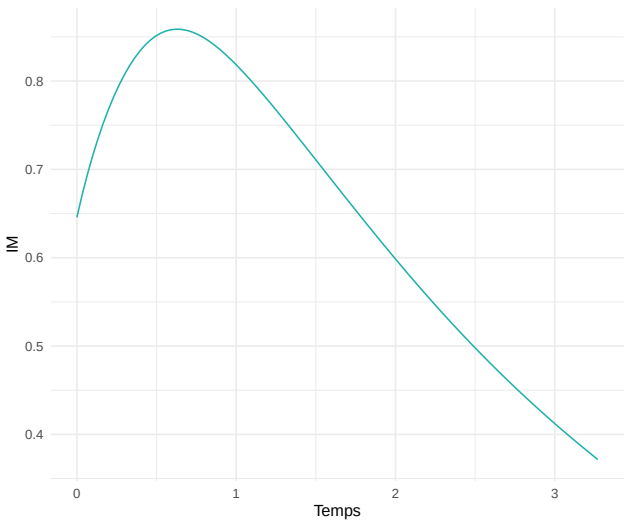

FIGURE 3.4 – IM associée aux taux de transition et aux fonctions d'intensités de la Figure 3.5

Notons que lorsque les intensités sont constantes, les deux formulations (calendaires et inter-temps) sont équivalentes. L'intensité du processus est donc définie par l'ensemble de ces taux de transition. La Figure 3.5 propose une illustration de taux de transition pour lesquels chaque évènement augmente les probabilités d'en faire un suivant. A partir de ces taux de transition, des individus ont été simulés et l'intensité tracée pour chacun d'entre eux. Dans ce cas, la fonction d'intensité est continue par morceaux. L'IM, en tant que quantité marginale, correspond ici à une moyenne pondérée par la probabilité d'appartenance à chaque état (nombre d'évènements), présenté en Figure 3.4. Nous reviendrons plus formellement sur le lien entre l'IM et l'intensité dans ce cas particulier.

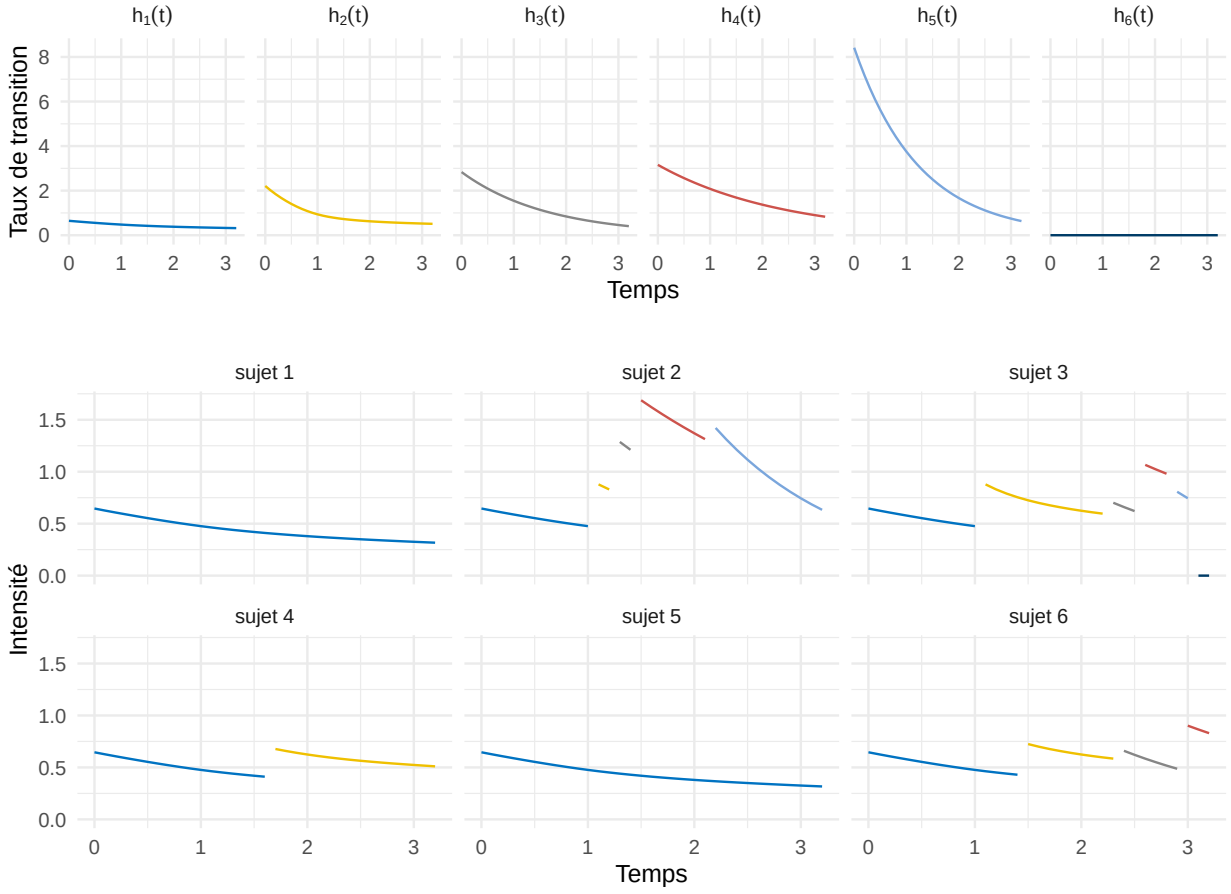

FIGURE 3.5 – Taux de transitions et intensités individuelles chez des sujets simulés. *Les taux de transition sont définis selon les temps calendaires.*

Lien avec le cas à un seul évènement

Il est possible de définir le cadre de survie classique à un unique évènement dans le formalisme du processus de comptage. Dans ce cas particulier, N(t) ne présente qu'un seul saut dont le temps est défini par la variable aléatoire T. Tant que l'évènement n'est pas survenu à un instant t, l'historique du processus peut être résumé par le fait qu'aucun évènement ne s'est produit c'est-à-dire que $T \leq t$.

Ainsi, l'intensité du processus, tant que l'évènement ne s'est pas produit, est :

$$\begin{aligned}\lambda_0(t) &= \lim_{dt \to 0} \frac{P(dN(t) = 1 | H(t))}{dt} \\ &= \lim_{dt \to 0} \frac{P(t \geq T \geq t + dt | T \leq t)}{dt} \\ &= h(t)\end{aligned}$$

où, $h(t)$ est le taux tel que défini dans la partie précédente.

Après la survenue de l'évènement, la probabilité de faire un second évènement est nulle donc :

$$\lambda_1(t) = 0$$

En toute rigueur, l'intensité ne correspond donc pas au taux, même si les deux sont confondus tant que l'évènement étudié n'a pas eu lieu.

Calculer ρ revient à calculer la moyenne pondérée des intensités entre ceux qui ont fait l'évènement et ceux qui ne l'ont pas fait. Ainsi :

$$\begin{aligned}\rho(t) &= \lambda_0(t)S(t) + \lambda_1(t)(1 - S(t)) \\ &= h(t)S(t) \\ &= f(t)\end{aligned}$$

où, $S(t)$ est la fonction de survie et $f(t)$ est la densité associée à la variable aléatoire T. Ainsi, l'IM correspond à la densité de T.

Enfin, $\mu(t)$ est l'intégrale de cette densité entre 0 et t soit, par définition, la fonction de répartition.

1.2 Modélisation de l'intensité

Une première approche de modélisation d'un processus à évènements récurrents est de faire de la régression sur l'intensité. Dans la première partie de cette thèse, nous avons évoqué deux modèles de régression semi-paramétriques basés sur l'intensité : le modèle d'Andersen-Gill (AG) et le modèle de Prentice-William-Petersen (PWP). Nous revenons plus en détails sur ces modèles dans cette partie. A ce stade, nous traiterons uniquement le cadre inférentiel en présence de censure non informative. La modélisation en présence d'un évènement terminal sera discutée ultérieurement.

1.2.1 Le modèle d'Andersen-Gill

Pour un individu i, nous observons m_i évènements aux temps $t_{i1}, t_{i2}, ..., t_{im_i}$. Il peut être soumis à censure (nous ne considérons ici que la censure à droite ($\tau_0 = 0$) mais la troncature à gauche est également possible). On note τ_i le dernier temps observé pour l'individu i et on définit $\tau = \max_i \tau_i$. Nous observons le processus d'observation $y_i(t)$ prenant la valeur 1 sur $[0, \tau_i]$ et 0 sur $[\tau_i, \tau]$. Dans la suite, nous noterons $\mathbf{X}$ la matrice de design du modèle pour l'ensemble des individus d'un échantillon et $\mathbf{X}_i$ la ligne de cette matrice correspondant à l'individu i.

Le modèle d'Andersen and Gill 1982 (AG) est un modèle semi-paramétrique qui suppose que le processus de comptage est un processus de Poisson. De plus, les effets des covariables sont supposés proportionnels à une fonction d'intensité de base λ_0 laissée non spécifiée. Le modèle d'Andersen-Gill peut donc s'écrire comme suit :

$$\lambda(t, \mathbf{X}, \boldsymbol{\beta}) = \lambda_0(t) \exp(\mathbf{X}\boldsymbol{\beta}) \tag{1.9}$$

Comme dans la partie précédente, nous sommes intéressés par la description de l'intensité de base par une spline. Nous considérons donc le modèle plus général suivant :

$$\{\lambda(t, \mathbf{X}_i(t), \boldsymbol{\beta})\}_{i=1...n} = \exp(\mathbf{X}(t)\boldsymbol{\beta}) \tag{1.10}$$

où, $\mathbf{X}(t)$ est la matrice de design, qui dépend à la fois des covariables et du temps.

En utilisant la densité de probabilité (1.3), on peut écrire la log-vraisemblance du modèle comme suit :

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\sum_{j=1}^{m_i} log\left(\lambda(t_{ij}, \mathbf{X}_i(t_{ij}), \boldsymbol{\beta})\right) \right] - \int_0^{\tau} y_i(u)\lambda(u, \mathbf{X}_i(u), \boldsymbol{\beta})du$$

que l'on peut réécrire, en notant $t_{im_i+1} = \tau$ et $t_{i0} = 0$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m_i+1} \left[I(j \neq m_i + 1).\log(\lambda(t_{ij}, \mathbf{X}_i\,(t_{ij}), \boldsymbol{\beta})) - \int_{t_{i(j-1)}}^{t_{ij}} y_i(u)\lambda(u, \mathbf{X}_i(u), \boldsymbol{\beta})du \right] \tag{1.11}$$

A travers la contribution d'un individu i à la vraisemblance (1.11), on reconnaît la vraisemblance d'un modèle de survie, dans lequel on observerait des individus distincts, tronqués à gauche au temps de l'occurrence précédente (ou l'origine) et un individu ne présentant pas l'évènement sur la période entre la dernière occurrence et la date de censure τ_i. L'estimation ponctuelle et l'inférence sur $\boldsymbol{\beta}$ peut être réalisée par toute implémentation du modèle de Cox, moyennant un travail de mise en forme du jeu de données. L'intervalle $[t_{i(j-1)}, t_{i(j)}]$ du j^e évènement sera spécifié en utilisant un temps de troncature à gauche et de censure à droite.

En dérivant (1.11), nous obtenons la fonction score :

$$\begin{aligned}\mathbf{U}(\boldsymbol{\beta},\tau) &= \sum_{i=1}^{n}\mathbf{U}_i(\boldsymbol{\beta},\tau)\\ &= \sum_{i=1}^{n}\left[\sum_{j=1}^{m_i}\mathbf{X}_i(t_{ij}) - \int_0^{\tau} y_i(u).\mathbf{X}_i(u).\exp(\mathbf{X}_i(u)\boldsymbol{\beta})du\right]\end{aligned}$$

et comme $dN_i(t)$ vaut 0 sauf sur les sauts $\sum_{j=1}^{m_i}\mathbf{X}_i(t_{ij}) = \int_0^{\tau} y_i(s)\mathbf{X}_i(s)dN_i(s)$, ce qui donne :

$$\mathbf{U}(\boldsymbol{\beta},\tau) = \sum_{i=1}^{n}\int_0^{\tau} y_i(u)\mathbf{X}_i(u)\left[dN_i(u) - \exp(\mathbf{X}_i(u)\boldsymbol{\beta})du\right] \tag{1.12}$$

On note $\hat{\boldsymbol{\beta}}$ la solution de l'équation de score suivante :

$$U(\boldsymbol{\beta},\tau) = 0 \tag{1.13}$$

Par la suite, pour simplifier les notations, on notera $U(\boldsymbol{\beta},\tau) = U(\boldsymbol{\beta})$.

Soit $\boldsymbol{\beta}_0$, le vecteur des vraies valeurs de $\boldsymbol{\beta}$. On se place dans les conditions de régularité suivantes :

— (a) $\{N_i(.), Y_i(.), \mathbf{X}_i(.)\}$ sont indépendants et identiquement distribués,
— (b) $P(Y_i \geq \tau) > 0, i = 1, ..., n$, où τ est une constante prédéterminée,
— (c) $N_i(\tau), i = 1, ..., n$ est borné par une constante,
— (d) $\mathbf{X}_i(.), i = 1, ..., n$ a ses variations totales bornées.
— (e) $A = \mathbb{E}\left[\int_0^{\tau}\mathbf{X}_i(u))\mathbf{X}_i^T(u)\exp(\mathbf{X}_i(u)\boldsymbol{\beta}_0)du\right]$ est définie positive.

Alors, la distribution de l'estimateur $\hat{\boldsymbol{\beta}}$ est asymptotiquement normale :

$$\hat{\boldsymbol{\beta}} \underset{n\to+\infty}{\sim} \mathcal{N}(\boldsymbol{\beta}_0, \mathcal{I}(\boldsymbol{\beta}_0)^{-1})$$

où, $\mathcal{I}$ est la matrice de Fisher définie par :

$$\mathcal{I}(\boldsymbol{\beta}) = -\mathbb{E}(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T})$$

En pratique, comme on ne connaît pas $\mathcal{I}(\boldsymbol{\beta}_0)$, on utilise la matrice d'information de Fisher observée, estimée au maximum de vraisemblance $I(\hat{\boldsymbol{\beta}})$ pour construire des intervalles de confiance. Cette dernière est définie par :

$$I(\hat{\boldsymbol{\beta}}) = -\frac{\partial^2 l(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$$

Dans notre cas, $\lambda(t) = \exp(\mathbf{X}(t)\boldsymbol{\beta})$, on peut donc calculer le vecteur score et la matrice d'information observée :

$$I(\boldsymbol{\beta},t) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\sum_{i=1}^{n}\int_0^{\tau} y_i(u).\mathbf{X}_{\ i}(u).\mathbf{X}_i^T(u).\exp(\mathbf{X}_i(u)\boldsymbol{\beta})du \tag{1.14}$$

En pratique, l'hypothèse d'indépendance du nombre d'évènements sur des intervalles disjoints est peu réaliste. Il est cependant possible d'utiliser une approche robuste pour corriger la variance dans un cadre mal spécifié (Lin and Wei, 1989).

1.2.2 Le processus de Poisson avec effet aléatoire

Dans le cadre du processus de Poisson (et donc du modèle d'Andersen-Gill), il est supposé que l'intensité ne dépend pas de l'historique du processus, ce qui est peu réaliste. Une manière d'incorporer cet historique est d'utiliser un terme de fragilité individuelle b_i (Rondeau, 2010). L'intensité conditionnellement à b_i prend alors la forme suivante pour l'individu i :

$$\lambda_i(t, \mathbf{X}_i(t), \boldsymbol{\beta}, b_i) = \lambda_0(t) \exp\left(\mathbf{X}_i(t)\boldsymbol{\beta}\right) b_i$$

où, les b_i sont des effets aléatoires non-observés communs à toutes les observations du sujet i, indépendants et identiquement distribués. Cette distribution est classiquement une loi Gamma ou une loi log-normale, dont on notera la densité $f(b, \theta)$, paramétrée par θ.

Nous cherchons à estimer $\boldsymbol{\beta}, \theta, \lambda_0$. Comme en section 1.2.1, nous allons procéder par vraisemblance marginale :

$$f(\mathbf{t}|\boldsymbol{\beta}, \theta) = \int_b f(\mathbf{t}, \mathbf{b}|\boldsymbol{\beta}, \theta) d\mathbf{b} \tag{1.15}$$

Pour un individu i,

$$\begin{aligned} f(\mathbf{t}_i, b_i|\boldsymbol{\beta}, \theta) &= f(\mathbf{t}_i|\boldsymbol{\beta}, \theta, b_i) f(b_i|\theta) \\ &= \left[\prod_{k=1}^{m_i} \lambda_i(t_{ik})\ b_i\right] \exp\left(-\int_0^{\tau} y_i(s)\lambda_i(s)b_i ds\right) f(b_i|\theta) \\ &= \left[\prod_{k=1}^{m_i} \lambda_i(t_{ik})\ b_i\ \exp\left(-\int_{t_{i(k-1)}}^{t_{ik}} y_i(s)\lambda_i(s)b_i ds\right)\right] f(b_i|\theta) \end{aligned} \tag{1.16}$$

Ce qui donne finalement :

$$f(\mathbf{t}|\boldsymbol{\beta}, \theta) = \int_b \left\{ \prod_{i=1}^{n} \left[\prod_{k=1}^{m_i} \lambda_i(t_{ik})\ b\ \exp\left(-\int_{t_{i(k-1)}}^{t_{ik}} y_i(s)\lambda_i(s)b ds\right)\right]\right\} f(b_i|\theta) db \tag{1.17}$$

L'implémentation de cette vraisemblance est disponible via le package R `frailtypack` (Rondeau et al., 2012). L'intensité cumulée y est modélisée par une spline pénalisée (M-spline) et l'intensité est obtenue par dérivation. Il a également été proposé de directement modéliser l'intensité par une spline pénalisée en utilisant un modèle exponentiel par morceaux (Ramjith et al., 2024) via le package R `pammtools`. L'ajustement peut être ensuite réalisé en utilisant un modèle de Poisson via le package `mgcv`.

Interprétation

Comme dans le cadre de la survie classique, une part d'hétérogénéité entre les patients ou fragilité est toujours présente du fait de l'absence d'une ou plusieurs variables (Balan and Putter, 2020). Dans le cadre d'un processus récurrent, une dépendance entre les évènements d'un même sujet peut également être présente (Xu and Cheung, 2018). Dans le modèle de Poisson avec effet aléatoire, le terme de fragilité peut alors capter ces deux sources d'hétérogénéité (Balan and Putter, 2020).

Vraisemblance et troncature à gauche

En présence de troncature à gauche dans le modèle classique du taux avec fragilité, la vraisemblance doit tenir compte du fait que les délais d'entrée sont conditionnels à la valeur de l'effet aléatoire (dont la distribution est définie en t=0) (Charvat and Belot, 2021; Balan and Putter, 2020). Dans ce cas, on observe également $\mathbf{t}_0$, le vecteur composé des t_{0ij}, le début d'observation de l'individu j du groupe i. En supposant que la distribution des temps de troncature à gauche ne dépend pas des $\boldsymbol{\beta}$, $f(\mathbf{y}|\theta, \boldsymbol{\beta}, b_i)$ devient :

$$f(\mathbf{t}, \boldsymbol{\delta}, \mathbf{t}_0, \mathbf{b}|\theta, \boldsymbol{\beta}) \propto f(\mathbf{t}, \boldsymbol{\delta}|\theta, \boldsymbol{\beta}, \mathbf{b}, \mathbf{T} > \mathbf{t}_0) f(\mathbf{b}|\theta, \mathbf{T} > \mathbf{t}_0) \tag{1.18}$$

Pour un groupe i, on peut écrire les densités $f(\mathbf{t}_i, \boldsymbol{\delta}_i|\theta, \boldsymbol{\beta}, b_i, \mathbf{T}_i > \mathbf{t}_{0i})$ et $f(b_i|\theta, \mathbf{T}_i > \mathbf{t}_{0i})$ de la façon suivante (van den Berg and Drepper, 2016) :

$$\begin{aligned}
f(\mathbf{t}_i, \boldsymbol{\delta}_i|\theta, \boldsymbol{\beta}, b_i, \mathbf{T}_i > \mathbf{t}_{0i}) &= \prod_{j=1}^{n_i} \{h(t_{ij}) b_i\}^{\delta_{ij}} \exp(-\int_{t_{0ij}}^{t_{ij}} h(s) b_i ds) \\
f(b_i|\mathbf{T}_i > \mathbf{t}_{0i}, \theta) &= \frac{P(\mathbf{T}_i > \mathbf{t}_{0i}|b_i) f(b_i|\theta)}{P(\mathbf{T}_i > \mathbf{t}_{0i})} \\
&= \frac{\prod_{j=1}^{n_i} \exp\{-\int_0^{t_{0ij}} h(\mathbf{s}, b_i) ds\}}{\int_b \prod_{j=1}^{n_i} \exp\{-\int_0^{t_{0ij}} h(\mathbf{s}, b) ds\} f(b|\theta) db} f(b_i|\theta)
\end{aligned}$$

Pour l'ensemble des groupes, on obtient donc la vraisemblance suivante :

$$\begin{aligned}
f(\mathbf{t}, \boldsymbol{\delta}, \mathbf{t}_0|\theta, \boldsymbol{\beta}) &= \prod_{i=1}^{m} \int_b \frac{\prod_{j=1}^{n_i} \{h(t_{ij}) b_i\}^{\delta_{ij}} \exp(-\int_{t_{0ij}}^{t_{ij}} h(s) v ds) \exp\{-\int_0^{t_{0ij}} h(\mathbf{s}, v) ds\}}{\int_b \prod_{j=1}^{n_i} \exp\{-\int_0^{t_{0ij}} h(\mathbf{s}, u) ds\} f(u|\theta) du} f(v|\theta) dv \\
&= \prod_{i=1}^{m} \int_b \frac{\prod_{j=1}^{n_i} \{h(t_{ij}) v\}^{\delta_{ij}} \exp\{-\int_0^{t_{ij}} h(\mathbf{s}, v) ds\}}{\int_b \prod_{j=1}^{n_i} \exp\{-\int_0^{t_{0ij}} h(\mathbf{s}, u) ds\} f(u|\theta) du} f(v|\theta) dv
\end{aligned} \tag{1.19}$$

Dans le cas du processus de Poisson, nous utilisions la troncature à gauche du modèle de taux pour définir le début de l'intervalle d'un évènement. Cela permet de calculer la vraisemblance du processus recurrent en utilisant une implémentation de la vraisemblance du modèle de taux. En présence d'un terme de fragilité, la vraisemblance (1.18) ne correspond donc pas à celle de la vraisemblance du modèle de taux flexible avec troncature à gauche, que l'on trouvera dans certains packages R comme `mexhaz` (Charvat and Belot, 2021), par exemple. Une implémentation spécifique au modèle récurrent est nécessaire dans ce cas.

1.2.3 Le modèle multi-états

Pour évaluer l'impact de covariables sur une intensité d'unn processus de type multi-états, Prentice et al. 1981 ont proposé un modèle semi-paramétrique connu sour le nom de Prentice-Williams-Petersen (PWP), extension du modèle de Cox. La période d'observation du patient est découpée en plusieurs lignes définissant les périodes entre deux évènements. L'ajustement se fait comme un modèle de Cox classique mais en stratifiant sur le nombre d'évènements passés. Plus formellement, le modèle s'écrit, en supposant un ratio d'intensités indépendant du nombre

d'occurrences d'évènements :

$$\lambda(t, k, \mathbf{X}) = \lambda_{0k}(t)\ exp(\mathbf{X}\boldsymbol{\beta}) \tag{1.20}$$

Dans le modèle PWP, conditionnellement au nombre d'évènements passés, les observations sont supposées indépendantes. Cependant, il peut rester une hétérogénéité individuelle créant de la corrélation entre les évènements d'un sujet. Il a donc été proposé d'ajouter un effet aléatoire au niveau du sujet pour tenir compte de cette corrélation individuelle (Box-Steffensmeier and De Boef, 2006).

2 Proposition d'un cadre de modélisation flexible pour l'intensité marginale

2.1 Avant-propos

Cette partie propose un cadre d'estimation flexible de l'intensité marginale (IM). Nous nous plaçons dans un contexte de censure non-informative. Nous discuterons néanmoins des possibilités de modélisation en cas d'occurrence d'un évènement terminal. Dans ce cadre, nous rappelons la définition de l'IM, quel que soit le processus sous-jacent :

$$\rho(t) = \lim_{dt \to 0} \frac{P(dN(t) = 1)}{dt} = \mathbb{E}\left[dN(t)\right] \tag{2.1}$$

En s'affranchissant du conditionnement à l'historique du processus, l'IM présente plusieurs atouts intéressants. D'une part, son estimation ne requiert pas d'hypothèse sur les caractéristiques du processus sous-jacent, évitant des erreurs de spécification. L'inférence peut être obtenue à l'aide d'un estimateur robuste, ne nécessitant pas davantage d'hypothèses. D'autre part, elle propose une description synthétique du processus au niveau de la population (elle perd, cependant, en même temps que son conditionnement, son interprétation au niveau individuel). Son utilisation est donc destinée à décrire le processus ou à évaluer l'effet d'une variable à l'échelle d'une population (épidémiologie) plutôt qu'à faire de la prédiction individuelle.

L'identification de la quantité ciblée par la modélisation, c'est-à-dire marginale ou conditionnelle, dans la littérature sur les événements récurrents est souvent floue ou ambiguë. L'IM est fréquemment désignée sous le nom *rate*, en particulier par Lin et al. 2000 puis par Cook and Lawless 2007. Le manque de spécificité de ce terme (*hazard rate*, *mortality rate*, etc) nous semble être une première source de confusion. C'est pourquoi le terme intensité marginale (*marginal intensity*) lui sera préféré dans ce développement. La modélisation explorée dans ces travaux vise à aller au-delà du cadre d'analyse traditionnel, qui semble être en grande partie responsable des confusions observées au cours de ces recherches. Ce cadre traditionnel comprend les éléments suivants :

Le modèle semi-paramétrique

A notre connaissance, le premier modèle de régression sur l'IM avec une inférence basée sur des équations d'estimation, a été introduit par Pepe and Cai 1993. L'article proposait de décrire le processus avec deux modèles : un modèle de taux jusqu'au premier évènement et un modèle d'intensité marginale sur les évènements subséquents. Cette première approche offrait une visualisation des indicateurs au cours du temps en se basant sur des polynômes cubiques. C'est ensuite à Lin et al. 2000 que l'on doit la formalisation rigoureuse du modèle d'intensité marginale dans un cadre semi-paramétrique. La perte d'intérêt pour la visualisation

de l'indicateur, dans un contexte aussi complexe qu'un processus récurrent, ne semble pas favoriser la compréhension du sujet. Ce travail revisite la modélisation flexible de l'IM, basée sur les équations d'estimation et que l'on appellera l'approche directe. Il l'étend également à un cadre pénalisé permettant de lisser la forme estimée du taux afin de faciliter la construction du modèle.

Le processus de Poisson

L'omniprésence du processus de Poisson dans la littérature et la juxtaposition des deux intensités (conditionnelle et marginale) dans ce cas particulier ne rend pas justice à l'IM, qui perd dans ce cadre tout son intérêt. A titre d'exemple, bien qu'un chapitre entier lui soit dédié dans *The statistical analysis of recurrent events* (Cook and Lawless, 2007), la définition de l'intensité marginale y est introduite dans le cadre du processus de Poisson. Par ailleurs, une manière classique de prendre en compte la corrélation entre évènements intra-sujet est d'introduire un effet aléatoire au niveau individuel. Conditionnellement à cet effet aléatoire, le processus d'évènements individuel est alors de Poisson, dont les intensités conditionnelles et marginales sont donc confondues. Dans ce contexte, il est donc courant de définir l'IM au niveau individuel (Cook and Lawless 2007 (p 76-77), Chiou et al. 2023). Cette définition de l'IM ne coïncide cependant pas avec la définition (2.1) (au niveau de la population). Nous proposons d'explorer une seconde manière d'estimer l'IM, que nous appellerons approche indirecte. Cette approche est, cette fois-ci, basée sur un modèle d'intensité conditionnelle de type processus de Poisson avec effets aléatoires. L'IM est calculée en intégrant l'intensité conditionnelle du modèle par rapport à la distribution des effets aléatoires.

L'hypothèse de proportionnalité

Il existe de nombreuses études de simulation concernant le modèle de Lin et al. 2000. Ce modèle semi-paramétrique, extension du modèle de Cox, repose donc sur une hypothèse de proportionnalité. La simulation d'un processus nécessite de définir des intensités conditionnelles. Revenir à l'intensité marginale à partir des intensités conditionnelles est une tâche complexe, sauf dans le cas du processus de Poisson (y compris en présence d'un effet aléatoire). En outre, en dehors de ces cas, la proportionnalité des intensités conditionnelles n'implique plus la proportionnalité des intensités marginales. Les études de simulation de l'intensité marginale sont donc réalisées dans le cas spécifique ou intensité conditionnelles et marginales sont confondues (ou presque) (Metcalfe and Thompson, 2006). Cette thèse propose donc une exploration de l'IM au-delà des processus de Poisson (et ses dérivés) à travers le processus multi-états. Ce travail est notamment rendu possible par la modélisation utilisée, gérant les effets non-proportionnels des covariables.

2.2 Cadre de modélisation non pénalisé

2.2.1 Méthode directe

Dans le cadre du processus de Poisson, le modèle d'IM est un modèle d'intensité conditionnelle puisque les deux quantités sont confondues. On peut donc écrire la log-vraisemblance (1.11) pour estimer et faire de l'inférence sur les paramètres du modèle. En dehors du processus de Poisson, spécifier un modèle sur l'IM ne suffit pas à décrire la loi du processus (il faut revenir au niveau de l'intensité conditionnelle pour cela), le modèle n'a donc pas de vraisemblance. Il est

cependant possible d'utiliser la théorie des équations d'estimation pour estimer les paramètres et construire leurs intervalles de confiance.

En reprenant les notations du modèle sur le logarithme du taux (1.1.2), (1.7), on considère le modèle sur le logarithme de l'IM suivant :

$$\begin{aligned} \log(\rho(t,\mathbf{x})) &= \sum_l f_l(t,\mathbf{x}) \\ \{log(\rho(t,\mathbf{x}_i))\}_{i=1,\dots,n} &= \mathbf{X}(t)\boldsymbol{\beta} \end{aligned} \tag{2.2}$$

f_l est une fonction multidimensionnelle du temps et des covariables et $\mathbf{X}(t)$ la matrice de design du modèle (incluant la représentation des covariables sous forme de splines).

Pour un design donné (nombre et emplacement des noeuds des splines fixés), on appelle $\boldsymbol{\beta}_0$ le vecteur de paramètres qui approxime le mieux la forme théorique de l'IM. On notera $\log(\rho(t,\mathbf{x})) = \mathbf{X}(t)\boldsymbol{\beta}_0$.

On considère la fonction de score du processus de Poisson (1.12) comme fonction d'estimation sous le modèle (2.2) :

$$\begin{aligned} \mathbf{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \int_0^\tau y_i(u)\mathbf{X}_i(u)\Big[dN_i(u) - \exp(\mathbf{X}_i(u)\boldsymbol{\beta})du\Big] \\ &= \sum_{i=1}^{n} U_i(\boldsymbol{\beta}) \end{aligned} \tag{2.3}$$

On note $\hat{\boldsymbol{\beta}}$ la solution de l'équation d'estimation : $U(\boldsymbol{\beta}) = 0$. L'estimateur $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}_0$ a les propriétés suivantes :

Convergence de $\hat{\boldsymbol{\beta}}$

On suppose que les conditions (a)-(e) définies pour le modèle d'AG, détaillées en section 1.2.1 sont vérifiées, alors :

$\hat{\boldsymbol{\beta}}$ *converge presque-sûrement vers* $\boldsymbol{\beta_0}$.

Démonstration. Ce résultat est obtenu en constatant que si $Y_i(t)$ et $N_i(t)$ sont indépendants (censure non informative) et que sous le modèle (2.2), $\mathbb{E}\left[dN_i(t)|\mathbf{X}_i(t)\right] = \exp(\mathbf{X}_i(u)\boldsymbol{\beta}_0)$, la fonction d'estimation est non biaisée :

$$\mathbb{E}\left(\mathbf{U}(\boldsymbol{\beta}_0)|\mathbf{X}(t)\right) = 0$$

De cette propriété, on déduit donc que l'estimateur $\hat{\boldsymbol{\beta}}$ solution de l'équation d'estimation est un estimateur convergent de $\boldsymbol{\beta}_0$ (Pepe and Cai, 1993) (Appendix A.1). □

Distribution asymptotique de $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \to \mathcal{N}(0, \mathbf{A}^{-1}\boldsymbol{\Sigma}\mathbf{A}^{-1}) \tag{2.4}$$

où,

$$\mathbf{\Sigma} = \mathbf{E}\left[U_i(\boldsymbol{\beta}_0)U_i(\boldsymbol{\beta}_0)\right]$$

$$\mathbf{A} = -\mathbf{E}\left[\frac{\partial U_i^T(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right]$$

que l'on estime respectivement avec :

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_i^n U_i(\hat{\boldsymbol{\beta}})U_i^T(\hat{\boldsymbol{\beta}}) \tag{2.5}$$

$$\hat{\mathbf{A}} = -\frac{1}{n}\frac{\partial U(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \tag{2.6}$$

Démonstration. Pour montrer cela, on commence par poser $M_i(t) = N_i(t) - \int_0^t \exp(\mathbf{X}_i(u)\boldsymbol{\beta}_0)du$, on peut alors réécrire (1.12) :

$$U(\boldsymbol{\beta}_0) = \sum_{i=1}^n \int_0^\tau y_i(u)\mathbf{X}_i(u)dM_i(u)$$

On note que $\mathbb{E}(dM_i(t))|\mathbf{X}_i(t)) = 0$ et comme on fait l'hypothèse que y_i et N_i sont indépendants $\mathbb{E}(y_i(u)\mathbf{X}_i(u)dM_i(u)) = 0$.

Pour $i = 1, .., n$, les $U_i(\boldsymbol{\beta}_0) = \int_0^\tau y_i(u)\mathbf{X}_i(u)dM_i(u)$ sont indépendants et identiquement distribués à τ fixé et centrés en 0. En utilisant le théorème central limite, on peut donc dire que :

$$n^{-1/2}U(\boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{\Sigma}), \tag{2.7}$$

la fonction de covariance étant obtenue par :

$$\mathbf{\Sigma} = \mathbb{E}\left[U_i(\boldsymbol{\beta}_0)U_i(\boldsymbol{\beta}_0)^T\right]$$

En utilisant des développements de Taylor, on peut écrire :

$$U(\hat{\boldsymbol{\beta}}) \simeq \frac{\partial U(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = 0$$

$$U(\boldsymbol{\beta}_0) \simeq \frac{\partial U(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*)$$

et donc :

$$n^{1/2}(\hat{\beta} - \beta_0) \simeq \hat{A}^{-1}(\beta^*)n^{-1/2}U(\beta_0)$$

où, β^* est compris entre $\hat{\beta}$ et β_0 et $\hat{A}(\beta) = -\frac{1}{n}\frac{\partial U(\beta)}{\partial \beta}$.

On remarque que $\hat{\mathbf{A}}$ est un estimateur convergent de $\mathbf{A}$ (défini en hypothèse (e) en section 1.2.1). De plus, en utilisant la convergence en loi de $n^{-1/2}U(\boldsymbol{\beta}_0)$, on en déduit la convergence en loi de $n^{1/2}(\hat{\beta} - \beta_0)$ vers un vecteur aléatoire gaussien de moyenne nulle et de matrice de variance-covariance :

$$\mathbf{\Gamma} = \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-1}.$$

□

En pratique, on pourra donc procéder comme si on ajustait un processus de Poisson. La matrice $\mathbf{A}^{-1}$ est la variance fournie par le modèle et on pourra calculer les U_i de la matrice $\mathbf{\Sigma}$ en utilisant une quadrature de Gauss-Legendre pour approximer l'intégrale :

$$\begin{aligned} U_i &= \sum_{j=1}^{m_i} \mathbf{X}_i(t_{ij}) - \int_0^\tau y_i(u).\mathbf{X}_i(u).\exp(\mathbf{X}_i(u)\boldsymbol{\beta})du \\ &= \sum_{j=1}^{m_i} \mathbf{X}_i(t_{ij}) - \sum_{k=1}^{K} w_k^{GL} y_i(t_k^{GL}).\mathbf{X}_i(t_k^{GL}).\exp(\mathbf{X}_i(t_k^{GL})\boldsymbol{\beta}) \end{aligned}$$

où, w_k^{GL} et t_k^{GL} sont respectivement les poids et les noeuds de la quadrature de Gauss-Legendre.

On définit ainsi l'estimateur robuste de la variance associé à $\log(\rho)$:

$$\hat{V}_{\text{Robuste}} = \mathbf{X}^T \hat{\mathbf{\Gamma}} \mathbf{X} \tag{2.8}$$

Nous définissons également l'estimateur de la variance de $log(\rho)$ sous les hypothèses du processus de Poisson par :

$$\hat{V}_{\text{Naïf}} = \mathbf{X}^T \hat{\mathbf{A}}^{-1} \mathbf{X} \tag{2.9}$$

2.2.2 Méthode indirecte

Il existe un lien direct entre l'intensité conditionnelle obtenue à partir d'un modèle basé sur le processus de Poisson avec terme de fragilité et l'intensité marginale. En considérant que l'effet aléatoire capture à la fois la fragilité et la dépendance entre les évènements intra-sujets, l'intensité marginale s'obtiendrait en intégrant sur la distribution des effets aléatoires :

$$\rho(t, \mathbf{x}, \boldsymbol{\beta}) = \int_b \lambda(t, \mathbf{x}, \boldsymbol{\beta}, b) f(b) db \tag{2.10}$$

$f(b)$ étant la densité de la distribution des effets aléatoires, associée à un paramètre de variance θ^2.

Cette méthode indirecte est donc basée sur un modèle d'intensité, ce qui nous permet de faire l'estimation des paramètres et l'inférence en utilisant une vraisemblance (définie en 1.15). Pour les distributions Gamma et log-Normale, l'IM et l'intensité conditionnelle sont reliées par un simple facteur multiplicatif[1] (voir encart).

Pour calculer la variance associée à l'estimateur de variance de $\log(\hat{\rho}(t, \mathbf{X}(t), \hat{\boldsymbol{\beta}}))$, on peut écrire :

$$\hat{Var}(\log(\hat{\rho}(t, \mathbf{X}(t), \hat{\boldsymbol{\beta}})) = \hat{Var}(\log(\hat{\lambda}(t, \mathbf{X}(t), \hat{\boldsymbol{\beta}}))) + \hat{Var}\left(\frac{\hat{\theta}^2}{2}\right) - 2\ Cov\left(\log(\hat{\lambda}(t, \mathbf{X}(t), \hat{\boldsymbol{\beta}})), \frac{\hat{\theta}^2}{2}\right)$$

En utilisant la Δ-méthode, il vient :

$$\begin{aligned} \hat{Var}\left(\frac{\hat{\theta}^2}{2}\right) &= \frac{1}{4}\hat{Var}(\exp(2\log(\hat{\theta}))) \\ &= \hat{\theta}^4\ \hat{Var}(\log(\hat{\theta})). \end{aligned}$$

1. Cette propriété ne tient, en général, plus lorsque l'on se trouve en présence d'un évènement terminal.

En outre, on a :

$$
\begin{aligned}
2\ \hat{Cov}\left(\log(\hat{\lambda}(t,\mathbf{X}(t),\hat{\boldsymbol{\beta}}),\frac{\hat{\theta}^2}{2}\right) &= \hat{Cov}(\mathbf{X}(t)\hat{\boldsymbol{\beta}},\log(\hat{\theta})))\\
&= \mathbf{X}(t)^T Cov(\hat{\boldsymbol{\beta}},\exp(2\log(\hat{\theta}))),\\
&= 2\mathbf{X}(t)^T\hat{\theta}^2 Cov(\hat{\boldsymbol{\beta}},\log(\hat{\theta}))
\end{aligned}
$$

Finalement, en notant $\hat{V}_{\text{Mixte}} = \hat{Var}(\log(\hat{\rho}(t,\mathbf{X}(t),\hat{\boldsymbol{\beta}}))$:

$$
\hat{V}_{\text{Mixte}} = \hat{Var}(\log(\hat{\lambda}(t,\mathbf{X}(t),\hat{\boldsymbol{\beta}}))) + \hat{\theta}^4\ \hat{Var}(\log(\hat{\theta})) - 2\mathbf{X}(t)^T\hat{\theta}^2 Cov(\hat{\boldsymbol{\beta}},\log(\hat{\theta})) \tag{2.11}
$$

Processus de Poisson et effets aléatoires - Distributions Log-Normale et Gamma

Établissons le lien entre intensité marginale et intensité conditionnelle lorsque les effets aléatoires suivent une distribution Gamma et Log-Normale.

Frailty Gamma On suppose que $b \sim \Gamma(\alpha,\zeta)$, de telle sorte que : $\mathbf{E}(b) = \alpha/\zeta = 1$ et $Var(b) = \alpha/\zeta^2 = \theta^2$. L'équation (2.10) devient :

$$
\begin{aligned}
\rho(t,\mathbf{X}(t),\boldsymbol{\beta}) &= \int_b \lambda(t,\mathbf{X}(t),\boldsymbol{\beta}) b\ b^{\alpha-1}\frac{\zeta^\alpha exp(-\zeta b)}{\Gamma(\alpha)}db\\
&= \int_b \lambda(t,\mathbf{X}(t),\boldsymbol{\beta}) b^\alpha \frac{\zeta^{\alpha+1}exp(-\zeta b)}{\Gamma(\alpha+1)}\frac{\Gamma(\alpha+1)}{\zeta\Gamma(\alpha)}db\\
&= \lambda(t,\mathbf{X}(t),\boldsymbol{\beta})\frac{\Gamma(\alpha+1)}{\zeta\Gamma(\alpha)}\\
&= \lambda(t,\mathbf{X}(t),\boldsymbol{\beta})
\end{aligned} \tag{2.12}
$$

Frailty Gaussien

Dans ce cas, il est plus pratique de reparamétrer le modèle en $\lambda(t,\mathbf{X}(t),\boldsymbol{\beta})\exp(\tilde{b})$ avec $\tilde{b} \sim \mathcal{N}(0,\theta^2)$. L'équation (2.10) devient :

$$
\begin{aligned}
\rho(t,\beta) &= \int_b \lambda(t,\beta) exp(b)\frac{1}{\sqrt{2\pi}\theta}\exp\left(-\frac{b^2}{2\theta^2}\right)db\\
&= \int_b \lambda(t,\beta)\frac{1}{\sqrt{2\pi\theta^2}}\exp\left(-\frac{(b-\theta^2)^2}{2\theta^2}\right)exp\left(\frac{\theta^4}{2\theta^2}\right)db\\
&= \lambda(t,\beta)\exp\left(\frac{\theta^2}{2}\right)
\end{aligned} \tag{2.13}
$$

2.3 Cadre de modélisation pénalisé

Dans cette section, nous introduisons une pénalisation quadratique dans les modèles, au même titre que celle discutée en partie II.

2.3.1 Méthode directe

La pénalisation a déjà été proposée dans le modèle semi-paramétrique d'IM pour faire de la sélection de variables. Tong et al. 2009 ont considéré le cas des fonctions de pénalisation non-concaves comme le Lasso ou le SCAD dans le modèle semi-paramétrique de l'IM et Cai et al. 2020 ont proposé une méthode de sélection de variables groupées. La pénalisation pour des objectifs de lissage n'est en revanche, à notre connaissance, pas abordée pour ce modèle. Comme la vraisemblance du modèle n'est pas disponible, la pénalisation est appliquée sur l'équation d'estimation (2.3) :

$$U_p(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) - \mathbf{S}^{\boldsymbol{\kappa}}\boldsymbol{\beta} \tag{2.14}$$

Comme dans le cadre non pénalisé, on cherche $\hat{\boldsymbol{\beta}}_{PEN}$ tel que $U_p(\hat{\boldsymbol{\beta}}_{PEN}) = 0$.

Pour l'estimation des paramètres de lissage, en l'absence de vraisemblance, Tong et al. 2009 ont proposé d'utiliser un critère GCV basé sur la vraisemblance d'un processus de Poisson. Comme précédemment, nous proposons d'utiliser le critère LAML, basé lui aussi sur la vraisemblance du processus de Poisson.

On note $\hat{\boldsymbol{\beta}}_{PEN}$ la solution de l'équation d'estimation : $U_p(\boldsymbol{\beta}) = 0$. L'estimateur $\hat{\boldsymbol{\beta}}_{PEN}$ de $\boldsymbol{\beta}_0$ a les propriétés suivantes :

Convergence de $\hat{\boldsymbol{\beta}}_{PEN}$

On suppose que les conditions (a)-(e) définies pour le modèle d'AG, détaillées en section 1.2.1 sont vérifiées, alors :

$\hat{\boldsymbol{\beta}}_{PEN}$ *converge presque-sûrement vers* $\boldsymbol{\beta}_0$.

Démonstration. Lorsque $\boldsymbol{\kappa}$ est fixé, $U_p(\boldsymbol{\beta})$ ne diffère de $U(\boldsymbol{\beta})$ que par le terme $\mathbf{S}^{\boldsymbol{\kappa}}\boldsymbol{\beta}$. $U_p(\boldsymbol{\beta})$ et $U(\boldsymbol{\beta})$ partagent donc les mêmes propriétés de continuité et de dérivabilité (inversibles et bornées, existence de dérivées partielles). Ainsi, en utilisant les mêmes arguments que Pepe and Cai 1993 (Appendix A.1), on en déduit la convergence de $\hat{\boldsymbol{\beta}}_{PEN}$ vers $\boldsymbol{\beta}_0$.

Lorsque $\boldsymbol{\kappa}$ est obtenu par estimation, pour préserver la convergence, il faut s'assurer que le terme de pénalisation reste petit par rapport à $U(\boldsymbol{\beta})$ quand la taille de l'échantillon augmente. On reprend alors le raisonnement de Wood et al. 2016 (*Supplementary Material* B.2), qui montre qu'avec une estimation du paramètre de lissage par LAML, "*the penalty is unlikely to alter the consistency*". Bien que Wood et al. 2016 travaillent dans le cas d'une vraisemblance régulière, la démonstration se base sur l'équation de score et pas la vraisemblance en elle-même (voir formule (2) du *Supplementary Material* B.2). Les résultats sont donc transposables à l'équation d'estimation (2.14). On en déduit que $\hat{\boldsymbol{\beta}}_{PEN}$ converge presque-sûrement vers $\boldsymbol{\beta}_0$ lorsque $\boldsymbol{\kappa}$ est estimé par LAML. □

Distribution asymptotique de $\hat{\boldsymbol{\beta}}_{PEN}$

$$\sqrt{n}\left(\hat{\mathbf{A}}+\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\right)\left(\hat{\boldsymbol{\beta}}_{PEN}-\boldsymbol{\beta}_0+\left(\hat{\mathbf{A}}+\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\right)^{-1}\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\boldsymbol{\beta}_0\right)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\boldsymbol{\Sigma}) \tag{2.15}$$

avec $\hat{\boldsymbol{A}}$ et $\boldsymbol{\Sigma}$ définis respectivement en (2.6) *et* (2.5).

Démonstration. En faisant des expansions de Taylor au voisinage de $\boldsymbol{\beta}_0$ et de $\hat{\boldsymbol{\beta}}_{PEN}$:

$$U_p(\hat{\boldsymbol{\beta}}_{PEN})\simeq U_p(\boldsymbol{\beta}^*)+\left(\frac{\partial U(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}-\mathbf{S}^{\boldsymbol{\kappa}}\right)(\hat{\boldsymbol{\beta}}_{PEN}-\boldsymbol{\beta}^*)=0 \tag{2.16}$$

$$U_p(\boldsymbol{\beta}_0)\simeq U_p(\boldsymbol{\beta}^*)+\left(\frac{\partial U(\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}-\mathbf{S}^{\boldsymbol{\kappa}}\right)(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*) \tag{2.17}$$

(2.17) - (2.16) donne :

$$U_p(\boldsymbol{\beta}_0)\simeq(n\hat{\mathbf{A}}+\mathbf{S}^{\boldsymbol{\kappa}})(\hat{\boldsymbol{\beta}}_{PEN}-\boldsymbol{\beta}_0)$$

De plus,

$$U_p(\boldsymbol{\beta}_0)=U(\boldsymbol{\beta}_0)-\mathbf{S}^{\boldsymbol{\kappa}}\boldsymbol{\beta}_0$$

En combinant tout cela,

$$\begin{aligned}\frac{U(\boldsymbol{\beta}_0)}{\sqrt{n}}&=1\sqrt{n}\left[(n\hat{\mathbf{A}}+\mathbf{S}^{\boldsymbol{\kappa}})(\hat{\boldsymbol{\beta}}_{PEN}-\boldsymbol{\beta}_0)+\mathbf{S}^{\boldsymbol{\kappa}}\boldsymbol{\beta}_0\right]\\&=\frac{1}{\sqrt{n}}\left(\hat{\mathbf{A}}+\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\right)\left[(\hat{\boldsymbol{\beta}}_{PEN}-\boldsymbol{\beta}_0)+(\hat{\mathbf{A}}+\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n})^{-1}\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\boldsymbol{\beta}_0\right]\end{aligned}$$

De plus, nous avons établi la distribution de $\frac{1}{\sqrt{n}}U(\boldsymbol{\beta}_0)$ en (2.7). On en déduit la distribution asymptotique de $\hat{\beta}_{PEN}$. □

Nous définissons la variance robuste de $\log(\rho)$ par :

$$\hat{V}^*_{\text{Robuste}}=\mathbf{X}^T\left(\hat{\mathbf{A}}+\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\right)^{-1}\boldsymbol{\Sigma}\left(\hat{\mathbf{A}}+\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\right)^{-1}\mathbf{X} \tag{2.18}$$

On définit également l'estimateur empirique bayésien sous l'hypothèse que le processus est de Poisson :

$$\hat{V}^*_{\text{Naïf}}=\mathbf{X}^T\left(\hat{\mathbf{A}}+\frac{\mathbf{S}^{\boldsymbol{\kappa}}}{n}\right)^{-1}\mathbf{X} \tag{2.19}$$

Intervalle de confiance par bootstrap

Il est également possible de calculer les intervalles de confiance par échantillonnage bootstrap. Afin de prendre en compte la corrélation des évènements intra-individuels, on peut procéder par ré-échantillonnage en groupes (Deen and de Rooij, 2020). De plus, l'incertitude du paramètre de lissage est prise en compte en le réévaluant à chaque échantillon bootstrap.

2.3.2 Méthode indirecte

Nous souhaitons ajuster le modèle d'intensité conditionnelle d'un processus de Poisson avec effets aléatoires mais dans un cadre pénalisé. Le passage de l'intensité conditionnelle à l'IM (en intégrant sur la distribution des effets aléatoires) n'est pas impacté par la pénalisation. Le cadre d'estimation pénalisé, défini par Wood et al. 2016, avec lequel nous avons travaillé jusqu'à présent, permet également d'ajuster des modèles flexibles avec effets aléatoires. Comme détaillé en partie II section 1.2.2, les paramètres du modèles sont estimés par maximum *a posteriori*, c'est-à-dire en maximisant la loi *a posteriori* du modèle sous une loi *a priori* gaussienne (méthode pénalisée). La variance de la distribution des effets aléatoires peut être estimée avec le critère LAML, au même titre que les paramètres de lissage du modèle pénalisé. Ce cadre ne permet de considérer que la distribution log-Normale pour les effets aléatoires.

Remarque

Dans cette thèse, nous sommes intéressés par l'exploration du cadre pénalisé défini par Wood et al. 2016, qui a l'avantage d'estimer les paramètres du modèles et de lissage de façon simultanée. Cependant, il est possible d'envisager d'autres cadres pénalisés. Par exemple, les paramètres du modèle peuvent être obtenus par maximisation de la vraisemblance marginale pénalisée à paramètres de lissage fixés (Rondeau et al., 2003, 2007). Le choix de ces derniers peut être réalisé en fixant le nombre de degrés de liberté ou par LCV pour des splines unidimensionnelles. Ce cadre permet, notamment, de spécifier une distribution des effets aléatoires log-Normale ou Gamma.

3 | Explorations du modèle

3.1 Simulation d'évènements récurrents et modèle d'intensité marginale

Afin d'observer le comportement des modèles proposés, une étude de simulation a été réalisée en section 3.2. Une attention particulière a été portée afin de varier le processus de génération des données entre les différents scénarios (Metcalfe and Thompson, 2006). En effet, l'un des intérêts du modèle d'IM est qu'il ne nécessite pas d'hypothèses fortes comme pour les modèles basés sur l'intensité conditionnelle. Il devrait donc s'ajuster correctement sur divers types de processus de comptage.

3.1.1 Simulation d'évènements récurrents

Processus de renouvellement

Le processus de comptage le plus naturel à simuler au niveau individuel est le processus de renouvellement. Dans ce cas, les inter-temps $W_1, W_2, ..., W_K$ sont indépendants et il est donc possible de les simuler de façon successive à partir de la loi des W_k. Pour un individu i, on pourra ainsi procéder comme suit :

- Simuler le temps d'évènement 1, $T_{i1} = W_{i1}$, selon la distribution choisie. Pour une distribution générale, on pourra utiliser la méthode d'inversion uniforme décrite en (Partie 2 - Chapitre 2.2.1) ;
- Simuler le temps inter-évènement W_{i2} selon la distribution choisie. Le second temps d'évènement est alors $T_{i2} = T_{i1} + W_{i2}$.
- On reproduit le schéma de simulation jusqu'à atteindre un temps maximum déterminé τ, le premier T_{ik} dépassant cette valeur étant censuré.

Il est possible de s'écarter du cadre stricte du processus de renouvellement en autorisant la loi de W_k à varier en fonction de k. Dans ce cas, on est dans le cadre du processus multi-états type "inter-temps" (voir section 1.1.2). En faisant varier la loi en fonction de k, on crée de la corrélation entre évènements intra-sujets. La simulation de ce type de processus est proposée par le package `survsim` (Moriña and Navarro, 2014). Plusieurs distributions sont disponibles pour les inter-temps : Weibull, Log-Normal et Log-Logistique.

Processus de Poisson

Dans le cadre du processus de Poisson, W_k dépend uniquement du temps de survenue T_{k-1}. Nous souhaitons donc caractériser la loi de $W_k|T_{k-1} = t_{k-1}$ (Jahn-Eimermacher et al., 2015).

On note $\tilde{h}$, le taux associé à la variable aléatoire $W_k|T_{k-1} = t_{k-1}$ et λ l'intensité du processus de Poisson :

$$\begin{aligned}\tilde{h}_t(w)dw &= \mathcal{P}(w \leq W_k < w + dw|T_{k-1} = t, W_k \geq w) \\ &= \mathcal{P}(w + t \leq T_k < w + t + dw|T_{k-1} = t, T_k \geq w + t) \\ &= \mathcal{P}(dN(w + t) = 1|T_{k-1} = t, T_k \geq w + t) \\ &= \lambda(w + t)dw \end{aligned} \tag{3.1}$$

Le taux obtenu permet de caractériser la loi de $W_k|T_{k-1} = t_{k-1}$ et donc d'appliquer l'algorithme de simulation ci-dessus.

Processus multi-états type "temps calendaires"

De la même manière que pour le processus de renouvellement, on peut relaxer le cadre stricte du processus de Poisson, en faisant dépendre l'intensité du nombre d'évènements survenus. On se retrouve alors sur un processus multi-états de type "temps calendaire" (voir section 1.1.2). La relation (3.1) devient donc :

$$\tilde{h_{kt}}(w)du = \lambda_k(w + t)du \qquad\qquad k = 1...K \tag{3.2}$$

où, $\lambda_k(t)dt = \mathcal{P}(dN(t) = 1|N(t^-) = k - 1)$

Censure et effets aléatoires

Pour plus de réalisme, (Jahn-Eimermacher et al., 2015) proposent d'ajouter une variabilité individuelle au taux de base, par l'ajout d'un effet aléatoire W (*frailty*), tel que $\mathbf{E}(W) = 1$ et $Var(W) = \theta$:

$$\lambda_i(t) = \lambda_0(t)\ w_i\ \exp(X_i\boldsymbol{\beta})$$

Il est également possible d'appliquer une censure non informative en simulant C selon une loi exponentielle par exemple. Si $C < \tau$, on censure tous les évènements survenant après C.

3.1.2 Calcul de l'IM à partir des intensités du modèle multi-états

Comme évoqué en section 2.1, nous souhaitons évaluer le modèle d'IM en dehors du cas particulier du processus de Poisson (pouvant inclure un effet aléatoire au niveau individuel). Nous avons vu comment simuler le processus multi-états à partir des intensités conditionnelles. Cette section propose une manière de calculer l'IM dans ce cas.

On considère un modèle multi-états de type "temps calendaires". Les intensités conditionnelles et marginales peuvent être reliées par la relation suivante :

$$\begin{aligned}\rho(t) &= \lim_{dt\to 0} \frac{P(dN(t) = 1)}{dt} \\ &= \sum_k \lim_{dt\to 0} \frac{P(dN(t) = 1|N(t) = k)}{dt} P(N(t) = k) \\ &= \sum_k \lambda_k(t) P_{0k-1}(0, t) \end{aligned} \tag{3.3}$$

On note $P_{0k}(0, t) = P(N(t) = k)$, les probabilités de passage de l'état 0 (pas d'évènement) à l'état k (k évènements) entre le début du suivi et le temps t. De façon générale, on définit

$P_{mn}(s,t)$, la probabilité de passage de l'état m à l'état n entre les temps s et t pour $m \in \{0,..,K\}$ et $n \in \{m,..,K\}$. Pour des intensités générales, il n'est pas possible de calculer ces probabilités de façon analytique. Il est cependant possible de les obtenir par intégration numérique avec une quadrature de Gauss-Legendre par exemple. Voici les expressions de certaines de ces probabilités :

— La probabilité de rester dans l'état k entre s et t :

$$P_{kk}(s,t) = exp\left(-\int_s^t \lambda_{k+1}(u)du \right)$$

— Pour $k = 1$: la probabilité de passer de 0 à 1 évènement

$$P_{01}(0,t) = \int_0^t P_{00}(0,u)\lambda_1(u)P_{11}(u,t)du$$

— Pour $k = 2$: la probabilité de passer de 0 à 2 évènements

$$P_{02}(0,t) = \int_0^t \int_u^t P_{00}(0,u)\lambda_1(u)P_{11}(u,v)\lambda_2(v)P_{22}(v,t)dvdu$$

On comprend donc que les coûts de calcul augmentent rapidement avec le nombre d'évènements.

En ajoutant un intercept aléatoire au niveau du sujet sur l'intensité, la relation (3.3) devient :

$$\rho(t) = \int \rho(t|b)f(b)db = \int \sum_k \lambda_k(t|b)P_{0k}(t|b)f(b)db \tag{3.4}$$

Il est donc possible de calculer l'IM par intégration numérique en utilisant une quadrature de Gauss-Hermite dans ce cas (Charvat and Belot, 2021).

3.2 Etude de simulation

Dans cette section, nous explorons les propriétés des approches flexibles proposées au chapitre précédent.

3.2.1 Scénarios

Nous considérons 6 scénarios, dont les intensités sont décrites dans la Table 3.1. Les scénarios Poiss, PGauss et PGamma sont respectivement simulés à partir d'intensités conditionnelles d'un processus de Poisson sans effet aléatoire et avec effet aléatoire de distribution log-Normale et Gamma. Le fonction d'intensité a été définie par la fonction théorique : $\lambda(t) = \exp\left[-(1+\frac{t^2}{16})\right]$. Les autres scénarios sont basés sur des processus multi-états. Pour assurer un certain réalisme aux modèles théoriques considérés, les intensités conditionnelles théoriques ont été obtenues en ajustant des modèles d'intensité sur des réelles. Ces modèles théoriques ont été ajustés en utilisant une spline du temps à 5 noeuds, avec le package `survPen`. Le scénario ColRec a été construit à partir des données colorectal du package R `frailtypack`. Les scénarios Heart et HeartGauss ont été élaborés à partir des données hfaction_cpx12 (Heart Failure data) du package R `WA`. Pour chaque scénario, l'intensité conditionnelle dépend d'une covariable binaire x.

Scenario	*Type*	*Effet aléatoire*	*Définition*
Poiss	Poisson		$\lambda_{Poiss}(t)\exp(0.3\ I(x=1))$
PGauss	Poisson	$b_i \sim log-\mathcal{N}(0,0.4^2)$	$\lambda_{Poiss}(t)\exp(0.3\ I(x=1))\ \ b_i$
PGamma	Poisson	$b_i \sim \Gamma(2,2)$	$\lambda_{Poiss}(t)\exp(0.3\ I(x=1))\ \ b_i$
ColRec	Multi-états		$\lambda_{Multi}(t,k)\exp(0.3\ I(x=1))$
Heart	Multi-états		$\lambda_{Multi}(t,k)\exp(0.3\ I(x=1))$
HeartGauss	Multi-états	$b_i \sim log-\mathcal{N}(0,0.4^2)$	$\lambda_{Multi}(t,k)\exp(0.3\ I(x=1))\ \ b_i$

TABLE 3.1 – Liste des scénarios de l'étude de simulation

$$\lambda_{Poiss}(t) = \lim_{dt\to 0} \frac{P(dN(t)=1)}{dt} \qquad \lambda_{Multi}(t,k) = \lim_{dt\to 0} \frac{P(dN(t)=1|N(t^-)=k-1)}{dt}$$

$k-1$ étant le nombre d'évènements survenus juste avant le temps t.

Les intensités conditionnelles sont représentées en Figure 3.6 et les IM associées, obtenues par intégration numérique en Figure 3.7. Notons que l'effet de la covariable x est proportionnel sur les intensités dans chaque scénario. Cependant, en dehors des scénarios de type Poisson, la proportionnalité n'est plus respectée lorsque l'on passe à l'IM comme le montre la Figure 3.8.

Les données sont simulées comme détaillé au chapitre précédent, en retenant comme origine le début du suivi. Pour chaque scénario, nous considérons 3 tailles d'échantillons : 100, 400 et 800 sujets. Pour chaque sous-scénario, $D = 1000$ échantillons sont simulés.

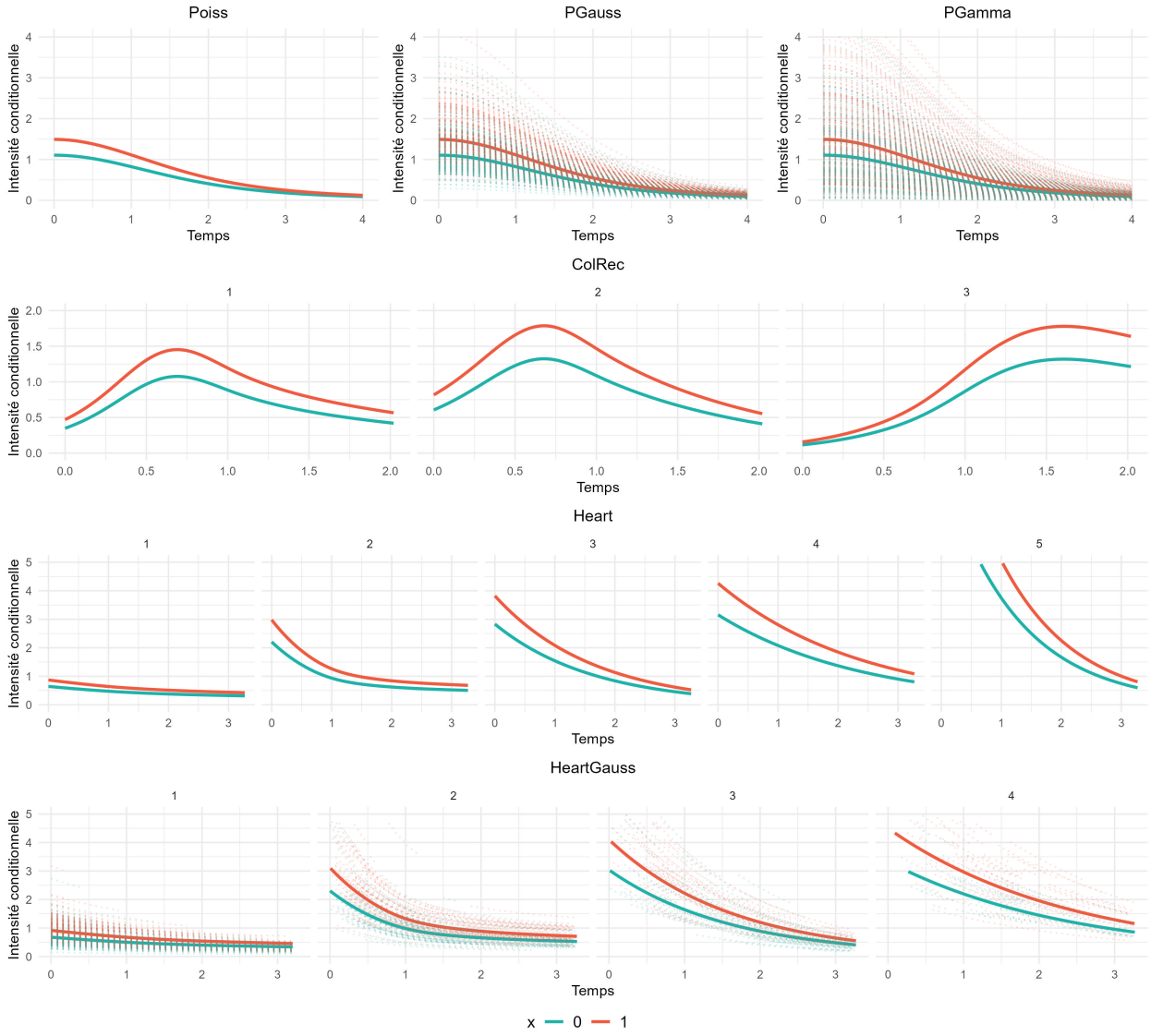

FIGURE 3.6 – Intensités théoriques des différents scénarios. *Pour les scénarios PGauss, PGamma et HeartGauss, les courbes en pointillés larges correspondent à l'intensité pour un effet aléatoire égal à 1 alors que celles en pointillés fins sont obtenues pour un tirage aléatoire de 500 valeurs de b_i dans sa distribution.*

3.2.2 Spécification des modèles

Les caractéristiques des modèles ajustés sur les échantillons simulés des différents scénarios sont résumées en Table 3.2. Les modèles associés aux scénarios Poiss, PGauss et PGamma sont spécifiés avec un effet proportionnel de x :

$$\log(\rho(t,x)) = s_c(t) + \beta x \tag{3.5}$$

s_c est une spline cubique naturelle du temps commune aux deux niveaux de la covariable x.

Dans les scénarios ColRec, Heart et HeartGauss, l'hypothèse de proportionnalité n'est pas valide, le modèle suivant est donc spécifié :

$$\log(\rho(t,x)) = s_0(t)I(x=0) + s_1(t)I(x=1) \tag{3.6}$$

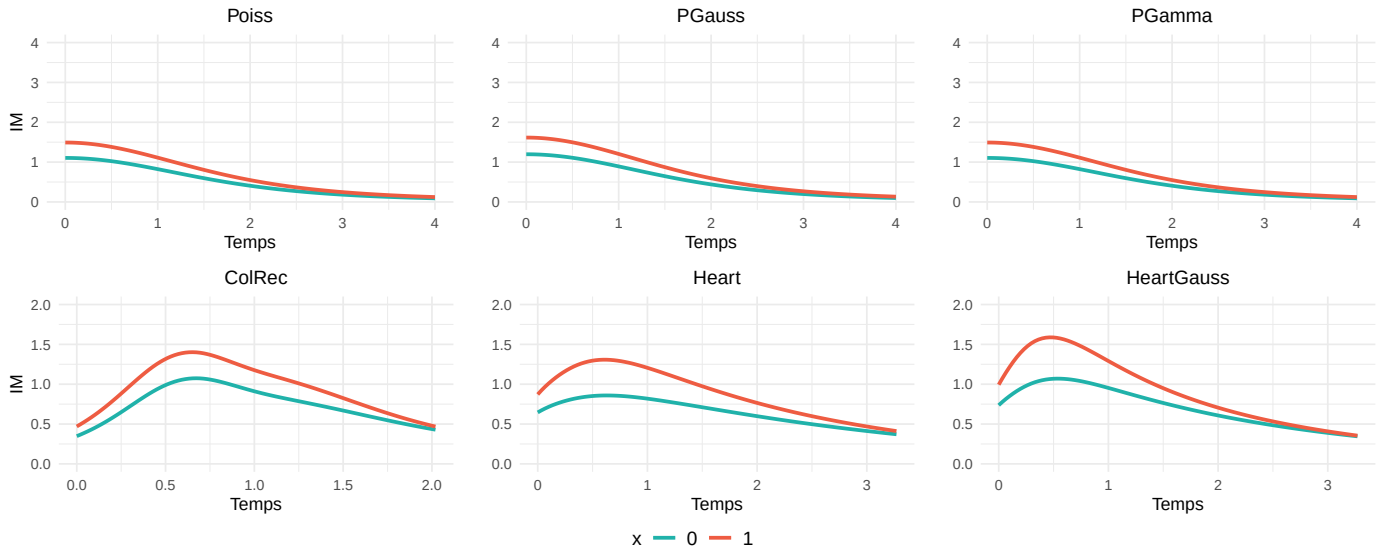

FIGURE 3.7 – IM théoriques des différents scénarios.

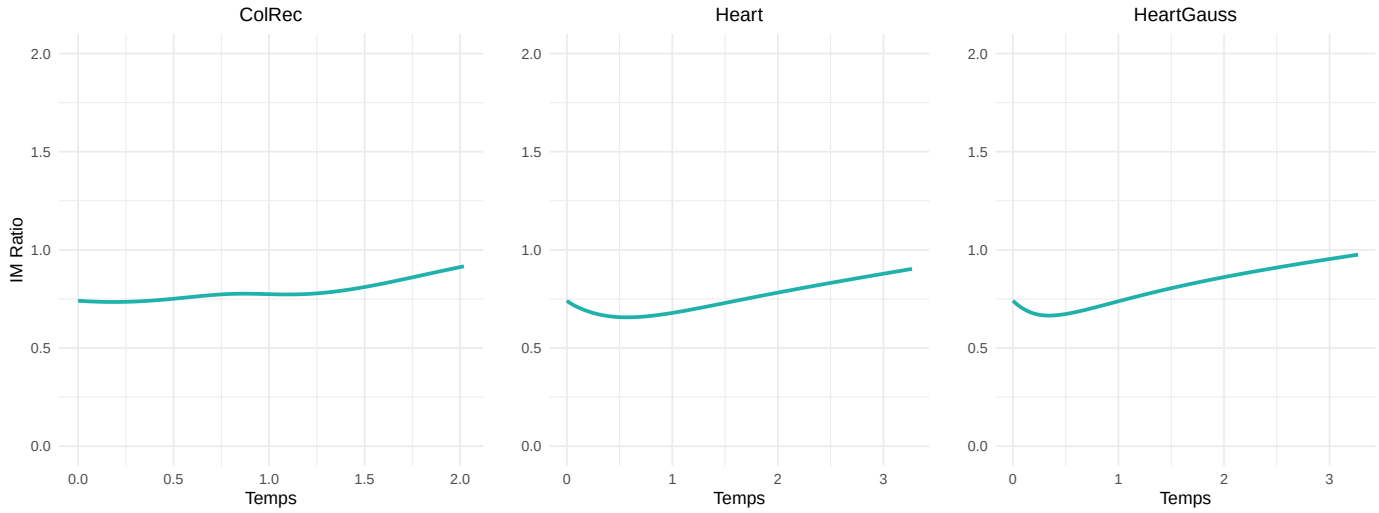

FIGURE 3.8 – Ratios d'IM des scénarios ColRec, Heart et HeartGauss. *Référence : $x = 1$*

Cette formulation implique donc une spline du temps pour chaque niveau de la covariable x.

Tous les modèles ont été ajustés avec le package R `survPen`, excepté le modèle non pénalisé avec effets aléatoires qui a été ajusté avec le package R `mexhaz`. L'ajustement du modèle mixte pour des évènements récurrents a nécessité une modification de la fonction `mexhaz` qui implémente la vraisemblance tenant compte de la troncature à gauche (voir section 1.2.2). Cette implémentation est disponible à partir de la version 2.6. Le modèle semi-paramétrique d'Andersen-Gill avec estimateur robuste de la variance a également été ajusté à titre de référence avec le package R `survival`.

Dans le cadre non pénalisé, les variances V_{Robuste} et $V_{\text{Naïf}}$ ont été estimées pour $log(\rho)$ pour le modèle direct. Le modèle indirect a été ajusté avec la variance V_{Mixte}. La spline cubique a été spécifiée avec 5 noeuds pour les modèles non pénalisés. Un modèle semi-paramétrique avec une variance robuste (que nous appellerons "modèle de Cox") a également été ajusté accompagné

Scénario	Nb de noeuds	Effet de x	Modèles	Package
Non pénalisé				
Poiss, PGauss, PGamma	5	Proportionnel	Direct	survPen
			Indirect	mexhaz
ColRec, Heart, HeartGauss	5	Non proportionnel	Direct	survPen
			Indirect	mexhaz
Pénalisé				
Poiss	10	Proportionnel	Direct	survPen
PGauss, PGamma	10	Proportionnel	Direct	survPen
			Indirect	survPen
ColRec, Heart, HeartGauss	10	Non proportionnel	Direct	survPen

TABLE 3.2 – Caractéristiques des modèles ajustés sur les échantillons simulés.

du test des résidus du Schoenfeld.

Dans le cadre pénalisé, les variances robuste V_{Robuste^*} et bayésienne empirique sous les hypothèses du processus de Poisson $V_{\text{Naïf}^*}$ ont été estimées pour les modèles directs. Un intervalle de confiance par échantillonnage bootstrap a également été calculé pour le modèle direct. Les bornes de l'intervalle ont été déterminées par les percentiles de la distribution des 300 échantillons de boostrap. Du fait de sa mauvaise performance et des coûts de calculs très importants du modèle pénalisé indirect, ce dernier n'a été ajusté que pour les scénarios PGauss et PGamma. Les modèles pénalisés ont été ajustés en utilisant des splines de régression cubiques à 10 noeuds

3.2.3 Comparaison des modèles

La performance des modèles a été évaluée en estimant l'IM en différents temps (100 points également répartis sur l'espace de définition) pour les deux modalités de la covariable x.

On note $\rho(t,x)$ la valeur théorique de l'IM au temps t pour la covariable x et $\hat{\rho}_d(t,x)$ son estimation dans le d^e échantillon simulé. Les indicateurs suivants sont considérés :

— Les probabilités de couverture empiriques, définies comme la proportion d'intervalles de confiance à 95% qui incluent la valeur théorique ;

— Le biais relatif moyen dans l'estimation de l'IM : $\frac{1}{D}\sum_{d=1}^{D}\frac{\hat{\rho}_d(t,x)-\rho(t,x)}{\rho(t,x)}$;

— Le ratio d'erreurs-types estimées versus empiriques, défini par :

$$SER = \frac{\sigma_{est}}{\sigma_{emp}}$$

où,

— $\sigma_{est} = \frac{1}{D}\sum_{d=1}^{D} \sigma_{\log(\hat{\rho}_d(t,x))}$;

— $\sigma_{emp} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left[\log(\hat{\rho}_d(t,x) - \overline{\log(\rho)}(t,x)\right]^2}$;

— $\overline{\log(\rho)}(t,x) = \frac{1}{D}\sum_{d=1}^{D}\log(\hat{\rho}_d)(t,x)$

— La racine de l'erreur quadratique moyenne(RMSE), définie par :

$$\sqrt{\frac{1}{D}\sum_{d=1}^{D}\left[\hat{\rho}_d(t,x) - \rho(t,x)\right]^2}$$

3.2.4 Résultats

Les caractéristiques des échantillons simulés sont présentées en Table 3.3.

		Proportions					
Scenario	*Moyenne*	0	1	2	3	4+	Max
Poiss	2.4	10.0%	22.4%	25.7%	20.2 %	21.8 %	13
PGauss	2.6	12.0%	21.7 %	22.4 %	17.5 %	26.5%	21
PGamma	2.4	21.3 %	22.9 %	18.3 %	13.1%	24.4 %	27
ColRec	1.7	19.0 %	26.9 %	16.3 %	37.8%	0.0 %	3
Heart	2.5	19.5 %	18.4 %	17.4 %	8.3%	36.5 %	5
HeartGauss	2.7	19.5 %	16.7 %	15.0 %	7.0%	41.8%	5

TABLE 3.3 – Moyennes, proportions d'individus selon leur nombre total d'évènements et nombre maximum d'évènements par sujet pour les différents scénarios simulés

Cadre non pénalisé

Comme attendu, la méthode directe et la méthode indirecte sont équivalentes en termes de biais (Figure 3.9). Ce dernier est faible ($< 5\%$) dans tous les scénarios. Les probabilités de couverture des prédictions de l'IM en différents temps pour une valeur de la covariable x de 0 sont présentées en Figure 3.10. Les modèles Direct avec variance robuste et Indirect sont proches du nominal dans tous les scénarios. Pour les scénarios Heart (processus de type multi-états), les intervalles de confiance du modèle Indirect présentent une légère sous-couverture sur le début de la période de temps. La sous-couverture est aggravée par l'ajout de l'effet aléatoire (HeartGauss).

Dans le scénario ColRec, le modèle Indirect n'estime aucun effet aléatoire (voir Figure 3.11). Cela peut s'expliquer par la faible corrélation introduite sur l'IM et par le faible nombre maximum d'évènements par patient. A l'inverse, dans le scénario Poiss, un effet aléatoire est parfois estimé malgré l'absence de corrélation simulée.

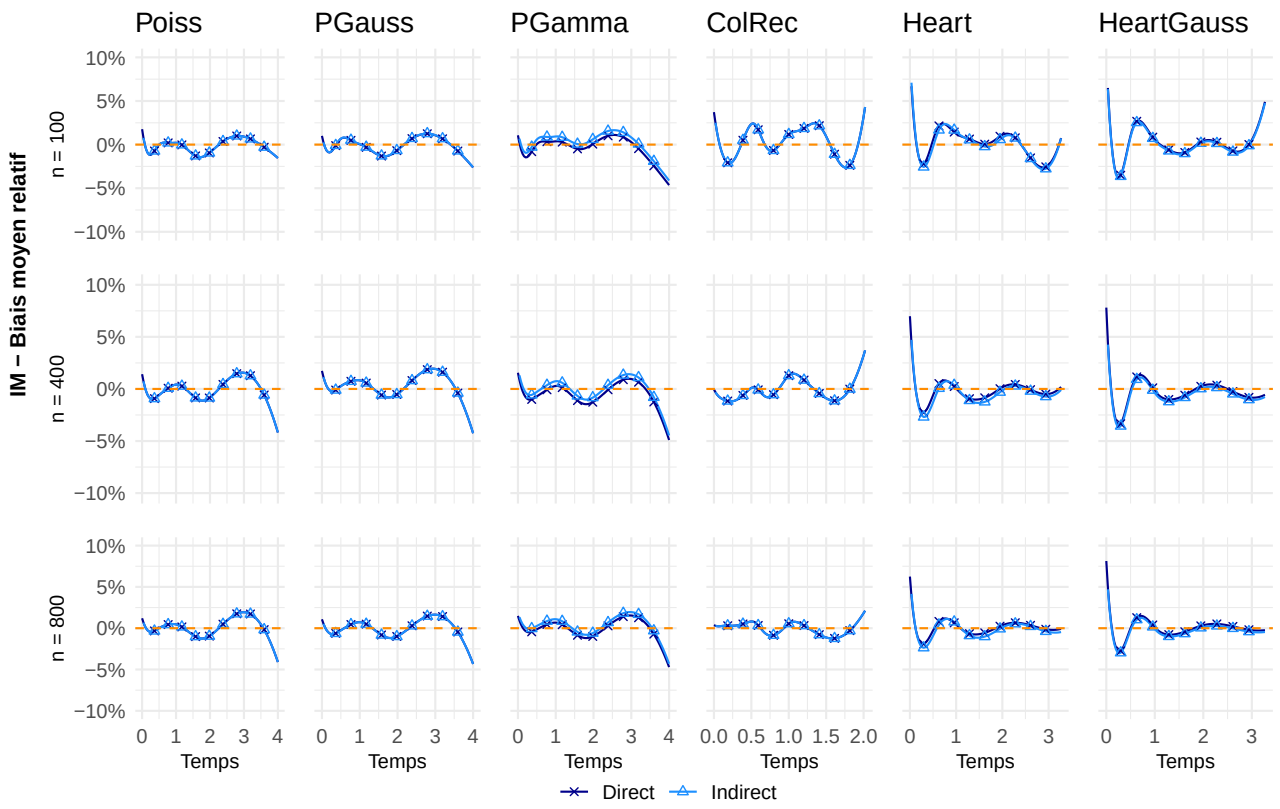

FIGURE 3.9 – **Cadre non pénalisé** - Biais moyen relatif sur les estimations de l'IM (pour x=0)

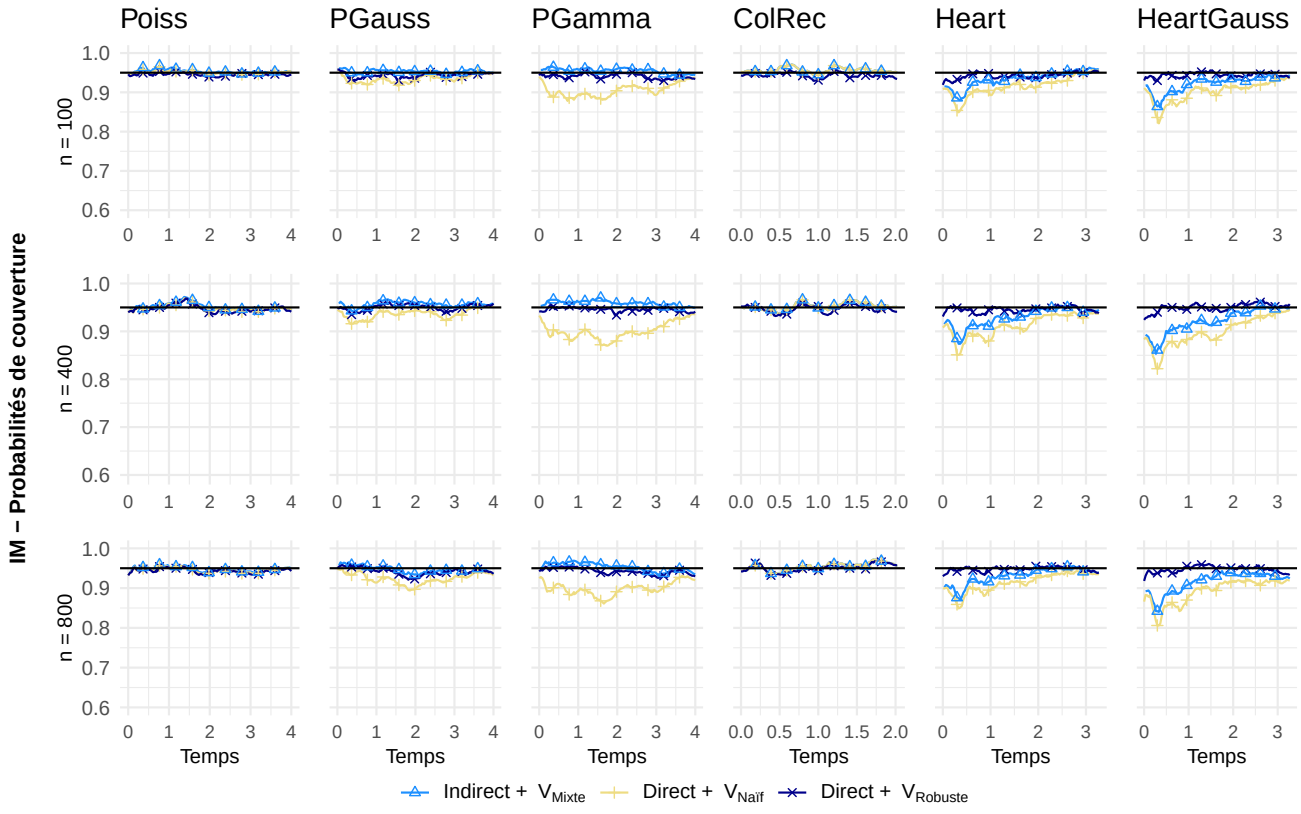

FIGURE 3.10 – **Cadre non pénalisé** - Probabilités de couverture de l'IM prédit en fonction du temps (pour x=0)

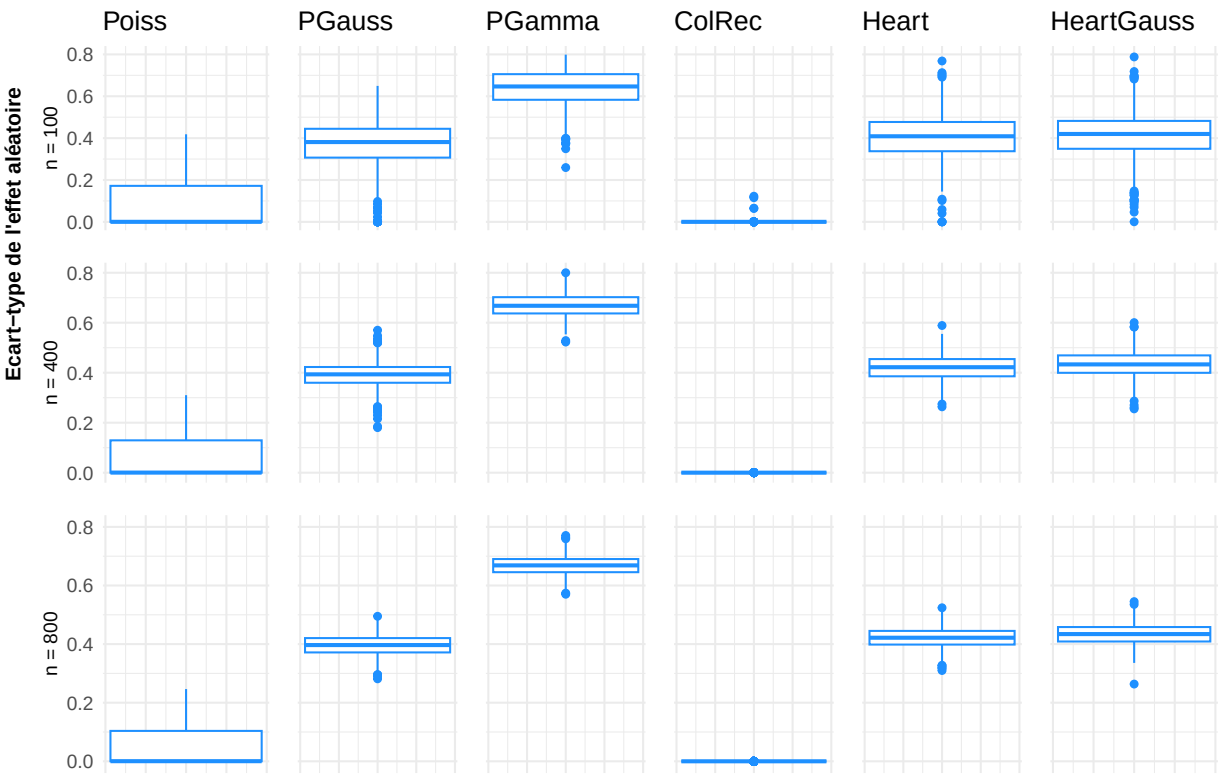

FIGURE 3.11 – **Cadre non pénalisé** - Ecarts-types des effets aléatoires estimés dans le modèle indirect.

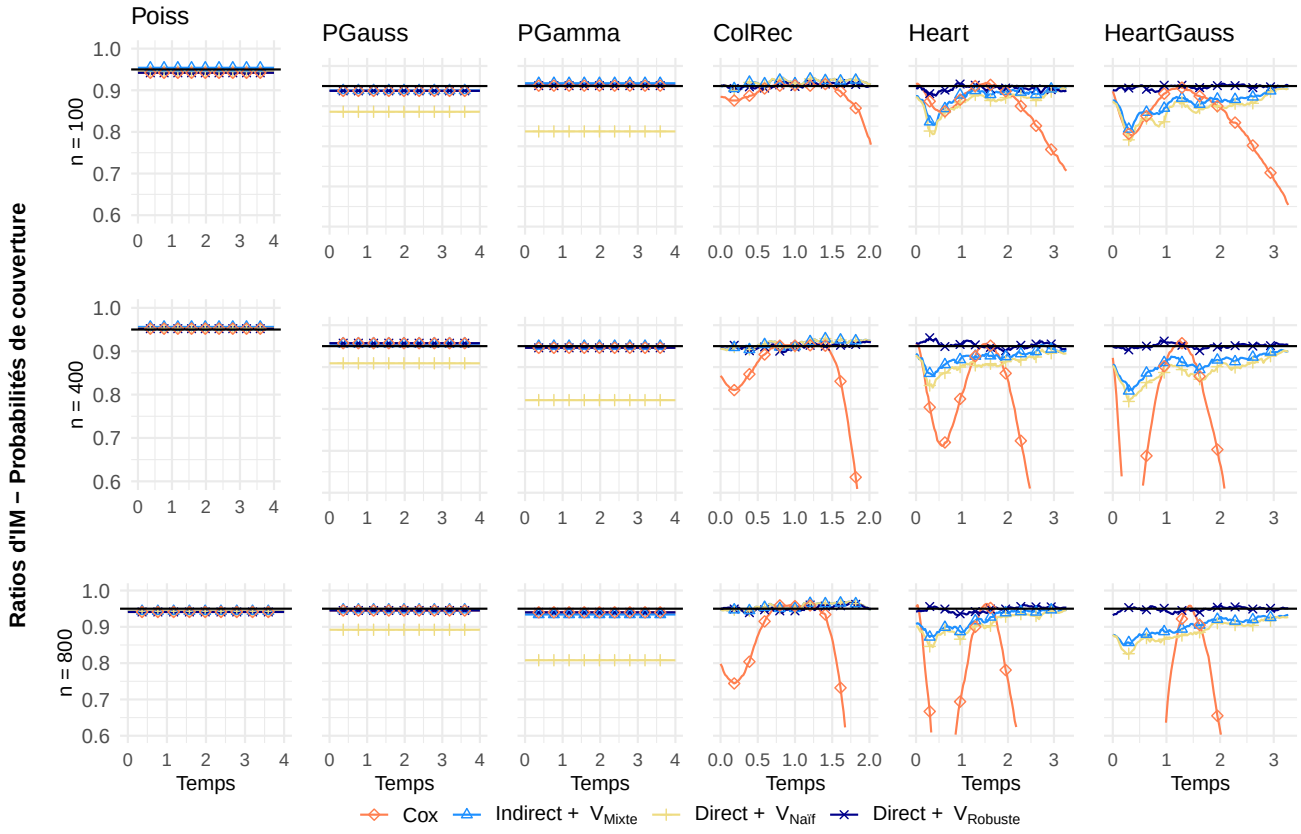

FIGURE 3.12 – **Cadre non pénalisé** - Probabilités de couverture des ratios d'IM en fonction du temps (x=1 vs x=0).

La Figure 3.12 présente les probabilités de couverture pour les ratios IM des modèles Direct et Indirect mais aussi pour le modèle de Cox (avec variance robuste). Les résultats sont similaires à ceux observés sur l'IM. Les probabilités de couverture pour le modèle Direct avec variance robuste sont les mêmes que dans le modèle de Cox malgré l'estimation additionnelle de la dynamique de l'IM. Notons que les résidus de Schoenfeld en Table 3.4 échouent à rejeter l'hypothèse de proportionnalité à un seuil de 5% pour les scénarios ColRec, Heart et HeartGauss malgré le biais très important observé dans le modèle de Cox.

	Proportions de rejet		
Scenario	*n=100*	*n=400*	*n=800*
Poiss	4.4 %	5.1%	6.0%
PGauss	4.8 %	5.7%	5.0 %
PGamma	5.3 %	5.5 %	4.2 %
ColRec	6.9 %	10.1 %	11.9 %
Heart	18.4 %	25.9 %	36.9 %
HeartGauss	25.3 %	42.1 %	58.0 %

TABLE 3.4 – Proportion de rejet du test de proportionnalité de Schoenfeld

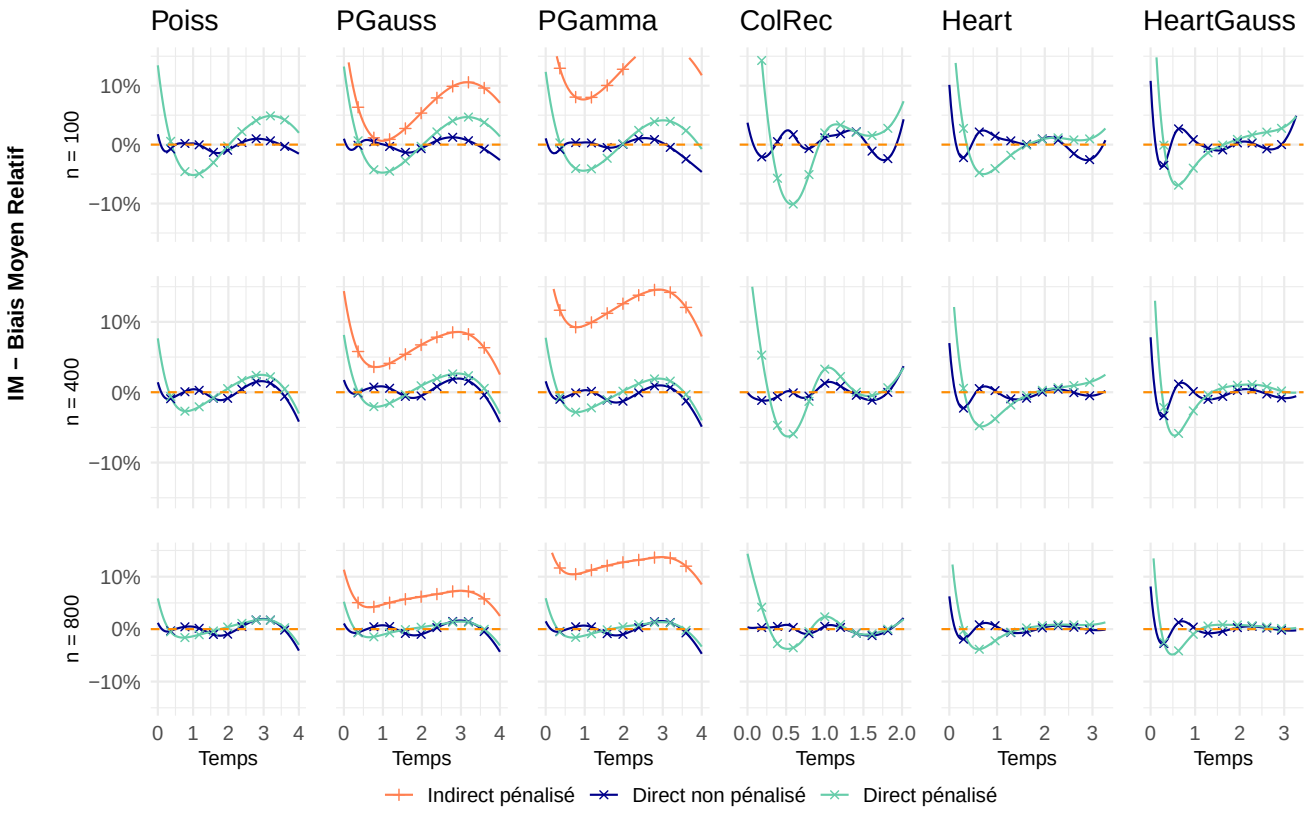

FIGURE 3.13 – **Cadre pénalisé** - Biais moyen relatif

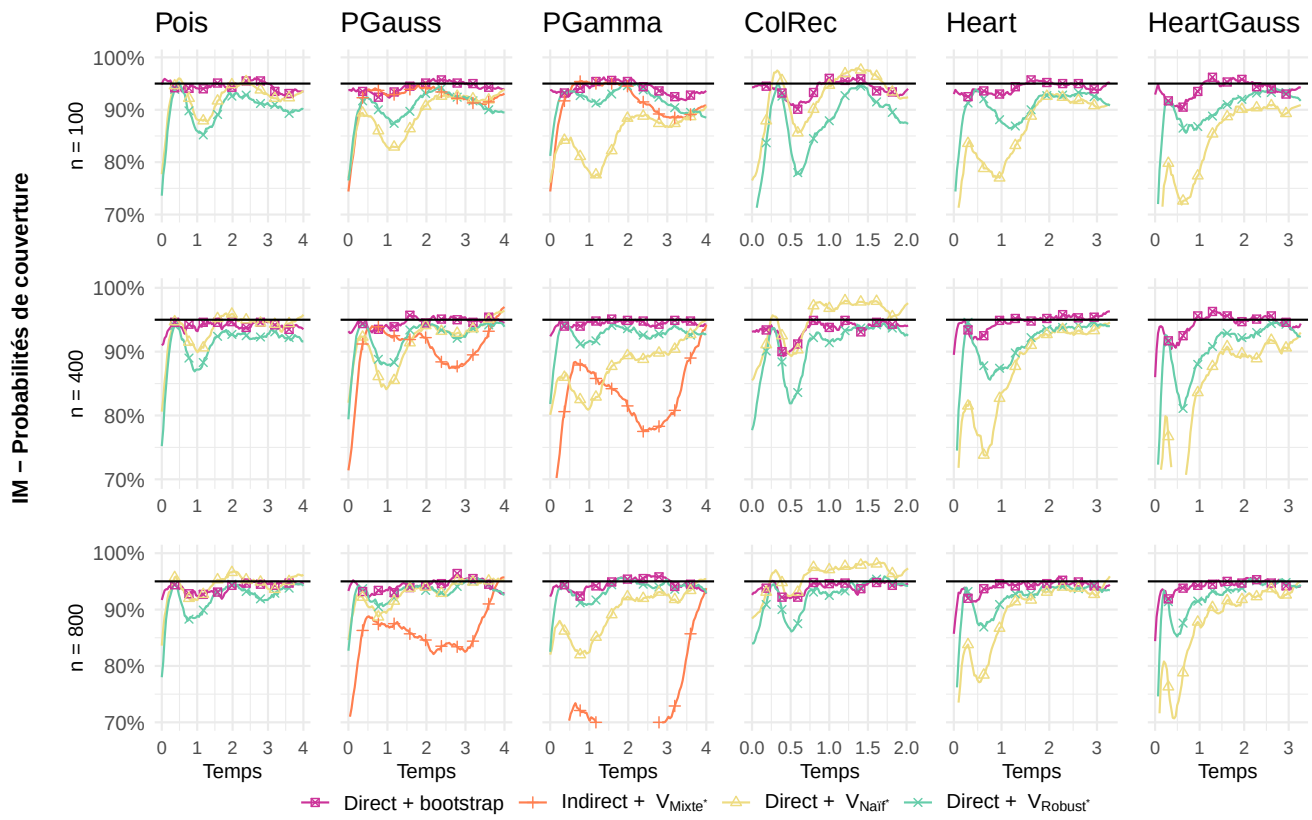

FIGURE 3.14 – **Cadre pénalisé** - Probabilités de couverture des estimations de l'IM

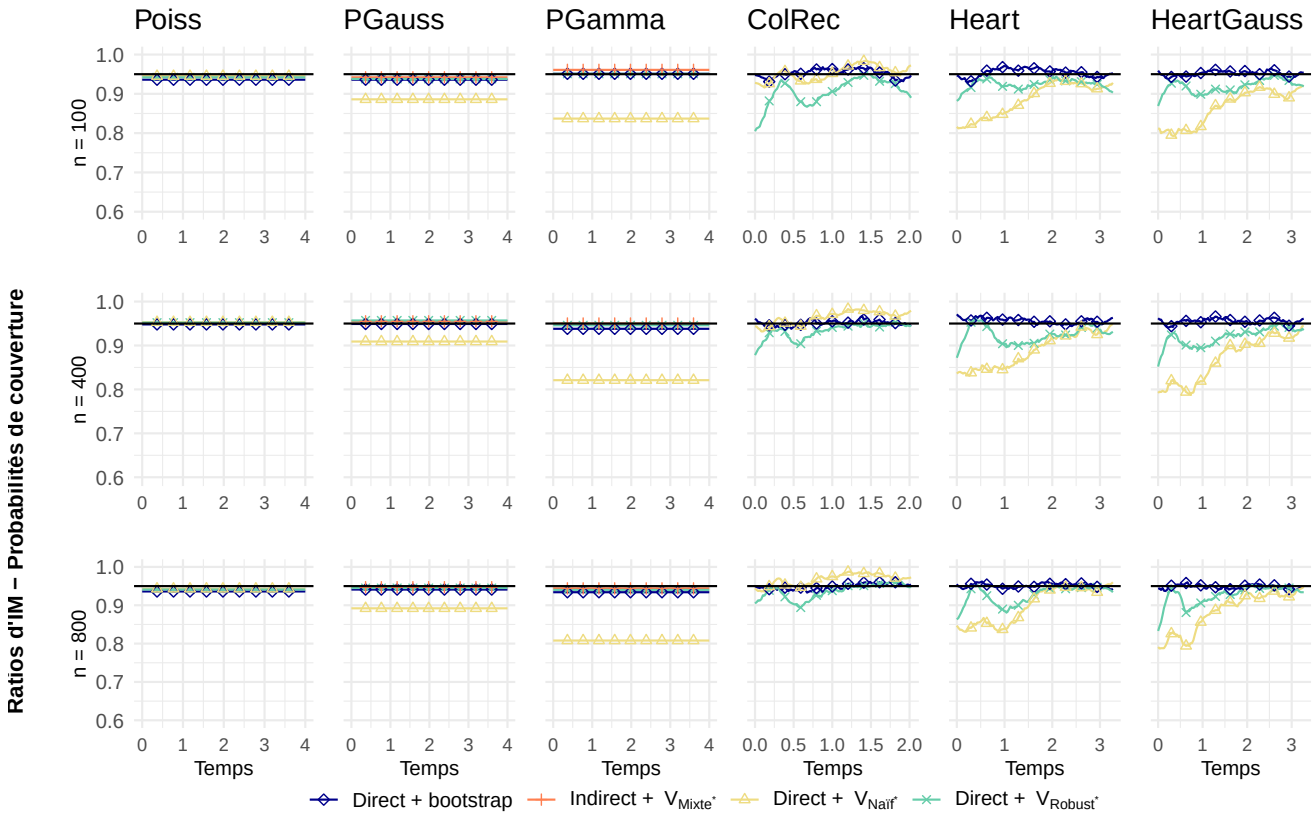

FIGURE 3.15 – **Cadre pénalisé** - Probabilités de couverture des estimations des ratios d'IMs

Modèles pénalisés

La Figure 3.14 présente les probabilités de couverture de l'IM pour les modèles pénalisés. Le modèle Direct + V_{Robuste^*} est plus proche du nominal que le modèle Indirect + $V_{\text{Naïf}^*}$ sauf pour les scénarios Poiss et ColRec. Cependant, les probabilités de couverture sont systématiquement inférieures à 95% dans tous les scénarios. Les intervalles de confiance obtenus par boostrap atteignent des probabilités de couvertures proches du nominal dans tous les scénarios. Dans les scénarios PGauss et PGamma, les probabilités de couverture du modèle Indirect + V_{Mixte^*} ont de mauvaises probabilités de couverture du fait d'un biais important, visible sur la Figure 3.13. De plus, le biais s'aggrave avec l'augmentation de la taille des échantillons. Il n'est, en revanche, pas observé sur l'estimation des ratios d'IM en Figure 3.15. Enfin, le modèle est également beaucoup plus long à ajuster que le modèle Direct (même avec les intervalles de confiance de type bootstrap). Les probabilités de couverture des ratios d'IM sont présentés en Figure 3.15. Dans les scénario Poiss, PGauss et PGamma, la pénalisation n'impacte pas les probabilités de couverture des modèles car ils sont proportionnels. De plus le biais important observé sur la prédiction de l'IM dans le modèle mixte pénalisé n'impacte pas les ratios d'IM. Dans les scénarios ColRec, Heart et HeartGauss, l'estimateur robuste conduit à une sous-couverture des ratios d'IM. En revanche, les intervalles obtenus par échantillonnage bootstrap sont proches du nominal.

3.3 Application : Jeu de données *Staphylococcus aureus*

Le modèle Direct d'IM a été appliqué au jeu de données *Staphylococcus aureus* du package `pammtools` (Abdulgader et al., 2019). Ces données font partie de la *Drakenstein Child Health study*, une cohorte d'étude de population de 1143 couples mères-enfants de la communauté péri-urbaine du Cap en Afrique du Sud (Zar et al., 2015). Des échantillons nasopharyngés (NP) ont été collectés dans le but de déterminer la dynamique de portage nasopharyngé de *Staphylococcus aureus*. Le jeu de données comprend 137 couples mère-enfants dans lesquels les enfants ont atteint leur première année de vie avec un minimum de 18 échantillons NP. Ces derniers ont servi à définir des évènements d'acquisition qui sont ici l'évènement récurrent d'intérêt. Le statut HIV, est défini par le fait d'être un enfant non-infecté par le VIH, né d'une mère infectée par le VIH. Il est étudié comme étant un potentiel prédicteur de l'acquisition. Un total de 242 évènements d'acquisition ont été recensés dans le jeu de données.

Nous considérons les modèle d'IM pénalisé suivant :

$$\log(\rho(t)) = s(t) + s_1(t)I(\text{hiv} = 1) \tag{3.7}$$

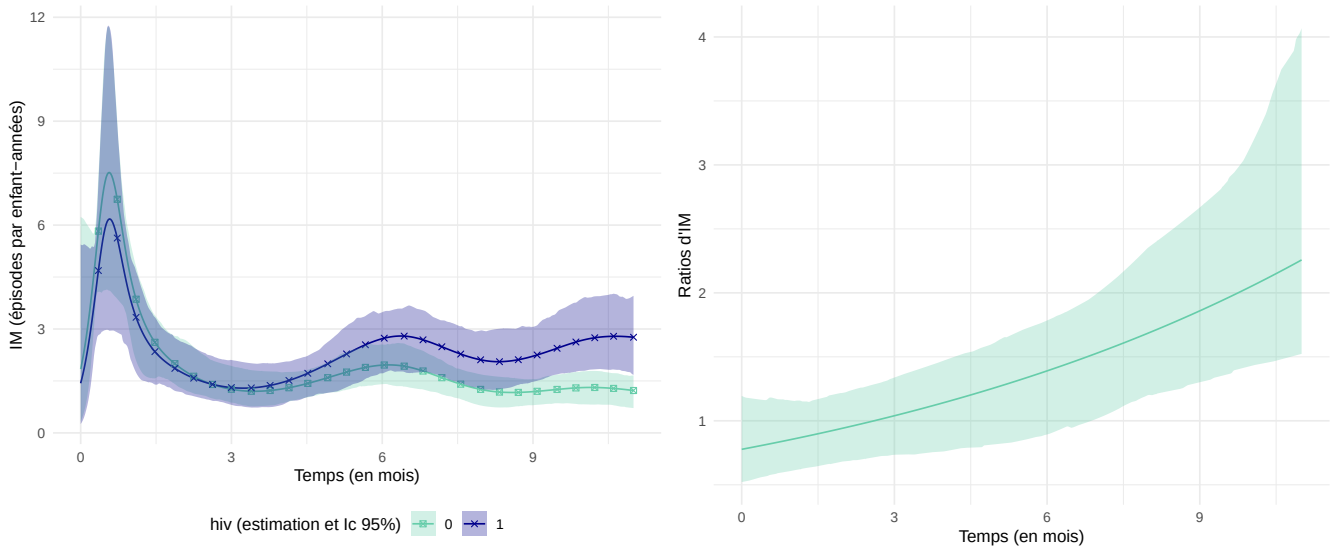

FIGURE 3.16 – *(gauche) :* IM d'acquisition d'évènements au cours du temps, *(droite) :* Ratios d'IM en fonction du temps

Dans cette situation, s est la spline sur l'effet du temps pour $hiv = 0$, alors que s_1 est la spline pour la différence, en termes de dynamique de l'effet du temps, entre les catégories 1 et 0. Un paramètre de lissage est associé avec s, et un autre avec s_1. Cela signifie que le modèle pénalise la différence en termes de dynamique de l'effet du temps, entre les deux catégories de *hiv*. En passant à l'exponentielle, on en déduit que c'est le ratio d'IM qui est en fait pénalisé. Cette formulation du modèle est donc différente de celle que nous avons utilisée dans l'étude de simulation, dans laquelle la pénalisation est appliquée sur chaque spline de façon séparée. Dans cet étude, nous choisissons de pénaliser les différences car nous avons un *a priori* sur le fait que la dynamique soit similaire entre les deux groupes. On utilise des splines cubiques naturelles avec 10 noeuds, les noeuds extérieurs étant les temps maximum et minimum et les noeuds intérieurs sont localisés du 10 au 90^e percentiles de la distribution des temps d'évènements. Les intervalles de confiance sont construits par bootstrap avec 300 tirages.

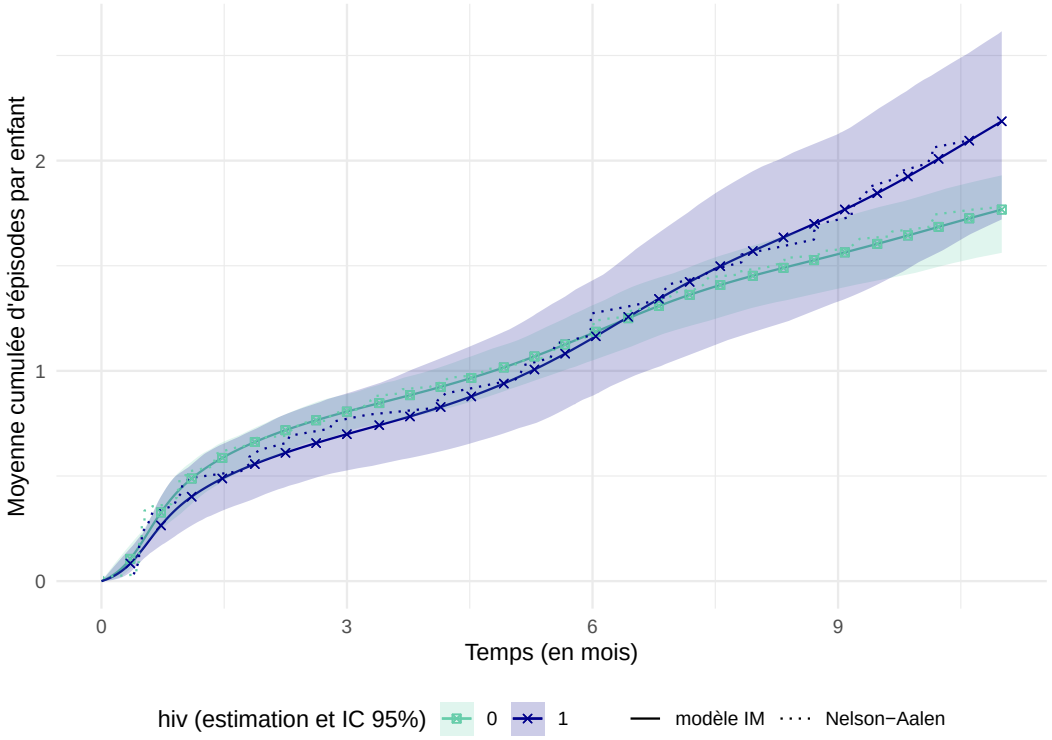

FIGURE 3.17 – Moyenne cumulée d'évènements estimés à partir du modèle d'IM flexible (ligne continue) et par l'estimateur non paramétrique de Nelson-Aalen (ligne pointillée). *L'IC à 95% est obtenu à partir du modèle flexible par échantillonnage bootstrap.*

La Figure 3.16 (gauche) décrit la dynamique de l'IM au cours du temps pour les groupes HIV/non HIV. Les trajectoires sont similaires dans le début du suivi mais l'IM semble plus haut dans le groupe HIV après 4 mois. On observe cela sur la courbe des ratios d'IM, en Figure 3.16 (droite), qui semble décrire un effet non proportionnel (test des résidus de Schoenfeld[1] - p-value : 0.039). Pour ce qui est de l'impact de l'exposition à HIV sur l'apparition du premier évènement, Abdulgader et al. 2019 ont conclu qu'il n'y avait pas de différence. En considérant tous les évènements, une différence est notable entre les exposés et non exposés, avec des ratios d'IM significativement plus élevés dans le groupe HIV après 7 mois. Une IM plus élevée est observée au cours des deux premières semaines, comme décrit par Abdulgader et al. 2019. Un second pic d'acquisitions est dépeint par le modèle à 6 mois. La Figure 3.17 décrit la moyenne cumulée d'évènements par enfant au cours du temps, estimée à partir du modèle flexible de l'IM et par le modèle non-paramétrique de Nelson-Aalen. Les deux courbes sont proches l'une de l'autre.

1. Obtenu en ajustant un modèle de Cox sur le jeu de données

3.4 Application : Evènements indésirables dermatologiques dans la cohorte des mélanomes

3.4.1 Contexte & Objectifs

Nous allons reproduire l'analyse réalisée dans la partie II, section 2.4, sur les évènements dermatologiques des patients de la cohorte des mélanomes, mais en considérant plusieurs épisodes de toxicité au lieu du premier. Pour cela, nous considérons les modèles sur l'IM tels qu'introduits dans le chapitre précédent.

3.4.2 Méthode

Un premier modèle est ajusté sans autre variable que le temps avec et sans pénalisation sur le taux au premier évènement et sur l'IM (modèle temps).

Ensuite, les mêmes variables qu'en section 2.4 sont introduites dans le modèle. Les mêmes trois modèles candidats ont été ajustés comparés par critère AICc (i) sans interaction, (ii) interaction entre l'âge et le temps (iii) interaction entre le NLR et le temps.

Les intervalles de confiance ont été estimés par bootstrap pour les modèles pénalisés et avec l'estimateur robuste pour les modèles non pénalisé.

3.4.3 Résultats

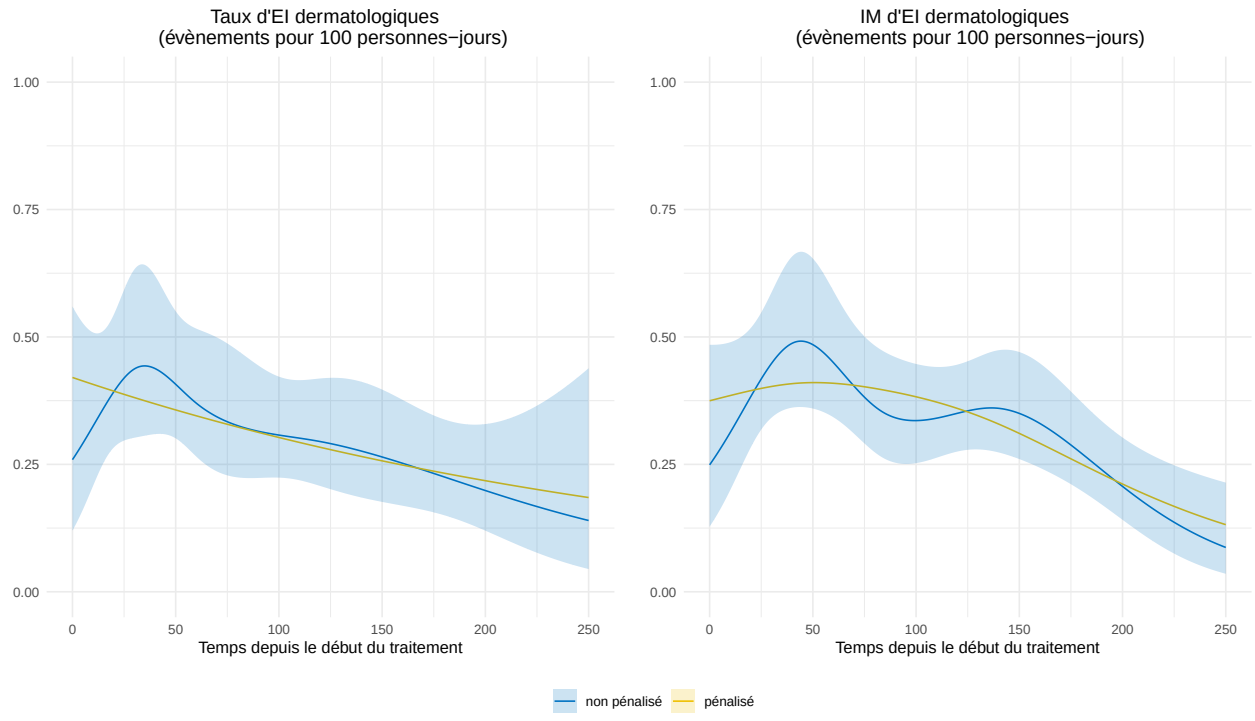

FIGURE 3.18 – (*gauche)* : Taux au premier évènement, (*droite)* : IM. *Pour des raisons de lisibilité seuls les intervalles de confiance non pénalisés sont présentés sur cette figure.*

La Figure 3.18 présente le taux jusqu'au premier évènement et l'IM des épisodes d'EI dermatologiques chez les patients en cours de traitement (modèles temps). Une seconde vague d'évènements semble survenir autour des 150 jours.

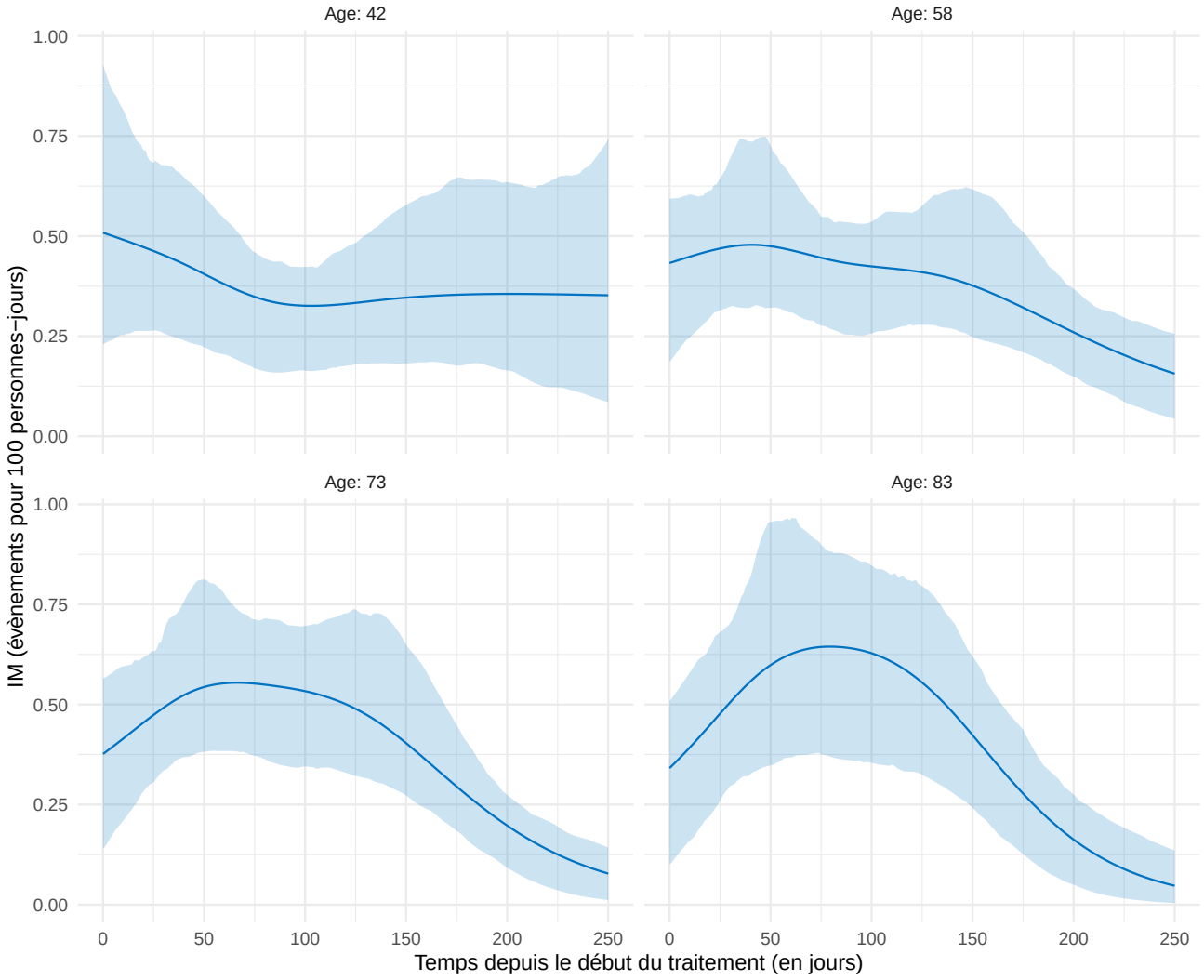

FIGURE 3.19 – Effet de l'âge sur l'IM d'EI dermatologiques sur l'IM d'EI dermatologiques. *Les courbes d'IM sont prédites pour différents âges : 42, 58, 73 et 83 ans (quantiles 0.1, 0.33, 0.66 et 0.9)*

Dans le cas de l'IM, le meilleur modèle selon l'AICc est le modèle avec une interaction entre le temps et l'âge (Modèle 2) (AICc respectifs : 2270, 2265, 2272). La Figure 3.19 présente l'IM pour 4 valeurs d'âges. Les courbes sont prédites pour les valeurs de référence sur les autres variables. La dynamique d'IM semble dépendre de l'âge. Les plus jeunes présentent une IM relativement plate alors que les plus âgés présentent une forme en cloche avec un maximum à 3 mois après le début du traitement. En Figure 3.20, l'effet du NLR est non significatif, de même que pour les variables à effet non proportionnel (excepté le pays) comme dans le modèle à un évènement.

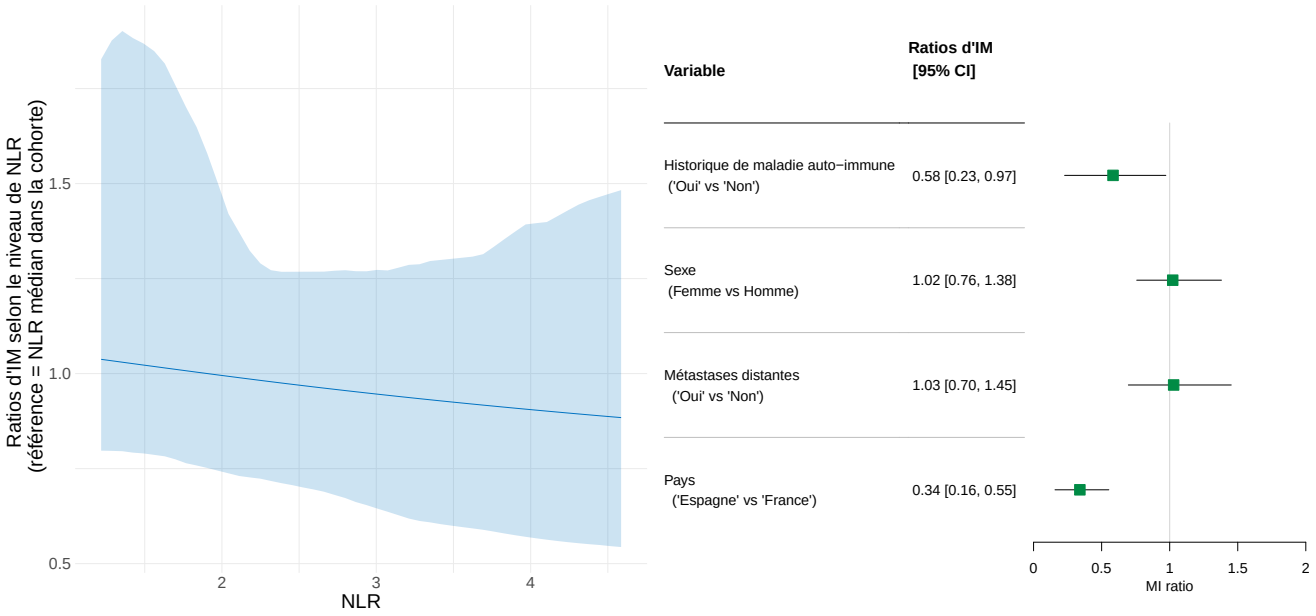

FIGURE 3.20 – Effets du NLR (*gauche)* et des variables avec effet proportionnel (*droite)* sur l'IM d'EI dermatologiques.

4 | Discussion et Conclusion

L'intensité marginale (IM)

Ces travaux ont permis l'exploration de la modélisation flexible de l'IM dans le but de décrire la survenue d'évènements récurrents ainsi que son association avec des covariables. Deux approches ont été considérées, l'une directe basée sur une équation d'estimation et l'autre indirecte basée sur un modèle d'intensité classiquement utilisé dans la littérature. D'un point de vue pratique, l'estimation de l'IM est directement réalisable à partir d'un logiciel de survie gérant la modélisation flexible comme `survPen` ou `mexhaz`. L'estimateur robuste de la variance a néanmoins dû être développé.

Les principaux avantages de l'IM en tant qu'indicateur est qu'il est très synthétique, ce qui en fait un bon candidat pour faire une description générale d'un processus à l'échelle d'une population. En l'absence de censure informative, l'IM est directement reliée au nombre moyen cumulé d'évènements, qui est une quantité plus familière que l'IM pour des non-statisticiens. Présenter le nombre moyen cumulée peut ainsi favoriser la communication des résultats.

Lorsque la question de recherche vise des trajectoires individuelles, l'IM n'est évidemment pas adaptée. De plus, elle tend à être assez plate lorsque la proportion de patients sans évènement est importante. Dans ce cas, des alternatives pourraient être envisagés comme le modèle à évènements récurrents avec excès de zéros (Ma and Crimin, 2024). L'approche à deux modèles (un taux jusqu'au premier évènement et une IM pour les évènements suivants) proposée par Pepe and Cai 1993 peut également être considérée, et le cadre d'estimation détaillé dans cette thèse peut tout à fait s'accommoder de cette double modélisation.

Etude de simulation

L'une des forces de l'étude de simulation réside dans l'effort déployé pour prendre en compte un processus de génération des données distinct de celui du processus de Poisson. Le processus multi-états (avec ou sans effets aléatoires) introduisent une corrélation intra-sujet liée au processus lui-même, un aspect qui, à notre connaissance, n'a jamais été pris en compte pour évaluer un modèle d'IM. La nécessité de définir des intensités conditionnelles a toutefois rendu indispensable le calcul numérique de l'IM théorique. Ce calcul est exigeant en termes de ressources informatiques, et cette exigence augmente avec le nombre maximal d'événements par individu. En outre, les intégrales sur l'historique du processus deviennent impossibles à réaliser lorsque le processus devient trop complexe.

L'approche directe

Dans le cadre non pénalisé, l'approche directe avec l'estimateur robuste de la variance a de bonnes performances, meilleure que l'approche indirecte en termes de probabilités de couver-

ture. Dans les scénarios proportionnels, estimer la dynamique de l'IM ne dégrade pas la précision des estimations ponctuelles des effets des covariables et de leurs intervalles de confiance. Les ratios d'IM sont aussi précis que ceux du modèle semi-paramétrique. Si la dynamique de l'IM est complexe, augmenter le nombre de noeuds conduit à un fort risque de sur-ajustement, comme observé dans la seconde partie de cette thèse. La pénalisation est une solution intéressante pour faire de la sélection de modèle automatique et limiter ce phénomène de sur-ajustement

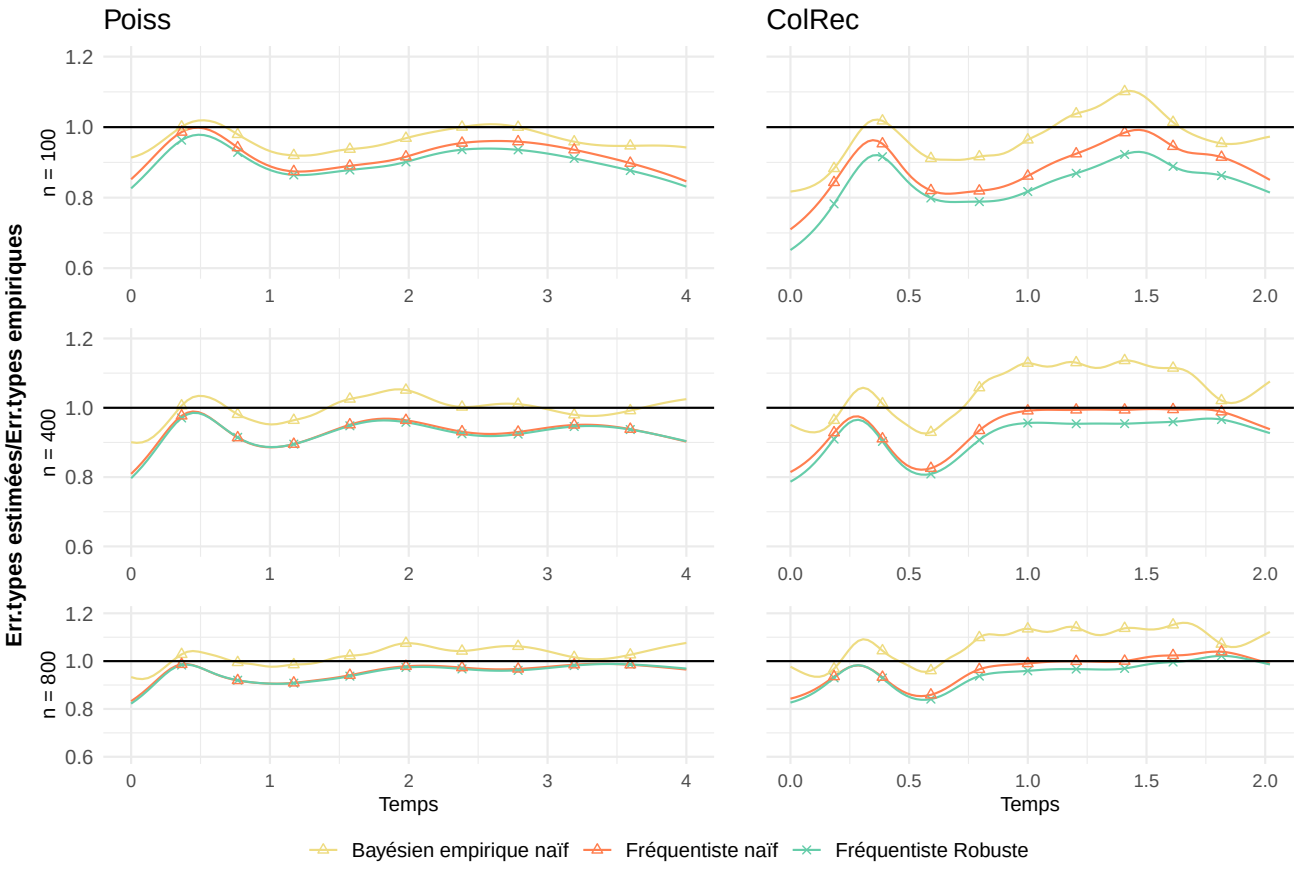

FIGURE 3.21 – Ratios d'erreur-type estimée sur erreur-type empirique

Dans le cadre pénalisé, l'estimateur robuste conduit à des variances systématiquement sous-évaluées. La pénalisation réduit la RMSE mais introduit un biais, ce qui complique l'inférence. De façon alternative, les intervalles de confiance obtenus par bootstrap produisent des probabilités de couvertures très satisfaisantes.

Nous avons observé que les probabilités de couverture de l'estimateur de la variance naïf étaient meilleures que celles obtenues avec l'estimateur de la variance robuste dans les scénarios Poiss et ColRec. Ce comportement s'explique par deux effets. Tout d'abord, nous avons constaté qu'il n'existe pas, ou très peu (dans le cas de ColRec), de corrélation entre les événements intra-individuels dans ces scénarios. Ainsi, l'estimateur robuste n'apporte aucune correction par rapport à l'estimateur naïf. Le second effet découle de la manière dont les estimateurs sont construits. L'estimateur naïf est un estimateur Bayésien empirique (Wahba, 1985) contrairement à l'estimateur sandwich robuste qui est un estimateur fréquentiste (Gray, 1992). Lorsque l'on calcule les ratios d'erreur-type estimée sur erreur-type empirique (Figure 3.21), on re-

marque que celui obtenu par l'estimateur naïf et l'estimateur robuste sont très proches. La meilleure performance de l'estimateur naïf par rapport au robuste dans ces scénarios est donc principalement expliquée par la différence de paradigme Bayésien/Fréquentiste. En présence de pénalisation, les intervalles de confiance doivent tenir compte du biais de lissage. Cette composante est, en quelque sorte, prise en compte dans l'intervalle de confiance Bayésien (Nychka, 1988), ce qui explique ses bonnes propriétés de couverture. Ainsi, constuire des intervalles de confiance bayésiens robustes (Armstrong et al., 2022) pourrait améliorer les estimateurs de la variance et éviter le bootstrap.

Toujours dans le cadre pénalisé, le choix des paramètres de lissage est basé sur le critère LAML. Ce dernier est construit à partir de la vraisemblance de Poisson, et ne tient donc pas compte de la corrélation entre les évènements intra-sujets. Pour cette raison, le choix des paramètres pourrait ne pas être optimal. Cependant, en pratique, le modèle donne des résultats satisfaisants.

L'approche indirecte

La possibilité de modéliser l'IM à partir d'un modèle d'intensité de type processus de Poisson avec effet aléatoire a été étudiée comme approche alternative à l'approche par équations d'estimations (directe). Dans le cadre non pénalisé, le package `mexhaz` a été utilisé pour ajuster le modèle via une maximisation de la vraisemblance marginale. L'intégrale sur la distribution des effets aléatoires est réalisée avec une quadrature adaptative de Gaus-Hermite. Dans ce cas, les estimations de l'IM sont non biaisées et les intervalles de confiance obtenus sont plutôt robustes à la mispécification, malgré une sous-couverture dans le cas multi-états. En revanche, dans le cadre pénalisé, le package `survPen` utilise une méthode pénalisée pour estimer les paramètres du modèle mixte (Ripatti and Palmgren, 2000) (voir partie II, section 1.2.2). Du fait du faible nombre d'évènements par individu (groupe), estimer le maximum *a posteriori* à la place du maximum de vraisemblance, conduit à un biais. L'annexe B propose une illustration de ce phénomène par une simulation issue d'une distribution de Poisson. Les estimations de la dynamique des IM sont fortement biaisées en raison du manque de précision de la méthode pénalisée lorsque le nombre d'observations par individu est faible. Par conséquent, cette méthode ne semble pas recommandée dans ce contexte. En particulier, ce problème pourrait concerner les modèles d'intensité à effets aléatoires ajustés par le package R `mgcv` (Wood, 2017) (e.g. modèles exponentiels par morceaux (Ramjith et al., 2024)). Cependant, comme observé dans l'étude de simulation, le biais introduit par le mode de calcul ne semble pas affecter les ratios.

IM et évènement terminal

Dans ces travaux, nous avons supposé que le mécanisme de censure éventuel était indépendant du processus récurrent. Cependant, en présence d'un évènement terminal, cette hypothèse ne tient plus.

Dans le cadre du modèle direct d'IM, l'équation d'estimation 2.3 est biaisée ($\mathbb{E}\left(\mathbf{U}(\boldsymbol{\beta}_0)|\mathbf{X(t)}\right) \neq 0$). Il est cependant possible de redéfinir l'indicateur conditionnellement à l'absence de survenue de l'évènement terminal (Lin et al., 2000; Cook and Lawless, 2007)(p 222). En notant D, le temps de survenue de l'évènement terminal :

$$\bar{\rho}(t) = \mathbf{E}[dN(t)|D \geq t] \tag{4.1}$$

L'indicateur s'interprète donc maintenant comme une probabilité d'évènement par unité de temps chez les individus n'ayant pas encore présenté l'évènement terminal à l'instant t. En redéfinissant ainsi l'intensité marginale, l'équation d'estimation est à nouveau non biaisée.

$$\mathbb{E}\left(\mathbf{U}(\boldsymbol{\beta}_0)|\mathbf{X}(t), D \geq t\right) = 0$$

Le cadre direct d'estimation et d'inférence proposé précédemment tient donc toujours.

En revanche, de façon analogue au cadre des risques compétitifs avec le taux, l'IM n'est plus directement reliée à la moyenne cumulée d'évènements mais par la relation suivante (Cook and Lawless, 2007) :

$$\mu(t) = \int_0^t \bar{S}(s)\bar{\rho}(s)ds \tag{4.2}$$

où, $\bar{S}(t) = \exp\left(-\int_0^t \bar{h}(s)ds\right)$ et $\bar{h}(t) = \lim_{dt\to 0} \frac{P(t\leq D<t+dt|D\geq t)}{dt}$ est le taux de décès marginalisé par rapport à l'historique du processus N(t). Notons que l'on ne fait aucune hypothèse d'indépendance entre N(t) et D.

Avec l'approche indirecte, l'écriture de la vraisemblance nécessite de modéliser la dépendance entre l'évènement terminal et le processus récurrent. Cette dépendance peut, par exemple, être prise en compte par un effet aléatoire partagé mais cela requiert une modélisation jointe du taux d'évènement terminal et de l'intensité du processus récurrent (Rondeau et al., 2007).

Conclusions et perspectives

Conclusions

L'objectif de cette thèse était de modéliser le risque de toxicité chez des patients traités par immunothérapie anti-cancéreuse et d'identifier des profils à risque. Du fait de la complexité des données de toxicité, la notion même de risque et d'indicateur permettant de le mesurer a été un *leitmotiv* au cours de ces travaux. Le sujet a été abordé de façon très générale au travers d'une revue de littérature des modèles de régression proposés pour modéliser le risque de toxicité. De nombreuses mesures de risque sont employées, offrant la possibilité de couvrir diverses caractéristiques des événements, telles que la temporalité, la sévérité ou la récurrence. La conclusion est qu'il semble illusoire d'évaluer et de comparer de façon fiable le risque au travers d'une seule mesure comme par exemple une probabilité globale pour un évènement au cours de l'étude. En particulier, la dimension temporelle est fondamentale afin d'améliorer la comparabilité des études. Dans un premier temps, nous nous sommes donc focalisés sur la modélisation des mesures de risque suivantes en fonction du temps : le taux, la probabilité d'incidence cumulée, et les taux relatifs afin de modéliser un seul événement dans un contexte de risques compétitifs. Dans un second temps, nous avons considéré l'intensité marginale et son ratio pour le cadre des évènements récurrents. En ce qui concerne l'intensité marginale, au-delà des contributions théoriques à sa modélisation, ces travaux ont proposé de revisiter cette mesure à travers des simulations s'écartant du cadre classique du processus de Poisson, soulignant ainsi son utilité dans des contextes aussi complexes que ceux des processus récurrents.

La relative simplicité d'application des modèles proposés dans ces travaux permettrait de les envisager en utilisation courante pour décrire l'occurrence des évènements indésirables au cours du temps et d'enrichir leur description, dans le cadre d'un essai clinique par exemple. La multiplicité des événements indésirables collectés nécessite, en pratique, l'utilisation d'outils descriptifs clés en main. Le cadre pénalisé avec estimation du paramètre de lissage permet de faciliter la construction des modèles et de réduire la variabilité, particulièrement lorsque le nombre d'évènements observés est modeste. La représentation de la dynamique de survenue d'évènements récurrents au cours du temps est encore moins souvent réalisée que pour un unique évènement. Le modèle d'intensité marginale proposé offre une représentation synthétique et nécessite peu d'hypothèses, ce qui en fait un bon candidat pour la description de ce types d'évènements.

Cependant, obtenir une tendance suffisamment précise requiert un nombre d'évènements qui n'est pas toujours disponible pour des données de toxicité, souvent décrites comme sous-dimensionnées. En particulier, nous avons vu que l'estimation des taux relatifs associés à un effet non proportionnel d'une covariable ne permet pas de définir une tendance fiable pour un faible nombre d'évènements. De plus, nous avons rencontré des dynamiques de taux complexes pour lesquelles le lissage était trop important. Enfin, si l'intérêt de la représentation de la toxicité au cours du temps fait aujourd'hui consensus, il est à noter que la définition du temps d'évènement n'est pas toujours évidente, particulièrement dans le cadre de données observationnelles. La

visualisation de la répartition des événements au fil du temps a révélé la présence de censure par intervalle, qui n'est pas prise en compte par le modèle. Ce type de censure n'affecte pas tous les événements de la même manière. Enfin, si l'on souhaite décrire la récurrence des évènements, la collecte et la structure des données sont également cruciaux et doivent être pensés en amont, par exemple pour pouvoir distinguer facilement un nouvel épisode de l'évolution (e.g. aggravation) d'un épisode antérieur.

Perspectives

Pour encourager l'utilisation des méthodes proposées dans cette thèse, le développement de modules pour le package `survPen` serait une ressource précieuse. Cela inclurait un premier module pour le calcul des probabilités d'incidence cumulées dans un cadre compétitif et un second pour l'estimateur robuste de la variance de l'intensité marginale.

Concernant le modèle du taux flexible, une piste intéressante pour améliorer le modèle dans le contexte des données de toxicité serait le lissage adaptatif, qui pourrait aider à éviter le lissage excessif observé dans certaines situations, même lorsque le nombre d'événements est conséquent. La question de la censure par intervalles mérite également d'être explorée davantage.

Le cadre de modélisation de l'intensité marginale proposé demanderait à être davantage exploré en pratique et notamment dans un cadre avec évènement terminal. Trouver l'équivalent de l'estimateur Bayésien empirique dans un cadre d'équations d'estimation reste également une question ouverte, même si l'intervalle de confiance obtenu par bootstrap reste très satisfaisant.

Annexes

A | Valorisation scientifique

Articles publiés

1. E. **Coz**, M. Fauvernier, D. Maucort-Boulch. (2023) An Overview of Regression Models for Adverse Events Analysis. *Drug Safety*, doi :10.1007/s40264-023-01380-7.

2. C.H. Vacheron, A. Friggeri, B. Allaouchiche, D. Maucort-Boulch, E. **Coz** (2021) Quiet scandal : variable selection in three major intensive care medicine journals, *Intensive Care Medicine*, doi : 10.1007/s00134-021-06535-7.

3. P.C. Vinke, M. Combalia, G.H. de Bock, C. Leyrat, A.M. Spanjaart, S. Dalle, M. Gomes da Silva, A. Fouda Essongue, A. Rabier, M. Pannard, M.S. Jalali, A. Elgammal, M. Papazoglou, M.S. Hacid, C. Rioufol, M.J. Kersten, M. GH van Oijen, E. Suazo-Zepeda, A. Malhotra, E. Coquery, A. Anota, M. Preau, M. Fauvernier, E. **Coz**, S. Puig, D. Maucort-Boulch (2023) Monitoring multidimensional aspects of quality of life after cancer immunotherapy : protocol for the international multicentre, observational QUALITOP cohort study, *BMJ Open*, doi :10.1136/bmjopen-2022-069090

Communications orales

1. E. **Coz***, M. Fauvernier, D. Maucort-Boulch, Utilisation des données de vie réelle pour l'identification des facteurs de risques associés aux effets indésirables de l'immunothérapie anti-cancéreuse », Forum de la recherche en cancérologie, 29-30th March 2022, Lyon (France)
2. E. **Coz***, M. Fauvernier, D. Maucort-Boulch, Les modèles flexibles du taux pour l'analyse des évènements indésirables, EPICLIN 10-12th May, 2023, Nancy, France
3. E. **Coz**, H. Charvat, D. Maucort-Boulch, M. Fauvernier*, Smooth marginal events rate models for recurrent event data, 45th Annual Conference of the International Society for Clinical Biostatistics (ISCB 2024) 21th – 25st July 2024, Thessaloniki, Greece

* orateur

Posters

1. E. **Coz**, M. Fauvernier, D. Maucort-Boulch, Flexible log-hazard models for adverse events analysis, 44th Annual Conference of the International Society for Clinical Biostatistics (ISCB 2024) 27th – 31st August 2023, Milano, Italy

2. E. **Coz**, M. Fauvernier, D. Maucort-Boulch, Quantifying the risk of treatment related adverse events in oncology : data, issues and models, 44th Annual Conference of the International Society for Clinical Biostatistics (ISCB 2024) 27th – 31st August 2023, Milano, Italy

B | Utilisation du MAP pour estimer les paramètres du modèle de Poisson à effets mixtes avec un petit nombre d'observations par groupe

On se propose d'illustrer l'impact des différentes approximations permettant d'estimer l'intégrale dans la vraisemblance marginale dans un modèle mixte non-linéaire. On considère le modèle théorique suivant :

$$\begin{aligned} y_{ij} &\sim \mathcal{P}(\mu_i) \\ \mu_i &= \beta_0 + 0.7x - 0.8x^2 + 0.8\text{ttt} + b_i \\ b_i &\sim \mathcal{N}(0,1) \end{aligned}$$

où, β_0 est l'intercept du modèle que l'on fera varier (respectivement 0.1 et 1), ttt est une covariable binaire répartie de façon équilibrée, b_i est un effet aléatoire suivant une loi normale centrée réduite et x est une covariable continue comprise entre 0 et 1 suivant une loi uniforme.

On simule un échantillon de $n = 300\ 000$ individus en considérant deux tailles de groupes pour l'effet aléatoire (respectivement 2 et 20). On ajuste le modèle sur l'échantillon avec le package `lme4`, en utilisant plusieurs méthodes pour estimer la vraisemblance marginale : (i) Quadrature de Gauss-Hermite adaptative à 10 noeuds (`nAGQ = 10`), (ii) Approximation de Laplace (`nAGQ = 1`), (iii) Méthode pénalisée (`nAGQ = 0`).

La Figure B.1 présente les écarts du prédicteur linéaire entre les différents modèles avec pour référence le modèle ajusté avec une quadrature de Gauss-Hermite à 10 noeuds. Les estimations obtenues par l'approximation de Laplace et par la quadrature de Gauss-Hermite sont très proches dans tous les scénarios. En revanche, on observe un écart entre la méthode pénalisée et la quadrature de Gauss-Hermite. Cet écart est accru lorsque la taille des groupes est faible. L'écart concerne principalement l'intercept. Un écart est observé sur les HR (graphique A par exemple) mais qui est très modeste.

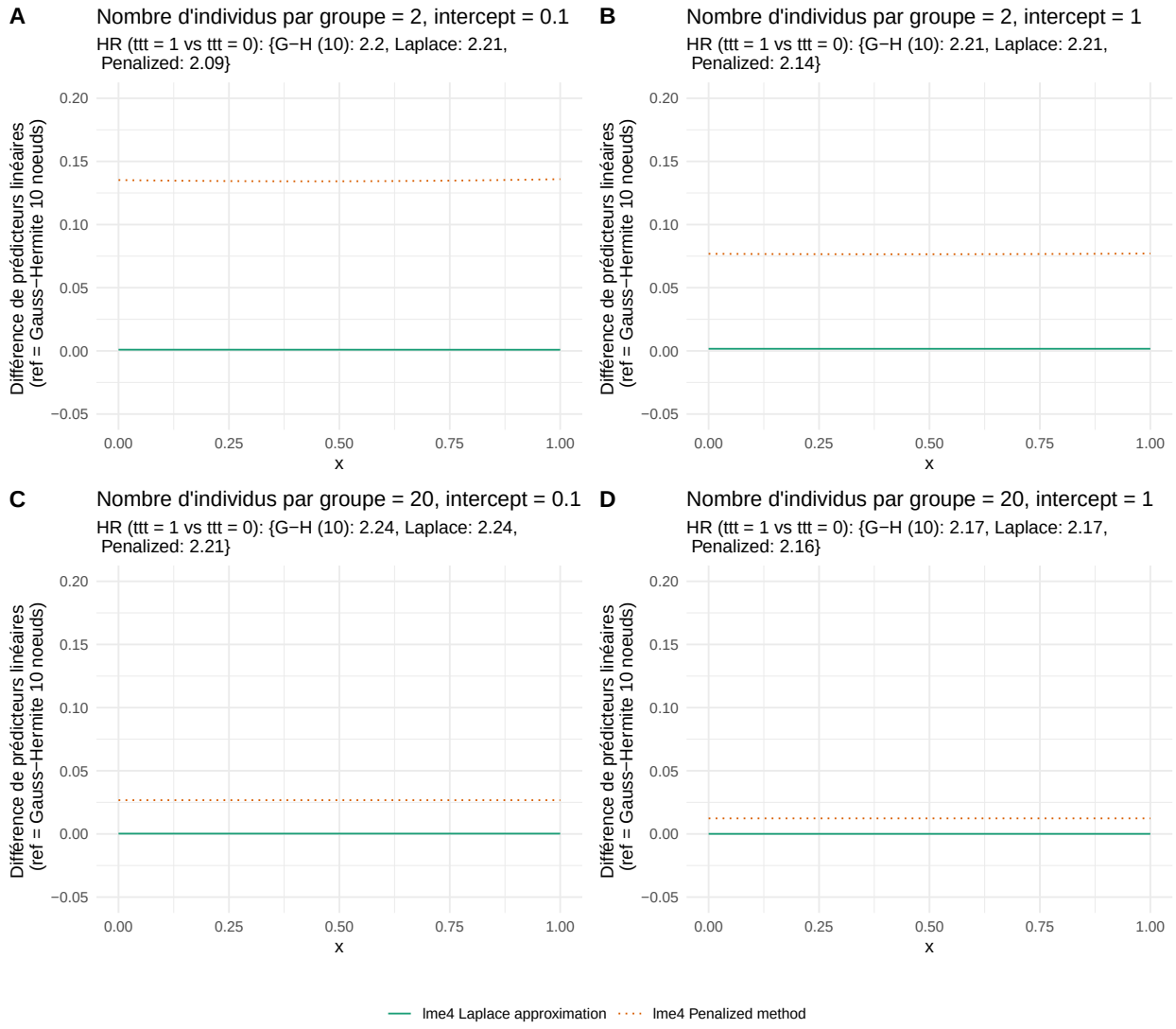

FIGURE B.1 – Ecart de prédicteur linéaire par rapport au modèle ajusté avec une quadrature de Gauss-Hermite. *(A) 2 individus par groupe et intercept de 0.1, (B) 2 individus par groupe et intercept de 1, (C) 20 individus par groupe et intercept de 0.1, (D) 20 individus par groupe et intercept de 1*

Références

[1] Abdulgader, S. M., Robberts, L., Ramjith, J., Nduru, P. M., Dube, F., Gardner-Lubbe, S., Zar, H. J., and Nicol, M. P. (2019). Longitudinal Population Dynamics of Staphylococcus aureus in the Nasopharynx During the First Year of Life. *Frontiers in Genetics*, 10. Publisher : Frontiers.

[2] Abu-Sbeih, H., Ali, F. S., and Wang, Y. (2020). Immune-checkpoint inhibitors induced diarrhea and colitis : a review of incidence, pathogenesis and management. *Current Opinion in Gastroenterology*, 36(1) :25.

[3] Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes : A Large Sample Study. *Annals of Statistics*, 10(4) :1100–1120.

[4] Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31(11-12) :1074–1088.

[5] Antoniadis, A., Gijbels, I., and Nikolova, M. (2011). Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Ann Inst Stat Math*, 63 :585–615.

[6] Armstrong, T. B., Kolesár, M., and Plagborg-Møller, M. (2022). Robust Empirical Bayes Confidence Intervals. *Econometrica*, 90(6) :2567–2602.

[7] Atkinson, T. M., Ryan, S. J., Bennett, A. V., Stover, A. M., Saracino, R. M., Rogak, L. J., Jewell, S. T., Matsoukas, K., Li, Y., and Basch, E. (2016). The association between clinician-based common terminology criteria for adverse events (CTCAE) and patient-reported outcomes (PRO) : a systematic review. *Supportive Care in Cancer*, 24(8) :3669–3676.

[8] Balan, T. A. and Putter, H. (2020). A tutorial on frailty models. *Statistical Methods in Medical Research*, 29(11) :3424–3454. Publisher : SAGE Publications Ltd STM.

[9] Barroso-Sousa, R., Barry, W. T., Garrido-Castro, A. C., Hodi, F. S., Min, L., Krop, I. E., and Tolaney, S. M. (2018). Incidence of Endocrine Dysfunction Following the Use of Different Immune Checkpoint Inhibitor Regimens. *JAMA Oncology*, 4(2) :173–182.

[10] Basch, E., Reeve, B. B., Mitchell, S. A., Clauser, S. B., Minasian, L. M., Dueck, A. C., Mendoza, T. R., Hay, J., Atkinson, T. M., Abernethy, A. P., Bruner, D. W., Cleeland, C. S., Sloan, J. A., Chilukuri, R., Baumgartner, P., Denicoff, A., St Germain, D., O'Mara, A. M., Chen, A., Kelaghan, J., Bennett, A. V., Sit, L., Rogak, L., Barz, A., Paul, D. B., and Schrag, D. (2014). Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *Journal of the National Cancer Institute*, 106(9) :dju244.

[11] Belot, A., Abrahamowicz, M., Remontet, L., and Giorgi, R. (2010). Flexible modeling of competing risks in survival analysis. *Statistics in Medicine*, 29(23) :2453–2468.

[12] Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11) :1713–1723.

[13] Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6) :956–971.

[14] Bilen, M. A., Martini, D. J., Liu, Y., Lewis, C., Collins, H. H., Shabto, J. M., Akce, M., Kissick, H. T., Carthon, B. C., Shaib, W. L., Alese, O. B., Pillai, R. N., Steuer, C. E., Wu, C. S., Lawson, D. H., Kudchadkar, R. R., El-Rayes, B. F., Master, V. A., Ramalingam, S. S., Owonikoko, T. K., and Harvey, R. D. (2019). The prognostic and predictive impact of inflammatory biomarkers in patients who have advanced-stage cancer treated with immunotherapy. *Cancer*, 125(1) :127–134.

[15] Bottomley, A., Coens, C., Mierzynska, J., Blank, C. U., Mandalà, M., Long, G. V., Atkinson, V. G., Dalle, S., Haydon, A. M., Meshcheryakov, A., Khattak, A., Carlino, M. S., Sandhu, S., Puig, S., Ascierto, P. A., Larkin, J., Lorigan, P. C., Rutkowski, P., Schadendorf, D., Koornstra, R., Hernandez-Aya, L., Di Giacomo, A. M., van den Eertwegh, A. J. M., Grob, J.-J., Gutzmer, R., Jamal, R., van Akkooi, A. C. J., Krepler, C., Ibrahim, N., Marreaud, S., Kicinski, M., Suciu, S., Robert, C., Eggermont, A. M. M., and EORTC Melanoma Group (2021). Adjuvant pembrolizumab versus placebo in resected stage III melanoma (EORTC 1325-MG/KEYNOTE-054) : health-related quality-of-life results from a double-blind, randomised, controlled, phase 3 trial. *The Lancet. Oncology*, 22(5) :655–664.

[16] Box-Steffensmeier, J. M. and De Boef, S. (2006). Repeated events survival models : the conditional frailty model. *Statistics in Medicine*, 25(20) :3518–3533.

[17] Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference : Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2) :261–304. Publisher : SAGE Publications Inc.

[18] Cai, K., Shen, H., and Lu, X. (2020). Group variable selection in the Andersen–Gill model for recurrent event data. *Journal of Statistical Planning and Inference*, 207 :99–112.

[19] Carlino, M. S., Larkin, J., and Long, G. V. (2021). Immune checkpoint inhibitors in melanoma. *Lancet (London, England)*, 398(10304) :1002–1014.

[20] Charvat, H. and Belot, A. (2021). mexhaz : An R Package for Fitting Flexible Hazard-Based Regression Models for Overall and Excess Mortality with a Random Effect. *Journal of Statistical Software*, 98 :1–36.

[21] Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., Belot, A., and Group, t. C. W. S. (2016). A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine*, 35(18) :3066–3084.

[22] Chiou, S. H., Xu, G., Yan, J., and Huang, C.-Y. (2023). Regression Modeling for Recurrent Events Possibly with an Informative Terminal Event Using R Package reReg. *Journal of Statistical Software*, 105 :1–34.

[23] Cook, R. and Lawless, J. (2007). The Statistical Analysis of Recurrent Events. *Statistics for Biology and Health*.

[24] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4) :377–403.

[25] Crowe, B. J., Xia, H. A., Berlin, J. A., Watson, D. J., Shi, H., Lin, S. L., Kuebler, J., Schriver, R. C., Santanello, N. C., Rochester, G., Porter, J. B., Oster, M., Mehrotra, D. V., Li, Z., King, E. C., Harpur, E. S., and Hall, D. B. (2009). Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development : a report of the safety planning, evaluation, and reporting team. *Clinical Trials*, 6(5) :430–440. Publisher : SAGE Publications.

[26] Danieli, C. and Abrahamowicz, M. (2019). Competing risks modeling of cumulative effects of time-varying drug exposures. *Statistical Methods in Medical Research*, 28(1) :248–262. Publisher : SAGE Publications Ltd STM.

[27] Deen, M. and de Rooij, M. (2020). ClusterBootstrap : An R package for the analysis of hierarchical data using generalized linear models with the cluster bootstrap. *Behavior Research Methods*, 52(2) :572–590.

[28] Eun, Y., Kim, I. Y., Sun, J.-M., Lee, J., Cha, H.-S., Koh, E.-M., Kim, H., and Lee, J. (2019). Risk factors for immune-related adverse events associated with anti-PD-1 pembrolizumab. *Scientific Reports*, 9(1) :14039.

[29] Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., Remontet, L., and and the Challenges in the Estimation of Net Survival Working Survival Group (2019). Multi-dimensional penalized hazard model with continuous covariates : applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 68(5) :1233–1257.

[30] FDA (2019). ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials.

[31] Fine, J. P. and Gray, R. J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446) :496–509.

[32] Foulley, J.-L., Delmas, C., and Robert-Granié, C. (2002). Méthodes du maximum de vraisemblance en modèle linéaire mixte. *Journal de la société française de statistique*, 143(1-2), 5-52.

[33] Fukihara, J., Sakamoto, K., Koyama, J., Ito, T., Iwano, S., Morise, M., Ogawa, M., Kondoh, Y., Kimura, T., Hashimoto, N., and Hasegawa, Y. (2019). Prognostic Impact and Risk Factors of Immune-Related Pneumonitis in Patients With Non-Small-Cell Lung Cancer Who Received Programmed Death 1 Inhibitors. *Clinical Lung Cancer*, 20(6) :442–450.e4.

[34] Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data : Consistency and computation. *Biometrika*, 81(3) :618–623.

[35] Gray, R. J. (1992). Flexible Methods for Analyzing Survival Data Using Splines, with Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association*, 87(420) :942–951.

[36] Guthrie, G. J. K., Charles, K. A., Roxburgh, C. S. D., Horgan, P. G., McMillan, D. C., and Clarke, S. J. (2013). The systemic inflammation-based neutrophil–lymphocyte ratio : Experience in patients with cancer. *Critical Reviews in Oncology/Hematology*, 88(1) :218–230.

[37] Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, 1973(Symposium) :10–41.

[38] Jahn-Eimermacher, A., Ingel, K., Ozga, A.-K., Preussler, S., and Binder, H. (2015). Simulating recurrent event data with hazard functions defined on a total time scale. *BMC Medical Research Methodology*, 15(1) :16.

[39] Jin, S. and Andersson, B. (2020). A note on the accuracy of adaptive Gauss–Hermite quadrature. *Biometrika*, 107(3) :737–744.

[40] Joly, P., Commenges, D., and Letenneur, L. (1998). A Penalized Likelihood Approach for Arbitrarily Censored and Truncated Data : Application to Age-Specific Incidence of Dementia. *Biometrics*, 54(1) :185–194. Publisher : [Wiley, International Biometric Society].

[41] Kipourou, D., Charvat, H., Rachet, B., and Belot, A. (2019). Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. *Statistics in Medicine*, 38(20) :3896–3910.

[42] Lalani, A.-K. A., Xie, W., Martini, D. J., Steinharter, J. A., Norton, C. K., Krajewski, K. M., Duquette, A., Bossé, D., Bellmunt, J., Van Allen, E. M., McGregor, B. A., Creighton, C. J., Harshman, L. C., and Choueiri, T. K. (2018). Change in Neutrophil-to-lymphocyte ratio (NLR) in response to immune checkpoint blockade for metastatic renal cell carcinoma. *Journal for Immunotherapy of Cancer*, 6(1) :5.

[43] Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, 66(6) :648–653.

[44] Lee, P. Y., Oen, K. Q. X., Lim, G. R. S., Hartono, J. L., Muthiah, M., Huang, D. Q., Teo, F. S. W., Li, A. Y., Mak, A., Chandran, N. S., Tan, C. L., Yang, P., Tai, E. S., Ng, K. W. P., Vijayan, J., Chan, Y. C., Tan, L. L., Lee, M. B.-H., Chua, H. R., Hong, W. Z., Yap, E. S., Lim, D. K., Yuen, Y. S., Chan, Y. H., Aminkeng, F., Wong, A. S. C., Huang, Y., and Tay, S. H. (2021). Neutrophil-to-Lymphocyte Ratio Predicts Development of Immune-Related Adverse Events and Outcomes from Immune Checkpoint Blockade : A Case-Control Study. *Cancers*, 13(6) :1308.

[45] Leung, K.-M., Elashoff, R. M., and Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*, 18(Volume 18, 1997) :83–104. Publisher : Annual Reviews.

[46] Lin, D. Y. and Wei, L. J. (1989). The Robust Inference for the Cox Proportional Hazards Model. *Journal of the American Statistical Association*, 84(408) :1074–1078.

[47] Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). 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.

[48] Lineberry, N., Berlin, J. A., Mansi, B., Glasser, S., Berkwits, M., Klem, C., Bhattacharya, A., Citrome, L., Enck, R., Fletcher, J., Haller, D., Chen, T.-T., and Laine, C. (2016). Recommendations to improve adverse event reporting in clinical trial publications : a joint pharmaceutical industry/journal editor perspective. *BMJ*, 355 :i5078. Publisher : British Medical Journal Publishing Group Section : Research Methods &amp ; Reporting.

[49] Liu, Q. and Pierce, D. A. (1994). A Note on Gauss-Hermite Quadrature. *Biometrika*, 81(3) :624–629. Publisher : [Oxford University Press, Biometrika Trust].

[50] Long, G. V., Atkinson, V., Ascierto, P. A., Robert, C., Hassel, J. C., Rutkowski, P., Savage, K. J., Taylor, F., Coon, C., Gilloteau, I., Dastani, H. B., Waxman, I. M., and Abernethy, A. P. (2016). Effect of nivolumab on health-related quality of life in patients with treatment-naïve advanced melanoma : results from the phase III CheckMate 066 study. *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, 27(10) :1940–1946.

[51] Ma, C. and Crimin, K. (2024). Joint Analysis of Longitudinal Data and Zero-Inflated Recurrent Events. *Statistics in Biopharmaceutical Research*, 16(1) :40–46.

[52] Malkhasyan, K. A., Zakharia, Y., and Milhem, M. (2017). Quality-of-life outcomes in patients with advanced melanoma : A review of the literature. *Pigment Cell & Melanoma Research*, 30(6) :511–520.

[53] Metcalfe, C. and Thompson, S. G. (2006). The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Statistics in Medicine*, 25(1) :165–179.

[54] Michael Friedman (1982). Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics*, 10(1) :101–113.

[55] Moriña, D. and Navarro, A. (2014). The R Package survsim for the Simulation of Simple and Complex Survival Data. *Journal of Statistical Software*, 59 :1–20.

[56] Nakanishi, Y., Masuda, T., Yamaguchi, K., Sakamoto, S., Horimasu, Y., Nakashima, T., Miyamoto, S., Tsutani, Y., Iwamoto, H., Fujitaka, K., Miyata, Y., Hamada, H., Okada, M., and Hattori, N. (2019). Pre-existing interstitial lung abnormalities are risk factors for immune checkpoint inhibitor-induced interstitial lung disease in non-small cell lung cancer. *Respiratory Investigation*, 57(5) :451–459.

[57] Nychka, D. (1988). Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Association*, 83(404) :1134–1143. Publisher : [American Statistical Association, Taylor & Francis, Ltd.].

[58] Owen, D. H., Wei, L., Bertino, E. M., Edd, T., Villalona-Calero, M. A., He, K., Shields, P. G., Carbone, D. P., and Otterson, G. A. (2018). Incidence, Risk Factors, and Effect on Survival of Immune-related Adverse Events in Patients With None–Small-cell Lung Cancer. *Clinical lung cancer*, 19(6) :e893–e900.

[59] Ozenne, B., Sørensen, Lyngholm, A., Scheike, T., Torp-Pedersen, C., and Gerds, Alexander, T. (2017). riskRegression : Predicting the Risk of an Event using Cox Regression Models. *The R Journal*, 9(2) :440.

[60] Pepe, M. S. and Cai, J. (1993). Some Graphical Displays and Marginal Regression Analyses for Recurrent Failure Times and Time Dependent Covariates. *Journal of the American Statistical Association*, 88(423) :811–820.

[61] Perperoglou, A., Sauerbrei, W., and Abrahamowicz, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, page 16.

[62] Pintilie, M. (2006). *Competing Risks : A Practical Perspective*. Wiley–Blackwell, Chichester, England ; Hoboken, NJ, 1er édition edition.

[63] Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2) :373–379.

[64] Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics : competing risks and multi-state models. *Statistics in Medicine*, 26(11) :2389–2430.

[65] Ramjith, J., Bender, A., Roes, K. C. B., and Jonker, M. A. (2024). Recurrent events analysis with piece-wise exponential additive mixed models. *Statistical Modelling*, 24(3) :266-287. Publisher : SAGE Publications India.

[66] Reiss, P. T. and Todd Ogden, R. (2009). Smoothing Parameter Selection for a Class of Semiparametric Linear Models. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 71(2) :505–523.

[67] Remontet, L., Uhry, Z., Bossard, N., Iwaz, J., Belot, A., Danieli, C., Charvat, H., and Roche, L. (2019). Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables : Performance of this multidimensional penalized spline approach in net survival trend analysis. *Statistical Methods in Medical Research*, 28(8) :2368–2384. Publisher : SAGE Publications Ltd STM.

[68] Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4) :1016–1022.

[69] Rondeau, V. (2010). Statistical models for recurrent events and death : Application to cancer events. *Mathematical and Computer Modelling*, 52(7) :949–955.

[70] Rondeau, V., Commenges, D., and Joly, P. (2003). Maximum Penalized Likelihood Estimation in a Gamma-Frailty Model. *Lifetime Data Analysis*, 9(2) :139–153.

[71] 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.

[72] Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation : application on cancer events. *Biostatistics*, 8(4) :708–721.

[73] Sauerbrei, W., Royston, P., and Look, M. (2007). A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal. Biometrische Zeitschrift*, 49(3) :453–473.

[74] Schadendorf, D., Dummer, R., Hauschild, A., Robert, C., Hamid, O., Daud, A., van den Eertwegh, A., Cranmer, L., O'Day, S., Puzanov, I., Schachter, J., Blank, C., Salama, A., Loquai, C., Mehnert, J. M., Hille, D., Ebbinghaus, S., Kang, S. P., Zhou, W., and Ribas, A. (2016). Health-related quality of life in the randomised KEYNOTE-002 study of pembrolizumab versus chemotherapy in patients with ipilimumab-refractory melanoma. *European Journal of Cancer (Oxford, England : 1990)*, 67 :46–54.

[75] Suzman, D. L., Pelosof, L., Rosenberg, A., and Avigan, M. I. (2018). Hepatotoxicity of immune checkpoint inhibitors : An evolving picture of risk associated with a vital class of immunotherapy agents. *Liver International*, 38(6) :976–987.

[76] Tang, S., Qin, C., Hu, H., Liu, T., He, Y., Guo, H., Yan, H., Zhang, J., Tang, S., and Zhou, H. (2022). Immune Checkpoint Inhibitors in Non-Small Cell Lung Cancer : Progress, Challenges, and Prospects. *Cells*, 11(3) :320.

[77] Tong, X., Zhu, L., and Sun, J. (2009). Variable selection for recurrent event data via nonconcave penalized estimating function. *Lifetime Data Analysis*, 15(2) :197–215.

[78] Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America*, 72(1) :20–22.

[79] Uhry, Z., Chatignoux, E., Dantony, E., Colonna, M., Roche, L., Fauvernier, M., Defossez, G., Leguyader-Peyrou, S., Monnereau, A., Grosclaude, P., Bossard, N., and Remontet, L. (2020). Multidimensional penalized splines for incidence and mortality-trend analyses and validation of national cancer-incidence estimates. *International Journal of Epidemiology*, 49(4) :1294–1306.

[80] Unkel, S., Amiri, M., Benda, N., Beyersmann, J., Knoerzer, D., Kupas, K., Langer, F., Leverkus, F., Loos, A., Ose, C., Proctor, T., Schmoor, C., Schwenke, C., Skipka, G., Unnebrink, K., Voss, F., and Friede, T. (2019). On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. *Pharmaceutical Statistics*, 18(2) :166–183.

[81] van den Berg, G. J. and Drepper, B. (2016). Inference for Shared-Frailty Survival Models with Left-Truncated Data. *Econometric Reviews*, 35(6) :1075–1098.

[82] Vinke, P. C., Combalia, M., de Bock, G. H., Leyrat, C., Spanjaart, A. M., Dalle, S., Gomes da Silva, M., Fouda Essongue, A., Rabier, A., Pannard, M., Jalali, M. S., Elgammal, A., Papazoglou, M., Hacid, M.-S., Rioufol, C., Kersten, M.-J., van Oijen, M. G., Suazo-Zepeda, E., Malhotra, A., Coquery, E., Anota, A., Preau, M., Fauvernier, M., Coz, E., Puig, S., and Maucort-Boulch, D. (2023). Monitoring multidimensional aspects of quality of life after cancer immunotherapy : protocol for the international multicentre, observational QUALITOP cohort study. *BMJ open*, 13(4) :e069090.

[83] Wahba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. *The Annals of Statistics*, 13(4) :1378–1402.

[84] Whittaker, C. F., Tom, S. E., Bivens, A., and Klein-Schwartz, W. (2017). Evaluation of an Educational Intervention on Knowledge and Awareness of Medication Safety in Older Adults with Low Health Literacy. *American Journal of Health Education*, 48(2) :100–107.

[85] Wood, S. N. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association*, 99(467) :673–686.

[86] Wood, S. N. (2017). *Generalized additive models : an introduction with R*. Chapman & Hall/CRC texts in statistical science. CRC Press/Taylor & Francis Group, Boca Raton, second edition edition.

[87] Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, 111(516) :1548–1563.

[88] Wu, Y. L., Fulgenzi, C. A. M., D'Alessio, A., Cheon, J., Nishida, N., Saeed, A., Wietharn, B., Cammarota, A., Pressiani, T., Personeni, N., Pinter, M., Scheiner, B., Balcar, L., Huang, Y.-H., Phen, S., Naqash, A. R., Vivaldi, C., Salani, F., Masi, G., Bettinger, D., Vogel, A., Schönlein, M., von Felden, J., Schulze, K., Wege, H., Galle, P. R., Kudo, M., Rimassa, L., Singal, A. G., Sharma, R., Cortellini, A., Gaillard, V. E., Chon, H. J., Pinato, D. J., and Ang, C. (2022). Neutrophil-to-Lymphocyte and Platelet-to-Lymphocyte Ratios as Prognostic Biomarkers in Unresectable Hepatocellular Carcinoma Treated with Atezolizumab plus Bevacizumab. *Cancers*, 14(23) :5834. Number : 23 Publisher : Multidisciplinary Digital Publishing Institute.

[89] Xie, T., Zhang, Z., Qi, C., Lu, M., Zhang, X., Li, J., Shen, L., and Peng, Z. (2021). The Inconsistent and Inadequate Reporting Of Immune-Related Adverse Events in PD-1/PD-L1 Inhibitors : A Systematic Review of Randomized Controlled Clinical Trials. *The Oncologist*, page onco.13940.

[90] Xu, Y. and Cheung, Y. B. (2018). Frailty Models and Frailty-mixture Models for Recurrent Event Times : Update. *The Stata Journal*, 18(2) :477–484. Publisher : SAGE Publications.

[91] Yin, Q., Wu, L., Han, L., Zheng, X., Tong, R., Li, L., Bai, L., and Bian, Y. (2023). Immune-related adverse events of immune checkpoint inhibitors : a review. *Frontiers in Immunology*, 14 :1167975.

[92] Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., and Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39(8) :1199–1236.

[93] Zar, H. J., Barnett, W., Myer, L., Stein, D. J., and Nicol, M. P. (2015). Investigating the early-life determinants of illness in Africa : the Drakenstein Child Health Study. *Thorax*, 70(6) :592–594.