Université de Montréal

Une procédure de sélection automatique de la discrétisation optimale de la ligne du temps pour des méthodes longitudinales d'inférence causale

par

Steve Ferreira Guerra

Faculté de Pharmacie

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Sciences pharmaceutiques, option Médicament et santé des populations

12 juillet 2017

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Une procédure de sélection automatique de la discrétisation optimale de la ligne du temps pour des méthodes longitudinales d'inférence causale

présenté par

Steve Ferreira Guerra

a été évalué par un jury composé des personnes suivantes :

Marie-Pierre Sylvestre, PhD

(président-rapporteur)

Mireille Schnitzer, PhD

(directrice de recherche)

Lucie Blais, PhD

(codirectrice)

Geneviève Lefebvre, PhD

(membre du jury)

Mémoire accepté le

31 août 2017

RÉSUMÉ

Lors d'études observationnelles longitudinales, les caractéristiques des sujets sont mesurées et suivies dans le temps. Bien que la distribution sous-jacente de ces variables puisse être continue dans le temps, en pratique, les mesures sont observées à des temps discrets. De fait, dans les bases de données réelles, telles les bases de données administratives, il est raisonnable de supposer que les processus d'exposition sous-jacents se produisent en temps réel continu. Cependant, les méthodes d'inférence causale standard pour les données longitudinales supposent généralement que les expositions sous-jacentes sont des processus discrets et que les temps où les expositions sont observées en pratique sont les seuls moments où celles-ci peuvent changer de valeur. Une problématique de cette approche traditionnelle réside dans la manière de discrétiser cette ligne du temps continue sous-jacente.

L'approche traditionnelle consiste à discrétiser arbitrairement la ligne du temps en plusieurs points temporels qui séparent les données en intervalles. Ainsi, d'une part, la difficulté réside dans la sélection d'une discrétisation appropriée qui permet de suffisamment saisir toutes les relations entre les variables dans l'optique de contrôler de manière appropriée pour le biais de confusion dépendant du temps. D'autre part, une autre difficulté intrinsèque à ce problème est que, à mesure que le nombre de points temporels augmente, la dimensionnalité augmente et le volume de l'espace longitudinal augmente si rapidement que les données peuvent

devenir rares. En particulier, les méthodes d'inférence causale standard nécessitent un support raisonnable pour l'exposition d'intérêt. Sans un tel support, les estimateurs seront mal définis ou extrêmement variables. Bien que l'on puisse répondre à ce défi en créant une discrétisation plus grossière, une telle approche introduit un biais en ne contrôlant pas pour tous les facteurs de confusion variant dans le temps. Ceci ouvre le débat du choix d'un degré optimal de discrétisation. Par conséquent, un nombre arbitraire de points temporels peut entraîner un biais non catégorisable ou augmenter la variance.

Nous proposons une méthode novatrice qui sélectionne de manière adaptative une discrétisation optimale. Cela se fera par la validation croisée d'une fonction de perte basée sur l'estimation par maximum vraisemblance ciblée longitudinale groupée. Nous effectuons une étude de simulation dans laquelle nous générons des données avec confusion dépendant du temps afin d'évaluer le compromis biais-variance et la performance de la procédure de sélection. Nous appliquons également notre procédure de sélection à une application de données réelles portant sur l'effet de médicaments antiasthmatiques pendant la grossesse sur la durée de gestation.

Mots-clés : bases de données administratives ; TMLE ; estimation semi-paramétrique ; apprentissage machine

SUMMARY

In longitudinal observational studies, subject characteristics are measured and followed over time. Although the underlying evolution of such response variables may be continuous in time, in practice the measurements are observed at discrete time points. In real world data sets such as administrative databases, it is reasonable to assume that the underlying observed exposure processes happen in real time. However, standard causal inference methods for longitudinal data usually assume that the underlying exposures are discrete time processes, and that the observational time points are the only time points when the exposures may change values. A crucial pitfall of this traditional approach lies in the manner to discretize this underlying continuous timeline.

The common approach is to arbitrarily discretize the timeline into several time-points that separate the data into intervals. The problem at hand is then selecting the appropriate discretization that sufficiently captures all of the intricate relationships between the variables in the optic of appropriately controlling the time-dependent confounding bias. On the other hand, the common theme of this problem is that as the number of time-points increases, the dimensionality increases, and the volume of the space increases so rapidly that the available data may become sparse. In particular, standard causal inference methods rely on reasonable data support for the exposure of interest. Without such support, estimators will be ill-defined or extremely variable. Although one might respond to this challenge by creating a coarser discretization, such an approach introduces

bias by failing to capture all the time-dependent confounding and leaves open the question of how to choose an optimal degree of coarsening. Hence, an arbitrary number of time-points may result in unpredictable bias or inflated variance.

We propose a novel method that data-adaptively selects an optimal discretization. This will be done through the cross-validation of a loss function based on pooled Longitudinal Targeted Maximum Likelihood Estimation. We conduct a simulation study in which we generate time-dependent confounded data to evaluate the bias-variance trade-off and the performance of the selection procedure. We also apply our selection procedure to a real-world data application of the effect of asthma medication during pregnancy on pregnancy duration.

Keywords : administrative databases ; coarsening ; TMLE ; semi-parametric estimation ; machine learning

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

ATE : Effet causal moyen (de l'anglais : «*Average Treatment Effect* »)

BDA : Base de données administratives

EIC : Courbe d'influence efficace (de l'anglais : «*Efficient Influence Curve* »)

GINA : «*Global Initiative for Asthma* »

ICS : Corticostéroïdes inhalés (de l'anglais : «*Inhaled corticosteroids* »)

IPTW : Pondération par l'inverse de probabilité de traitement (de l'anglais : «*Inverse Probability of Treatment Weighting* »)

LABA : Bêta-agonistes à longue durée d'action (de l'anglais : «*Long-acting beta-agonists* »)

LTMLE : Estimation par maximum de vraisemblance ciblée longitudinale (de l'anglais : «*Longitudinal Targeted Maximum Likelihood Estimation* »)

LTRA : Antagonistes des récepteurs aux leucotriènes (de l'anglais : «*Leukoriene Receptor Antagonist* »)

MED-ECHO : Maintenance et exploitation des données pour l'étude de la clientèle hospitalière

MSM : Modèle structurel marginal

NAEPP : « *National Asthma Education and Prevention Program* »

NCS : Corticosteroïdes nasaux (de l'anglais : «*Nasal corticosteroids* »)

OCS : Corticosteroïdes oraux (de l'anglais : «*Oral corticosteroids* »)

PLTMLE : Estimation par maximum de vraisemblance ciblée longitudinale groupée (de l'anglais : «*Pooled Longitudinal Targeted Maximum Likelihood Estimation* »)

RAMQ : Régie de l'assurance maladie du Québec

SABA : Bêta-agonistes à courte durée d'action (de l'anglais : «*Short-acting beta-agonists* »)

SUTVA : Hypothèse de valeur de traitement des unités stable (de l'anglais : «*Stable Unit Treatment Value Assumption* »)

TMLE : Estimation par maximum de vraisemblance ciblée (de l'anglais : « *Targeted Maximum Likelihood Estimation* »)

REMERCIEMENTS

Je souhaite tout d'abord remercier ma directrice de recherche, Dr.Mireille Schnitzer, sans qui ce mémoire n'aurait pas eu lieu. Je profite donc de ces courtes lignes pour lui démontrer toute ma reconnaissance envers son aide constante dans ce projet. Elle a su m'apprendre à toujours viser plus haut et à toujours croire en mon potentiel. Je lui en serai éternellement reconnaissant tout au long de ma carrière. Je souhaite la remercier spécialement pour sa compréhension et son support inconditionnel dans ce parcours parfois difficile. Merci.

Je souhaite aussi remercier ma codirectrice, Dr.Lucie Blais, pour son soutien, sa disponibilité et ses conseils opportuns, et surtout pour ce qu'elle est, un exemple de succèx de carrière, de rigueur et de vivacité d'esprit.

Je tiens à remercier tous ceux que j'ai côtoyés au cours de cette merveilleuse expérience, qui, par leur bonne humeur et leur aide, ont contribué à rendre mes journées plus vives. Une vie est tissée de milliers de liens et vous faites désormais partie de ma toile. Je souhaite adresser un remerciement particulier à Amélie, qui a su très bien m'épauler dans l'extraction des données au début de ma maîtrise, et à Geneviève, car elle connaît tout le monde et tous les recoins de cette Faculté, et qui m'a particulièrement aidé à relativiser mes problèmes et à toujours garder le sourire !

Je souhaite aussi exprimer mon infinie gratitude à ma famille et à tous mes amis qui m'ont de près ou de loin encouragé. À mes parents qui se sont énormément sacrifiés pour que je puisse accomplir mes buts. Je vous suis éternellement

reconnaissant. Je suis ce que vous êtes et vous êtes ce que je suis. Aussi, à ma soeur, qui fut toujours une source de motivation et de dépassement dans mes études.

Finalement, à ma belle Marie, qui m'a rejoint en cours de route, pour tout ce qu'elle est, fait et continue de faire. Tu es un exemple de persévérance, de résilience et une source de motivation constante. Je t'aime !

Finalement, je remercie CNODES, pour leur appui financier tout au long de mes études.

À Marie,

Não tenhas medo, tout va bien aller...

AVANT-PROPOS

Ce mémoire porte sur un sujet de recherche inédit introduit au moyen d'un manuscrit prêt à être soumis dans une revue à comité de révision par des pairs.

L'auteur principal a rédigé le corps de l'article et a réalisé l'étude de simulation, de même que l'extraction et l'analyse de données présentées dans le manuscrit. En accord avec Dr.Schnitzer, il a procédé aux développements des nouveaux concepts et des nouvelles méthodes décrites dans ledit manuscrit. Dr.Schnitzer a conseillé la rédaction théorique contenue dans l'article. Dr.Schnitzer et Dr.Blais ont participé à la mise en œuvre et à la supervision de l'analyse de données. Amélie Forget a supervisé l'extraction de données. Tous les coauteurs ont procédé à une révision du manuscrit final présenté dans le chapitre 3.

INTRODUCTION

Il va sans nul doute que l'évaluation de l'efficacité et de l'innocuité de médicaments, de traitements ou d'interventions est un sujet de la plus haute importance dans plusieurs domaines reliés à la santé. Pour ce type de questions, les études randomisées contrôlées sont généralement considérées comme le *gold standard*, car elles permettent, sous certaines conditions idéales, d'identifier l'effet réel dû au traitement [**1**]. En effet, puisque les traitements sont assignés au hasard, et lorsque le traitement est donné à l'aveugle et sous une adhésion parfaite au traitement, le traitement reçu est indépendant des caractéristiques des patients à l'étude et les résultats ne sont pas affectés par le biais de confusion [**2**]. Toutefois, pour des raisons éthiques, économiques ou de faisabilité, de telles études peuvent souvent être irréalisables. Notons par exemple la sous-représentation de femmes enceintes dans les études randomisées contrôlées où la sécurité néonatale et foetale représentent des questions éthiques complexes limitant l'inclusion de ce type de participantes pour tester des médicaments utilisés pour des maladies aussi courantes que l'asthme, la dépression et l'épilepsie [**3**].

Une solution de remplacement est ainsi d'entreprendre une étude observationnelle, tirant des inférences à partir d'un échantillon de la population où les sujets sont naturellement exposés au traitement à l'étude. Conséquemment, comme les caractéristiques des sujets peuvent influencer leur exposition au traitement, les

études observationnelles peuvent être affectées par le biais de confusion. Plus particulièrement, s'il est d'intérêt de conduire une étude longitudinale afin d'observer l'effet d'une exposition longitudinale variant dans le temps, ces études peuvent être sujettes au biais de confusion variant dans le temps [**4**]. Les méthodes traditionnelles de contrôle pour la confusion, telles que les méthodes de régression, ne s'appliquent pas à la confusion variant dans le temps, car elles introduisent un biais, que l'on ajuste ou non pour les facteurs de confusion variant dans le temps [**5**].

En conséquence, plusieurs méthodes en inférence causale ont été proposées pour adresser ce type de biais ; l'inférence causale étant un cadre statistique visant à identifier le lien de cause à effet entre l'exposition et l'issue à l'étude. Ce mémoire se concentre principalement sur l'utilisation d'une de ces méthodes statistiques pour l'estimation de paramètres causaux en présence de confusion variant dans le temps : l'estimation par maximum de vraisemblance ciblée longitudinale (LTMLE) [**6**]. Brièvement, la LMTLE est une méthode produisant des estimateurs semi-paramétriques doublement robustes et localement efficaces pour divers paramètres causaux. La LTMLE développe deux modèles d'ajustement pour la confusion : l'un modélisant l'issue longitudinale en fonction des facteurs de confusion et l'autre modélisant l'exposition longitudinale en fonction des facteurs de confusion. Par conséquent, la LTMLE hérite de la propriété de double robustesse, car il suffit de bien contrôler pour la confusion dans l'un ou l'autre des modèles. Notamment, la LTMLE peut être employée afin d'estimer des paramètres causaux provenant de modèles structurels marginaux (MSM) [**7**, **8**], une classe de modèles causaux dont les paramètres représentent l'effet causal entre l'exposition et l'issue. Ces modèles seront particulièrement d'intérêt dans ce mémoire.

Tout comme la LTMLE, la plupart des méthodes longitudinales d'inférence causale supposent que les données sont mesurées à des temps discrets. Cette approche est tout à fait naturelle si l'information des sujets est collectée à des intervalles réguliers, comme dans une étude planifiée où les patients sont réexaminés à chaque visite. Or, dans le cas de certaines études observationnelles longitudinales, il est plutôt raisonnable de supposer que les données sont générées à partir d'un processus en temps continu, mais ne sont observables qu'à des temps discrets. Ceci est le cas lorsque les données proviennent de bases de données administratives (BDA), où les données ne sont pas collectées à des intervalles réguliers pour tous les patients, mais plutôt en temps continu. Dans ce mémoire, nous aborderons l'utilisation de méthodes d'inférence causale lorsque la ligne du temps continue est discrétisée.

Les principales contributions de ce mémoire sont de définir la problématique de la discrétisation arbitraire pour les méthodes causales longitudinales et la mise sur pied d'un cadre pour l'estimation de paramètres causaux longitudinaux pour des données discrétisées. De plus, ce mémoire propose une méthode novatrice de sélection automatique de la discrétisation optimale de la ligne du temps intégrée à l'estimation par maximum de vraisemblance ciblée longitudinale groupée (PLTMLE) [**8**]. D'un point de vue appliqué, ce mémoire présente une première analyse longitudinale de données de grossesses asthmatiques à l'aide de méthodes d'inférence causale.

Dans le chapitre 1, nous présentons une revue de la littérature couvrant : l'exemple de l'effet de médicaments antiasthmatiques durant la grossesse qui a justifié le développement de la méthodologie proposée, les bases de données administratives et leur utilisation en santé, ainsi que le cadre de l'inférence causale et les méthodes utilisées. Au chapitre 2, nous introduisons les objectifs de ce

mémoire. Le chapitre 3 consiste en un article présentant le développement méthodologique ainsi que les résultats d'une étude de simulation et d'une application de la méthode développée à l'asthme durant la grossesse. Le chapitre 4 se veut un complément du manuscrit du chapitre 3. Finalement, le dernier chapitre sera voué à la discussion des résultats obtenus, ainsi qu'à la pertinence de cette méthode.

Chapitre 1

REVUE DE LA LITTÉRATURE SCIENTIFIQUE

La revue de la littérature est divisée en trois sections principales. La première section porte sur une application qui a servi de motivation au développement méthodologique subséquent et englobe le contexte de l'effet de médicaments antiasthmatiques durant la grossesse sur la durée de gestation. La seconde section révise les forces et faiblesses qui justifient l'utilisation de BDA, leur utilisation accrue dans le domaine de la santé, de même que leur structure, notamment au niveau des BDA en santé au Québec. Enfin, la dernière section abordera le cadre et les hypothèses de l'inférence causale, de même que les méthodes utilisées dans ce mémoire.

1.1. Exemple motivateur – Asthme durant la grossesse

Ce mémoire aborde des problèmes qui peuvent être retrouvés dans bon nombre d'applications longitudinales. L'application d'intérêt dans ce mémoire porte sur une étude longitudinale de l'effet de médicaments antiasthmatiques pendant la grossesse.

1.1.1. Définition et prévalence

L'asthme est une maladie chronique des voies respiratoires, caractérisée, entre autres, par des symptômes tels l'essoufflement, une respiration sifflante, la production d'expectorations et la toux [**9**, **10**]. En plus de symptômes respiratoires, l'asthme peut entraîner une limitation de l'activité et des exacerbations qui nécessitent parfois des soins de santé urgents et qui peuvent être fatales si elles ne sont pas traitées [**11**]. En 2014, au Canada, environ 8,1% de la population âgée de 12 ans et plus ont été diagnostiqués par un professionnel de la santé comme ayant de l'asthme [**12**]. Quant aux femmes canadiennes, 9,2% sont atteintes d'asthme [**12**]. Qui plus est, chez les femmes enceintes, une étude américaine montre que la prévalence de l'asthme, évalué au moyen d'enquêtes nationales, se situe entre 3,7% et 8,4% [**13**]. Selon une étude un peu plus récente par les mêmes auteurs, visant à mettre à jour les résultats précédemment obtenus, entre 8,4 % et 8,8% de femmes enceintes ont rapporté souffrir d'asthme [**14**], ce qui en fait une des maladies chroniques les plus fréquentes durant la grossesse. Dans un ordre d'idées similaire, une étude menée entre 2001 et 2007 à travers 11 régimes d'assurance maladie des États-Unis indique que 9,7% de femmes ont reçu des médicaments antiasthmatiques pendant la grossesse [**15**], ce qui vient corroborer que l'asthme est une des maladies chroniques les plus prévalentes durant la grossesse.

1.1.2. Contrôle de l'asthme et l'effet sur le foetus

Selon les lignes directrices nationales, le contrôle de l'asthme est défini au sens large par la prévention d'exacerbations et de symptômes reliés à l'asthme par un traitement adéquat [**16**]. Le contrôle de l'asthme est mesuré par la fréquence et la sévérité de symptômes asthmatiques tels les réveils nocturnes, la perturbation

d'activités quotidiennes, le recours à des médicaments de secours, les exacerbations et par des tests de fonction pulmonaire.

Un large éventail d'études démontrent que l'asthme non contrôlé a un effet néfaste tant sur la santé maternelle que celle du foetus [**17**]. Entre autres, relativement à la santé du foetus, il a été démontré que l'asthme mal contrôlé peut augmenter le risque de faible poids à la naissance [**18**], de petite taille pour l'âge gestationnel [**18**] et de malformations congénitales [**18**,**19**]. Plus particulièrement, le mauvais contrôle de l'asthme est aussi associé à une augmentation du risque de naissance prématurée [**20**]. Dans un même ordre d'idées, deux méta-analyses réalisées par *Murphy et al.* [**21**, **22**], regroupant des études comparant un groupe de femmes asthmatiques à un groupe de femmes non-asthmatiques, confirment qu'une association existe entre l'asthme et le faible poids à la naissance, le petit poids pour l'âge gestationnel, la prématurité et les malformations congénitales. Par ailleurs, des études concluent que le traitement approprié menant à un asthme bien contrôlé n'aurait pas d'effets nocifs sur le foetus [**23**, **24**]. En particulier, une revue systématique menée par *Lim et al.* en 2011 [**25**] conclut que l'utilisation de médicaments préventifs pour la prise en charge de l'asthme pendant la grossesse doit être basée sur une évaluation des avantages de l'utilisation de médicaments au détriment des risques d'un asthme mal contrôlé.

Conséquemment, les lignes directrices actuelles recommandent le maintien du traitement pour l'asthme durant la grossesse [**11**, **16**, **26**] étant donné qu'il a été suggéré que les avantages du traitement de l'asthme pendant la grossesse l'emportent généralement sur les risques négatifs potentiels des médicaments [**25**, **27**].

1.1.3. Traitement de l'asthme

Dans le traitement de l'asthme, nous retrouvons deux grandes catégories de médicaments : les médicaments de secours pris afin d'interrompre rapidement des exacerbations d'asthme ainsi que d'atténuer les symptômes d'asthme et des médicaments de contrôle, administrés quotidiennement dans le but d'atteindre et de maintenir le contrôle de l'asthme. Les médicaments de secours incluent les bêta-agonistes à courte durée d'action (SABA), les anticholinergiques, tous deux des bronchodilatateurs inhalés, et les corticostéroïdes oraux (OCS) ou nasaux (NCS). Les médicaments de contrôle à long terme incluent quant à eux les corticostéroïdes inhalés (ICS), les bêta-agonistes à longue durée d'action (LABA), les antagonistes des récepteurs aux leucotriènes (LTRA) et des OCS ou NCS prescrits en utilisation quotidienne. Parmi cette dernière classe de médicaments, les ICS feront l'objet d'étude dans ce mémoire.

1.1.4. Lignes directrices pour le traitement de l'asthme durant la grossesse

Plusieurs lignes directrices font état d'approches pour guider la prise de décision clinique nécessaire pour répondre aux besoins individuels des patients, mais peu sont celles qui émettent directement des lignes directrices pour les femmes enceintes. Notamment, les lignes directrices de la Global Initiative for Asthma (GINA) relatives à la grossesse sont similaires à celles pour la population générale et non adaptées aux femmes enceintes [**16**]. Par contre, des traitements propres aux femmes enceintes sont rapportés dans un rapport du National Asthma Education and Prevention Program (NAEPP) [**26**] (Figure 1.1). Toutefois, ces lignes directrices datent de 2004 et, faute de données probantes sur lesquelles baser leurs recommandations, sont potentiellement inadéquates pour les femmes enceintes.

Appendix B: Figures

Figure 1 Stepwise Approach for Managing Asthma During Pregnancy and Lactation: Treatment

Classify Severity: Clinical Features Before Treatment or Adequate Control	Symptoms/ Day / Symptoms/ Night	PEF or FEV_1 / PEF Variability	Medications Required To Maintain Long-Term Control: Daily Medications
Step 4 Severe Persistent	Continual / Frequent	≤60% / >30%	• Preferred treatment: - High-dose inhaled corticosteroid AND - Long-acting inhaled $beta_2$-agonist AND, if needed, - Corticosteroid tablets or syrup long term (2 mg/kg per day, generally not to exceed 60 mg per day). (Make repeat attempts to reduce systemic corticosteroid and maintain control with high-dose inhaled corticosteroid.*) • Alternative treatment: - High-dose inhaled corticosteroid* AND - Sustained release theophylline to serum concentration of 5–12 mcg/mL.
Step 3 Moderate Persistent	Daily / >1 night/week	>60%–<80% / >30%	• Preferred treatment: EITHER - Low-dose inhaled corticosteroid* and long-acting inhaled $beta_2$-agonist OR - Medium-dose inhaled corticosteroid.* If needed (particularly in patients with recurring severe exacerbations): - Medium-dose inhaled corticosteroid* and long-acting inhaled $beta_2$-agonist. • Alternative treatment: - Low-dose inhaled corticosteroid* and either theophylline or leukotriene receptor antagonist.† If needed: - Medium-dose inhaled corticosteroid* and either theophylline or leukotriene receptor antagonist.†
Step 2 Mild Persistent	>2 days/week but <daily / >2 nights/month	≥80% / 20%–30%	• Preferred treatment: - Low-dose inhaled corticosteroid.* • Alternative treatment (listed alphabetically): cromolyn, leukotriene receptor antagonist† OR sustained-release theophylline to serum concentration of 5–12 mcg/mL.
Step 1 Mild Intermittent	≤2 days/week / ≤2 nights/month	≥80% / <20%	• No daily medication needed. • Severe exacerbations may occur, separated by long periods of normal lung function and no symptoms. A course of systemic corticosteroid is recommended.
Quick Relief All Patients	• Short-acting bronchodilator: 2–4 puffs **short-acting inhaled $beta_2$-agonist**‡ as needed for symptoms. • Intensity of treatment will depend on severity of exacerbation; up to 3 treatments at 20-minute intervals or a single nebulizer treatment as needed. Course of systemic corticosteroid may be needed. • Use of short-acting inhaled $beta_2$-agonist‡ >2 times a week in intermittent asthma (daily, or increasing use in persistent asthma) may indicate the need to initiate (increase) long-term-control therapy.		

Step down
Review treatment every 1–6 months; a gradual stepwise reduction in treatment may be possible.

Step up
If control is not maintained, consider step up. First, review patient medication technique, adherence, and environmental control.

Goals of Therapy: Asthma Control

- Minimal or no chronic symptoms day or night
- Minimal or no exacerbations
- No limitations on activities; no school/work missed
- Maintain (near) normal pulmonary function
- Minimal use of short-acting inhaled $beta_2$-agonist‡
- Minimal or no adverse effects from medications

Notes

- The stepwise approach is meant to assist, not replace, the clinical decisionmaking required to meet individual patient needs.
- Classify severity: assign patient to most severe step in which any feature occurs (PEF is percent of personal best; FEV_1 is percent predicted).
- Gain control as quickly as possible (consider a short course of systemic corticosteroid), then step down to the least medication necessary to maintain control.
- Minimize use of short-acting inhaled $beta_2$-agonist‡ (e.g., use of approximately one canister a month even if not using it every day indicates inadequate control of asthma and the need to initiate or intensify long-term-control therapy).
- Provide education on self-management and controlling environmental factors that make asthma worse (e.g., allergens, irritants).
- Refer to an asthma specialist if there are difficulties controlling asthma or if Step 4 care is required. Referral may be considered if Step 3 care is required.

* There are more data on using budesonide during pregnancy than on using other inhaled corticosteroids.
† There are minimal data on using leukotriene receptor antagonists in humans during pregnancy, although there are reassuring animal data submitted to FDA.
‡ There are more data on using albuterol during pregnancy than on using other short-acting inhaled $beta_2$-agonists.

FIGURE 1.1. Lignes directrices pour le traitement de l'asthme durant la grossesse - *NAEPP expert panel report : managing asthma during pregnancy : recommendations for pharmacologic treatment—2004 update*

Il est à remarquer que bien que ces lignes directrices soient la référence pour le traitement de l'asthme pendant la grossesse, une étude récente [**28**] montre qu'environ 50% des femmes enceintes choisissent d'arrêter ou de diminuer leurs doses d'ICS, particulièrement si les doses d'ICS sont faibles. Curieusement, dans cette étude, l'arrêt d'ICS était associé à une diminution du risque d'exacerbations. Cependant, les auteurs soutiennent que ces résultats peuvent avoir été causés par un biais de confusion résiduel dû à la sévérité de l'asthme. Il est donc naturel de se questionner sur l'efficacité et l'innocuité de la prise d'ICS pendant la grossesse sur la santé foetale, plus précisément dans notre cas sur la durée de gestation.

1.1.5. Association entre les ICS et la durée de gestation

Aucune étude n'évalue précisément l'effet des ICS sur la durée de gestation. Une seule étude descriptive portant seulement sur deux molécules d'ICS a été réalisée, n'ayant pas trouvé d'association significative entre ces molécules et la durée de gestation [**29**]. On retrouve plutôt des études évaluant l'effet des ICS sur la prématurité, qui peut être considérée comme un analogue de la durée de gestation.

Plusieurs études comparant des femmes asthmatiques enceintes exposées aux ICS à des femmes asthmatiques enceintes non exposées ne sont pas arrivées à démontrer l'existence d'une association significative entre la prise d'ICS lors de la grossesse et la prématurité [**30**, **31**]. De plus, une étude longitudinale récente plus rigoureuse, prenant en compte les doses des ICS, ne démontre pas d'association significative entre des doses faibles d'ICS et la prématurité [**32**]. Ainsi, il est généralement consensuel dans les lignes directrices que des doses faibles à modérées d'ICS n'ont pas d'effets néfastes sur la prématurité.

Par conséquent, les recommandations internationales actuelles pour le contrôle d'un asthme léger durant la grossesse approuvent soit une dose faible d'ICS , soit

aucun médicament de contrôle, selon la sévérité de l'asthme. Ainsi, bien que les études ci-haut citées et le restant de la littérature abondent dans le sens d'un effet nul des ICS sur la prématurité, aucune des études de nature longitudinale n'a pris en considération que le contrôle de l'asthme est un facteur de confusion variant dans le temps. Cette omission pourrait créer un biais potentiel dans les résultats obtenus et doit être prise en compte dans l'analyse. Aucune étude observationnelle longitudinale utilisant des méthodes d'inférence causale pouvant tenir compte de ce type de biais n'a été publiée en ce moment. La réalisation d'une telle étude est donc nécessaire afin de fournir des données probantes pouvant venir infirmer ou confirmer les résultats publiés dans la littérature. Ceci a justifié l'emploi des méthodes d'inférence causale ci-dessous exploitées.

1.2. Bases de données administratives

Plusieurs études ci-haut citées sont des études observationnelles dont les données proviennent de bases de données administratives (BDA). En effet, étant donné la grande complexité de la réalisation d'études randomisées contrôlées sur les femmes enceintes, les études observationnelles deviennent alors essentielles pour établir l'innocuité des médicaments pris pendant la grossesse. Une source de données particulièrement utile dans la conduite de ce type d'études sont les BDA, dont l'utilisation en santé a connu une augmentation marquée au cours des dernières décennies [**33**]. Ainsi, de nombreux domaines liés à l'épidémiologie, particulièrement la pharmacoépidémiologie, ont adopté les BDA comme leur principale source de données observationnelles [**33**].

1.2.1. Bases de données administratives en santé

Les données administratives sont des données collectées de façon régulière par des organisations gouvernementales, des entreprises et d'autres organismes

dans un nombre de domaines variés tels l'impôt, la santé ou l'éducation, et ce, principalement à des fins d'archivage du service fourni pour l'enregistrement de transactions et la tenue de dossiers.

Plus particulièrement, en santé, les BDA sont des dépôts massifs de données collectées à travers les divers soins de santé. Ces données sont généralement enregistrées dans des bases fédérales ou provinciales des ministères de la Santé et des bases d'assureurs ou de compagnies privées de services en santé afin d'être utilisées par les assureurs et les prestataires de soins [**34**].

La plupart des BDA ayant été initialement mises en place comme outils de suivi des systèmes de santé d'un point de vue administratif et financier, elles diffèrent généralement d'autres données électroniques relatives à la santé, par exemple des dossiers de santé électroniques. Ces derniers sont principalement utilisés par les cliniciens pour documenter l'état clinique des patients et correspondent à une version numérique des dossiers patients, alors que les BDA tirent plutôt leurs données de demandes de remboursement de services de santé, de procédures médicales, d'ordonnances servies et de diagnostics [**35**, **36**]. Ainsi, les données provenant de BDA sont généralement générées au moment de la sortie de l'hôpital ou au moment de la prestation des services de santé. Elles contiennent habituellement des diagnostics primaires et secondaires, des informations sur les procédures effectuées et les médicaments reçus et contiennent ou peuvent être liés à des fichiers contenant des informations démographiques. À l'inverse des dossiers de santé électroniques, les données administratives ne contiennent généralement pas de résultats de laboratoire ou des mesures cliniques telles que la pression artérielle, la taille et le poids [**33**, **35**, **37**].

1.2.2. BDA en santé au Québec

Au Canada, aux États-Unis et en Europe, de nombreuses BDA ont été construites par les gouvernements et les assureurs privés chargés du financement des soins de santé. Les plus fréquemment utilisées dans la littérature pour la recherche pharmacoépidémiologique sont la *Saskatchewan Health* de la province de Saskatchewan au Canada [**38**], dont l'utilisation remonte au début des années 1970, *Medicaid*, *Medicare*, *Kaiser Permanente* et la *Veteran Affairs Clinical Database* aux États-Unis [**33**, **39**] et la base de données *General Practice Research Database* (aujourd'hui *Clinical Practice Research Datalink*) au Royaume-Uni, qui, de toutes les BDA européennes, a été la plus utilisée pour la recherche pharmacoépidémiologique [**40**].

Relativement au Québec, les données administratives proviennent de deux sources principales : les bases de données de la régie de l'assurance maladie du Québec (RAMQ) et la banque de données ministérielle «Maintenance et exploitation des données pour l'étude de la clientèle hospitalière »(MED-ECHO). La RAMQ est un régime d'assurance maladie et médicaments public. Les bases de données MED-ECHO viennent compléter les informations de la RAMQ et regroupent des informations relatives aux séjours hospitaliers de courte durée. En somme, les principaux éléments du système de données administratives du Québec sont des données sur les consultations médicales, les hospitalisations et les visites aux services d'urgence, ainsi que de l'information sur les médicaments d'ordonnance délivrés en pharmacie pour les Québécois couverts par l'assurance médicaments de la RAMQ. Le régime public d'assurance médicaments de la RAMQ couvre les personnes de 65 ans ou plus, les prestataires d'aide financière, les personnes non admissibles à un régime privé et les enfants des assurés du régime public. Les différentes bases disponibles sont répertoriées dans le Tableau 1.1.

Tableau 1.1. BDA au Québec

	Base de données	Description
RAMQ	1. Fichier d'inscription des personnes assurées	Informations sociodémographiques à propos des personnes assurées
	2. Période d'admissibilité au régime public d'assurance médicaments du Québec	Périodes de couverture par le régime de la RAMQ
	3. Services pharmaceutiques	Informations sur les médicaments délivrés à la pharmacie
	4. Services rémunérés à l'acte	Informations liées à la facturation à l'acte des professionnels de santé
MED -ECHO	1. Séjours hospitaliers	Informations sur les séjours hospitaliers de courte durée
	2. Diagnostics	Diagnostics à l'admission, principal et secondaires
	3. Services	Informations sur les divers services hospitaliers visités
	4. Soins intensifs	Informations sur les séjours aux soins intensifs
	5. Interventions	Interventions pratiquées : actes thérapeutiques, diagnostics, chirurgicaux et obstétricaux

Par conséquent, il est évident que ces systèmes de données fournissent une grande variété d'informations cliniques et démographiques. Bien qu'en général la collecte d'information dans ces bases de données ne soit pas prévue à des fins de recherche, les BDA québécoises peuvent être utilisées comme un outil de recherche robuste [**41**], particulièrement en ce qui a trait aux variables liées à la grossesse chez les femmes asthmatiques [**42**].

1.2.3. Avantages

La pharmacoépidémiologie est l'un des domaines ayant le plus bénéficié de la croissante utilisation des données administratives. De fait, la pharmacoépidémiologie étudie souvent des effets indésirables de médicaments à long terme ou avec une incidence faible et nécessite ainsi des tailles d'échantillon considérables, de

même que des périodes de suivi prolongées. Étant donné que la recherche scientifique est souvent limitée en ressources, les BDA offrent des alternatives rentables à ces problèmes.

Nombre sont les études ayant fait l'objet d'évaluer l'utilité et la qualité des données administratives [**33**, **37**, **43**–**45**]. Comme ce sujet n'est pas l'objet de ce mémoire, nous tenterons de synthétiser ces informations de façon succincte. En résumé, les BDA constituent un énorme potentiel pour la recherche, car elles fournissent une grande variété de données déjà stockées avec un investissement minimal de logistique et de temps et qu'elles sont relativement facilement accessibles, peu coûteuses et collectées sur de larges populations pendant une longue période de temps. Ceci permet notamment de pallier les problèmes de puissance statistique faible et d'impossibilité d'effectuer des analyses de sous-groupes et d'évènements rares. Par ailleurs, les BDA permettent plus facilement la conception d'études longitudinales nécessitant un suivi de cohortes à long terme. Étant un sous-produit de systèmes administratifs existants, les données sont enregistrées indépendamment des patients et des chercheurs, ce qui permet non seulement de minimiser certains biais de sélection et d'information, tels les biais de non-réponse, de rappel et d'intervieweur, mais aussi d'éviter la nécessité d'obtenir le consentement des patients afin d'avoir accès à leurs informations. De surcroît, en raison de leur représentativité des soins cliniques dans la population générale, les BDA offrent facilement la possibilité d'étudier l'efficacité et l'innocuité en contextes réels de post-commercialisation, ce qui augmente la validité externe des études.

1.2.4. Limites

Bien que les données administratives puissent présenter de précieux avantages pour mener des études observationnelles, les données n'ont été en aucun cas générées spécifiquement à des fins de recherche. Par conséquent, cela implique que ce

type de données présente plusieurs limites non négligeables, qui seront brièvement abordées ci-après.

Une des allégations couramment formulées à l'encontre des données administratives concerne la généralisabilité des études utilisant des BDA, puisque les personnes y figurant ne sont souvent pas représentatives de la population d'intérêt [**37**]. Par exemple, certaines BDA peuvent contenir en moyenne des personnes plus âgées ou à statut socio-économique plus faible que la population générale. Une seconde limite porte également sur le manque de continuité dans la couverture du régime d'assurance, ce qui conduit à une instabilité au sein de la population suivie [**46**]. Par conséquent, la possibilité d'analyses longitudinales peut être entravée par les inscriptions et désinscriptions continuelles des participants aux régimes d'assurance. Un autre défi considérable des BDA réside dans l'évaluation de l'adhésion et de la conformité au traitement. En effet, les bases de données pharmaceutiques fournissent des informations à savoir si un patient a reçu une prescription de médicament et si celle-ci a été délivrée. Cependant, aucune information n'est disponible pour évaluer si le patient a effectivement ingéré le médicament conformément à la dose prescrite [**47**], d'où les nombreuses études de validation de mesures d'adhésion pour les données d'ordonnances [**48**]. De plus, plusieurs BDA ne contiennent pas les médicaments pris à l'hôpital, sans compter qu'il n'existe pas de données pour des médicaments prescrits, mais non délivrés. Un sujet supplémentaire qui doit être pris en compte lors de l'utilisation de grandes bases de données est celui de la signification statistique et de son interprétation. En général, étant donné une taille d'échantillon importante, toutes les comparaisons peuvent donner des différences statistiquement significatives, même celles de plus faible ampleur. Par conséquent, l'utilisation de la signification statistique comme facteur discriminatoire peut dans de nombreux cas devenir peu pratique [**49**, **50**].

Les informations manquantes sont considérées comme des limites importantes des études de BDA. Les BDA ne contiennent généralement pas de covariables reliées au tabagisme, à l'alcool, etc. Ces variables sont souvent essentielles dans l'ajustement pour la confusion et, vu leur omission, peuvent créer un effet résiduel dû à la confusion [**43**]. Enfin, le biais de classification, dû au fait que les codes diagnostiques ne représentent pas toujours le diagnostic exact sous-jacent, est l'une des plus grandes critiques relatives aux BDA. Bien que ce biais ne soit pas intrinsèque aux BDA, il y est souvent présent [**33**,**51**]. Par conséquent, l'exactitude des données est un aspect particulier des BDA qui a été étudié exhaustivement ; les résultats variant d'une étude à l'autre, selon la BDA et le contexte analysé [**52**,**53**]. Par ailleurs, une revue d'articles portant sur la validité des BDA [**54**] conclut que les résultats d'études de validation sont mitigés, mais tendent à soutenir un niveau de validité élevé.

En résumé, lors de l'utilisation de BDA, les chercheurs devraient reconnaître de façon transparente les limites inhérentes aux données administratives en examinant comment celles-ci peuvent influencer leurs résultats [**55**].

1.2.5. Particularité des BDA

Une particularité des données administratives peu abordée dans la littérature est que les données sont recueillies sur une base continue en temps réel. Ceci explique pourquoi les BDA sont définies comme une source de données réelles. De nombreuses définitions de données réelles existent. Selon le groupe de travail du *Relative Effectiveness* du Forum Européen, il s'agit «d'une mesure permettant de comprendre les données sur les soins de santé recueillies dans les circonstances de la vie réelle» [**56**]. Communément, ces données sont définies comme tout ce qui n'est pas interventionnel [**56**].

Ainsi, dans le but de mieux comprendre comment les BDA sont structurées, il est intéressant de se pencher sur le processus de collecte de données. Les données administratives sont recueillies en temps continu, ce qui signifie que les données sont enregistrées à tout moment lorsque les services se produisent, c'est-à-dire lorsque les patients reçoivent des soins hospitaliers, renouvellent une prescription, etc. De fait, ces données sont enregistrées de façon continue chaque jour, impliquant que pour chaque patient, un suivi quotidien des évènements survenus peut être extrapolé à partir des données recueillies dans les BDA. Or, lors d'une étude longitudinale cette nature riche et complexe propre aux BDA est à la source même d'un problème reconnu pour les méthodes longitudinales standard d'inférence causale [**8**, **57**, **58**] : la discrétisation arbitraire de la ligne du temps. Cette limite des méthodes d'inférence causale, qui a mené au développement méthodologique dont fait l'objet ce mémoire, sera plus amplement décrite dans le manuscrit du chapitre 3.

1.3. Inférence causale

Au sens large, l'inférence causale est un concept englobant des méthodes cherchant à estimer la relation de cause à effet entre deux variables, par exemple l'effet d'une exposition sur une issue. Ainsi, la présente section vise tout d'abord à décrire le modèle causal de Neyman-Rubin, un cadre permettant d'identifier statistiquement des effets causaux, ainsi que les hypothèses causales nécessaires à l'identification d'une vaste gamme de paramètres causaux. Par la suite, des méthodes permettant l'estimation de paramètres causaux, particulièrement pour des données longitudinales d'analyse de survie, seront décrites. Finalement, l'estimation par maximum de vraisemblance ciblée (TMLE) sera abordée, avec un accent particulier sur la LTMLE et la PLTMLE. Cependant, avant de se lancer dans le cadre de l'inférence causale, il est important de mentionner certains concepts

d'analyse de survie qui seront utilisés subséquemment.

1.3.1. Concepts de base en analyse de survie

L'analyse de survie cherche à modéliser le temps jusqu'à l'occurrence d'un certain évènement. Par exemple, dans notre cas, on pourrait s'intéresser à la modélisation du temps jusqu'à l'accouchement, et comment il pourrait être affecté par l'exposition à certains médicaments antiasthmatiques.

Soit la variable aléatoire T le temps jusqu'à l'évènement d'intérêt. En analyse de survie, plusieurs fonctions d'intérêt existent afin de décrire T, notamment la fonction de survie et la fonction de risque. La fonction de survie représente la probabilité que l'évènement d'intérêt ne se soit pas produit jusqu'au temps t et est donnée par $S(t) = P(T > t)$, où $P(\cdot)$ est la fonction de probabilité. $S(t)$ est une fonction monotone décroissante, puisqu'au fil du temps de plus en plus d'évènements surviennent. On définit $S(t) = 1$ au temps $t = 0$ sous l'hypothèse qu'aucun évènement ne puisse se produire au début du suivi. La fonction de risque (souvent dénommée taux de risque ou risque de survie) représente le taux instantané d'un évènement pour un individu ayant survécu jusqu'au temps t :

$$\lambda(t) = \lim_{\Delta\, t \to 0} \frac{P(t < T \leq t + \Delta\, t | T > t)}{\Delta\, t}$$

Une particularité des méthodes d'analyse de survie est la possibilité de ne pas observer l'évènement d'intérêt pour certains individus. Le temps de suivi, d'un sujet i est dit censuré lorsque son évènement n'est pas observé, de sorte que le temps jusqu'à l'évènement correspondant, T_i, ne peut être déterminé. Les méthodes d'analyse de survie sont spécialement développées pour l'analyse de ce type de données.

Le modèle de risques proportionnels de Cox [**59**] est la méthode statistique prenant en compte la censure la plus couramment utilisée afin d'examiner la relation entre T et des variables explicatives. Il est fondé sur l'hypothèse que les risques restent proportionnellement constants dans le temps et est une méthode semi-paramétrique comportant un terme non paramétrique représentant le risque de base et une fonction paramétrique des covariables. Le modèle de Cox décrit donc le taux de risque conditionnel en fonction d'un vecteur de covariables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ et est défini par :

$$\lambda(t|\mathbf{X}) = \lambda_0(t)exp(\boldsymbol{\alpha}\mathbf{X}) \tag{1.3.1}$$

où $\boldsymbol{\alpha}$ est le vecteur de coefficients associés à $\mathbf{X}$, $\lambda_0(t)$ est le risque de base qui correspond au risque au temps t lorsque $\mathbf{X} = 0$ et $exp(\cdot)$ représente la fonction exponentielle. L'association entre une covariable et le risque de survie dans un modèle de Cox est mesurée en termes de rapport de risques. Le rapport de risques est tout simplement le rapport des taux de risque pour différentes valeurs de $\mathbf{X}$:

$$\frac{\lambda_i(t|\mathbf{X})}{\lambda_j(t|\mathbf{X})} = \frac{\lambda_0(t)exp(\boldsymbol{\alpha}\mathbf{X}_i)}{\lambda_0(t)exp(\boldsymbol{\alpha}\mathbf{X}_j)} = \frac{exp(\boldsymbol{\alpha}\mathbf{X}_i)}{exp(\boldsymbol{\alpha}\mathbf{X}_i)} = exp(\boldsymbol{\alpha}\mathbf{X}_i - \boldsymbol{\alpha}\mathbf{X}_j).$$

En particulier, si on cherche à comparer deux groupes indiqués par une variable dichotomique $X = \{0, 1\}$, le rapport de risques devient $exp(\alpha)$.

Parfois, lors d'études longitudinales les caractéristiques des individus sont mesurées à répétition pendant le suivi. Ainsi, des raffinements du modèle de Cox incluant des variables dépendantes du temps existent [**60**]. Une variable dépendante du temps (ou variant dans le temps) est une variable dont les valeurs peuvent

changer à travers le temps. Par exemple, $\mathbf{X}(t)$ correspond au vecteur de covariables $\mathbf{X}$ au temps t auquel est lié le modèle de Cox avec variables dépendantes du temps :

$$\lambda(t|\mathbf{X}(t)) = \lambda_0(t)exp(\boldsymbol{\alpha}\mathbf{X}(t)). \tag{1.3.2}$$

Ce dernier modèle est particulièrement d'intérêt dans notre exemple motivateur où l'on s'intéresse à comment la durée de gestation pourrait être affectée par des traitements pour l'asthme qui peuvent varier au cours de la grossesse.

1.3.2. Une note sur l'approximation du modèle de Cox par régression logistique groupée

Dans les études longitudinales prospectives, un groupe d'individus est suivi pendant une certaine période de temps et les données sont habituellement recueillies de façon répétée à plusieurs intervalles dans le temps. S'il est d'intérêt d'évaluer l'effet de mesures répétées d'exposition sur le temps jusqu'à un certain évènement, une alternative au modèle de Cox est la régression logistique groupée [**61**]. Au lieu de traiter la période de suivi sous-jacente comme continue, tel un modèle de Cox, cette méthode divise le suivi en intervalles de temps et traite chaque intervalle comme une étude distincte où l'évènement d'intérêt serait évalué à la fin du suivi. Dans ce cas, à chaque temps discret, chaque personne encore à risque de subir un évènement représente une nouvelle observation. Dans l'analyse, toutes les observations personne-temps de tous les intervalles sont regroupées dans un échantillon et une régression logistique est effectuée sur l'ensemble de données groupées, suivant le modèle :

$$logit[P(D(t) = 1|D(t-1) = 0, \mathbf{X})] = \beta_0(t) + \boldsymbol{\beta}\mathbf{X}(t) \tag{1.3.3}$$

où t désigne des unités d'intervalles de temps depuis le début du suivi, $D(t)$ est un indicateur de décès par le temps t et $\beta_0(t)$ est un intercept relatif au temps t. L'intercept $\beta_0(t)$ est souvent considéré comme constant à travers le temps, auquel cas $\beta_0(t) = \beta_0$ [**62**]. Lorsque les intervalles entre les temps de mesures sont petits et que le taux d'évènements est petit dans chaque intervalle, *D'Agostino et al.* [**62**] ont montré que les paramètres $\boldsymbol{\beta}$ de la régression logistique groupée sont asymptotiquement équivalents aux paramètres $\boldsymbol{\alpha}$ obtenus du modèle de régression de Cox avec variables dépendantes du temps.

Dans la littérature d'inférence causale, la régression logistique groupée est fréquemment utilisée pour approximer les modèles de Cox, plus particulièrement dans les cas où une régression pondérée est effectuée, ce qui est commun avec certaines méthodes d'inférence causale [**63**–**65**]. Cette approximation fut avant tout suggérée parce que les logiciels disponibles à l'époque du développement de ces méthodes étaient incapables d'incorporer des poids variant dans le temps au modèle de Cox [**63**]. Par contre, comme mentionné précédemment, cette approche est inadéquate lorsque l'évènement n'est pas rare. Ainsi, plus récemment, les modèles de Cox pondérés ont fait l'objet d'études de simulation qui démontrèrent que les approximations par régression logistique groupée augmentaient la variabilité de l'estimation [**66**].

Ces méthodes s'englobent dans une approche statistique traditionnelle pour l'estimation d'effets, mais ne définissent pas exactement ce que l'on entend par un effet ou quel impact une nouvelle intervention peut avoir sur la population à l'étude. Pour ce faire, nous commencerons par définir un cadre régissant l'inférence causale, nous permettant de définir et d'estimer des effets causaux.

1.3.3. Le cadre des issues potentielles

Le cadre des issues potentielles, aussi connu sous le modèle causal de Neyman-Rubin [**67**] a été introduit par Neyman [**68**], dans le contexte des études expérimentales, et établi par Rubin [**2**, **69**] pour les études observationnelles comme une base pour l'inférence causale. Ainsi, le modèle de Neyman-Rubin est l'un des cadres les plus utilisés pour décrire les effets causaux d'une manière assez intuitive et directe.

Supposons que l'on s'intéresse à l'effet causal d'une exposition A ponctuelle, c'est-à-dire évaluée à un point fixe dans le temps, sur une issue Y. Soit A_i et Y_i l'exposition et l'issue d'intérêt pour le sujet i, respectivement. À des fins de simplicité, et à moins d'avis contraire, nous supposons à partir de maintenant que A est une exposition binaire prenant la valeur $A_i = 1$ si le sujet i est exposé et 0 sinon. Afin de définir l'effet causal de A sur Y, pensons à une intervention hypothétique qui force A à prendre la valeur a pour chaque sujet et définissons Y_i^a comme l'issue obtenue par le sujet i dans le cadre de cette intervention hypothétique, c'est-à-dire si l'exposition avait été forcée à $A = a$. Les variables Y^a sont appelées issues potentielles ou issues contrefactuelles puisqu'elles décrivent la valeur de l'issue qui aurait été observée en fonction d'une valeur d'exposition potentielle qui pourrait ne pas être celle que le sujet a réellement reçue. Par exemple, dans le cas où un individu est exposé au traitement, $A = 1$, ou non-exposé, $A = 0$, ce traitement dichotomique implique que chaque individu a deux issues potentielles (Y_i^1, Y_i^0). Ici, l'effet causal individuel est la différence entre l'issue potentielle d'une personne avec et sans l'exposition : $Y_i^1 - Y_i^0$. Évidemment, pour estimer un effet causal pour le sujet i, il faut observer les deux issues potentielles. Le problème trivial de l'utilisation de cette définition hypothétique est que nous ne pouvons observer qu'un individu sous une version de l'exposition à un moment donné, c'est-à-dire que nous ne pouvons observer que Y_i^1 ou Y_i^0 :

le problème fondamental de l'inférence causale. Cependant, cette difficulté peut être résolue sous un ensemble d'hypothèses nous permettant d'identifier un effet causal.

1.3.4. Hypothèses d'identifiabilité

Soit $\boldsymbol{O}_i = (\boldsymbol{L}_i, A_i, Y_i)$ un vecteur de données observées à un point dans le temps pour un individu i, où $\boldsymbol{L}$ est un vecteur de covariables nécessaires et suffisantes pour le respect des hypothèses d'identifiabilité. Dans ce contexte, un paramètre qui pourrait être d'intérêt est l'effet causal moyen (ATE) d'une exposition A sur une issue Y, soit la différence entre les résultats espérés si toute la population est exposée ($A = 1$) versus si toute la population n'est pas exposée ($A = 0$), de telle sorte que :

$$ATE = E[Y^1 - Y^0].$$

Les hypothèses nécessaires pour identifier l'ATE à partir de données observées sont : l'hypothèse d'ordre temporel, l'hypothèse de «valeur de traitement des unités stable» (SUTVA), l'hypothèse de positivité et l'hypothèse d'interchangeabilité conditionnelle. Ces hypothèses sont souvent appelées les conditions d'identifiabilité [**70**].

L'hypothèse d'ordre temporel stipule que les données sont toujours observées dans le même ordre, de telle sorte que $\boldsymbol{L}$ précède A, qui à son tour précède Y, justifiant ainsi le lien de cause à effet de A vers Y. SUTVA [**71**] est une hypothèse qui permet de définir l'issue contrefactuelle. Elle stipule que le résultat potentiel de l'individu i, Y_i^a, ne devrait pas être affecté par l'exposition d'autres unités $j \neq i$. Ceci est parfois défini sous l'hypothèse d'aucune interaction

entre les individus [**72**]. SUTVA suppose également qu'il n'y a pas de versions différentes d'une même exposition, c'est-à-dire que les issues potentielles pour chaque individu sous chaque exposition possible sont bien définies et prennent une valeur unique. S'il existe plusieurs versions d'une même exposition et si ces différentes versions donnent lieu à des issues potentielles différentes, cette hypothèse sera violée et un effet causal ne peut être identifié. Cette dernière partie de l'hypothèse est aussi parfois dénommée l'hypothèse de cohérence [**73**, **74**], notée $Y_i^a = Y_i$ lorsque $A_i = a$. Dans la littérature, l'hypothèse de cohérence est connue sous le nom d'«aucune version du traitement» [**75**] ou d'«interventions bien définies» [**76**]. Pour les expositions binaires, elle est parfois formulée comme : $Y = AY^1 + (1 - A)Y^0$, de sorte qu'elle indique de manière évidente que l'issue contrefactuelle, Y^a, sous la valeur d'exposition potentielle a est égale à l'issue réellement observée Y. Par conséquent, l'hypothèse de cohérence garantit que si un sujet possède $A = a$ et $Y = y$ en l'absence d'intervention, le sujet atteindrait également le niveau $Y = y$ lorsqu'il serait forcé au niveau $A = a$.

Une autre hypothèse requise est l'hypothèse de positivité [**77**] :

$$P(A = a|\boldsymbol{L} = \boldsymbol{l}) > 0 \ \forall a \in \mathcal{A}, \boldsymbol{l} \in \boldsymbol{\mathcal{L}} \text{ t.q. } P(\boldsymbol{L} = \boldsymbol{l}) \neq 0$$

où $P(\boldsymbol{L} = \boldsymbol{l})$ est la distribution marginale de $\boldsymbol{L}$, $P(A = a|\boldsymbol{L} = \boldsymbol{l})$ est la distribution conditionnelle de A chez les sujets avec $\boldsymbol{L} = \boldsymbol{l}$ et $\mathcal{A}$ et $\boldsymbol{\mathcal{L}}$ sont les domaines de valeurs possibles de A et $\boldsymbol{L}$, respectivement. Afin d'identifier Y^a cela signifie que chaque sujet dans la population doit avoir une probabilité non nulle de recevoir chaque valeur d'exposition dans chaque strate de $\boldsymbol{L}$. L'hypothèse de positivité est également appelée hypothèse de traitement expérimental [**78**] puisque cette condition est garantie dans des expériences randomisées stratifiées. Notons que cette hypothèse est parfois appelée positivité théorique en opposition à un

manque de positivité en pratique [**79**] qui se produit lorsque la probabilité estimée d'une valeur d'exposition est zéro dans certains sous-groupes de covariables dans les données observées. Ceci peut être dû à une faible taille d'échantillon ou des expositions rares. Dans ce cas, nous sommes confrontés à un non-respect pratique de l'hypothèse de positivité.

La dernière hypothèse nécessaire est l'hypothèse d'interchangeabilité conditionnelle [**2**, **70**] :

$$Y^a \perp A | \boldsymbol{L} = \boldsymbol{l} \; \forall a \in \mathcal{A}, \boldsymbol{l} \in \boldsymbol{\mathcal{L}}.$$

Ceci indique que l'exposition A est indépendante de l'issue potentielle dans les strates de $\boldsymbol{L}$. L'interchangeabilité conditionnelle implique qu'à l'intérieur des niveaux des variables mesurées $\boldsymbol{L}$, les exposés, s'ils n'avaient pas été exposés, auraient eu la même distribution de l'issue que ceux non exposés, c'est-à-dire que les distributions des covariables dans les groupes exposés et non exposés sont égales. Dans la littérature, sous l'hypothèse d'interchangeabilité conditionnelle, on dit que l'attribution de l'exposition peut être ignorée (*ignorable assignment mechanism*) [**80**]. Par exemple, l'hypothèse d'interchangeabilité conditionnelle est trivialement respectée dans les études randomisées où l'exposition est attribuée complètement au hasard. Conséquemment, l'hypothèse d'interchangeabilité conditionnelle est aussi connue sous l'hypothèse d'absence de facteurs de confusion non mesurés [**81**]. En présence de confusion, l'exposition dépend des covariables observées. En l'absence de toutes les covariables nécessaires afin d'éliminer la confusion, l'hypothèse d'interchangeabilité conditionnelle ne tient plus.

Sous ces conditions, un effet causal peut être estimé à partir de données observationnelles en dépit du problème fondamental de l'inférence causale. Ainsi, l'ATE peut être identifié comme suit :

$$\begin{aligned} ATE = E[Y^1 - Y^0] &= E[Y^1] - E[Y^0] \\ &= E[E[Y^1|\boldsymbol{L} = \boldsymbol{l}]] - E[E[Y^0|\boldsymbol{L} = \boldsymbol{l}]] \\ &= E[E[Y^1|\boldsymbol{L} = \boldsymbol{l}, A = 1]] - E[E[Y^0|\boldsymbol{L} = \boldsymbol{l}, A = 0]] \\ &\quad \text{par interchangeabilité conditionnelle} \\ &= E[E[Y|\boldsymbol{L} = \boldsymbol{l}, A = 1]] - E[E[Y|\boldsymbol{L} = \boldsymbol{l}, A = 0]] \\ &\quad \text{par SUTVA} \end{aligned}$$

où $E[Y|\boldsymbol{L} = \boldsymbol{l}, A = a]$ est estimable à partir des données observées.

Lorsque les conditions d'identifiabilité sont vérifiées, un effet causal peut être estimé par des méthodes statistiques standard. Cependant, pour les structures de données longitudinales, des conditions d'identifiabilité différentes doivent être utilisées.

1.3.5. Données longitudinales et confusion dépendante du temps

Dans ce scénario, supposons que l'on observe des données dépendantes du temps de la forme longitudinale générale

$$\boldsymbol{O} = (\boldsymbol{L}(0), A(0), Y(1), \boldsymbol{L}(1), A(1), \ldots, Y(K), \boldsymbol{L}(K), A(K), Y(K+1))$$

où $Y(t)$ représente une issue variant dans le temps, $\boldsymbol{L}(t)$ est un vecteur de covariables et $A(t)$ est la valeur d'exposition, tous mesurés au temps t dans cet ordre. En particulier, $\boldsymbol{L}(0)$ contient toutes les variables confondantes de base. Une notation communément utilisée est le trait suscrit qui dénote l'historique d'une

variable. Ainsi, $\bar{A}(t) = (A(0), A(1), \ldots, A(t))$ et $\bar{\boldsymbol{L}}(t) = (\boldsymbol{L}(0), \boldsymbol{L}(1), \ldots, \boldsymbol{L}(t))$ représente l'historique des expositions et des covariables jusqu'au temps t, respectivement. Par exemple, des représentations de différents historiques fixes d'exposition, aussi dénommés régimes d'exposition, sont les cas où un individu est toujours exposé, $\bar{a}(t) = (1, 1, \ldots, 1)$ ou jamais exposé $\bar{a}(t) = (0, 0, \ldots, 0)$ à chaque point dans le temps. De plus, soit $Y^{\bar{a}}(t)$ l'issue contrefactuelle au temps t sous l'historique de traitement fixé à $\bar{A}(t-1) = \bar{a}(t-1)$. Ainsi, soit $\bar{\mathcal{A}}$ l'ensemble de tous les régimes d'exposition possibles et $\bar{\mathcal{A}}(t)$ l'ensemble de toutes les valeurs possibles de l'exposition jusqu'au temps t.

Avec des données observationnelles, l'hypothèse d'interchangeabilité conditionnelle n'est pas assurée compte tenu des déséquilibres dans la répartition des covariables entre les groupes d'exposition, ce qui invalide l'estimation d'un effet causal par de simples mesures d'association. Les variables créant ces déséquilibres sont souvent nommées des facteurs de confusion. Afin de permettre l'estimation d'effets causaux par des mesures d'association provenant d'études observationnelles, il faut donc contrôler pour tous ces facteurs de confusion. Cependant, lorsque ces covariables sont des facteurs de confusion dépendants du temps affectés par l'exposition antérieure Figure 1.2, des méthodes d'ajustement traditionnelles telles que la stratification ou des modèles de régression, dans lesquelles l'exposition et les facteurs de confusion sont inclus comme covariables, produisent des estimations biaisées de l'effet causal d'une exposition variant dans le temps [**4**, **5**].

Un facteur de confusion dépendant du temps est une covariable variant dans le temps qui cause le(s) exposition(s) et le(s) issue(s) future(s). En particulier, un facteur de confusion dépendant du temps affecté par une exposition antérieure est une variable affectant à la fois l'exposition et les issues ultérieures, tout en étant affectée par une exposition antérieure, étant donc une variable intermédiaire dans

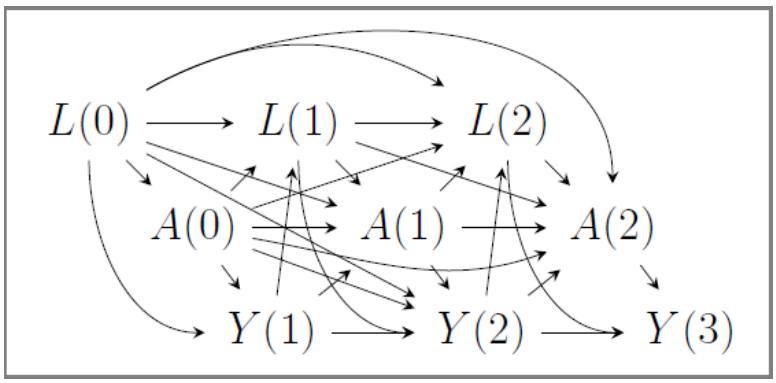

FIGURE 1.2. Exemple de confusion variant dans le temps pour trois points temporels

le chemin causal de l'exposition et de l'issue, également dénommé un médiateur. Dans de nombreuses études longitudinales observationnelles, il peut exister un facteur de confusion dépendant du temps qui est aussi une variable médiatrice. Tel est le cas dans notre exemple motivateur d'asthme durant la grossesse, où le contrôle de l'asthme est un tel facteur de confusion pour la relation entre les médicaments antiasthmatiques et la durée de gestation. En effet, le contrôle de l'asthme affecte le traitement ultérieur de l'asthme et la durée de gestation, tout en étant affecté par le traitement antérieur de l'asthme [**16**].

En résumé, ce type de facteur de confusion pose problème, car il biaise l'estimation de l'effet causal d'une exposition variant dans le temps. En effet, il faut contrôler pour les facteurs de confusion dépendants du temps, sinon l'estimateur de l'effet de traitement sera biaisé en raison de confusion résiduelle non contrôlée [**4**,**5**]. En revanche, on ne devrait pas ajuster pour un médiateur, ce qui faussera l'estimation de l'effet total de l'exposition, puisqu'ajuster pour de telles variables, s'il existe des variables latentes qui prédisent à la fois les facteurs de confusion et l'issue, produira de fausses associations entre l'exposition et l'issue [**4**, **5**]. On se retrouve alors avec un estimateur biaisé, que l'on ajuste ou non pour les facteurs de confusion.

1.3.6. Modèles structurels marginaux

Les modèles structurels marginaux (MSM) appartiennent à une classe de modèles causaux [**5**, **82**] proposés dans le but d'estimer les effets causaux d'une exposition variant dans le temps en présence de facteurs de confusion dépendants du temps affectés par une exposition antérieure. L'effet est modélisé à l'aide d'issues contrefactuelles. Les MSM sont structurels, car ils modélisent des variables contrefactuelles, un terme qui a été inventé dans les sciences sociales et économiques [**82**]. Il est dit marginal parce qu'il décrit l'effet de l'exposition sur la distribution marginale des issues contrefactuelles, $E[Y^{\bar{a}}(t)]$, par opposition à la modélisation habituelle d'une distribution conditionnelle par les modèles traditionnels.

Dans le cas d'études comportant des expositions mesurées à de multiples points dans le temps, il existe de nombreux modèles différents pouvant illustrer diverses relations exposition-issue. Par conséquent, une définition générale d'un MSM est :

$$E[Y^{\bar{A}}(t)|t, \boldsymbol{W}] = \eta(\boldsymbol{\beta}, \bar{A}, t, \boldsymbol{W}) = f^{-1}[\sum_{j=1}^{J} \boldsymbol{\beta}_j m_j(\bar{A}, t, \boldsymbol{W})],\ t = 1, \ldots, K+1 \quad (1.3.4)$$

où $m_j(\cdot)$ est une fonction de l'historique d'exposition, par exemple l'exposition cumulative au temps t, $cum(\bar{A}(t))$, du temps t et d'un éventuel sous-ensemble de covariables de base $\boldsymbol{W}$, qui pourraient par exemple être incluses pour quantifier un effet causal dans un sous-ensemble de la population d'intérêt. $f(\cdot)$ est une fonction de lien permettant à l'issue d'être reliée au modèle de façon linéaire. Les effets causaux de l'exposition sont alors les coefficients des variables d'exposition dans le modèle linéaire.

1.3.6.1. *Une précision à propos de la collapsibilité*

Un avantage des MSM par rapport aux modèles linéaires généralisés traditionnels concerne la collapsibilité car, pour de nombreux paramètres, le paramètre marginal n'est pas équivalent au paramètre conditionnel, ce qu'on appelle la non-collapsibilité [**83**].

Dans un contexte de modèles linéaires généralisés, on dit qu'un modèle est *collapsible* si le coefficient γ_1 dans la régression de Y sur A et W :

$$E[Y|A,W] = f^{-1}(\gamma_1 A + \gamma_2 W)$$

est égal à γ_1^* dans la régression sans W :

$$E[Y|A] = f^{-1}(\gamma_1^* A).$$

Cette condition est seulement respectée lorsque la fonction de lien est de nature linéaire ou log-linéaire [**83**] et lorsqu'on suppose que A et W sont indépendants [**84**]. En conséquence, si deux enquêteurs modélisaient les mêmes données, mais contrôlaient pour différentes variables, même si le contrôle de confusion était suffisant dans les deux analyses, les paramètres estimés à l'aide d'une telle méthode de régression pourraient ne pas être les mêmes. Ceci illustre l'utilité des MSM pour estimer des effets causaux.

1.3.7. Paramètre d'intérêt

Les données observées se composent de n copies indépendantes et identiquement distribuées de $\boldsymbol{O}$ tirées d'une distribution $P_0(\boldsymbol{O}) \in \mathcal{M}$. C'est-à-dire

$\boldsymbol{O} \sim P_0(\boldsymbol{O})$, où $\mathcal{M}$ est un modèle statistique pour $P_0(\boldsymbol{O})$ avec les hypothèses sur ce modèle statistique limitées de façon à s'assurer que $\mathcal{M}$ contient $P_0(\boldsymbol{O})$. Formellement, $\mathcal{M}$ représente une collection de distributions de probabilité potentielles dont $P_0(\boldsymbol{O})$ est le membre particulier qui génère les données en tant que variable aléatoire $\boldsymbol{O}$.

La vraisemblance de $P_0(\boldsymbol{O})$ est la densité conjointe des variables aléatoires qui composent $\boldsymbol{O}$. L'ordre de toutes les variables dans $\boldsymbol{O}$ tel que proposé implique la factorisation suivante de la vraisemblance des données observées en composantes orthogonales indexées par l'ordre temporel t :

$$P_0(\boldsymbol{O}) = \underbrace{\underbrace{\prod_{t=1}^{K+1} P_0((Y(t)|\bar{A}(t-1), \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1))}_{Q_{0Y}(\boldsymbol{O})} \underbrace{\prod_{t=0}^{K} P_0((\boldsymbol{L}(t)|\bar{Y}(t), \bar{A}(t-1), \bar{\boldsymbol{L}}(t-1))}_{Q_{0\boldsymbol{L}}(\boldsymbol{O})}}_{Q_0(\boldsymbol{O})} \underbrace{\prod_{t=0}^{K} P_0((A(t)|\bar{\boldsymbol{L}}(t), \bar{Y}(t), \bar{A}(t-1))}_{g_0(\boldsymbol{O})}.$$

Ainsi, $P_0(\boldsymbol{O})$ est factorisée en deux composantes distinctes $Q_0(\boldsymbol{O})$, les densités conditionnelles de l'issue et des covariables, correspondant à la partie non interventionnelle de la vraisemblance, et $g_0(\boldsymbol{O})$, la densité conditionnelle du processus d'exposition, qui est la contribution à la vraisemblance des variables sur lesquelles nous souhaitons intervenir.

Maintenant, nous pouvons considérer la distribution des données dans le cadre d'interventions sur les variables d'exposition. Dans un cadre contrefactuel où nous intervenons pour fixer les valeurs d'exposition à $\bar{A}(t) = \bar{a}(t)$ pour chaque individu, la factorisation des données contrefactuelles s'écrit :

$$P_0(\boldsymbol{O}^{\bar{a}}) \equiv Q_0(\boldsymbol{O}^{\bar{a}}) = \underbrace{\prod_{t=1}^{K+1} P_0((Y^{\bar{a}}(t)|\bar{Y}^{\bar{a}}(t-1), \bar{\boldsymbol{L}}^{\bar{a}}(t))}_{Q_{0Y}(\boldsymbol{O}^{\bar{a}})} \underbrace{\prod_{t=0}^{K} P_0((\boldsymbol{L}^{\bar{a}}(t)|\bar{Y}^{\bar{a}}(t), \bar{\boldsymbol{L}}^{\bar{a}}(t-1))}_{Q_{0\boldsymbol{L}}(\boldsymbol{O}^{\bar{a}})}$$

où $\boldsymbol{O}^{\bar{a}}$ sont les données contrefactuelles lorsque l'historique d'exposition a été fixé à $\bar{A}(t) = \bar{a}(t)$.

Cette distribution est connue sous le nom de formule-G [**4**]. Il est important de remarquer que tous les facteurs correspondant à $g_0(\boldsymbol{O})$ n'apparaissent plus dans la distribution contrefactuelle puisqu'ils sont maintenant fixés aux niveaux d'exposition intervenus avec probabilité 1. Toutes les autres distributions sont conditionnelles à $\bar{a}$, puisque $A(t)$ est fixée au niveau $a(t)$ pour tout t. Puisque la distribution des données contrefactuelles sous les interventions souhaitées est définie, nous pouvons maintenant proposer des paramètres d'intérêt sous cette distribution.

Par conséquent, soit le paramètre causal d'intérêt $\boldsymbol{\psi}$ défini par une fonction d'une composante de la vraie densité sous-jacente $P_0(\boldsymbol{O}^{\bar{a}})$. C'est-à-dire, soit $\boldsymbol{\psi} \equiv \Psi(Q_0(\boldsymbol{O}^{\bar{a}})) : \mathcal{M} \to \mathbb{R}^n$ où Ψ est une fonction qui prend comme argument un modèle d'un espace statistique, $\mathcal{M}$, et renvoie une valeur dans l'espace des réels. Ce paramètre est une fonction de la distribution des $\boldsymbol{L}^{\bar{a}}(t)$ et $Y^{\bar{a}}(t)$, qui sont générées dans un cadre interventionnel. Malheureusement, ces données contrefactuelles en cadre interventionnel ne sont pas toutes observables. Par conséquent, afin d'identifier $\boldsymbol{\psi}$ comme une fonction de la distribution génératrice des données observées $P_0(\boldsymbol{O})$, il suffit que les hypothèses d'identifiabilité généralisées ci-après mentionnées soient respectées. De fait, l'effet causal $\boldsymbol{\psi}$ peut être estimé au moyen d'un paramètre statistique exprimé en fonction de la distribution générant les données observées : $\boldsymbol{\psi}_n = \Psi(P_0(\boldsymbol{O})) : \mathcal{M} \to \mathbb{R}^n$. De plus, indépendamment de la véracité des hypothèses d'identifiabilité, ce paramètre statistique représente tout de même une mesure d'effet interprétable.

Dorénavant, considérons la même structure de données longitudinales $\boldsymbol{O}$ définie ci-haut, mais avec la particularité que $Y(t)$, $t = 1, \ldots K+1$, est maintenant un indicateur monotone d'un évènement par le temps t prenant la valeur $Y(t) = 1$ si l'évènement s'est produit avant ou au temps t, $Y(t) = 0$ sinon. Ainsi, s'il est d'intérêt d'estimer l'effet causal d'une fonction de A, t, et de covariables de base $\boldsymbol{W} \subset \boldsymbol{L}(0)$ sur la probabilité contrefactuelle de la survenue de l'évènement par le temps t, étant donné que l'évènement n'est pas survenu au temps $t-1$, il est possible d'employer un MSM correspondant à un modèle logistique groupé :

$$\eta(\boldsymbol{\beta}, \bar{A}, t, \boldsymbol{W}) \equiv logit[P(Y^{\bar{A}}(t) = 1 | Y^{\bar{A}}(t-1) = 0, t, \boldsymbol{W})] = \boldsymbol{\beta} m(\bar{A}, t, \boldsymbol{W}). \quad (1.3.5)$$

Par conséquent, notre paramètre causal d'intérêt peut être défini comme suit :

$$\boldsymbol{\psi} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, E \sum_{t=1}^{K+1} \sum_{\bar{a} \in \bar{\mathcal{A}}} \{ Y^{\bar{a}}(t) log(expit(\boldsymbol{\beta} m(\bar{a}, t, \boldsymbol{W}))) + (1 - Y^{\bar{a}}(t)) log(1 - expit(\boldsymbol{\beta} m(\bar{a}, t, \boldsymbol{W}))) \} \ . \quad (1.3.6)$$

En d'autres termes, notre paramètre causal d'intérêt est le vecteur des coefficients du MSM qui minimise l'espérance de la fonction de perte d'une régression logistique contrefactuelle.

1.3.8. Hypothèses d'identifiabilité pour données longitudinales

Les hypothèses d'identifiabilité généralisées pour les données longitudinales peuvent être définies comme suit :

(1) Cohérence :

Si $\bar{A}_i(t-1) = \bar{a}_i(t-1) \Rightarrow Y_i^{\bar{a}}(t) = Y_i(t)$

(2) Positivité :

$P(A(t) = a(t)|\bar{\boldsymbol{L}}(t), \bar{Y}(t), \bar{A}(t-1)) > 0 \ \forall a(t) \in \mathcal{A}(t), t = 0, \ldots, K$

t.q. $P(\bar{\boldsymbol{L}}(t) = \bar{\boldsymbol{l}}(t), \bar{A}(t-1) = \bar{a}(t-1)) \neq 0$

(3) Interchangeabilité conditionnelle :

$Y^{\bar{a}}(t) \perp A(t-1) \mid \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1), \bar{A}(t-2) \ t = 1, \ldots, K+1.$

L'hypothèse de positivité est généralisée de façon à ce qu'à chaque point temporel, la probabilité conditionnelle de recevoir chaque valeur de A dans chaque combinaison des covariables et de l'exposition antérieures soit supérieure à 0. De même, l'hypothèse d'interchangeabilité conditionnelle indique que les résultats potentiels sont indépendants de la valeur d'exposition précédente étant donné l'historique de covariables et d'exposition. Cela équivaut à une étude séquentiellement randomisée dans laquelle, à chaque temps t, l'exposition est attribuée de façon aléatoire avec des probabilités qui peuvent dépendre de l'exposition passée et des covariables. Conséquemment, cette hypothèse est parfois désignée sous le nom d'hypothèse de randomisation séquentielle [**4**]. En outre, nous supposons toujours aucune interférence pour compléter l'hypothèse SUTVA, c'est-à-dire qu'il existe seulement une version de l'issue contrefactuelle $Y_i^{\bar{a}}(t)$ indépendamment de l'historique de traitement d'un individu $j \neq i$. Finalement, l'ordre temporel tel que mentionné précédemment doit aussi être respecté.

1.3.9. Estimation de modèles structurels marginaux

Puisque seule une issue contrefactuelle peut être observée par individu, un MSM ne peut être estimé directement à partir des données observées. Cependant, sous les hypothèses d'identifiabilité longitudinales, même en présence de confusion dépendante du temps, les estimés provenant de modèles associationnels correspondant aux MSM peuvent être interprétés comme causaux à condition d'utiliser

des méthodes d'estimation qui tiennent compte de ce phénomène. Plusieurs méthodes existent pour estimer des MSM en présence de confusion dépendante du temps. Les plus connues sont : la pondération par l'inverse de probabilité de traitement (IPTW) [**5**, **63**], le calcul-G [**4**] et des méthodes d'estimation doublement robustes tels l'IPTW augmentée [**85**,**86**], une variation de l'IPTW augmentée utilisant une séquence de régressions imbriquées [**87**] et l'estimation par maximum de vraisemblance ciblée (TMLE) [**88**]. La TMLE est un cadre général pour la construction d'estimateurs de substitution doublement robustes et efficaces qui a aussi été développée avec succès pour les données longitudinales [**6**] et les modèles structurels marginaux [**7**, **8**].

1.3.9.1. *Estimation par inverse de probabilité de traitement*

L'une des méthodes les plus populaires, de par sa simplicité d'exécution et de compréhension, est l'IPTW. Grossièrement, l'IPTW est une méthode de pondération qui vise à recréer une population où tous les individus auraient reçu chaque régime d'exposition. Dans cette population, il n'existe pas de déséquilibres par rapport aux covariables, éliminant ainsi les effets de la confusion lors de l'estimation. De façon formelle, l'IPTW, également appelé pondération par score de propension dans la littérature, a d'abord été développé par Horvitz et Thompson [**89**] pour des études transversales dans le cadre de l'échantillonnage ; le but étant d'estimer la moyenne et le total d'une population à partir d'un échantillon aléatoire stratifié en tenant compte de différentes proportions spécifiques à chaque observation dans les strates de la population cible.

Dans le contexte de l'estimation d'un MSM pour données longitudinales, l'estimation par IPTW a été généralisée par Robins [**5**]. En pratique, cette méthode consiste à ajuster un modèle de régression correspondant au MSM aux données observées, où chaque observation est pondérée par l'inverse de la probabilité de

recevoir le régime d'exposition reçue étant donné les facteurs de confusion. En longitudinal, les MSM peuvent donc être estimés par IPTW avec des poids dépendants du temps qui sont calculés en prenant le produit de l'inverse des probabilités d'exposition à chaque temps [**5**]. Les poids variant dans le temps pour un individu i sont définis comme suit :

$$w_i(t) = \prod_{k=0}^{t} \frac{1}{P(A_i(k) \mid \bar{\boldsymbol{L}}_i(k), \bar{Y}(k), \bar{A}_i(k-1))}. \tag{1.3.7}$$

En pratique, ces poids, dits non stabilisés, peuvent avoir une variabilité très élevée. Ceci est particulièrement le cas lorsque le traitement est fortement associé avec les covariables, ce qui mène à des probabilités de recevoir l'exposition très fortes ou très faibles, pouvant causer le non-respect de l'hypothèse de positivité en pratique. Cette forte variabilité conduit à des estimateurs des paramètres causaux potentiellement biaisés et dont la variance est élevée [**90**]. Il est ainsi recommandé d'effectuer une stabilisation des poids [**5**] en introduisant un facteur de stabilisation au numérateur de l'équation (1.3.7). Le choix du numérateur ne doit pas affecter la convergence de l'estimateur de l'effet causal, mais en réduire la variabilité. Ainsi, les poids stabilisés sont préférables par rapport aux poids non-stabilisés puisque leur utilisation donne des estimateurs plus efficaces de l'effet causal [**90**]. En général, un facteur de stabilisation correspondant à la probabilité actuelle d'exposition en fonction de seulement les fonctions d'expositions antérieures utilisées dans le MSM peut être utilisé, de sorte que les poids stabilisés peuvent s'écrire :

$$sw_i(t) = \prod_{k=0}^{t} \frac{P(A_i(k) \mid \bar{Y}(k), m^*(\bar{A}, \boldsymbol{W}))}{P(A_i(k) \mid \bar{\boldsymbol{L}}_i(k), \bar{Y}(k), \bar{A}_i(k-1), \boldsymbol{W})} \tag{1.3.8}$$

où $m^*(\bar{A}, \boldsymbol{W})$ est une fonction de $\bar{A}$ et $\boldsymbol{W}$ définie par le MSM. Si l'on s'intéresse à analyser l'effet causal dans un sous-ensemble de covariables $\boldsymbol{W}$, par exemple

un modificateur d'effet, il est possible de conditionner aussi sur $\boldsymbol{W}$ tant dans le numérateur que dans le dénominateur de $sw_i(t)$. Il faut cependant noter que l'ajustement dans le numérateur ne devrait contenir que des mesures d'exposition présentes dans le MSM, car dans divers contextes où le MSM est spécifié de sorte à seulement considérer un historique d'exposition partiel, les poids stabilisés peuvent produire des inférences biaisées si le numérateur annule le contrôle de la confusion atteint par le dénominateur [**91**].

Essentiellement, la pondération crée une pseudo-population où $\boldsymbol{L}(t)$ ne prédit plus $A(t)$, au temps t, c'est-à-dire que $\boldsymbol{L}(t)$ n'est plus un facteur de confusion dans la relation exposition-issue. Sous les hypothèses d'identifiabilité, un estimateur sans biais du paramètre causal peut donc être estimé en ajustant directement le MSM aux données pondérées. Par ailleurs, afin d'assurer un estimateur IPTW sans biais, il est requis que les modèles utilisés afin d'estimer les poids soient correctement spécifiés. Des modèles paramétriques sont souvent préconisés par leur facilité d'approche. Or, il peut être préférable d'effectuer ces estimations en utilisant des techniques d'apprentissage automatique s'adaptant aux données, tel le *Super Learner* [**92**]. De plus, lors de l'utilisation de l'IPTW, les observations auront généralement des poids inégaux, ce qui introduit de la variabilité dans les données. Lorsque cela n'est pas pris en compte, la variance de l'estimateur de l'effet causal est sous-estimée. Par conséquent, lors de l'utilisation de l'IPTW pour estimer les paramètres d'un MSM, il est nécessaire d'utiliser un estimateur d'erreur standard robuste pour l'inférence [**63**].

Les estimateurs IPTW ont un certain nombre de qualités attrayantes. Entre autres, ils peuvent être intuitivement compris. Ils sont relativement faciles à mettre en oeuvre et fournissent une estimation accessible de l'erreur standard. Cependant, les estimateurs IPTW ont également des lacunes importantes. En

particulier, ils sont non convergents si le mécanisme d'exposition utilisé pour construire les poids est mal estimé. Par exemple, une mauvaise spécification due à des relations mal spécifiées peut provoquer un biais important dans le modèle de régression utilisant ces poids [**93**]. De plus, les estimateurs IPTW sont instables dans des contextes de forte confusion (quasi-violations pratiques de positivité) et le biais résultant dans les estimations ponctuelles et de variance peut entraîner de mauvaises conclusions [**79**].

1.3.10. Estimation par maximum de vraisemblance ciblée pour les paramètres d'un modèle structurel marginal

De façon générale, la TMLE est une méthode pour la construction d'estimateurs de substitution semi-paramétriques visant à constituer une amélioration par rapport à l'IPTW et à la modélisation du résultat. La TMLE combine les idées des deux principes afin de créer un estimateur doublement robuste [**88**,**94**] : convergent pour le paramètre d'intérêt si le mécanisme d'estimation de l'issue ou le mécanisme d'estimation de l'exposition sont estimés correctement. Lorsque les deux sont correctement estimés, la TMLE sera efficace, en ce sens que l'estimateur atteint la plus faible variance asymptotique parmi la classe des estimateurs réguliers asymptotiquement linéaires. Le concept clé de la TMLE est la réalisation d'une étape de mise à jour de la modélisation initiale du résultat. Cette mise à jour est réalisée grâce à un sous-modèle paramétrique construit de telle sorte que son score (dérivée de la log-vraisemblance) engendre la courbe d'influence efficace (EIC) du paramètre d'intérêt. En construisant la TMLE autour de l'EIC, la TMLE hérite ainsi de propriétés d'efficacité locale. Les avantages du TMLE ont été démontrés à maintes reprises par études de simulation et d'analyses appliquées [**79**, **95**–**97**].

Cette approche générale permet donc à la TMLE d'être appliquée à une classe très large de paramètres, y compris ceux qui ne peuvent pas être écrits comme la solution d'une équation d'estimation [**94**]. Bien que la TMLE soit un algorithme général pour une large gamme de paramètres, ce mémoire s'attarde principalement sur sa mise en oeuvre pour l'estimation des paramètres d'un MSM pour données longitudinales. Un estimateur TMLE pour les MSM de Cox utilisant une approche stratifiée a été proposé par Schnitzer et al. [**7**] basée sur la TMLE longitudinale (LTMLE) pour les effets moyens de traitements [**6**]. Cet estimateur supplante l'IPTW en termes de robustesse et d'efficacité. Cependant, dans les applications présentant de nombreux historiques d'expositions d'intérêt, la LTMLE stratifiée est vulnérable aux données rares pour certaines expositions, puisqu'elle ne cible pas directement les paramètres du MSM afin de tirer parti de l'extrapolation entre les expositions. Plus récemment, Petersen et al. [**8**], en s'appuyant sur les résultats de Robins [**85**] et Bang et Robins [**87**], ont présenté un estimateur groupé de la LTMLE (PLTMLE) pour les MSM longitudinaux, en particulier pour l'analyse de survie. Cet estimateur cible directement les paramètres du MSM et regroupe toutes les expositions d'intérêt dans l'étape de mise à jour, ce qui pourrait minimiser l'impact de données rares afin d'atteindre l'efficacité et la double robustesse.

Dans ce mémoire, nous abordons la PLTMLE pour les paramètres de MSM longitudinaux, spécifiquement pour le MSM (1.3.5).

1.3.10.1. *Reformulation du paramètre d'intérêt en termes d'espérances conditionnelles définies itérativement*

La LTMLE est un estimateur de substitution. Notons que la définition de notre paramètre d'intérêt comme une fonction qui ne dépend que des facteurs

$Q_0(\boldsymbol{O}^{\bar{a}})$ suggère immédiatement un estimateur de substitution [**98**] basé sur l'estimation de $Q_0(\boldsymbol{O}^{\bar{a}})$. Un estimateur de substitution est produit en estimant les composantes de la distribution $P_0(\boldsymbol{O}^{\bar{a}})$ par $P_n(\boldsymbol{O}^{\bar{a}})$ et en estimant $\Psi(P_0(\boldsymbol{O}^{\bar{a}}))$ par $\Psi(P_n(\boldsymbol{O}^{\bar{a}}))$. Par conséquent, l'absence de biais des estimateurs de substitution repose sur la modélisation correcte des composantes requises de la distribution des données.

Cela peut être clairement illustré en réécrivant notre paramètre d'intérêt en termes d'espérances conditionnelles définies itérativement. Cette redéfinition est très commode d'un point de vue de l'estimation ; l'estimation d'espérances conditionnelles étant, en général, beaucoup plus évidente que l'estimation de densités conditionnelles. Pour un t fixé, $t = 1, \ldots, K+1$, définissons récursivement pour $j = t, \ldots, 1$, $\bar{Q}_t^{\bar{a}}(j) = E(\bar{Q}_t^{\bar{a}}(j+1)|\bar{A}(j-1) = \bar{a}(j-1), \bar{\boldsymbol{L}}(j-1), \bar{Y}(j-1))$, où $\bar{Q}_t^{\bar{a}}(t+1) \equiv Y^{\bar{a}}(t)$. De telle sorte que, par la loi des espérances itérées, [**87**] :

$$\begin{aligned}
\bar{Q}_t^{\bar{a}}(1) &= E[\bar{Q}_t^{\bar{a}}(2) \mid \bar{A}(0) = \bar{a}(0), \bar{\boldsymbol{L}}(0)] \\
&= E[E[\bar{Q}_t^{\bar{a}}(3) \mid \bar{A}(1) = \bar{a}(1), \bar{\boldsymbol{L}}(1), \bar{Y}(1)] \mid \bar{A}(0) = \bar{a}(0), \bar{\boldsymbol{L}}(0)] \\
&= \qquad\qquad\qquad \vdots
\end{aligned}$$

et ainsi de suite jusqu'à ce que $\bar{Q}_t^{\bar{a}}(1)$ soit définie en termes de l'espérance conditionnelle la plus imbriquée, $\bar{Q}_t^{\bar{a}}(t) = E[Y^{\bar{a}}(t) \mid \bar{A}(t-1) = \bar{a}(t-1), \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1)]$.

De telle sorte que notre paramètre d'intérêt (1.3.6) est défini en fonction de ces espérances conditionnelles et peut être identifié par rapport à $\bar{Q}_t^{\bar{a}}(1)$ comme :

$$\begin{aligned}
\boldsymbol{\psi} \;=\; \text{argmin}_{\boldsymbol{\beta}}\, E \sum_{t=1}^{K+1} \sum_{\bar{a} \in \bar{\mathcal{A}}} \{ & \bar{Q}_t^{\bar{a}}(1) log(expit(\boldsymbol{\beta} m(\bar{a}, t, \boldsymbol{W}))) \;+ \\
& (1 - \bar{Q}_t^{\bar{a}}(1)) log(1 - expit(\boldsymbol{\beta} m(\bar{a}, t, \boldsymbol{W})))\}.
\end{aligned} \tag{1.3.9}$$

L'objectif de la PLTMLE est donc d'estimer $\bar{Q}_t^{\bar{a}}(1)$ par l'intermédiaire de l'estimation de chacune de ces espérances conditionnelles itérées. L'idée générale de la PLTMLE sera expliquée ci-dessous.

1.3.10.2. *Estimateur non ciblé*

Dans le but d'obtenir une première intuition de la procédure de la PLTMLE [**8**], nous présentons un estimateur non ciblé pour $\boldsymbol{\psi}$ basé sur l'estimation de régressions séquentielles pour estimer $\bar{Q}_t^{\bar{a}}(1)$.

Pour un t fixé, et pour $j = t, \ldots, 1$ soit $\bar{Q}_{t,n}^{\bar{a}}(j)$ un estimateur de $\bar{Q}_t^{\bar{a}}(j)$, c'est-à-dire $\bar{Q}_{t,n}^{\bar{a}}(j)$ est, par exemple, la régression de $\bar{Q}_{t,n}^{\bar{a}}(j)$ sur $\bar{A}(j-1) = \bar{a}(j-1)$ et $\bar{\boldsymbol{L}}(j-1)$ sur le sous-ensemble de sujets encore à risque de subir l'évènement au temps j, $\bar{Y}(j-1) = 0$. De plus, soit $\bar{Q}_{t,n}^{\bar{a}}(t+1) \equiv Y(t)$.

Pour chaque temps $t = 1, \ldots, K+1$ et pour $j = t$:

1) Un estimateur de $\bar{Q}_t^{\bar{a}}(j)$ peut être construit tout d'abord en estimant $\bar{Q}_{t,n}^{\bar{a}}(j)$ et en évaluant cette régression pour chaque individu sous chaque régime de traitement considéré, $\bar{a} \in \bar{\mathcal{A}}$.
2) Pour chaque $\bar{a} \in \bar{\mathcal{A}}$ séparément, étant donné $\bar{Q}_{t,n}^{\bar{a}}(j)$, estimons $\bar{Q}_{t,n}^{\bar{a}}(j-1)$. Ceci est exécuté en régressant $\bar{Q}_{t,n}^{\bar{a}}(j)$ sur $\bar{A}(j-2) = \bar{a}(j-2)$ et $\bar{\boldsymbol{L}}(j-2)$ étant donné les sujets encore à risque de subir l'évènement au temps $j-1$, $\bar{Y}(j-2) = 0$, et en évaluant cette régression pour chaque individu sous chaque régime de traitement considéré, $\bar{a} \in \bar{\mathcal{A}}$.
3) Pour chaque $\bar{a} \in \bar{\mathcal{A}}$ séparément, étant donné $\bar{Q}_{t,n}^{\bar{a}}(j-1)$, estimons $\bar{Q}_{t,n}^{\bar{a}}(j-2)$. Ceci est exécuté en régressant $\bar{Q}_{t,n}^{\bar{a}}(j-1)$ sur $\bar{A}(j-3) = \bar{a}(j-3)$ et $\bar{\boldsymbol{L}}(j-3)$ étant donné les sujets encore à risque de subir l'évènement au temps $j-1$, $\bar{Y}(j-3) = 0$, et en évaluant cette régression pour chaque individu

sous chaque régime de traitement considéré, $\bar{a} \in \bar{\mathcal{A}}$.

$\vdots$

Ceci est effectué itérativement jusqu'à l'obtention de $\bar{Q}^{\bar{a}}_{t,n}(1)$.

Par ce processus, on obtient un jeu de données final avec $\bar{Q}^{\bar{a}}_{t,n}(1)$ pour tout t et pour chaque $\bar{a} \in \bar{\mathcal{A}}$. L'estimateur de $\boldsymbol{\psi}$ est alors obtenu en régressant $\bar{Q}^{\bar{a}}_{t,n}(1)$ sur $m(\bar{a}, t, \boldsymbol{W})$ selon le MSM spécifié. Cette dernière étape constitue l'étape de substitution, car nous venons substituer $\bar{Q}^{\bar{a}}_{t}(1)$ par les $\bar{Q}^{\bar{a}}_{t,n}(1)$ obtenus.

Afin de faciliter la compréhension de l'estimateur non ciblé, nous présentons un exemple avec les données suivantes $\boldsymbol{O} = (\boldsymbol{L}(0), A(0), Y(1), \boldsymbol{L}(1), A(1), Y(2))$. Dans ce scénario, l'estimateur non ciblé se présenterait comme suit :

1. Commençons par exemple avec $t = 2$ et fixons $j = 2$:
   1) Estimons $\bar{Q}^{\bar{a}}_{2}(2)$ en régressant $Y(2)$ sur $\boldsymbol{L}(1)$ et $A(1)$, conduisant à l'estimation $\bar{Q}^{\bar{a}}_{2,n}(2)$ qui sera évaluée à chaque $\bar{a} \in \bar{\mathcal{A}}$. Pour une exposition binaire, $\bar{\mathcal{A}} = \{(1,1), (0,1), (1,0), (0,0)\}$. Donc, nous obtenons $\bar{Q}^{(1,1)}_{2,n}(2)$, $\bar{Q}^{(0,1)}_{2,n}(2)$, $\bar{Q}^{(1,0)}_{2,n}(2)$ et $\bar{Q}^{(0,0)}_{2,n}(2)$.
   2) L'estimateur imbriqué suivant, $\bar{Q}^{\bar{a}}_{2,n}(1)$, est obtenu en effectuant une régression distincte pour chaque $\bar{a} \in \bar{\mathcal{A}}$ de $\bar{Q}^{\bar{a}}_{2,n}(2)$ sur $\boldsymbol{L}(0)$ et $A(0) = a(0)$ et en l'évaluant pour le $\bar{a} \in \bar{\mathcal{A}}$ correspondant. Il en résulte $\bar{Q}^{\bar{a}}_{2,n}(1)$. Par exemple, pour $\bar{a} = (1,1)$, effectuer la régression de $\bar{Q}^{(1,1)}_{2,n}(2)$ sur $\boldsymbol{L}(0)$ et $A(0) = 1$ afin d'obtenir $\bar{Q}^{(1,1)}_{2,n}(1)$.
2. Pour $t = 1$, fixons $j = 1$:
   1) Estimons $\bar{Q}^{\bar{a}}_{1}(1)$ en régressant $Y(1)$ sur $\boldsymbol{L}(0), A(0)$ et en l'évaluant à chaque $\bar{a} \in \bar{\mathcal{A}}$ de telle sorte qu'il en résulte $\bar{Q}^{(1,1)}_{1,n}(1)$, $\bar{Q}^{(0,1)}_{1,n}(1)$, $\bar{Q}^{(1,0)}_{1,n}(1)$ et $\bar{Q}^{(0,0)}_{1,n}(1)$.

De telle sorte qu'un estimé de $\boldsymbol{\psi}$ peut être obtenu en effectuant une régression de $\bar{Q}^{\bar{a}}_{t,n}(1) = (\bar{Q}^{\bar{a}}_{1,n}(1), \bar{Q}^{\bar{a}}_{2,n}(1))$ sur $m(\bar{a}, t, \boldsymbol{W})$ selon les spécifications du MSM. Spécifiquement si le MSM est défini tel que :

$$\eta(\boldsymbol{\beta}, \bar{A}, t) \equiv logit[P(Y^{\bar{A}}(t+1) = 1 | Y^{\bar{A}}(t) = 0)] = \beta_0 + \beta_1 A(t) + \beta_2 t$$

on effectue la régression de $\bar{Q}^{\bar{a}}_{t,n}(1)$ sur $A(t)$ et t. Pour chaque individu, il en résulte la matrice suivante :

$$\begin{pmatrix} \bar{Q}^{(1,1)}_{2,n}(1) & 1 & 2 \\ \bar{Q}^{(0,1)}_{2,n}(1) & 1 & 2 \\ \bar{Q}^{(1,0)}_{2,n}(1) & 0 & 2 \\ \bar{Q}^{(0,0)}_{2,n}(1) & 0 & 2 \\ \bar{Q}^{(1,1)}_{1,n}(1) & 1 & 1 \\ \bar{Q}^{(0,1)}_{1,n}(1) & 0 & 1 \\ \bar{Q}^{(1,0)}_{1,n}(1) & 1 & 1 \\ \bar{Q}^{(0,0)}_{1,n}(1) & 0 & 1 \end{pmatrix}$$

où la première colonne correspond à $\bar{Q}^{\bar{a}}_{t,n}(1)$, la seconde à $a(t)$ et la dernière à t. Un exemple plus détaillé de cette étape est illustré au chapitre 3. Notons que dans cet exemple, on s'intéresse à toutes les combinaisons d'historique de traitement. En pratique, un sous-ensemble d'historiques de traitement pourrait être d'intérêt.

La principale différence entre la PLTMLE et l'estimateur non ciblé décrit ci-dessus est que les ajustements initiaux de $\bar{Q}^{\bar{a}}_{t,n}(t)$ sont mis à jour pour éliminer le biais dans une série d'étapes ciblées. C'est l'étape pour laquelle la TMLE a obtenu son nom et qui est la raison sous-jacente à leurs caractéristiques de double robustesse et d'efficacité.

1.3.10.3. *PLTMLE*

Le reste de cette section se concentre sur l'estimateur PLTMLE de $\boldsymbol{\psi}$.

Comme brièvement mentionné plus tôt, la convergence de l'estimateur IPTW repose sur l'estimation exacte du mécanisme $g_0(\boldsymbol{O})$, tandis que la convergence de l'estimateur non ciblé repose sur l'estimation exacte du mécanisme $Q_0(\boldsymbol{O})$. Dans cette section, nous abordons un estimateur semi-paramétrique efficace, la PLTMLE, qui est robuste contre la spécification incorrecte de l'un ou l'autre des mécanismes. Ces promesses théoriques reposent sur l'utilisation d'une composante importante, l'EIC pour $\boldsymbol{\psi}$.

La courbe d'influence d'un estimateur mesure à quel point une observation individuelle modifie la valeur d'un estimateur [**99**]. La courbe d'influence fournit des informations sur les propriétés de robustesse d'un estimateur et peut être utilisée pour dériver de nouveaux estimateurs avec des propriétés de robustesse prédéterminées [**99**]. Les estimateurs réguliers asymptotiquement linéaires possèdent une variance asymptotique pouvant être écrite en fonction de leur courbe d'influence. Or, il existe une fonction aléatoire unique appelée l'EIC du paramètre, c'est-à-dire la courbe d'influence avec une variance minimale. [**100**]. La variance de l'EIC s'apparente à l'information de Fisher dans un modèle paramétrique, en ce sens qu'elle définit la borne d'efficacité liée aux estimateurs réguliers de $\boldsymbol{\psi}$. Il s'agit d'une généralisation du théorème de la borne inférieure de Cramer-Rao, qui cite implicitement qu'il existe une variance minimale dans la classe des estimateurs semi-paramétriques réguliers asymptotiquement linéaires, qui ne peut être atteinte qu'à travers l'EIC [**94**]. L'EIC pour $\boldsymbol{\psi}$ peut être dérivée par la méthode delta, retrouvée dans [**7**, **8**], et est définie par :

$$
\begin{aligned}
D(\bar{Q}_t^{\bar{a}}, \bar{g}^{\bar{a}}) \quad = \quad & E\Big(\tfrac{\partial}{\partial \boldsymbol{\beta}} f\left[\eta(\boldsymbol{\beta}, \bar{a}, t)\right] expit(\eta(\boldsymbol{\beta}, \bar{a}, t))(1 - expit(\eta(\boldsymbol{\beta}, \bar{a}, t))) \tfrac{\partial}{\partial \boldsymbol{\beta}} f\left[\eta(\boldsymbol{\beta}, \bar{a}, t)\right]\Big)^{-1} \\
& \Big(\sum_{t=1}^{K+1} \sum_{\bar{a} \in \bar{\mathcal{A}}} \tfrac{\partial}{\partial \boldsymbol{\beta}} f\left[\eta(\boldsymbol{\beta}, \bar{a}, t)\right] (\bar{Q}_t^{\bar{a}}(1) - expit(\eta(\boldsymbol{\beta}, \bar{a}, t))) \\
& + \sum_{t=1}^{K+1} \sum_{j=1}^{t} \sum_{\bar{a} \in \bar{\mathcal{A}}} \frac{I(\bar{a}(j-1) = \bar{A}(j-1))}{\bar{g}^{\bar{a}}(j-1)} \tfrac{\partial}{\partial \boldsymbol{\beta}} f\left[\eta(\boldsymbol{\beta}, \bar{a}, t)\right] (\bar{Q}_t^{\bar{a}}(j+1) - \bar{Q}_t^{\bar{a}}(j))\Big).
\end{aligned}
$$

Le terme $\frac{\partial}{\partial \boldsymbol{\beta}} f[\eta(\boldsymbol{\beta}, \bar{a}, t)]$ est la dérivée du MSM par rapport à $\boldsymbol{\beta}$. Le terme $I(\bar{A}(j-1) = \bar{a}(j-1))$ est une fonction indicatrice égale à 1 si l'historique d'exposition observée au temps $j-1$ est égal au régime d'exposition $\bar{a}(j-1)$. Le terme $\bar{g}^{\bar{a}}(j) = \prod_{k=0}^{j} P(A(k) = a(k) | \bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}(k-1))$ représente la probabilité de recevoir l'exposition $\bar{A}(j) = \bar{a}(j)$ étant donné l'historique de covariables $\bar{\boldsymbol{L}}(j)$ et un historique d'exposition $\bar{A}(j-1) = \bar{a}(j-1)$.

Rappelons que la mise à jour de l'estimation initiale de $Q_0(\boldsymbol{O})$ vers une estimation ciblée du paramètre d'intérêt est l'objectif principal de la TMLE. Pour ce faire, il faut veiller à ce que les estimés de $\bar{Q}_t^{\bar{a}}(j)$ résolvent l'équation de l'EIC pour le paramètre d'intérêt. La théorie qui établit ce résultat est présenté dans [**88**] et [**94**]. En général, le processus de mise à jour repose sur l'utilisation d'une fonction de perte $L(\bar{Q}_t^{\bar{a}}(t))$ et d'un sous-modèle paramétrique $\bar{Q}_{t,n}^{\bar{a}}(t, \boldsymbol{\epsilon})$, choisi afin que l'étape de mise à jour présente le plus grand gain pour l'estimation sans biais du paramètre d'intérêt. Cela est fait en choisissant $\bar{Q}_{t,n}^{\bar{a}}(t, \boldsymbol{\epsilon})$ de sorte que $\frac{\partial}{\partial \boldsymbol{\epsilon}} L(\bar{Q}_{t,n}^{\bar{a}}(t, \boldsymbol{\epsilon}))\Big|_{\boldsymbol{\epsilon}=0}$ engendre l'EIC. Le sous-modèle est :

$$
logit(\bar{Q}_{t,n}^{\bar{a}}(t, \boldsymbol{\epsilon})) = logit(\bar{Q}_{t,n}^{\bar{a}}(t)) + \boldsymbol{\epsilon} \frac{I(\bar{A}(t-1) = \bar{a}(t-1))}{\bar{g}_n^{\bar{a}}(t-1)} \frac{\partial}{\partial \boldsymbol{\beta}} f\left[\eta(\boldsymbol{\beta}, \bar{A}, t)\right]\Big|_{\bar{A}(t-1) = \bar{a}(t-1)} \tag{1.3.10}
$$

où $\bar{g}_n^{\bar{a}}(t-1)$ est l'estimation des composantes de $g_0(\boldsymbol{O})$ de $P_0(\boldsymbol{O})$ évaluée à $\bar{A}(t-1) = \bar{a}(t-1)$, définie par :

$$
\bar{g}^{\bar{a}}(t) = \prod_{k=0}^{t} P(A(k) = a(k) | \bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}(k-1)).
$$

D'un point de vue procédural, la PLTMLE utilise le même algorithme de régression séquentielle décrit dans l'estimateur non ciblé pour estimer les ajustements initiaux de $Q_0(\boldsymbol{O})$. Cependant, ces estimations initiales sont séquentiellement mises à jour pour supprimer les biais dans une série d'étapes ciblées s'appuyant sur l'estimation de $g_0(\boldsymbol{O})$ par $\bar{g}_n^{\bar{a}}(t)$. Ces étapes impliquent le sous-modèle (1.3.10) ci-dessus dont le score engendre l'EIC.

Un algorithme plus détaillé de la procédure PLTMLE est décrit exhaustivement dans le chapitre 3 et ne sera pas répété ici à des fins de concision.

1.3.11. Note sur l'estimation des modèles pour $Q_0(\boldsymbol{O})$ et $g_0(\boldsymbol{O})$

Tout comme pour l'IPTW ou l'estimateur non ciblé, afin de tirer parti des propriétés de la TMLE, les modèles pour l'estimation de $Q_0(\boldsymbol{O})$ et $g_0(\boldsymbol{O})$ doivent être correctement spécifiés. La TMLE ne dicte pas l'utilisation d'une méthode d'estimation particulière, bien qu'il y ait des gains clairs dans la performance en échantillons finis si les modèles sont correctement spécifiés. La capacité à intégrer, sous certaines conditions [**101**, **102**], plusieurs méthodes d'estimation différentes (paramétriques, semi-paramétriques, apprentissage automatique, etc.) tout en conservant une inférence valable est une motivation supplémentaire pour la TMLE.

1.3.12. Note sur les méthodes d'inférence causale pour les données longitudinales en temps continu

Bien que le problème de discrétisation de la ligne du temps pour les méthodes longitudinales d'inférence causale soit reconnu [**8**, **57**, **58**], il faut noter que le mécanisme de l'exposition est encore conventionnellement modélisé en

temps discret, quoique l'historique d'exposition sous-jacent puisse être en temps continu. Quelques méthodes proposent la modélisation d'effets causaux en temps continu [**58**, **103**, **104**], mais n'ont pas gagné en popularité vu leur complexité d'exécution. En particulier, l'une d'entre elles propose une modélisation paramétrique des poids d'inverse de probabilité de traitement en temps continu. Or, cette méthode ne s'applique pour l'instant qu'à des données avec processus de comptage [**58**]. Ainsi, à notre connaissance, aucune méthode dans la littérature ne se penche sur le sujet de la discrétisation arbitraire de la ligne du temps. Le corps de ce mémoire, par l'entremise d'un manuscrit, y sera dévoué.

Chapitre 2

OBJECTIFS

L'objectif principal de ce mémoire est de développer une procédure sélectionnant automatiquement la discrétisation optimale de la ligne du temps. Cette procédure a pour but d'être intégrée à des méthodes longitudinales d'inférence causale. Un sous-objectif est par conséquent d'évaluer la performance de cette procédure de sélection.

Un objectif secondaire passe par spécifier les problèmes liés à l'estimation d'effets causaux en présence de discrétisation. Nous chercherons donc à présenter un cadre régissant l'estimation d'effets causaux en présence de discrétisation.

Finalement, un objectif tertiaire se veut d'appliquer la méthodologie développée à un contexte de données réelles. Cet objectif cherche à évaluer l'effet de la prise de faibles doses d'ICS pendant la grossesse sur la durée de gestation de femmes ayant un asthme léger. Ceci est réalisé en employant des méthodes longitudinales d'inférence causale dans un contexte où l'on sait qu'il existe des variables de confusion qui varient dans le temps.

Chapitre 3

MANUSCRIT

Les résultats des travaux de recherche réalisés dans le cadre de ce mémoire sont présentés dans l'article suivant :

A data-adaptive procedure for the optimal discretization of the timeline for longitudinal causal inference methods

Steve Ferreira Guerra[a], Lucie Blais[a,b], Amélie Forget[a,b] and Mireille E. Schnitzer[a]

[a] *Faculté de Pharmacie, Université de Montréal*

[b] *Centre de recherche, Hôpital du Sacré-Coeur de Montréal, Montréal, Québec, Canada*

Le présent article fût soumis pour révision à la revue *Statistics in Medicine* le 18 juillet 2017. Une déclaration d'accord de divulgation de cet article dans ce mémoire a préalablement été obtenue de tous les coauteurs.

Abstract

In health care research, administrative databases are being increasingly used to conduct analyses evaluating drug safety and effectiveness. In longitudinal settings, causal inference methods usually rely on a discretization of the patient

timeline that may not reflect the underlying continuous nature of administrative data and bias may result when coarsening is arbitrarily chosen by the researcher. This is due to the data coarsening resulting in both a modified definition of the parameter of interest and discarded information about time-dependent confounders. This paper investigates the estimation of causal parameters under discretized data. It presents the implicit assumptions practitioners make but do not acknowledge when discretizing data to assess various longitudinal causal parameters. We further propose a cross-validation procedure to select a timeline discretization for use with pooled Longitudinal Targeted Maximum Likelihood Estimation for the estimation of the parameters of a marginal structural model. We use a simulation study to evaluate the bias-variance trade-off of this method. We then apply our approach to a study on the relative effect of alternative asthma treatments during pregnancy on pregnancy duration. The results of the simulation study illustrate how coarsening changes the target parameter of interest as well as how it creates bias due to a lack of control for time-dependent confounders. We also demonstrate the performance of our selection procedure, observing how it converges to an unbiased estimate of the parameter of interest at the finest discretization as the sample size increases.

keywords : administrative databases ; coarsening ; TMLE ; semi-parametric estimation ; machine learning

3.1. Introduction

In many fields, such as epidemiology, public health, and pharmacoepidemiology, causal inference has become central when investigating questions about the effect of exposures in observational studies [**1**, **105**]. Due to their broad potential for analysis (e.g. relatively low cost, large size, longitudinal nature and long

follow-up period), one particular source of data that has rapidly and abundantly become a powerful tool for observational research is administrative data [**37**,**106**]. In spite of these benefits, such databases were not intended to be used for research as the data is mainly collected for administrative purposes. Hence, they often lack an appropriate structure for usage in research, which leads to inherent problems including poor data quality, absence of information on confounders, and comparability of data [**43**]. Furthermore, the considerable complexity of the data generates many additional statistical challenges involving confounding control, measurement error, and missing data [**107**, **108**].

One particular statistical problem derives from the fact that administrative data depicts real-world data. Indeed, in this type of data, the times at which exposures may change are not controlled by the study design, leading to an underlying exposure process that changes in continuous time. However, most existing methods for longitudinal causal inference assume that exposure changes only at common discrete time-points and therefore rely on a discretization of the timeline [**103**]. Consequently, an analytical issue arises when the coarsening is left to the possibly arbitrary choice of the researcher [**57**].

Arbitrary discretization can become problematic because the chosen discretization scale for the analysis can affect the definition of the target parameter and the assumptions required for estimating causal effects from longitudinal observational data. Hence, as for conditional parametric models [**109**], the Marginal Structural Model (MSM) parameters [**110**] may change with different discretizations. In addition, in the presence of time-dependent confounding, if one chooses a scale that is not fine enough to capture the intricate relations between exposure, time-dependent confounders, and outcome, the estimates may be biased. On the

other hand, an excessively narrow scale could lead to data sparsity at the individual time-points, potentially resulting in bias and inflated variance [**90**,**110**,**111**]. This leaves open the question of how to select an appropriate discretization that would balance this trade-off between increased bias and inflated variance.

In the presence of time-dependent confounding, traditional methods of adjusting for confounders fail to produce unbiased estimation of causal effects [**4**]. Several methods have been developed to estimate various causal parameters for longitudinal data [**4**, **110**–**113**]. Targeted Maximum Likelihood Estimation (TMLE) [**114**–**116**] is one such framework that constructs double-robust semiparametric estimators of causal parameters. Following work on double-robust estimation [**87**], van der Laan and Gruber [**117**] developed Longitudinal TMLE (LTMLE) based on the estimation of iterated conditional expectations of the outcome [**87**]. Subsequently, Petersen *et al.* [**8**] developed a pooled version of LTMLE for the estimation of the coefficients of a MSM [**81**, **118**], that has been shown to perform better than alternative implementations of LTMLE and the more commonly used Inverse of Probability of Treatment Weighted (IPTW) estimator [**7**, **8**, **110**, **111**]. Indeed, in part because of its double robustness property, LTMLE has been shown [**119**] to be less biased and result in smaller estimation variance than IPTW methods, especially in cases where data sparsity occurs.

In this paper, we propose a time-discretization selection procedure integrated with pooled LTMLE. The proposed method consists of a procedure that will sequentially analyze different discretizations of a certain dataset and then automatically select the "optimal" discretization. This optimal discretization is selected based on the minimization of a pooled LTMLE loss function that is evaluated through cross-validation [**120**]. This loss function is specifically aimed at balancing the bias-variance trade-off for estimation of the parameter of interest.

We then evaluate our approach through a simulation study and provide empirical evidence that the selected discretization converges to an unbiased estimate of the parameter of interest. Finally, we apply our selection procedure to an example of the evaluation of asthma treatment on pregnancy duration.

More specifically, in section 3.2, we will present a motivating example on the effect of asthma medication on gestational age. In section 3.3, we describe the general data structure and introduce the concept of a discretized dataset. In section 3.4, we introduce our causal parameter of interest. In section 3.5, we explore how discretization may affect the target parameter and its identifiability assumptions. In section 3.6, we present the pooled LTMLE algorithm and our corresponding selection procedure. Sections 3.7 and 3.8 refer, respectively, to the simulation study and the real data application, and accompanying results. Finally, we discuss possible limitations and extensions of the proposed methodology.

3.2. Motivating example

The proposed method was motivated by a study on the evaluation of the safety of asthma controller medications on pregnancy outcomes [**121**]. In the initial study, data on pregnant women with asthma with deliveries between 1998 and 2008 were extracted from the linkage of the RAMQ and MED-ECHO administrative databases in the province of Quebec, Canada. Information on pregnancy outcomes, exposure to asthma medication, and related confounders were assessed from prescription renewals in community pharmacies, hospitalizations, emergency room visits, and outpatient medical consultations. For additional information on the data and these administrative databases, refer to the Table 1 and the supplementary materials of [**121**]. Although asthma control is a time-dependent confounder of the longitudinal exposure to controller medication, initial analyses

failed to consider the time-dependent nature of the data. Indeed, asthma control affects subsequent asthma treatment and pregnancy outcomes while also being affected by previous controller medication [**16**]. Thus proper causal inference methods that allow for time-dependent confounding should be used [**110**].

It is known that uncontrolled asthma is associated with adverse effects for the fetus and that advantages of adequate control outweigh any potential risks of asthma medication and therefore that asthma should continue to be controlled with adequate medications during pregnancy [**16**, **20**]. However, it has been shown that about 50% of women tend to lower their controller medication during pregnancy [**28**], potentially due to fear of adverse medication effects. This is specifically the case for women with mild asthma, for which treatment may alternate between low daily doses of Inhaled Corticosteroids (ICS) or no controller medication [**16**,**122**]. The published literature tends to identify that there are no adverse effects of low ICS dosage on pregnancy outcomes. However, there are currently no data on the effectiveness and safety of asthma medications from studies considering the time-dependent confounded nature of asthma control. The proposed analysis is therefore aimed at estimating the longitudinal causal effect of ICS on time until delivery using a MSM to evaluate the short-term relative effect of low ICS dose versus no ICS dose on time to delivery in women with mild asthma.

Since it is known that the relationship between asthma control and treatment happens at a finer scale than trimester [**16**], interest lies in performing a time-dependent extraction of the administrative health data. While a finer scale would supposedly ensure adequate control for time-dependent confounding, limitations in the refinement of the data exist because of the nature of the methods used to control for said time-dependent confounding. We are then faced with the dilemma

on how to choose an appropriate discretization given the available data.

3.3. Data structure

In the above described example, consider a longitudinal data structure where, for every individual, we observe the following data

$$\boldsymbol{O} = (\boldsymbol{L}(0),\ A(0),\ Y(1),\ \boldsymbol{L}(1),\ A(1),\ \ldots,\ Y(K),\ \boldsymbol{L}(K),\ A(K),\ Y(K+1))\ .$$

Let t index the discrete times at which time-dependent variables are observed, $t = 0, \ldots, K+1$. Let $Y(t)$ denote a time-dependent outcome, $A(t)$ a time-dependent exposure, and $\boldsymbol{L}(t)$ a vector of time-dependent covariates, all measured at time-point t. In particular, let $\boldsymbol{L}(0)$ be the baseline covariates measured at the beginning of the study, and $Y(K+1)$ be the final outcome assessed at the end of the study after all exposures and covariates. We consider a binary exposure where $A(t) = 1$ indicates that a person was exposed at time-point t. Let the overbar represent a variable's history, such that, for example, $\bar{A}(t) = (A(0), A(1), \ldots, A(t))$ denotes an individual's exposure history up until time-point t. Examples of fixed regimes (i.e. possible values at which $\bar{A}(t)$ may be set) include "always exposed", $\bar{a}(t) = (1, 1, \ldots, 1)$, a regime where the subject is exposed to treatment at every time-point, and its counterpart, "never exposed", $\bar{a}(t) = (0, 0, \ldots, 0)$, where the subject is unexposed to treatment at every time-point. Furthermore, define $\bar{\mathcal{A}}$ as the set of all possible regimes.

3.3.1. Discretization

Formally, let us define the timeline $\mathcal{I} = [0, \tau]$ on the real line, where τ corresponds with the maximum follow-up period of a longitudinal study at which the last outcome is assessed. A partition $\mathcal{P}$ of $\mathcal{I}$ is a finite collection of points $t_0, t_1, \ldots, t_K, t_{K+1}$ such that the union formed by the disjoint intervals $J_k =$

$[t_k, t_{k+1}[$ is $\mathcal{I}$, where the t_k can be ordered such that $0 = t_0 < t_1 < \ldots < t_K < t_{K+1} = \tau$. It follows that each different partition $\mathcal{P}_r$, $r = r_0, r_1, \ldots, r_R$, of our timeline corresponds to a different discretized observed dataset O_r, where T_r is the set of time-points in $\mathcal{P}_r$ and R is the number of possible different discretizations.

In particular, let us denote $\mathcal{P}_{r_0}$ as the finest partition into which our timeline can be divided in practice, corresponding to the dataset O_{r_0} and time-points T_{r_0}. This finest discretized data structure is equal to the finest possible scale at which changes can be observed. In health administrative databases, this finest partition usually consists of the set of all days during the follow-up period. Any other partition $\mathcal{P}_r$ that does not include all of these points is said to be coarser than $\mathcal{P}_{r_0}$. Examples of coarser partitions could be a set of time-points where the intervals between time-points are of 1 week. Any O_r can then be defined from O_{r_0} as a function of its partition : $O_r := \Theta(\mathcal{P}_r)(O_{r_0})$. Figure 3.1 illustrates three nested discretized datasets, representing respectively a timeline where we consider the finest discretized data, a timeline where we only consider every other time-point, and a timeline where we consider every fourth time-point. Inversely, a discretization formed by partition $\mathcal{P}_{r_1}$ is finer than another discretization $\mathcal{P}_{r_2}$ if it includes all the points in $\mathcal{P}_{r_2}$ and at least one other point in I. $\mathcal{P}_{r_1}$ is then said a refinement of $\mathcal{P}_{r_2}$. It follows that there are no possible refinements of $\mathcal{P}_{r_0}$. Note that a partition need not only include points that form equally spaced intervals. Hence, in practice, from one undiscretized dataset one can obtain various different discretized datasets for the analysis. Consequently, let $\bar{a}_r(t)$ define an exposure regime up to time t on the discretized data O_r, with $\bar{a}_r(\tau) \equiv \bar{a}_r$ of same length as T_r. Correspondingly, $\bar{\mathcal{A}}_r$ can be defined as the set of all possible regimes on T_r, where $\bar{\mathcal{A}}_r \subseteq \bar{\mathcal{A}}_{r_0}$ (with some abuse of notation). From here on, with some abuse of notation, let us renumber the time-points T_r in any given partition $\mathcal{P}_r$

as $T_r = \{0, 1, 2, \ldots, K_r, K_r + 1\}$.

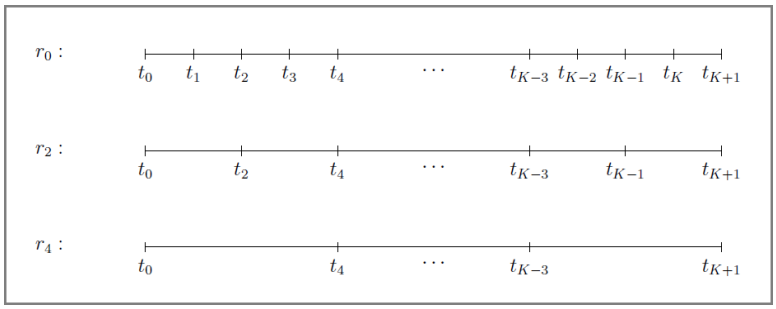

FIGURE 3.1. Illustration of the finest discretized timeline, and two coarser discretizations of the finest timeline

3.4. PARAMETER OF INTEREST

Following the Neyman-Rubin counterfactual framework [**2**, **71**, **123**], causal effects may be described using counterfactual outcomes. Hence, for a given discretization indexed by r, let $Y^{\bar{a}_r}(t)$ be a random variable representing a subject's counterfactual outcome at time t had he followed the exposure history $(\bar{A}(t-1), t \in T_r) = \bar{a}_r(t-1)$, where variables indexed by negative values and $Y(0)$ should be taken as the null set. $E[Y^{\bar{a}_r}(t)]$ designates the mean outcome $Y(t)$ had the exposure history been set to a specific $(\bar{A}(t-1), t \in T_r) = \bar{a}_r(t-1)$ for every subject in the population of interest. In such longitudinal settings, it is often of interest to describe the expectation of the time-dependent counterfactual outcome as a function of exposure history, time, and possibly a subset of baseline characteristics, $\boldsymbol{W} \subseteq \boldsymbol{L}(0)$. As in Robins *et al.* [**110**], this can be accomplished using a MSM that models the expected counterfactual outcome $Y^{\bar{A}}(t)$ under observed treatment regime $\bar{A}(t-1), t \in T_r$ as a function of desired components [**8**] :

$$E[Y^{\bar{A}}(t)|t, \boldsymbol{W}] = \eta(\boldsymbol{\beta}, \bar{A}, t, \boldsymbol{W}) = f^{-1}\left[\sum_{j=1}^{J} \beta_j m_j(\bar{A}, t, \boldsymbol{W})\right], t \in T_r.$$

Our vector causal parameter of interest, for a given discretization r, $\boldsymbol{\psi}_r$, is then defined as the coefficients of the MSM that satisfy :

$$\boldsymbol{\psi}_r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, E \sum_{t \in T_r} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \mathcal{L}(\eta(\boldsymbol{\beta}, \bar{a}_r, t, \boldsymbol{W}), Y^{\bar{a}_r}(t))$$

for a loss function $\mathcal{L}(\cdot)$ that allows the identification of the causal parameter of interest. Specifically, we require that $E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}(\eta(\boldsymbol{\beta}, \bar{a}_r, t, \boldsymbol{W}), Y^{\bar{a}_r}(t))\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0}\right] = 0$ at the true $\boldsymbol{\beta}^0$.

Like in our motivating example, when the outcome variable $Y(t)$ is a binary indicator of a failure event by time t, we may be interested in modeling the counterfactual probability of the event by time t given survival up until time $t-1$, $P(Y^{\bar{a}_r}(t)|Y^{\bar{a}_r}(t-1))$, as a function of most recent exposure to a certain medication. In this case, one could define the following MSM, evaluating the effect of most recent exposure on the probability of an event :

$$\eta(\boldsymbol{\beta}, \bar{A}, t) \equiv logit[P(Y^{\bar{A}}(t) = 1 | Y^{\bar{A}}(t-1) = 0), t] = \beta_0 + \beta_1 A(t-1) + \beta_2 t,\ t \in T_r\ . \quad (3.4.1)$$

Accordingly, our target parameter could be defined using the logistic loss function as :

$$\boldsymbol{\psi}_r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, E \sum_{t \in T_r} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \{Y^{\bar{a}_r}(t) log(expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t))) + (1 - Y^{\bar{a}_r}(t)) log(1 - expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t)))\}\ . \quad (3.4.2)$$

3.4.1. Note on data-adaptive parameters

The above MSM is defined on a fixed (arbitrary) discretization. In order to define a data-adaptive causal parameter [**124, 125**] over a set of alternative discretizations, consider the hypothetical experiment where exposure is set to some

fixed regime $\bar{a}_r(t)$ at time-points in T_r. The parameter of interest is defined on the resulting counterfactual data :

$$\boldsymbol{O}_r^{\bar{a}_r} = (Y^{\bar{a}_r}(t), \boldsymbol{L}^{\bar{a}_r}(t); t \in T_r)$$

where $\boldsymbol{L}^{\bar{a}_r}(t)$ is the counterfactual covariate value at time t had past exposures been set to $\bar{a}_r(t-1)$.

This counterfactual data consists of n independent, identically distributed observations from a true distribution $Q_0(\boldsymbol{O}_r^{\bar{a}_r})$, which can be decomposed according to the time-dependent distribution of the data as :

$$Q_0(\boldsymbol{O}_r^{\bar{a}_r}) = \underbrace{\prod_{T_r \setminus \{0\}} P_0(Y^{\bar{a}_r}(t) | \bar{\boldsymbol{L}}^{\bar{a}_r}(t), \bar{Y}^{\bar{a}_r}(t-1))}_{Q_{0Y}(\boldsymbol{O}_r^{\bar{a}_r})} \underbrace{\prod_{T_r \setminus \{K_r+1\}} P_0(\boldsymbol{L}^{\bar{a}_r}(t) | \bar{Y}^{\bar{a}_r}(t), \bar{\boldsymbol{L}}^{\bar{a}_r}(t-1))}_{Q_{0\boldsymbol{L}}(\boldsymbol{O}_r^{\bar{a}_r})}$$

Here we suppose that $Q_0(\boldsymbol{O}_r^{\bar{a}_r})$ is a member of a statistical model space $\mathcal{M}$, that can be decomposed into $\mathcal{Q}_Y$ and $\mathcal{Q}_{\boldsymbol{L}}$, the set of all possible values of Q_{0Y} and $Q_{0\boldsymbol{L}}$, respectively. For any given r, let $Q_0(\boldsymbol{O}_r^{\bar{a}_r})$ be the true underlying distribution of $\boldsymbol{O}_r^{\bar{a}_r}$. For every coarsened dataset, our causal parameter of interest can be defined as a mapping

$$\boldsymbol{\psi}_r \equiv \boldsymbol{\Psi}(Q_0(\boldsymbol{O}_r^{\bar{a}_r})) : \mathcal{M} \to \mathbb{R}^n$$

for some function $\boldsymbol{\Psi}$ that takes as argument a member from the statistical model space $\mathcal{M}$ into the parameter space $\mathbb{R}^n$.

As a result, we see that the true value of the parameter of interest $\boldsymbol{\psi}_r$ directly depends on the data discretization and consequently may change as the underlying data changes. These parameters may differ with the underlying data because the parameter is interpreted as if exposure were randomized at select

time-points $t \in T_r$. Since the target parameter mapping depends on the data it can be called data-adaptive [**124**, **125**].

3.5. Causal assumptions

To obtain consistent estimates of causal effects from longitudinal observational data one must assume some identifiability conditions. We further review these assumptions.

3.5.1. Causal assumptions for longitudinal data

A necessary assumption for causal inference is the so called no unmeasured confounders assumption [**70**, **77**]. Formally, for longitudinal time-varying exposures, on an arbitrary discretization indexed as $t = 0, 1, \ldots, K+1$, we assume the (weak) sequential randomization assumption (SRA) [**126**] :

$$Y^{\bar{a}}(t) \coprod A(t-1) | \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1), \bar{A}(t-2), t = 1, ..., K+1.$$

This means that at each time t, $Y^{\bar{a}}(t)$ and $A(t-1)$ are independent given past covariates $\bar{\boldsymbol{L}}(t-1)$ and past exposure history. This assumption can be thought of as in a sequential randomized trial, where at each follow-up time t exposure is randomly assigned conditional on the observed history. Here, it is assumed that the measured $\bar{\boldsymbol{L}}(t-1)$ is a sufficient set of confounders such that the SRA holds.

We further assume positivity [**77**, **87**] which requires that for every combination of the values of the confounders $\bar{\boldsymbol{L}}(t)$ and exposure history for which $P(\bar{\boldsymbol{L}}(t) = \bar{\boldsymbol{l}}(t), \bar{A}(t-1) = \bar{a}(t-1)) > 0$, at each time-point every individual has a non-null probability of receiving all values of exposure :

$$P(A(t) = a(t)|\bar{\boldsymbol{L}}(t), \bar{Y}(t), \bar{A}(t-1)) > 0, \forall a(t), t = 0, \ldots, K.$$

Even if the positivity assumption holds in theory, practical positivity violations may occur, perhaps due to data sparsity, in which case the estimated probabilities of receiving exposure approach zero. Hence, for estimation, the positivity assumption must also hold in practice.

Furthermore, in defining the counterfactual outcomes, we assume the no interference assumption [**72**] which indicates that an individual's counterfactual variable's value is not affected by another individual's exposure. And finally, we assume consistency : $Y^{\bar{a}}(t) = Y(t)$ and $\boldsymbol{L}^{\bar{a}}(t) = \boldsymbol{L}(t)$ when $\bar{A}(t-1) = \bar{a}(t-1)$ is observed [**71**, **73**]. Namely, if the past exposure of a certain subject is equal to $\bar{a}(t-1)$ the consistency assumption assures us that the potential outcome $Y^{\bar{a}}(t)$ is equal to the observed outcome $Y(t)$. This assumption also implies that the levels of exposure have to correspond to well defined interventions [**127**].

Below we will examine some of the implications of discretization on the plausibility of these causal assumptions.

3.5.2. Causal assumptions for discretized data

Excessively coarse discretizations may fail to capture sufficient relations between the variables to remove time-dependent confounding bias. Thus by imposing a discretization, we are potentially wrongfully assuming that the SRA still holds. Suppose we assume that the SRA holds at the finest discretization r_0 :

$$Y^{\bar{a}_{r_0}}(t) \coprod A(t-1)|\bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1), \bar{A}(t-2), t \in T_{r_0}.$$

As illustrated in the Directed Acyclic Graph (DAG) in Figure 3.2, depicting the causal relations between exposure $A(t)$, covariate $L(t)$, and outcome $Y(t+1)$ at three time-points, failing to choose a scale fine enough that would capture the scope at which these relations happen could result in poor adjustment for time-dependent confounders and subsequent bias. One can see that when we remove information about $A(1)$, $L(1)$, and $Y(1)$, an unmeasured confounder $U(1)$ is created for the relationship between $A(2)$ and $Y(3)$. By failing to adjust for $U(1)$, confounding bias may be introduced when estimating the effect of $A(2)$ on $Y(3)$.

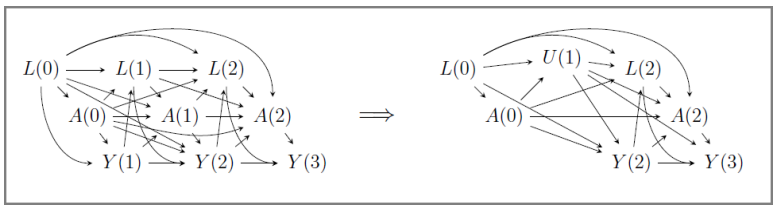

FIGURE 3.2. DAG illustrating the result of discretization on time-dependent confounding

In some very specific settings, there may exist some coarser discretization, $r^{\dagger}$, where the SRA still holds, i.e. :

$$Y^{\bar{a}_{r^{\dagger}}}(t) \coprod A(t-1) | \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1), \bar{A}(t-2), t \in T_{r^{\dagger}}.$$

For example, Figure 3.3 demonstrates a case where the SRA would hold when calculating the effect of most recent exposure, $A(2)$, on subsequent outcome, $Y(3)$, even in the presence of $U(1)$.

The positivity assumption must also hold for every discretization r, at each time-point $t \in T_r$:

$$P(A(t) = a(t) | \bar{\boldsymbol{L}}(t), \bar{Y}(t), \bar{A}(t-1)) > 0, \forall a(t), t \in T_r.$$

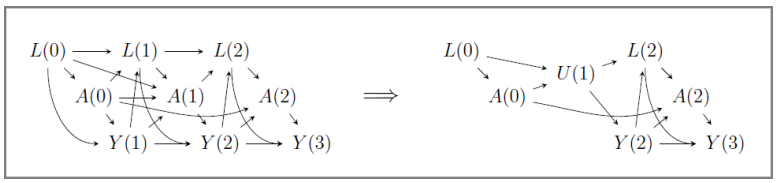

FIGURE 3.3. DAG illustrating the result of discretization on time-dependent confounding

Because the set of discretized regimes is included in the set of regimes at the finest scale, $\bar{A}_r \subseteq \bar{A}_{r_0}$, if positivity holds at the finest scale, then it is plausible that it would hold for any coarser discretization conditional on the true covariate history. However, discretization affects the practical positivity assumption in that, as the discretization gets coarser, the practical positivity assumption may be relaxed by only requiring that the estimated probabilities be larger than zero at fewer observed discretized times.

For consistency, note that the counterfactual intervention is defined relative to the chosen discretization. Consistency must therefore be interpreted as whether the hypothetical experiment, specific to the discretization, is represented by the observed data. Suppose that consistency holds at the finest discretization and consider a coarsening that removes time-point $t-1$. Coarsening will violate consistency when

$$Y^{\bar{a}_r(t-2),a(t-1),a(t)}(t) \neq Y^{\bar{a}_r(t-2),1-a(t-1),a(t)}(t)$$

since the removal of the time-point $t-1$ would imply multiple versions of $Y^{\bar{a}_r(t-2),a(t)}(t)$.

Finally, it is logical to assume that if no interference holds at T_{r_0} it is sustained at all T_r. If one subject's exposure does not affect another subject's potential

outcome under the finest discretization, then it will also not affect the potential outcome under a coarsened discretization.

3.6. Methods

Various implementations of Longitudinal TMLE have been developed for different causal quantities [**7**,**8**,**117**]. In this section, we focus on describing the steps of the pooled LTMLE algorithm developed by Petersen et. al [**8**] to estimate the parameters of a MSM. In section 3.6.2, we further introduce a selection procedure for the optimal discretization of the timeline integrated with pooled LTMLE.

3.6.1. Pooled LTMLE

For the estimation of the pooled LTMLE, let us define the working probability distribution of the observed discretized data $\boldsymbol{O}_r$ as :

$$P_0(\boldsymbol{O}_r) = \underbrace{\underbrace{\prod_{t=1}^{K_r+1} P_0(Y(t)|\bar{A}(t-1),\bar{\boldsymbol{L}}(t-1),\bar{Y}(t-1))}_{Q_{0Y}(\boldsymbol{O}_r)} \underbrace{\prod_{t=0}^{K_r} P_0(\boldsymbol{L}(t)|\bar{Y}(t),\bar{A}(t-1),\bar{\boldsymbol{L}}(t-1))}_{Q_{0\boldsymbol{L}}(\boldsymbol{O}_r)}}_{Q_0(\boldsymbol{O}_r)} \underbrace{\prod_{t=0}^{K_r} P_0(A(t)|\bar{\boldsymbol{L}}(t),\bar{Y}(t),\bar{A}(t-1))}_{g_0(\boldsymbol{O}_r)}$$

where $P_0(\boldsymbol{O}_r)$ is the distribution function of $\boldsymbol{O}_r$ which can be decomposed into Q_0, a mechanism for the outcome, and g_0 a mechanism for the exposure. Beforehand, in order to explain the algorithm, let us define the conditional expected value of the outcome at time t given past covariate history and a fixed exposure history $\bar{a}_r(t-1)$ as $\bar{Q}_t^{\bar{a}_r}(t) = E(Y^{\bar{a}_r}(t)|\bar{A}(t-1) = \bar{a}_r(t-1), \bar{\boldsymbol{L}}(t-1), \bar{Y}(t-1))$. Recursively, for $j = t, ..., 1$, $\bar{Q}_t^{\bar{a}_r}(j) = E(\bar{Q}_t^{\bar{a}_r}(j+1)|\bar{A}(j-1) = \bar{a}_r(j-1), \bar{\boldsymbol{L}}(j-1), \bar{Y}(j-1))$ where $\bar{Q}_t^{\bar{a}_r}(t+1) := Y^{\bar{a}_r}(t)$. Under the described causal assumptions, our causal parameter of interest can be specified with respect to these iterated conditional

expectations [**8**, **87**] as

$$\begin{aligned}\boldsymbol{\psi}_r &= \text{argmin}_{\boldsymbol{\beta}} E \sum_{t=1}^{K_r+1} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \{\bar{Q}_t^{\bar{a}_r}(1) log(expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t, \boldsymbol{W}))) \; + \\ &\qquad (1 - \bar{Q}_t^{\bar{a}_r}(1)) log(1 - expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t, \boldsymbol{W})))\}.\end{aligned}$$

$\boldsymbol{\psi}_r$ is a plug-in estimator and can be estimated by regressing an estimate of $\bar{Q}_t^{\bar{a}_r}(1)$ on $\bar{a}_r, t,$ and $\boldsymbol{W}$ according to our MSM. $\bar{Q}_t^{\bar{a}_r}(j)$ is sometimes conveniently referred to as the outcome model.

Likewise, let us define for every $j = t, \ldots, 1$, the estimation of the $g_0(\boldsymbol{O}_r)$ component of $P_0(\boldsymbol{O}_r)$ as $\bar{g}^{\bar{a}_r}(j) = \prod_{k=0}^{j} P(A(k) = a(k) | \bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}_r(k-1))$ which can be designated as the exposure model and represents the probability of receiving exposure $\bar{A}(j) = \bar{a}_r(j)$ given covariate history $\bar{\boldsymbol{L}}(j)$ and a fixed exposure history $\bar{A}(j-1) = \bar{a}_r(j-1)$.

Generally, for each time t, implementation of the pooled LTMLE involves the calculation of an initial estimate $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ and the update of this initial estimate to $\bar{Q}_{t,n}^{\bar{a}_r,*}(j)$ via a specific submodel defined as a function of the exposure probabilities. This is recursively carried out for $j = t, \ldots, 1$ to obtain $\bar{Q}_{t,n}^{\bar{a}_r,*}(1)$. This is then repeated for each t, after which we can compute our targeted substitution estimator $\boldsymbol{\psi}_r$ using a pooled logistic regression. The updates are carried out using a specifically constructed multidimensional covariate $\boldsymbol{h}_j(\bar{a}_r, t, \boldsymbol{W})$. This covariate is derived from the parameter of interest's Efficient Influence Curve (EIC). The update ensures that the generalized score of the submodel used for the update spans the EIC. Therefore, by construction, TMLE creates estimators that solve the estimating equation of the EIC, and are therefore doubly robust and locally semi-parametric efficient estimators in their class of Regular Asymptotically Linear estimators [**116**]. The estimation steps are detailed below using an example

where $Y(t)$ is a binary indicator of a failure type outcome at time t and the target parameters are the MSM coefficients defined in equation (3.4.1).

3.6.1.1. *Algorithm*

First, calculate the estimator $\bar{g}_n^{\bar{a}_r}(j)$ of $\bar{g}^{\bar{a}_r}(j) = \prod_{k=0}^{j} P(A(k) = a(k)|\bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}_r(k-1))$, by estimating every component $\bar{g}_{k,n}^{\bar{a}_r}(j) = P_n(A(k) = a(k)|\bar{\boldsymbol{L}}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}_r(k-1))$. Denote $\bar{g}_n^{\bar{a}_r}(j) = \prod_{k=0}^{j} \bar{g}_{k,n}^{\bar{a}_r}(j)$ as their product.

Define $\bar{Q}_{t,n}^{\bar{a}_r,*}(t+1) \equiv Y(t)$.

For $t = K_r + 1, \ldots, 1$ set $j = t$:

Step 1 : Initial estimation of $\bar{Q}_t^{\bar{a}_r}(j)$

Calculate the initial estimate by evaluating the conditional expectation $\bar{Q}_t^{\bar{a}_r}(j) = E(\bar{Q}_{t,n}^{\bar{a}_r,*}(j+1)|\bar{A}(j-1), \bar{\boldsymbol{L}}(j-1)), \bar{Y}(j-1))$. Then predict $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ for each possible value of $\bar{a}_r$ and every subject $i = 1, \ldots, n$. It follows that $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ is of length $n \times$ *number of potential regimes*.

Step 2 : Updating step

For every individual and for every exposure regime $\bar{a}_r \in \bar{\mathcal{A}}_r$ calculate :

$$\boldsymbol{h}_j(\bar{a}_r, t) = \frac{I(\bar{A}(j-1) = \bar{a}_r(j-1))}{\bar{g}_n^{\bar{a}_r}(j-1)} \frac{\partial}{\partial \boldsymbol{\beta}} f\left[\eta(\boldsymbol{\beta}, \bar{A}, t)\right]\Big|_{\bar{A}(t-1)=\bar{a}_r(t-1)} . \quad (3.6.1)$$

The term $\frac{d}{d\boldsymbol{\beta}} f\left[\eta(\boldsymbol{\beta}, \bar{A}, t)\right]$ is the derivative of the MSM formula with respect to $\boldsymbol{\beta}$ which is evaluated at $\bar{A}(t-1) = \bar{a}_r(t-1)$. The term $I(\bar{A}(j-1) = \bar{a}_r(j-1))$ is an indicator function that is equal to 1 if the observed exposure up until time $j-1$ is equal to the exposure regime $\bar{a}_r(j-1)$. For an individual and a specific exposure regime $\bar{a}_r$, the dimension of $\boldsymbol{h}_j(\bar{a}_r, t)$ is equal to the dimension of $\boldsymbol{\beta}$. Therefore, $\boldsymbol{h}_j(\bar{a}_r, t)$ is of dimension $n \times$ *number of regimes* $\times$ *number of MSM coefficients*.

The following submodel defines the update path for $\bar{Q}_{t,n}^{\bar{a}_r,*}(j)$:

$$logit(\bar{Q}_{t,n}^{\bar{a}_r,*}(j)) = logit(\bar{Q}_{t,n}^{\bar{a}_r}(j)) + \boldsymbol{\epsilon}\boldsymbol{h}_j(\bar{a}_r, t). \tag{3.6.2}$$

$\bar{Q}_{t,n}^{\bar{a}_r,*}(j)$ is obtained by fitting an intercept-free pooled logistic regression, over all $\bar{a}_r$, of $\bar{Q}_t^{\bar{a}_r,*}(j+1)$ on the covariates $\boldsymbol{h}_j(\bar{a}_r, t)$ using the previously calculated initial estimates $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ as offset. Subsequently, plug $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_n$ into equation (3.6.2) and evaluate it for each possible value of $\bar{a}_r$ to give $\bar{Q}_{t,n}^{\bar{a}_r,*}(j)$ for every subject under every possible $\bar{a}_r$.

For $j = t-1, \ldots, 1$:

Step 3 : Estimation of $\bar{Q}_t^{\bar{a}_r}(j)$

For every $\bar{a}_r \in \bar{\mathcal{A}}_r$ separately, calculate the conditional expectation $\bar{Q}_t^{\bar{a}_r}(j) = E(\bar{Q}_t^{\bar{a}_r,*}(j+1)|\bar{A}(j-1), \bar{\boldsymbol{L}}(j-1), \bar{Y}(j-1))$ and evaluate the obtained estimate $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ at each respective $\bar{a}_r$. Consequently, we obtain,

for every subject, a copy of $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ for each different regime $\bar{a}_r$. It follows that $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ is of length $n \times$ *number of regimes*.

Step 4 : Update $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ to $\bar{Q}_t^{\bar{a}_r,*}(j)$

Construct the covariate $\boldsymbol{h}_j(\bar{a}_r, t)$ as before using equation (3.6.1). Then estimate $\boldsymbol{\epsilon}$ by pooled regression over all regimes $\bar{a}_r$ using submodel (3.6.2) and update $\bar{Q}_{t,n}^{\bar{a}_r}(j)$ to $\bar{Q}_{t,n}^{\bar{a}_r,*}(j)$ as above.

Final Step :

The above steps provide $\bar{Q}_{t,n}^{\bar{a}_r,*}(1)$ for every t and every $\bar{a}_r$. The estimate of the pooled LTMLE estimator $\boldsymbol{\psi}_r$ is obtained by solving :

$$\begin{aligned}\boldsymbol{\psi}_{r,n} \;=\; \text{argmin}_{\boldsymbol{\beta}}\, E_n \sum_{t=1}^{K_r+1} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \{\bar{Q}_{t,n}^{\bar{a}_r,*}(1) log(expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t))) \;+ \\ (1 - \bar{Q}_{t,n}^{\bar{a}_r,*}(1)) log(1 - expit(\eta(\boldsymbol{\beta}, \bar{a}_r, t)))\}.\end{aligned}$$

where E_n is the empirical distribution. In practice, this estimate is obtained by regressing $\bar{Q}_t^{\bar{a}_r,*}(1)$ on $\bar{a}_r$ and t according to the linear specification of the MSM.

TMLE creates doubly robust estimators, which means that it is consistent if either group of models for the estimation of Q_0 or g_0 is correctly specified and is asymptotically efficient when both are correct. The variance of $\boldsymbol{\psi}_r$ can be approximated using a sandwich estimator based on the EIC. For derivation of the EIC refer to Petersen *et al.* [**8**]. Note that the EIC is a function of the inverse of

the exposure probabilities which can take large values in practice. Recall that the positivity assumption requires that at each time t within all levels of covariate history $\bar{\boldsymbol{L}}(t)$ every subject has a non-null probability of receiving every possible value of exposure according to $\bar{\mathcal{A}}_r$. So, even if in theory every combination can be received, estimated near practical positivity violations may occur when the support in the data for a given regime is not sufficient. Under near practical positivity violations, standard causal methods have been known to produce biased estimates and inflated variance [**128**]. In longitudinal contexts with many observational time-points, we may be more at risk of so called near practical positivity violations because the conditional exposure model is the product of all past time-point conditional exposure probabilities [**116**]. When the discretization is overly fine and the number of time-points is large these probabilities may vanish.

3.6.2. Selection procedure

We propose an automatic selection procedure to choose an optimal degree of coarsening that penalizes for both bias due to coarsening the timeline and variance which can increase particularly on a finer discretization. This optimal discretization, r^*, would therefore be the one allowing to adjust sufficiently for time-dependent confounding while minimizing the variance caused by unnecessary refining, if possible.

Our selection procedure incorporates the pooled LTMLE algorithm by using the following loss function :

$$\mathcal{L}_1(\bar{Q}_n^*) = -E_n \sum_{t=1}^{K_r+1} \sum_{j=1}^{t} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \frac{I(\bar{A}(j-1) = \bar{a}_r(j-1))(\bar{Q}_{t,n}^{\bar{a}_r,*}(j+1) log(\bar{Q}_{t,n}^{\bar{a}_r,*}(j)) + (1 - \bar{Q}_{t,n}^{\bar{a}_r,*}(j+1)) log(1 - \bar{Q}_{t,n}^{\bar{a}_r,*}(j)))}{\delta_{rj1}}$$

where $\delta_{rj1} = |I(\bar{A}(j-1) = \bar{a}_r(j-1))|$ is defined as the number of subjects with observed exposures up to time $j-1$ that are equal to the exposure regime $\bar{a}_r(j-1)$.

Below, we describe the algorithm of our selection procedure.

Step 1 : Select compared discretizations

Beforehand, the candidate discretizations that we want to compare must be chosen. This choice could be motivated by a priori knowledge or convention.

Step 2 : Cross-validation [120]

Our procedure uses the sum of V-fold cross-validated errors, based on the proposed loss function, to select the optimal discretization.

For each candidate discretization r, separate each discretized dataset into V folds of size $\frac{n}{V}$. Let each of these folds be indicated by $v = 1, \ldots, V$. Let the observations in a specific fold v constitute the validation sample and let the training sample comprise of the remaining $V-1$ folds. Let $P^0_{0,v}$ and $P^1_{0,v}$ represent the true probability distributions of the training and validation samples, respectively, and let $P^0_{n,v}$ and $P^1_{n,v}$ represent estimations of $P^0_{0,v}$ and $P^1_{0,v}$, respectively.

For each fold $v = 1, \ldots, V$, repeat steps 3 and 4 :

Step 3 : Pooled LTMLE

Fit the pooled LTMLE algorithm using the training sample, i.e. estimators $P^0_{n,v}$ of $P^0_{0,v}$ are built in the training sample.

Step 4 : Evaluation of $\mathcal{L}_1(\bar{Q}^{*1}_{n,v}(P^0_{n,v}))$

Define $\bar{Q}^{*1}_{n,v}(P^0_{n,v})$ as the pooled LTMLE updated $\bar{Q}_n$ values evaluated on the validation sample using the models $P^0_{n,v}$ built with the training sample in Step 3. Specifically, using the remaining sample v, calculate estimates of $\bar{Q}^*_n$ from the estimated fits of $P^0_{n,v}$. With the validation sample v, evaluate the error $\mathcal{L}_1(\bar{Q}^{*1}_{n,v}(P^0_{n,v}))$.

The optimal discretization, r^*, is the one minimizing $\sum_{v=1}^{V} \mathcal{L}_1(\bar{Q}^{*1}_{n,v}(P^0_{n,v}))$.

The resulting estimated parameter $\boldsymbol{\psi}_{r,n}$ is obtained using the selected optimal discretized dataset $\boldsymbol{O}_{r^*}$.

3.7. Simulation study

This simulation study aims to assess the impact of discretization on the estimation of a causal quantity of interest using pooled LTMLE and to evaluate the performance of the proposed loss function in selecting a discretization.

The simulated data consist of n *i.i.d.* observations with structure

$$\boldsymbol{O}_i = (L_i(0), A_i(0), Y_i(1), L_i(1), A_i(1), Y_i(2), \ldots, L_i(11), A_i(11), Y_i(12)).$$

These data mimic a study where individuals remain exposed after first exposure and where the interest lies in a discrete time-to-event outcome. Hence, defining all variables as binary, let $A(t) = 1$ indicate whether a person was exposed by time t, $L(t) = 1$ indicate whether a person had covariate L at time t, with $L(0)$ representing a sole baseline covariate, and $Y(t) = 1$ indicate the occurrence of the outcome by time t. The exposure and outcome processes are monotone (i.e. if $Y(t-1) = 1$ then $Y(t) = 1$, and similarly for $A(t)$) and all subjects were outcome-free at study entry (i.e. $Y(0) \equiv 0$). The data is generated at the finest scale, at fixed equally-spaced intervals, conditional on the past three time-points, $t-1, t-2, t-3$. All variables were generated according to the following Bernoulli (Bern) processes :

$$L(0) \sim Bern[0.3]$$

$$L(t) = \begin{cases} \text{NA} & \text{if } Y(t) = 1 \\ Bern[expit(-0.75 + log(1.5)L(t-1) + log(0.6)A(t-1) + \\ log(1.25)L(t-2) + log(0.8)A(t-2) + \\ log(1.1)L(t-3) + log(0.9)A(t-3))] & \text{otherwise} \end{cases}$$

$$A(t) = \begin{cases} \text{NA} & \text{if } Y(t) = 1 \\ 1 & \text{if } A(t-1) = 1 \\ Bern[expit(-3.5 + log(1.75)L(t) + log(1.5)L(t-1) + \\ log(1.25)L(t-2) + log(1.05)L(t-3))] & \text{otherwise} \end{cases}$$

$$Y(t+1) = \begin{cases} 1 & \text{if } Y(t) = 1 \\ Bern[expit(-4 + log(3.5)L(t) + log(0.75)A(t) + \\ log(2.5)L(t-1) + log(0.85)A(t-1) + log(1.5)L(t-2) & \text{otherwise} \\ + log(0.95)A(t-2))] \end{cases}$$

Candidate discretizations were created by sequentially removing a single time-point from the finest generated data. The candidate discretized datasets consist

of $\mathcal{O} = \{\boldsymbol{O}_{r_0}, \boldsymbol{O}_{r_1}, \boldsymbol{O}_{r_2}, \boldsymbol{O}_{r_3}, \boldsymbol{O}_{r_4}, \boldsymbol{O}_{r_5}, \boldsymbol{O}_{r_6}, \boldsymbol{O}_{r_7}, \boldsymbol{O}_{r_8}, \boldsymbol{O}_{r_9}, \boldsymbol{O}_{r_{10}}\}$. T_r indicates which time-points are included in the discretization, such that $T_{r_0} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, $T_{r_1} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12\}$, $T_{r_2} = \{0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 12\}$, $T_{r_3} = \{0, 1, 2, 4, 5, 6, 8, 9, 10, 12\}$, $T_{r_4} = \{0, 1, 2, 4, 5, 6, 8, 10, 12\}$, $T_{r_5} = \{0, 2, 4, 5, 6, 8, 10, 12\}$, $T_{r_6} = \{0, 2, 4, 6, 8, 10, 12\}$, $T_{r_7} = \{0, 2, 4, 6, 8, 12\}$, $T_{r_8} = \{0, 4, 6, 8, 12\}$, $T_{r_9} = \{0, 4, 8, 12\}$, $T_{r_{10}} = \{0, 4, 12\}$. In practice, a number of discretization strategies may be employed. We adopt a discretization method that consists of removing from the data the observed information of unselected time-points, as if it had never been observed. For example, $\boldsymbol{O}_{r_9} = (L(0), A(0), Y(4), L(4), A(4), Y(8), L(8), A(8), Y(12))$, where $(Y(t), L(t), A(t)) \in \boldsymbol{O}_{r_9}$ are identical to $(Y(t), L(t), A(t)) \in \boldsymbol{O}_{r_0}$ for $t \in \{0, 4, 8, 12\}$.

The target parameter is defined as in equation (3.4.2). Here the goal is therefore to analyze the effect of most-recent exposure on the probability of event by time t across different discretizations. The true value of the parameter of interest for the finest discretization was assessed numerically and is equal to -0.44 up to the second decimal. True values for every other discretization were also attained numerically and are presented in Table 3.1. In the simulation, the target parameter is estimated through pooled LTMLE. Pooled LTMLE code for the parameters of a MSM is currently available in R software [**129**] as the *ltmleMSM* function of the *ltmle* [**130**] package version 0.9-9, developed in [**8**]. We constructed pooled LTMLE code specific to this estimation problem, which can be found in the Appendix with a two time-point example. This was done in order to reduce the computational complexity of the more general function *ltmleMSM* and to more efficiently perform the cross-validation steps. The point estimates were found to be roughly identical between both codes with sample datasets.

It is often recommended to respond to practical positivity violations using ad-hoc methods such as truncation [**128**]. Since we propose an alternative, principled approach to performing a bias-variance trade-off, we do not employ such methods in the simulation study. In practice, however, they could be used in combination.

500 sets of simulated data were analyzed, and the results pooled. In Table 3.1, for each discretization, the mean estimates and Monte Carlo (MC) variance are reported for the corresponding pooled LTMLE estimate. Due to the nature of the generated data, which may present severe practical positivity violations at certain levels of discretization, the pooled LTMLE algorithm may not converge. In such situations, we report an estimated value NA. We therefore report mean estimates and MC variance from non-missing estimates only. Additionally in Table 3.1, we present these measures for the chosen optimal discretization as well as the percentage of times each discretization was chosen as the optimal discretization. In order to assess how the support in the data affects the discretization choice, data with sample sizes $n = 500$, 1000, 2500 were analyzed. Furthermore, an alternative loss $\mathcal{L}_2$ was analyzed and the results are also presented in Table 3.1 :

$$\mathcal{L}_2(\bar{Q}_n^*) = -\sum_{t=1}^{K_r+1} \sum_{\bar{a}_r \in \bar{\mathcal{A}}_r} \frac{I(\bar{A}(t-1) = \bar{a}_r(t-1))(\bar{Q}_{t,n}^{\bar{a}_r,*}(t+1) log(\bar{Q}_{t,n}^{\bar{a}_r,*}(1)) + (1 - \bar{Q}_{t,n}^{\bar{a}_r,*}(t+1)) log(1 - \bar{Q}_{t,n}^{\bar{a}_r,*}(1)))}{\delta_{rt2}}$$

where $\delta_{rt2} = |I(\bar{A}(t-1) = \bar{a}_r(t-1))|$ is the number of subjects with observed exposures up to time $t-1$ that are equal to the exposure regime $\bar{a}_r(t-1)$.

3.7.1. Simulation results

The simulation study displays the consequences of discretization on the estimation problem described in Sections 3.4 and 3.5. The results for each discretization and for the selection procedure are presented in Table 3.1, which is separated in three sections : one for each sample size. The first two rows of each section

TABLEAU 3.1. Mean estimate, MC variance, percentage selected for each loss and percentage NA for every discretization and the optimal discretization as selected using $\mathcal{L}_1$ and $\mathcal{L}_2$ for various sample sizes

	Discretization	True value†	Mean Est.*	MC variance*	% select $\mathcal{L}_1$	% select $\mathcal{L}_2$	% NA
n = 500	Optimal - $\mathcal{L}_1$	-0.44	-0.64	0.11	-	-	0.8
	Optimal - $\mathcal{L}_2$	-0.44	-0.62	0.08	-	-	0.8
	r_{10}	-0.65	-0.80	0.15	24.8	9.2	0.0
	r_9	-0.60	-0.66	0.08	34.8	48.8	0.4
	r_8	-0.52	-0.53	0.05	26.0	27.2	21.2
	r_7	-0.52	-0.51	0.04	11.2	11.2	37.4
	r_6	-0.51	-0.49	0.03	1.6	1.8	46.0
	r_5	-0.47	-0.43	0.02	1.0	1.2	60.6
	r_4	-0.49	-0.50	0.04	0.2	0.2	79.6
	r_3	-0.47	-0.50	0.03	0.2	0.2	81.6
	r_2	-0.44	-0.45	0.03	0	0	85.6
	r_1	-0.44	-0.44	0.03	0	0	84.8
	r_0	-0.44	-0.46	0.03	0	0	87.2
n = 1000	Optimal - $\mathcal{L}_1$	-0.44	-0.53	0.03	-	-	1.0
	Optimal - $\mathcal{L}_2$	-0.44	-0.52	0.03	-	-	1.0
	r_{10}	-0.65	-0.78	0.07	2.4	0	0.0
	r_9	-0.60	-0.65	0.04	7.4	9.8	0.0
	r_8	-0.52	-0.52	0.03	20.0	19.8	2.6
	r_7	-0.52	-0.52	0.02	44.2	42.8	7.2
	r_6	-0.51	-0.51	0.02	1.6	1.6	8.4
	r_5	-0.47	-0.43	0.01	17.4	18.8	22.6
	r_4	-0.49	-0.47	0.02	1.8	1.8	45.6
	r_3	-0.47	-0.47	0.01	1.6	2.2	47.2
	r_2	-0.44	-0.42	0.01	1.0	0.6	52.0
	r_1	-0.44	-0.41	0.01	1.6	1.6	63.6
	r_0	-0.44	-0.43	0.01	1.0	1.0	65.0
n = 2500	Optimal - $\mathcal{L}_1$	-0.44	-0.47	0.01	-	-	0.0
	Optimal - $\mathcal{L}_2$	-0.44	-0.47	0.01	-	-	0.0
	r_{10}	-0.65	-0.76	0.03	0	0	0.2
	r_9	-0.60	-0.64	0.02	0	0	0.0
	r_8	-0.52	-0.51	0.01	2.0	2.0	0.2
	r_7	-0.52	-0.51	0.01	6.6	5.8	0.2
	r_6	-0.51	-0.50	0.01	0.6	0.6	0.2
	r_5	-0.47	-0.44	0.01	22.0	22.8	0.8
	r_4	-0.49	-0.49	0.01	25.0	23.4	8.0
	r_3	-0.47	-0.48	0.01	2.8	3.4	9.0
	r_2	-0.44	-0.45	0.01	5.0	4.4	7.6
	r_1	-0.44	-0.44	< 0.005	34.0	35.4	13.6
	r_0	-0.44	-0.44	< 0.005	2.0	2.2	13.0

† indicates true value for every discretization
* computed using non missing values only

correspond with results from the selected optimal discretization, using $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Underneath are the results for every candidate discretization. The true values of the target parameter of interest for every discretization are displayed in the second column. It can be observed that the true value changes according to the underlying discretization of the data, with the true value at the finest discretization (-0.44) being far from the true value at the coarsest discretization (-0.65). For the optimal selected discretization, the reference true value is of the finest discretization.

The following two columns contain the mean pooled LTMLE estimates and MC variance for every discretization and for the optimal selected discretization. We can see from these results that at coarser discretizations the mean estimate was biased for the corresponding true value of the parameter of interest and that, at finer discretizations, the mean estimates were roughly unbiased for the true value. For example, with a sample size of $n = 500$, the mean estimate at the coarser discretization, r_{10}, was equal to -0.80 and the true value equalled -0.65. This indicates that coarser discretizations may not have fully controlled for time-dependent confounding and that corresponding estimates were biased. Note also that the Monte Carlo variance decreased with finer discretization. This corresponds with the logic that as the number of time-points increases our estimation becomes more precise since we have more pooled data. However, these measures were only computed on available estimates and, for finer discretizations and small sample sizes, most of the estimates were returned as NA, as displayed in the last column. For example, considering a sample size of $n = 500$, discretizations finer than r_4 produced an estimate for less than 20% of simulated sets. Consequently, these discretizations could only be considered as optimal less than 20% of times.

Missing estimates (denoted as NA and tabulated in the final column) were partly due to large weights that occurred frequently at finer discretizations. This was particularly problematic at smaller sample sizes, often resulting in the pooled LTMLE algorithm failing to converge. An inspection of the weights across discretizations for a fixed sample size of $n = 500$ (results not shown here) illustrated that for finer discretizations weights could average as high as $6.13\text{x}10^{14}$. Discretizations leading to such destabilizing weights were potentially responsible for the pooled LTMLE algorithm failing to converge, in which cases we systematically assigned infinite cross-validated error and so it could not be selected as optimal.

Regarding the cross-validated selection procedure incorporating $\mathcal{L}_1$, in the column % select $\mathcal{L}_1$, we first notice that, as the sample size increased, it tended to select finer discretizations. This is illustrated by the percentage of times each discretization was selected. Indeed, the most selected discretizations were r_9, r_7, and r_1 for sample sizes $n = 500, 1000,$ and 2500, respectively. We also notice that, as sample size increased, the mean optimal estimate appeared to converge to an unbiased estimate of the parameter of interest at the finest discretization. In particular, the optimally selected estimate at $n = 2500$ $(-0, 47)$ was approximately unbiased for the parameter of interest at the finest discretization (-0.44). In addition, as the sample size increased, the Monte Carlo variability of the optimally selected estimate was reduced. The small variability of these (0.01 for $n = 2500$) indicates that the chosen values were relatively close, regardless of the chosen discretization. Finally, we note that the cross-validated selection procedure occasionally chose optimal discretizations that, due to resampling variability, resulted in the pooled LTMLE algorithm failing to converge once employed with the full data. However, this rarely occurred (with prevalences of $0.8\%, 1.0\%,$ and 0.0% for sample sizes of $n = 500, 1000,$ and 2500, respectively). As shown in the column %

select $\mathcal{L}_2$, all results for the selection procedure employing $\mathcal{L}_2$ are almost identical.

3.8. Data application

This section revisits the motivating example of Section 3.2 on the comparison of the effect of low ICS dose versus no ICS on pregnancy duration. Data on pregnant women with asthma were longitudinally extracted from the linkage of administrative databases. To reflect the underlying nature of the data and to capture all possible changes in covariates, a day-by-day follow-up of asthma medication exposure, asthma related covariates and pregnancy related covariates was constructed for every pregnancy. The final cohort consists of pregnancies with no gaps in the woman's insurance plan coverage from 1 year before and throughout pregnancy, of women who were less than 45 years of age, and had a singleton delivery. The presence of asthma was established based on at least one diagnosis of asthma combined with at least one filled prescription for an asthma medication during the pregnancy or one year prior to pregnancy. In this cohort, women could only contribute to a maximum of two pregnancies. Women taking theophylline, cromoglycate, nedocromil, ketotifen, or LABA without an ICS were excluded. Finally, a subsetted cohort of women with mild asthma in the year prior to pregnancy with no use of ICS during the first trimester was created. Mild asthma was defined using a validated severity indicator developed in our research group [**131**]. The start of follow-up was established at 20 weeks since no deliveries occurred before week 20 by definition. In this application, candidate discretized datasets to be analyzed consisted of 3-week intervals ($\boldsymbol{O}_{r_3}$), 4-week intervals ($\boldsymbol{O}_{r_4}$), 5-week intervals ($\boldsymbol{O}_{r_5}$), and 6-week intervals ($\boldsymbol{O}_{r_6}$) from start of follow-up. The finest discretized dataset consisting of daily data was not considered as a candidate discretization due to the resulting data being far too voluminous and sparse. Discretized data were created from the finest daily data and

defined as $\boldsymbol{O}_r = (Y(t), \boldsymbol{L}(t), \boldsymbol{A}(t))$, $t \in T_r$, where $\boldsymbol{L}(t)$ is a vector of confounder variables measured during $[t-1, t[$, $\boldsymbol{A}(t) = (A_1(t), A_2(t))$ represents a multivariate "treatment" measured at t composed of an ongoing exposure indicator $A_1(t)$ and a censoring indicator $A_2(t)$, where $A_1(t) = 1$ indicates exposure to low daily doses of ICS and $A_1(t) = 0$ indicates no exposure to ICS. Identically, $A_2(t) = 1$ indicates that a subject has been censored by time t, $A_2(t) = 0$ otherwise. Specifically, a subject could be censored if their asthma treatment differed from one of the above defined regimes for mild asthma. For example, a subject could be censored if they begin receiving a higher ICS daily dose or the concomitant usage of LABA with ICS, which both indicate an increase in asthma severity. Finally, $Y(t) = 1$ represents a delivery occurring during $[t-1, t[$. Figure 3.4 displays the time-ordering of the observed data according to different discretizations.

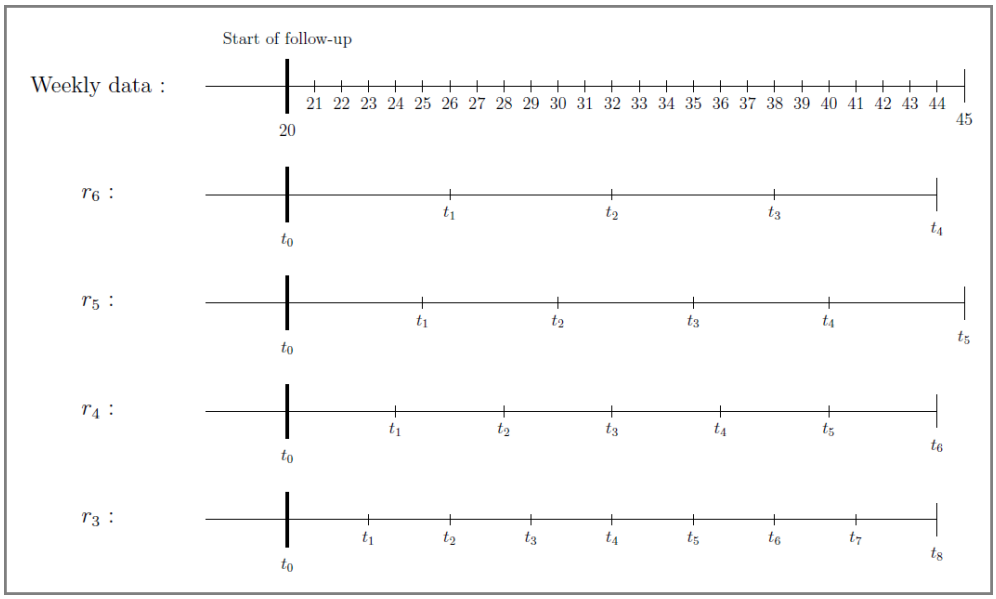

FIGURE 3.4. Candidate discretized timelines in the data application

The potential considered confounders were divided into four categories : characteristics of the mother, chronic maternal diseases, pathologies related to pregnancy, and maternal asthma control related variables. A complete list of confounders may be found in Table 3.2. For the final dataset, baseline characteristics are

presented in Table 3.2. In this application baseline characteristics were evaluated from the year prior to pregnancy to the start of follow-up at 20 weeks of gestation.

The parameter of interest was defined according to equation (3.4.2). In this study, because the algorithm to estimate ICS exposure is monotone (once a woman receives a medication, she is considered to be exposed for the remainder of the pregnancy, unless censored), regimes of interest were defined as initiation of low ICS doses at any time during pregnancy. For instance, for women who haven't initiated ICS treatment in the first two time periods, then start at the third, the regime would be written as $\bar{a}_r = (0, 0, 1, \ldots, 1)$.

The usual influence curve based sandwich estimator for the variance of pooled LTMLE is said to be anti-conservative when practical positivity violations occur [**8**]. In order to obtain a valid estimate of the variance, the original pooled LTMLE function from the *ltmleMSM* R package was used for each discretization to obtain an alternative robust variance estimate [**130**]. Efficiency in the estimation of the pooled LTMLE requires consistent estimation of both the models for the probabilities of exposure and outcome processes. Therefore, it may be preferred to estimate these quantities using machine learning techniques or Super Learner [**92**], an ensemble-learning approach. However, due to computational reasons, we opted not to use such data-adaptive methods for this application. Therefore, simple logistic regressions conditional on all past covariates were used to estimate the models in steps 1 and 2 of the pooled LTMLE algorithm. As in common practice, in order to avoid overly large weights, bounding at a level of 5% was applied.

3.8.1. Application Results

The final cohort of women with mild asthma in the year prior to pregnancy with no ICS use during the first trimester comprised of 2878 pregnancies. The

distribution of pregnancy duration in weeks is depicted by the histogram in Figure 3.5. The pregnancy duration in this cohort had a mean of 38.5 weeks and a range of 20 to 42 weeks, which corresponds roughly with the distribution in the general population of pregnant women with asthma.

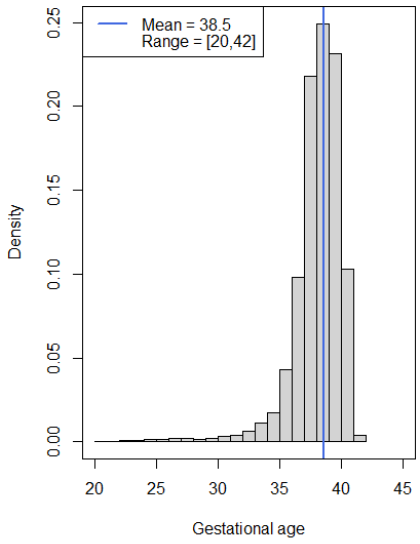

FIGURE 3.5. Histogram and mean of pregnancy duration for the final cohort

The baseline characteristics of the pregnancy cohort are given in Table 3.2 by exposure at the 20^{th} week of gestation, corresponding with the start of follow-up. Exposure at the start of follow-up includes no ICS doses or low doses of ICS. In the case that a woman is exposed to neither of these treatments, she was censored. For gestational hypertension / PECL / ECL, a missing value at baseline is presented since these characteristics could only be measured after the 20^{th} week of pregnancy.

The sample was primarily composed of women who had their asthma controlled in the year prior to pregnancy (80.40%). Women with no ICS usage at start of follow-up tended to have their asthma less well controlled in the year prior to

pregnancy. The women were also mostly aged between 18 and 34 years at delivery (85.79%), mostly living in urban areas (83.22%), with roughly half receiving social assistance (54.76%). With respect to mother characteristics, chronic maternal diseases, and pathologies related to pregnancy, no great disparities existed between exposure groups. On the other hand, asthma control related variables were dissimilar for women taking low ICS doses versus no ICS doses. Indeed, a higher proportion of women with low ICS doses had at least one hospitalization or emergency room visit for asthma. Identically, these women had also a higher proportion of oral and nasal corticosteroids, and had a higher number of doses per week of short-acting beta2-agonists. Markers for poor asthma control were much higher in the low ICS group due to differential indication since in the latent period between the end of first trimester and cohort entry their asthma may have been poorly controlled, which motivated the clinical decision to change asthma treatment to low ICS doses. This table also demonstrates potential practical positivity problems in the data as evidenced by the few counts in many cells.

Table 3.3 shows exposed women, women who changed treatment from no ICS to low doses of ICS, and new censored women at each time-point $t \in T_r$ for every candidate discretization. For example, for the discretized dataset $\boldsymbol{O}_{r_3}$, there were 275 exposed women at t_0, the start of follow-up. At t_1 there were now 366 women exposed, 100 of whom changed from no ICS at t_0 to low doses of ICS at t_1. In total, 36 women from either exposure group were censored at t_1. Missing values indicate end of exposure measurement. For example, for the discretized data $\boldsymbol{O}_{r_6}$, all values were missing after time-point t_3, since only four time-points were used. Generally, the results from this table show that as the discretizations get coarser, the data offers more support for every exposure regime of interest.

TABLEAU 3.2. Women's baseline characteristics per exposure status at 20 weeks of gestation- *n(%)*

	Exposure		
	No ICS	Low ICS	Neither (Censored)
	n = 2426	n = 275	n = 177
Characteristics of the mother			
Age at beginning of pregnancy			
< 18	30 (1.24)	8 (2.91)	0 (0.00)
18-34	2095 (86.36)	236 (85.82)	138 (77.97)
> 34	301 (12.41)	31 (11.27)	39 (22.03)
Social assistance benificiary in the year prior to pregnancy	1322 (54.49)	147 (53.45)	107 (60.45)
Rural location of residence at delivery	413 (17.02)	44 (16.00)	26 (14.69)
Chronic maternal diseases			
Chronic hypertension	67 (2.76)	7 (2.55)	3 (1.69)
Diabetes mellitus	77 (3.17)	3 (1.09)	7 (3.95)
Uterine disorders	44 (1.81)	4 (1.45)	1 (0.56)
Other chronic diseases	17 (0.70)	3 (1.09)	1 (0.56)
Pathologies related to pregnancy			
Gestational diabetes	31 (1.28)	9 (3.27)	2 (1.13)
Gestational hypertension / PECL / ECL	-	-	-
Placental complications	20 (0.82)	0 (0.00)	2 (1.13)
Other pregnancy-related pathologies	475 (19.58)	52 (18.91)	29 (16.38)
Maternal asthma control related variables			
Asthma control in the year prior to pregnancy	1991 (82.07)	218 (79.27)	105 (59.32)
SABA (doses/week)			
0	1159 (47.77)	44 (16.00)	20 (11.30)
> 0-3	1267 (52.23)	231 (84.00)	157 (88.70)
> 3	0 (0.00)	0 (0.00)	0 (0.00)
LTRA	30 (1.24)	1 (0.36)	9 (5.08)
OCS	179 (7.38)	55 (20.00)	41 (23.16)
NCS	266 (10.96)	53(19.27)	43 (24.29)
Hospitalization for asthma	20 (0.82)	9 (3.27)	13 (7.34)
Emergency room visit for asthma	207 (8.53)	77 (28.00)	32 (18.08)

In Table 3.4, Pooled LTMLE estimates of the parameter of interest and corresponding standard errors and 95 % confidence intervals (CI) are reported for every discretization. The corresponding values of the cross-validated loss are also presented for each candidate discretization. The reported values in the second column correspond to estimates of the β_1 parameter in equation (3.4.1), interpreted as the effect of most recent exposure on delivery. Consequently, $exp(\beta_1)$ corresponds to the odds of delivery for women on low ICS dose versus no ICS. Hence, regardless of discretization, the point estimates can be interpreted as a

TABLEAU 3.3. Number of exposures, treatment changes, and censorings at every time-point for each candidate discretization

	Disc.	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
	r_6	275	465	624	619	-	-	-	-	-
Exposure	r_5	275	422	576	684	271	-	-	-	-
(Prevalent)	r_4	275	395	517	627	692	270	-	-	-
	r_3	275	366	460	538	624	682	616	77	-
	r_6	275	200	177	117	-	-	-	-	-
Treatment change	r_5	275	158	167	128	33	-	-	-	-
(Incident)	r_4	275	130	131	116	98	21	-	-	-
	r_3	275	100	98	87	90	75	56	4	-
	r_6	177	45	47	19	-	-	-	-	-
Censoring	r_5	177	43	39	26	6	-	-	-	-
(Incident)	r_4	177	36	33	25	18	4	-	-	-
	r_3	177	31	18	31	17	13	7	0	-

reduced odds of delivery after switching to low ICS dose at any given time t, which is consistent with clinical hypotheses.

For all discretizations, the results were statistically non-significant, except for the estimate obtained from the discretized data $\boldsymbol{O}_{r_4}$. $\mathcal{L}_1$ and $\mathcal{L}_2$, were minimized at the discretization r_3, so that the optimal discretization selected by our procedure consisted of 3 week intervals. Hence, we obtain a point OR estimate of 0.773 with CI = [0.453,1.317]. We observe that the decision of "optimal" discretization sometimes depends on the choice of loss function. Had $\boldsymbol{O}_{r_3}$ not been a candidate dataset, $\mathcal{L}_1$ and $\mathcal{L}_2$ would not have agreed on the best optimal discretization. In contrast with the simulation results, the data application point estimates did not display clear convergence to a value as the discretization became finer. The fact that the candidate discretizations were not necessarily nested may have contributed to this. The standard error estimates decreased progressively for discretizations r_6, r_5, and r_4. This can be explained since additional data translates into

a gain in precision in the pooled model. Yet, at the finest discretization r_3, the standard error estimate increased. This suggests that this discretization led to large weights and standard errors.

TABLEAU 3.4. Pooled LTMLE estimates with corresponding standard error, 95% CI, and cross-validated loss values for every candidate discretization

Discretization	Estimate	Standard error	95% CI	$\mathcal{L}_1$	$\mathcal{L}_2$
r_6	-0.232	0.164	[-0.554, 0.090]	0.901	0.719
r_5	-0.171	0.136	[-0.437, 0.095]	1.061	0.751
r_4	-0.310	0.133	[-0.570, -0.050]	1.000	0.650
r_3	-0.258	0.272	[-0.792, 0.275]	0.856	0.526

3.9. DISCUSSION

Apart from a few methods that handle a continuous underlying data generating distribution [**58**, **103**, **104**, **132**], most causal inference methods have relied on arbitrary discretization of the patient timeline in longitudinal studies. We note that arbitrary discretization is analogous to the choice of follow-up time-points in an observational study where treatment can change between follow-ups. We have shown that such arbitrary timeline coarsening may result in bias and inflated variance and affects the value and existence of the underlying parameter of interest one is estimating. Furthermore, we proposed an automatic selection procedure of the timeline integrated with pooled LTMLE. This procedure relied on the cross-validated evaluation of a pooled LTMLE loss based function. This procedure is readily adaptable to any MSM and LTMLE specification.

One appeal of coarser discretization is that it may limit practical positivity violations by decreasing the number of observed time-points. Certain causal inference methods have been known to be sensitive to such violations which may

occur much more frequently in longitudinal contexts. Although pooled LTMLE has shown practical advantages over IPTW methods, notably in contexts with small support for certain regimes of interest, such methods may still be vulnerable to severe practical positivity violations. Such violations occurred in our simulation study when the number of time-points was large and the sample size small and in the data application, particularly when the number of time-points increased, given the relatively high confounder space.

In the simulation study, we observed that, as sample size increased, the selection procedure estimates converged towards the true value under the finest discretization. Recall that confounding was generated at the finest scale. Hence, in this scenario, at larger sample sizes, the selection procedure tended towards selecting discretizations that adequately controlled for confounding. In future work, it would be interesting to generate confounding at a coarser level to observe if the selection procedure converges to discretizations at which all confounding could be controlled rather than the finest discretization. Our results have also shown that the true value of our parameter of interest may change across discretizations. This indicates that the interpretation of said parameter should be made with respect to the chosen discretization. Additional future work could be aimed at overcoming the need for *a priori* defined candidate discretizations.

Practically, this procedure may be challenging to implement with such data-adaptive methods as Super Learner due to computational intensity and innovations may be needed to adapt the procedure with such methods. Finally, given the initial selection of a discretization and the definition of the target parameter as a data-adaptive parameter, valid asymptotic inference for this estimator remains to be investigated. One solution may be to use sample splitting in order to select

a discretization and obtain estimates on separate data splits [**133**].

As for the data application, our selection procedure selected 3 week intervals as the optimal discretization of the options provided. The results could not conclude that a protective effect of low ICS treatment versus no ICS treatment on early delivery in women with mild asthma exists. However, given the absence of adjustment for important confounders such as smoking, body mass index, etc., these estimates remain potentially biased. While we evaluated most recent exposure these methods may be employed for other exposure measures such as cumulative exposure or the effects of exposure at multiple time-points. Most importantly, the different results across discretizations illustrate how the underlying data discretization may alter the final conclusion, and the need for transparent data discretization. Additionally, the fact that the two losses could potentially not always select the same discretization may also be a concern for further studies.

In summary, this paper serves as an introduction to the causal inference problem of arbitrary data discretization and proposes a data-adaptive solution to possible *ad hoc* analytic decisions. On that account, given the widespread usage of arbitrary discretization, more investigations are needed to evaluate the potential related problems. Data-adaptive approaches to data extraction from administrative data may provide statistical advantages and an unambiguous decision making procedure.

3.10. Acknowledgements

The authors would like to thank Joshua Schwab for the insightful comments and help in the use of pooled LTMLE. MS and SFG would also like to thank Miguel Hernán for his helpful comments on the manuscript.

3.11. Supplementary material

Appendix A – Pooled LTMLE code for two time-points

This appendix contains code that executes pooled LTMLE for the following structured data :

$$\boldsymbol{O} = (L(0), A(0), Y(4), L(4), A(4), Y(12))$$

where $L(t)$, $A(t)$, and $Y(t)$ are described as in the simulation study of section 3.7.

```
PLTMLE2TP = function(dat){


t = 2
d = 3 # number of regimes


#### Indicators of treatment ####


#t = 2


I2_11 = ifelse(is.na(dat$A0) | is.na(dat$A4) ,NA,
ifelse(dat$A0 == 1 & dat$A4 == 1 , 1, 0))


I2_01 = ifelse(is.na(dat$A0) | is.na(dat$A4) ,NA,
ifelse(dat$A0 == 0 & dat$A4 == 1 , 1, 0))


I2_00 = ifelse(is.na(dat$A0) | is.na(dat$A4) ,NA,
ifelse(dat$A0 == 0 & dat$A4 == 0 , 1, 0))
```

```
I2 = c(I2_11,I2_01,I2_00)

#t = 1

I1_11 = ifelse(is.na(dat$A0),NA,
ifelse(dat$A0 == 1, 1, 0))

I1_01 = ifelse(is.na(dat$A0) ,NA,
ifelse(dat$A0 == 0, 1, 0))

I1_00 = ifelse(is.na(dat$A0),NA,
ifelse(dat$A0 == 0, 1, 0))

I1 = c(I1_11,I1_01,I1_00)

###### Set regimes data ########

dat_11 = dat
dat_11[,grep("A", names(dat_11))] = 1

dat_01 = dat
dat_01[,grep("A", names(dat_01))] = c(rep(0,nrow(dat_01)),
rep(1,nrow(dat_01)))

dat_00 = dat
dat_00[,grep("A", names(dat_00))] = 0

dat_rep = rbind(dat_11,dat_01,dat_00)
```

```
#Step 1 - Estimation of g models

#A4
g1 = glm(A4 ~ L0 + A0 + L4 , family = binomial,data = dat)
g1_d = ifelse(dat_rep$A4 == 1, predict(g1,type = "response",
 newdata = dat_rep),
1-predict(g1,type = "response", newdata = dat_rep))

#A0
g0 = glm(A0 ~ L0, family = binomial,data = dat)
g0_d = ifelse(dat_rep$A0 == 1, predict(g0,type = "response",
 newdata = dat_rep),
1-predict(g0,type = "response", newdata = dat_rep))

#Cumulative g

g10_d = g1_d*g0_d


#Step 2 - Q22*

# a) Initial estimation of Q22

Q22 = glm(Y12 ~ L0 + A0 + L4 + A4, family = "binomial",data = dat)
dat_rep$Q22 = predict(Q22,type = "response",newdata = dat_rep)

# b) repeated copy of Y2 = Q32*
```

```
Y = dat_rep$Y12

# c) created weighted covariate

S1 = 1
S2 = 2
S3 = c(rep(1,nrow(dat)),rep(1,nrow(dat)),rep(0,nrow(dat)))
off = logit(dat_rep$Q22)

update_data_22 = as.data.frame(cbind(Y,S1,S2,S3,off))
colnames(update_data_22) = c("Y","S1","S2","S3","off")

update_data_22$I2 = I2
update_data_22$HS1 = (update_data_22$S1/g10_d)*update_data_22$I2
update_data_22$HS2 = (update_data_22$S2/g10_d)*update_data_22$I2
update_data_22$HS3 = (update_data_22$S3/g10_d)*update_data_22$I2

# d) Estimation of epsilon

eps_22 = glm(Y ~ -1 + HS1 + HS2 + HS3 + offset(off) , family=
quasibinomial,data = update_data_22,subset = (I2 == 1))

# e) Generation of Q22*

Q22_star = predict(eps_22, type = "response", newdata =
update_data_22)
Q22_star[which(is.na(Q22_star))] = 1
```

```
# See if solved EIF -->if not NA

score = (((Y - Q22_star)*cbind(S1,S2,S3))/g10_d)[I2==1,]
if(sum(abs(colSums(score,na.rm = T))) > 0.001){

CalcScore <- function(e) {
Q22_star = plogis(off + cbind(S1,S2,S3)%*% e)
Q22_star[which(is.na(Q22_star))] = 1
score = (((Y - Q22_star[,1])*cbind(S1,S2,S3))/g10_d)[I2==1,]
return(sum(abs(colSums(score,na.rm = T))))
#each column has to add to zero
}

FindMin <- function(minimizer) {
num.tries <- 20
init.e <- numeric(3) #first try an initial estimate of epsilon=0
for (i in 1:num.tries) {
m <- nlminb(start=init.e, objective=CalcScore, control=list(
abs.tol=0.0001, eval.max=500, iter.max=500, x.tol=1e-14, rel.tol=1e-14))
e <- m$par
obj.val <- m$objective
if (obj.val < 0.0001) {
m$ltmle.msg <- "updating step using glm failed to solve score equation;
 solved using nlminb"
return(list(e=e, solved=TRUE, m=m))
}
init.e <- rnorm(3)
```

```
#if the first try didn't work, try a random initial estimate of epsilon
}
return(list(e=numeric(3), solved=FALSE, m="score equation not solved!"))
#nocov - return Q (not updated)
}

mm = FindMin(nlminb)

if (mm$solved){
Q22_star = plogis(off + cbind(S1,S2,S3)%*% mm$e)[,1]
Q22_star[which(is.na(Q22_star))] = 1
}   else {
paramlist = list("g10_d" = g10_d, "g0_d" =  g0_d, "coeff" = NA)
return(paramlist)
stop("Did not converge")
}
}

### Step 3 - Q12*

# a)

dat$Q22_star_11 = Q22_star[1:nrow(dat)]
dat$Q22_star_01 = Q22_star[(nrow(dat) + 1):((d-1)*nrow(dat))]
dat$Q22_star_00 = Q22_star[((d-1)*nrow(dat) + 1):(d*nrow(dat))]

QQ12_11 = glm(Q22_star_11 ~ L0 + A0 , family = "quasibinomial",data = dat)
QQ12_01 = glm(Q22_star_01 ~ L0 + A0 , family = "quasibinomial",data = dat)
```

```
QQ12_00 = glm(Q22_star_00 ~ L0 + A0 , family = "quasibinomial",data = dat)

Q12_11 = predict(QQ12_11,type = "response",newdata = dat_11)
Q12_01 = predict(QQ12_01,type = "response",newdata = dat_01)
Q12_00 = predict(QQ12_00,type = "response",newdata = dat_00)

dat_rep$Q12 = c(Q12_11,Q12_01,Q12_00)

# b) Construction of weighted covariate

Y = Q22_star
S1 = 1
S2 = 2
S3 = c(rep(1,nrow(dat)),rep(1,nrow(dat)),rep(0,nrow(dat)))
off = logit(dat_rep$Q12)

update_data_12 = as.data.frame(cbind(Y,S1,S2,S3,off))
colnames(update_data_12) = c("Y","S1","S2","S3","off")

update_data_12$I1 = I1
update_data_12$HS1 = (update_data_12$S1/g0_d)*update_data_12$I1
update_data_12$HS2 = (update_data_12$S2/g0_d)*update_data_12$I1
update_data_12$HS3 = (update_data_12$S3/g0_d)*update_data_12$I1

# c) MLE estimation of epsilon

eps_12 = glm(Y ~ -1 + HS1 + HS2 + HS3 + offset(off) , family=
quasibinomial,data = update_data_12, subset = (I1 == 1))
```

```
# e)

Q12_star = predict(eps_12, type = "response", newdata =
update_data_12)

# See if solved EIF -->if not NA

score = (((Y - Q12_star)*cbind(S1,S2,S3))/g0_d)[I1==1,]
if(sum(abs(colSums(score,na.rm = T))) > 0.01){

CalcScore <- function(e) {
Q12_star = plogis(off + cbind(S1,S2,S3)%*% e)
Q12_star[which(is.na(Q12_star))] = 1
score = (((Y - Q12_star[,1])*cbind(S1,S2,S3))/g0_d)[I1==1,]
return(sum(abs(colSums(score,na.rm = T))))
#each column has to add to zero
}

FindMin <- function(minimizer) {
num.tries <- 20
init.e <- numeric(3) #first try an initial estimate of epsilon=0
for (i in 1:num.tries) {
m <- nlminb(start=init.e, objective=CalcScore, control=list(
abs.tol=0.0001, eval.max=500, iter.max=500, x.tol=1e-14, rel.tol=1e-14))
e <- m$par
obj.val <- m$objective
if (obj.val < 0.0001) {
```

```
m$ltmle.msg <- "updating step using glm failed to solve score equation;
solved using nlminb"
return(list(e=e, solved=TRUE, m=m))
}
init.e <- rnorm(3)
#if the first try didn't work, try a random initial estimate of epsilon
}
return(list(e=numeric(3), solved=FALSE, m="score equation not solved!"))
#nocov - return Q (not updated)
}

mm = FindMin(nlminb)

if (mm$solved){
Q12_star = plogis(off + cbind(S1,S2,S3)%*% mm$e)[,1]
Q12_star[which(is.na(Q12_star))] = 1
}   else {
paramlist = list("g10_d" = g10_d, "g0_d" =  g0_d, "coeff" = NA)
return(paramlist)
stop("Did not converge")}

}

### Step 4 ###

# a) Initial estimation of Q11

Q11 = glm(Y4 ~ L0 + A0, family = "binomial",data = dat)
```

```
dat_rep$Q11 = predict(Q11,type = "response",newdata = dat_rep)

# b)

Y = dat_rep$Y4

# c)

S1 = 1
S2 = 1
S3 = c(rep(1,nrow(dat)),rep(0,nrow(dat)*2))
off = logit(dat_rep$Q11)

update_data_11 = as.data.frame(cbind(Y,S1,S2,S3,off))
colnames(update_data_11) = c("Y","S1","S2","S3","off")

update_data_11$I1 = I1
update_data_11$HS1 = (update_data_11$S1/g0_d)*update_data_11$I1
update_data_11$HS2 = (update_data_11$S2/g0_d)*update_data_11$I1
update_data_11$HS3 = (update_data_11$S3/g0_d)*update_data_11$I1

# d)

eps_11 = glm(Y ~ -1 + HS1 + HS2 + HS3 + offset(off) , family=
quasibinomial,data = update_data_11, subset = (I1 == 1))

# e)
```

```
Q11_star = predict(eps_11, type = "response", newdata =
update_data_11)

# See if solved EIF -->if not NA

score = (((Y - Q11_star)*cbind(S1,S2,S3))/g0_d)[I1==1,]
if(sum(abs(colSums(score,na.rm = T))) > 0.01){

CalcScore <- function(e) {
Q11_star = plogis(off + cbind(S1,S2,S3)%*% e)
Q11_star[which(is.na(Q11_star))] = 1
score = (((Y - Q11_star[,1])*cbind(S1,S2,S3))/g0_d)[I1==1,]
return(sum(abs(colSums(score,na.rm = T))))
#each column has to add to zero
}

FindMin <- function(minimizer) {
num.tries <- 20
init.e <- numeric(3) #first try an initial estimate of epsilon=0
for (i in 1:num.tries) {
m <- nlminb(start=init.e, objective=CalcScore, control=list(
abs.tol=0.0001, eval.max=500, iter.max=500, x.tol=1e-14, rel.tol=1e-14))
e <- m$par
obj.val <- m$objective
if (obj.val < 0.0001) {
m$ltmle.msg <- "updating step using glm failed to solve score equation;
solved using nlminb"
return(list(e=e, solved=TRUE, m=m))
```

```
}
init.e <- rnorm(3)
#if the first try didn't work, try a random initial estimate of epsilon
}
return(list(e=numeric(3), solved=FALSE, m="score equation not solved!"))
#nocov - return Q (not updated)
}

mm = FindMin(nlminb)

if (mm$solved){
Q11_star = plogis(off + cbind(S1,S2,S3)%*% mm$e)[,1]
Q11_star[which(is.na(Q11_star))] = 1
}   else {
paramlist = list("g10_d" = g10_d, "g0_d" =  g0_d, "coeff" = NA)
return(paramlist)
stop("Did not converge")}

}

Q1_star = c(Q11_star, Q12_star)


### Step 5 - MSM

S1 = 1
S2 = c(rep(1,nrow(dat)*d), rep(2,nrow(dat)*d))
S3 = c(rep(1,nrow(dat)),rep(0,nrow(dat)*2),rep(1,nrow(dat)*(2)),
```

```
rep(0,nrow(dat)))

msm_data = as.data.frame(cbind(Q1_star,S1,S2,S3))
colnames(msm_data) = c("Q1_star","S1","S2","S3")

msm = glm(Q1_star ~ -1 + S1 + S2 + S3, family = "quasibinomial",
data = msm_data)
#summary(msm)

paramlist = list("g10_d" = g10_d, "g0_d" =  g0_d, "coeff" = coef(msm))

return(paramlist)

}
```

Chapitre 4

PRÉCISIONS MÉTHODOLOGIQUES

Ce chapitre se veut un complément du manuscrit principal. Il vise principalement à reprendre le contenu de certains sujets abordés dans le manuscrit du chapitre 3 et à y apporter certaines précisions. Il se divise en trois sections.

La première section se veut un exemple détaillé de la méthode de la PLTMLE employée dans ce mémoire. Plus particulièrement, elle sert de support à l'explication de la méthode de la PLTMLE, dont un algorithme de calcul détaillé est fourni dans le manuscrit.

La seconde section apporte des précisions quant à l'étude de simulation.

Finalement, la dernière section décrit principalement les sources de données et la constitution de la cohorte utilisée dans l'application.

4.1. Illustration de l'algorithme de la PLTMLE

Rappelons que le paramètre d'intérêt, $\boldsymbol{\psi}$, que nous cherchons à estimer, correspond aux coefficients du MSM de l'équation (1.3.5). Par ailleurs, s'il est d'intérêt d'estimer la probabilité contrefactuelle d'un évènement par temps t en fonction de, par exemple, l'exposition la plus récente à un certain médicament, nous pouvons définir le MSM :

$$\eta(\boldsymbol{\beta}, \bar{A}, t) \equiv logit[P(Y^{\bar{A}}(t) = 1 | Y^{\bar{A}}(t-1) = 0), t] = \beta_0 + \beta_1 A(t-1) + \beta_2 t,\ t = 1, \ldots K+1. \quad (4.1.1)$$

Rappelons aussi que ce paramètre d'intérêt peut être redéfini en l'équation (1.3.6), de telle sorte qu'il suffit d'estimer $\bar{Q}_t^{\bar{a}}(1)$ par l'intermédiaire d'espérances conditionnelles itérées.

ID	L(0)	A(0)	Y(1)	L(1)	A(1)	Y(2)
1	0	1	1	NA	NA	1
2	1	1	0	1	0	1
3	1	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	1	0	1	0	0

FIGURE 4.1. Données observées

Supposons que nous ayons une structure de données observées longitudinales de la forme $\boldsymbol{O} = (L(0), A(0), Y(1), L(1), A(1), Y(2))$, illustrée à la Figure 4.1, où toutes les données sont évidemment fictives. Rappelons que $Y(t)$ est un évènement de type survie. Ainsi $L(t)$ et $A(t)$ ne sont pas observées après la survenue d'un évènement (ici identifiées par la valeur NA). Supposons aussi qu'il soit d'intérêt d'estimer $\boldsymbol{\psi}$ pour tous les régimes d'exposition possibles, soit $\bar{a} \in \bar{\mathcal{A}} = \{(1,1), (0,1), (1,0), (0,0)\}$. Ceci conduit au jeu de données répétées, où l'historique d'exposition de chaque individu a été fixé à chaque régime d'exposition possible, illustré par la Figure 4.2

Pour ces données, le but de l'algorithme PLTMLE est alors d'estimer

$$\begin{aligned} \bar{Q}_2^{\bar{a}}(1) &= E[\bar{Q}_2^{\bar{a}}(2) \mid \bar{A}(0) = \bar{a}(0), \bar{L}(0)] \\ &= E[E[Y^{\bar{a}}(2) \mid \bar{A}(1) = \bar{a}(1), \bar{L}(1)] \mid \bar{A}(0) = \bar{a}(0), \bar{L}(0)] \end{aligned}$$

ID	L(0)	a(0)	Y(1)	L(1)	a(1)	Y(2)
1	0	1	1	NA	1	1
1	0	0	1	NA	1	1
1	0	1	1	NA	0	1
1	0	0	1	NA	0	1
2	1	1	0	1	1	1
2	1	0	0	1	1	1
2	1	1	0	1	0	1
2	1	0	0	1	0	1
3	1	1	0	0	1	0
3	1	0	0	0	1	0
3	1	1	0	0	0	0
3	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	1	0	1	1	0
n	0	0	0	1	1	0
n	0	1	0	1	0	0
n	0	0	0	1	0	0

FIGURE 4.2. Données répétées

et

$$\bar{Q}_1^{\bar{a}}(1) \quad = \quad E[Y^{\bar{a}}(1) \mid \bar{A}(0) = \bar{a}(0), \bar{L}(0)].$$

Dans cet exemple, nous nous concentrerons sur l'illustration de l'estimation de $\bar{Q}_2^{\bar{a}}(1)$ (pour $t = 2$), qui est facilement généralisable pour $\bar{Q}_1^{\bar{a}}(1)$ ($t = 1$).

Algorithme PLTMLE pour l'estimation de $\bar{Q}_2^{\bar{a}}(1)$ (pour $t = 2$) :

1) **Estimation initiale de $\bar{Q}_2^{\bar{a}}(2)$:**

   $\bar{Q}_2^{\bar{a}}(2)$ peut être estimée en effectuant une régression logistique de $Y(2)$ sur $L(1), A(1), L(0)$ et $A(0)$ à partir des données observées. Ensuite, prédire $\bar{Q}_{2,n}^{\bar{a}}(2) = E_n[Y(2)|L(1), \bar{A}(1) = \bar{a}, L(0)]$ pour chaque individu et chaque $\bar{a}$ (correspondant à une ligne du jeu de données répétées de la Figure 4.2, constitué d'une observation par individu par valeur fixée de chaque régime d'exposition). Il est important de noter que pour un contexte de survie,

$E[Y(2)|Y(1) = 1] = 1$. Ainsi, les valeurs ne sont prédites que pour les sujets dont $Y(1) = 0$.

2) **Mise à jour de $\bar{Q}^{\bar{a}}_{2,n}(2)$ vers $\bar{Q}^{\bar{a},*}_{2,n}(2)$:**

Rappelons que la mise à jour se fait en estimant le paramètre $\boldsymbol{\epsilon}$ du modèle (1.3.10). Ce modèle requiert l'estimation des composantes $\bar{g}^{\bar{a}}(t) = \prod_{k=0}^{t} P(A(k) = a(k)|\bar{L}(k), \bar{Y}(k), \bar{A}(k-1) = \bar{a}(k-1))$. Dans cet exemple, une estimation de $\bar{g}^{\bar{a}}(2) = P(A(0) = a(0)|L(0)) \cdot P(A(1) = a(1)|L(1), Y(1), A(0) = a(0), L(0))$ peut être obtenue en estimant chaque modèle de probabilité à partir des données observées et en prédisant $g^{\bar{a}}_n(1)$ pour chaque $\bar{a}$ à l'aide du jeu de données répétées.

Ce modèle requiert également le calcul de $\frac{\partial}{\partial \boldsymbol{\beta}} f[\eta(\boldsymbol{\beta}, \bar{A}, t)]\Big|_{\bar{A}(t-1)=\bar{a}(t-1)} = \frac{\partial}{\partial \boldsymbol{\beta}}(\beta_0 + \beta_1 A(t-1) + \beta_2 t) \Big|_{\bar{A}(t-1)=\bar{a}(t-1)}$ et de $I(\bar{A}(t-1) = \bar{a}(t-1))$. Pour le MSM (4.1.1), cette dérivée est égale à $(1, A(t-1), t)$ pour la dérivée par rapport à $(\beta_0, \beta_1, \beta_2)$. Dans cet exemple (pour $t = 2$), $(1, A(t-1), t)\Big|_{\bar{A}(t-1)=\bar{a}(t-1)} = (1, a(1), 2)$. Tous ces éléments sont alors regroupés dans un jeu de données utilisé pour calculer la mise à jour de $\bar{Q}^{\bar{a}}_{2,n}(2)$ vers $\bar{Q}^{\bar{a},*}_{2,n}(2)$, illustré dans la Figure 4.3. Finalement, l'estimateur $\boldsymbol{\epsilon}_n$ de $\boldsymbol{\epsilon}$ est calculé en effectuant, sur les données groupées, une seule régression logistique groupée de $Y(2)$ sur $\frac{I(\bar{A}(1) = \bar{a}(1))}{g^{\bar{a}}_n(1)}(1, a(1), 2)$ avec décalage $logit(\bar{Q}^{\bar{a}}_{2,n}(2))$. $\bar{Q}^{\bar{a},*}_{2,n}(2)$ peut être prédit pour chaque $\bar{a}$ du jeu de données groupées par $\bar{Q}^{\bar{a},*}_{2,n}(2) = expit(logit(\bar{Q}^{\bar{a}}_{2,n}(2)) + \boldsymbol{\epsilon}_n \frac{I(\bar{A}(1) = \bar{a}(1))}{g^{\bar{a}}_n(1)}(1, a(1), 2))$. Il est à noter que pour les individus ayant $Y(1) = 1$ on fixe $\bar{Q}^{\bar{a},*}_{2,n}(2) = 1$.

3) **Estimation initiale de $\bar{Q}^{\bar{a}}_2(1)$:**

Une fois $\bar{Q}^{\bar{a},*}_{2,n}(2)$ estimé, une estimation initiale de $\bar{Q}^{\bar{a}}_2(1)$ pour chaque $\bar{a}$ séparément est donné par $\bar{Q}^{\bar{a}}_{2,n}(1) = E_n[\bar{Q}^{\bar{a},*}_{2,n}(2)|A(0) = a(0), L(0)]$. Ceci peut être estimé en effectuant une régression de $\bar{Q}^{\bar{a},*}_{2,n}(2)$ sur $A(0)$ et $L(0)$ et prédire la variable dépendante pour chaque $\bar{a}$ séparément selon les jeux de données simples présentés à la Figure 4.4.

ID	L(0)	$a(0)$	Y(1)	L(1)	$a(1)$	Y(2)	$Q^{a}_{2,n}(2)$	$g^{\bar{a}}_{n}(1)$	$\frac{d}{d\beta_0}f[\eta(\boldsymbol{\beta},\bar{A},t)]$	$\frac{d}{d\beta_1}f[\eta(\boldsymbol{\beta},\bar{A},t)]$	$\frac{d}{d\beta_2}f[\eta(\boldsymbol{\beta},\bar{A},t)]$	$I(\bar{A}(1)=\bar{a}(1))$
1	0	1	1	NA	1	1						0
1	0	0	1	NA	1	1						0
1	0	1	1	NA	0	1						0
1	0	0	1	NA	0	1						0
2	1	1	0	1	1	1	$\bar{Q}^{(1,1)}_{2,n}(2)$	$g^{(1,1)}_{n}(1)$	1	1	2	0
2	1	0	0	1	1	1	$\bar{Q}^{(0,1)}_{2,n}(2)$	$g^{(0,1)}_{n}(1)$	1	1	2	0
2	1	1	0	1	0	1	$\bar{Q}^{(1,0)}_{2,n}(2)$	$g^{(1,0)}_{n}(1)$	1	0	2	1
2	1	0	0	1	0	1	$\bar{Q}^{(0,0)}_{2,n}(2)$	$g^{(0,0)}_{n}(1)$	1	0	2	0
3	1	1	0	0	1	0	$\bar{Q}^{(1,1)}_{2,n}(2)$	$g^{(1,1)}_{n}(1)$	1	1	2	0
3	1	0	0	0	1	0	$\bar{Q}^{(0,1)}_{2,n}(2)$	$g^{(0,1)}_{n}(1)$	1	1	2	1
3	1	1	0	0	0	0	$\bar{Q}^{(1,0)}_{2,n}(2)$	$g^{(1,0)}_{n}(1)$	1	0	2	0
3	1	0	0	0	0	0	$\bar{Q}^{(0,0)}_{2,n}(2)$	$g^{(0,0)}_{n}(1)$	1	0	2	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮						
n	0	1	0	1	1	0	$\bar{Q}^{(1,1)}_{2,n}(2)$	$g^{(1,1)}_{n}(1)$	1	1	2	0
n	0	0	0	1	1	0	$\bar{Q}^{(0,1)}_{2,n}(2)$	$g^{(0,1)}_{n}(1)$	1	1	2	0
n	0	1	0	1	0	0	$\bar{Q}^{(1,0)}_{2,n}(2)$	$g^{(1,0)}_{n}(1)$	1	0	2	1
n	0	0	0	1	0	0	$\bar{Q}^{(0,0)}_{2,n}(2)$	$g^{(0,0)}_{n}(1)$	1	0	2	0

FIGURE 4.3. Données groupées pour la mise à jour de $\bar{Q}^{\bar{a}}_{2,n}(2)$

ID	L(0)	a(0)	Y(1)	L(1)	a(1)	Y(2)	$\bar{Q}^{\bar{a},*}_{2,n}(2)$
1	0	1	1	NA	1	1	$\bar{Q}^{(1,1),*}_{2,n}(2)$
2	1	1	0	1	1	1	$\bar{Q}^{(1,1),*}_{2,n}(2)$
3	1	1	0	0	1	0	$\bar{Q}^{(1,1),*}_{2,n}(2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	1	0	1	1	0	$\bar{Q}^{(1,1),*}_{2,n}(2)$

$\bar{Q}^{(1,1)}_{2,n}(1)$

ID	L(0)	a(0)	Y(1)	L(1)	a(1)	Y(2)	$\bar{Q}^{\bar{a},*}_{2,n}(2)$
1	0	0	1	NA	1	1	$\bar{Q}^{(0,1),*}_{2,n}(2)$
2	1	0	0	1	1	1	$\bar{Q}^{(0,1),*}_{2,n}(2)$
3	1	0	0	0	1	0	$\bar{Q}^{(0,1),*}_{2,n}(2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	0	0	1	1	0	$\bar{Q}^{(0,1),*}_{2,n}(2)$

$\bar{Q}^{(0,1)}_{2,n}(1)$

ID	L(0)	a(0)	Y(1)	L(1)	a(1)	Y(2)	$\bar{Q}^{\bar{a},*}_{2,n}(2)$
1	0	1	1	NA	0	1	$\bar{Q}^{(1,0),*}_{2,n}(2)$
2	1	1	0	1	0	1	$\bar{Q}^{(1,0),*}_{2,n}(2)$
3	1	1	0	0	0	0	$\bar{Q}^{(1,0),*}_{2,n}(2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	1	0	1	0	0	$\bar{Q}^{(1,0),*}_{2,n}(2)$

$\bar{Q}^{(1,0)}_{2,n}(1)$

ID	L(0)	a(0)	Y(1)	L(1)	a(1)	Y(2)	$\bar{Q}^{\bar{a},*}_{2,n}(2)$
1	0	0	1	NA	0	1	$\bar{Q}^{(0,0),*}_{2,n}(2)$
2	1	0	0	1	0	1	$\bar{Q}^{(0,0),*}_{2,n}(2)$
3	1	0	0	0	0	0	$\bar{Q}^{(0,0),*}_{2,n}(2)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	0	0	0	1	0	0	$\bar{Q}^{(0,0),*}_{2,n}(2)$

$\bar{Q}^{(0,0)}_{2,n}(1)$

FIGURE 4.4. Données pour l'estimation de $\bar{Q}^{\bar{a}}_{2,n}(1)$

4) **Mise à jour de $\bar{Q}^{\bar{a}}_{2,n}(1)$ vers $\bar{Q}^{\bar{a},*}_{2,n}(1)$:**

La mise à jour est effectuée de façon analogue à l'étape de mise à jour décrite précédemment. $\bar{g}^{\bar{a}}(0) = P(A(0) = a(0)|L(0))$ est estimé sur le jeu de données observées et prédit $g^{\bar{a}}_{n}(0)$ pour chaque $\bar{a}$ sur le jeu de données répétées.

Toujours pour $t = 2$, $\frac{\partial}{\partial\boldsymbol{\beta}}f[\eta(\boldsymbol{\beta},\bar{A},t)]\Big|_{\bar{A}(t-1)=\bar{a}(t-1)} = (1, a(1), 2)$. L'estimateur $\boldsymbol{\epsilon}_n$ de $\boldsymbol{\epsilon}$ est calculé en effectuant, sur les données groupées décrites à la Figure 4.5, une seule régression logistique groupée de $\bar{Q}^{\bar{a}}_{2,n}(1)$ sur $\frac{I(\bar{A}(0)=\bar{a}(0))}{g^{\bar{a}}_{n}(0)}(1, a(1), 2)$ avec décalage $logit(\bar{Q}^{\bar{a}}_{2,n}(1))$. $\bar{Q}^{\bar{a},*}_{2,n}(1)$ est prédit pour chaque $\bar{a}$ sur le jeu de données groupées selon $\bar{Q}^{\bar{a},*}_{2,n}(1) = expit($

ID	L(0)	a(0)	Y(1)	L(1)	a(1)	Y(2)	$\bar{Q}^{\bar{a}}_{2,n}(1)$	$g^{\bar{a}}_n(0)$	$\frac{d}{d\beta_0} f[\eta(\boldsymbol{\beta}, \bar{A}, t)]$	$\frac{d}{d\beta_1} f[\eta(\boldsymbol{\beta}, \bar{A}, t)]$	$\frac{d}{d\beta_2} f[\eta(\boldsymbol{\beta}, \bar{A}, t)]$	$I(\bar{A}(0) = \bar{a}(0))$
1	0	1	1	NA	1	1	$\bar{Q}^{(1,1)}_{2,n}(1)$	$g^{(1,1)}_n(0)$	1	1	2	1
1	0	0	1	NA	1	1	$\bar{Q}^{(0,1)}_{2,n}(1)$	$g^{(0,1)}_n(0)$	1	1	2	0
1	0	1	1	NA	0	1	$\bar{Q}^{(1,0)}_{2,n}(1)$	$g^{(1,0)}_n(0)$	1	0	2	1
1	0	0	1	NA	0	1	$\bar{Q}^{(0,0)}_{2,n}(1)$	$g^{(0,0)}_n(0)$	1	0	2	0
2	1	1	0	1	1	1	$\bar{Q}^{(1,1)}_{2,n}(1)$	$g^{(1,1)}_n(0)$	1	1	2	1
2	1	0	0	1	1	1	$\bar{Q}^{(0,1)}_{2,n}(1)$	$g^{(0,1)}_n(0)$	1	1	2	0
2	1	1	0	1	0	1	$\bar{Q}^{(1,0)}_{2,n}(1)$	$g^{(1,0)}_n(0)$	1	0	2	1
2	1	0	0	1	0	1	$\bar{Q}^{(0,0)}_{2,n}(1)$	$g^{(0,0)}_n(0)$	1	0	2	0
3	1	1	0	0	1	0	$\bar{Q}^{(1,1)}_{2,n}(1)$	$g^{(1,1)}_n(0)$	1	1	2	0
3	1	0	0	0	1	0	$\bar{Q}^{(0,1)}_{2,n}(1)$	$g^{(0,1)}_n(0)$	1	1	2	1
3	1	1	0	0	0	0	$\bar{Q}^{(1,0)}_{2,n}(1)$	$g^{(1,0)}_n(0)$	1	0	2	0
3	1	0	0	0	0	0	$\bar{Q}^{(0,0)}_{2,n}(1)$	$g^{(0,0)}_n(0)$	1	0	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮						
n	0	1	0	1	1	0	$\bar{Q}^{(1,1)}_{2,n}(1)$	$g^{(1,1)}_n(0)$	1	1	2	1
n	0	0	0	1	1	0	$\bar{Q}^{(0,1)}_{2,n}(1)$	$g^{(0,1)}_n(0)$	1	1	2	0
n	0	1	0	1	0	0	$\bar{Q}^{(1,0)}_{2,n}(1)$	$g^{(1,0)}_n(0)$	1	0	2	1
n	0	0	0	1	0	0	$\bar{Q}^{(0,0)}_{2,n}(1)$	$g^{(0,0)}_n(0)$	1	0	2	0

FIGURE 4.5. Données groupées pour la mise à jour de $\bar{Q}^{\bar{a}}_{2,n}(1)$

$logit(\bar{Q}^{\bar{a}}_{2,n}(1)) + \boldsymbol{\epsilon}_n \frac{I(\bar{A}(0) = \bar{a}(0))}{g^{\bar{a}}_n(0)}(1, a(1), 2))$. On obtient ainsi $\bar{Q}^{\bar{a},*}_{2,n}(1)$ pour chaque individu pour chaque régime d'exposition.

L'algorithme ci-haut est établi de façon similaire pour l'obtention de $\bar{Q}^{\bar{a},*}_{1,n}(1)$, auquel cas nous obtenons $\bar{Q}^{\bar{a},*}_{1,n}(1)$ pour chaque individu pour chaque régime d'exposition.

5) Étape finale : Estimation du paramètre d'intérêt

Les paramètres du MSM sont estimés en effectuant la régression de $\bar{Q}^{\bar{a},*}_{2,n}(1)$ et $\bar{Q}^{\bar{a},*}_{1,n}(1)$ par rapport à $A(t-1) = a(t-1)$ et t, conformément à la spécification linéaire du MSM. Ceci est exécuté à l'aide des données groupées pour l'estimation du MSM, présentées à la Figure 4.6.

4.2. SIMULATION

Dans cette section, nous introduisons brièvement l'étude de simulation réalisée et apportons quelques précisions supplémentaires à l'information contenue

ID	$\bar{Q}_{t,n}^{\bar{a},*}(1)$	$a(t-1)$	t
1	$\bar{Q}_{2,n}^{(1,1),*}(1)$	1	2
1	$\bar{Q}_{2,n}^{(0,1),*}(1)$	1	2
1	$\bar{Q}_{2,n}^{(1,0),*}(1)$	0	2
1	$\bar{Q}_{2,n}^{(0,0),*}(1)$	0	2
1	$\bar{Q}_{1,n}^{(1,1),*}(1)$	1	1
1	$\bar{Q}_{1,n}^{(0,1),*}(1)$	1	1
1	$\bar{Q}_{1,n}^{(1,0),*}(1)$	0	1
1	$\bar{Q}_{1,n}^{(0,0),*}(1)$	0	1
2	$\bar{Q}_{2,n}^{(1,1),*}(1)$	1	2
2	$\bar{Q}_{2,n}^{(0,1),*}(1)$	1	2
2	$\bar{Q}_{2,n}^{(1,0),*}(1)$	0	2
2	$\bar{Q}_{2,n}^{(0,0),*}(1)$	0	2
2	$\bar{Q}_{1,n}^{(1,1),*}(1)$	1	1
2	$\bar{Q}_{1,n}^{(0,1),*}(1)$	1	1
2	$\bar{Q}_{1,n}^{(1,0),*}(1)$	0	1
2	$\bar{Q}_{1,n}^{(0,0),*}(1)$	0	1
3	$\bar{Q}_{2,n}^{(1,1),*}(1)$	1	2
3	$\bar{Q}_{2,n}^{(0,1),*}(1)$	1	2
3	$\bar{Q}_{2,n}^{(1,0),*}(1)$	0	2
3	$\bar{Q}_{2,n}^{(0,0),*}(1)$	0	2
3	$\bar{Q}_{1,n}^{(1,1),*}(1)$	1	1
3	$\bar{Q}_{1,n}^{(0,1),*}(1)$	1	1
3	$\bar{Q}_{1,n}^{(1,0),*}(1)$	0	1
3	$\bar{Q}_{1,n}^{(0,0),*}(1)$	0	1
⋮	⋮	⋮	⋮
n	$\bar{Q}_{2,n}^{(1,1),*}(1)$	1	2
n	$\bar{Q}_{2,n}^{(0,1),*}(1)$	1	2
n	$\bar{Q}_{2,n}^{(1,0),*}(1)$	0	2
n	$\bar{Q}_{2,n}^{(0,0),*}(1)$	0	2
n	$\bar{Q}_{1,n}^{(1,1),*}(1)$	1	1
n	$\bar{Q}_{1,n}^{(0,1),*}(1)$	1	1
n	$\bar{Q}_{1,n}^{(1,0),*}(1)$	0	1
n	$\bar{Q}_{1,n}^{(0,0),*}(1)$	0	1

FIGURE 4.6. Données groupées pour l'estimation du MSM

dans le manuscrit.

Les données simulées ont été générées afin de représenter des données avec de nombreux points dans le temps et où l'on retrouve des problèmes de quasi-violations pratiques de positivité, qui sont atténués par une discrétisation des données. Les données sont de la forme :

$$O_i = (L_i(0), A_i(0), Y_i(1), L_i(1), A_i(1), Y_i(2), \ldots, L_i(11), A_i(11), Y_i(12))$$

où $L(t)$ est une variable confondante binaire, $A(t)$ une variable d'exposition binaire et $Y(t+1)$ un indicateur monotone d'évènement, générés pour $t = 0, \ldots, 11$, dans cet ordre. La génération des données est entièrement décrite dans le manuscrit.

Le paramètre d'intérêt est encore une fois défini comme les coefficients du MSM (4.1.1). La vraie valeur du paramètre d'intérêt à la discrétisation la plus fine a été calculée numériquement. L'évaluation numérique de la vraie valeur des coefficients d'un MSM consiste à générer un très large nombre de données contrefactuelles et à évaluer le MSM sur ces données contrefactuelles. Les données contrefactuelles sont générées en conservant les mêmes mécanismes de génération de données pour l'issue et les covariables, mais où l'exposition est fixée à des valeurs correspondant aux expositions d'intérêt. Dans cette simulation, nous avons généré 25 échantillons contrefactuels de taille n = 1 000 000. Les vraies valeurs des paramètres d'intérêt n'ayant pas changé jusqu'à la deuxième décimale pour chacun des 25 échantillons, nous avons utilisé la moyenne de ces valeurs. Nous avons également calculé les vraies valeurs pour chaque discrétisation en utilisant la même approche.

Finalement, comme il n'était pas d'intérêt d'évaluer la performance de la PLTMLE, déjà évaluée par [**8**], tous les modèles ont été bien spécifiés dans l'algorithme de la PLTMLE en fonction de la génération des données, limitant ainsi d'autres biais possibles dus à une spécification erronée.

4.3. Application

Dans cette application, on s'intéresse à l'effet de médicaments antiasthmatiques sur la durée de gestation. Plus particulièrement, on s'intéresse à l'effet de la prise d'ICS chez des femmes dont l'asthme est léger durant l'année précédant la grossesse.

4.3.1. Source et extraction des données

Cette étude a été menée en utilisant des bases de données de la RAMQ qui sont des registres de soins de santé de personnes assurées par les régimes publics d'assurance maladie et d'assurance médicaments. Cela permet de créer un suivi longitudinal individuel reconstitué en liant les diverses bases de données décrites à la section 1.2.2. Ces bases de données contiennent des renseignements détaillés sur les prescriptions de médicaments servies et renouvelées, des renseignements sur les visites médicales, dont le diagnostic, les procédures effectuées et la date de visite, des renseignements sur toutes les hospitalisations ayant lieu au Québec et des renseignements démographiques, comme l'âge, la période de couverture et la prestation d'aide sociale.

Il est à noter que dans ces bases de données, ces renseignements peuvent varier dans le temps et que la fréquence à laquelle ces renseignements varient diffère d'un individu à l'autre. La fréquence la plus fine à laquelle il est possible de détecter

un changement à partir des données est d'une journée à l'autre. Ainsi, une extraction de données longitudinale reconstituant un suivi journalier des individus a été effectuée.

4.3.2. Population à l'étude et conception de la cohorte

Une étude de cohorte de grossesses de femmes asthmatiques identifiées à partir des BDA de la RAMQ a été réalisée. Pour être admissibles à l'étude, les mères devaient être assurées de façon continue par la RAMQ un an avant la grossesse jusqu'à la date d'accouchement. Les mères sont dites asthmatiques pour une grossesse en particulier, si elles avaient au moins un diagnostic d'asthme et au moins un renouvellement de médicaments contre l'asthme pendant l'année précédant la grossesse jusqu'à l'accouchement. Le diagnostic d'asthme a été identifié à l'aide des codes de diagnostic ICD-9 et ICD-10 correspondants.

Les femmes enceintes et asthmatiques âgées de moins de 45 ans avec au moins un accouchement entre le 1er novembre 1998 et le 31 décembre 2008 ont été incluses. Un accouchement comprenait toutes les naissances vivantes simples (à terme et prématurées) et les mortinaissances. Une femme ayant eu plusieurs accouchements au cours de la période d'étude, ne contribue qu'avec les deux premiers. Il se pourrait cependant qu'ils n'aient pas été les premiers accouchements de la femme.

En accord avec notre question de recherche, seules les femmes ayant un asthme léger durant l'année avant la grossesse ont été incluses. La sévérité de l'asthme est basée sur un algorithme validé développé et employé précédemment par notre groupe de recherche [**131**].

4.3.3. Définition de l'exposition aux ICS et de la durée de gestation

La variable d'exposition comporte deux catégories d'intérêt soit des doses nulles ou faibles d'ICS ($\leq$ 250 μg en équivalence de fluticasone). La quantité d'ICS a été définie en suivant un algorithme précédemment développé dans notre groupe de recherche [**32**].

En cas de non-adhésion à une des deux expositions d'intérêt, les femmes ont été censurées. Les cas où les femmes ont été censurées sont si elles passent à des doses modérées ou hautes d'ICS (plus de 250μg de fluticasone) ou lors de l'ajout d'un LABA aux ICS.

L'issue d'intérêt est un indicateur d'accouchement. Cette mesure a été obtenue à partir de la durée de gestation établie dans une étude précédente menée par notre groupe de recherche [**32**]. La durée de gestation y avait été extraite du dossier MED-ECHO de la mère. En cas de valeur manquante pour la mère, la durée de gestation avait été extraite du dossier du bébé.

4.3.4. Liste des variables confondantes

Nous reprenons les mêmes variables confondantes identifiées dans [**32**] qui se divisent en quatre catégories : les caractéristiques de la mère, les pathologies chroniques de la mère, les pathologies liées à la grossesse et les variables liées à l'asthme maternel. Ces variables peuvent être retrouvées dans le Tableau 4.1.

Pour cette analyse, en raison de problèmes de positivité pratique, les variables «Syndrome antiphospholipide » et «Maladie cardiaque cyanotique » ont été regroupées en «Autres maladies chroniques». De même, «Hypertension gestationnelle»

Tableau 4.1. Tableau des variables confondantes

Caractéristiques de la mère
Âge au début de la grossesse
Prestataire de la sécurité du revenu
Lieu de résidence
Pathologies chroniques de la mère
Hypertension chronique
Diabète mellitus
Syndrome antiphospholipide
Maladie cardiaque cyanotique
Désordres utérins
Pathologies liées à la grossesse
Diabète gestationnel
Hypertension gestationnelle
Éclampsie / pré-éclampsie
Infection de la mère
Anémie
Saignements vaginaux
Hémorragie foeto-maternelle
Complications placentaires
Décollement placentaire
Prise de bêta-bloqueur
Variables liées à l'asthme
Contrôle de l'asthme dans l'année précédant la grossesse
SABA
LTRA
OCS
NCS
Visite à l'urgence
Hospitalisation

et «Éclampsie / pré-éclampsie» ont été regroupées en une seule variable et «Décollement placentaire » a été incluse dans «Complications placentaires » . Finalement, «Anémie » , «Saignements vaginaux » , «Hémorragie foeto-maternelle » et «Prise de bêta-bloqueur » , ont été groupées en «Autres pathologies liées à la

grossesse » .

Toutes les variables sauf les caractéristiques de la mère et le contrôle de l'asthme durant l'année précédant la grossesse sont des variables qui varient dans le temps. Les pathologies chroniques de la mère sont mesurées pendant toute la période de suivi et sont des variables monotones, c'est-à-dire qu'une fois que les femmes sont atteintes de cette pathologie, elles restent atteintes. Les pathologies liées à la grossesse sont mesurées durant la grossesse seulement. Parmi celles-ci seules le diabète gestationnel et l'hypertension gestationnelle sont monotones. Un facteur de confusion variant dans le temps important dans la relation entre le traitement de l'asthme et la durée de gestation est le contrôle de l'asthme. Un algorithme validé défini précédemment dans notre groupe de recherche [**131**] nous a permis de calculer le contrôle de l'asthme dans l'année précédant la grossesse. Durant la grossesse, le contrôle de l'asthme est plutôt identifié au moyen des proxys suivants : la dose de SABA, la prise de LTRA, OCS, NCS, les visites à l'urgence pour l'asthme et les hospitalisations pour asthme.

Chapitre 5

CONCLUSION

Dans ce mémoire, nous avons abordé la discrétisation de la ligne du temps pour des méthodes longitudinales d'inférence causale et avons proposé une procédure de sélection de la discrétisation optimale afin de répondre à la discrétisation arbitraire. Plus précisément, nous nous sommes intéressés à l'estimation des paramètres de modèles structurels marginaux par estimation par maximum de vraisemblance ciblée longitudinale groupée. Comme vu au chapitre 1, l'utilisation de la PLTMLE est justifiée par ses nombreuses propriétés attrayantes. Ce chapitre offre donc un résumé des principaux résultats obtenus et de leurs implications, ainsi que de possibles avenues de recherche futures afin de venir complémenter les résultats obtenus.

5.1. Résumé des résultats principaux et implications

Dans le manuscrit du chapitre 3, nous avons tout d'abord proposé une définition claire de la discrétisation et d'un jeu de données discrétisé. Ensuite, nous avons défini notre paramètre d'intérêt en vertu de la discrétisation choisie comme un paramètre dit adaptable aux données. Ceci illustre l'importance de bien comprendre et d'identifier avec soin le paramètre d'intérêt qui est estimé dans l'analyse. Par la suite, nous avons établi les hypothèses d'identifiabilité pour

des données longitudinales discrétisées. Nous avons détaillé comment ces hypothèses pouvaient être affectées par la discrétisation et dans quels contextes elles pourraient ne pas être respectées. La recherche proposée dans ce mémoire est donc une des premières publications à aborder et à démontrer les problèmes pouvant résulter de la discrétisation arbitraire.

Puis, en plus de présenter un algorithme détaillé de la PLTMLE, nous avons développé une procédure de sélection de la discrétisation optimale de la ligne du temps. Cette procédure est basée sur la minimisation d'une fonction de perte relative à la LTMLE groupée et est calculée par validation croisée.

Nous avons réalisé une étude de simulation dont les résultats montrèrent qu'à mesure que la taille de l'échantillon augmentait, la procédure de sélection convergeait vers la vraie valeur du paramètre d'intérêt à la discrétisation la plus fine, c'est-à-dire la discrétisation où le contrôle de la confusion était adéquat. De plus, ces résultats ont montré que la vraie valeur du paramètre que l'on cherche à estimer peut varier en fonction de la discrétisation choisie, confirmant ainsi la nature adaptable aux données du paramètre d'intérêt. Cette étude de simulation a démontré que la procédure proposée offre une meilleure alternative à la discrétisation arbitraire de la ligne du temps, qui est la méthode communément employée dans la littérature.

Nous avons par ailleurs présenté un exemple appliqué qui requérait une discrétisation de la ligne du temps. Dans cette application, les données provenaient de bases de données administratives et avaient été extraites de manière à reconstituer un suivi quotidien pour chaque grossesse de femmes enceintes asthmatiques. Vu la nature de ces données, pouvant provoquer des problèmes de données rares, une discrétisation plus grossière que des données quotidiennes était nécessaire. La

procédure de sélection de la discrétisation a sélectionné des intervalles de trois semaines comme discrétisation optimale parmi les options fournies, soit des intervalles de trois, quatre, cinq et six semaines. Ainsi, les résultats finaux n'ont pas démontré un effet protecteur significatif d'une faible dose d'ICS par rapport à aucune dose d'ICS sur la durée de gestation chez les femmes souffrant d'asthme léger. L'application servit à démontrer l'utilité de la procédure de sélection lors de l'analyse en présence de données devant être discrétisées, car la discrétisation optimale est choisie indépendamment du résultat final.

5.2. Recherche future

Bien que la recherche présentée dans ce mémoire mette en lumière le problème de la discrétisation de la ligne du temps pour les méthodes d'inférence causale, plusieurs avenues de recherche future restent à être explorées. Si bien que les résultats présentent certaines limites. Par exemple, il serait intéressant de développer une procédure de sélection automatique de la discrétisation qui ne nécessite pas de fournir des discrétisations candidates. D'autres travaux pourraient impliquer l'optimisation du choix des discrétisations candidates, dans le but d'aider les analystes à choisir quelles discrétisations candidates pourraient être les plus appropriées pour différents types de données. De plus, les résultats de cette procédure demeurent inconnus sous différents scénarios de simulation tels que des modèles de confusion générés à diverses discrétisations, différentes méthodes d'estimation ou d'autres paramètres d'intérêt. Par ailleurs, il serait intéressant d'évaluer différentes fonctions de pertes. Entre autres, l'utilisation de l'EIC comme fonction de perte [**120**] devrait permettre à la procédure de converger vers la discrétisation où le contrôle de la confusion est adéquat.

Finalement, à plus grande échelle, le développement et l'utilisation plus répandue de méthodes telles que la nôtre pourraient faire place à de la recherche plus transparente et à une amélioration de la pratique statistique.

Bibliographie

[1] Hernán MA. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health* 2004 ; **58**(4) :265–271.

[2] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 1974 ; **66**(5) :688.

[3] Goodrum LA, Hankins GD, Jermain D, Chanaud CM. Conference report : complex clinical, legal, and ethical issues of pregnant and postpartum women as subjects in clinical trials. *Journal of Women's Health* 2003 ; **12**(9) :857–867.

[4] Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 1986 ; **7**(9-12) :1393–1512.

[5] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology 2000.

[6] Gruber S, van der Laan MJ, *et al.*. Targeted minimum loss based estimation of a causal effect on an outcome with known conditional bounds. *The International Journal of Biostatistics* 2012 ; **8**(1) :1–18.

[7] Schnitzer ME, Moodie EE, van der Laan MJ, Platt RW, Klein MB. Modeling the impact of hepatitis c viral clearance on end-stage liver disease in an hiv co-infected cohort with targeted maximum likelihood estimation. *Biometrics* 2014 ; **70**(1) :144–152.

[8] Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference* 2014 ; **2**(2) :147–185.

[9] WHO. World health organization. Http ://www.who.int/respiratory/asthma/.

[10] NHLBI. National heart, lung and blood institute. Http ://www.nhlbi.nih.gov/health/health-topics/topics/asthma.

[11] Lougheed MD, Lemiere C, Ducharme FM, Licskai C, Dell SD, Rowe BH, FitzGerald M, Leigh R, Watson W, Boulet LP. Canadian thoracic society 2012 guideline update : diagnosis and management of asthma in preschoolers, children and adults. *Canadian respiratory journal* 2012 ; **19**(2) :127–164.

[12] CANSIM. Statistique canada - tableau 105-0501, catalogue no. 82-221-x.

[13] Kwon HL, Belanger K, Bracken MB. Asthma prevalence among pregnant and childbearing-aged women in the united states : estimates from national health surveys. *Annals of epidemiology* 2003 ; **13**(5) :317–324.

[14] Kwon HL, Triche EW, Belanger K, Bracken MB. The epidemiology of asthma during pregnancy : prevalence, diagnosis, and symptoms. *Immunology and allergy clinics of North America* 2006 ; **26**(1) :29–62.

[15] Hansen C, Joski P, Freiman H, Andrade S, Toh S, Dublin S, Cheetham C, Cooper W, Pawloski P, Li DK, *et al.*. Medication exposure in pregnancy risk evaluation program : the prevalence of asthma medication use during pregnancy. *Maternal and child health journal* 2013 ; **17**(9) :1611–1621.

[16] GINA. Global strategy for asthma management and prevention 2016. URL : http ://ginasthma.org/2016-gina-report-global-strategy-for-asthma-management-and-prevention/ (Accessed 20 Sept. 2016).

[17] Schatz M. Interrelationships between asthma and pregnancy : a literature review. *Journal of allergy and clinical immunology* 1999 ; **103**(2) :S330–S336.

[18] Liu S, Wen SW, Demissie K, Marcoux S, Kramer MS. Maternal asthma and pregnancy outcomes : a retrospective cohort study. *American journal of obstetrics and gynecology* 2001 ; **184**(2) :90–96.

[19] Blais L, Kettani F, Elftouh N, Forget A. Effect of maternal asthma on the risk of specific congenital malformations : A population-based cohort study. *Birth defects research. Part A, Clinical and molecular teratology* 2010 ; **88**(4) :216.

[20] Murphy V, Clifton V, Gibson P. Asthma exacerbations during pregnancy : incidence and association with adverse pregnancy outcomes. *Thorax* 2006 ; **61**(2) :169–176.

[21] Murphy V, Namazy J, Powell H, Schatz M, Chambers C, Attia J, Gibson P. A meta-analysis of adverse perinatal outcomes in women with asthma. *BJOG : An International Journal of Obstetrics & Gynaecology* 2011 ; **118**(11) :1314–1323.

[22] Murphy V, Wang G, Namazy J, Powell H, Gibson P, Chambers C, Schatz M. The risk of congenital malformations, perinatal mortality and neonatal hospitalisation among pregnant women with asthma : a systematic review and meta-analysis. *BJOG : An International Journal of Obstetrics & Gynaecology* 2013 ; **120**(7) :812–822.

[23] Yawn B, Knudtson M. Treating asthma and comorbid allergic rhinitis in pregnancy. *The Journal of the American Board of Family Medicine* 2007 ; **20**(3) :289–298.

[24] Schatz M, Zeiger RS, Hoffman CP, Harden K, Forsythe A, Chilingar L, Saunders B, Porreco R, Sperling W, Kagnoff M. Perinatal outcomes in the pregnancies of asthmatic women : a prospective controlled analysis. *American journal of respiratory and critical care medicine* 1995 ; **151**(4) :1170–1174.

[25] Lim A, Stewart K, König K, George J. Systematic review of the safety of regular preventive asthma medications during pregnancy. *Annals of Pharmacotherapy* 2011 ; **45**(7-8) :931–945.

[26] Busse WW. Naepp expert panel report : managing asthma during pregnancy : recommendations for pharmacologic treatment—2004 update. *Journal of Allergy and Clinical Immunology* 2005 ; **115**(1) :34–46.

[27] Murphy VE, Gibson PG. Asthma in pregnancy. *Clinics in chest medicine* 2011 ; **32**(1) :93–110.

[28] Blais L, Firoozi F, Kettani FZ, Ducharme FM, Lemière C, Beauchesne MF, Bérard A. Relationship between changes in inhaled corticosteroid use and markers of uncontrolled asthma during pregnancy. *Pharmacotherapy : The Journal of Human Pharmacology and Drug Therapy* 2012 ; **32**(3) :202–209.

[29] Clifton VL, Rennie N, Murphy VE. Effect of inhaled glucocorticoid treatment on placental 11β-hydroxysteroid dehydrogenase type 2 activity and neonatal birthweight in pregnancies complicated by asthma. *Australian and New Zealand journal of obstetrics and gynaecology* 2006 ; **46**(2) :136–140.

[30] Rahimi R, Nikfar S, Abdollahi M. Meta-analysis finds use of inhaled corticosteroids during pregnancy safe : a systematic meta-analysis review. *Human & experimental toxicology* 2006 ; **25**(8) :447–452.

[31] Hodyl NA, Stark MJ, Osei-Kumah A, Bowman M, Gibson P, Clifton VL. Fetal glucocorticoid-regulated pathways are not affected by inhaled corticosteroid use for asthma during pregnancy. *American journal of respiratory and critical care medicine* 2011 ; **183**(6) :716–722.

[32] Cossette B, Forget A, Beauchesne MF, Rey É, Lemière C, Larivée P, Battista MC, Blais L. Impact of maternal use of asthma-controller therapy on perinatal outcomes. *Thorax* 2013 ; :thoraxjnl–2012.

[33] Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *Journal of epidemiology and community health* 2013 ; :jech–2013.

[34] Zhan C, Miller M. Administrative data based patient safety research : a critical review. *Quality and Safety in Health Care* 2003 ; **12**(suppl 2) :ii58–ii63.

[35] Miriovsky BJ, Shulman LN, Abernethy AP. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. *Journal of Clinical Oncology* 2012 ; **30**(34) :4243–4248.

[36] Yu PP. The evolution of oncology electronic health records. *The Cancer Journal* 2011 ; **17**(4) :197–202.

[37] Mazzali C, Duca P. Use of administrative data in healthcare research. *Internal and Emergency Medicine* 2015 ; **10**(4) :517–524.

[38] Malcolm E, Downey W, Strand L, McNutt M. Saskatchewan health's linkable data bases and pharmacoepidemiology. *Post Marketing Surveillance* 1993 ; **6**(3) :175–175.

[39] Hennessy S. Use of health care databases in pharmacoepidemiology. *Basic & clinical pharmacology & toxicology* 2006 ; **98**(3) :311–313.

[40] García Rodríguez LA, Pérez Gutthann S. Use of the uk general practice research database for pharmacoepidemiology. *British journal of clinical pharmacology* 1998 ; **45**(5) :419–425.

[41] Tamblyn R, Lavoie G, Petrella L, Monette J. The use of prescription claims databases in pharmacoepidemiological research : the accuracy and comprehensiveness of the prescription claims database in quebec. *Journal of clinical epidemiology* 1995 ; **48**(8) :999–1009.

[42] Vilain A, Otis S, Forget A, Blais L. Agreement between administrative databases and medical charts for pregnancy-related variables among asthmatic women. *Pharmacoepidemiology and drug safety* 2008 ; **17**(4) :345–353.

[43] Suissa S, Garbe E. Primer : administrative health databases in observational studies of drug effects – advantages and disadvantages. *Nature Reviews Rheumatology* 2007 ; **3**(12) :725–732.

[44] Crystal S, Akincigil A, Bilder S, Walkup JT. Studying prescription drug use and outcomes with medicaid claims data strengths, limitations, and strategies. *Medical care* 2007 ; **45**(10 SUPL) :S58.

[45] Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annual review of public health* 2011 ; **32** :91–108.

[46] Riley GF. Administrative and claims records as sources of health care cost data. *Medical care* 2009 ; **47**(7_Supplement_1) :S51–S55.

[47] Briesacher BA, Andrade SE, Fouayzi H, Chan KA. Comparison of drug adherence rates among patients with seven different medical conditions. *Pharmacotherapy : The Journal of Human Pharmacology and Drug Therapy* 2008 ; **28**(4) :437–443.

[48] Hess LM, Raebel MA, Conner DA, Malone DC. Measurement of adherence in pharmacy administrative databases : a proposal for standard definitions and preferred measures. *Annals of pharmacotherapy* 2006 ; **40**(7-8) :1280–1288.

[49] Grimes DA. Epidemiologic research using administrative databases : garbage in, garbage out. *Obstetrics & Gynecology* 2010 ; **116**(5) :1018–1019.

[50] Grimes DA, Schulz KF. False alarms and pseudo-epidemics : the limitations of observational epidemiology. *Obstetrics & Gynecology* 2012 ; **120**(4) :920–927.

[51] Freitas J, Silva-Costa T, Marques B, Costa-Pereira A. Implications of data quality problems within hospital administrative databases. *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*, Springer, 2010 ; 823–826.

[52] Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. *Medical care* 2004 ; **42**(11) :1066–1072.

[53] Levy A, O'Brien B, Sellors C, Grootendorst P, Willison D. Coding accuracy of administrative drug claims in the ontario drug benefit database. *The Canadian journal of clinical pharmacology= Journal canadien de pharmacologie clinique* 2002 ; **10**(2) :67–71.

[54] Virnig BA, McBean M. Administrative data for public health surveillance and planning. *Annual review of public health* 2001 ; **22**(1) :213–230.

[55] Haut ER, Pronovost PJ, Schneider EB. Limitations of administrative databases. *JAMA* 2012 ; **307**(24) :2589–2590.

[56] Annemans L, Aristides M, Kubin M. Real-life data : a growing need. *ISPOR connections* 2007 ; **13**(5) :8–12.

[57] Neugebauer R, Silverberg MJ, van der Laan MJ. Observational study and individualized antiretroviral therapy initiation rules for reducing cancer incidence in hiv-infected patients. *U.C. Berkeley Division of Biostatistics Working Paper Series* 2010 ; (Working Paper 272).

[58] Saarela O, Liu ZA. A flexible parametric approach for estimating continuous-time inverse probability of treatment and censoring weights. *Statistics in medicine* 2016 ; **35**(23) :4238–4251.

[59] Cox DR. Regression models and life-tables. *Breakthroughs in statistics*. Springer, 1992 ; 527–541.

[60] Fisher LD, Lin DY. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health* 1999 ; **20**(1) :145–157.

[61] Cupples LA, D'Agostino RB, Anderson K, Kannel WB. Comparison of baseline and repeated measure covariate techniques in the framingham heart study. *Statistics in medicine* 1988 ; **7**(1-2) :205–218.

[62] D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent cox regression analysis : the framingham heart study. *Statistics in medicine* 1990 ; **9**(12) :1501–1515.

[63] Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men 2000.

[64] Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research : analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources : the ispor good research practices for retrospective database analysis task force report—part iii. *Value in Health* 2009 ; **12**(8) :1062–1073.

[65] Fewell Z, Hernán MA, Wolfe F, Tilling K, Choi H, Sterne JA, *et al.*. Controlling for time-dependent confounding using marginal structural models. *Stata J* 2004 ; **4**(4) :402–420.

[66] Xiao Y, Abrahamowicz M, Moodie EE. Accuracy of conventional and marginal structural cox model estimators : a simulation study. *The international journal of biostatistics* 2010 ; **6**(2).

[67] Holland PW. Statistics and causal inference. *Journal of the American statistical Association* 1986 ; **81**(396) :945–960.

[68] Neyman J. Sur les applications de la theorie des probabilites aux experiences agricoles : essai des principes (masters thesis) ; justification of applications of the calculus of probabilities to the solutions of certain questions in agricultural experimentation. excerpts english translation (reprinted). *Stat Sci* 1923 ; **5** :463–472.

[69] Rubin DB. Causal inference using potential outcomes : Design, modeling, decisions. *Journal of the American Statistical Association* 2005 ; **100**(469) :322–331.

[70] Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology* 1986 ; **15**(3) :413–419.

[71] Rubin DB. Randomization analysis of experimental data : The fisher randomization test comment. *Journal of the American Statistical Association* 1980 ; **75**(371) :591–593.

[72] Cox DR. Planning of experiments. 1958 ; .

[73] Cole SR, Frangakis CE. The consistency statement in causal inference : a definition or an assumption ? *Epidemiology* 2009 ; **20**(1) :3–5.

[74] VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009 ; **20**(6) :880–883.

[75] VanderWeele TJ, Hernan MA, *et al.*. Causal inference under multiple versions of treatment. *J Causal Inference* 2013 ; **1**(1) :1–20.

[76] Hernán MA, Taubman SL. Does obesity shorten life ? the importance of well-defined interventions to answer causal questions. *International journal of obesity* 2008 ; **32** :S8–S14.

[77] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983 ; **70**(1) :41–55.

[78] Wang Y, Petersen ML, Bangsberg D, van der Laan MJ. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment 2006 ; .

[79] Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research* 2012 ; **21**(1) :31–54.

[80] Rubin DB. Basic concepts of statistical inference for causal effects in experiments and observational studies. *Cambridge, MA : Harvard University, Department of Statistics* 2003 ; .

[81] Robins JM. Marginal structural models. 1997.

[82] Robins JM. Marginal structural models versus structural nested models as tools for causal inference. *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000 ; 95–133.

[83] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999 ; :29–46.

[84] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984 ; :431–444.

[85] Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association*, vol. 1999, 2000 ; 6–10.

[86] Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. *Aids Epidemiology*. Springer, 1992 ; 297–331.

[87] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005 ; **61** :962–972.

[88] van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *International Journal of Biostatistics* 2007 ; .

[89] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 1952 ; **47**(260) :663–685.

[90] Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health* 2006 ; **60**(7) :578–586.

[91] Talbot D, Atherton J, Rossi AM, Bacon SL, Lefebvre G. A cautionary note concerning the use of stabilized weights in marginal structural models. *Statistics in medicine* 2015 ; **34**(5) :812–823.

[92] Van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Statistical applications in genetics and molecular biology* 2007 ; **6**(1).

[93] Lefebvre G, Delaney JA, Platt RW. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in medicine* 2008 ; **27**(18) :3629–3642.

[94] Rose S, van der Laan MJ. *Targeted Learning : Causal Inference for Observational and Experimental Data*. Springer New York :, 2011.

[95] Stitelman OM, De Gruttola V, van der Laan MJ. A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. uc berkeley division of biostatistics working paper series. *Technical Report*, Working Paper 281. Available at : http ://biostats. bepress. com/ucbbiostat/paper281 2011.

[96] Porter KE, Gruber S, Van Der Laan MJ, Sekhon JS. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics* 2011 ; **7**(1) :1–34.

[97] Gruber S. An overview of targeted maximum likelihood estimation 2013 ; .

[98] Putter H, Van Zwet WR. Resampling : consistency of substitution estimators. *Selected Works of Willem van Zwet*. Springer, 2012 ; 245–266.

[99] Hampel FR. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 1974 ; **69**(346) :383–393.

[100] Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer Series in Statistics, Springer, 2006.

[101] Kennedy EH. Semiparametric theory and empirical processes in causal inference. *Statistical Causal Inferences and Their Applications in Public Health Research*. Springer, 2016 ; 141–167.

[102] Díaz I, Carone M, van der Laan MJ. Second-order inference for the mean of a variable missing at random. *The international journal of biostatistics* 2016 ; **12**(1) :333–349.

[103] Zhang M, Joffe MM, Small DS. Causal inference for continuous-time processes when covariates are observed only at discrete times. *Annals of Statistics* 2011 ; **39**(1) :131–173.

[104] Lok JJ. Statistical modeling of causal effects in continuous time. *The Annals of Statistics* 2008 ; :1464–1507.

[105] Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. *Annual Review of Public Health* 2013 ; **34** :61–75.

[106] Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *Journal of Epidemiology & Community Health* 2014 ; **68** :283–287.

[107] Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research : challenges and potential approaches. *Medical Care* 2010 ; **48**(6 0) :S114–S120.

[108] Kapteyn A, Ypma JY. Measurement error and misclassification : A comparison of survey and administrative data. *Journal of Labor Economics* 2007 ; **25**(3) :513–551.

[109] Grøn R, Gerds TA, Andersen PK. Misspecified poisson regression models for large-scale registry data : inference for 'large n and small p'. *Statistics in Medicine* 2016 ; **35**(7) :1117–1129.

[110] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in Epidemiology. *Epidemiology* 2000 ; **11**(5) :550–560.

[111] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000 ; **11**(5) :561–570.

[112] Robins JM. Structural nested failure time models. *Encyclopedia of Biostatistics* 1998 ; .

[113] Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association*, vol. 1999, 2000 ; 6–10.

[114] van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2006 ; **2**(1) :Article 11.

[115] van der Laan MJ, Gruber S. Targeted minimum loss based estimation of an intervention specific mean outcome. *The International Journal of Biostatistics* 2012 ; **8**(1) :Article 9.

[116] van der Laan MJ, Rose S. *Targeted Learning : Causal Inference for Observational and Experimental Data.* Springer Series in Statistics, Springer, 2011.

[117] van der Laan MJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics* 2012 ; **8**(1).

[118] Robins JM. Association, causation, and marginal structural models. *Synthese* 1999 ; **121**(1) :151–179.

[119] Schnitzer ME, van der Laan MJ, Moodie EE, Platt RW. Effect of breastfeeding on gastrointestinal infection in infants : a targeted maximum likelihood approach for clustered longitudinal data. *The annals of applied statistics* 2014 ; **8**(2) :703.

[120] Van Der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator : Finite sample oracle inequalities and examples 2003 ; .

[121] Cossette B, Forget A, Beauchesne MF, Rey É, Lemière C, Larivée P, Battista MC, Blais L. Impact of maternal use of asthma-controller therapy on perinatal outcomes. *Thorax* 2013 ; :thoraxjnl–2012.

[122] Busse WW. Naepp expert panel report : managing asthma during pregnancy : recommendations for pharmacologic treatment - 2004 update. *Journal of Allergy and Clinical Immunology* 2005 ; **115**(1) :34–46.

[123] Splawa-Neyman J, Dabrowska D, Speed T, *et al.*. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science* 1990 ; **5**(4) :465–472.

[124] van der Laan MJ, Hubbard AE, Pajouh SK. Statistical inference for data adaptive target parameters 2013 ; .

[125] Hubbard AE, Kherad-Pajouh S, van der Laan MJ. Statistical inference for data adaptive target parameters. *The international journal of biostatistics* 2016 ; **12**(1) :3–19.

[126] Robins JM. Causal inference from complex longitudinal data. *Latent variable modeling and applications to causality*. Springer, 1997 ; 69–117.

[127] Rubin DB. Comment on : "statistics and causal inference" by p. holland. *Journal of the American Statistical Association* 1983 ; **81** :961–962.

[128] Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research* 2012 ; :0962280210386 207.

[129] R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011. URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

[130] Schwab J, Lendle S, Petersen M, van der Laan M. *LTMLE : longitudinal targeted maximum likelihood estimation* 2013. URL `https://CRAN.R-project.org/package=ltmlel`, r package.

[131] Firoozi F, Lemière C, Beauchesne MF, Forget A, Blais L. Development and validation of database indexes of asthma severity and control. *Thorax* 2007 ; **62**(7) :581–587.

[132] Røysland K, *et al.*. A martingale approach to continuous-time marginal structural models. *Bernoulli* 2011 ; **17**(3) :895–915.

[133] van der Laan MJ, Luedtke AR, Díaz I. Discussion of identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data, by jessica young, miguel hernán, and james robins. *Epidemiologic Methods* 2014 ; **3**(1) :21–31.