Prédiction contrefactuelle pour l'estimation causale à partir de données de vie réelle

Arthur Chatton

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE NANTES

THESE DE DOCTORAT DE

L'UNIVERSITE DE NANTES

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : *Biostatistique*

Par
Arthur CHATTON

Prédiction contrefactuelle pour l'estimation causale à partir de données de vie réelle

Thèse présentée et soutenue à Nantes, le 13 décembre 2021
Unité de recherche : UMR INSERM 1246 - SPHERE « methodS in Patient-centered outcomes & HEalth ResEarch »

Rapporteurs avant soutenance :

Raphaël PORCHER	Professeur des universités – Praticien hospitalier, Université de Paris, France
Cécile PROUST-LIMA	Directrice de recherche, INSERM, Université de Bordeaux, France

Composition du Jury :

Président :	
Cécile PROUST-LIMA	Directrice de recherche, INSERM, Université de Bordeaux, France
Examinateurs :	
Anne-Laure BOULESTEIX	Professeur des universités, Université Ludwig-Maximilian de Munich, Allemagne
Julie JOSSE	*Advanced researcher*, INRIA Sophia-Antipolis, Montpellier, France
Directeur de thèse :	
Yohann FOUCHER	Maître de conférences des universités, Université de Nantes, France
Co-encadrant de thèse :	
Florent LE BORGNE	Ingénieur de recherche, Entreprise IDBC, Pacé, France

Remerciements

Cher lecteur,

Tu es la première personne que je souhaite remercier, car après tout, ce manuscrit ne servirait proprement à rien s'il n'était pas lu. Quelle que soit la folie subite t'ayant pris quant à la causalité statistique, je te souhaite une bonne lecture.

Bien que mon nom soit en évidence sur ce manuscrit, son écriture n'aurait pas été possible sans une foultitude de personnes ayant participé de près ou de loin à mes recherches.

Premièrement, j'exprime toute ma gratitude envers les membres du jury qui me font l'honneur de juger la qualité scientifique de cette thèse. Merci aux Professeurs Raphaël Porcher et Cécile Proust-Lima d'avoir accepté de rapporter cette thèse. Merci également aux Professeures Anne-Laure Boulesteix et Julie Josse pour leur participation à ce jury comme examinatrices.

Je souhaiterais ensuite remercier Véronique Sébille et Bruno Giraudeau pour m'avoir accueilli au sein de l'unité SPHERE. Ayant la chance de vous avoir eu comme enseignants, je peux témoigner que votre gentillesse n'a d'égal que votre pédagogie. Le premier cours de Véronique sur les facteurs de confusion fût une réelle révélation pour moi. Quelques années plus tard, me voici en train de défendre ma thèse sur ce même sujet. Quel chemin parcouru ! Pour l'anecdote, je suis allé voir Véronique à la fin de ce cours pour lui demander un stage sur ce sujet. Ne travaillant pas directement sur cette problématique, elle a partagé mon CV au sein du laboratoire. Il tomba ainsi, un peu par hasard finalement, dans les mains de Yohann qui me proposa un stage de M1 sur les scores prédictifs. Il m'avoua lors de l'entretien qu'il n'avait jamais lu ledit CV. Mais comme le monde est bien fait, j'ai pu bifurquer vers la causalité dès le M2 toujours avec Yohann.

J'en viens donc naturellement à remercier Yohann, ainsi que Florent, qui m'encadrent depuis toutes ces années. J'ai énormément de chance d'avoir pu évoluer et apprendre à vos côtés. Merci pour tout le temps que vous m'avez consacré malgré vos nombreuses responsabilités. J'aimerais associer à ces remerciements Clémence Leyrat pour ses encouragements, ses explications et sa disponibilité constante au cours de ces trois années. Tu as fait de moi un bien meilleur scientifique.

Je souhaiterais maintenant remercier Cyrille Loncle et Noël Minard, dirigeants d'IDBC-A2COM. Vous m'avez fait suffisamment confiance pour financer mon sujet de thèse et m'accueillir au sein de votre entreprise pendant ces trois années. Je vous en suis reconnaissant.

Ces dernières années ont permis de belles rencontres sur le plan humain également. Victor Hugo disait que "*l'esprit s'enrichit de ce qu'il reçoit, le cœur de ce que qu'il donne*". Je tiens donc à remercier toutes les personnes avec qui j'ai pu partager un bureau ou un goûter (important les goûters !) à SPHERE. Dans le désordre : Élodie, Bastien, Maxime et Maxime, Jeanne, Line, Rémi, Camille, Yseulys, Marion, Jérôme, Odile, Lucas, Julie et Julie (une plus longtemps que l'autre certes), Marie-Cécile, Priscilla, Solène et tous ceux que j'oublie à ma grande honte. Merci pour la bonne ambiance constante, les pauses cafés, les petits déjeuners et les goûters au labo dont seule la COVID a fini par avoir raison. Ma thèse n'aurait pas été la même sans vous.

J'ai également une pensée pour Gabriel. Tu m'as fait confiance pour t'encadrer dans le cadre de ton stage de M2. J'ai énormément appris pendant ces six mois, j'espère que ce fut également ton cas. Cher lecteur, tu pourras voir le résultat de cette belle collaboration au chapitre 6. Dans

la même veine, merci aux personnes m'ayant permis d'enseigner : Yohann, Cyrille, Véronique, Étienne et Jean-Benoit. Ce fut une expérience extrêmement plaisante qui m'apporta beaucoup. Je souhaite bien continuer dans le futur !

Bien qu'étant passionné de sciences depuis tout petit, le déclic envers la recherche se fit grâce à Émilie et Laurianne. Merci pour ce stage de L3 et pour l'amour de la bibliographie que vous m'avez transmises.

Enfin, je tiens à remercier mes proches. Merci papa, merci maman, pour votre soutien sans faille depuis toujours. J'ai une pensée particulière pour ma grand mère Christiane qui est mon modèle. Tu n'as malheureusement pas faire d'études alors que tu étais faite pour ça. Cette thèse est un peu pour toi.

Clarisse, merci pour ta présence constante, dans le bonheur et la maladie, dans la richesse (pas en tant que doctorant...) et la pauvreté. Tu es la co-autrice de ma plus belle réalisation à ce jour, Alban. Merci à toi aussi, mon petit bonhomme. Tu es né un peu avant le début de cette thèse et tu fus donc présent tout au long de cette aventure. Merci pour tout le bonheur que tu m'apportes chaque jour, tu as transformé ma vie à jamais. J'ai vu à quel point ce fut dur pour toi de me voir travailler alors que tu voulais qu'on joue ensemble. Enfin, mes derniers mots seront pour toi. Toi, que je ne connais pas encore. Toi, que j'ai hâte de rencontrer dans quelques mois. Je t'aime déjà.

Valorisations scientifiques

Publications issues de la thèse

Les auteurs ayant une contribution équivalente (co-premiers ou co-derniers auteurs) à une publication sont indiqués avec un †.

1. **Chatton A**, Le Borgne F, Leyrat C, Gillaizeau F, Rousseau C, Barbin L, Laplaud D-A, Léger M, Giraudeau B et Foucher Y. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets : a comparative simulation study (2020). *Scientific Reports* 10, 9219. DOI : 10.1038/s41598-020-65917-x
2. Le Borgne F†, **Chatton A**†, Léger M, Lenain R et Foucher Y. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes (2021). *Scientific Reports* 11, 1435. DOI : 10.1038/s41598-021-81110-0
3. **Chatton A**, Le Borgne F, Leyrat C et Foucher Y. G-computation and doubly robust standardisation for continuous-time data : a comparison with inverse probability weighting (2021). *Statistical Methods in Medical Research* [Accepté] DOI : 10.1177/09622802211047345
4. Danelian G, Foucher Y, Léger M, Le Borgne F et **Chatton A**. Identifying positivity violations using decision trees : introducing the P-CART algorithm. [En préparation]

Autres publications en lien avec la thèse

1. Lejeune F, **Chatton A**, Laplaud D-A, Le Page E, Wiertlewski S, Edan G, Kerbrat A, Veillard D, Hamonic S, Jousset N, Le Frère F, Ouallet J-C, Brochet B, Ruet A, Foucher Y et Michel L. SMILE : a predictive model for Scoring the severity of relapses in MultIple scLErosis (2021). *Journal of Neurology* 268, 669–679. DOI : 10.1007/s00415-020-10154-5
2. Féray C, Taupin J-L, Sebagh M, Allain V, Demir Z, Allard M-A, Desterke C, Coilly A, Saliba F, Vibert E, Azoulay D, Guettier C, **Chatton A**, Debray D, Caillat-Zucman S et Samuel D. Donor HLA Class 1 Evolutionary Divergence Is a Major Predictor of Liver Allograft Rejection (2021). *Annals of Internal Medicine* [Accepté] DOI : 10.7326/M20-7957
3. Léger M, **Chatton A**, Le Borgne F, Pirracchio R, Sigismond L, Foucher Y. Causal inference in case of near-violation of positivity : comparison of methods (2021). *Biometrical Journal* [Accepté]
4. Haber NA, Wieten SE, Rohrer JM, Arah OA, Tennant PWG, Stuart EA, Murray EJ, Pilleron S, Lam ST, Riederer E, Howcutt SJ, Simmons AE, Leyrat C, Schoenegger P, Booman A, Dufour M-SK, O'Donoghue AL, Baglini R, Do S, Takashima MDLR, Evans TR, Rodriguez-Molina D, Alsalti TM, Dunleavy DJ, Meyerowitz-Katz G, Antonietti A, Calvache JA, Kelson MJ, Salvia MG, Parra CO, Khalatbari-Soltani S, McLinden T, **Chatton A**, Seiler J, Steriu A, Alshihayb TS, Twardowski SE, Dabravolskaj J, Au E, Hoopsick RA, Suresh S, Judd N, Peña S, Axfors C, Khan P, Aguirre AER, Odo NU, Schmid I et Fox MP. Causal and Associational Linking Language From Observational Research and Health Evaluation Literature in Practice : A systematic language evaluation. medRxiv :2021.08.25.21262631. [En révision dans *American Journal of Epidemiology*]

Communications orales

1. **Chatton A**, Le Borgne F, Léger M, Lenain R, Foucher Y. Machine learning, G-computation and small sample sizes : a simulation study. *42nd Conference of the International Society for Clinical Biostatistics*, 18-22 Juillet 2021, Lyon, France (online), 2021.
2. **Chatton A** (invité). G-computation et score de propension. *Séminaire "Méthodes de l'approche causale en épidémiologie"*, 6-8 Juillet 2021, Rennes, France (online), 2021.
3. **Chatton A**, Le Borgne F, Léger M, Lenain R, Foucher Y. G-computation et intelligence artificielle en inférence causale. *15ème Conférence Francophone d'EPIdémiologie CLINique et 28èmes journées des Statisticiens des Centre de Lutte Contre le Cancer*, 8-11 Juin 2021, Marseille, France, 2021.
4. **Chatton A**, Le Borgne F, Foucher Y. G-computation and machine learning for causal inference. *The 8th Channel Network Conference of the International Biometric Society*, 7-9 Avril 2021, Paris, France (online), 2021.
5. **Chatton A**, Le Borgne F, Leyrat C, Foucher Y. G-computation et pondération sur le score de propension en analyse de survie. *14ème Conférence Francophone d'EPIdémiologie CLINique et 27èmes journées des Statisticiens des Centre de Lutte Contre le Cancer*, 15-16 Septembre 2020, Angers, France (online), 2020.
6. **Chatton A**, Le Borgne F, Leyrat C, Foucher Y. G-computation and Inverse Probability Weighting for time-to-event analyses. *41st Annual Conference of the International Society of Clinical Biostatistics*, 23-27 Août 2020, Kraków, Pologne (online), 2020.
7. **Chatton A**, Le Borgne F, Leyrat C, Foucher Y. G-computation et Inverse-Probability-Weighting en analyse de survie. *Journées 2019 du GDR « Statistiques & Santé », de la Société Française de Biométrie et du groupe « Biopharmacie » de la Société Française de Statistique*, 10-11 Octobre 2019, Paris, France, 2019.

Table des matières

Liste des figures

Liste des tableaux

Abréviations

AIPW : *Augmented Inverse Probability Weighting*
ATE : *Average Treatment effect in the Entire population*
ATT : *Average Treatment effect on the Treated*
ATU : *Average Treatment effect on the Untreated*
BDP : *Back-Door Path*
DAG : *Directed Acyclic Graph* - Diagramme acyclique orienté
DML : *Double Machine Learning*
ECR : Essai Contrôlé Randomisé
EDR : Estimateur doublement robuste
FDP : *Front-Door Path*
GC : G-computation
IDBC : Informatique et Données Biomédicales à la Carte
IPW : *Inverse Probability Weighting* - Pondération sur l'inverse du score de propension
ML : *Machine Learning* - Apprentissage automatique
MSM : Modèle Structurel Marginal
SDR : Standardisation Doublement Robuste
SL : *Super Learner*
SP : Score de propension
TMLE : *Targeted Maximim Likelihood Estimateur*

Chapitre 1

Contexte

> « *On ne peut pas toujours tout savoir. Et une partie de ce qu'on sait est toujours fausse. Peut-être la partie la plus importante... Il y a une part de sagesse à savoir ça. Et une part de courage à continuer malgré tout.* »
>
> Robert Jordan, *La roue du temps*

Sommaire

1.1 IDBC et Plug-Stat®

L'entreprise de services du numérique IDBC (*Informatique et Données Biomédicales à la Carte*) est historiquement spécialisée dans la mise en place d'applications web destinées à la collecte des données et à la gestion des consultations. En 2016, IDBC crée un partenariat public-privé avec l'Université de Nantes via l'unité mixte de recherche INSERM 1246 SPHERE (*methodS for Patients-centered outcomes and HEalth REsearch*) dans le but de développer et commercialiser une application web d'analyse statistique sur-mesure : Plug-Stat® [1].

Destinée à l'analyse de données observationnelles, Plug-Stat®est dotée d'interfaces intuitives afin de faciliter et de réduire le temps nécessaire à la réalisation des analyses statistiques. L'utilisateur obtient un rapport complet, rédigé comme une trame pour une publication dans une revue internationale à comité de lecture. A ce jour, cinq études de causalité ont été réalisées avec Plug-Stat® [2–6].

Cependant, cette l'application rencontre des difficultés de commercialisation potentiellement dues à un nombre d'étapes non-automatisables trop important requérant l'implication d'experts en analyse de données observationnelles. Pour ce premier frein, IDBC envisage le développement d'une seconde version de Plug-Stat®utilisable directement par le professionel de santé. Ainsi, la valorisation des données observationnelles en santé serait grandement facilitée et l'investissement pour collecter et garantir la qualité des données mieux rentabilisé.

Actuellement, Plug-Stat®propose des analyses causales. Un second frein à sa commercialisation est l'impossibilité de réaliser des analyses prédictives. La seconde version de Plug-Stat®devra donc inclure un module d'analyses prédictives afin de répondre à ce besoin.

1.2 Objectifs et plan du manuscrit

Dans ce contexte, le premier objectif de cette thèse est d'automatiser un maximum d'étapes liées à l'estimation causale, telles que les hypothèses de validité, afin de rendre l'application encore plus autonome. Différents verrous technologiques devront être levés et feront l'objet d'un travail de recherche approfondi.

Le second objectif consiste au développement et à l'implémentation du module prédictif dans Plug-Stat®. Ce développement sera brièvement décrit dans l'annexe A.

Ce manuscrit est organisé comme suit. Le chapitre 2 consistera en un état de l'art de l'estimation causale et se conclura par les problématiques spécifiques à l'automatisation. Les chapitres 3 à 6 présenteront les contributions scientifiques correspondantes. Enfin, le dernier chapitre consistera en une discussion générale résumant les travaux réalisés et ouvrant sur de nouvelles perspectives tant d'un point de vue scientifique qu'industriel avec le futur potentiel de Plug-Stat®.

Chapitre 2

État de l'art

« Je serais végétarien si le bacon poussait dans les arbres. »

Homer Simpson

Sommaire

2.1 Association et causalité

Pour répondre à la question de l'effet d'un traitement[1], un essai contrôlé randomisé (ECR) permet l'obtention de deux groupes comparables où la seule explication possible en cas de différences post-randomisation serait le traitement [7]. Néanmoins, les ECR sont sujets à certains biais et ne peuvent parfois pas être conduits pour des raisons éthiques ou de faisabilité [8]. Ainsi, des données dites observationnelles peuvent être utilisées, la principale différence étant l'absence de comparabilité directe des groupes de traitement. L'inférence causale peut être vue comme une tentative de recréer un ECR à partir de données observationnelles [9, 10].

Supposons que nous ayons une large population d'individus traités. Nous pouvons les suivre et savoir combien vont décéder d'ici un an. Imaginons que ce soit 20%. Maintenant remontons le temps, ne donnons pas le traitement à ces patients et regardons combien sont décédés à la fin de l'année. Supposons que ce soit 50%. De cette façon, nous pourrions prouver que le traitement a, en moyenne, un effet causal bénéfique sur cette population puisqu'il réduit la mortalité. Évidemment, il est impossible de remonter le temps. Quantifier l'effet causal requiert l'opposition de deux populations, aussi proches que possible, traitée et non-traitée [11].

Supposons maintenant que nous ayons une population d'individus dont certains sont traités et d'autres non. Imaginons que 30% des traités décèdent contre seulement 10% des non-traités à la fin de l'année. Nous ne pourrions pas conclure à un effet causal du traitement si les deux populations ne sont pas similaires. Par exemple, le surplus de mortalité peut s'expliquer par le fait que les patients traités sont plus âgés et donc plus à risque que les non-traités. En revanche, nous pouvons dire que le traitement et la mortalité sont *associés* parce que le risque de décès est différent selon le groupe de traitement. Quantifier une association requiert seulement l'opposition de deux groupes d'individus sous différents traitements. Il peut donc y avoir association sans causalité.

2.2 Critères de Bradford Hill

En 1965, Hill [12] définissait les neuf critères suivant pour différencier les notions de causalité et d'association dans le cadre de la controverse sur le tabac comme cause du cancer du poumon :

1. Force : Une association forte est plus susceptible d'avoir une composante causale que qu'une association modeste.
2. Stabilité : Une répétition observée dans le temps et l'espace est plus susceptible d'être causale.
3. Cohérence : Une conclusion causale ne doit pas contredire fondamentalement les connaissances actuelles connaissance substantielle.
4. Spécificité : Une cause produit un effet donné dans une certaine population en l'absence d'autres explications.
5. Temporalité : La cause précède la conséquence.
6. Relation dose-effet : L'effet augmente avec la dose.
7. Plausibilité : Il est possible d'expliquer les mécanismes impliqués.
8. Expérimentation : La randomisation renforce l'idée de causalité.
9. Analogie : L'association est semblable à des relations causales existantes.

Abondamment discutés [13], ces critères ont servi (et servent parfois encore) à l'inférence causale dans le domaine médical. Hill lui-même précisait que l'ensemble de ces critères n'étaient, ni nécessaires, ni suffisants pour établir une relation causale [12]. Le critère le plus important est sans conteste la temporalité. En effet, la temporalité était une des premières définitions de la causalité en tant que telle [14]. Si Ioannidis [15] juge que seuls deux autres critères (la stabilité et

1. Terme interchangeable avec action, intervention et exposition.

l'expérimentation) sont réellement importants en santé, plusieurs auteurs [16, 17] ajoutent que la plausibilité est un critère clef qui peut être investigué par les graphiques acycliques orientés (DAG, *Directed acyclic graphs*) [18].

2.3 Diagrammes acycliques orientés

2.3.1 Principe

Un DAG est une conceptualisation des connaissances d'experts de la problématique étudiée sous forme de nœuds et de flèches. Chaque nœud représente une variable qu'elle soit mesurée ou non. Ces nœuds sont reliés par des flèches dès lors que nous soupçonnons qu'une variable a un effet causal sur une autre, quelque soit la forme fonctionnelle de cette relation (*e.g.*, quadratique, exponentielle...) [19]. Ainsi, les DAG représentent à la fois les relations causales par des flèches directes et les associations par des chemins composés d'une ou plusieurs flèches. Ceci est illustré dans la Figure 2.1. Soit l'ensemble de variables $\{A, Y, W\}$ où A correspond au traitement, Y à l'évènement et $W = \{U, V\}$ un ensemble de covariables pouvant être mesurées (V) ou non (U). Ici, la variable mesurée V a un effet causal sur A et sur Y représenté par deux flèches pointant sur ces nœuds. Puisqu'aucune flèche ne relie A et Y, A n'a pas d'effet causal sur Y et inversement. Cependant, A et Y sont associées puisqu'il y a un chemin reliant A et Y via $V : A \leftarrow V \rightarrow Y$.

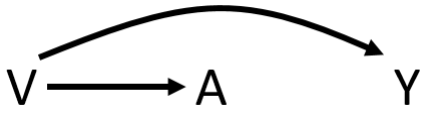


FIGURE 2.1 – Exemple de diagramme causal

Popularisés de façon concomitante par Pearl [20–22] et par Spirtes, Glymour et Scheines [23], les DAG sont maintenant couramment utilisés en épidémiologie pour identifier de potentiels problèmes dans la planification d'une étude, guider l'analyse statistique et permettre une discussion scientifique précise et efficace [24]. Notons qu'un DAG n'est finalement qu'une représentation figurée d'un système d'équations non-paramétriques, dit modèle structurel causal [21]. Par exemple, pour la figure 2.1 :

$$\begin{cases} V = f_V(U_V) \\ A = f_A(V, U_A) \\ Y = f_Y(A, V, U_Y) \end{cases}$$

Où U_X correspond au terme d'erreur de la variable X, c'est-à-dire à l'ensemble de ses causes non-mesurées [21].

2.3.2 Composants particuliers

Les variables W peuvent être classées en six composants différents selon le chemin sur lequel elles se trouvent (Figure 2.2). Les deux cas les plus simples sont l'*instrument* (ou variable instrumentale) qui est une cause du traitement A mais pas de l'évènement Y et le facteur de risque (ou exposition compétitive [25]) qui est une cause de Y mais pas de A. Un *médiateur* est une variable étant sur le chemin causal de deux autres variables. Ici, il s'agit d'une cause de Y et d'une conséquence de A. Si un *collider* est une conséquence commune à deux autres variables, un *facteur de confusion* est au contraire une cause commune à deux autres variables. Enfin, un *proxy* est une conséquence (mesurée) d'une variable non-mesurée.

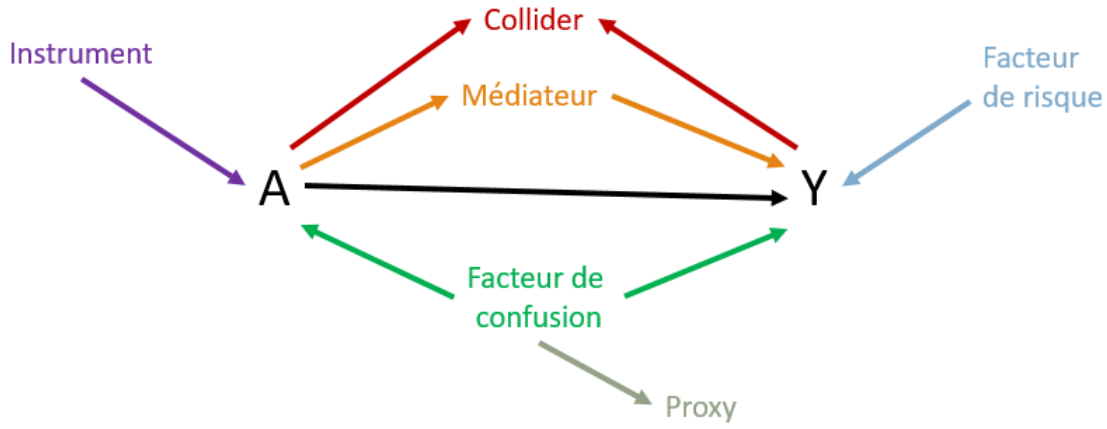


FIGURE 2.2 – Illustration des différents composants possibles d'un diagramme causal

Il est important de noter qu'aucune approche statistique ne permet de différencier un facteur de confusion d'un collider ou d'un médiateur (et vice-versa). En effet, des associations entre A et W et entre W et Y seront observées, que W soit un facteur de confusion, un médiateur ou un collider. [2] Seule une réflexion sur la plausibilité de la relation et de la temporalité de survenue des variables permet de donner des éléments de réponse [26, 28]. Il semble donc d'emblée que l'automatisation complète d'une analyse d'inférence causale soit un objectif impossible à atteindre.

En réalité, la notion de facteur de confusion est un peu plus complexe et nécessite d'introduire les notions de *back-door path* (BDP) et *front-door path* (FDP). Considérons le DAG représenté par la Figure 2.3. En s'intéressant à l'effet causal de A sur Y, le FDP est simplement le chemin liant A et Y dans le sens des flèches, ici $A \rightarrow Y$. Ce type de chemin peut inclure des médiateurs et peut être utile dans les analyses de médiation [29]. Un BDP est un chemin commençant par une flèche pointant vers A et se finissant par une flèche pointant vers Y. En d'autres termes, la présence d'un BDP est lié à la présence d'un facteur commun affectant à la fois l'exposition et l'évènement [19]. Ainsi sur la Figure 2.3, nous pouvons voir le BDP suivant : $A \leftarrow U \rightarrow V \rightarrow Y$ dû à une cause commune non-mesurée U. Les variables A et Y sont donc statistiquement associées à la fois par le FDP et par le BDP. Or, l'effet causal de A sur Y correspond seulement au FDP. Il est nécessaire de bloquer tous les BDP pour estimer correctement l'effet causal ciblé.

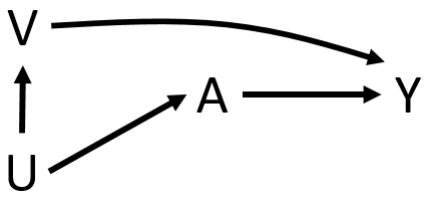


FIGURE 2.3 – Illustration des *back-door* et *front-door paths*.

2.3.3 Biais de confusion et de sélection

Le biais de confusion est sûrement l'un des biais les plus documentés dans la littérature. Ce type de biais est causé par la présence d'une cause commune à A et Y non-contrôlée. Autrement dit, un biais de confusion peut survenir lorsqu'un BDP entre A et Y est ouvert. L'estimation d'un effet causal nécessite donc que tous les BDP soient bloqués pour que l'association observée

2. C'est le paradoxe de Simpson, contrairement à l'idée répandue selon laquelle il correspondrait aux situations de confusion extrême [26, 27]

entre A et Y soit causale. Il existe un ensemble de quatre règles, connu sous le nom de *D(irected)-separation* (Tableau 2.1), permettant de savoir si un chemin est ouvert ou bloqué [30] :

1. Sans conditionnement [3], le chemin est bloqué si et seulement si un collider se trouve sur le chemin
2. Un chemin se bloque si l'on conditionne sur un non-collider
3. Un chemin s'ouvre si l'on conditionne sur un collider
4. Un chemin s'ouvre si l'on conditionne sur une conséquence d'un collider

TABLEAU 2.1 – Résumé des chemins possibles et de l'impact du conditionnement sur L.

Chemin	Description	Terminologie pour V	Avant[a]	Après[a]
$A \rightarrow V \rightarrow Y$	A cause Y via V	Médiateur	Ouvert	Bloqué
$A \leftarrow V \rightarrow Y$	A et Y ont une cause commune V	Facteur de confusion	Ouvert	Bloqué
$A \rightarrow V \leftarrow Y$	A et Y causent V	collider	Bloqué	Ouvert

[a] Avant ou après avoir conditionné sur V.

Reprenons la Figure 2.3. En appliquant les règles précédentes, nous voyons que le BDP $A \leftarrow U \rightarrow V \rightarrow Y$ est ouvert puisqu'il n'y a ni conditionnement, ni collider. Un biais de confusion est alors introduit. Pour bloquer ce chemin, il est nécessaire de conditionner sur un non-collider présent sur le BDP, donc sur U ou V. Puisque U n'est pas mesuré, nous ne pouvons bloquer ce chemin qu'en conditionnant sur V. C'est dans ce sens que VanderWeele et Shpitser [31] ont proposé une nouvelle définition du facteur du confusion selon sa capacité à bloquer les BDP ouverts. Ainsi, un médiateur (mesuré) de la relation entre un facteur de confusion non-mesuré et soit l'évènement, soit l'exposition, peut être considéré comme un facteur de confusion, même s'il n'est pas une cause commune à ces deux variables. Dans le cas où aucun médiateur n'est mesuré, ils proposent de conditionner sur un proxy du facteur de confusion pour réduire le biais dû au BDP ouvert. Notons que pour cet exemple, V peut survenir après A sans que cela ne pose de problème. Le conditionnement sur des variables post-traitement est à éviter lorsque que ce sont des conséquences du traitement.

Les DAG sont le plus souvent vus comme un outil permettant de déterminer l'ensemble de variables $L \subseteq V$ sur lequel conditionner pour éviter les biais de confusion [25]. Or, ils peuvent également représenter les biais de sélection [32, 33]. [4] La Figure 2.4 illustre une situation où aucune association ne sera observée entre A et Y puisqu'un collider, V_1, se trouve sur le BDP. En conditionnant sur ce collider V_1, un biais de sélection sera créé et une association non-causale désormais sera observée entre A et Y via la chemin $A \leftarrow U \rightarrow V_1 \leftarrow V_2 \rightarrow Y$. De façon analogue aux biais de confusion, les biais de sélection peuvent être contrôlés en bloquant les chemins allant du collider conditionné vers Y [11], par exemple en conditionnant sur V_2. Enfin, conditionner sur un médiateur de la relation causale d'intérêt bloque le chemin causal et empêche l'estimation de l'effet total de l'exposition [34]. Notons que deux autres types de biais, non-abordés dans ce manuscrit, peuvent être représentés dans les DAG : la présence de données manquantes et l'erreur de mesure [7].

2.3.4 Limites

Les DAG sont des outils indubitablement utiles. Ils permettent de résumer de grandes quantités d'informations dans une seule figure, facilitant ainsi la planification des études et la discussion

3. Par conditionnement, on entend stratification, ajustement, appariement, etc.
4. Attention, en sciences humaines et sociales le terme biais de sélection est utilisé pour la confusion.

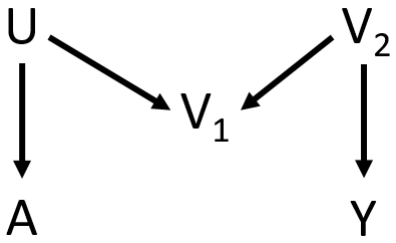


FIGURE 2.4 – Structure en M

autour des biais éventuels pouvant survenir. Les DAG ont notamment facilité la visualisation de certains paradoxes tels que celui de l'obésité [35] ou du poids de naissance [36].

Néanmoins, les DAG sont des représentations qualitatives de la réalité supposées d'après l'expertise du domaine [26,37,38]. Ils ne donnent aucune information sur la présence éventuelle d'interactions ou sur la forme des relations inter-variables [24]. Si d'un côté des tentatives d'inclure les interactions existent [39], van der Laan et Rose [40] arguent que tous les modèles paramétriques sont biaisés par définition puisque la forme de ces relations est impossible à connaître dans des domaines complexes telles que la santé ou les sciences sociales. De façon analogue, les relations elles-mêmes sont complexes à déterminer. Existe-t-il une flèche entre ces deux variables ? Dans quel sens cette flèche pointerait-elle ? Existe-t-il une cause commune non-mesurée entre ces deux variables ? Rohrer [19] propose de tracer plusieurs DAG représentant nos différentes hypothèses afin de réaliser des analyses de sensibilité. Ceci est malheureusement de moins en moins réalisable lorsque le nombre de variables incluses dans le DAG augmente [41]. Ferguson *et al.* [17] proposent une méthode basée sur la littérature pour construire les DAG. De plus, certaines sources de confusion susceptibles pour des petits échantillons (même randomisés), telle que la confusion aléatoire, ne peuvent être représentées graphiquement [42]. Un autre champ de recherche est la découverte causale où les relations causales sont recherchées à partir des données. Glymour *et al.* [43] proposent une revue des méthodes de découverte causale couramment employées.

Comme mentionné précédemment, les DAG sont principalement vus comme un moyen de trouver (via la *D-separation*) l'ensemble L minimal sur lequel conditionner pour contrôler la confusion et ainsi estimer un effet causal [25]. Nous venons de voir qu'obtenir un DAG résumant parfaitement le mécanisme de génération des données est quelque peu irréaliste. C'est dans ce sens que VanderWeele [44] proposait un critère plus large que la *D-separation* pour définir l'ensemble d'ajustement L. Basée sur les DAG et l'expertise clinique, son idée est de sélectionner les variables selon la règle suivante :

1. Sélectionner toutes les variables pouvant être une cause de l'allocation du traitement ou de l'évènement.
2. Exclure de cet ensemble les instruments connus.
3. Ajouter dans l'ensemble tout proxy d'une variable non-mesurée causant à la fois l'allocation du traitement et l'évènement.

Le principal problème de cette approche est la présence possible de variables colinéaires [45]. Vanderweele [44] proposait ainsi de combiner son critère à une approche d'apprentissage automatique (ou *machine learning*, ML) pour restreindre l'ensemble de covariables aux plus importantes mais également obtenir une correcte spécification du modèle (*e.g.*, [46–48]).

Il est important de noter que la méthode de conditionnement utilisée peut conduire à l'estimation d'effets causaux différents [24]. Ainsi, un cadre théorique additionnel est nécessaire pour déterminer l'effet causal ciblé.

2.4 Mondes contrefactuels

2.4.1 Évènements potentiels

Rubin [49] proposait dès 1974 un concept alternatif pour l'estimation causale : les mondes contrefactuels. Inspiré des travaux de Neyman [50] et adapté aux données observationelles, ce cadre théorique fut ensuite étendu à un contexte temps-dépendant par Robins [51].

Ce modèle conceptuel se base sur la question suivante. Que se passerait-il si le patient recevait le traitement A = 1 au lieu de A = 0?. Comme expliqué précédemment, avec une machine à remonter le temps, il serait possible de répondre à cette question. Il serait ainsi possible de donner chaque traitement au patient, de le suivre et de regarder s'il finissait par faire l'évènement d'intérêt ou non, toutes choses égales par ailleurs. Ainsi, avec un traitement binaire, chaque individu possède un couple d'évènements potentiels associés au traitement étudié. Ces évènements potentiels peuvent être notés $Y^{A=1}$ et $Y^{A=0}$ (abrégés en Y^1 et Y^0). C'est-à-dire les évènements faits en ayant reçu les modalités de traitement A = 1 et A = 0, respectivement [52]. Seul un évènement peut être observé, l'autre est dit contrefactuel.

2.4.2 Formalisation des effets causaux

Les évènements potentiels sont des variables conceptuelles, non mesurables, qui permettent de définir l'effet causal ciblé : l'estimand théorique [53, 54]. Il est à différencier de l'estimand empirique qui correspond à un effet estimable à partir des données observées [53]. Par exemple, l'effet causal du traitement dans la population entière (ou *average treatment effect on the entire population*, ATE) peut être égal à la différence moyenne des évènements potentiels des individus composant l'échantillon : $E(Y^1 - Y^0)$. Cette quantité n'est pas directement estimable puisque ces variables ne sont pas disponibles. Au contraire, l'estimand empirique correspondra à la différence moyenne des évènements observés dans deux groupes comparables : $E(Y|A = 1 - Y|A = 0)$. L'estimand théorique est dit *identifiable* lorsqu'il correspond à l'estimand empirique. Pour ceci, il nécessite le respect de plusieurs hypothèses décrites peu après. Deux sortes de biais peuvent ainsi survenir : le *biais d'identification* correspondant au non-respect d'au moins une des conditions d'identifiabilité, et le *biais d'estimation* correspondant à un problème de modélisation [55]. Ce premier biais est commun à toutes les méthodes utilisées pour estimer un effet causal et requiert une expertise du domaine clinique pour éviter son apparition [56]. Au contraire, le second type de biais est spécifique à la méthode d'estimation causale employée et peut nécessiter des hypothèses supplémentaires [57].

TABLEAU 2.2 – Exemples d'estimands théoriques causaux

Dénomination	Contraste causal	Population cible	Estimand théorique
Effet causal individuel	$Y^1 - Y^0$	Individu i	$Y_i^1 - Y_i^0$
ATE (Différence de risque)	$Y^1 - Y^0$	Critères d'éligibilité	$E(Y^1 - Y^0)$
ATT (Différence de risque)	$Y^1 - Y^0$	Individus traités $(A = 1)$	$E(Y^1 - Y^0 \mid A = 1)$
ATU (Différence de risque)	$Y^1 - Y^0$	Individus non-traités $(A = 0)$	$E(Y^1 - Y^0 \mid A = 0)$
ATE (Odds-ratio marginal)	$\frac{P(Y^1=1)/(1-P(Y^0=1))}{P(Y^0=1)/(1-P(Y^1=1))}$	Critères d'éligibilité	$\frac{E(Y^1)/(1-E(Y^0))}{E(Y^0)/(1-E(Y^1))}$
Odds-ratio conditionnel	$\frac{P(Y^1=1)/(1-P(Y^0=1))}{P(Y^0=1)/(1-P(Y^1=1))}$	Sous-population défini par L	$\frac{E(Y^1\|L)/(1-E(Y^0\|L))}{E(Y^0\|L)/(1-E(Y^1\|L))}$

L'estimand théorique possède deux composants. Le premier correspond à un contraste particulier (voir Tableau 2.2 pour des exemples) entre les évènements potentiels clarifiant ainsi l'intervention contrefactuelle étudiée. Il peut s'agir d'une différence de risque, d'un rapport de risque ou

même d'un contraste entre des fonctions des évènements potentiels [52,58]. Le second composant de l'estimand théorique correspond à la population cible [53]. Comme les évènements potentiels se situent en dehors de tout modèle statistique, l'estimand théorique se définit selon la pertinence clinique. Reprenons l'exemple de l'ATE. Le contraste peut correspondre à la moyenne des différences entre les évènements potentiels individuels. La population cible est ainsi celle définie par les critères d'éligibilité. Pour un même contraste d'évènements contrefactuels (*i.e.*, $E(Y^1 - Y^0)$), des estimands alternatifs peuvent être définis. L'effet causal moyen chez les traités (ou *Average treatment effect on the treated*, ATT) correspond à ce contraste dans une sous-population d'individus traités (A = 1). L'effet causal moyen chez les non-traités (ou *Average treatment effect on the untreated*, ATU) est quant à lui défini dans la sous-population des individus non-traités (A = 0) [59]. Enfin, il est également possible de cibler un effet causal conditionnel correspondant à une sous-population particulière définie par l'ensemble des variables d'ajustement L. Il est également possible de définir différents estimands théoriques dans la même population cible, par exemple l'ATT et un odds-ratio spécifique à la sous-population traitée.

2.4.3 Identifiabilité

Une fois l'estimand théorique défini, un lien avec un paramètre statistique (*i.e.*, l'estimand empirique) doit être fait au moyen des conditions d'identifiabilité [53, 54]. Décrite différemment selon les auteurs [7, 11, 59, 60], nous définirons l'identifiabilité au moyen de trois hypothèses [5] :

1. Consistance : $Y = Y^a | A = a$
2. Échangeabilité : $Y^a \perp\!\!\!\perp A$, où $\perp\!\!\!\perp$ désigne l'indépendance [62]
3. Positivité : $0 < P(A = a|L) < 1$

La consistance implique la correspondance des évènements observés et potentiels. Elle nécessite d'abord une définition précise du traitement [63, 64], et ensuite l'absence d'interférence (*i.e.*, *spillover effect* ou contamination) entre les évènements des différents patients [65,66]. L'échangeabilité implique que les individus traités et non-traités aient le même risque moyen de faire l'évènement avant de recevoir une modalité de traitement, ces deux groupes étant donc *échangeables* dans le sens où le même effet causal aurait été observé si les modalités de traitement avaient été inter-changées. Elle implique l'absence de biais de confusion et de sélection, soit formellement $P(Y^a|A = 1) = P(Y^a|A = 0)$ [67]. Si ces biais sont présents, l'*échangeabilité conditionnelle* peut être considérée à la place : $Y^a \perp\!\!\!\perp A|L$ en considérant que toutes les sources de ces deux sortes de biais soient contrôlées par le conditionnement sur L. Il en découle que les groupes sont échangeables entre les différentes strates de la population cible plutôt que directement dans la population. La positivité implique, quant à elle, que tous les individus puissent théoriquement recevoir n'importe quelle modalité de traitement. Deux sortes de violations co-existent. Une violation *structurelle* sera due à la présence d'individus n'ayant théoriquement aucune chance de recevoir l'une des modalités du traitement, par exemple des femmes ménopausées pour une étude sur l'efficacité d'une pilule contraceptive. Une définition précise des critères d'éligibilité est donc nécessaire pour éviter ce type de violation. Cependant, des violations *aléatoires* peuvent également survenir. Elles seront dues à la présence de strates d'individus n'ayant qu'une modalité de traitement sans raison particulière si ce n'est un nombre trop faible d'individus dans ladite strate. Ce type de violation est étroitement lié à l'échangeabilité conditionnelle puisque seules les strates définies par les variables de l'ensemble d'ajustement L peuvent biaiser l'estimation [7]. Si la robustesse aux problèmes de positivité varie selon les méthodes d'estimation causale [68, 69], les analyses de sensibilité restent le seul moyen de s'assurer de la robustesse quant à l'hypothèse d'échangeabilité conditionnelle [70, 71].

5. Des extensions de ces définitions existent dans des cas plus complexes *i.e.*, traitement dépendant du temps ou médiation [61].

Comme indiqué précédemment, les estimands théorique et empirique correspondent si ces trois conditions sont réunies. Gardons l'exemple de l'ATE :

$$\begin{aligned} \text{ATE} &= \text{E}(\text{Y}^1 - \text{Y}^0) \\ &= \text{E}(\text{Y}^1) - \text{E}(\text{Y}^0) \end{aligned}$$

Par échangeabilité :

$$= \text{E}(\text{Y}^1|\text{A}=1) - \text{E}(\text{Y}^0|\text{A}=0)$$

Par consistance :

$$= \text{E}(\text{Y}=1|\text{A}=1) - \text{E}(\text{Y}=0|\text{A}=0)$$

Où les quantités $\text{E}(\text{Y}=1|\text{A}=1)$ et $\text{E}(\text{Y}=0|\text{A}=0)$ sont estimables à partir des données, puisqu'elles n'impliquent pas de variables contrefactuelles non-mesurables par définition.

2.4.4 Interprétation des estimands théoriques

L'ATE permet de déterminer ce qu'il se passerait si tous les patients recevaient ou non le traitement. Au contraire, l'ATT répond à la question de l'efficacité d'un traitement existant chez les patients le recevant [72, 73]. Ces questions ont des implications différentes selon l'objectif visé. Le but est-il de convaincre de la pertinence d'élargir l'indication du traitement ? Ou, au contraire, le but est-il de montrer que le traitement fonctionne chez ceux qui le reçoivent ? Dans le premier cas, l'ATE sera un effet causal plus pertinent, tandis que l'ATT serait à privilégier pour le second.

La notion d'effet causal est également souvent associée à celle d'effet marginal [74, 75] vraisemblablement puisqu'il s'agit de l'effet estimé au travers de l'ECR, considéré comme la référence pour l'estimation causale [10, 22, 74]. Cependant, le terme d'effet causal regroupe les deux notions marginale et conditionnelle, respectivement populationnelle et individuelle [52]. Bien que l'effet causal individuel ne puisse être estimé à proprement dit, l'effet conditionnel peut être un proxy correct dès lors qu'un nombre suffisant de covariables est considéré. Le choix dépend encore de la question étudiée. L'investigateur cherche-t-il à savoir l'effet du traitement sur toute la population incluse ? Ou veut-il estimer l'effet pour un patient particulier ? Ainsi, il voudra respectivement estimer l'effet causal marginal ou l'effet causal conditionnel. Ce choix peut donc être vu comme un compromis entre santé publique et médecine personnalisée. Néanmoins, certains estimands sont équivalents, qu'ils soient conditionnels ou marginaux, ils sont alors dits *collapsibles* [76]. La non-collapsibilité est due à l'inégalité de Jensen : la moyenne d'une fonction non-linéaire n'est pas égale à cette fonction appliquée aux moyennes [77]. Ainsi, lors d'un conditionnement sur L, des mesures non-collapsibles vont différer dès lors que L est associée à Y [78]. Citons comme exemples d'estimands non-collapsibles l'odds-ratio [26, 76], le hazard ratio [79] et la différence de survie [80]. Au contraire, les différences et rapports de risque seront collapsibles. Seuls les effets marginaux seront considérés dans la suite de ce manuscrit.

2.5 Méthodes d'estimation causale

Différentes alternatives de modélisation sont possibles (Figure 2.5, adaptée d'après Schuler et Rose [81]). D'un côté, le score de propension (SP) [82] a pour but de se rapprocher d'un ECR via la modélisation de l'allocation du traitement. Alternativement, la g-computation (GC) est une approche d'imputation/prédiction de l'évènement contrefactuel [83]. Enfin, ces deux approches peuvent être combinées conduisant à des estimateurs dits *doublement robustes* (EDR) [40, 84]. [6]

6. D'autres méthodes existent, par exemple les variables instrumentales [85, 86], la différence de différences [87] ou la g-estimation [88], mais ne seront pas abordées dans ce manuscrit.

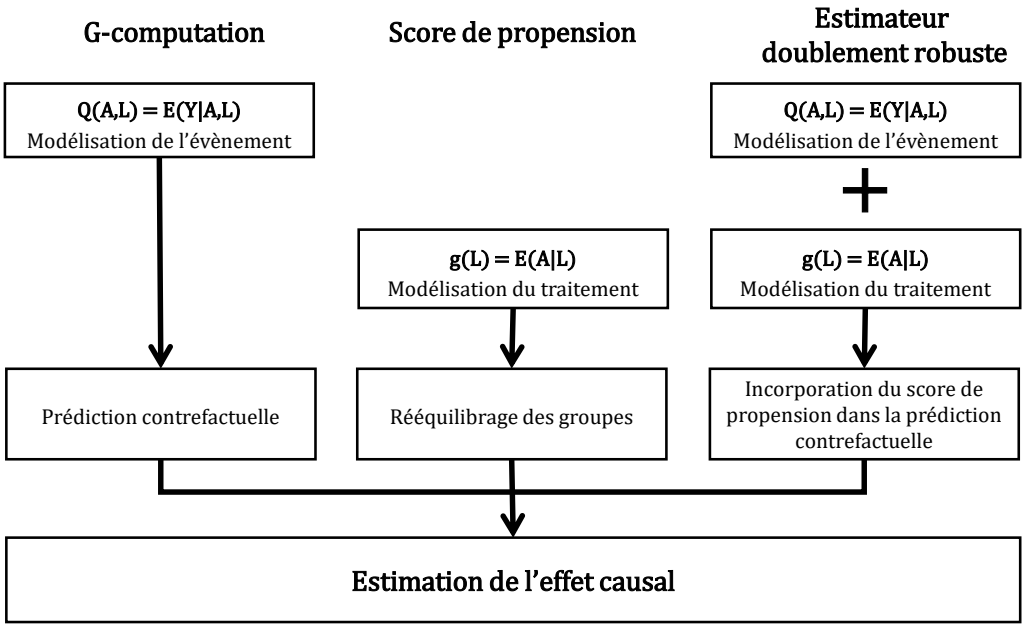


FIGURE 2.5 – Principales méthodes d'estimation causale, adaptée de Schuler et Rose (2017)

2.5.1 Score de propension

Principe

Les méthodes basées sur le SP gagnent en popularité, que ce soit en chirurgie [89], en sciences médicales [90], en épidémiologie [91], en sciences humaines et sociales [92] ou de façon plus générale [93]. Le SP fut défini en 1983 par Rosenbaum et Rubin [82] comme une fonction des variables d'ajustement L permettant d'obtenir deux groupes de traitement équilibrés :

$$g(\mathrm{L}) = \mathrm{P}(\mathrm{A} = 1|\mathrm{L})$$

Ainsi, le SP permet de résumer l'ensemble des potentiels facteurs de confusion mesurés en une seule variable [94]. Ceci permet à la fois d'éviter les problèmes de dimensionalité [95] et le *Table 2 fallacy* (*i.e.*, interprétation erronée des coefficients d'ajustement comme étant causaux) [96]. Le SP est également un score d'équilibre : conditionnellement au SP, la distribution de L sera similaire entre les individus traités et non-traités [94]. Il permet donc de recréer une situation de pseudo-randomisation plus habituelle pour l'estimation causale.

Le but du SP est de rééquilibrer les groupes sur les variables permettant de vérifier l'échangeabilité conditionnelle, donc les facteurs de confusion. Bien qu'étant une modélisation de l'allocation du traitement, l'inclusion d'instruments (cf Figure 2.2, page 6) induit une inflation de la variance [97–99] et peut également conduire à une augmentation du biais en présence de confusion résiduelle [100, 101]. Au contraire, inclure des expositions compétitives peut permettre de diminuer la variance [97, 102]. Puisque le but n'est pas de prédire parfaitement l'allocation traitement, le choix des variables ne devrait pas être guidé par des mesures de discrimination [103]. Plus généralement, la difficulté est que l'estimation de $g(\mathrm{L})$ devrait être guidée par l'équilibre des covariables dans la population contrefactuelle [104], alors qu'elle est le plus souvent guidée par des métriques relatives à ses capacités prédictives du mécanisme d'allocation du traitement.

Une fois l'ensemble de covariables L défini, le modèle de travail $g(\mathrm{L})$ doit être choisi. [7] Il s'agit le plus souvent d'une régression logistique mais des approches de ML plus complexes peuvent

7. Modèle de travail : modèle ne présentant pas un intérêt immédiat mais étant utilisé pour l'analyse des paramètres d'intérêt, ici pour aboutir à une situation de pseudo-randomisation.

être envisagées [105–107]. Le SP alors estimé peut être utilisé de quatre façons : ajustement, stratification, appariement et pondération [82, 108].

Ajustement

Première des trois méthodes initialement proposées par Rosenbaum et Rubin [82] avec la stratification et l'appariement, l'ajustement sur le score de propension consiste à modéliser la relation entre le traitement et l'évènement en incluant le SP comme covariable d'ajustement : $E(Y|A,L) = \beta_0 + \beta_1 A + \beta_2 g(L)$.

Cette approche souffre néanmoins de plusieurs problèmes. Premièrement, elle est particulièrement sensible aux mauvaises spécifications des modèles puisqu'il suffit qu'un seul des deux modèles soit mal spécifié pour biaiser l'estimation [109]. Deuxièmement, seuls certains estimands théoriques peuvent être ciblés. Ainsi, β_1 correspondra à l'ATE avec un évènement continu et à l'odds-ratio conditionnel avec un évènement binaire [109]. Williamson *et al.* [60] notent qu'il est possible d'estimer l'ATT et l'ATU avec un évènement continu en incluant une interaction entre le SP et le traitement dans le modèle final : $E(Y|A,L) = \beta_0 + \beta_1 A + \beta_2 g(L) + \beta_3 A g(L)$. L'ATT et l'ATU correspondront alors à $\beta_1 + \beta_3 \times E(g(L)|A = a)$ avec respectivement $a = 1$ et $a = 0$. Enfin, l'ajustement sur le SP ne permet généralement pas d'atteindre un équilibre entre les groupes suffisant pour l'estimation causale [110]. Ainsi, de nombreux auteurs recommandent d'éviter cette approche [102, 111, 112].

Stratification

Aussi appelée sous-classification, la stratification sur le SP peut être vue comme une sorte de méta-analyse de plusieurs essais (quasi) randomisés [113]. L'idée est de créer plusieurs strates où les individus ont des valeurs de SP proches, puis d'estimer l'effet du traitement dans chaque strate avant d'en faire la moyenne inter-strates [60]. L'ATT peut être estimé en pondérant les strates par la prévalence du traitement spécifique à la strate. De façon analogue, l'ATU est estimable en pondérant par la fraction d'individus non-traités de la strate [59]. La moyenne des variances spécifiques aux strates permet d'obtenir la variance de l'effet estimé [113]. Rosenbaum et Rubin [82] suggéraient l'utilisation des quintiles du SP comme seuils de stratification. Néanmoins, plus fine est la strate, plus faible sera la confusion résiduelle intra-strate [91].

Appariement

Méthode basée sur le SP la plus usitée [89–91], elle permet d'obtenir un pseudo-échantillon équilibré sur L en appariant des individus traités à des individus non-traités ayant une métrique basée sur SP proche, souvent le logit du PS [113]. Ainsi, le SP est utilisé comme une mesure de la distance entre les individus traités et non-traités [114]. Parmi les nombreux algorithmes d'appariement existant [115], nous pouvons distinguer l'appariement selon les plus proches voisins, où un individu traité est tiré aléatoirement de l'échantillon et est apparié à l'individu non-traité ayant le SP le plus proche, de l'appariement optimal où la distance moyenne entre les paires d'appariement sera minimisée automatiquement [91]. L'appariement selon les plus proches voisins implique qu'un individu apparié ne sera plus utilisé par la suite même s'il se serait apparié plus précisément avec un autre individu plus tard [113]. L'appariement peut se faire avec ou sans remise. L'appariement peut également constituer des paires (un individu traité et un non-traité) ou des groupes plus larges tels qu'un traité pour plusieurs non-traités et réciproquement. Enfin, il peut être nécessaire de définir un seuil maximal de dissimilarité, appelé *caliper*, afin d'éviter d'apparier des individus trop dissemblables ce qui conduirait à de la confusion résiduelle [116]. Austin conseille un caliper égal à 0.2 écart-type du logit du SP [117]. De façon générale, l'appariement selon les plus proches voisins avec un tel caliper et sans remise semble conduire aux meilleures

performances [118]. Néanmoins, des approches plus récentes comme le *full matching*[8] semblent prometteuses [119]. Cette méthode est particulièrement intéressante car elle permet d'estimer aussi l'ATE au contraire des autres approches d'appariement sur le SP qui tendent à se rapprocher de l'ATT ou l'ATU [119].

Une fois le sous-échantillon apparié obtenu, les distributions des covariables L doivent être équilibrées entre les deux groupes. La méthode la plus courante pour s'en assurer est la comparaison des différences standardisées entre les deux groupes pour l'échantillon initial et le pseudo-échantillon, une différence inférieure à 10% étant considérée comme nécessaire [90]. Notons que d'autres métriques sont également utilisables [120]. L'utilisation des p-valeurs ne permet pas d'évaluer l'équilibre entre les deux groupes et doit être évitée [121, 122].

Une fois l'appariement effectué, l'effet causal peut être estimé aussi simplement que dans un ECR mais l'estimation de la variance doit néanmoins tenir compte de l'appariement [113], par exemple via l'utilisation d'une matrice de variance robuste de type *sandwich* [123]. Bien que l'utilisation de l'appariement ait été justifié par le fait qu'il permettrait de résoudre d'éventuels problèmes de positivité en supprimant les individus non-appariés, cette suppression conduit principalement à un changement de la population cible et donc d'estimand théorique [116]. Pour une estimation correcte, la redéfinition des critères d'éligibilité devrait être envisagée. King et Nielsen [93] arguent que le changement de population cible (problème de positivité ou non) est un défaut suffisamment important pour éviter l'appariement sur le SP en présence d'alternative crédible telle que la pondération.

Pondération

Méthode plus récente [108], la pondération sur l'inverse du SP (*inverse probability weighting*, IPW) cherche à créer un pseudo-échantillon similaire à celui d'un ECR où les deux groupes seraient échangeables. Chaque individu se voit attribuer un poids correspondant à une fonction du SP [108, 124]. Cette approche peut ainsi être vue comme une extension de l'estimateur d'Horvitz-Thompson [125]. Desai et Franklin [126] arguent que l'IPW est l'approche basée sur le SP la plus flexible puisque l'utilisation de différents systèmes de pondération permettent de cibler différents estimands théoriques (Tableau 2.3). Les poids ciblant l'ATE peuvent être stabilisés, en multipliant le numérateur par la prévalence de la modalité de traitement $P(A = a)$, pour réduire la variance de l'estimation et obtenir un pseudo-échantillon pondéré d'une taille similaire à l'échantillon initial [108, 127]. Notons que le système de pondération ciblant l'ATT (ou l'ATU) est parfois nommé *poids du taux standardisé de mortalité* [128].

De façon analogue à l'appariement, les distributions des covariables L doivent être équilibrées entre les deux groupes de traitement. Austin et Stuart [129] ont étendu l'approche des différences standardisées aux pseudo-échantillons pondérés.

TABLEAU 2.3 – Poids courants utilisés pour la pondération sur le score de propension et estimand ciblé

Système	Poids des traités	Poids des non-traités	Estimand ciblé
ATE classique	$1/g(L)$	$1/(1-g(L))$	ATE
ATT classique	1	$g(L)/(1-g(L)$	ATT
ATU classique	$(1-g(L))/g(L)$	1	ATU
ATE stabilisé	$P(A=1)/g(L)$	$P(A=0)/(1-g(L))$	ATE

8. Décrite comme méthode d'appariement, elle combine en réalité l'appariement, la stratification et la pondération sur le SP.

L'effet causal peut être estimé de différentes façons. Se basant sur les poids classiques et de façon concomitante, Hirano *et al.* [130] ainsi que Lunceford et Davidian [131] ont proposé l'estimateur de l'ATE suivant : $n^{-1}\sum_i^n \frac{A_i Y_i}{g(L_i)} - n^{-1}\sum_i^n \frac{(1-A_i)Y_i}{1-g(L_i)}$, où n dénote le nombre d'individus i. Robins *et al.* [108], Joffe *et al.* [132] ainsi que Cole et Hernán [133] proposent de pondérer un *modèle structurel marginal* (MSM) [9] par le SP afin de pouvoir cibler une gamme plus large d'estimands théoriques. Le MSM prend la forme générale suivante : $E(Y^a) = \beta_0 + \beta_1 \times a$ [11]. Notons que l'évènement considéré dans ce modèle est l'évènement potentiel Y^a. Ainsi, si l'évènement est binaire, l'exponentielle du coefficient $\hat{\beta}_1$ correspondra à une estimation de l'odds-ratio marginal. Avec un évènement continu, le coefficient $\hat{\beta}_1$ correspondra à une estimation de l'ATE ou l'ATT selon les poids utilisés. Une propriété particulièrement intéressante du MSM est qu'il ne sera jamais mal-spécifié lorsque l'effet causal est nul [11]. Néanmoins, le terme *marginal* peut être source de confusion. Un MSM est marginal au sens où il modélise la distribution marginale des évènements potentiels, non pas parce qu'il conduit forcement à l'estimation d'un effet marginal [108, 134]. Comme pour l'appariement sur le SP, il est possible d'ajuster le MSM sur des variables encore déséquilibrées entre les groupes pour contrôler la confusion résiduelle en résultant. Malheureusement, cette approche conduit de nouveau à un changement d'estimand théorique [116, 134]. Lors de l'estimation de la variance, il est nécessaire de tenir compte de la nature pondérée de l'échantillon via une matrice de variance robuste de type *sandwich* [108] ou par bootstrap. Cette dernière approche semble plus performante selon une étude de simulation récente [135].

2.5.2 G-computation

Peu de temps après la parution de l'article princeps sur le SP de Rosenbaum et Rubin [82], Robins publiait à son tour un article proposant une méthode d'estimation causale [51]. Cet article posait les bases de l'inférence causale en présence de confusion dépendante du temps et plus particulièrement d'une rétroaction entre le traitement et un facteur de confusion [10] où les autres méthodes ne fonctionnent pas [11]. [11] Plus particulièrement, Robins [51] y proposait une formule générale pour estimer un effet causal : la *g(eneral)-formula*. Extension de la standardisation [138], cette nouvelle approche peut être décrite comme une succession de prédictions contrefactuelles de l'évènement sous une série de traitements imposés par l'analyste [139] suivie d'une étape de standardisation permettant d'obtenir un effet marginal plutôt que conditionnel [140].

Prenons le cas le plus simple sans confusion dépendante du temps et avec une seule allocation du traitement. L'idée de la *g-formula* est d'estimer la probabilité de faire l'évènement sous une intervention hypothétique, possiblement contrairement aux faits, donc d'estimer la probabilité des deux évènements potentiels $P(Y^a = 1)$ où $a \in \{0, 1\}$.

Notons que la *g-formula* non-paramétrique est mathématiquement équivalente à la version non-paramétrique de l'IPW [141] et très proche du *do-calculus* proposé par Pearl [21]. En pratique, une estimation non-paramétrique est invraisemblable avec le nombre important de covariables à considérer, sans compter la présence possible de covariables continues conduisant à des strates de trop petite taille pour une telle estimation [141]. Robins [51] proposait donc une alternative paramétrique : la GC.

Dans un contexte où le traitement est invariant au cours du temps, la GC [12] se déroule en quatre étapes [144] :

1. Modéliser la survenue de l'évènement conditionnellement aux potentiels facteurs de confusion mesurés via une fonction $Q(A, L)$ sur tous les individus de l'échantillon

9. Attention, ce terme est régulièrement retrouvé dans la littérature comme synonyme d'IPW, à tort [134].

10. Un facteur de confusion a un effet sur l'allocation du traitement qui aura lui-même un effet sur la prochaine valeur du facteur de confusion.

11. L'IPW sera ensuite étendue à ce contexte également, voir [136, 137] pour des introductions.

12. Dans ce contexte, plusieurs synonymes co-existent : *parametric g-formula* [11], *g-standardisation* [142] ou *regression standardisation* [143]

2. Dupliquer l'échantillon en deux échantillons "contrefactuels" identiques excepté sur A. Fixer A = 1 dans un échantillon et A = 0 dans l'autre.
3. Appliquer la fonction Q(A = a, L) dans chaque échantillon pour calculer les probabilités individuelles de faire les évènements contrefactuels pour chaque individu (*i.e.*, prédiction contrefactuelle).
4. Calculer les deux moyennes des probabilités individuelles quand A = 1 et A = 0 pour obtenir l'estimation de l'effet causal d'intérêt.

Notons que la variance est estimable par bootstrap ou simulations paramétriques. Wang *et al.* [145] présentent un tirage avec remise depuis l'échantillon initial pour constituer les deux échantillons contrefactuels de la seconde étape. Néanmoins, cette approche est plus gourmande en temps de calcul sans réel bénéfice [145]. Westreich *et al.* [83] proposent d'estimer seulement l'évènement contrefactuel au lieu des deux potentiels pour augmenter la précision de l'estimation. Snowden *et al.* [144] présentent, comme alternative à l'étape 4, de régresser l'ensemble des prédictions contrefactuelles sur le traitement afin d'estimer l'effet causal au travers d'un MSM. L'estimation de l'ATT ou de l'ATU peut se faire simplement en restreignant l'échantillon aux individus traités ou non-traités, respectivement [141, 145]. Keil *et al.* [146] ont récemment proposé une approche de GC permettant d'estimer l'effet causal de mélange d'expositions. De plus, des versions bayésiennes [147, 148], de médiation [149, 150] ou fonctionnant en présence d'interférence [151] ont été récemment publiées. Néanmoins, la GC reste une méthode quasi-exclusive au domaine de la santé. [13]

2.5.3 Estimateurs doublement robustes

Principe

Si les méthodes basées sur le SP cherchent à recréer un ECR, la GC modélise les deux mondes hypothétiques que l'on voudrait idéalement comparer. La grande différence entre ces approches correspond donc aux hypothèses sous-jacentes de modélisation, le SP nécessitant une bonne spécification du modèle de travail g(L) expliquant l'allocation du traitement tandis que la GC requiert une bonne spécification du modèle de travail Q(A, L) expliquant la survenue de l'évènement. Les EDR utilisent à la fois Q(A, L) et g(L) afin d'obtenir une estimation non-biaisée si au moins un de ces deux modèles de travail est bien spécifié [153, 154].

Pour leur construction, deux philosophies existent : incorporer la GC dans l'IPW ou utiliser l'IPW pour améliorer l'estimation initiale de la GC [153, 155].

Augmented-IPW (AIPW)

Initialement proposée par Robins *et al.* [156], l'AIPW consiste en l'utilisation de Q(A, L) pour améliorer une estimation obtenue avec l'IPW [84]. Ainsi, la probabilité de faire l'évènement potentiel Y^a peut s'écrire comme une combinaison analytique des prédictions issues des deux modèles de travail [157] :

$$P(Y^a = 1) = n^{-1} \sum_{i=1}^{n} \left(\frac{Y_i \mathbb{1}(A_i = a)}{g(L_i)} - \frac{\mathbb{1}(A_i = a) - g(L_i)}{g(L_i)} \times Q(A_i = a, L_i) \right)$$

Bien que cet estimateur soit non-biaisé dès lors qu'au moins un modèle de travail est bien spécifié [14], une mauvaise spécification des deux modèles de travail conduit à une amplification

13. Exception étonnante, cette approche a également été utilisée par Vock et Vock pour étudier le monde contrefactuel où un joueur de baseball donné adopterait les techniques d'un autre joueur [152].
14. Voir annexes de Glynn et Quinn [158] pour les démonstrations.

du biais, donc à un EDR plus biaisé que la GC ou l'IPW [159]. Joffe, dans un manuscrit qui ne fut jamais publié [156], proposait plutôt d'estimer l'effet causal avec un modèle Q(A, L) pondéré par des poids issus de g(L). Il s'agit de la *standardisation doublement robuste* (SDR) [142]. Cette approche à l'avantage de ne pas être sujette à une telle amplification du biais [153, 160]. D'autres alternatives existent comme estimer deux modèles Q(A = 1, L) et Q(A = 0, L) [158] ou inclure $1/g$(L) comme covariable dans Q(A, L) [161].

Si Lunceford et Davidian [131] et Funk *et al.* [84] proposaient respectivement l'utilisation d'une matrice sandwich et du bootstrap pour l'estimation de la variance, le développement des EDR repose sur la théorie des fonctions d'influence (ou gradients canoniques) [162, 163]. [15] Ainsi, ces fonctions peuvent être utilisées pour obtenir une estimation plus sûre et rapide de la variance [164, 166].

Targeted Maximum Likelihood Estimator (TMLE)

En 2006, van der Laan et Rubin [167] introduisaient la TMLE, un EDR utilisant le SP pour réduire une éventuelle confusion résiduelle présente dans l'estimation obtenue par GC [40]. Plus performante que l'AIPW [168], la TMLE est composée de quatre étapes [81, 166, 169] :

1. Estimation initiale de Q(A, L) (étape 1 de la GC)
2. Estimation de g(L)
3. Optimisation : Utilisation de g(L) pour obtenir une estimation "débiaisée" de Q(A, L) notée $Q^*(A, L)$
   (a) Création de deux fonctions contrefactuelles de g(L) : H^a ($a \in \{0, 1\}$)
   (b) Estimation des paramètre de fluctuation δ^a et du modèle optimisé :

$$\text{logit}[Q^*(A, L)] = \text{logit}[Q(A, L)] + \delta^0 \times H^0 + \delta^1 \times H^1$$

4. Estimation de l'effet causal : Étapes 2 à 4 de la GC

Les fonctions H^a sont appelées *clever covariates* et ressemblent aux poids utilisés en IPW. Par exemple, pour cibler l'ATE, les *clever covariates* correspondent au système de pondération décrit dans le Tableau 2.3 (page 14). Les paramètres de fluctuation δ^a représentent quant à eux la force de l'association entre ces *clever covariates*, donc le SP, et les résidus de Q(A, L) [169]. Ces coefficients tendront vers zéro lorsque Q(A, L) est correctement spécifié ou lorsque que g(L) ne permet plus d'améliorer l'estimation. L'étape d'optimisation peut se faire de manière itérative en réutilisant g(L) pour optimiser l'estimation $Q^*(A, L)$ obtenue précédemment. Notons que pour l'ATE, la convergence est garantie en une seule étape [157]. Van der Laan et Rose [40] proposent une procédure alternative utilisant les *clever covariates* pour pondérer le modèle d'optimisation plutôt que comme variables d'ajustement.

Comme l'AIPW, la TMLE est basée sur une fonction d'influence, spécifique à l'estimand, définissant les *clever covariates* [40]. La variance peut être obtenue mathématiquement avec ces fonctions d'influence ou par bootstrap [81]. En présence d'un évènement continu, il est nécessaire de borner les valeurs de l'évènement, appliquer la TMLE puis finalement ré-échelonner l'estimation finale ainsi que sa variance [166].

Les EDR permettent d'avoir plus de chance d'obtenir une estimation sans biais dû à une mauvaise spécification. Kreif *et al.* [170] montraient par simulations que la TMLE ne souffre pas des problèmes d'amplification du biais lorsque les deux modèles de travail sont mal-spécifiés. Cependant, la variance n'est correctement estimée que si les deux modèles de travail sont bien spécifiés. Benkeser *et al.* [171] ont récemment étendu la TMLE afin d'obtenir une inférence correcte lorsqu'un seul des deux modèles de travail est correctement spécifié.

15. L'idée générale est que la fonction d'influence nous informe de l'effet dû au changement d'une observation pour un couple estimateur-estimand donné, voire [164, 165] pour des introductions.

2.6 Problématiques scientifiques dans la perspective d'une plus grande automatisation de Plug-Stat®

2.6.1 Choix de la méthode d'estimation causale et covariables à considérer

L'IPW est la méthode d'estimation causale actuellement implémentée dans Plug-Stat®. Ce choix faisait suite aux nombreuses études de simulation comparant les approches basées sur le SP [172–176], confirmées ensuite par d'autres auteurs [177, 178]. Néanmoins, nous avons vu que d'autres méthodes peuvent également être envisagées. Ces approches, puisque basées sur la modélisation de Q(A, L), permettraient de simplifier l'étape d'estimation causale en évitant l'étape de vérification de l'équilibre entre les deux groupes du pseudo-échantillon. De plus, ces approches peuvent être plus robustes aux violations de l'hypothèse de positivité que l'IPW grâce à leur capacité d'extrapolation [69] (cf annexe C pour un travail collaboratif étudiant ce cas précis). Enfin, il est difficile d'envisager une automatisation de la construction de g(L) sachant que le ML est plus aisé sur des métriques permettant de maximiser des capacités prédictives alors que le but est d'atteindre l'équilibre des covariables. Similairement, l'automatisation pourrait aboutir à l'inclusion d'instruments.

Bien que plusieurs études de simulation comparaient l'IPW, la GC et un ou plusieurs EDR avec un évènement binaire [81, 154, 179], elles ne considéraient que des estimands ciblant la population entière. Seuls Colson *et al.* [180] comparaient les performances de ces méthodes pour estimer l'ATT mais avec un évènement continu. De plus, aucune étude de simulation n'investiguait l'impact du choix des covariables (inclusion d'instruments ou de facteurs de risque) pour des méthodes non-basées sur le SP.

Dans le chapitre 3, nous chercherons à combler ce manque dans la littérature en comparant à la fois les principales méthodes d'estimation causale et différents ensembles d'ajustement pour estimer l'ATE et l'ATT.

2.6.2 Automatisation de la construction de Q(A, L)

Les résultats présentés dans le chapitre 3 montrent de bonnes performances de la GC lorsque les facteurs de risque de l'évènement sont intégrés dans L. Cela ouvre les possibilités d'une automatisation de Q(A, L) pour des métriques liées à ses capacités prédictives. Puisque cette approche est basée sur la modélisation de la survenue de l'évènement, elle possède le potentiel pour éviter l'inclusion néfaste d'instruments. Le ML associé à la GC se pose ainsi comme une alternative aux EDR pour éviter les biais liés à la mauvaise spécification du modèle de travail est l'utilisation de ML [181]. L'avantage est que ces approches ne font pas d'hypothèses quant à la structure des données. L'utilisation de ML permet ainsi de modéliser des relations inter-variables complexes en minimisant une fonction de perte pour s'approcher au mieux des données [182].

Bien que des travaux aient étudié l'intérêt du ML pour construire g(L) [105–107, 183–188], seul Austin [188] s'est spécifiquement intéressé au ML pour construire Q(A, L). Parmi les nombreuses approches de ML existantes, le *super learner* (SL) [189] semble particulièrement intéressant. En effet, cette approche pondère les prédictions obtenues par les différents modèles et algorithmes considérés pour obtenir une estimation finale qui sera moins biaisée que la moins biaisée des méthodes incluses [190].

Dans le chapitre 4, nous étudierons les performances de différentes approches de ML, dont un SL, pour estimer Q(A, L) à partir d'un ensemble pré-défini de covariables potentiellement confondantes.

2.6.3 Estimation causale pour données de temps d'évènement censurés

Les résultats des chapitres 3 et 4 confirment l'intérêt de la GC. Mais son implémentation en analyse de survie n'est pas directe. Trois difficultés ont été identifiées. Premièrement, un biais de sélection intervient au cours du temps (Figure 2.6) [191, 192]. Soit Y_t l'indicatrice d'observation de l'évènement au temps t. Par définition, l'observation de l'évènement au temps 2 dépend de l'observation faite au temps 1. Ce type d'évènement induit ainsi un conditionnement, indiqué par l'encadrement, sur les observations intercurrentes de l'évènement. En présence d'un facteur de risque de l'évènement V, un chemin sera ouvert entre A et Y_τ, τ étant le dernier temps de suivi, introduisant un biais de sélection. Bien que les caractéristiques moyennes des deux groupes de traitement étaient semblables à l'inclusion, elles pourront différer au cours du temps. En effet, le chemin $A \rightarrow Y_1 \leftarrow V \rightarrow Y_2$ est initialement bloqué par Y_1 qui est un collider. L'auto-conditionnement inhérent à la nature de l'évènement ouvre donc ce chemin, biaisant l'estimation de l'effet d'intérêt qui correspond à la flèche allant de A vers Y_2. Une solution pour contrôler ce biais est d'ajuster sur l'ensemble des facteurs de risque V de l'évènement. En présence de facteurs de risques non-mesurés, un biais de sélection résiduel subsistera néanmoins.

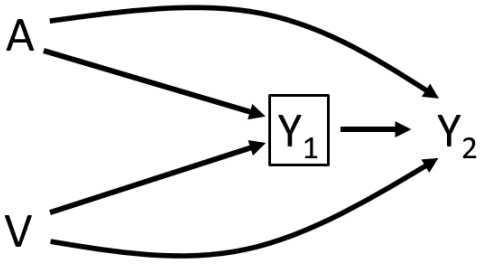


FIGURE 2.6 – Biais de sélection auto-induit en analyse de survie représenté par le chemin ouvert rouge

Deuxièmement, les algorithmes de GC et les EDR actuels pour des analyses de survie sont en temps discret [67,193–196] alors que la grande majorité des problématiques cliniques sont a temps continu [197, 198].

Troisièmement, puisque la GC modélise le mécanisme de survenue de l'évènement, des situations de forte censure à droite pourraient affecter négativement ses performances, au contraire des méthodes basées sur le SP.

Dans le chapitre 5, nous proposons un nouvel estimateur de GC permettant l'utilisation de données en temps continu et censurées à droite, ainsi qu'une extension doublement robuste via l'incorporation d'une étape d'IPW. Nous comparons également ces trois méthodes dans des scénarios faisant varier la taille d'échantillon, le taux de censure et les ensembles d'ajustement.

2.6.4 Identification de patients induisant une situation de non-positivité

L'IPW a l'avantage de permettre une visualisation du respect de l'hypothèse de positivité, nécessaire pour l'identifiabilité de l'effet causal. La distribution du SP dans chaque groupe de traitement peut être représentée graphiquement pour vérifier la présence d'un chevauchement suffisant [60, 199]. De façon analogue, des valeurs de SP extrêmes indiquent une potentielle violation de cette hypothèse [69, 199]. Mais ces approches sont subjectives, nécessitent une spécification correcte du SP et ne permettent pas d'identifier les strates de patients induisant ces violations.

Westreich et Cole [200] proposaient de créer un ensemble de tableaux de contingence représentant l'ensemble des strates possibles, telles que définies par les variables d'ajustement, pour vérifier l'absence de cellules vides. Cette approche n'est réaliste qu'en présence d'un très faible nombre de facteurs de confusion discrets. Un outil basé sur le bootstrap a été proposé par Petersen *et al.* [69] puis modifié par Bahamyirou *et al.* [201]. Il s'agit de la seule approche à ce jour permettant de quantifier le biais dû à une violation de cette hypothèse.

Dans le chapitre 6, nous développons un outil permettant d'identifier les strates d'individus induisant une violation de l'hypothèse de positivité afin de pouvoir redéfinir les critères d'éligibilité. Cette approche a l'avantage de pouvoir être intégrée dans Plug-Stat® quelle que soit la méthode d'estimation causale employée.

Chapitre 3

Comparaison des principales méthodes d'estimation causale selon les covariables incluses

« *Welcome to the real world! It sucks. You're gonna love it.* »

Monica Geller, *Friends*

Les *Supplementary Materials* peuvent être trouvés dans l'Annexe D de ce manuscrit.

SCIENTIFIC REPORTS
nature research

OPEN

G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study

Arthur Chatton[1,2], Florent Le Borgne[1,2], Clémence Leyrat[1,3], Florence Gillaizeau[1,4], Chloé Rousseau[1,4,5], Laetitia Barbin[4], David Laplaud[4,6], Maxime Léger[1,7], Bruno Giraudeau[1,8] & Yohann Foucher[1,4]✉

Controlling for confounding bias is crucial in causal inference. Distinct methods are currently employed to mitigate the effects of confounding bias. Each requires the introduction of a set of covariates, which remains difficult to choose, especially regarding the different methods. We conduct a simulation study to compare the relative performance results obtained by using four different sets of covariates (those causing the outcome, those causing the treatment allocation, those causing both the outcome and the treatment allocation, and all the covariates) and four methods: g-computation, inverse probability of treatment weighting, full matching and targeted maximum likelihood estimator. Our simulations are in the context of a binary treatment, a binary outcome and baseline confounders. The simulations suggest that considering all the covariates causing the outcome led to the lowest bias and variance, particularly for g-computation. The consideration of all the covariates did not decrease the bias but significantly reduced the power. We apply these methods to two real-world examples that have clinical relevance, thereby illustrating the real-world importance of using these methods. We propose an R package *RISCA* to encourage the use of g-computation in causal inference.

The randomised controlled trial (RCT) remains the primary design for evaluating the marginal (population average) causal effect of a treatment, *i.e.*, the average treatment effect between two hypothetical worlds where: i) everyone is treated and ii) everyone is untreated[1]. Indeed, a well-designed RCT with a sufficient sample size ensures the baseline comparability between groups, thus allowing the estimation of a marginal causal effect. Nevertheless, it is well established that RCT is performed under optimal circumstances (*e.g.*, over-representation of treatment-adherent patients, low frequency of morbidity), which may be different from real-life practices[2]. Observational studies have the advantage of limiting the issue of external validity, but treated and untreated patients are often non-comparable, leading to a high risk of confounding bias.

To reduce such confounding bias, the vast majority of observational studies have been based on multivariable models (mainly linear, logistic, or Cox models), allowing for the direct estimation of conditional (subject-specific) effects, *i.e.*, the average effect across sub-populations of subjects who share the same characteristics. Several

[1]INSERM UMR 1246 - SPHERE, Université de Nantes, Université de Tours, Nantes, France. [2]A2COM-IDBC, Pacé, France. [3]Department of Medical Statistics & Cancer Survival Group, London School of Hygiene and Tropical Medicine, London, UK. [4]Centre Hospitalier Universitaire de Nantes, Nantes, France. [5]INSERM CIC1414, CHU Rennes, Rennes, France. [6]Centre de Recherche en Transplantation et Immunologie INSERM UMR1064, Université de Nantes, Nantes, France. [7]Département d'Anesthésie-Réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France. [8]INSERM CIC1415, CHRU de Tours, Tours, France. ✉e-mail: Yohann.Foucher@univ-nantes.fr

methods have been proposed to estimate marginal causal effects in observational studies, amongst which propensity score (PS)-based methods are increasingly used in epidemiology and medical research[3].

Propensity score-based methods make use of the PS in four different ways to account for confounding, namely matching, stratification, conditional adjustment[4] and inverse probability of treatment weighting (IPTW)[5]. Stratification and conditional adjustment on PS are associated with the highest bias[6–8], because the two methods estimate the conditional treatment effect rather than the marginal causal effect. Matching on PS remains the most common approach with a usage rate of 83.8% in 303 surgical studies using PS-based methods[9] and 68.9% in 296 medical studies (without restriction regarding the field) also using PS-methods[10]. The IPTW appears to be less biased and associated with a lower variance than matching in several studies[8,11–14]. Nevertheless, in particular settings, full matching (FM) was associated with lower mean square error (MSE) in other studies[15–17].

Multivariable models, even non-linear ones, can also be used to indirectly estimate the marginal causal effect with g-computation (GC)[18]. This method is also called the parametric g-formula[1] or (g-)standardisation[19] in the literature. Snowden *et al.*[20] and Wang *et al.*[21] detailed the corresponding methodology for estimating the average treatment (*i.e.*, marginal causal) effect on the entire population (ATE) or only on the treated (ATT), respectively. The ATE is the average effect, at the population level, of moving an entire population from untreated to treated. The ATT is the average effect of treatment on those subjects who ultimately received the treatment[22]. Furthermore, some authors[23,24] have proposed combinations of GC and PS to improve the estimation of the marginal causal effect. These methods are known as doubly robust estimators (DRE) because they require the specification of both the outcome (for GC) and treatment allocation (for PS) mechanisms to minimise the impact of model misspecification. Indeed, these estimators are consistent as long as either the outcome model or the treatment model is estimated correctly[25].

Each of these methods carries out the adjustment in different ways, but all of these methods rely on the same condition: a correct specification of the PS or the outcome model[1]. In practice, a common issue is choosing the set of covariates to include to obtain the best performance in terms of bias and precision. Three simulation studies[7,26,27] have investigated this issue for PS-based methods. They studied four sets of covariates: those causing the outcome, those causing the treatment allocation, those are a common cause of both the treatment allocation and the outcome, and all the covariates. For the rest of this paper, we called these strategies the *outcome set*, the *treatment set*, the *common set* and the *entire set*, respectively. These studies argued in favour of the outcome or common sets for PS-based methods, but it is not immediately clear that such works will generalise to other methods of causal inference. Brookhart *et al.*[26] and Lefebvre *et al.*[27] focused on count and continuous outcomes. Austin *et al.*[7] investigated binary outcomes on matching, stratification and adjustment on PS. However, GC and DRE also require the correct specification of the outcome model with a potentially different set of covariates. Recent works have shown that efficiency losses can accompany the inclusion of unnecessary covariates[28–31]. De Luna *et al.*[32] also highlighted the variance inflation caused by the treatment set. In contrast, VanderWeele and Shpitser[33] suggested the inclusion of both the outcome and the treatment sets.

Before selecting the set of covariates, one needs to select the method to employ. Several studies have compared the performances of GC, PS-based methods and DRE in a point treatment study to estimate the ATE[13,23,25,34–36]. Half of these studies investigated a binary outcome[13,25,34]. Only Colson *et al.*[17] studied the ATT, but they focused on a continuous outcome. Except in Neugebauer and van der Laan[25], these studies only investigated the ATE (or ATT) defined as a risk difference. The CONSORT recommended the presentation of both the absolute and the relative effect sizes for a binary outcome, "*as neither the relative measure nor the absolute measure alone gives a complete picture of the effect and its implications*"[37]. None of these studies was interested in the set of covariates necessary to obtain the best performance.

In our study, we sought to compare different sets of covariates to consider to estimate a marginal causal effect. Moreover, we compared GC, PS-based methods and DRE for both the ATE and ATT, either in terms of risk difference or marginal causal OR. Three main types of outcome are used in epidemiology and medical research: continuous, binary and time-to-event outcomes. We focused on a binary outcome because i) a continuous outcome is often appealing for linear regression where the two conditional and marginal causal effects are collapsible[38], and ii) time-to-event analyses present additional methodological difficulties, such as the time-dependant covariate distribution[39]. We also limit our study to a binary treatment, as in the current literature, and the extension to three or more modalities is beyond the scope of our study.

The paper is structured as follows. In the next section, the methods are detailed. The third section presents the design and results of the simulations. In the fourth section, we consider two real data sets. Finally, we discuss our results in the last section.

Methods

Setting and notations. Let A denote the binary treatment of interest ($A = 1$ for treated patients and 0 otherwise), Y denote the binary outcome ($Y = 1$ for events and 0 otherwise), and L denote a set of baseline covariates. Consider a sample of size n in which one can observe the realisations of these random variables: a, y, and l, respectively. Define $\pi_a = E(P(Y = 1|do(A = a), L))$ or $\pi_a = E(P(Y = 1|do(A = a), L)|A = 1)$ as the expected proportions of event if the entire (ATE) or the treated (ATT) populations were treated ($do(A = 1)$) or untreated ($do(A = 0)$), respectively[40]. From these probabilities, the risk difference can be estimated as $\Delta\pi = \pi_1 - \pi_0$ and the log of the marginal causal OR estimated as $\theta = \text{logit}(\pi_1)/\text{logit}(\pi_0)$, where $\text{logit}(\bullet) = \log(\bullet/(1-\bullet))$. The methods described bellow allow for the estimation of both the ATE and the ATT effects.

Causal inference requires the three following assumptions, called *identifiability conditions*: i) The values of exposure under comparisons correspond to well-defined interventions that, in turn, correspond to the versions of treatment in the data. ii) The conditional probability of receiving every value of treatment, though not decided by the investigators, depends only on the measured covariates. iii) The conditional probability of receiving

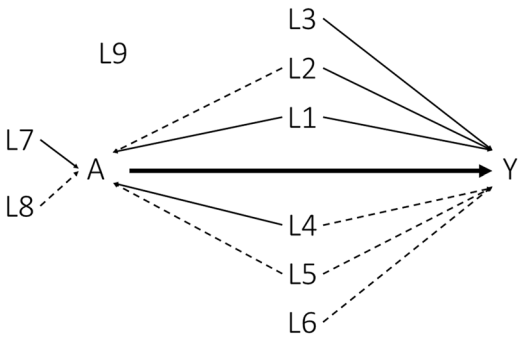


Figure 1. Causal diagram. Solid lines corresponded to a strong association (OR = 6.0) and dashed lines to a moderate one (OR = 1.5).

every value of the treatment is greater than zero, *i.e.*, is positive. These assumptions are known as *consistency*, *(conditional) exchangeability* and *positivity*, respectively[1]. However, PS-based methods rely on treatment allocation modelling to obtain a pseudo-population in which the confounders are balanced across treatment groups. Covariate balance can be checked by computing the standardised difference of the covariates included in the PS between the two treatment groups[10]. In contrast, GC relies on outcome modelling to predict hypothetical outcomes for each subject under each treatment regimen. Note that one can ignore the lack of positivity if one is willing to rely on Q-model extrapolation[1]. As is the case for standard regression models, these methods also require the assumptions of no interference, no measurement error and no model misspecification.

Weighting on the inverse of the propensity score. Formally, the PS is $p_i = P(A_i = 1|L_i)$, *i.e.* the probability that subject i ($i = 1, \ldots, n$) will be treated according to his or her characteristics L_i at the time of the treatment allocation[4]. It is often estimated using a logistic regression. The IPTW makes it possible to reduce confounding by correcting the contribution of each subject i by a weight ω_i. For ATE, Xu *et al.*[41] defined $\omega_i = A_i P(A_i = 1)/p_i + (1 - A_i)P(A_i = 0)/(1 - p_i)$. The use of stabilised weights has been shown to produce a suitable estimate of the variance even when there are subjects with extremely large weights[5,41]. For ATT, Morgan and Todd[42] defined $\omega_i = A_i + (1 - A_i)p_i/(1 - p_i)$. Based on ω_i, the following weighted univariate logistic regression can be fitted: $\text{logit}\{P(Y = 1|A)\} = \hat{\alpha}_0 + \hat{\alpha}_1 A$, resulting in $\hat{\pi}_0 = (1 + \exp(-\hat{\alpha}_0))^{-1}$, $\hat{\pi}_1 = (1 + \exp(-\hat{\alpha}_0 - \hat{\alpha}_1))^{-1}$, and $\hat{\theta} = \hat{\alpha}_1$. To obtain $\widehat{var}(\hat{\theta})$, we used a robust sandwich-type variance estimator[5] with the R package *sandwich*[43].

Full Matching on the propensity score. The FM minimises the average within-stratum differences in the PS between treated and untreated subjects[16]. Then, two weighting systems can be applied in each stratum, making it possible to estimate either the ATE or the ATT unlike other matching methods which can only estimate the ATT[44]. If t and u denote the number of treated and untreated subjects in a given stratum, one can define the weight for a subject i in this stratum as $\omega_i = A_i P(A = 1)(t + u)/u + (1 - A_i)(1 - P(A = 1))(t + u)/t$ for ATE and $\omega_i = A_i + (1 - A_i)t/u$ for ATT[16]. In the latter case, the weights of untreated subjects are rescaled such that the sum of the untreated weights across all the matched sets is equal to the number of untreated subjects: $\widetilde{\omega}_i = \omega_i \times \sum_{j=1}^{n}(1 - A_j)/\sum_{j=1}^{n}\omega_j(1 - A_j)$[45]. From the resulting paired data set, we fitted a weighted univariate logistic regression, and the rest of the data analysis is tantamount to IPTW. We used the R package *MatchIt*[45] to generate the pairs.

G-computation. Consider the following multivariable logistic regression $\text{logit}\{P(Y = 1|A, L)\} = \gamma A + \beta L$. This regression is frequently called the *Q-model*[20]. Once fitted, one can compute for all subjects $\hat{P}(Y_i = 1|do(A_i = 1), L_i)$ and $\hat{P}(Y_i = 1|do(A_i = 0), L_i)$, *i.e.* the two expected probabilities of events if they were treated or untreated[20]. For ATE, one can then obtain $\hat{\pi}_a = n^{-1}\sum_i \hat{P}(Y_i = 1|do(A_i = a), L_i)$. The same procedure can be performed amongst the treated patients for ATT[21]. For implementation in practice, consider a treated subject ($A_i = 1$) included in the fit of the Q-model. Thanks to this model, one can then compute for this subject his or her predicted probabilities of the event if he or she received the treatment ($do(A_i = 1)$) or not ($do(A_i = 0)$). Computing these predicted probabilities for all the subjects, one can obtain two vectors of probabilities if the entire sample were treated or not. The corresponding means correspond to $\hat{\pi}_1$ and $\hat{\pi}_0$, respectively. We obtained $\widehat{var}(\hat{\theta})$ by simulating the parameters of the multivariable logistic regression assuming a multinormal distribution[46]. Note that we could have used bootstrap resampling instead. However, regarding the computational burden of bootstrapping and the similar results obtained by Aalen *et al.*[46], the variance estimates in the simulation study were only based on parametric simulations. We used both bootstrap resampling and parametric simulations in the applications.

Targeted Maximum Likelihood Estimator. Amongst the several existing DREs, we focused on the targeted maximum likelihood estimator (TMLE)[24], for which estimators of ATE and ATT have been proposed[47]. The TMLE begins by fitting the Q-model to estimate the two expected hypothetical probabilities of events $\hat{\pi}_1$ and $\hat{\pi}_0$. An additional "targeting" step involves estimation of the treatment allocation mechanism, *i.e.*, the PS

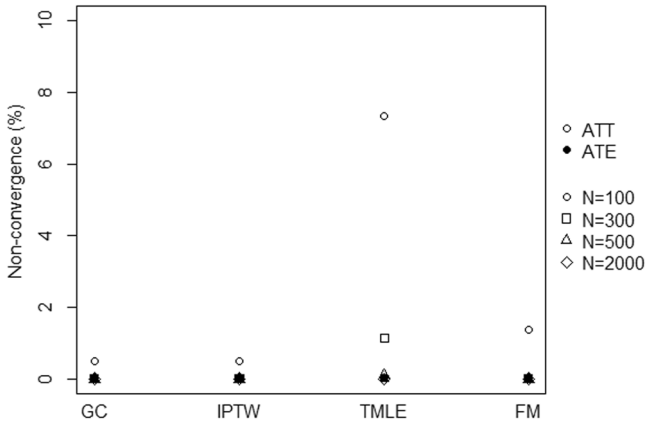


Figure 2. Percentage of simulation iterations which did not converge according to the methods.

$P(A_i = 1|L_i)$, which is then used to update the initial estimates obtained by GC. In the presence of residual confounding, the PS provides additional information to improve the initial estimates. Finally, the updated estimates of $\hat{\pi}_1$ and $\hat{\pi}_0$ are used to generate $\widehat{\Delta\pi}$ or $\hat{\theta}$. We used the efficient influence curve to obtain standard errors[47,48]. A recent tutorial provides a step-by-step guided implementation of TMLE[49].

Simulation study

Design. We used a close data generating procedure from previous studies on PS models[7,50]. We generated the data in three steps. i) Nine covariates ($L_1, \ldots, L_9$) were independently simulated from a Bernoulli distribution with a parameter equal to 0.5 for all covariates. ii) We generated the treatment A according to a Bernoulli distribution with a probability obtained by the logistic model with the following linear predictor: $\gamma_0 + \gamma_1 L_1 + \cdots + \gamma_9 L_9$. We fixed the parameter γ_0 at −3.3 or −5.2 to obtain a percentage of treated patients equal to 50% for scenarios related to ATE and 20% for ATT, respectively. iii) We simulated the event Y using a Bernoulli distribution with a probability obtained by the logistic model with the following linear predictor: $\beta_0 + \beta_1 A + \beta_2 L_1 + \cdots + \beta_{10} L_9$. We set the parameter β_1 for a conditional OR at 0 (the null hypothesis is no treatment effect) or 2 (the alternative hypothesis is a negative impact of treatment). We also fixed the parameter β_0 at −3.65 and −3.5 to obtain a percentage of the event close to 50% in ATE and ATT, respectively. Figure 1 presents the values of the regression coefficients γ_1 to γ_9 and β_1 to β_{10}. We considered four covariates sets as explained in the introduction: the outcome set included the covariates L_1 to L_6, the treatment set included the covariates L_1, L_2, L_4, L_5, L_7, L_8, the common set included the covariates L_1, L_2, L_4, L_5, and the entire set included the covariates L_1 to L_9. For each of the four methods and the four covariate sets, we studied the performance under different sample sizes: $n = 100$, 300, 500 and 2000. For each scenario, we randomly generated 10 000 data sets. We computed the theoretical values of π_1 and π_0 by averaging the values of π_1 and π_0 obtained from univariate logistic models (treatment as the only covariate) fitted from data sets simulated as above, except that the treatment A was simulated independently of the covariates L[50]. We reported the following criteria: i) the percentage of non-convergence, ii) the mean absolute bias (*e.g.*, $E(\hat{\theta}) - \theta$), iii) the MSE ($E[(\hat{\theta} - \theta)^2]$), the variance estimation bias $\left[\text{VEB} = 100 \times \left(\sqrt{E[\widehat{V}ar(\hat{\theta})]}/\sqrt{Var(\hat{\theta})} - 1\right)\right]$[51], the empirical coverage rate of the nominal 95% confidence intervals (CIs), defined as the percentage of 95% CI including the theoretical value, the type I error, defined as the percentage of rejection of the null hypothesis under the null hypothesis, and the statistical power, defined as the percentage of rejections of the null hypothesis under the alternative hypothesis. The MSE was our primary performance measure of interest because it combines bias and variance. We assumed that the identifiability conditions hold in these scenarios. We further performed the same simulations by omitting L_1 in the PS or in the Q-model to evaluate the impact of an unmeasured confounder. We performed all the analyses using R version 3.6.0[52].

Results

Convergence. Non-convergence only occurred for ATT estimation when sample sizes were lower or equal to 300 subjects (see Fig. 2). The GC, IPTW and FM had a minimal convergence percentage higher than 98%, even under small sample size (n = 100). Similarly, TMLE experienced some difficulty in converging for ATT estimation in the medium-sized sample (n = 300). However, they experienced severe difficulty in converging in the small sample with a convergence percentage of approximately 92%.

Mean bias. As expected with the common set, the mean absolute bias of θ was close to zero for GC, IPTW and TMLE when the three identifiability assumptions hold with a maximum at −0.028 given moderate sample size (n = 300) under the alternative hypothesis for ATT estimation (Table 1). Note that the three other covariate sets led to a bias close to zero with a maximum of 0.053 for TMLE with the entire set given small sample size (n = 100) under the alternative hypothesis for ATE estimation (Table 2). Furthermore, FM was also associated with a similar bias with a maximum of 0.082 given a small sample size (n = 100), with the treatment set under the alternative hypothesis for the ATE estimation. With an unmeasured confounder, the bias increased in all scenarios with a

n	method	selection strategy	mean bias				log OR				
			π_0	π_1	$\Delta\pi$	log OR	MSE	MSE*	VEB (%)	coverage (%)	power (%)
100	GC	outcome	0.000	−0.001	−0.001	0.012	0.526	0.716	−6.2	94.1	17.7
		treatment	0.002	−0.001	−0.003	0.006	0.580	0.786	−5.7	94.1	14.0
		common	0.002	−0.001	−0.003	0.006	0.552	0.735	−4.2	94.8	15.1
		entire	−0.001	−0.001	−0.001	0.013	0.558	0.768	−8.8	93.3	16.9
	IPTW	outcome	0.000	−0.001	−0.001	0.008	0.578	0.727	10.8	97.3	7.8
		treatment	−0.000	−0.001	−0.001	0.000	0.716	0.837	−1.2	95.1	9.8
		common	0.002	−0.001	−0.003	0.003	0.587	0.743	6.6	96.8	8.8
		entire	−0.003	−0.001	0.002	0.005	0.741	0.838	−1.5	95.2	9.6
	TMLE	outcome	−0.001	−0.001	0.000	0.002	0.694	0.794	30.0	95.7	5.8
		treatment	0.000	−0.001	−0.001	−0.020	0.876	0.955	183.3	98.8	1.0
		common	−0.000	−0.001	−0.001	−0.001	0.702	0.794	10.4	95.3	7.3
		entire	−0.003	−0.001	0.001	−0.013	0.886	0.953	412.2	98.8	0.5
	FM	outcome	−0.004	−0.001	0.003	0.022	0.665	0.787	−16.7	90.1	18.9
		treatment	−0.006	−0.001	0.004	0.017	0.822	0.911	−32.3	81.3	25.2
		common	−0.001	−0.001	−0.000	0.010	0.653	0.795	−15.3	91.0	17.5
		entire	−0.008	−0.001	0.006	0.022	0.842	0.921	−33.8	80.3	26.7
300	GC	outcome	0.001	−0.001	−0.002	−0.021	0.283	0.555	−1.6	94.5	43.6
		treatment	0.002	−0.001	−0.003	−0.024	0.319	0.606	−2.3	94.3	35.2
		common	0.002	−0.001	−0.003	−0.023	0.304	0.561	−1.5	94.8	38.5
		entire	0.001	−0.001	−0.002	−0.022	0.297	0.600	−2.6	94.0	39.9
	IPTW	outcome	0.002	−0.001	−0.003	−0.027	0.301	0.556	16.4	97.9	24.0
		treatment	0.001	−0.001	−0.002	−0.026	0.372	0.628	6.6	96.2	21.4
		common	0.003	−0.001	−0.004	−0.028	0.318	0.563	9.1	96.8	26.1
		entire	0.001	−0.001	−0.002	−0.025	0.361	0.622	11.7	97.2	20.0
	TMLE	outcome	0.000	−0.001	−0.001	−0.023	0.358	0.577	−2.3	93.6	29.0
		treatment	0.002	−0.001	−0.003	−0.035	0.454	0.683	51.2	99.1	6.8
		common	0.001	−0.001	−0.002	−0.023	0.378	0.582	−3.5	93.0	26.5
		entire	0.002	−0.001	−0.003	−0.035	0.432	0.674	81.8	99.3	4.4
	FM	outcome	−0.000	−0.001	−0.001	−0.020	0.351	0.579	−11.7	91.9	37.2
		treatment	−0.001	−0.001	−0.000	−0.022	0.444	0.656	−30.2	82.7	38.9
		common	0.001	−0.001	−0.002	−0.024	0.363	0.587	−14.6	90.4	36.9
		entire	−0.001	−0.001	0.000	−0.020	0.439	0.662	−29.3	83.2	39.1
500	GC	outcome	0.001	−0.001	−0.002	−0.014	0.217	0.509	−1.1	94.7	64.5
		treatment	0.001	−0.001	−0.002	−0.014	0.245	0.556	−1.5	94.4	53.6
		common	0.001	−0.001	−0.002	−0.015	0.233	0.618	−0.8	94.8	57.6
		entire	0.001	−0.001	−0.002	−0.014	0.228	0.552	−2.0	94.2	60.5
	IPTW	outcome	0.002	−0.001	−0.003	−0.019	0.230	0.509	16.5	97.9	43.3
		treatment	0.000	−0.001	−0.001	−0.013	0.285	0.574	6.8	96.6	35.4
		common	0.002	−0.001	−0.003	−0.018	0.244	0.514	9.2	96.8	43.7
		entire	0.000	−0.001	−0.001	−0.014	0.274	0.571	12.3	97.2	33.9
	TMLE	outcome	0.001	−0.001	−0.002	−0.015	0.272	0.521	−4.7	93.4	48.5
		treatment	0.001	−0.001	−0.002	−0.018	0.347	0.618	35.0	99.1	15.9
		common	0.000	−0.001	−0.001	−0.013	0.289	0.527	−4.8	93.1	43.7
		entire	0.001	−0.001	−0.002	−0.019	0.328	0.611	51.1	99.3	12.9
	FM	outcome	0.001	−0.001	−0.002	−0.015	0.265	0.525	−9.9	92.4	53.0
		treatment	−0.001	−0.001	−0.000	−0.011	0.346	0.597	−31.0	82.7	51.7
		common	0.001	−0.001	−0.001	−0.014	0.283	0.530	−15.8	90.1	52.3
		entire	−0.002	−0.001	0.001	−0.008	0.340	0.596	−29.8	83.2	52.6
2000	GC	outcome	0.000	0.000	−0.000	−0.002	0.108	0.479	−1.7	94.7	99.6
		treatment	0.001	0.000	−0.000	−0.003	0.122	0.524	−1.2	94.8	98.6
		common	0.001	0.000	−0.000	−0.003	0.116	0.480	−0.9	94.7	99.1
		entire	0.000	0.000	−0.000	−0.002	0.113	0.523	−1.8	94.5	99.4
	IPTW	outcome	0.002	0.000	−0.001	−0.006	0.113	0.478	16.3	97.6	98.1
		treatment	0.000	0.000	−0.000	−0.002	0.138	0.539	7.9	96.4	93.0
		common	0.002	0.000	−0.001	−0.006	0.120	0.480	9.4	97.0	97.7
		entire	0.000	0.000	−0.000	−0.002	0.131	0.537	13.9	97.4	93.6

Continued

n	method	selection strategy	mean bias				log OR				
			π_0	π_1	$\Delta\pi$	log OR	MSE	MSE*	VEB (%)	coverage (%)	power (%)
2000	TMLE	outcome	0.001	0.000	−0.000	−0.002	0.132	0.483	−5.9	93.3	97.5
		treatment	0.000	0.000	0.000	−0.002	0.169	0.568	18.2	98.2	71.8
		common	−0.000	0.000	0.000	−0.000	0.142	0.486	−5.6	93.6	95.5
		entire	0.001	0.000	−0.000	−0.004	0.158	0.565	23.5	98.6	75.3
	FM	outcome	0.000	0.000	−0.000	−0.002	0.134	0.484	−12.0	91.6	97.7
		treatment	0.001	0.000	−0.000	−0.005	0.203	0.548	−41.6	74.6	89.9
		common	0.001	0.000	−0.000	−0.003	0.149	0.485	−20.5	88.5	96.7
		entire	0.000	0.000	0.000	−0.002	0.162	0.543	−26.9	84.5	94.8

Table 1. Simulation results comparing the ATT estimation under the alternative hypothesis. *MSE in the presence of an unmeasured confounder. Theoretical values: $\pi_1 = 0.701$, $\pi_0 = 0.589$, $\theta = 0.492$.

minimum of 0.456 for GC with the common set given a large sample size for the ATT estimation (see Online Supporting Information (OSI) for complete results). The results were similar under the null hypothesis (see OSI).

Variance. For all methods, the outcome set led to the lowest MSE, followed closely by the common set. G-computation led to the lowest MSE and FM to the highest. In ATT, IPTW had lower MSE than TMLE. Note that the VEB was particularly high for FM in all ATE scenarios with a minimum of −17.5% (n = 500 with the outcome set). For the ATT, FM also had a higher VEB than other methods, apart from TMLE with the treatment or entire sets in sample sizes of fewer than 2000 subjects. In the presence of an unmeasured confounder, the MSE increased in all scenarios in agreement with the increase in bias. The VEBs did not change notably with an unmeasured confounder.

Coverage and error rates. G-computation produced coverage rates close to 95%, except for ATE in a small sample size leading to an anti-conservative 95% CIs with a minimum of 91.7% with the entire set under the null hypothesis. Anti-conservatives 95% CIs were also produced by FM in all scenarios, and by TMLE given a small sample size. Conversely, conservative 95% CIs were obtained when using TMLE for the ATT with the entire or the treatment sets, and when using IPTW for ATT or ATE with the outcome or the common sets.

Lending confidence to these results, the type I error was close to 5% for GC in all scenarios and may vary for other methods. The power was more impacted by the choice of the covariate set. The outcome set led to the highest power for GC.

Applications

We illustrated our findings by using two real data sets. First, we compared the efficiency of two treatments, *i.e.*, Natalizumab and Fingolimod, sharing the same indication for active relapsing-remitting multiple sclerosis. Physicians preferentially use Natalizumab in practice for more active disease, indicating possible confounders. Given the absence of a clinical trial with a direct comparison of their efficacy, Barbin *et al.*[53] recently conducted an observational study. We reused their data. Second, we sought to study barbiturates that can lead to a reduction of the patient functional status. Indeed, barbiturates are suggested in Intensive Care Units (ICU) for the treatment of refractory intracranial pressure increases. However, the use of barbiturates is associated with haemodynamic repercussions that can lead to brain ischaemia and immunodeficiency, which may contribute to the occurrence of infection. These applications were conducted in accordance with the French law relative to clinical noninterventional research. According to the French law on Bioethics (July 29, 1994; August 6, 2004; and July 7, 2011, Public Health Code), the patients' written informed consent was collected. Moreover, data confidentiality was ensured in accordance with the recommendations of the French commission for data protection (Commission Nationale Informatique et Liberté, CNIL decisions DR-2014-558 and DR-2013-047 for the first and the second application, respectively).

To define the four sets of covariates, we asked experts (D.L. for multiple sclerosis and M.L. for ICU) which covariates were causes of the treatment allocation and which were causes of the outcome, as proposed by VanderWeele and Shpitser[33]. We checked the positivity assumption and the covariate balance (see OSI). We applied B-spline transformations for continuous variables when the log-linearity assumption did not hold.

Natalizumab versus Fingolimod to prevent relapse in multiple sclerosis patients. The outcome was at least one relapse within one year of treatment initiation. Six hundred and twenty-nine patients from the French national cohort OFSEP were included (www.ofsep.org). The first part of Table 3 presents a description of their baseline characteristics.

All included patients could have received either treatment. Therefore, we sought to estimate the ATE. The first part of Table 4 presents the results according to the different possible methods and covariate sets. The GC, IPTW and TMLE yield similar results regardless of the covariate sets considered. Thus, Fingolimod exhibits lower efficacy than Natalizumab with an OR [95% CI] ranging from 1.50 [1.02; 2.21] for IPTW with the entire set to 1.55 [1.06; 2.28] for GC with the common set. When using FM, the OR ranged from 1.73 [1.19; 2.51] with the outcome set to 1.78 [1.23; 2.56] with the common set. Note that, unlike IPTW, FM does not to balance all covariates in the outcome set with standardised differences higher than 10%.

Overall, the confounder-adjusted proportion of patients with at least one relapse within the first year of treatment was lower in the hypothetical world where all patients received Natalizumab (approximately 20% and

n	method	set	mean bias				log OR				
			π_0	π_1	$\Delta\pi$	log OR	MSE	MSE*	VEB (%)	coverage (%)	power (%)
100	GC	outcome	−0.001	−0.002	−0.001	−0.003	0.404	0.634	−7.3	93.2	24.7
		treatment	−0.002	−0.001	0.000	0.004	0.477	0.727	−9.5	92.4	19.9
		common	−0.001	−0.002	−0.001	−0.002	0.434	0.650	−6.6	93.5	22.1
		entire	−0.002	−0.001	0.001	0.003	0.450	0.714	−11.4	91.8	22.6
	IPTW	outcome	−0.003	−0.001	0.001	0.011	0.464	0.646	12.1	97.4	12.1
		treatment	−0.006	0.002	0.008	0.046	0.633	0.769	−7.6	93.8	16.7
		common	−0.002	−0.001	0.001	0.010	0.480	0.657	6.3	96.3	13.5
		entire	−0.006	0.003	0.009	0.053	0.647	0.773	−7.2	94.7	16.4
	TMLE	outcome	−0.001	−0.002	−0.000	0.003	0.438	0.642	−14.3	89.5	26.9
		treatment	−0.004	0.002	0.006	0.039	0.572	0.757	−24.9	84.3	27.5
		common	−0.001	−0.002	−0.001	0.002	0.469	0.657	−10.7	90.9	21.2
		entire	−0.005	0.003	0.007	0.043	0.544	0.748	−30.7	80.9	34.3
	FM	outcome	−0.005	0.002	0.006	0.039	0.549	0.710	−24.3	87.1	28.5
		treatment	−0.009	0.005	0.014	0.082	0.677	0.832	−37.7	78.0	35.1
		common	−0.005	0.001	0.006	0.038	0.563	0.713	−26.3	85.8	29.1
		entire	−0.007	0.006	0.014	0.082	0.674	0.830	−37.3	78.1	34.8
300	GC	outcome	−0.000	−0.000	0.000	0.001	0.221	0.532	−1.9	94.5	59.8
		treatment	−0.000	−0.000	0.000	0.001	0.259	0.608	−2.8	94.3	47.4
		common	−0.000	−0.000	0.000	0.001	0.237	0.539	−1.2	94.8	53.5
		entire	−0.000	−0.000	0.000	0.001	0.241	0.600	−3.4	94.0	53.0
	IPTW	outcome	−0.001	−0.000	0.001	0.006	0.239	0.533	20.2	98.0	34.7
		treatment	−0.002	0.000	0.003	0.014	0.330	0.615	4.6	96.0	29.5
		common	−0.001	−0.000	0.001	0.006	0.252	0.541	13.3	97.4	36.5
		entire	−0.002	0.000	0.002	0.013	0.326	0.607	7.9	96.6	28.5
	TMLE	outcome	−0.000	−0.001	−0.000	0.000	0.233	0.532	−3.0	93.9	54.2
		treatment	−0.001	0.000	0.002	0.009	0.310	0.612	−10.4	90.6	40.2
		common	−0.001	−0.001	0.000	0.001	0.249	0.540	−1.5	94.6	48.1
		entire	−0.001	0.000	0.001	0.008	0.290	0.603	−13.2	89.6	46.1
	FM	outcome	−0.002	0.000	0.002	0.010	0.294	0.552	−20.2	88.7	51.6
		treatment	−0.003	0.003	0.006	0.032	0.389	0.652	−39.3	77.0	53.3
		common	−0.001	−0.000	0.001	0.008	0.315	0.588	−25.5	86.2	51.3
		entire	−0.003	0.003	0.006	0.032	0.377	0.644	−37.4	77.8	52.2
500	GC	outcome	−0.000	0.000	0.001	0.003	0.168	0.501	−0.4	94.8	81.1
		treatment	−0.000	0.000	0.001	0.002	0.198	0.573	−1.0	94.8	69.0
		common	−0.000	0.000	0.000	0.002	0.183	0.505	−0.7	94.9	75.0
		entire	−0.000	0.000	0.001	0.004	0.183	0.569	−1.0	94.8	75.3
	IPTW	outcome	−0.001	0.000	0.001	0.005	0.180	0.501	22.2	98.3	58.5
		treatment	−0.001	0.001	0.001	0.007	0.248	0.573	8.1	96.5	42.3
		common	−0.001	0.000	0.001	0.005	0.193	0.505	13.8	97.3	58.6
		entire	−0.001	0.000	0.001	0.006	0.239	0.569	13.1	97.2	41.3
	TMLE	outcome	−0.000	0.000	0.000	0.002	0.177	0.501	−0.8	94.7	76.8
		treatment	−0.000	0.000	0.000	0.003	0.234	0.571	−5.9	92.7	56.1
		common	−0.000	0.000	0.000	0.002	0.190	0.505	−0.5	94.7	69.7
		entire	−0.000	0.000	0.000	0.003	0.218	0.566	−7.5	91.8	63.1
	FM	outcome	−0.001	0.000	0.001	0.005	0.219	0.518	−17.5	89.8	70.1
		treatment	−0.002	0.002	0.003	0.018	0.302	0.598	−39.8	76.2	65.5
		common	−0.001	−0.000	0.001	0.005	0.266	0.555	−31.8	82.3	66.4
		entire	−0.002	0.002	0.004	0.019	0.289	0.592	−37.1	78.3	66.2
2000	GC	outcome	−0.000	−0.000	−0.000	−0.001	0.085	0.482	−0.6	94.6	100.0
		treatment	0.000	−0.001	−0.001	−0.003	0.099	0.550	−0.6	94.7	99.8
		common	0.000	−0.001	−0.001	−0.003	0.092	0.483	−0.8	94.7	99.9
		entire	−0.000	−0.000	−0.000	−0.001	0.091	0.550	−0.6	94.7	99.9
	IPTW	outcome	−0.000	−0.000	0.000	0.002	0.090	0.482	21.2	98.2	99.8
		treatment	0.000	−0.001	−0.001	−0.002	0.122	0.547	9.3	96.7	95.1
		common	−0.000	−0.000	0.000	0.001	0.096	0.483	13.5	97.3	99.7
		entire	0.000	−0.000	−0.001	−0.002	0.117	0.546	14.3	97.5	95.6

Continued

n	method	set	mean bias				log OR				
			π_0	π_1	$\Delta\pi$	log OR	MSE	MSE*	VEB (%)	coverage (%)	power (%)
2000	TMLE	outcome	−0.000	−0.000	−0.000	−0.001	0.088	0.482	−0.6	94.8	100.0
		treatment	0.000	−0.001	−0.001	−0.003	0.116	0.545	−2.2	94.4	98.7
		common	0.000	−0.000	−0.001	−0.002	0.095	0.483	−0.3	94.8	99.9
		entire	0.000	−0.000	−0.001	−0.002	0.108	0.544	−2.6	94.1	99.4
	FM	outcome	−0.000	−0.000	−0.000	0.000	0.129	0.497	−29.9	82.9	99.0
		treatment	−0.001	−0.000	0.000	0.003	0.169	0.569	−46.6	70.6	96.2
		common	0.000	−0.000	−0.001	−0.001	0.205	0.534	−55.9	61.1	92.7
		entire	−0.000	−0.000	0.000	0.002	0.145	0.549	−37.7	77.9	98.2

Table 2. Simulation results comparing the ATE estimation under the alternative hypothesis. *MSE in the presence of an unmeasured confounder. Theoretical values: $\pi_1 = 0.557$, $\pi_0 = 0.441$, $\theta = 0.466$.

varying slightly depending on method and set of covariates) than one in which all patients received Fingolimod (approximately 28%). This difference of approximately 8% is clinically meaningful and suggests the superiority of Natalizumab over Fingolimod to prevent relapses at one year. This result was concordant with the recent clinical literature[53,54].

Impact of barbiturates in the ICU on the functional status at three months. We define an unfavourable functional outcome by a 3-month Glasgow Outcome Scale (GOS) lower than or equal to 3. We used the data from the French observational cohort AtlanREA (www.atlanrea.org) to estimate the ATT of barbiturates because physicians recommended these drugs to a minority of severe patients. The second part of Table 3 presents the baseline characteristics of the 252 included patients.

The second part of Table 4 presents the results according to the different possible methods and covariate sets. G-computation and TMLE lead to the conclusion of a significant negative effect of barbiturates regardless of the covariate set considered with an OR [95% CI] ranging from 0.43 [0.25; 0.76] for GC with the common set to 0.51 [0.29; 0.90] for TMLE with the entire set. By contrast, the results were discordant when using different covariate sets for IPTW and FM. We report, for instance, OR estimates obtained by FM ranging from 1.520 with the outcome set to 2.300 with the common set. In line with the simulation study, the estimated standard errors were higher for these methods (0.294 and 0.293 for GC and TMLE when the outcome set was considered, respectively) leading to lower power. Note also that standardised differences were higher than 10% for the IPTW with the entire set (see OSI) and for FM with the outcome, the treatment and the entire sets.

Depending on the methods and sets of covariates included, we estimated that from 18% to 20% of patients treated with barbiturates had an unfavourable GOS at three months. If these patients had not received barbiturates, the methods estimate that from 30% to 35% would have had an unfavourable GOS at three months. For the patients, this difference is meaningful but full clinical relevance depends also on the effect of barbiturates on other clinically relevant outcomes, such as death or ventilator-associated pneumonia. However, the results obtained by GC or TMLE differ with those obtained by Majdan *et al.*[55], who did not find any significant effect of barbiturates on the GOS at six months. Two main methodological reasons can explain this difference: the GOS was at six months rather than three months post-initiation, and the authors used multivariate logistic regression leading to a different estimand.

Discussion

The aim of this study was to better understand the different sets of covariates to consider when estimating the marginal causal effect.

The results of our simulation study, limited to the studied scenarios, highlight that the use of the outcome set was associated with the lower bias and variance, principally when associated with GC, for both ATE and ATT. As expected, an unmeasured confounder led to increased bias, regardless of method employed. Although we do not report an impact on the variance, the effect's over- or under-estimation leads to the corresponding over- or under-estimation of power and compromises the validity of the causal inference.

The performance of FM is lower than that of the other studied methods, especially for the variance. Our results were in line with King and Nielsen[56], who argued for halting the use of PS matching for many reasons such as covariate imbalance, inefficiency, model dependence and bias. Nonetheless, Colson *et al.*[17] found slightly higher MSE for GC than FM. Their more simplistic scenario, with only two simulated confounders leading to little covariate imbalance, could explain the difference with our results. Moreover, is unclear whether they accounted for the matched nature of the data, as recommended by Austin and Stuart[16] or Gayat *et al.*[50].

While DRE offers protection against model misspecification[23,34,36], our simulation study resulted in the finding that GC was more robust to the choice of the covariate set than the other methods, TMLE included. This result was particularly important when the treatment set was taken into account, which fits with the results of Kang and Schafer[35]: when both the PS and the Q-model were misspecified, DRE had lower performance than GC. Furthermore, GC was associated with lower variance than DRE in several simulation studies[13,17,35], which accords with our results.

The first application to multiple sclerosis (ATE) illustrated similar results between the studied methods. In contrast, the second application (ATT) to severe trauma or brain-damaged patients showed different results between the methods. In agreement with simulations, the estimations obtained with GC or TMLE were similar

A - Multiple sclerosis	Overall (n = 629)		First line treatment					Relapse at 1 year				
			Ntz (n = 326)		Fng (n = 303)		*p*	No (n = 478)		Yes (n = 151)		*p*
Patient age, years (mean, sd)	37.0	9.6	36.8	9.9	37.2	9.2	0.6505	37.1	9.7	36.6	9.2	0.5849
Female patient (n, %)	479.0	76.2	254.0	77.9	225.0	74.3	0.2822	367.0	76.8	112.0	74.2	0.5124
Disease duration, years (mean, sd)	8.5	6.4	8.0	6.1	9.0	6.8	0.0505	8.6	6.6	8.2	6.0	0.4809
At least one relapse (n, %)	526.0	83.6	293.0	89.9	233.0	76.9	<0.0001	391.0	81.8	135.0	89.4	0.0277
Gd-enhancing lesion on MRI (n, %)	311.0	49.4	185.0	56.7	126.0	41.6	0.0001	240.0	50.2	71.0	47.0	0.4944
EDSS score >3 (n, %)	288.0	45.8	166.0	50.9	122.0	40.3	0.0074	212.0	44.4	76.0	50.3	0.1986
Previous immunomodulatory treatment (n, %)	556.0	88.4	293.0	89.9	263.0	86.8	0.2284	424.0	88.7	132.0	87.4	0.6672
B – ICU	**Overall (n = 252)**		**Barbiturates treatment**					**Favourable GOS at 3 months**				
			No (n = 178)		Yes (n = 74)		*p*	No (n = 180)		Yes (n = 72)		*p*
Patient age, years (mean, sd)	47.4	17.4	48.7	17.9	44.1	15.7	0.0565	50.8	16.4	38.7	16.9	<0.0001
Female patient (n, %)	89.0	35.3	58.0	32.6	31.0	41.9	0.1592	68.0	37.8	21.0	29.2	0.1963
Diabetes (n, %)	17.0	6.7	15.0	8.4	2.0	2.7	0.0989	15.0	8.3	2.0	2.8	0.1122
Nosological entity: Severe trauma (n, %)	124.0	49.2	95.0	53.4	29.0	39.2	0.0403	77.0	42.8	47.0	65.3	0.0012
SAP ≤90 mmHg before admission (n, %)	56.0	22.2	36.0	20.2	20.0	27.0	0.2368	46.0	25.6	10.0	13.9	0.0442
Evacuation of subdural or extradural hematoma (n, %)	41.0	16.3	33.0	18.5	8.0	10.8	0.1301	27.0	15.0	14.0	19.4	0.3878
External ventricular drain (n, %)	64.0	25.4	39.0	21.9	25.0	33.8	0.0486	48.0	26.7	16.0	22.2	0.4640
Evacuation of cerebral hematoma or lobectomy (n, %)	42.0	16.7	28.0	15.7	14.0	18.9	0.5362	34.0	18.9	8.0	11.1	0.1345
Decompressive craniectomy (n, %)	27.0	10.7	15.0	8.4	12.0	16.2	0.0686	21.0	11.7	6.0	8.3	0.4396
Blood transfusion before admission (n, %)	34.0	13.5	25.0	14.0	9.0	12.2	0.6903	26.0	14.4	8.0	11.1	0.4841
Pneumonia before increased ICP (n, %)	29.0	11.5	16.0	9.0	13.0	17.6	0.0519	19.0	10.6	10.0	13.9	0.4538
Osmotherapy (n, %)	112.0	44.4	75.0	42.1	37.0	50.0	0.2525	89.0	49.4	23.0	31.9	0.0115
GCS score ≥8	62.0	24.6	39.0	21.9	23.0	31.1	0.1237	37.0	20.6	25.0	34.7	0.0183
Hemoglobin, g/dL (mean, sd)	11.8	2.3	11.7	2.2	12.1	2.5	0.1824	11.8	2.4	11.9	1.9	0.7373
Platelets, counts/mm^3 (mean, sd)	206.7	78.0	207.4	79.7	205.1	74.2	0.8312	209.0	83.8	200.9	61.1	0.4589
Serum creatinine, mmol/L (mean, sd)	71.1	29.3	71.1	27.6	71.1	33.3	0.9853	72.4	32.6	67.9	18.7	0.2732
Arterial pH (mean, sd)	7.3	0.1	7.3	0.1	7.3	0.1	0.0978	7.3	0.1	7.3	0.1	0.6317
Serum proteins, g/L (mean, sd)	58.2	10.4	57.7	10.6	59.6	9.7	0.1662	58.0	10.7	58.8	9.7	0.5963
Serum urea, mmol/L (mean, sd)	5.0	2.5	5.2	2.7	4.7	1.8	0.1827	5.2	2.3	4.5	2.9	0.0505
PaO_2/FiO_2 ratio (mean, sd)	302.7	174.0	292.7	154.7	326.6	212.9	0.1595	282.1	172.4	354.2	168.4	0.0028
SAPS II score (mean, sd)	47.6	11.4	47.6	10.7	47.6	12.9	0.9847	49.9	10.8	41.8	10.7	<0.0001

Table 3. Baseline characteristics of patients of the two studied cohorts. Ntz: Natalizumab, Fng: Fingolimod, Gd: Gadolinium, MRI: Magnetic Resonance Imaging, EDSS: Expanded Disability Status Scale, SAP: Systolic Arterial Pressure, ICP: Intra-Cranial Pressure, GCS: Glasgow Coma Scale, PaO_2/FiO_2: arterial partial Pressure of Oxygen/Fraction of Inspired Oxygen, SAPS II: Simplified Acute Physiology Score II.

in terms of logOR estimation and variance regardless of the covariate set considered. Estimations obtained with IPTW or FM were highly variable, depending on the covariate set employed: some indicated a negative impact of barbiturates and others did not. These results also tended to demonstrate that GC or TMLE had the highest statistical power. Variances obtained by parametric simulations or by bootstrap resampling were similar (results not displayed).

One can, therefore, question the relative predominance of the PS-based approach compared to GC, although there are several potential explanations. First, there appears to be a pre-conceived notion according to which multivariable non-linear regression cannot be used to estimate marginal absolute and relative effects[57]. Indeed, under logistic regression, the mean sample probability of an event is different from the event probability of a subject with the mean sample characteristics. Second, while there is an explicit variance formula for the IPTW[58], the equivalent is missing for the GC. The variance must be obtained by bootstrapping, simulation or the delta method. Third, several didactic tutorials on PS-based methods can be found, for instance[59–61].

We still believe that PS-based methods may have value when multivariate modelling is complex, for instance, for multi-state models[62]. In future research, it would be interesting to examine whether the use of potentially better settings would provide equivalent results, such as the Williamson estimator for IPTW[58], the Abadie-Imbens estimator for PS matching[63], or bounded the estimation of TMLE, which can also be updated several times[36]. We also emphasise that we did not investigate these methods when the positivity assumption does not hold. Several authors have studied this problem[13,25,35,36,64]. G-computation was less biased than IPTW or DRE except in Porter *et al.*[36], where the violation of the positivity assumption was also associated with model misspecifications. The robustness of GC to non-positivity could be due to a correct extrapolation into the missing sub-population, which is not feasible with PS[1]. Other perspectives of this work are to extend the problem to i) time-to-event, continuous or multinomial outcomes and ii) multinomial treatment. However, implementing GC using continuous treatment raises many important considerations concerning the research question and resulting inference[64].

application	method	set	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\theta}$	SE	95% CI
A - Multiple sclerosis	GC	outcome	20.3	28.2	0.432	0.189	[0.062, 0.802]
		treatment*	20.3	28.3	0.436	0.195	[0.054, 0.819]
		common*	20.3	28.3	0.436	0.195	[0.054, 0.819]
		entire	20.3	28.2	0.431	0.191	[0.056, 0.806]
	IPTW	outcome	21.2	28.8	0.406	0.195	[0.023, 0.789]
		treatment*	20.3	28.2	0.433	0.191	[0.059, 0.808]
		common*	20.3	28.2	0.433	0.191	[0.059, 0.808]
		entire	21.3	28.9	0.406	0.196	[0.022, 0.791]
	TMLE	outcome	21.2	28.8	0.407	0.195	[0.025, 0.790]
		treatment*	20.3	28.2	0.433	0.190	[0.061, 0.806]
		common*	20.3	28.2	0.433	0.190	[0.061, 0.806]
		entire	21.1	28.9	0.410	0.196	[0.026, 0.794]
	FM	outcome	19.1	29.0	0.549	0.189	[0.178, 0.921]
		treatment*	19.9	30.6	0.575	0.187	[0.210, 0.941]
		common*	19.9	30.6	0.575	0.187	[0.210, 0.941]
		entire	21.1	31.9	0.561	0.183	[0.201, 0.920]
B - ICU	GC	outcome	66.3	81.1	0.778	0.294	[0.201, 1.354]
		treatment	65.3	81.1	0.824	0.298	[0.240, 1.407]
		common	65.0	81.1	0.836	0.289	[0.270, 1.402]
		entire	66.5	81.1	0.769	0.295	[0.191, 1.347]
	IPTW	outcome	31.0	81.1	0.656	0.356	[−0.042, 1.354]
		treatment	68.2	81.1	0.693	0.355	[−0.002, 1.388]
		common	67.4	81.1	0.729	0.353	[0.038, 1.421]
		entire	69.2	81.1	0.645	0.362	[−0.064, 1.354]
	TMLE	outcome	66.2	79.6	0.692	0.293	[0.118, 1.266]
		treatment	65.4	80.2	0.758	0.288	[0.194, 1.322]
		common	64.8	79.9	0.769	0.298	[0.185, 1.354]
		entire	66.4	79.4	0.668	0.285	[0.109, 1.228]
	FM	outcome	73.8	81.1	0.419	0.342	[−0.252, 1.090]
		treatment	67.2	81.1	0.739	0.337	[0.078, 1.399]
		common	65.1	81.1	0.831	0.336	[0.173, 1.490]
		entire	66.2	81.1	0.782	0.336	[0.123, 1.442]

Table 4. Results of the two applications. *Treatment and common sets contain same covariates. π_0: Percentage of event in the Natalizumab (or control) group, π_1: Percentage of event in the Fingolimod (or Barbiturates) group, SE: standard error.

To facilitate its use in practice, we have implemented the estimation of both ATE and ATT, and their 95% CI, from a logistic model in the existing R package entitled *RISCA* (available at cran.r-project.org/web/packages/RISCA). We provide an example of R code in the appendix. Note that the package did not consider the inflation of the type I error rate due to the modelling steps of the Q-model. Users also have to consider novel strategies for post-model selection inference.

In the applications, we classified covariates into sets based on experts knowledge[33]. However, several statistical methods can be useful when no clinical knowledge is available. Heinze *et al.*[65] proposed a review of the most used, while Witte and Didelez[66] reviewed strategies specific to causal inference. Alternatively, data-adaptive methods have recently been developed, such as the outcome-adaptive LASSO[67] to select covariates associated with both the outcome and the treatment allocation. Nevertheless, according to our results, it may be preferable to focus on constructing the best outcome model based on the outcome set. For instance, the consideration of a super learner[68,69], merging models and modelling machine learning algorithms may represent an exciting perspective[70].

Finally, we emphasise that the conclusions from our simulation study cannot be generalised to all situations. They are consistent with the current literature on causal inference, but theoretical arguments are missing for generalisation. Notably, our results must be considered in situations where both the PS and the Q-model are correctly specified and where positivity holds.

To conclude, we demonstrate in a simulation study that adjusting for all the covariates causing the outcome improves the estimation of the marginal causal effect (ATE or ATT) of a binary treatment in a binary outcome. Considering only the covariates that are a common cause of both the outcome and the treatment is possible when the number of potential confounders is large. The strategy consisting of considering all available covariates, *i.e.*, no selection, did not decrease the bias but significantly decreased the power. Amongst the different studied methods, GC had the lowest bias and variance regardless of covariate set considered. Consequently, we recommend that the use of the GC with the outcome set, because of its highest power in all the simulated scenarios. For

instance, at least 500 individuals were necessary to achieve a power higher than 80% in ATE, with a theoretical OR at 2, and a percentage of treated subjects at 50%. In ATT, we needed larger sample size to reach a power of 80% because the estimation considers only the treated patients. With 2000 individuals, all the studied methods with the outcome set led to a bias close to zero and a statistical power superior to 95%.

Received: 9 July 2019; Accepted: 26 April 2020;
Published online: 08 June 2020

References

1. Hernan, M. A. & Robins, J. M. *Causal Inference: What if*? (Chapman & Hall/CRC, 2020).
2. Zwarenstein, M. & Treweek, S. What kind of randomized trials do we need? *Journal of Clinical Epidemiology* **62**, 461–463, https://doi.org/10.1016/j.jclinepi.2009.01.011 (2009).
3. Gayat, E. *et al*. Propensity scores in intensive care and anaesthesiology literature: a systematic review. *Intensive Care Medicine* **36**, 1993–2003, https://doi.org/10.1007/s00134-010-1991-5 (2010).
4. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55, https://doi.org/10.2307/2335942 (1983).
5. Robins, J. M., Hernán, M. A. & Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560, https://doi.org/10.1097/00001648-200009000-00011 (2000).
6. Lunceford, J. K. & Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* **23**, 2937–2960, https://doi.org/10.1002/sim.1903 (2004).
7. Austin, P. C., Grootendorst, P., Normand, S.-L. T. & Anderson, G. M. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine* **26**, 754–768, https://doi.org/10.1002/sim.2618 (2007).
8. Abdia, Y., Kulasekera, K. B., Datta, S., Boakye, M. & Kong, M. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal* **59**, 967–985, https://doi.org/10.1002/bimj.201600094 (2017).
9. Grose, E. *et al*. Use of propensity score methodology in contemporary high-impact surgical literature. *Journal of the American College of Surgeons* **230**, 101–112.e2, https://doi.org/10.1016/j.jamcollsurg.2019.10.003 (2020).
10. Ali, M. S. *et al*. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology* **68**, 112–121, https://doi.org/10.1016/j.jclinepi.2014.08.011 (2015).
11. Le Borgne, F., Giraudeau, B., Querard, A. H., Giral, M. & Foucher, Y. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. *Statistics in Medicine* **35**, 1103–1116, https://doi.org/10.1002/sim.6777 (2016).
12. Hajage, D., Tubach, F., Steg, P. G., Bhatt, D. L. & De Rycke, Y. On the use of propensity scores in case of rare exposure. *BMC Medical Research Methodology* **16**, https://doi.org/10.1186/s12874-016-0135-1 (2016).
13. Lendle, S. D., Fireman, B. & van der Laan, M. J. Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology* **66**, S91–S98, https://doi.org/10.1016/j.jclinepi.2013.02.017 (2013).
14. Austin, P. C. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* **29**, 2137–2148, https://doi.org/10.1002/sim.3854 (2010).
15. Austin, P. C. & Stuart, E. A. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical Methods in Medical Research* **26**, 2505–2525, https://doi.org/10.1177/0962280215601134 (2017).
16. Austin, P. C. & Stuart, E. A. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research* **26**, 1654–1670, https://doi.org/10.1177/0962280215584401 (2017).
17. Colson, K. E. *et al*. Optimizing matching and analysis combinations for estimating causal effects. *Scientific Reports* **6**, https://doi.org/10.1038/srep23222 (2016).
18. Robins, J. M. 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* **7**, 1393–1512, https://doi.org/10.1016/0270-0255(86)90088-6 (1986).
19. Vansteelandt, S. & Keiding, N. Invited commentary: G-computation-lost in translation? *American Journal of Epidemiology* **173**, 739–742, https://doi.org/10.1093/aje/kwq474 (2011).
20. Snowden, J. M., Rose, S. & Mortimer, K. M. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology* **173**, 731–738, https://doi.org/10.1093/aje/kwq472 (2011).
21. Wang, A., Nianogo, R. A. & Arah, O. A. G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology* **17**, https://doi.org/10.1186/s12874-016-0282-4 (2017).
22. Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* **86**, 4–29, https://doi.org/10.1162/003465304323023651 (2004).
23. Bang, H. & Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973, https://doi.org/10.1111/j.1541-0420.2005.00377.x (2005).
24. van der Laan, M. J. & Rubin, D. B. Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**, https://doi.org/10.2202/1557-4679.1043 (2006).
25. Neugebauer, R. & van der Laan, M. J. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* **129**, 405–426, https://doi.org/10.1016/j.jspi.2004.06.060 (2005).
26. Brookhart, M. A. *et al*. Variable Selection for Propensity Score Models. *American Journal of Epidemiology* **163**, 1149–1156, https://doi.org/10.1093/aje/kwj149 (2006).
27. Lefebvre, G., Delaney, J. A. C. & Platt, R. W. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine* **27**, 3629–3642, https://doi.org/10.1002/sim.3200 (2008).
28. Schisterman, E. F., Cole, S. R. & Platt, R. W. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* **20**, 488–495, https://doi.org/10.1097/EDE.0b013e3181a819a1 (2009).
29. Rotnitzky, A., Li, L. & Li, X. A note on overadjustment in inverse probability weighted estimation. *Biometrika* **97**, 997–1001, https://doi.org/10.1093/biomet/asq049 (2010).
30. Schnitzer, M. E., Lok, J. J. & Gruber, S. Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *The International Journal of Biostatistics* **12**, 97–115, https://doi.org/10.1515/ijb-2015-0017 (2016).
31. Myers, J. A. *et al*. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* **174**, 1213–1222, https://doi.org/10.1093/aje/kwr364 (2011).
32. De Luna, X., Waernbaum, I. & Richardson, T. S. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* **98**, 861–875, https://doi.org/10.1093/biomet/asr041 (2011).
33. VanderWeele, T. J. & Shpitser, I. A new criterion for confounder selection. *Biometrics* **67**, 1406–1413, https://doi.org/10.1111/j.1541-0420.2011.01619.x (2011).

34. Schuler, M. S. & Rose, S. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology* **185**, 65–73, https://doi.org/10.1093/aje/kww165 (2017).
35. Kang, J. D. Y. & Schafer, J. L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539, https://doi.org/10.1214/07-STS227 (2007).
36. Porter, K. E., Gruber, S., van der Laan, M. J. & Sekhon, J. S. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics* **7**, https://doi.org/10.2202/1557-4679.1308 (2011).
37. Moher, D. *et al.* Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, c869, https://doi.org/10.1136/bmj.c869 (2010).
38. Greenland, S., Robins, J. M. & Pearl, J. Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29–46, https://doi.org/10.1214/ss/1009211805 (1999).
39. Aalen, O. O., Cook, R. J. & Røysland, K. Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis* **21**, 579–593, https://doi.org/10.1007/s10985-015-9335-y (2015).
40. Pearl, J., Glymour, M. & Jewell, N. P. *Causal Inference in Statistics: A Primer* (John Wiley & Sons, 2016).
41. Xu, S. *et al*. Use of Stabilized Inverse Propensity Scores as Weights to Directly Estimate Relative Risk and Its Confidence Intervals. *Value in Health* **13**, 273–277, https://doi.org/10.1111/j.1524-4733.2009.00671.x (2010).
42. Morgan, S. L. & Todd, J. J. A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology* **38**, 231–282, https://doi.org/10.1111/j.1467-9531.2008.00204.x (2008).
43. Zeileis, A. Object-oriented computation of sandwich estimators. *Journal of Statistical Software* **16**, 1–16, https://doi.org/10.18637/jss.v016.i09 (2006).
44. Austin, P. C. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments: Propensity scores and survival analysis. *Statistics in Medicine* **33**, 1242–1258, https://doi.org/10.1002/sim.5984 (2014).
45. Ho, D., Imai, K., King, G. & Stuart, E. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42**, 1–28, https://doi.org/10.18637/jss.v042.i08 (2011).
46. Aalen, O. O., Farewell, V. T., De Angelis, D., Day, N. E. & Gill, O. N. A markov model for hiv disease progression including the effect of hiv diagnosis and treatment: application to aids prediction in england and wales. *Statistics in Medicine* **16**, 2191–2210, https://doi.org/10.1002/(sici)1097-0258(19971015)16:19<2191::aid-sim645>3.0.co;2-5 (1997).
47. van der Laan, M. J. & Rose, S. *Targeted learning: causal inference for observational and experimental data*. Springer series in statistics (Springer, 2011).
48. Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393, https://doi.org/10.2307/2285666 (1974).
49. Luque-Fernandez, M. A., Schomaker, M., Rachet, B. & Schnitzer, M. E. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine* **37**, 2530–2546, https://doi.org/10.1002/sim.7628 (2018).
50. Gayat, E., Resche-Rigon, M., Mary, J.-Y. & Porcher, R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharmaceutical Statistics* **11**, 222–229, https://doi.org/10.1002/pst.537 (2012).
51. Morris, T. P., White, I. R. & Crowther, M. J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* **38**, 2074–2102, https://doi.org/10.1002/sim.8086 (2019).
52. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2014).
53. Barbin, L. *et al*. Comparative efficacy of fingolimod vs natalizumab. *Neurology* **86**, 771–778, https://doi.org/10.1212/WNL.0000000000002395 (2016).
54. Kalincik, T. *et al*. Switch to natalizumab versus fingolimod in active relapsing-remitting multiple sclerosis. *Annals of Neurology* **77**, 425–435, https://doi.org/10.1002/ana.24339 (2015).
55. Majdan, M. *et al*. Barbiturates Use and Its Effects in Patients with Severe Traumatic Brain Injury in Five European Countries. *Journal of Neurotrauma* **30**, 23–29, https://doi.org/10.1089/neu.2012.2554 (2012).
56. King, G. & Nielsen, R. Why propensity scores should not be used for matching. *Political Analysis* **27**, 435–454, https://doi.org/10.1017/pan.2019.11 (2019).
57. Nieto, F. J. & Coresh, J. Adjusting survival curves for confounders: a review and a new method. *American Journal of Epidemiology* **143**, 1059–1068, https://doi.org/10.1093/oxfordjournals.aje.a008670 (1996).
58. Williamson, E. J., Forbes, A. & White, I. R. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* **33**, 721–737, https://doi.org/10.1002/sim.5991 (2014).
59. Williamson, E. J., Morley, R., Lucas, A. & Carpenter, J. Propensity scores: From naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research* **21**, 273–293, https://doi.org/10.1177/0962280210394483 (2012).
60. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research* **46**, 399–424, https://doi.org/10.1080/00273171.2011.568786 (2011).
61. Haukoos, J. S. & Lewis, R. J. The propensity score. *JAMA* **314**, 1637–1638, https://doi.org/10.1001/jama.2015.13480 (2015).
62. Gillaizeau, F. *et al*. Inverse probability weighting to control confounding in an illness-death model for interval-censored data. *Statistics in Medicine* **37**, 1245–1258, https://doi.org/10.1002/sim.7550 (2018).
63. Abadie, A. & Imbens, G. W. Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–267, https://doi.org/10.1111/j.1468-0262.2006.00655.x (2006).
64. Moore, K. L., Neugebauer, R., van der Laan, M. J. & Tager, I. B. Causal inference in epidemiological studies with strong confounding. *Statistics in Medicine* **31**, 1380–1404, https://doi.org/10.1002/sim.4469 (2012).
65. Heinze, G., Wallisch, C. & Dunkler, D. Variable selection - a review and recommendations for the practicing statistician. *Biometrical Journal* **60**, 431–449, https://doi.org/10.1002/bimj.201700067 (2018).
66. Witte, J. & Didelez, V. Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal* **61**, 1270–1289, https://doi.org/10.1002/bimj.201700294 (2019).
67. Shortreed, S. M. & Ertefaie, A. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* **73**, 1111–1122, https://doi.org/10.1111/biom.12679 (2017).
68. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**, Article 25, https://doi.org/10.2202/1544-6115.1309 (2007).
69. Naimi, A. I. & Balzer, L. B. Stacked generalization: An introduction to super learning. *European journal of epidemiology* **33**, 459–464, https://doi.org/10.1007/s10654-018-0390-z (2018).
70. Pirracchio, R. & Carone, M. The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research* **27**, 2504–2518, https://doi.org/10.1177/0962280216682055 (2018).

Acknowledgements

The authors would like to thank the members of AtlanREA and OFSEP Groups for their involvement in the study, the physicians who helped recruit patients and all patients who participated in this study. We also thank the clinical research associates who participated in the data collection. The analysis and interpretation of these data are the responsibility of the authors. This work was partially supported by a public grant overseen by the

French National Research Agency (ANR) to create the Common Laboratory RISCA (Research in Informatic and Statistic for Cohort Analyses, www.labcom-risca.com, reference: ANR-16-LCV1-0003-01). The funder had no role in study design; analysis, and interpretation of data; writing the report; and the decision to submit the report for publication.

Author contributions

A.C. and Y.F. designed and conceptualised the study, conducted statistical analyses, analysed the data and drafted the manuscript for intellectual content, F.L.B., F.G. and B.G. designed and conceptualised the study, analysed the data and revised the manuscript for intellectual content, C.L. and C.R. analysed the data and revised the manuscript for intellectual content, L.B., D.L. and M.L. had a significant role in the acquisition of data and revised the manuscript for intellectual content. All authors approved the final version of the manuscript.

Competing interests

Dr. Y. Foucher has received speaking honoraria from Biogen and Sanofi. Pr. D. Laplaud has received Funding for travel or speaker honoraria from Biogen, Novartis, and Genzyme. He has participated in advisory boards in the past years Biogen-Idec, TEVA Pharma, Novartis, and Genzyme. The other authors declared no conflict of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-65917-x.

Correspondence and requests for materials should be addressed to Y.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapitre 4

Apport de l'apprentissage automatique en g-computation

> « *En science, la phrase la plus excitante que l'on peut entendre, celle qui annonce de nouvelles découvertes, ce n'est pas "Eurêka!" mais "Tiens, c'est drôle."* »
>
> Isaac Asimov

Les *Supplementary Materials* peuvent être trouvés dans l'Annexe E de ce manuscrit.

scientific reports

Check for updates

OPEN

G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes

Florent Le Borgne[1,2,6], Arthur Chatton[1,2,6], Maxime Léger[1,3], Rémi Lenain[1,4] & Yohann Foucher[1,5]✉

In clinical research, there is a growing interest in the use of propensity score-based methods to estimate causal effects. G-computation is an alternative because of its high statistical power. Machine learning is also increasingly used because of its possible robustness to model misspecification. In this paper, we aimed to propose an approach that combines machine learning and G-computation when both the outcome and the exposure status are binary and is able to deal with small samples. We evaluated the performances of several methods, including penalized logistic regressions, a neural network, a support vector machine, boosted classification and regression trees, and a super learner through simulations. We proposed six different scenarios characterised by various sample sizes, numbers of covariates and relationships between covariates, exposure statuses, and outcomes. We have also illustrated the application of these methods, in which they were used to estimate the efficacy of barbiturates prescribed during the first 24 h of an episode of intracranial hypertension. In the context of GC, for estimating the individual outcome probabilities in two counterfactual worlds, we reported that the super learner tended to outperform the other approaches in terms of both bias and variance, especially for small sample sizes. The support vector machine performed well, but its mean bias was slightly higher than that of the super learner. In the investigated scenarios, G-computation associated with the super learner was a performant method for drawing causal inferences, even from small sample sizes.

Machine learning (ML) is a set of mathematical and statistical methods that computer systems use to perform tasks without specific instructions. In medical research, there is an increasing interest in these methods for prediction and, more recently, for causality[1] There is a large intersection between these fields since the first step of causal modelling consists of predicting the exposure for propensity score (PS)-based methods[2,3] or the outcome for G-computation (GC)[4,5].

Several recent methodological studies have therefore studied the potential applicability of ML for causal inference. A large number simulation-based studies have compared several ML methods to obtain PSs[1,6–10]. While the corresponding PS-based results were very encouraging, GC was compared to PS-based methods in the context of classical regression models and showed several advantages in terms of statistical power[11–14] and robustness of the estimates regardless of the set of included covariates[11]. However, simulation-based studies related to the use of ML for predicting outcomes in GC are infrequent. Austin examined the use of ensemble-based methods (bagged classification and regression trees (CART), random forests, and boosted CART (BCART)) and concluded that BCART was the highest performing algorithm[15]. He also concluded that BCART had a lower bias when it was used to impute potential outcomes than when it was used to estimate the PS for inverse probability treatment weighting.

In this paper, we studied the performances of GC in combination with different ML algorithms, including a super learner (SL), through simulations to estimate causal effects. Many of the previous studies were based on

[1]INSERM UMR 1246 - SPHERE, Nantes University, Tours University, 22 Boulevard Bénoni Goullin, 44200 Nantes, France. [2]IDBC-A2COM, Pacé, France. [3]Département D'Anesthésie Réanimation, Centre Hospitalier Universitaire D'Angers, Angers, France. [4]Lille University Hospital, Lille, France. [5]Nantes University Hospital, Nantes, France. [6]These authors contributed equally: Florent Le Borgne and Arthur Chatton. ✉email: yohann.foucher@univ-nantes.fr

large samples. Therefore, we made sure to include scenarios with small sample sizes. We limited our study to case where both the exposure and outcome were binary and to small-medium sample sizes. We also focused on ML techniques that are applicable in daily practice, i.e., with reasonable computation times on modern laptops or workstations.

Methods

G-computation. Let $Y(1)$ and $Y(0)$ be the two potential outcomes under the exposure and the non-exposure, respectively[16]. Let (Z, X) denote the random variables related to the exposure statuses of individuals ($Z = 1$ for exposed individuals and 0 otherwise) and the k covariates ($X = X_1, \ldots, X_k$) measured before exposure, respectively. The average causal effect is $ACE = E[Y(1) - Y(0)]$. It represents the mean difference between the outcomes of individuals if they had been exposed or unexposed[17].

Suppose (Y_i, Z_i, X_i) a dataset for analysis consists of n independent realisations of (Y, Z, X). The first step of GC is to fit $f(Y|Z, X)$, and this outcome model is frequently referred to as the Q-model[5]. Once estimated, the Q-model aims to predict, for each individual i ($i = 1, \ldots, n$), the two potential outcomes under each exposure status by maintaining her/his covariates X_i at the observed values and setting Z_i to 1 and 0: $\widehat{Y}_i(1) = \widehat{f}(Y|1, X_i)$ and $\widehat{Y}_i(0) = \widehat{f}(Y|0, X_i)$. The average causal effect is then estimated by $\widehat{ACE} = n^{-1}\sum_{i=1}^{n}\left[\widehat{Y}_i(1) - \widehat{Y}_i(0)\right]$.

Covariates selection. One of the main differences between prediction and causality is the selection of covariates. Knowledge of the causal relationship structure is essential for conducting causal inference[18]. This knowledge consists of excluding the mediators, colliders[19], and instrumental variables[20,21]. Note that a benefit of GC over PS-based methods is that it more effectively prevents instrumental variables, which are often included in the PS. In this context, the advantages and limits of ML algorithms have been well described[22,23]. As noted by VanderWeele and Shpitser[24], investigators can identify the causes of exposure statuses or outcomes as potential covariates.

Unfortunately, full knowledge of causal relationships is often unavailable. There is a growing literature about the best set of covariates to consider, and it recommends including all the covariates that cause the outcome[11,21,25]. The corresponding data-driven selection procedure for GC is straightforward since it corresponds to the predictors of the Q-model.

ML techniques. In contrast with PS-based methods, which consist of predicting exposure statuses, the Q-model must keep the exposure status as one of the predictors. This is not possible for several ML techniques, such as random forests, except by estimating $f(.)$ separately for the exposed and unexposed individuals. Nevertheless, this solution is not reasonable for small sample sizes (we have tested it, and the results confirm its deficient performances for $n < 1000$; data not shown). Below, we briefly describe the ML methods that we included in our simulations. For more details on these ML techniques, see McNeish for the penalized methods[26], and Bi et al. for the other methods[27]. We performed all the analyses using R version 3.6.1.

Lasso logistic regression (LLR). L1 regularisation allows for the selection of the predictors. To obtain a flexible model, we considered all the possible interactions between the exposure status Z and covariates X. Moreover, we used b-splines for the quantitative variables of the vector X. We used the *glmnet* function included in the *glmnet* package.

Elasticnet logistic regression (ELR). We used the same flexible logistic regression as previously defined, but with both the L1 and L2 regularisations (two tuning parameters).

Neural network (NN). We chose a neural network with one hidden layer, as this is probably the most common network architecture[3]. Its size constitutes the single tuning parameter. We used the *nnet* function of the *nnet* package.

Support vector machine (SVM). We chose the radial basis function kernel to flex the linear assumption. We used the *svmRadial* function of the *kernlab* package with two tuning parameters: the cost penalty of misclassification and the flexibility of the classification.

Boosted CART (BCART). This ML technique is an ensemble method, that is, a method that averages the percentages of events in the terminal nodes of several tree partitions. Four tuning parameters must be chosen: the number of trees, the highest level of covariate interactions, the learning rate, and the minimum number of observations in the terminal nodes. We used the *gbm* function included in the *gbm* package.

For the five methods listed above (LLR, ELR, NN, SVM, and BCART), we chose their respective tuning parameters by maximising the average area under the receiver operating characteristic curve (AUC) of tenfold cross-validation. We used the *caret* package with a tuning grid of length equals 20.

Super learner (SL). We included the previous ML techniques in the SL, with the exception of BCART due to the resulting computational burden. The SL consists of averaging the predictions obtained from the four approaches by using a weighted linear predictor[28]. In agreement with our previous choice, we estimated the weights by maximising the average AUC of tenfold cross-validation. We used the *SuperLearner* package.

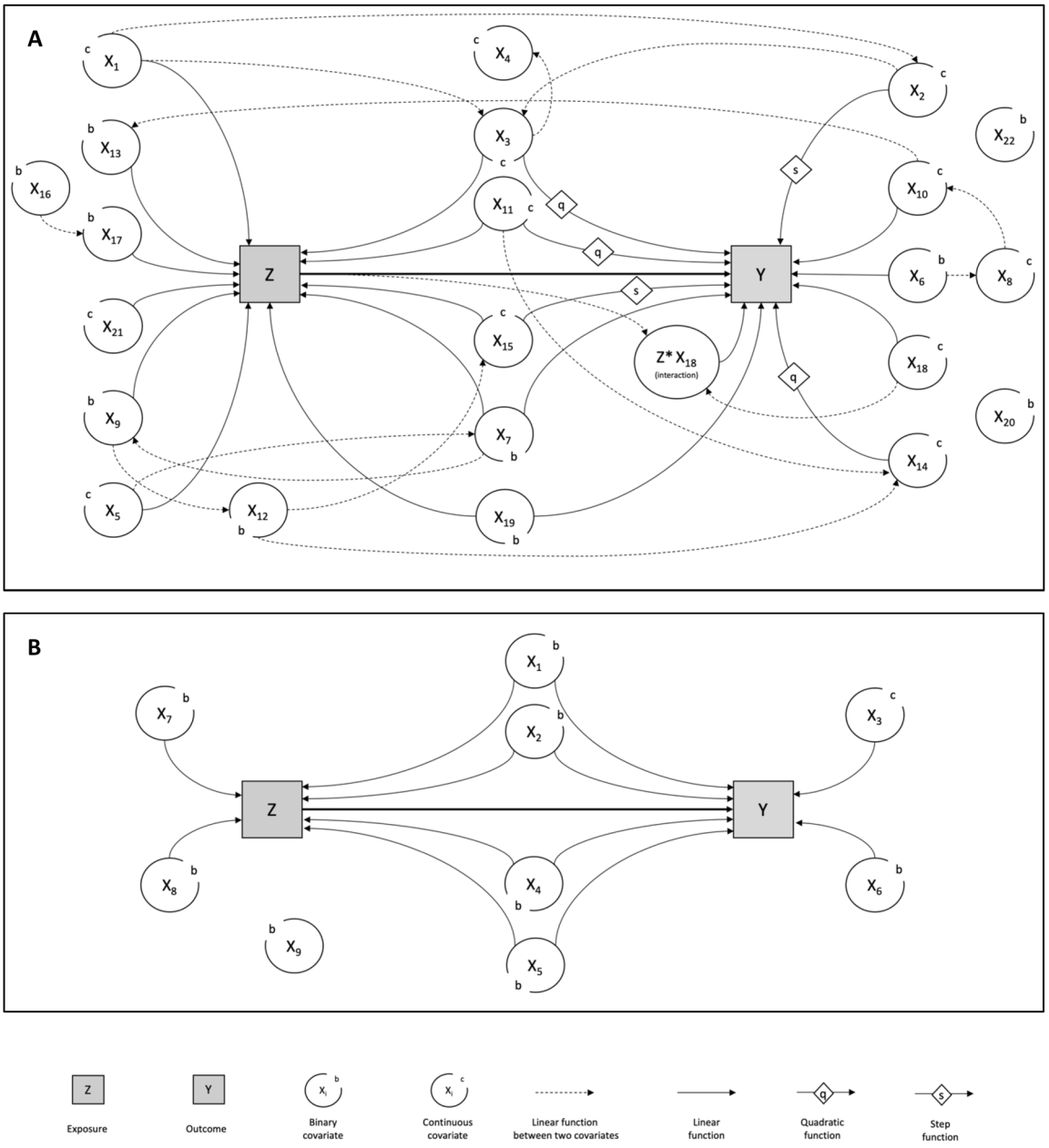


Figure 1. Directed acyclic graphs associated with the two simulated scenarios. (**A**) The realistic scenario with 22 covariates, linear and nonlinear relationships, and one interaction. (**B**) The simplistic scenario with nine covariates, linear relationships, and no interaction.

Variance estimation. By bootstrapping the entire procedure[29], one can obtain the standard error and the confidence interval of the *ACE*. Regarding the corresponding computational burden, a compromise consists of choosing the tuning parameters based on the entire sample and then using these values in the subsequent bootstrap samples[30,31]. Moreover, to consider the possible overfitting associated with such ML techniques, we performed a bootstrap cross-validation procedure. We trained the ML algorithms from the bootstrap sample, while we estimated the *ACE* from the individuals not included in the bootstrap sample. In this paper, we performed 500 iterations.

Simulation-based study

Data generation. We considered two main scenarios, as illustrated in Fig. 1 (the related models are in Supplementary Tables S1 and S2). First, we simulated the continuous and binary covariates from X_1 to X_k, allowing for dependences between the simulated covariate and those already generated. Second, we obtained Z and Y with Bernouilli distributions. The logit of the corresponding probabilities equaled the linear functions of X and (X, Z).

We choose two contrasting scenarios. We defined a realistic situation (Fig. 1A, Supplementary Table S1) with 22 correlated covariates at baseline. Nine covariates were included in the outcome model, among which one

covariate interacted with the exposure effect, two effects were step functions, three were quadratic functions, and four were linear. In contrast, we defined a simplistic situation (Fig. 1B, Supplementary Table S2) with nine independent covariates. Six covariates were included in the outcome model with linear effects and no interaction.

We simulated all the covariates X as variables measured before exposure. We did not consider mediators and colliders. As previously stated, the investigator must exclude these variables from the set of covariates. We studied different sample sizes: $n = 100$, 500, and 1000. For each scenario, we randomly generated 10,000 datasets.

Performance criteria. We computed the theoretical *ACE* by averaging the *ACE* estimations obtained from the univariate logistic models (with Z as the only explanatory variable) fitted based on datasets that were simulated as above, except that Z was generated independently of X[11,32]. We reported the following criteria (the formulae can be found in the Supplementary Materials): the mean bias (MB), the root mean square error (RMSE), the empirical standard deviation (ESD), the asymptotic standard deviation (ASD), the variance estimation bias (VEB), the empirical coverage rate of the nominal 95% confidence interval (95% CI), and the statistical power. We compared the performances of the previous ML techniques. In addition, we examined the results and compared them with those obtained by a perfectly specified LR, i.e., a LR with the same linear predictor as the one defined in the last lines of Supplementary Tables S1 and S2, in which we only estimated the corresponding regression coefficients.

Comparison of the ML techniques in terms of bias. *Overall results.* To evaluate the calibration of the ML methods for the simulated data, we added calibration plots of 10 simulated datasets for each combination of methods (LLR, ELR, NN, SVM, SL), complexity (simplistic, realistic), and sample size ($n = 100, 500, 1000$) to the Supplementary Materials (Figures S1-10). One can observe an overfitting of the ELR, SVM, and SL when $n = 100$, and this can be explained by the fact that the number of parameters was too large compared to the sample size.

We report the simulation results in Figs. 2, 3 and 4 for the realistic and simplistic scenarios (the numerical details can be found in Supplementary Tables S3 and S4). Independent of the sample size and the complexity of the relationships between the covariates and the outcome, BCART was associated with a significant level of bias, with the MB being higher than 3%.

The impact of the sample size in the realistic situation. To differentiate between the other methods, one can compare the MBs obtained when the relationships between the covariates and outcome are difficult for the analyst to manage, i.e., a realistic situation. When the learning support is small ($n = 100$), the penalized methods (ELR and LLR) and the NN resulted in unacceptable MBs higher than 4%. In contrast, the two remaining methods (SVM and SL) were associated with values lower than 1%. With large sample sizes ($n \geq 500$), the four methods performed correctly with MBs less than 3%, and the lowest MB was obtained with the SL (MB < 1% for all sample sizes). To further discriminate between the SVM and SL in this realistic situation, one can notice that the MB remained negligible for the SL regardless of the sample size, while for the SVM, the MB increased with the sample size (values between 1 and 2% when $n \geq 500$).

The impact of the sample size in the simplistic situation. Except when $n = 1000$, for which they were outperformed by the SL (MB < 1%), the penalized methods were associated with the smallest biases in the simplistic situation, with MBs less than 1% regardless of the sample size. The penalized methods were even the only methods such low values when $n = 100$. The NN was the only method with no significant variations according to the sample size (i.e., MBs between 1 and 2% for all three sample sizes).

Comparison of the ML techniques in terms of variance. *Overall results.* Regardless of the scenarios and the sample sizes used, one can observe an underestimation of the variance using BCART. Its VEB ranged from − 2 to − 56%.

The impact of the sample size in the realistic situation. To differentiate between the other methods, one can first consider the smallest sample size (n =100). The penalized approaches (LLR and ELR) resulted in the highest estimations of the variance, with ASDs close to 0.10. The SVM and NN were associated with the smallest variances, with ASDs close to 0.6 (the VEBs were − 6.4% and 8.8%, respectively). Compared with the two previous ML techniques, the SL resulted in a slightly higher ASD at 0.7, but a lower VEB at − 3.7%. For larger sample sizes ($n \geq 500$), the results in terms of variance were close for the four following approaches: LLR, ELR, SVM and SL. The NN was associated with an unacceptable overestimation of the variance (VEB = 19.0% and 31.1% for $n = 500$ and 1000, respectively).

The differences between the realistic and simplistic situations. The results were similar when the relationships between the covariates and the outcome were easier for the analyst to model (i.e., the simplistic situation). However, one can underline an exception: when $n = 100$, the NN resulted in an ASD close to those of the penalized approaches.

Synthesis of bias and variance in terms of the root mean square error and coverage. Even if BCART resulted in a critical level of bias, its RMSEs were reasonable, and this is mainly because of the previously reported underestimation of the variance. This bias associated with an underestimated variance resulted in coverage ranging from 57.2 to 82.2%, and the upper bound of this range is considerably lower than the nominal

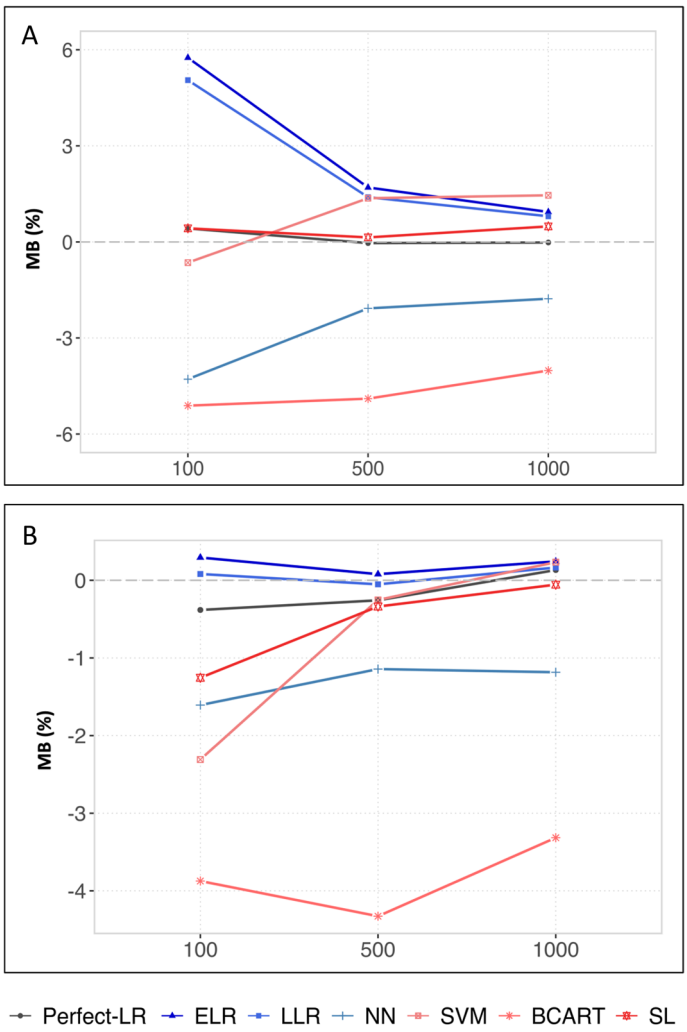


Figure 2. Mean biases (MBs) of G-computation in realistic (**A**) and simplistic (**B**) situations with the following Q-models: the theoretical logistic regression, elasticnet logistic regression, lasso logistic regression, neural network, support vector machine, boosted CART and super learner.

value of 95%. For the smallest sample size, in both the realistic and simplistic situations, the RMSEs of the penalized methods were among the highest because of their high-level of variance (simplistic situation) or high levels of bias (realistic situation).

When $n \geq 500$, the RMSEs of the penalized methods were close to those observed for the ML-based methods (NN, SVM and SL). However, for these two approaches, one can observe slightly anti-conservative 95% CIs in the realistic situation, because of their slight biases. For the remaining ML-based methods, the RMSEs were comparable for the three sample sizes and in the two situations, but the results of the NN should be interpreted with caution. Indeed, for $n = 100$, the NN was associated with a significant bias, but a low variance estimation, resulting in a CI of 86.6%, lower than the nominal value of 95%.

As previously reported, the two remaining methods (SL and SVM) were the two ML techniques associated with the smallest MBs. For each scenario, the MB of the SL was even lower than the value of SVM. This explains why the nominal coverage was slightly higher when using the SL. For instance in the realistic scenario, the coverage values associated with the SVM were 92.6%, 93.7% and 91.4% for n =100, 500 and 1000, respectively, while they were 93.1%, 95.2% and 94.6% for the SL.

Power of the unbiased methods. We only consider the methods and the scenarios in which the MB were lower than 1% due to the problems encountered when interpreting the power in the presence of bias.

The realistic situation. When $n = 100$, the SVM and SL had MBs lower than 1%. Of the two methods, the best power was achieved by the SVM (36.5% vs 30.8% for the SL). When $n = 1000$, the ELR, LLR and SL had MBs lower than 1%, and the best power values were achieved by the penalized methods (92.4% for the ELR, 91.5% for the LLR and 89.3% for the SL).

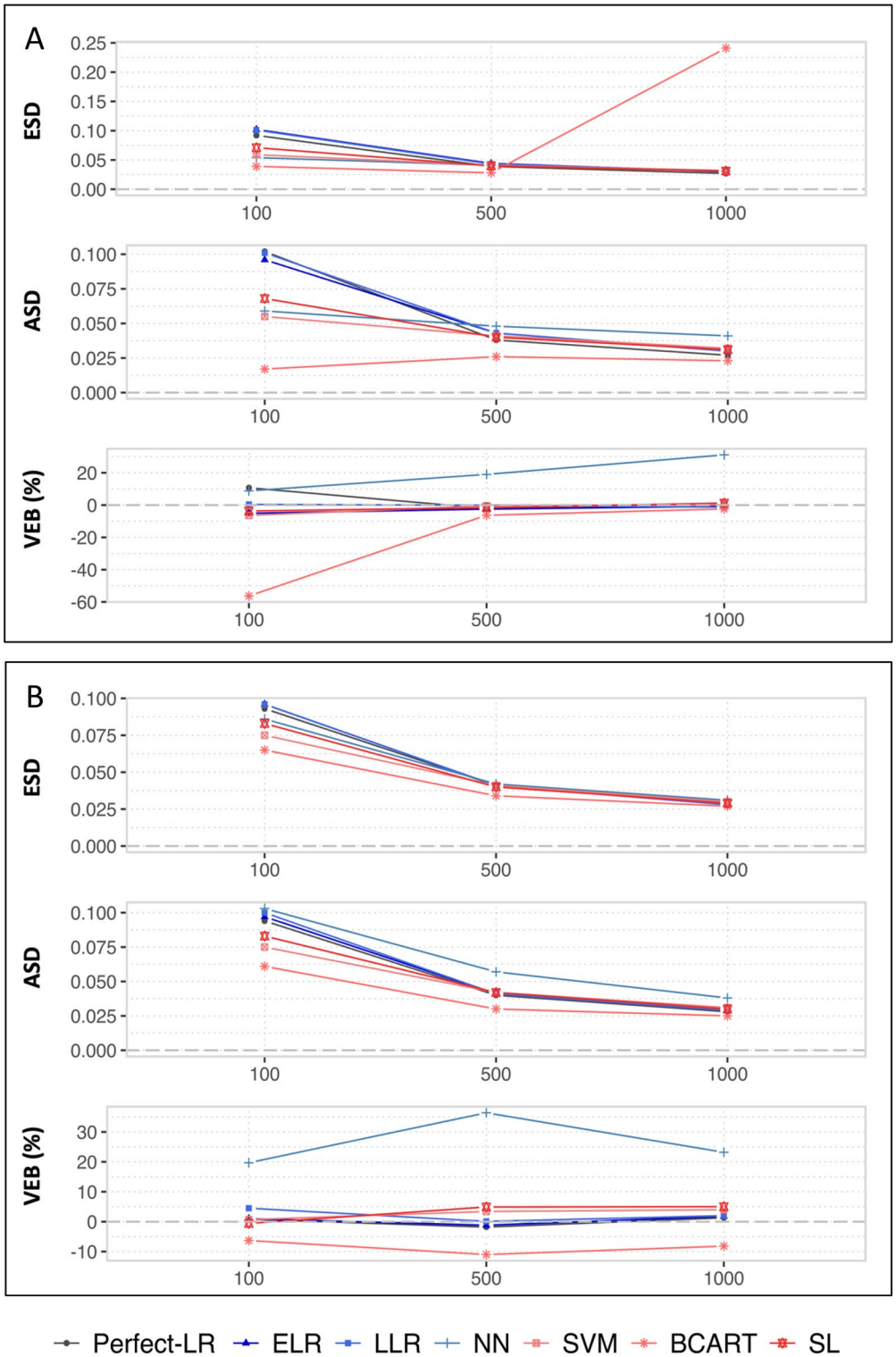


Figure 3. Empirical and asymptotic standard deviations (ESDs and ASDs, respectively) and variance estimation biases (VEBs) of G-computation in realistic (**A**) and simplistic (**B**) situations with the following Q-models: the theoretical logistic regression, elasticnet logistic regression, lasso logistic regression, neural network, support vector machine, boosted CART and super learner.

The simplistic situation. When n =100, only the penalized methods had MBs lower than 1%. The best power was obtained by the ELR (20.2% versus 18.0% for the LLR). When $n \geq 500$, we additionally observed MBs lower than 1% for the SVM and SL. The penalized methods were always associated with the best powers when compared with those of the two ML techniques with a gain between 1 and 4% depending on the scenarios.

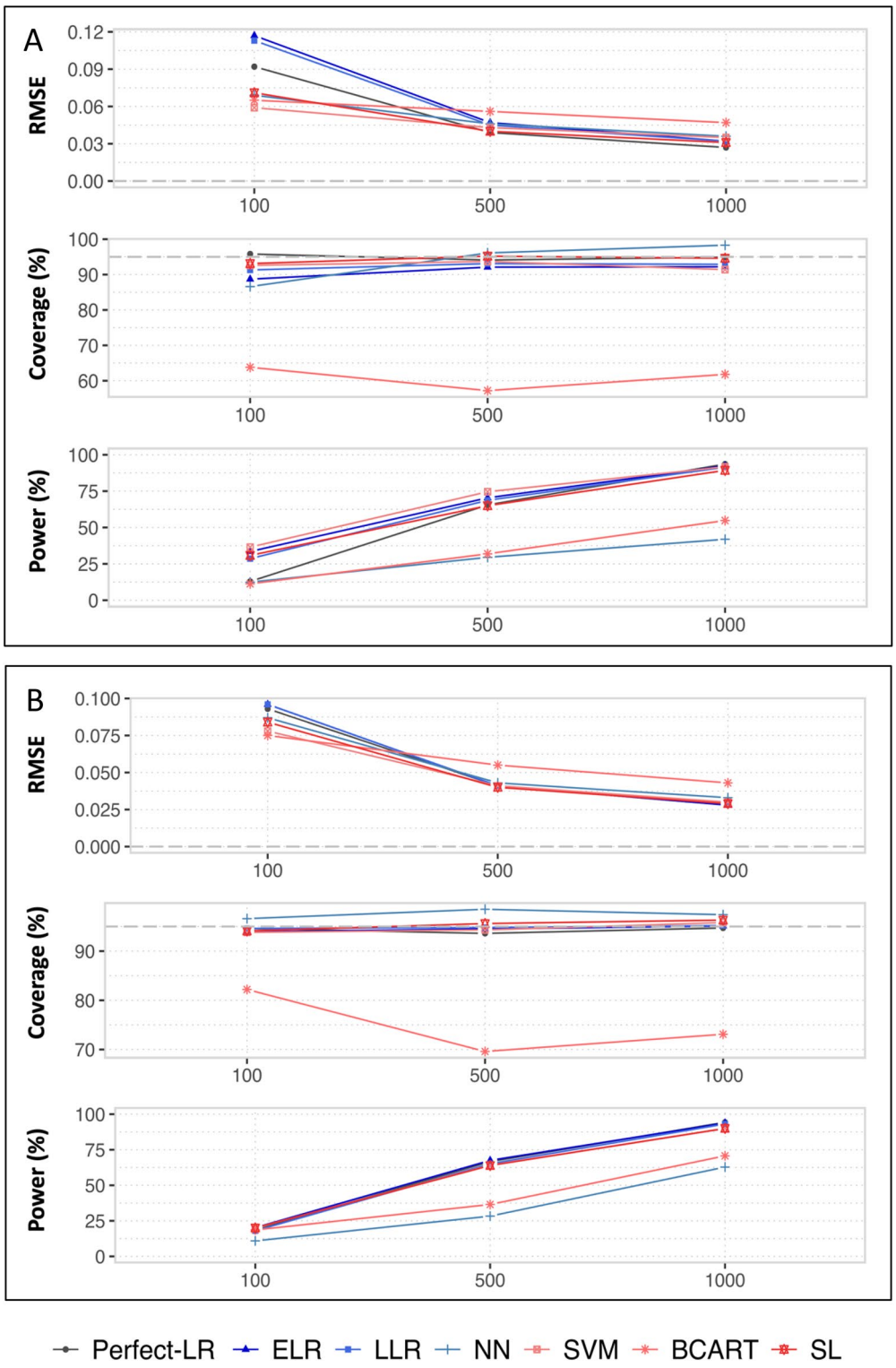


Figure 4. Root mean square errors (RMSEs), coverages and powers of G-computation in realistic (**A**) and simplistic (**B**) situations with the following Q-models: the theoretical logistic regression, elasticnet logistic regression, lasso logistic regression, neural network, support vector machine, boosted CART and super learner.

ML techniques versus the perfectly specified LR. The performances of the perfectly specified LR were better than those of the ML techniques for large sample sizes (n =1000). One can observe mean bias values close to 0%, and variance bias values close to 1%. Nevertheless, when the sample size decreased in the realistic situation, the performances of the perfectly specified LR decreased more than those of several ML techniques. When n =500, the variance bias associated with the perfectly specified LR was − 2.1% versus − 0.1% for the LLR,

− 0.5% for SVM and − 1.6% for the SL. When n =100, the variance bias associated with the perfectly specified LR was 10.7% versus 0.4% for the LLR, − 3.7% for the SL, and − 6.4% for the SVM. In this latter scenario, these three ML techniques resulted in higher statistical powers than the one obtained with the perfectly specified LR.

Application

Context. We applied the methods to evaluate the efficacy of barbiturates prescribed during the first 24 h of an episode of intracranial hypertension. The control group included patients without barbiturates at 24 h. One can use this treatment to decrease refractory intracranial pressure, but its effectiveness remains debated due to the associated adverse events (*e.g.*, haemodynamic impacts or infectious complications).

We used data from the French prospective cohort AtlanREA. We considered patients with intracranial pressures higher than 20 mmHg. We conducted this study following French law relative to non-interventional clinical research. Written informed consent was collected. Moreover, the French commission for data protection approved the collection (CNIL DR-2013-047). The study was approved by the AtlanREA scientific council (www.atlanrea.org) and the ethics committee of the French Society of Anesthesia and Intensive Care (SFAR, https://sfar.org/).

Implementation of the methods. We reduced the set of covariates to the possible causes of the outcome without considering the consequences of barbiturate use. We described this selection in detail in Supplementary Table S5. For the ML-based methods, we considered all the covariates before exposure and the corresponding interactions with the exposure status. As in the previous simulations, we used b-splines for the continuous covariates in the penalized methods. For the investigator-based method, all the outcome causes previously listed were included (Supplementary Table S5). The log-linearity assumption for continuous covariates seemed to be satisfied. We assumed that there was no interaction because of the absence of clinical relevance.

Results

Table 1 describes the 252 patients. Seventy-four patients were in the treatment group. The outcome was the proportion of patients with a favourable Glasgow Outcome Scale (GOS ≤ 3) at three months after admission to the intensive care unit. Figure 5 presents the confounder-adjusted estimates. The investigator-based approach resulted in a 17.5% decrease in the percentage of patients with favourable 3-month GOS due to barbiturates (95% CI from 6.6 to 28.4%). We observed similar results for the ELR and LLR, in terms of both the estimates and the 95% CIs. The other ML techniques resulted in lower associations, and the one fpr the NN was even nonsignificant (ACE =0.4%, 95% CI from − 3.1 to 2.4%). The SL resulted in a small but significant association (ACE =6.2%, 95% CI from 0.6% to 11.8%).

For a MacBook pro with a 2.6 GHz Intel Core i7 processor (6 cores), the results were available in 6.5 min for the ELR, 16.3 min for the LLR, 7.1 min for the NN, 2.3 min for the BCART, 2.6 min for the SVM, and 7 min for the SL.

Discussion

When modelling the outcome model for the GC in the presence of small to medium sample sizes, the results of our simulations tended to demonstrate that ML techniques allow for accurate estimations of causal effects. Overall, the SL remained robust in all situations and achieved a relevant compromise between both bias reduction and variance estimation. In contrast, the performances of the other methods tended to vary more significantly according to the complexity of the relationships between the covariates and the outcome (simplistic versus realistic situations) and the sample size. Nevertheless, in some situations, the other methods obtained better performances than those of the SL. When the sample size was small (n = 100) in the realistic scenario, the SVM had a larger MB but a smaller ASD, with an overall smaller RMSE. In this situation, the two ML techniques (SL and SVM) were even associated with lower variances than that of the perfectly specified LR. For instance, the variance bias was − 3.7% for the SL versus 10.7% for the perfectly specified LR. One can explain this result by the sample-to-sample fluctuation, which can lead to an observed structure that is different from the theoretical one. When the sample size was small in the simplistic scenario, the penalized methods (ELR and LLR) had lower MBs and similar RMSEs.

The use of ML techniques for causal inference does not preclude human intervention. In addition to the choice of the Q-model, we need to exclude the mediators, colliders and instrumental variables by considering the underlying causal structure. The use of directed acyclic graphs can help with this task[33]. We also emphasise that ML techniques do not serve as a cure-all for poor study designs or poor data quality. It is of primary importance to investigate the identifiability conditions: the exposure levels correspond to well-defined interventions, the corresponding conditional probabilities depend only on the measured covariates, and must be higher than zero. These assumptions are consistency, exchangeability, and positivity, respectively[34]. In this paper, we focused on the estimation of a causal effect given that the identifiability conditions were satisfied. In practice, the predictive performance of the Q-model is not sufficient to ensure the absence of bias in the estimation of the causal effect, which requires a precise conceptual knowledge of the causal model[35].

Perfect knowledge of the causal structure is impossible to obtain in practice. Therefore, the analyst and the investigator construct the Q-model to approximate the causal structure as closely as possible. This may involve different steps such as the transformation of the continuous covariates to respect the log-linearity assumption, the selection of the covariates, or the choice of relevant interaction(s). While the steps performed by the analyst are data-driven and stochastic, they are systematically ignored in the estimation of the effect variance[36]. The widespread interest in (human-free) ML stems from the possibility of considering a valid post-selection inference by bootstrapping the entire estimation procedure[29].

	Overall ($n = 252$)		Barbiturates treatment					Favourable GOS at three months				
			No ($n = 178$)		Yes ($n = 74$)		p	No ($n = 180$)		Yes ($n = 72$)		p
Female patient (n, %)	89	35.3	58	32.6	31	41.9	0.1592	68	37.8	21	29.2	0.1963
Diabetes (n, %)	17	6.7	15	8.4	2	2.7	0.0989	15	8.3	2	2.8	0.1122
No sological entity: severe trauma (n, %)	124	49.2	95	53.4	29	39.2	0.0403	77	42.8	47	65.3	0.0012
SAP ≤ 90 mmHg before admission (n, %)	56	22.2	36	20.2	20	27.0	0.2368	46	25.6	10	13.9	0.0442
Evacuation of subdural or extradural hematoma (n, %) (*)	41	16.3	33	18.5	8	10.8	0.1301	27	15.0	14	19.4	0.3878
External ventricular drain (n, %)	64	25.4	39	21.9	25	33.8	0.0486	48	26.7	16	22.2	0.4640
Evacuation of cerebral hematoma or lobectomy (n, %) (*)	42	16.7	28	15.7	14	18.9	0.5362	34	18.9	8	11.1	0.1345
Decompressive craniectomy (n, %) (*)	27	10.7	15	8.4	12	16.2	0.0686	21	11.7	6	8.3	0.4396
Blood transfusion before admission (n, %)	34	13.5	25	14.0	9	12.2	0.6903	26	14.4	8	11.1	0.4841
Pneumonia (n, %) (*)	29	11.5	16	9.0	13	17.6	0.0519	19	10.6	10	13.9	0.4538
Osmotherapy (n, %) (*)	112	44.4	75	42.1	37	50.0	0.2525	89	49.4	23	31.9	0.0115
GCS score ≥ 8 (n, %)	62	24.6	39	21.9	23	31.1	0.1237	37	20.6	25	34.7	0.0183
Patient age, years (mean, sd)	47.4	17.4	48.7	17.9	44.1	15.7	0.0565	50.8	16.4	38.7	16.9	0.0000
Haemoglobin, g/dL (mean, sd)	11.8	2.3	11.7	2.2	12.1	2.5	0.1824	11.8	2.4	11.9	1.9	0.7373
Platelets, counts/mm^3 (mean, sd)	206.7	78.0	207.4	79.7	205.1	74.2	0.8312	209.0	83.8	200.9	61.1	0.4589
Serum creatinine, mmol/L (mean, sd)	71.1	29.3	71.1	27.6	71.1	33.3	0.9853	72.4	32.6	67.9	18.7	0.2732
Arterial pH (mean, sd)	7.3	0.1	7.3	0.1	7.3	0.1	0.0978	7.3	0.1	7.3	0.1	0.6317
Serum proteins, g/L (mean, sd)	58.2	10.4	57.7	10.6	59.6	9.7	0.1662	58.0	10.7	58.8	9.7	0.5963
Serum urea, mmol/L (mean, sd)	5.0	2.5	5.2	2.7	4.7	1.8	0.1827	5.2	2.3	4.5	2.9	0.0505
PaO_2/FiO_2 ratio (mean, sd)	302.7	174.0	292.7	154.7	326.6	212.9	0.1595	282.1	172.4	354.2	168.4	0.0028
SAPS II score (mean, sd)	47.6	11.4	47.6	10.7	47.6	12.9	0.9847	49.9	10.8	41.8	10.7	0.0000

Table 1. Baseline characteristics of patients according to the treatment group ($n = 252$) and the GOS at three months after the treatment initiation. GOS score was dichotomised into favourable outcomes (good recovery or moderate disability) or unfavourable outcomes (severe disability, vegetative state or death). GOS, Glasgow outcome Scale; SAP, systolic arterial pressure; HICP, high intracranial pressure; GCS, Glasgow Coma Scale; PaO_2, partial arterial pressure of oxygen; FiO_2, fraction of inspired oxygen; SAPS, Simplified Acute Physiology Score. *Before HICP.

ML techniques are often associated with big data, especially in the field of causal inference[8,37,38]. Nevertheless, we described the acceptable properties of the SL used in a GC framework to provide causal inference conclusions from databases including several hundred subjects. To obtain this result, we first selected several simple ML techniques. We excluded deep learning techniques, such as neural networks with multiple hidden layers. Second, we retained the ML techniques that allow for maintaining the exposure as one of the predictors. Third, we included two parametric models. Fourth, we used bootstrap cross-validation to prevent overfitting. Fifth, we used two ML techniques (NN and SVM) for which there was no selection of predictors. Consequently, all covariates were also included in the SL, even those with low contributions due to having no association. The removal of confounders in GC can result in confounding bias, which can explain the poor performances of the penalized methods in realistic situations. These choices participated in the lower bias of the SL versus that of BCART. Our GC results are in agreement with the conclusions of Gruber et al., which concerned PS-based analyses[8]. Indeed, BCART is an ensemble learning method that avoids cross-validation by a single partitioning of the data into training and validation sets. It allows us to reduce the computational time, but it should be used with caution for small sample sizes.

Our study suffered from limitations. First, the results from the simulations cannot be generalised to all situations. Even if they are consistent with the current literature related to the use of ML in PS-based analyses, theoretical arguments are missing for generalisation purposes. Second, one perspective of our work is to improve the proposed SL with additional ML techniques or differently tuned techniques. For instance, we fixed the length of the tuning grid at 20; a lower value may be acceptable for reducing the computational time. The V-fold cross-validation is also an important parameter. We fixed $V = 10$, as conventionally used. A more appropriate choice could also be studied. For example, Naimi and Balzer recommended increasing V as the sample size decreases[22]. Third, we focused on the comparison of the ML techniques used in GC. We did not perform comparisons with other methods used for causal inference, such as the influence function-based or doubly robust estimators. In particular, the double/debiased machine learning and targeted maximum likelihood estimator allow for the unrestricted use of data-adaptive methods[38]. The principle is to combine the modelling of the outcome and exposure mechanisms to obtain an unbiased estimate when at least one of the two models is well-specified. However, such doubly robust estimators also have several drawbacks. If both models are misspecified, the estimation is more biased than that of a single-robust estimator such as GC[14]. The inclusion of a mediator also leads to more bias than that of GC[39]. Several studies have additionally reported that GC has a lower variance than those of doubly robust estimators[11–14]. As previously stated, the use of GC also represents a partial solution for preventing the selection of instrumental variables since it is independent of the exposure modelling. Fourth, our study focused

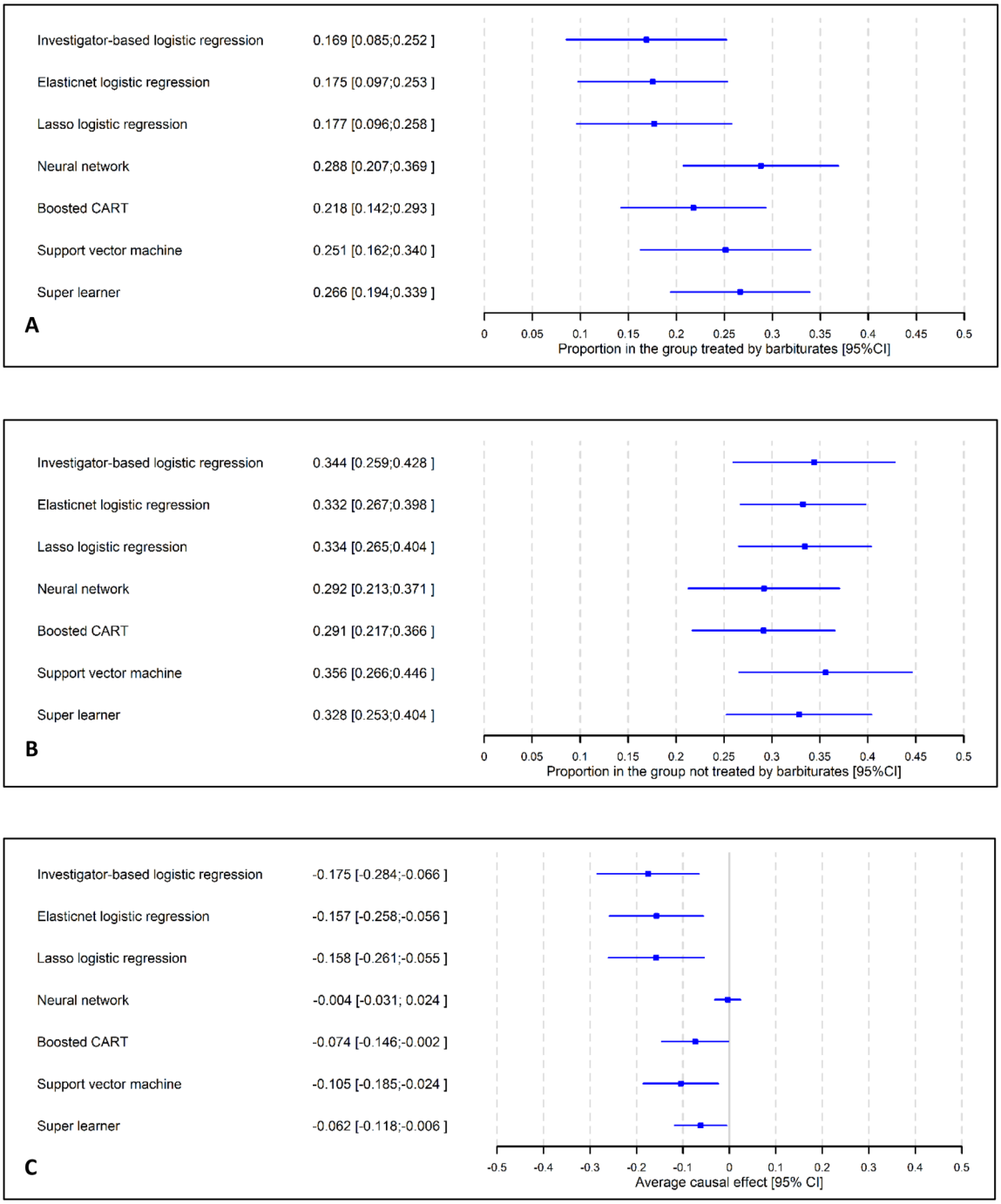


Figure 5. Estimations of the confounder-adjusted proportions of patients with favourable GOS among the patients treated with barbiturates (**A**), patients not treated with barbiturates during the first 24 h postadmission (**B**), and the corresponding average causal effects (**C**).

on the situation where both the exposure status and the outcome are binary. The generalisation of our approach to other contexts, especially for time-to-event outcomes, represents a short-term goal. Finally, we focused on the *ACE* if the entire sample had been exposed and if it had not been exposed. Additional analyses are needed to confirm these results to estimate the average causal effect only for the exposed individuals[40].

In conclusion, the super-learned G-computation is a promising method for causal inference, even with only several hundred subjects. The SVM represents an interesting alternative for small sample sizes with one hundred subjects when the relationships between the covariates and the outcome are complex. For such a small sample size, penalized methods appeared to be the best alternatives when the relationships were simplistic (few covariates with linear relationships and without interactions). The computation times of these ML techniques associated with GC were reasonable. Note that GC with the SL as the Q-model is implemented in the *RISCA* package (cran.r-project.org, version ≥ 0.82). The user can set the number of splits for cross-validation and the number of

parameter combinations to be evaluated. This is a particular solution, but it is not recommended for analysing any type of data using the same algorithm. We believe that such ML techniques constitute an opportunity for analysts to save some of their time used for repetitive modelling steps and use it for applying prior knowledge of the medical field and improving their comprehension of the given data structure.

Received: 6 July 2020; Accepted: 24 December 2020
Published online: 14 January 2021

References

1. Blakely, T., Lynch, J., Simons, K., Bentley, R. & Rose, S. Reflection on modern methods: when worlds collide: prediction, machine learning and causal inference. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyz132.
2. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
3. Westreich, D., Lessler, J. & Funk, M. J. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* **63**, 826–833 (2010).
4. Robins, J. M. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7**, 1393–1512 (1986).
5. Snowden, J. M., Rose, S. & Mortimer, K. M. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am. J. Epidemiol.* **173**, 731–738 (2011).
6. Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J. & Cook, E. F. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol. Drug Saf.* **17**, 546–555 (2008).
7. Lee, B. K., Lessler, J. & Stuart, E. A. Improving propensity score weighting using machine learning. *Stat. Med.* **29**, 337–346 (2010).
8. Gruber, S., Logan, R. W., Jarrín, I., Monge, S. & Hernán, M. A. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat. Med.* **34**, 106–117 (2015).
9. Pirracchio, R., Petersen, M. L. & van der Laan, M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am. J. Epidemiol.* **181**, 108–119 (2015).
10. Cannas, M. & Arpino, B. A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biom. J.* **61**, 1049–1072 (2019).
11. Chatton, A. *et al.* G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Sci. Rep.* (i**n press**) (2020).
12. Lendle, S. D., Fireman, B. & van der Laan, M. J. Targeted maximum likelihood estimation in safety analysis. *J. Clin. Epidemiol.* **66**, S91-98 (2013).
13. Colson, K. E. *et al.* Optimizing matching and analysis combinations for estimating causal effects. *Sci. Rep.* **6**, 23222 (2016).
14. Kang, J. D. Y. & Schafer, J. L. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat. Sci.* **22**, 523–539 (2007).
15. Austin, P. C. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivar. Behav. Res.* **47**, 115–135 (2012).
16. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701 (1974).
17. Hernán, M. A. A definition of causal effect for epidemiological research. *J. Epidemiol. Commun. Health* **58**, 265–271 (2004).
18. Lin, S.-H. & Ikram, M. A. On the relationship of machine learning with causal inference. *Eur. J. Epidemiol.* https://doi.org/10.1007/s10654-019-00564-9 (2019).
19. VanderWeele, T. J. Principles of confounder selection. *Eur. J. Epidemiol.* **34**, 211–219 (2019).
20. Myers, J. A. *et al.* Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am. J. Epidemiol.* **174**, 1213–1222 (2011).
21. Brookhart, M. A. *et al.* Variable selection for propensity score models. *Am. J. Epidemiol.* **163**, 1149–1156 (2006).
22. Naimi, A. I. & Balzer, L. B. Stacked generalization: an introduction to super learning. *Eur. J. Epidemiol.* **33**, 459–464 (2018).
23. Keil, A. P. & Edwards, J. K. You are smarter than you think: (super) machine learning in context. *Eur. J. Epidemiol.* **33**, 437–440 (2018).
24. VanderWeele, T. J. & Shpitser, I. A new criterion for confounder selection. *Biometrics* **67**, 1406–1413 (2011).
25. Lefebvre, G., Delaney, J. A. C. & Platt, R. W. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Stat. Med.* **27**, 3629–3642 (2008).
26. McNeish, D. M. Using lasso for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. *Multivar. Behav. Res.* **50**, 471–484 (2015).
27. Bi, Q., Goodman, K. E., Kaminsky, J. & Lessler, J. What is machine learning? A primer for the epidemiologist. *Am. J. Epidemiol.* https://doi.org/10.1093/aje/kwz189 (2019).
28. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, Article25 (2007).
29. Efron, B. Estimation and accuracy after model selection. *J. Am. Stat. Assoc.* **109**, 991–1007 (2014).
30. Schumacher, M., Binder, H. & Gerds, T. Assessment of survival prediction models based on microarray data. *Bioinformatics* **23**, 1768–1774 (2007).
31. Foucher, Y. & Danger, R. Time dependent ROC curves for the estimation of true prognostic capacity of microarray data. *Stat. Appl. Genet. Mol. Biol.* **11**, Article 1 (2012).
32. Gayat, E., Resche-Rigon, M., Mary, J.-Y. & Porcher, R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm. Stat.* **11**, 222–229 (2012).
33. Hernán, M. A., Hsu, J. & Healy, B. A second chance to get causal inference right: a classification of data science tasks. *Chance* **32**, 42–49 (2019).
34. Hernán, M. A. & Taubman, S. L. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int. J. Obes.* **32**, S8–S14 (2008).
35. Díaz, I. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics* **21**, 353–358 (2020).
36. Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. Valid post-selection inference. *Ann. Stat.* **41**, 802–837 (2013).
37. Wyss, R. *et al.* Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology* **29**, 96–106 (2018).
38. Karim, M. E., Pang, M. & Platt, R. W. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm?. *Epidemiology* **29**, 191–198 (2018).
39. Keil, A. P. *et al.* Resolving an apparent paradox in doubly robust estimators. *Am. J. Epidemiol.* **187**, 891–892 (2018).
40. Pirracchio, R. *et al.* Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Stat. Methods Med. Res.* **25**, 1938–1954 (2016).

Acknowledgements
The authors would like to thank the members of AtlanREA group for their involvement in the study, the physicians who helped recruit patients and all patients who participated in this study. We also thank the clinical research associates who participated in the data collection. The analysis and interpretation of these data are the responsibility of the authors.

Author contributions
Y.F. supervised this work and performed the simulations and other statistical analyses. F.L.B. and A.C. participated in the design of the simulation-based study. All the authors were engaged in the writing of the final proposal.

Funding
The French National Research Agency (ANR) partially supported this work (Research in Informatic and Statistic for Cohort Analyses, www.labcom-risca.com, reference: ANR-16-LCV1-0003-01). The funder had no role in study design; analysis, and interpretation of data; writing the report; and the decision to submit the report for publication.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information is available for this paper at https://doi.org/10.1038/s41598-021-81110-0.

Correspondence and requests for materials should be addressed to Y.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapitre 5

G-computation et standardisation doublement robuste pour temps d'évènement censuré

> « *Placez votre main sur un poêle une minute et ça vous semble durer une heure.*
> *Asseyez vous auprès d'une jolie fille une heure et ça vous semble durer une minute.* »
>
> Albert Einstein

Les *Supplementary Materials* peuvent être trouvés dans l'Annexe F de ce manuscrit.

Journal Title
XX(X):1–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

G-computation and doubly robust standardisation for continuous-time data: a comparison with inverse probability weighting

Arthur Chatton[1,2], Florent Le Borgne[1,2], Clémence Leyrat[3,4] and Yohann Foucher[1,5]

Abstract

In time-to-event settings, g-computation and doubly robust estimators are based on discrete-time data. However, many biological processes are evolving continuously over time. In this paper, we extend the g-computation and the doubly robust standardisation procedures to a continuous-time context. We compare their performance to the well-known inverse-probability-weighting (IPW) estimator for the estimation of the hazard ratio and restricted mean survival times difference, using a simulation study. Under a correct model specification, all methods are unbiased, but g-computation and the doubly robust standardisation are more efficient than inverse probability weighting. We also analyse two real-world datasets to illustrate the practical implementation of these approaches. We have updated the `R` package `RISCA` to facilitate the use of these methods and their dissemination.

Keywords

Causal inference, Parametric g-formula, Propensity score, Restricted mean survival time, Simulation study.

1 Introduction

Real-world evidence is scientific evidence obtained from data collected outside the context of randomised clinical trials.[1] The absence of randomisation complicates the estimation of the marginal causal effect (hereafter referred to merely as causal effect) of exposure (including treatment or intervention) due to a potential risk of confounding.[2] Rosenbaum and Rubin[3] introduced the propensity score (PS) as a tool for causal inference in the presence of measured confounders. In a binary exposure setting, it has been shown that the estimated PS is a balancing score, meaning that conditional on the estimated PS, the distribution of covariates is similar for exposed and unexposed patients. Following this property, the PS can be used in four ways to provide estimates of the causal exposure effect: matching, stratification, adjustment, and inverse-probability-weighting (IPW).[4] Stratification leads to residual confounding and adjustment relies on strong modelling assumptions.[5,6] Although matching on PS has long been the most popular,[7] IPW appears to be less biased and more precise in several studies.[8–10] Moreover, King and Nielsen[11] argued for halting the use of PS matching for many reasons, including covariate imbalance, inefficiency, model dependence, and bias. Indeed, matching on the PS is limited by the exclusion of patients without a suitable match leading to a non-representative population because of a change in the covariates distribution and a loss of statistical power.

[1]INSERM UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France
[2]IDBC-A2COM, Pacé, France
[3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[4]Inequalities in Cancer Outcomes Network (ICON), London School of Hygiene and Tropical Medicine, London, UK
[5]Centre Hospitalier Universitaire de Nantes, Nantes, France

Corresponding author:
Arthur Chatton, INSERM UMR 1246 - SPHERE, Institut de Recherche en Santé 2, 22 Boulevard Benoni-Goullin, 44200 NANTES, France
Email: arthur.chatton@univ-nantes.fr

Causal effects can also be estimated using the g-computation (GC), a maximum likelihood substitution estimator of the g-formula.[12,13] While IPW is based on exposure modelling, the GC relies on the prediction of the potential outcomes for each subject under each exposure status. Extensions of GC with time-to-event outcomes were recently proposed in a discrete-time setting.[14–16] Only Breskin *et al.*[16] noted an extension for continuous-time setting, *i.e.*, with infinitely short time intervals,[17] but without investigating its properties. Discrete-time models lead to estimates that depend of the length of the intervals of time, may generate interval censoring, and are often biologically implausible.[18] Furthermore, the non-collapsibility of estimands due to the self-induced selection bias increases with the length of the intervals of time.[17]

With time-to-event outcomes, the presence of right-censoring and its magnitude is of prime importance since a small number of observed events due to censoring may impact the estimation of the outcome model involved in the GC. By contrast, the IPW may perform well as long as the number of exposed patients is sufficient to estimate the PS and the total sample size is sufficiently large to limit variability in the estimated weights. To overcome potential model misspecifications, doubly robust estimators (DREs) were proposed. DREs combine both the GC and PS to obtain an unbiased estimate when at least one of the two working/nuisance models (*i.e.*, a model needed to estimate the target parameter but not estimating it itself[19]) is well-specified.[20,21] Similarly to GC, current implementation of DREs focus on discrete-time data. Therefore, we present an extension as proposed by Vansteelandt and Keiding.[22]

Several studies (see[23] and references therein) compared the IPW, GC and DRE in different contexts. They reported a lower variance for the GC than both the IPW and DRE. Nevertheless, to the best of our knowledge, no study has focused on time-to-event outcomes.

In the present paper, we aimed to detail the statistical framework for using the GC in time-to-event analyses. We restricted our developments to time-invariant confounders, and we refer the readers to Wen *et al.*[15] for a recent study of GC with time-to-event outcomes and time-varying exposure. An equivalent framework for IPW can be found in Hernán *et al.*.[24] We also compared the performances of the GC, IPW and DRS. The rest of this paper is structured as follows. In section 2, we detail the methods. Section 3 presents the design and findings of a simulation study. In section 4, we propose a practical comparison with two real-world applications related to treatment evaluations in multiple sclerosis and kidney transplantation. Finally, we discuss the results and provide practical recommendations to help analysts to choose the appropriate analysis method.

2 Methods

2.1 Notations

Let $(T_i, \delta_i, A_i, L_i)$ be the random variables associated with subject i $(i = 1, ..., n)$. n is the sample size, T_i is the participating time, δ_i is the censoring indicator (0 if right-censoring and 1 otherwise), A_i is a binary time-invariant exposure initiated at time $T = 0$ (1 for exposed subjects and 0 otherwise), and $L_i = \{L_{1i}, ..., L_{pi}\}$ is the set of the p measured time-invariant confounders. Let $S_a(t)$ be the survival function of group $A = a$ at time t, and let $\lambda_a(t)$ be the corresponding instantaneous hazard function. Suppose D_a is the number of different observed times of event in group $A = a$. At time t_j $(j = 1, ..., D_a)$, the number of events d_{ja} and the number of at-risk subjects Y_{ja} in group $A = a$ can be defined as $d_{ja} = \sum_{i:t_i=t_j} \delta_i \mathbb{1}(A_i = a)$ and $Y_{ja} = \sum_{i:t_i \geq t_j} \mathbb{1}(A_i = a)$.

2.2 Estimands

The hazard ratio (HR) has become the estimand of choice in confounder-adjusted studies with time-to-event outcomes. However, it has also been contested,[25,26] mainly because the time-varying distribution of the baseline characteristics among the corresponding at-risk populations leads to selection biases. To better understand this pitfall that differs from the concept of individual time-varying covariate(s), consider the data-generating process illustrating in Figure 1 (panel A). If the analyst controls for confounding by adjusting or stratifying on L_2 and L_5, there is no confounding at baseline. Additionally, suppose that the individual values of the quantitative covariate L_4 are constant over the time, and that both the random variables L_4 and A are independently associated with a higher risk of death. In this situation, a difference between the average L_4 values among the survivors appears over time (Figure 1, panel B). Then, even when the conditional HR between an exposed and an unexposed with the same characteristics L is constant over time, the marginal (population) HR varies over time. This is also referred to the non-collapsibility of the HR.[25] Instead of HR, one can estimate the average over time of the different time-specific HRs: $AHR = \int [\lambda_1(t)/\lambda_0(t)] f(t) \mathrm{d}t$.[27]

Nevertheless, Aalen *et al.*[25] concluded that it is difficult to draw causal conclusions from such a relative estimand. Hernán[26] advocated the use of the adjusted survival curves and related differences. For instance, the restricted mean survival time (RMST) allows us to summarise a survival curve for a specific time-window and to compare two curves by looking at the difference in RMST.[28] The RMST difference up to time τ is formally defined as :

$$\Delta(\tau) = \int_0^\tau [S_1(t) - S_0(t)]\mathrm{d}t \tag{1}$$

This value corresponds to the difference in terms of mean event-free time between two groups of exposed and unexposed individuals followed up to time τ. A further advantage of the RMST difference is its usefulness for public health decision making.[29] Note that other alternatives that might avoid this problem exist, such as the attributable fraction or the number needed to treat.[30]

Hereafter, we considered AHR and $\Delta(\tau)$.

2.3 Weighting on the inverse of propensity score

Formally, the PS is defined by $g(L_i) = P(A_i = 1|L_i)$, *i.e.*, the probability that subject i is exposed according to her/his characteristics L_i. In practice, analysts often use a logistic regression such that $g(L_i) = \exp(\alpha_0 + \alpha L_i)/(1 + \exp(\alpha_0 + \alpha L_i))$, where α_0 and α are the intercept and the regression coefficient associated with the exposure, respectively. The individual PSs are then the predictions from this model. Let ω_i be the stabilised weight of subject i. Xu *et al.*[31] defined $\omega_i = A_i P(A_i = 1)/g(L_i) + (1 - A_i)P(A_i = 0)/(1 - g(L_i))$ to obtain a pseudo-population in which the distribution of covariates is balanced between exposure groups, enabling estimation of the causal effect in the entire population.[2] The use of stabilised weights has been shown to produce a suitable estimate of the variance even when there are subjects with extremely large weights.[4,31] The weighted numbers of events and at-risk subjects at time t_j in group $A = a$ are $d^\omega_{ja} = \sum_{i:t_i=t_j} \omega_i \delta_i \mathbb{1}(A_i = a)$ and $Y^\omega_{ja} = \sum_{i:t_i \geq t_j} \omega_i \mathbb{1}(A_i = a)$, respectively. Cole and Hernán[32] proposed a weighted Kaplan-Meier estimator defined as:

$$\hat{S}_a(t) = \prod_{t_j \leq t} \left[1 - d^\omega_{ja}/Y^\omega_{ja}\right] \tag{2}$$

To estimate the corresponding AHR, they suggested the use of a weighted univariate Cox PH model, in which exposure is the single explanatory variable. We use equation (1) to estimate the corresponding $\Delta(\tau)$.

2.4 G-computation

Akin to the IPW, the GC involves two steps. The first step consists of estimating the working model $Q(A, L)$.[13] When suitable, it can consist of a proportional hazard (PH) regression: $h_0(t)\exp(\gamma A_i + \beta L_i)$ where $h_0(t)$ is the baseline hazard function at time t, and γ and β are the regression coefficients. Estimates of the cumulative baseline hazard $\hat{H}_0(t)$ and the regression coefficients $(\hat{\gamma}, \hat{\beta})$ can be obtained by the joint likelihood approach proposed by Breslow.[33] The second step consists of predicting the counterfactual mean survival function if all subjects would have been exposed $(do(A = 1))$ or unexposed $(do(A = 0))$:

$$\hat{S}_a(t) = n^{-1} \sum_{i=1}^{n} \exp\left[- \hat{H}_0(t) \times \exp(\hat{\gamma} \times do(A_i = a) + \hat{\beta} L_i)\right] \tag{3}$$

Then, $\widehat{AHR}$ can be computed as the mean of the individual counterfactual hazard ratios at the observed event times:[27]

$$\widehat{AHR} = \left[\sum_{i=1}^{n} \delta_i\right]^{-1} \sum_{i=1}^{n} \delta_i \left[\hat{\lambda}_1(t_i)/\hat{\lambda}_0(t_i)\right], \tag{4}$$

where $\hat{\lambda}_a(t) = -\partial \log \hat{S}_a(t)/\partial t$, which is obtained from equation (3) by numerical differentiation. We use equation (1) to estimate the corresponding $\Delta(\tau)$.

2.5 Doubly robust standardisation

DREs combine $Q(A,L)$ and $g(L)$ to obtain a consistent estimate when at least one of these working models is well-specified.[20,21] In most cases, $g(L)$ is used to update $Q(A,L)$, such as in Targeted Maximum Likelihood Estimator (TMLE).[34] However, it can seem more intuitive to first use the IPW approach to reduce the imbalance between exposure groups and to then apply GC to control for residual confounding.[35] Therefore, we proposed the following doubly robust standardisation (DRS).[22] First, $g(L)$ is fitted to obtain the individual stabilised weights ω_i, as defined in the subsection 2.3 to balance the exposure groups on L. Second, a weighted Q-model $Q(A,L)$ is fitted using the aforementioned weights ω_i in the procedure described in the subsection 2.4 to achieve the double robustness property.

2.6 Identifiability conditions

As for standard regression models, the IPW and the GC require assumptions of non-informative censoring, no measurement error, no model misspecification, and no interference.[36] Three additional assumptions, called *identifiability conditions*, are necessary for causal inference. (i) The values of exposure under comparisons correspond to well-defined interventions that, in turn, correspond to the versions of exposure in the data. (ii) The conditional probability of receiving every value of exposure depends only on the measured covariates. (iii) The conditional probability of receiving every value of exposure is greater than zero. These assumptions are known as *consistency*, *conditional exchangeability* and *positivity*, respectively.

3 Simulation study

3.1 Data generation

We generated data in three steps following the data-generating process illustrated in Figure 1 (panel A). (i) We simulated three covariates (L_1 to L_3) from a Bernoulli distribution with parameter equal to 0.5 and three covariates (L_4 to L_6) from a standard normal distribution. (ii) We generated the exposure A according to a Bernoulli distribution with probability obtained by the logistic model with the following linear predictor: $-0.5 + \log(2) \cdot L_2 + \log(1.5)L_3 + \log(1.5)L_5 + \log(2)L_6$. We set the intercept to obtain the prevalence of exposed individuals at 50%. (iii) We generated the times-to-event from a Weibull PH model. We set the scale and shape parameters to 40.0 and 2.0, respectively. Based on a random variable U_i drawn from a standard uniform distribution, we then computed the time-to-event from a Weibull PH model as $40.0 \times [(1 - \log(1-U_i)\exp(-\gamma A_i - \log(1.3)L_1 - \log(1.8)L_2 - \log(1.8)L_4 - \log(1.3)L_5)) - 1]^{-2.0}$, where $\gamma = \log(1.0)$ under the null hypothesis or $\log(1.3)$ under the alternative hypothesis. We subsequently censored the times-to-event using a uniform distribution on [0,70] or [0,15], leading to approximately 40% and 90% censored observations, respectively. For each scenario, we randomly generated 10,000 datasets.

3.2 Performance criteria

To compute the difference in $\Delta(\tau)$, we defined τ in each dataset as the time at which at least 10% of the individuals in each group (exposed or unexposed) were still at risk. We computed the theoretical values of the AHR and $\Delta(\tau)$ by averaging the estimations obtained, respectively, from univariate Cox PH models (A as the only explanatory covariate) and by equation (1) where the survival functions were estimated by the Kaplan-Meier estimator, fitted from datasets simulated as above, except A was simulated independently of L.[23] We reported the following criteria: (i) the percentage of datasets without convergence ; (ii) the bias either as $\mathrm{E}(\hat{\theta}) - \theta$ or $100 \times \mathrm{E}(\hat{\theta}/\theta - 1)$, where θ is the estimand of interest; (iii) the mean square error $MSE = \mathrm{E}[(\hat{\theta} - \theta)^2]$; (iv) the variance estimation bias $VEB = 100 \times (SD(\hat{\theta})/\mathrm{E}[\widehat{SD}(\hat{\theta})] - 1)$, where $\mathrm{E}(\widehat{SD}(\bullet))$ is the asymptotic standard deviation and $SD(\bullet)$ is the empirical standard deviation; (v) the empirical coverage rate of the nominal 95% confidence interval (CI), defined as the percentage of 95% CIs including θ; (vi) the type I error, defined as the percentage of times the null hypothesis is rejected when the null hypothesis is true; and (vii) the statistical power, defined as the percentage of times the null hypothesis is rejected when the alternative hypothesis is true. We obtained the variances by bootstrap (1000 iterations), as recently recommended by Austin.[37] We computed the Monte Carlo standard errors for each performance measure.[38]

3.3 Scenarios

In addition to the two censoring rates and the two effect sizes, we explored three sample sizes: $n = 100$, 500, and 2000. When the censoring rate was 90%, we did not investigate the smallest sample size due to the reduced number of events.

In the main simulations, we considered two sets of covariates: $L = \{L_1, L_2, L_4, L_5\}$ the risk factors of the outcome, or $L = \{L_2, L_5\}$ the true confounders. Therefore, we fitted $g(L)$ as a logistic model and $Q(A, L)$ as a Cox PH model.

In a second set of simulations, we aimed to investigate the impact of an omitted confounder on the bias of each method. In addition to the correct set of covariates $L = \{L_1, L_2, L_4, L_5\}$, we defined the incorrect sets of covariates as either $L_{inc} = \{L_1, L_4, L_5\}$ or $\{L_1, L_2, L_4\}$ by respectively omitting L_2 or L_5, two confounders weakly or strongly associated with both exposure and the outcome. We investigated three scenarios in which the methods used are: (i) $g(L)$ and $Q(A, L_{inc})$, (ii) $g(L_{inc})$ and $Q(A, L)$, and (iii) $g(L_{inc})$ and $Q(A, L_{inc})$.

3.4 Software

We performed all the analyses using `R` version 4.0.3.[39] Source code to reproduce the results is available as Supporting Information on the journal's web page. To facilitate their use in practice, we have implemented the previous methods in the `R` package entitled `RISCA` (versions $\geq$ 0.8.1), which is available at `cran.r-project.org`.

3.5 Results

The Monte Carlo errors were weak, and we did not encounter any convergence problems. Figure 2 presents the results under the alternative hypotheses for Δ. The results for (i) Δ under the null hypothesis, and (ii) AHR under the null and alternative hypotheses were comparable and can be found in the supplementary material available online.

The bias associated with IPW, GC and DRS were similar and close to zero in all scenarios in which we considered all the risk factors or only the true confounders. Nevertheless for GC, the bias of Δ under the alternative hypothesis was lower considering all the risk factors rather than only the true confounders: 0.007 versus 0.111 for $n = 2000$, respectively, but only for a censoring rate of 40%. In small sample sizes, the bias was higher for IPW than GC and DRS. For instance, when $n = 100$ with a censoring rate of 40% and the risk factors, the bias was 0.100 for GC versus 0.053 and 0.065 for GC and DRS, respectively.

The GC, when considering all outcome causes, produced the best results in terms of MSE, especially for small sample sizes. For instance, when $n = 100$ with a censoring rate of 40%, the MSE related to the AHR was 0.054 for GC versus 0.066 and 0.056 for IPW and DRS, respectively. When considering only true confounders, these values were 0.074, 0.077 and 0.059, respectively.

Regarding the VEB, the results were slightly better with the GC than IPW, except when $n = 100$. DRS underestimated the variance of Δ under the alternative hypothesis, especially with a censoring rate of 40%. However, DRS led to a more accurate variance of Δ with a censoring rate of 90%. For the AHR, DRS led to similar VEBs than GC and IPW.

All scenarios broadly respected the nominal coverage value of 95% and the type I error of 5%. The power was the highest for the GC and DRS, especially when considering all the risk factors, regardless of the scenario.

As expected, omitting a confounder in $g(L)$ or $Q(A, L)$ led to an important bias for IPW and GC, respectively (Figure 3). In contrast, DRS remained unbiased when the set of confounders is complete in at least $g(L)$ or $Q(A, L)$. Interestingly, the omission in both $g(L)$ and $Q(A, L)$ did not lead to a higher bias for DRS than GC and IPW. The magnitude of the bias due to the omission of a confounder was similar across methods, ranging from 31.4% to 35.7%, for the different estimands and the different strengths of association. Similarly, the sample size and the censoring rate did not significantly change the amplitude of bias (data not shown).

4 Applications

We used data from two studies performed for multiple sclerosis and for kidney transplantation.[40,41] We conducted these studies following the French law relative to clinical noninterventional research. Written informed consent was obtained. Moreover, the French commission for data protection approved the collection (CNIL decisions DR-2014-327 and 914184). To guide variable selection, we asked experts which covariates were causes of the exposure or the outcome prognosis to

define the causal structure.[42] We checked the positivity assumption and the considered covariates balance (see supplementary materials available online). The log-linearity hypothesis of continuous covariates was confirmed in the univariate analysis if the Bayesian information criterion was not reduced using natural spline transformation compared to the inclusion of the covariate in its natural scale. In case of violation, we used a natural spline transformation. We also assessed the PH assumption via the Grambsch-Therneau test at a significance level of 5%. For simplicity, we performed complete case analyses.

4.1 Dimethylfumarate versus Teriflunomide to prevent relapse in multiple sclerosis

With the increasing number of available drugs for preventing relapses in multiple sclerosis and the lack of head-to-head randomised clinical trials, Laplaud *et al.*[40] aimed to compare Teriflunomide (TRF) and Dimethylfumarate (DMF) using data from the multicentric cohort OFSEP. We reanalysed the primary outcome, defined as the time-to-first relapse. We presented the cohort characteristics of 1770 included patients in Table 1: 1057 patients were in the DMF group (59.7%) versus 713 in the TRF group (40.3%). Approximately 39% of patients (40% in the DMF group versus 38% in the TRF group) had at least one relapse during follow-up.

We presented the confounders-adjusted results in the left panel of Figure 4. The different set of covariates did not significantly change the results. For the difference in RMST, the width of the 95% CI was larger for the IPW. For instance, when we considered all the risk factors, the CI of IPW had a width of 36.6 days versus 33.6 days and 33.2 days for GC and DRS, respectively.

The conclusion of no significant difference between TRF and DMF was unaffected by the method for the AHR. In contrast, the IPW led to a protective effect of the TRF compared to DMF in terms of Δ at two years, while the GC and the DRS remained consistent with the AHR conclusions. By taking into account the risk factors, the use of IPW concluded to a gain of 21 days without relapse at two years versus 6.5 days and 7.4 days for GC and DRS, respectively. Owing to the similar performances of the three methods under the null hypothesis and because unmeasured confounding cannot be an issue here, we suppose that the difference can be explained by a misspecification of $g(L)$.

4.2 Basiliximab versus Thymoglobulin to prevent post-transplant complications

Amongst non-immunised kidney transplant recipients, one can expect similar rejection risk between Thymoglobulin (ATG) and Basiliximab (BSX), two possible immunosuppressive drugs proposed as induction therapy. However, ATG may be associated with higher serious events, especially in the elderly. We aimed to reassess the difference in cardiovascular complications in ATG versus BSX patients.[41] Table 2 describes the 383 included patients from the multicentric DIVAT cohort: 204 patients were in the BSX group (53.3%) versus 179 in the ATG group (46.7%). Approximately 30% of patients (29% in the BSX group and 31% in the ATG group) had a least one cardiovascular complication during follow-up. The median follow-up time was 1.8 years (min: 0.0; max: 8.2).

In the right panel of Figure 4, we presented the confounders-adjusted RMST differences for a cohort followed up to three years. The results obtained were similarly sensitive to the considered set of covariates. Indeed, the upper (respectively lower) bound of the 95% CIs of the AHR (respectively Δ) was closer to zero when only the risk factors were considered. Nevertheless, the conclusion remained identical whatever the method used and the estimand targeted: we were unable to conclude to a difference in cardiovascular complications between the ATG and BSX patients.

5 Discussion

We aimed to explain and compare the performances of the GC, IPW and DRS for estimating causal effects in time-to-event analyses with time-invariant confounders. We focused on the average HR and the RMST difference. The results of the simulations showed that the three methods performed similarly in terms of bias, VEB, coverage rate, and type I error rate. Nevertheless, both the GC and the DRS outperformed the IPW in terms of statistical power, even when the censoring rate was high. Furthermore, the simulations showed that $Q(A, L)$ should preferentially include all the risk factors to ensure a smaller bias due to the self-induced selection. The main advantage of using the GC is the gain in statistical power. DRS is also an interesting method due to its double robustness property at the cost of a small loss of efficiency.

We have overcome the self-induced selection in HR *i.e.*, the non-collapsibility due to the presence of a cause of the time-to-event independent of the exposure, by averaging the time-dependant HR. The AHR is furthermore a valid causal estimand

because it creates a contrast between a function of the potential outcomes under the two exposures $A = 1$ or $A = 0$.[43] Surprisingly, we reported a higher bias of the RMST difference by GC when we only considered the true confounders. The difference in survival, such as RMST, is only collapsible in a fully continuous-time context, *i.e.*, with infinitely short time intervals.[17] For the GC, we computed the survival probability for each observed event time leading to small intervals. Therefore, we could observe slight, but still present, bias due to non-collapsibility when the risk factors were not considered. We did not observe such bias with the higher censoring rate due to the scarcity of events.[25]

While the first application to multiple sclerosis highlighted differences between the duo GC/DRS and the IPW, the second one in kidney transplantation illustrated the importance of the set of covariates to consider. With all risk factors, we concluded with more confidence that there was not a significant difference between Basiliximab and Thymoglobulin. In contrast, when using only the true confounders, the bound of the 95% CIs were close to zero. This again highlights the fact that even if a risk factor is balanced between the exposure groups at baseline, it could become unbalanced over time (as illustrated in Figure 1, panel B).

Nevertheless, the higher power of the GC is counterbalanced by three points. First, IPW allows us to easily check positivity near-violations by plotting the individual weights.[44] Second, although the lack of positivity directly affects the estimation of $g(L)$, the GC is also impacted by the resulting lack of support to properly estimate $Q(A, L)$ which can be qualified as an extrapolation issue. An extension of our work is to explore the robustness of the previous methods in the presence of such near-violations. Third, the need for bootstrapping to estimate its variance, analytic estimators that are available for the IPW.[45,46] In practice, we must emphasise that bootstrapping the entire estimation procedure has the advantage of valid post-selection inference.[47] Furthermore, data-driven methods for variables selection, such as the super learner, have recently been developed and may represent a promising perspective when full clinical knowledge is unavailable.[48,49] With such a data-driven covariates selection, the use of GC also represents a partial solution to prevent the selection of instrumental variables since it is independent of the exposure modelling, but DRS can avoid potential residual confounding due to the omission of a confounder weakly associated with the outcome in $Q(A, L)$.[19] While DREs have been criticised because they can amplify the bias when the two working models are misspecified,[20,50] DRS and TMLEs do not cause such bias amplification.[22,51] Some TMLEs have been proposed to estimate either the RMST difference or survival functions.[52,53] Unfortunately, they require the discretisation of the times exacerbating the non-collapsibility of the estimands.[17]

The methods studied here are not the only available methods to estimate the causal effect. For instance, Conner *et al.*[45] compared the performances of IPW with that of other regression-based methods. Overall, the statistical performances were similar. However, the advantage of the studied methods compared to other ones is the visualisation of the confounder-adjusted results in terms of the survival curve or an indicator such as RMST.

Our study has several limitations. First, the results of the simulations and applications are not a theoretical argument for generalising to all situations. Second, we studied only logistic and Cox PH regression: other working models could be applied. Keil and Edwards[54] proposed a review of possible models for $Q(A, L)$ with a time-to-event outcome. Third, we considered only a reduced number of covariates, which could explain the abovementioned equivalence between the GC and the IPW with the extreme censoring rate. Last, we did not consider competing events or time-varying confounders that require specific estimation methods.[55,56]

To conclude, by means simulation and two applications on real datasets, this study tended to show the lower power of the IPW compared to GC and DRS to estimate the causal effect with time-to-event outcomes. All the risk factors should be considered in GC to overcome the self-induce selection bias. Our work is a continuation of the emerging literature that questions the near-exclusivity of PS-based methods in causal inference.

Acknowledgements

The authors would like to thank the members of DIVAT and OFSEP Groups for their involvement in the study, the physicians who helped recruit patients, the clinical research associates who participated in the data collection and all patients who participated in this study. We also thank David Laplaud and Magali Giral for their clinical expertise as well as Gabriel Danelian for language corrections. The analysis and interpretation of these data are the responsibility of the authors. This work was partially supported by a public grant overseen by the French National Research Agency (ANR) to create the Common Laboratory RISCA (Research in Informatic and Statistic for Cohort Analyses, `www.labcom-risca.com`, reference: ANR-16-LCV1-0003-01) involving the development of Plug-Stat software.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Arthur Chatton obtained a grant from IDBC for this work. Other authors received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

References

1. Sherman RE, Anderson SA, Dal Pan GJ et al. Real-world evidence – what is it and what can it tell us? *New England Journal of Medicine* 2016; 375(23): 2293–2297. DOI:10.1056/NEJMsb1609216.
2. Hernán MA. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health* 2004; 58(4): 265–271. DOI:10.1136/jech.2002.006361.
3. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70(1): 41–55. DOI:10.1093/biomet/70.1.41.
4. Robins JM, Hernán MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11(5): 550–560. DOI:10.1097/00001648-200009000-00011.
5. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 2004; 23(19): 2937–2960. DOI:10.1002/sim.1903.
6. Vansteelandt S and Daniel RM. On regression adjustment for the propensity score. *Statistics in Medicine* 2014; 33(23): 4053–4072. DOI:10.1002/sim.6207.
7. Ali MS, Groenwold RHH, Belitser SV et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology* 2015; 68(2): 112–121. DOI:10.1016/j.jclinepi.2014.08.011.
8. Le Borgne F, Giraudeau B, Querard AH et al. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. *Statistics in Medicine* 2016; 35(7): 1103–1116. DOI: 10.1002/sim.6777.
9. Hajage D, Tubach F, Steg PG et al. On the use of propensity scores in case of rare exposure. *BMC Medical Research Methodology* 2016; 16(1): 38. DOI:10.1186/s12874-016-0135-1.
10. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine* 2013; 32(16): 2837–2849. DOI:10.1002/sim.5705.
11. King G and Nielsen R. Why propensity scores should not be used for matching. *Political Analysis* 2019; : 1–20DOI:10.1017/pan.2019.11.
12. Robins J. 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): 1393–1512. DOI:10.1016/0270-0255(86)90088-6.
13. Snowden JM, Rose S and Mortimer KM. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology* 2011; 173(7): 731–738. DOI:10.1093/aje/kwq472.
14. Keil AP, Edwards JK, Richardson DR et al. The parametric g-formula for time-to-event data: towards intuition with a worked example. *Epidemiology* 2014; 25(6): 889–897. DOI:10.1097/EDE.0000000000000160.
15. Wen L, Young JG, Robins JM et al. Parametric g-formula implementations for causal survival analyses. *Biometrics* 2020; Published ahead of print. DOI:10.1111/biom.13321.
16. Breskin A, Edmonds A, Cole SR et al. G-computation for policy-relevant effects of interventions on time-to-event outcomes. *International Journal of Epidemiology* 2021; 49(6): 2021–2029. DOI:10.1093/ije/dyaa156.
17. Sjölander A, Dahlqwist E and Zetterqvist J. A Note on the Noncollapsibility of Rate Differences and Rate Ratios. *Epidemiology* 2016; 27(3): 356–359. DOI:10.1097/EDE.0000000000000433.
18. Gollob HF and Reichardt CS. Taking Account of Time Lags in Causal Models. *Child Development* 1987; 58(1): 80–92. DOI: 10.2307/1130293.

19. Kreif N and DiazOrdaz K. Machine Learning in Policy Evaluation: New Tools for Causal Inference. *Oxford Research Encyclopedia of Economics and Finance* 2019. DOI:10.1093/acrefore/9780190625979.013.256.
20. Tan, Z. Comment: Understanding OR, PS and DR. *Statistical Science* 2007; 22(4): 560–568. DOI:10.1214/07-STS227A.
21. Lendle SD, Fireman B and van der Laan MJ. Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology* 2013; 66: S91–S98. DOI:10.1016/j.jclinepi.2013.02.017.
22. Vansteelandt S and Keiding N. Invited Commentary: G-Computation-Lost in Translation? *American Journal of Epidemiology* 2011; 173(7): 739–742. DOI:10.1093/aje/kwq474.
23. Chatton A, Le Borgne F, Leyrat C et al. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific Reports* 2020; 10(11): 9219. DOI: 10.1038/s41598-020-65917-x.
24. Hernán MA, Brumback B and Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* 2001; 96(454): 440–448. DOI:10.1198/016214501753168154.
25. Aalen OO, Cook RJ and Røysland K. Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis* 2015; 21(4): 579–593. DOI:10.1007/s10985-015-9335-y.
26. Hernán MA. The Hazards of Hazard Ratios. *Epidemiology* 2010; 21(1): 13–15. DOI:10.1097/EDE.0b013e3181c1ea43.
27. Schemper M, Wakounig S and Heinze G. The estimation of average hazard ratios by weighted cox regression. *Statistics in Medicine* 2009; 28(19): 2473–2489. DOI:10.1002/sim.3623.
28. Royston P and Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology* 2013; 13(1): 152. DOI:10.1186/1471-2288-13-152.
29. Poole C. On the origin of risk relativism. *Epidemiology* 2010; 21(1): 3–9. DOI:10.1097/EDE.0b013e3181c30eba.
30. Sjölander A. Estimation of causal effect measures with the R-package stdReg. *European Journal of Epidemiology* 2018; 33(9): 847–858. DOI:10.1007/s10654-018-0375-y.
31. Xu S, Ross C, Raebel MA et al. Use of Stabilized Inverse Propensity Scores as Weights to Directly Estimate Relative Risk and Its Confidence Intervals. *Value in Health* 2010; 13(2): 273–277. DOI:10.1111/j.1524-4733.2009.00671.x.
32. Cole SR and Hernán MA. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine* 2004; 75(1): 45–49. DOI:10.1016/j.cmpb.2003.10.004.
33. Breslow N. Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society Series B* 1972; 34(2): 216–217. DOI: 10.1111/j.2517-6161.1972.tb00900.x.
34. Luque-Fernandez MA, Schomaker M, Rachet B et al. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine* 2018; 37(16): 2530–2546. DOI:10.1002/sim.7628.
35. Hernán MA, Hernández-Díaz S and Robins JM. Randomized trials analyzed as observational studies. *Annals of Internal Medicine* 2013; 159(8):560-2. DOI:10.7326/0003-4819-159-8-201310150-00709.
36. Hudgens MG and Halloran ME. Toward causal inference with interference. *Journal of the American Statistical Association* 2008; 103(482): 832–842. DOI:10.1198/016214508000000292.
37. Austin PC. Variance estimation when using inverse probability of treatment weighting (iptw) with survival analysis. *Statistics in Medicine* 2016; 35(30): 5642–5655. DOI:10.1002/sim.7084.
38. Morris TP, White IR and Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019; 38(11): 2074–2102. DOI:10.1002/sim.8086.
39. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.
40. Laplaud DA, Casey R, Barbin L et al. Comparative effectiveness of teriflunomide vs dimethyl fumarate in multiple sclerosis. *Neurology* 2019; 93: 1–12. DOI:10.1212/WNL.0000000000007938.
41. Masset C, Boucquemont J, Garandeau C et al. Induction therapy in elderly kidney transplant recipients with low immunological risk. *Transplantation* 2019; : 1DOI:10.1097/TP.0000000000002804.
42. VanderWeele TJ and Shpitser I. A new criterion for confounder selection. *Biometrics* 2011; 67(4): 1406–1413. DOI:10.1111/j.1541-0420.2011.01619.x.
43. Martinussen T, Vansteelandt S and Andersen PK. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Analysis* 2020; 26(4): 833–855. DOI:10.1007/s10985-020-09501-5.

44. Petersen ML, Porter KE, Gruber S et al. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 2012; 21(1): 31–54. DOI:10.1177/0962280210386207.
45. Conner SC, Sullivan LM, Benjamin EJ et al. Adjusted restricted mean survival times in observational studies. *Statistics in Medicine* 2019; 1-29. DOI:10.1002/sim.8206.
46. Hajage D, Chauvet G, Belin L et al. Closed-form variance estimator for weighted propensity score estimators with survival outcome. *Biometrical Journal* 2018; 60(6): 1151–1163. DOI:10.1002/bimj.201700330.
47. Efron B. Estimation and accuracy after model selection. *Journal of the American Statistical Association* 2014; 109(507): 991–1007. DOI:10.1080/01621459.2013.823775.
48. Blakely T, Lynch J, Simons K et al. Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. *International Journal of Epidemiology* 2019; 1-7. DOI:10.1093/ije/dyz132.
49. Le Borgne F, Chatton A, Léger M et al. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. *Scientific Reports* 2021; 11: 1435. DOI:10.1038/s41598-021-81110-0.
50. Kang JDY and Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; 22(4): 523–539. DOI:10.1214/07-STS227.
51. Kreif N, Gruber S, Radice R et al. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. *Statistical Methods in Medical Research* 2016; 25(5): 2315–2336. DOI:10.1177/0962280214521341.
52. Díaz I, Colantuoni E, Hanley D et al. Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime Data Analysis* 2019; 25(3): 439–468. DOI:10.1007/s10985-018-9428-5.
53. Benkeser D, Carone M and Gilbert PB. Improved estimation of the cumulative incidence of rare outcomes. *Statistics in Medicine* 2018; 37(2): 280–293. DOI:10.1002/sim.7337.
54. Keil AP and Edwards JK. A review of time scale fundamentals in the g-formula and insidious selection bias. *Current epidemiology reports* 2018; 5(3): 205–213. DOI:10.1007/s40471-018-0153-0
55. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ et al. A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine* 2020; 39(8): 1199–1236. DOI:10.1002/sim.8471.
56. Daniel R, Cousens S, De Stavola B et al. Methods for dealing with time-dependent confounding. *Statistics in Medicine* 2013; 32(9): 1584–1618. DOI:10.1002/sim.5686.

Table 1. Description of the multiple sclerosis cohort according to the treatment group.

	Overall (n=1770)		TRF (n=713)		DMF (n=1057)		p-value
	n	*%*	*n*	*%*	*n*	*%*	
Male recipient	485	27.4	202	28.3	283	26.8	0.4713
Disease modifying therapy before initiation	1004	56.7	395	55.4	609	57.6	0.3560
Including Interferon			237		369		
Glatiramer Acetate			158		240		
Relapse within the year before initiation	981	55.4	346	48.5	635	60.1	<0.0001
Relapse within the two years before initiation	1227	69.3	444	62.3	783	74.1	<0.0001
Gado. Positive lesion on MRI at baseline	601	34.0	207	29.0	394	37.3	0.0003
Center with more than 50 included patients	1612	91.1	653	91.6	959	90.7	0.5354
At least one relapse at two-year post-initiation	527	29.8	200	28.1	327	30.9	0.1928
	mean	*sd*	*mean*	*sd*	*mean*	*sd*	
Patient age at multiple sclerosis onset (years)	31.7	9.7	32.9	9.8	30.9	9.5	<0.0001
Patient age at initiation (years)	39.3	10.7	41.3	10.8	38.0	10.5	<0.0001
Disease duration (years)	7.6	7.4	8.4	7.8	7.1	7.0	0.0003
EDSS level at initiation	1.7	1.3	1.7	1.3	1.7	1.2	0.9885
Number of relapses in the previous year	0.7	0.8	0.6	0.7	0.8	0.8	<0.0001
Number of relapses in the two previous years	1.0	1.0	0.9	0.9	1.1	1.0	<0.0001

No variable have missing data.
Abbreviations: DMF, Dimethylfumarate; EDSS, Expanded Disability Status Scale; Gado, Gadolinium; MRI, Magnetic resonance imaging; MS, Multiple sclerosis; sd, Standard deviation; and TRF, Teriflunomide.

Table 2. Description of the kidney's transplantation cohort according to the induction therapy.

	Overall (n=383)			ATG (n=179)			BSX (n=204)			p-value
	missing	*n*	*%*	*missing*	*n*	*%*	*missing*	*n*	*%*	
Male recipient	0	284	74.2	0	137	76.5	0	147	72.1	0.3180
Recurrent causal nephropathy	0	63	16.4	0	29	16.2	0	34	16.7	0.9024
Preemptive transplantation	1	61	16.0	1	18	10.1	0	43	21.1	0.0035
History of diabetes	0	123	32.1	0	64	35.8	0	59	28.9	0.1530
History of hypertension	0	327	85.4	0	150	83.8	0	177	86.8	0.4124
History of vascular disease	0	109	28.5	0	53	29.6	0	56	27.5	0.6405
History of cardiac disease	0	153	39.9	0	75	41.9	0	78	38.2	0.4651
History of cardiovascular disease	0	203	53.0	0	99	55.3	0	104	51.0	0.3973
History of malignancy	0	94	24.5	0	42	23.5	0	52	25.5	0.6457
History of dyslipidemia	0	220	57.4	0	92	51.4	0	128	62.7	0.0250
Positive recipient CMV serology	5	230	60.8	4	119	68.0	1	111	54.7	0.0082
Male donor	0	187	48.8	0	93	52.0	0	94	46.1	0.2510
ECD donor	1	372	97.4	1	172	96.6	0	200	98.0	0.5244
Use of machine perfusion	12	208	54.3	6	86	48.0	6	122	59.8	0.0684
Vascular cause of donor death	0	275	71.8	0	126	70.4	0	149	73.0	0.5655
Donor hypertension	11	224	60.2	9	103	60.6	2	121	59.9	0.8927
Positive donor CMV serology	0	240	62.7	0	115	64.2	0	125	61.3	0.5486
Positive donor EBV serology	1	370	96.9	1	172	96.6	0	198	97.1	0.8102
HLA-A-B-DR incompatibilities >4	5	97	25.7	3	41	23.3	2	56	27.7	0.3256
		mean	*sd*		*mean*	*sd*		*mean*	*sd*	
Recipient age (years)	0	70.8	4.8	0	70.5	4.8	0	71.0	4.8	0.3733
Recipient BMI (kg/m^2)	3	26.7	4.0	3	26.9	4.2	0	26.5	3.9	0.2796
Duration on waiting list (months)	16	16.5	19.0	11	17.9	18.9	5	15.4	19.1	0.2082
Donor age (years)	1	72.7	8.8	1	72.1	10.0	0	73.1	7.5	0.2739
Donor creatininemia (μmol/L)	1	82.9	39.5	0	85.5	41.0	1	80.7	38.0	0.2331
Cold ischemia time (hours)	3	15.6	5.0	1	15.9	5.2	2	15.3	4.8	0.2820

Abbreviations: ATG, Thymoglobulin; BMI, Body mass index; BSX, Basiliximab; CMV, Cytomegalovirus; EBV, Epstein-Barr virus; ECD, Expanded criteria donor; HLA, Human leucocyte antigen; and sd, Standard deviation.

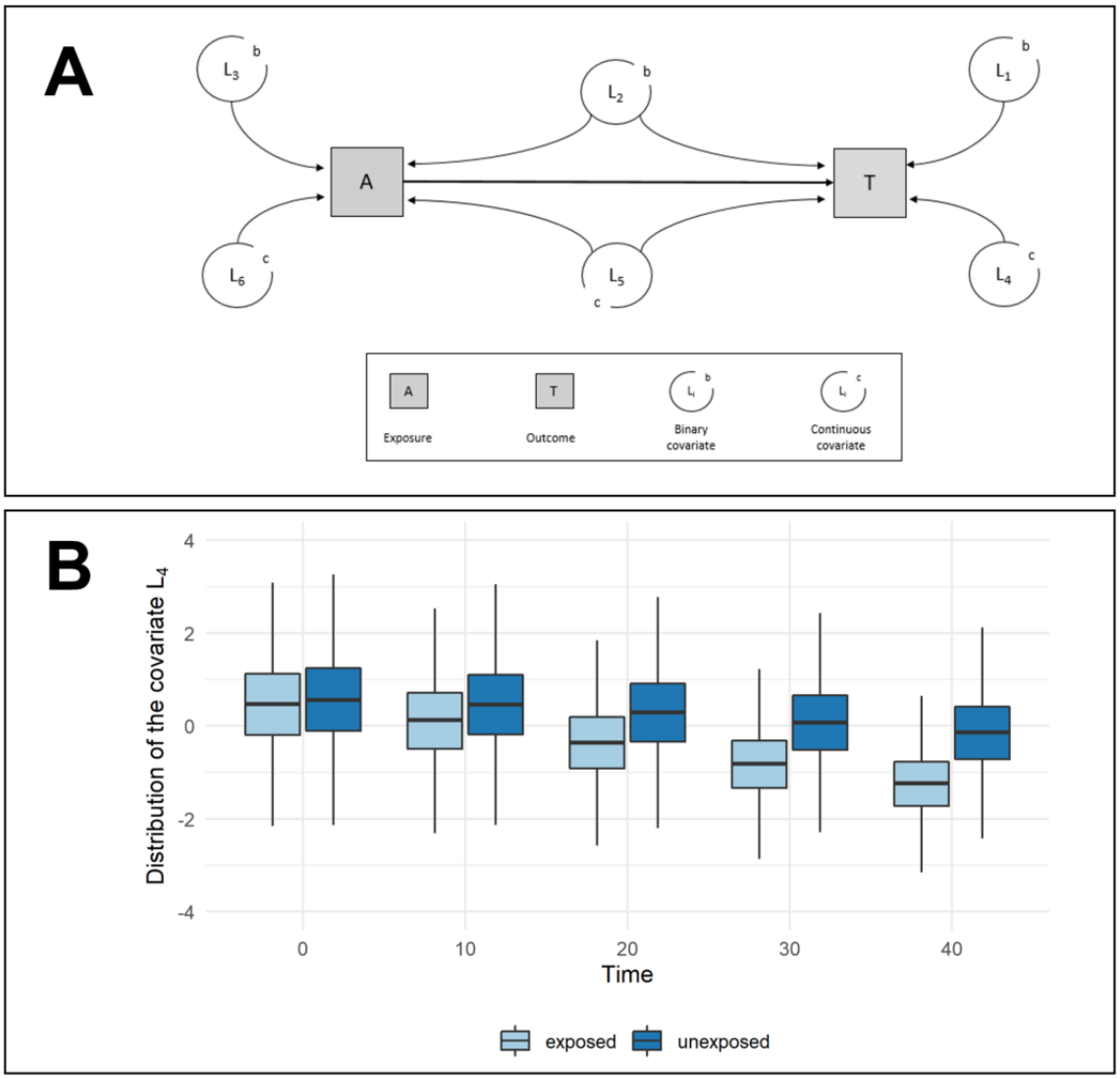


Figure 1. (**A**) Causal diagram illustrating the data-generating process. (**B**) Distribution of the baseline covariate L_4 over time according to exposure status in a simulated population of one million people. L_4 is moderately associated (HR = 1.8) with the outcome.

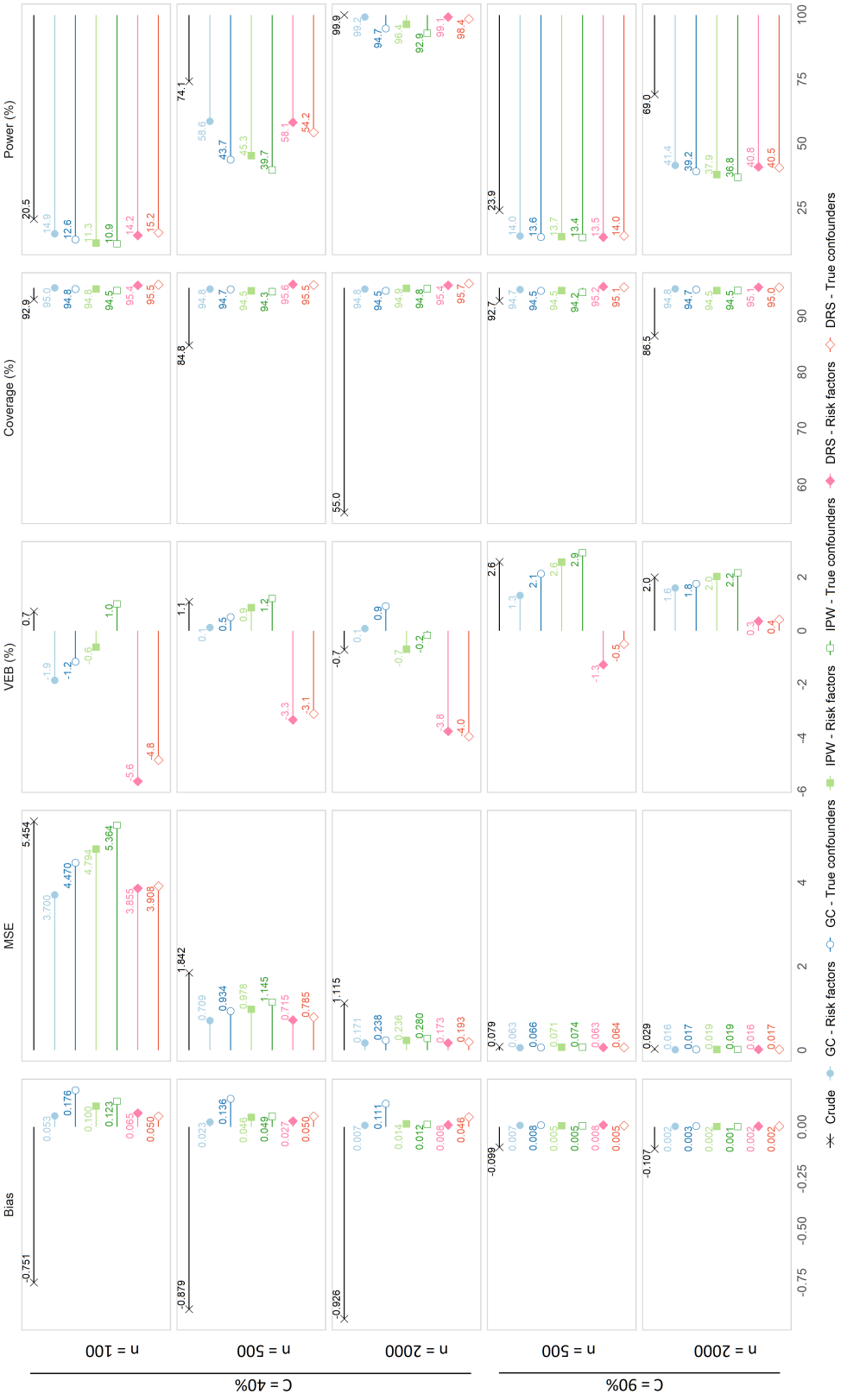


Figure 2. Performances of the g-computation (GC), inverse probability weighting (IPW) and Doubly Robust Standardisation (DRS) under the alternative hypothesis to estimate the restricted mean survival times difference at time τ. τ equals to 36.6 and 12.9 for censoring rates of 40% and 90%, respectively. Theoretical values of restricted mean survival times difference equal to -1.877 and -0.218 for censoring rates of 40% and 90%, respectively. Abbreviations: C, censoring rate; n, sample size.

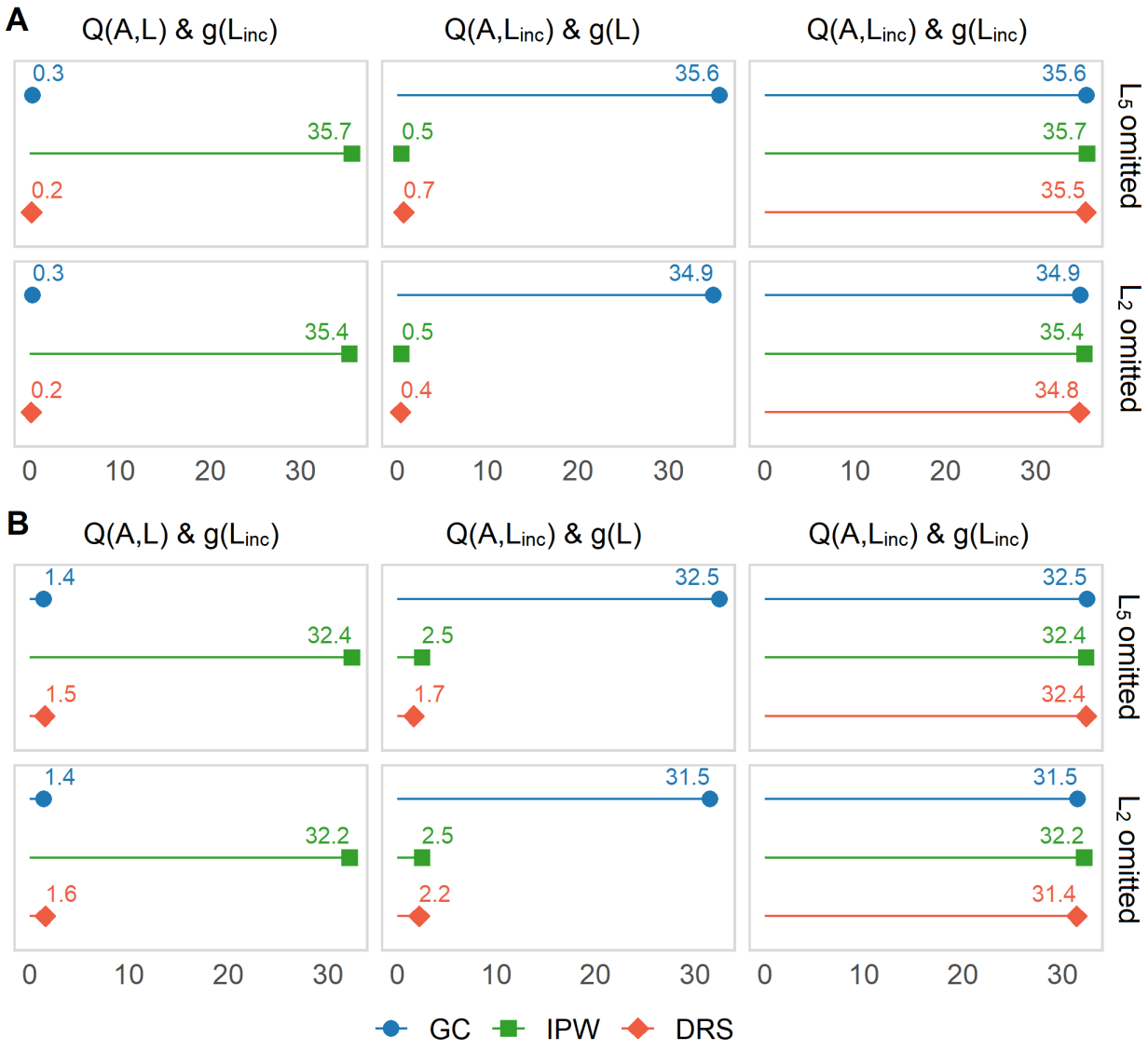


Figure 3. Relative bias (%) of the g-computation (GC), inverse probability weighting (IPW) and Doubly Robust Standardisation (DRS) with an ommitted confounder to estimate: **A** - the log average hazard ratio; **B** - the restricted mean survival times difference.

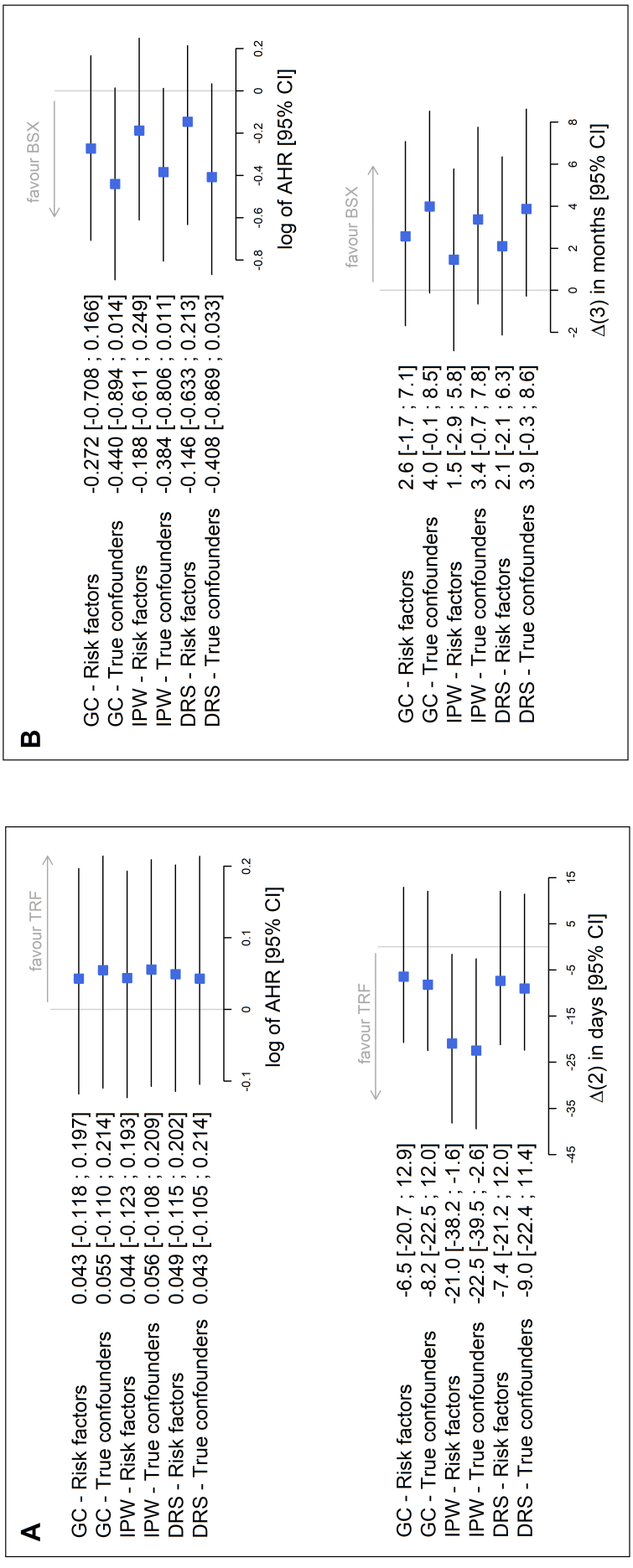


Figure 4. Comparison of: **A** - Dimethylfumarate and Teriflunomide (TRF) for the time-to-first relapse of multiple sclerosis; **B** - Basiliximab (BSX) and Thymoglobulin for the occurrence of a cardiovascular complication after a kidney's transplantation.

Chapitre 6

Identification automatisée de potentielles violations de l'hypothèse de positivité

« *En apprenant, tu enseigneras.*
En enseignant, tu apprendras. »

Phil Collins, *Son of Man*

Les *Supplementary Materials* peuvent être trouvés dans l'Annexe G de ce manuscrit.

Identifying positivity violations using decision trees: introducing the P-CART algorithm

Gabriel Danelian,[1,2] Yohann Foucher,[3,4] Maxime Léger,[3,5] Florent Le Borgne,[2,3] and Arthur Chatton[2,3*]

[1] Université de Lille, Lille, France

[2] IDBC/A2COM, Pacé, France.

[3] UMR INSERM 1246 - SPHERE, Université de Nantes, Université de Tours, Nantes, France.

[4] Centre Hospitalier Universitaire de Nantes, Nantes, France.

[5] Département d'anesthésie-réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France.

* Corresponding author: Arthur Chatton, UMR INSERM 1246 - SPHERE, Institut de Recherche en Santé 2, 22 Boulevard Benoni-Goullin, 44200 NANTES, France. Mail: arthur.chatton@univ-nantes.fr

Word count: 2592

Abstract

Background: The positivity assumption is crucial in observational studies where design does not guarantee ambivalence, but is often overlooked in practice. A structural (non-random) infraction of positivity is due to a subgroup of individuals being unable to receive one of the levels of exposure. To correctly estimate the causal effect, we need to identify and exclude this subgroup of individuals. We suggest a CART-based algorithm to help achieve this.

Development: Based on a succession of decision trees, the algorithm searches for combinations of covariate values that result in subgroups of individuals with a very low probability of being exposed or unexposed.

Application: To illustrate the method's usefulness, we applied it to four datasets from recently published studies. The algorithm found the two previously reported subgroups with a structural violation of the positivity, and it identified nine additional subgroups suspected of presenting a violation. According to the medical experts, three were considered false positives (probably due

to sample-to-sample fluctuation, i.e., random violation) and six were structural violations of positivity.

C*onclusions:* The suggested algorithm allows one to easily and rapidly detect potential violations of the positivity assumption. We implemented the algorithm in the R package RISCA to facilitate its use and dissemination. Easy to use and independent of the inference method, we hope that it will improve the verification of the positivity assumption in causal studies.

Keywords: causal inference; decision tree; eligibility criteria; identifiability; positivity; target population.

Key Messages

- Causal inference requires the positivity assumption: each individual must have a non-zero probability of having each level of exposure.
- Statistical approaches were commonly employed to deal with positivity violations, leading to a shift in inference population and thus in estimand.
- We suggest a CART-based method for identifying subgroups of individuals potentially responsible for random or structural violations.
- We illustrate its usefulness with the reanalysis of four recently published studies, finding that the inclusion criteria could have been better defined in three of them.
- PCART is available in the R package RISCA.

Introduction

The positivity assumption, also known as experimental treatment assignment or common support, is a cornerstone to drawing causal inferences. The positivity assumption means that each individual has a non-zero probability of being exposed and unexposed.[1] Its violation may take two forms: a structural form or a random one. A structural violation is deterministic due to the inclusion of individuals who can theoretically never receive one of the levels of the studied exposition, resulting in an incorrect definition of the eligibility criteria. Contrariwise, a random violation occurs by chance, *i*.e., without any underlying cause.[2] Such a random violation is more likely in small samples or with rare exposure.

Alongside consistency and exchangeability, positivity is essential to the identifiability of the causal effect, *i*.e., the mapping of the statistical parameter obtained from the data and the causal quantity of interest (hereafter the estimand).[3] The target population is an integral part of the estimand.[4,5] Epidemiologists are commonly interested in the population average causal effect and approximate it by the sample average causal effect.[6] Random violations of positivity can be addressed by statistical procedures, such as trimming or truncating the propensity scores if these have been used.[7] Alternatively, decision trees have been used to estimate a causal effect in a subpopulation with respect to positivity.[8,9] However, all of these approaches change the characteristics of the sample, shift the population of inference, and target a different sample average causal effect.[5,10] The difference between the population defined by the eligibility criteria and the inference population casts doubt upon the external validity of the results.[11] Therefore, Platt *et al.*[12] argued that the exclusion of the subjects unlikely to be (un)exposed, by redefining the eligibility criteria, is transparent and leads to clear statements about the generalisability of the results. This approach is closely related to a well-defined target trial.[13,14]

Rather than using decision trees to directly estimate the causal effect in an ill-defined population free of positivity violations, we suggest using classification and regression trees (CART) to identify the subgroups of individuals potentially causing such violations. Since a decision tree consists of

a set of nodes (*i.e.*, binary decision rules) themselves subdivided into other nodes (and so on, see Supplementary Figure 1), using the exposure as the outcome and one or several covariates as predictors allows one to obtain an estimation of the treatment probability in each subgroup represented by the nodes.[15,16] In the presence of a node with an extreme exposure probability, we can therefore suspect a violation of the positivity assumption.

We sought to develop a CART-based algorithm to check for potential violations of the positivity assumption and help redefine the eligibility criteria. To facilitate the use of this method, we implemented it as a function in the R package *RISCA*.[17]

The paper proceeds as follows. We present the algorithm in the next section, then apply it to several datasets from recently published studies, and present the results in the third section. The last section offers discussions and practical recommendations.

The P(ositivity)-CART algorithm

The algorithm definition.

Consider (α, β, γ) the three parameters allowing one to define the positivity violation. The parameter α is the minimal percentage of the whole sample size needed to define a problematic subgroup. A positivity violation is defined by an exposure probability lower than β% or greater than 100-β% in a subgroup.[18] The parameter γ refers to the maximal number of predictors used to define a subgroup.

Based on these parameters, we suggest using CART to identify such subgroups. Figure 1 presents an overview of the proposed algorithm, which proceeds in several steps. The first step involves estimating one tree for each predictor and memorising the leaves (*i.e.*, the final nodes) corresponding to problematic subgroups according to the parameters α and β. If γ=1, the algorithm stops. Otherwise, if at least one problematic subgroup is identified in the first step, the

corresponding predictor(s) is(are) not considered in the next step (to limit the number of identified subgroups), which estimates one tree for all possible couples of predictors and memorises the leaves corresponding to problematic subgroups according to the parameters α and β. If γ=2, the algorithm stops. Otherwise, the third step consists of one tree for all possible trio of remaining covariates not involved in the previously identified subgroups, and so on.

The default values of (α, β, γ).

To the best of our knowledge, there is no consensus on the precise definition a subgroup presenting a positivity violation. Nevertheless, one can set α=5%, as several authors suggest trimming the weights at this level in inverse probability weighing analysis.[19–21] Following D'Amour *et al.*,[18] we set β=5%. Note that this value is also consistent with the propensity score literature.[7,22] By default, the parameter γ is arbitrarily set to two since one can question the relevance of the results for higher levels.

Of note, we do not use the pruning step of CART in order to create the vastest possible number of divisions, and therefore of subgroups. Previous research also showed that using such large trees allows one to overcome a time-consuming optimisation of the CART-specific parameters.[23] Therefore, we used the default values provided in the R package *rpart* (version 4.1-15).[24]

Applications

Context.

We applied the P-CART algorithm by reanalysing datasets from four recently published observational studies.[25–28] We aimed to validate the algorithm's capacity to re-identify the positivity violations previously reported based on expert knowledge and identify potential new

ones that were previously missed. We used the adjustment set of covariates as predictors in the PCART algorithm (Supplementary Table 1) since positivity is relevant only for them.[1,2] The authors restricted the studied populations in order to respect the positivity assumption. Therefore, we did not apply these specific eligibility criteria when evaluating whether the PCART algorithm allowed us to identify the positivity violations. We conducted complete case analyses. We categorized each continuous variable according to meaningful cut-offs before running the algorithm to avoid clinically insignificant subgroups. We performed the main analysis with the default values (α=5%, β=5%, and γ=2). We considered other values (α=1%, α=10%, β=1% and β=10%) as sensitivity analyses (Supplementary Tables 2-6). In a second time, we reanalysed the data by (i) excluding the clinically plausible subgroups identified by P-CART and (ii) applying the inverse probability weighting approach performed initially by the authors (see Austin and Stuart[29] for an introduction). We performed all the analyses using R version 3.6.0.[30]

Barbiturates during intensive care for patients with traumatic brain injury.

Léger *et al.*[25] investigated the impact of barbiturates on mortality by using a cohort made up of 1088 patients admitted into intensive care units for traumatic brain injury. This treatment may be offered to reduce patients' intracranial hypertension and the corresponding consequences in terms of brain damage. The authors excluded individuals older than 70 because the therapy is contraindicated for the elderly. We applied the P-CART algorithm to the entire cohort regardless of patient age. We identified nine subgroups potentially associated with positivity violations (Table 1). First, the P-CART algorithm confirmed the patients older than 75 as problematic, a greater threshold value compared to the one suggested by the authors. Second, it indicated that patients without osmotherapy at admission had a probability of receiving barbiturates lesser than 5%. This is clinically coherent since barbiturates are a last-line therapy and should only be offered after osmotherapy. The authors had not identified this issue, presumably because these patients could have received another second-line therapy. Third, P-CART identified four

subgroups composed of patients without intracranial hypertension at admission. This characteristic was not identified in the first iteration of P-CART because the treatment probability for patients without intracranial hypertension was equal to 5.1%, just above the threshold β. Note that these individuals can receive barbiturates in a preventive way if they are identified at a high risk of the event. Fourth, the algorithm detected that a lactatemia lower than 1mmol/L, an SAPS II score from 40 to 44 without severe trauma, and an SAPS II score from 25 to 55 along with a creatinine level in between 50 and 60mmol/L were associated with a probability of barbiturates under 5%. These three subgroups do not seem to have a clinical explanation and can be qualified as random violations due to sample-to-sample fluctuation.

Figure 2 shows the results of the initial analysis performed by the authors and the reanalysis based on the restricted sample. The exclusion of individuals without intracranial hypertension at admission alone reduced the sample by almost 70%. In the final sample (N=173), the odds ratio is 1.9 (rather than 2.2 in the whole sample) and became closer to statistical insignificance with a confidence interval at 95% (CI95%) from 1.0 to 3.5.

Kidney transplantations from marginal donors.

Querard et al.[26] compared grafts from standard and marginal donors, as defined by the expanded donor criteria. The two kinds of grafts differ in their intrinsic quality. Because of the shortage of kidneys for transplantation, the grafts from standard donors are preferentially attributed to young recipients, thus implying a positivity issue. The authors did not consider the recipient age among the eligibility criteria. In contrast, P-CART detected a problematic subgroup consisting of 391 recipients younger than 30 years (8.1% of the whole sample) with a probability of receiving a marginal graft of 3.1%. The exclusion of these individuals did not change the magnitude of the effect.

The hypothermic perfusion machine for marginal donors.

Foucher et al.[27] compared the hypothermic perfusion machine to static cold storage in kidney transplantations from expanded donors. The authors reduced the studied cohort to individuals younger than 45 because of a potential structural violation: the old-to-old graft allocation policy results in a lower susceptibility of younger candidates receiving marginal grafts. By using the entire cohort with no restriction on patient age (N=1978), the P-CART algorithm did not detect this issue. Indeed, only 32 and 44 individuals were less than 45 years old in the perfusion machine and cold storage groups, respectively. In contrast, P-CART suggested that individuals transplanted before 2015 in four anonymised centres could be problematic. The centres A, B, C and D included 126, 81, 218 and 41 patients during this period, respectively. These centres respectively attributed a perfusion machine with a probability of 0.8%, 3.7%, 4.7% or 2.4%. This violation seems plausible given that some hospitals were slower to adopt hypothermic perfusion machines, which were only introduced in 2010. Similarly, patients with a cold ischemia time under 20 hours and transplanted before 2013 (N=101) had a probability of receiving a graft under static cold storage lower than 5%. This second violation also seems plausible because the perfusion machines were preferentially attributed to grafts transplanted away from the harvesting site (*i.e.*, with a high cold ischemia time). This preferential attribution became less strict with the wider availability of perfusion machines over time. The exclusion of these two subgroups shifted the HR from 0.9 (CI95% 0.7 – 1.1) in the whole sample to 0.8 (CI95% 0.6 – 1.1) in the restricted sample. Furthermore, restricting the sample halved the 10-years survival probability difference.

Induction therapy in elderly kidney transplant recipients.

Masset et al.[28] compared the risk of adverse events following a kidney transplant in elderly recipients depending on their induction therapy: anti-thymocyte globulins versus basiliximab. We did not detect any positivity violations with P-CART, confirming the authors' statement.

Discussion

By reanalysing the data of four published studies, the P-CART algorithm confirmed the two subgroups with a structural positivity violation, as stated by the authors. It also identified nine new subgroups which may be problematic. Of the nine subgroups, six were likely structural positivity violations, which were missing from the authors' results. It illustrated that the PCART algorithm would have put them on the track of potential violations of positivity, although expert knowledge will always be necessary to distinguish random and structural violations of positivity.

As illustrated by the applications, the restriction of the sample can lead to a different estimation. Although the clinical conclusions were unaffected, the effects' magnitude was smaller in the restricted samples. Interestingly, the estimate of Querard *et al.*[26] was unaffected by the restriction. These authors targeted the average causal effect on the exposed, for which a weaker positivity assumption is necessary. Indeed, only the exposed individuals should have a non-null probability of being unexposed.[31] Therefore, their inference was not affected by this subgroup composed of unexposed individuals.

The P-CART algorithm we suggest is based on three tuning parameters that define the subgroup affected by the positivity violation: α the minimum percentage of patients included in the subgroup, β the maximum probability of being exposed or unexposed in the subgroup, and γ the maximum number of predictors outlining the subgroups. In our applications, we set the parameter α at 5%, in agreement with the literature. However, it could be increased when the sample size decreases as random violations are more likely to occur. We set the parameter β at 5%, again in line with the literature. Alternative values can be considered. For instance, Petersen *et al.*[7] and Crump *et al.*[32] argued in favour of using a β = 1% and 10%, respectively. Note that the exposure prevalence can also influence this choice, a small prevalence leading to a tighter definition of the positivity, thus a smaller β. The application on the dataset of Léger *et al.*[25] outlined the advantage of increasing the parameter γ to identify structural violations when

β is close to exposure prevalence. The parameter γ may also be increased with the sample size to refine the definition of the subgroups. Nevertheless, it seems unnecessary to define structural violations with a combination of more than three variables intended to represent the exposure's allocation process.

Other approaches can be employed for diagnosing positivity violations. When a propensity score is used, Cole and Hernán warned against the presence of extreme weights.[33] Petersen *et al.* provided a bootstrap-based tool to quantify the amount of bias due to such violations.[7] Westreich and Cole suggested computing all possible contingency tables to check for the presence of empty cells.[2] Unlike P-CART, the two first approaches cannot identify the subgroups causing a positivity violation, while the last approach is only feasible in lower-dimensional settings and cannot deal with continuous variables. We stress that positivity is not solely related to propensity score-based approaches, even if a vast amount of the positivity literature is found in this context. Outcome regression-based approaches also require this assumption, albeit they can extrapolate to the problematic subgroups in the case of random violations.[7,25] Randomized clinical trials may also present a lack of positivity when the sample size is not large enough to prevent sample-to-sample fluctuations.

Some limitations need to be mentioned. First, we only considered binary treatments, and further development is needed to consider multimodal or even continuous exposure. Second, we did not perform any simulation study to illustrate the predictive performances of the algorithm. We believe that the P-CART algorithm is a practical tool to help investigators and analysts to precise the target population. Last, we did not investigate the optimal set of tuning parameters, another perspective to be addressed in future work.

In conclusion, we illustrated that P-CART represents a helpful decision-making tool for checking for the targeted population in causal studies. It enables essential discussions about the susceptible structural violations before causal analysis (e.g., propensity scores, G-computation, doubly robust estimators or multiple regressions). We recommend using the P-CART algorithm with different

tuning parameters to get an overview of the potential violations. P-CART can lead to better transparency of the research process, a better generalisability of the results, and better implementation of the findings in real life for the ultimate benefit of patients.

Author Contributions

G.D. developed the algorithm, analysed the data, and drafted the manuscript. A.C. designed the study, analysed the data, and wrote the manuscript. All authors interpreted the data, revised the manuscript, as well as read and approved the final version of the manuscript. A.C. is the guarantor of the article.

Acknowledgements

The authors would like to thank the members of AtlanREA and DIVAT groups for their involvement in the study, the physicians who helped recruit patients and all patients who participated in this study. We also thank the clinical research associates who participated in the data collection. The analysis and interpretation of these data are the responsibility of the authors.

Funding

This work was partially supported by a public grant overseen by the French National Research Agency (ANR) to create the Common Laboratory RISCA (Research in Informatic and Statistic for Cohort Analyses, www.labcom-risca.com, reference: ANR-16-LCV1-0003-01). The funder had no role in study design; analysis, and interpretation of data; writing the report; and the decision to submit the report for publication.

Conflict of interest: None declared.

References

1. Hernán M, Robins JM. Causal Inference: What if? Boca Raton: Chapman & Hall/CRC; 2020.

2. Westreich D, Cole SR. Invited Commentary: Positivity in Practice. *Am J Epidemiol*. 2010 Mar 15;**171**(6):674–677.

3. Petersen ML, Laan MJ van der. Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology*. 2014 May;**25**(3):418–426.

4. Maldonado G, Greenland S. Estimating causal effects. *International Journal of Epidemiology*. 2002 Apr 1;**31**(2):422–429.

5. Lundberg I, Johnson R, Stewart BM. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *Am Sociol Rev*. 2021 Jun 1;**86**(3):532–565.

6. Balzer LB. 'All Generalizations Are Dangerous, Even This One.'-Alexandre Dumas. *Epidemiology*. 2017;**28**(4):562–566.

7. Petersen ML, Porter KE, Gruber S, Wang Y, Laan MJ van der. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012 Feb;**21**(1):31–54.

8. Hill J, Su Y-S. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann Appl Stat*. 2013 Sep;**7**(3):1386–1420.

9. Kang J, Chan W, Kim M-O, Steiner PM. Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores. *Commun Stat Appl Methods*. 2016 Jan;**23**(1):1–20.

10. Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches. *Pharmacoepidemiol Drug Saf*. 2021 Nov;**30**(11):1471–1485.

11. Nethery RC, Mealli F, Dominici F. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *Ann Appl Stat*. 2019 Jun;**13**(2):1242–1267.

12. Platt RW, Delaney JAC, Suissa S. The positivity assumption and marginal structural models: the example of warfarin use and risk of bleeding. *Eur J Epidemiol*. 2012 Feb;**27**(2):77–83.

13. Didelez V. Commentary: Should the analysis of observational data always be preceded by specifying a target experimental trial? *International Journal of Epidemiology*. 2016 Dec 1;**45**(6):2049–2051.

14. García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol*. 2017 Jun 1;**32**(6):495–500.

15. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*. 2010 Aug 1;**63**(8):826–833.

16. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol*. 2019 Dec;**188**(12):2222–2239.

17. Foucher Y, Le Borgne, Dantan E, Gillaizeau F, Chatton A, Combescure C. RISCA: Causal Inference and Prediction in Cohort-Based Analyses. 2019. https://CRAN.R-project.org/package=RISCA (20 October 2021, date last accessed).

18. D'Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*. 2021 Apr 1;**221**(2):644–654.

19. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution—A Simulation Study. *Am J Epidemiol*. 2010 Oct 1;**172**(7):843–854.

20. Glynn RJ, Lunt M, Rothman KJ, Poole C, Schneeweiss S, Stürmer T. Comparison of alternative approaches to trim subjects in the tails of the propensity score distribution. *Pharmacoepidemiol Drug Saf*. 2019 Oct;**28**(10):1290–1298.

21. Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. *PLOS ONE*. 2011 Mar 31;**6**(3):e18174.

22. Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. *Stat Methods Med Res*. 2020 Dec 1;**29**(12):3721–3756.

23. Mantovani RG, Horváth T, Cerri R, Junior SB, Vanschoren J, Leon Ferreira de Carvalho ACP de. An empirical study on hyperparameter tuning of decision trees. *arXiv:1812.02207*. 2019 Feb 12;**preprint: not peer reviewed**.

24. Therneau TM, Atkinson B. rpart: Recursive Partitioning and Regression Trees. 2019. https://CRAN.R-project.org/package=rpart (20 October 2021, date last accessed).

25. Léger M, Chatton A, Le Borgne F, Pirracchio R, Sigismond L, Foucher Y. Causal inference in case of near-violation of positivity: comparison of methods. *Biom J*. 2021:In press.

26. Querard A-H, Foucher Y, Combescure C, et al. Comparison of survival outcomes between Expanded Criteria Donor and Standard Criteria Donor kidney transplant recipients: a systematic review and meta-analysis. *Transpl Int*. 2016 Apr 1;**29**(4):403–415.

27. Foucher Y, Fournier M-C, Legendre C, et al. Comparison of machine perfusion versus cold storage in kidney transplant recipients from expanded criteria donors: a cohort-based study. *Nephrol Dial Transplant*. 2020 Jun 1;**35**(6):1043–1070.

28. Masset C, Boucquemont J, Garandeau C, et al. Induction Therapy in Elderly Kidney Transplant Recipients With Low Immunological Risk. *Transplantation*. 2020 Mar;**104**(3):613–622.

29. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015 Dec 10;**34**(28):3661–3679.

30. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/

31. Pirracchio R, Carone M, Rigon MR, Caruana E, Mebazaa A, Chevret S. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Stat Methods Med Res*. 2016 Oct;**25**(5):1968–1954.

32. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009 Mar 1;**96**(1):187–199.

33. Cole SR, Hernan MA. Constructing Inverse Probability Weights for Marginal Structural Models. *Am J Epidemiol*. 2008 Jul 15;**168**(6):656–664.

Table 1: Subgroups of patients identified by the PCART algorithm as potential sources of non-positivity.

Authors	*Sample size*	Problematic subgroup (*n*, %)	Identified by the authors	Clinically plausible	Violation type
Léger *et al.*[25]	1088	Age ≥ 75 years (73, 6.7)	Yes	Yes	Structural
		No osmotherapy at admission (732, 67.3)	No	Yes	Structural
		25 ≤ SAPS II score < 55 & 50 ≤ Creatinine < 60 (135, 12.4)	No	No	Random
		No IH at admission nor history of head trauma (710, 65.3)	No	Yes	Structural
		No IH & severe trauma at admission (385, 35.4)	No	Yes	Structural
		No IH at admission & Creatinine < 150 (740, 68.0)	No	Yes	Structural
		No IH at admission & SAPS II score < 55 (532, 48.9)	No	Yes	Structural
		Lactatemia < 1 (197, 18.1)	No	No	Random
		40 ≤ SAPS II score < 45 & No severe trauma (69, 6.3)	No	No	Random
Querard *et al.*[26]	3422	Recipient age < 30 years (391, 8.1)	Yes	Yes	Structural
Foucher *et al.*[27]	1978	Transplant before 2014 & A - D centres[a] (376, 19.0)	No	Yes	Structural
		Transplant before 2012 & CIT ≥ 20h (101, 5.1)	No	Yes	Structural
Masset *et al.*[28]	383	None detected[b]	-	-	-

Abbreviations: BMI, Body Mass Index; CIT, Cold Ischemia Time and IH, Intracranial Hypertension. Creatinine and lactatemia were in mmol/L.

[a] The centres were anonymised. [b] Regardless of the sample and the adjustment set.

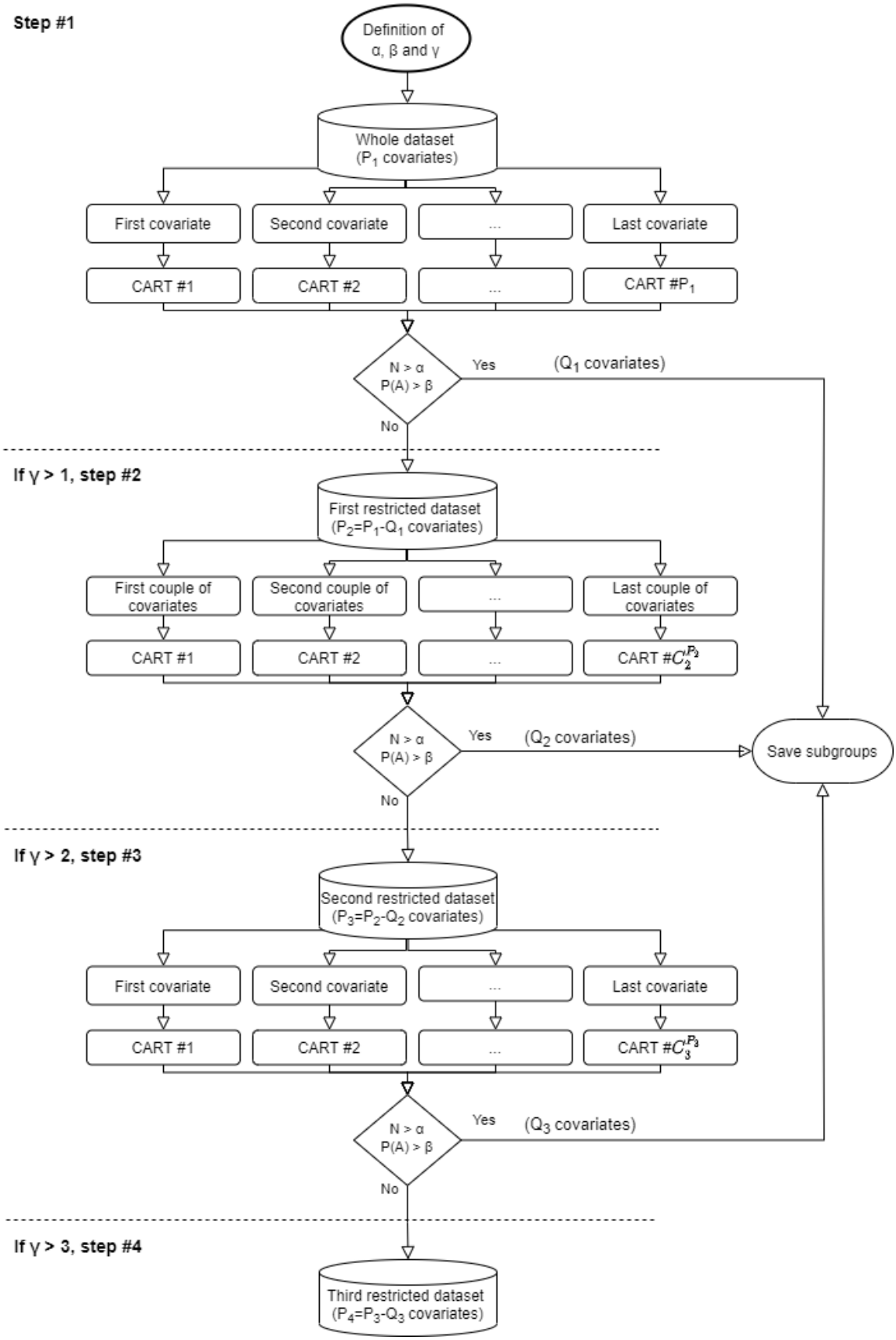


Figure 1: Flow-chart presenting the PCART algorithm. A: Treatment; N: Percentage of the whole sample contained in the subgroup; α, β and γ were user-supplied hyperparameters.

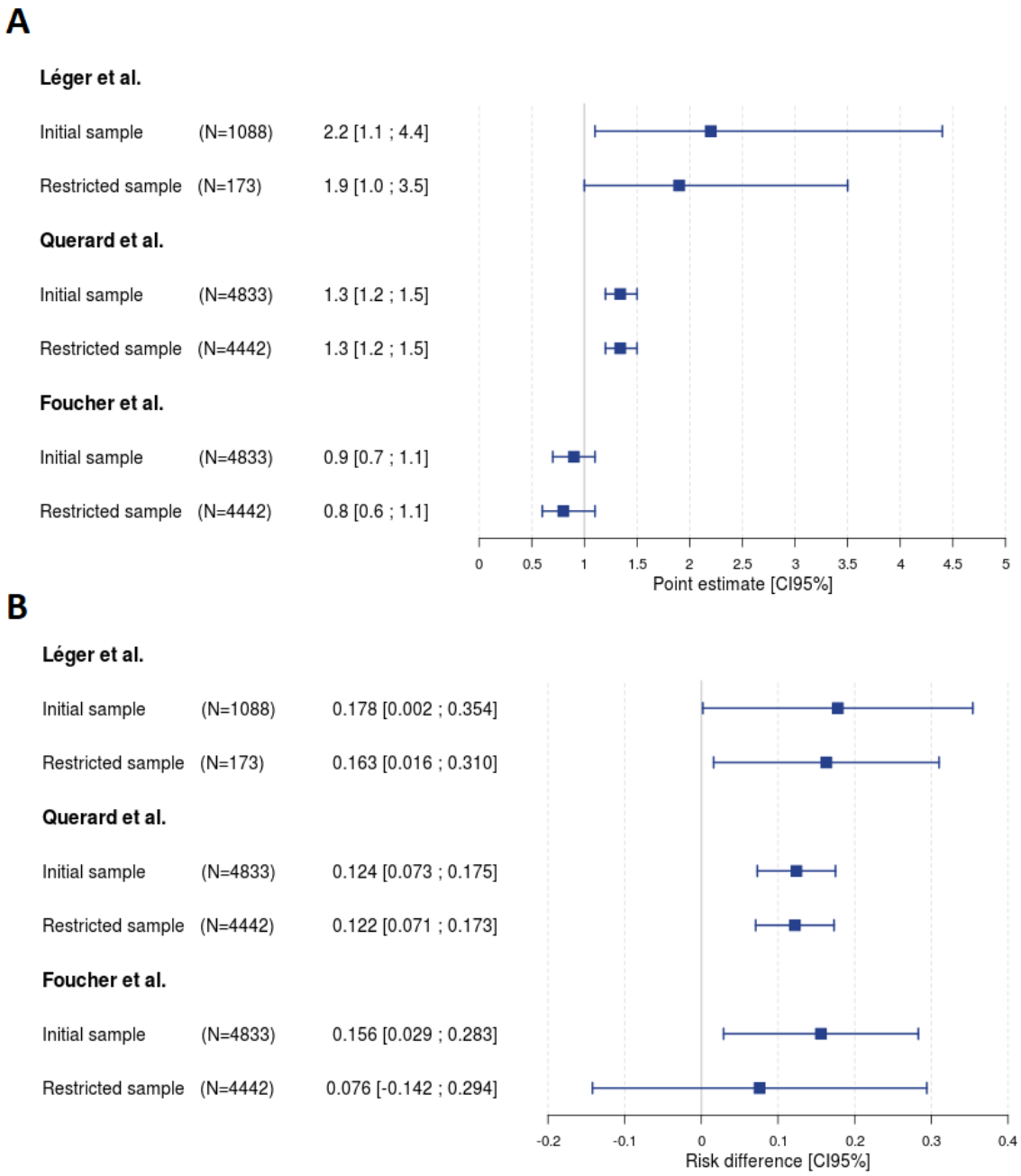


Figure 2: Results of the reanalyses considering the restricted datasets on the relative (A) and absolute scales (B). Léger et al. presents an odds-ratio and a risk difference, while the others present a hazard ratio and a 10years survival difference.

Chapitre 7

Discussion générale

« Ce qui est étrange dans l'acquisition du savoir, c'est que plus j'avance, plus je me rends compte que je ne savais même pas que ce que je ne savais pas existait. »

D. Keyes, *Des fleurs pour Algernon*

Sommaire

7.1 Résumé des travaux réalisés

Avec un évènement binaire, une première étude de simulation a été réalisée pour comparer les performances des principales méthodes d'estimation causale avec différents ensembles d'ajustement pouvant contenir des facteurs de risque et/ou des instruments. Cette étude a montré qu'inclure des facteurs de risque permet de diminuer la variance, que la GC est plus robuste à l'inclusion d'instruments que les autres approches étudiées (TMLE incluse) et qu'elle possède la variance la plus faible quel que soit l'estimand théorique considéré.

Toujours avec un évènement binaire, une seconde étude de simulation a investigué le potentiel du ML pour obtenir une spécification correcte du modèle de travail Q(A, L). Le SL surpassait les autres techniques étudiées en terme de biais et de variance. Associé à la GC, le SL conduisait à une estimation correcte, même dans un échantillon composé d'une centaine d'individus. Notons que l'utilisation du *Support Vector Machine* aboutissait à une erreur quadratique moyenne plus faible que celle du SL avec une centaine d'individu. Dans ce scénario particulier, ces deux approches sont plus précises qu'une régression logistique parfaitement spécifiée. En effet, la fluctuation d'échantillonage inhérente à une faible taille d'échantillon fait différer la structure observée de celle théorique.

Un troisième travail se plaçait dans la cadre de l'étude de temps d'évènement en présence de censure à droite pour développer une approche de GC et son extension doublement robuste (*i.e.*, la SDR). Ces approches ont été comparées à l'IPW via une étude de simulation. Les trois approches présentaient un biais équivalent mais l'IPW conduisait à une variance plus élevée. L'inclusion de facteurs de risque en plus des facteurs de confusion dans la modélisation a permis de diminuer le biais dû au phénomène de sélection auto-induite. Enfin, la SDR n'était pas sensible au phénomène d'amplification du biais lorsque les deux modèles de travail étaient mal-spécifiés.

Le dernier travail consistait à développer un outil permettant de vérifier le respect de l'hypothèse de positivité et d'identifier les individus problématiques le cas échéant. Au moyen d'une étude empirique sur des travaux préalablement publiés par notre équipe, l'algorithme basé sur les arbres de décision identifiait les deux strates d'individus problématiques décrites dans les papiers ainsi que neuf supplémentaires, dont sept sont vraisemblablement structurelles après expertise clinique.

7.2 Avantages et inconvénients des approches considérées

Les méthodes basées sur le SP modélisent l'allocation du traitement plutôt que la survenue de l'évènement. Elles semblent donc appropriées en présence d'évènements rares. La GC reposant sur les hypothèses de modélisation inverses, son utilisation est plus naturelle en cas d'allocation déséquilibrée du traitement. Les EDR combinent les deux modèles de travail et sont aisément utilisables dans ces deux contextes.

La principale différence est liée à l'hypothèse de positivité. Les approches estimant $g(L)$ peuvent vérifier d'éventuelles violations de cette hypothèse contrairement à la GC [69, 202]. Une violation de cette hypothèse est particulièrement problématique pour l'IPW à cause de l'apparition de poids extrêmes induisant une inflation de la variance et pouvant biaiser l'estimation [91, 114]. Au contraire, la GC est capable d'extrapoler la prédiction des évènements potentiels dans les strates problématiques lorsque la violation est aléatoire [69]. Un telle extrapolation conduit à des résultats corrects à partir du moment où Q(A, L) est correctement spécifié (cf annexe C). De façon analogue à l'IPW, la présence de poids extrêmes est possible pour les EDR en absence de positivité [69]. Néanmoins, les EDR sont capables d'extrapoler dans les strates problématiques comme la GC. Cette capacité d'extrapolation est perdue lorsque Q(A, L) est mal-spécifié [153, 201]. De plus, l'utilisation des fonctions d'influence pour la TMLE peut induire une sous-estimation de la variance

en cas de problème de positivité [203] ou de petite taille d'échantillon [48]. Dans ce cas, le bootstrap peut être préféré [203]. Enfin, de nombreuses études de simulation montraient une variance plus faible pour la GC en absence de positivité [68, 154, 159, 168, 179].

Outre les situations de non-positivité, la GC est généralement plus précise que les autres approches considérées [154, 155, 168, 180]. Bien que les EDR soient plus sensibles que la GC à l'inclusion de médiateurs dans l'ensemble d'ajustement L [204], ils possèdent la propriété de double robustesse maximisant les chances d'obtenir une estimation finale sans biais dû à une mauvaise spécification des modèles de travail. Notons que la GC, la TMLE et la SDR permettent d'estimer à la fois un effet marginal et conditionnel. L'effet conditionnel est estimable à partir de Q(A, L) directement, sans passer par l'étape de prédiction contrefactuelle. Un autre avantage des méthodes basées sur Q(A, L) est la possibilité de faire de la prédiction contrefactuelle dans une autre population afin d'évaluer la validité externe des résultats.

Bien que non-considérés dans ce manuscrit, les traitements continus (*e.g.*, doses) sont plus facilement modélisables avec les approches basées sur le SP [205]. Les approches de GC et les EDR nécessitent la catégorisation du traitement pouvant conduire à une perte d'information.

7.3 Réflections autour du ML

Le ML permet de modéliser de façon très fine les relations inter-variables (*e.g.*, interactions, formes fonctionelles, etc.). Mais cette flexibilité a un coût. Ces approches convergent plus lentement (*i.e.*, nécessitent une taille d'échantillon plus importante) que les estimateurs standards [206] et peuvent introduire deux autres sortes de biais dus au sur-apprentissage et à la régularisation (*i.e.*, présence d'une certaine contrainte inhibant les modèles trop complexes, l'exemple classique étant le paramètre de pénalisation des approches LASSO) [207]. La régularisation permet de limiter le sur-apprentissage en réduisant la variance au prix d'une augmentation du biais. Malheureusement, lorsque utilisés en GC, certains algorithmes de ML induisent une diminution du sur-apprentissage par régularisation trop lente pour compenser la diminution correspondante de la variance, ce qui peut conduire à une inférence incorrecte (*i.e.*, des intervalles de confiance très larges et pouvant ne pas contenir la vraie valeur) [208]. De plus, il n'a pas été prouvé que le théorème central limite puisse s'appliquer avec du ML (même avec du bootstrap [209]), ce qui complique l'estimation de la variance [157, 208, 210, 211]. Ainsi, de nombreux auteurs mettent en garde contre l'utilisation du ML avec des estimateurs n'étant pas doublement robustes [55, 157, 206, 208, 210, 212, 213].

Les EDR ont, quant à eux, été construits de façon à surmonter ces problèmes et permettre l'utilisation de ces approches de ML plus flexibles [40, 214]. Pour surmonter le biais dû à la régularisation, l'orthogonalisation est employée. L'idée est d'estimer les deux modèles de travail g(L) et Q(A, L) par ML puis de régresser linéairement les résidus de Q(A, L) sur les résidus de g(L) pour obtenir une estimation sans biais de régularisation [215]. De plus, les estimations des deux modèles de travail sont combinés de façon à ce que le taux de convergence de l'EDR corresponde au produit des taux de convergence de chaque modèle de travail permettant alors l'utilisation d'approches de ML ayant des taux de convergence moindres [157, 206]. En d'autres termes, puisque l'erreur globale correspond à la multiplication des résidus des deux modèles de travail, l'erreur résiduelle globale diminuera plus rapidement. La TMLE peut surmonter le sur-apprentissage via l'utilisation de SL incluant des approches plus simples [213]. En effet, les approches de ML incluses dans le SL conduisant au sur-apprentissage d'un des modèles de travail seront pondérées de façon à réduire leur contribution, évitant ainsi le sur-apprentissage. N'inclure que des approches hautement flexibles dans un SL conduira alors à un sur-apprentissage comme dans Naimi *et al.* [206]. La TMLE peut également être couplée à de la validation croisée afin de réduire davantage le sur-apprentissage [40]. De plus, l'utilisation des courbes d'influence permet d'obtenir un taux de convergence plus rapide rendant alors possible l'utilisation du ML pour la construction d'intervalles de confiance valides [157, 216].

Chernozhukov *et al.* [214] ont introduit un autre EDR permettant l'utilisation de ML : le double machine learning (DML). Comme la TMLE, la DML cherche d'une part à réduire le biais et d'autre part à obtenir un taux de convergence suffisant pour l'utilisation du ML et la construction d'intervalles de confiance valides. Les auteurs introduisaient la notion d'orthogonalisation de Neyman, où les modèles de travail sont utilisés pour définir l'effet causal. Ainsi, la dernière étape de l'orthogonalisation correspond à une solution analytique plutôt qu'à une simple régression des résidus. Par exemple, pour l'ATE :

$$\text{ATE} = n^{-1} \sum_{i=1}^{n} \left(\left[\text{Q}(\text{A}=1, \text{L}_i) - \text{Q}(\text{A}=0, \text{L}_i) \right] + \frac{\text{A}_i \left[\text{Y}_i - \text{Q}(\text{A}=1, \text{L}_i) \right]}{g(\text{L}_i)} - \frac{(1-\text{A}_i) \left[\text{Y}_i - \text{Q}(\text{A}=0, \text{L}_i) \right]}{1-g(\text{L}_i)} \right)$$

où la première partie de la somme (entre crochets) et les deux fractions correspondent respectivement à une estimation biaisée de l'estimand empirique et aux termes de "débiaisement". Notons que dans ce cas particulier, la DML correspond à un estimateur AIPW [157]. Chernozhukov *et al.* [214] utilisent une procédure d'apprentissage croisé (*cross-fitting*) pour éviter les problèmes de sur-apprentissage. Le jeu de données est aléatoirement séparé en deux, $g(\text{L})$ et $\text{Q}(\text{A}, \text{L})$ sont estimés sur la première partition et la dernière étape de l'orthogonalisation se fait dans la seconde partition. La même procédure se fait en échangeant les partitions, $g(\text{L})$ et $\text{Q}(\text{A}, \text{L})$ sont estimés sur la seconde partition, puis la régression entre les résidus se fera dans la première partition. Deux estimations de l'effet causal sont ainsi obtenues. L'estimation finale, sans biais de régularisation et de sur-apprentissage, se calcule comme la médiane des deux estimations [157, 214]. Zivich et Breskin [210] proposent une approche d'apprentissage doublement croisée où l'échantillon est séparé en trois partitions plutôt qu'en deux. La différence étant que les modèles $g(\text{L})$ et $\text{Q}(\text{A}, \text{L})$, estimés sur l'une des partitions, seront appliqués sur des partitions différentes de façon mutuellement exclusive afin que les prédictions issues des modèles de travail ne proviennent pas du même ensemble de données. L'apprentissage doublement croisé sacrifie donc du temps de calcul pour optimiser le taux de convergence. Néanmoins, un échantillon de grande taille est nécessaire pour utiliser cette approche. Comme pour la TMLE, l'estimation de la variance est basée sur les courbes d'influence [214]. Notons également que la DML semble particulièrement sensible à l'inclusion de médiateurs ou de colliders [217].

Nous avons pu voir dans le chapitre 6 que l'application du SL proposé en GC aboutissait à de bonnes performances statistiques, incluant un biais moyen proche de zéro, un taux de couverture proche de 95% et l'absence de biais lors de l'estimation de la variance (*i.e.*, les variances empirique et asymptotique étaient proches, cf Tableau B.1 Annexe B pour les formules). Ces résultats, à première vue inconsistants avec la littérature énoncée précédemment, peuvent être expliqués de la façon suivante. Le biais lié à la régularisation et la diminution du taux de convergence ont été limités par l'exclusion de méthodes trop flexibles telles que les arbres de décision [208]. En effet, l'utilisation d'approches de *boosting* (BCART) et de forêts aléatoires conduisaient effectivement à un biais important, des problèmes d'estimation de variance et à une couverture sous-optimale. Ces approches ont malheureusement été considérées dans les autres études de simulation étudiant les performances de la GC associée au ML [206, 210]. Le sur-apprentissage a été contrôlé au moyen d'une validation croisée par bootstrap [218]. Le fait de pouvoir tenir compte de l'ensemble de l'échantillon dans le processus d'estimation peut expliquer les bonnes performances de notre approche dans des échantillons relativement petits. Ces faibles tailles d'échantillon compliquent l'utilisation de la DML ou de la TMLE. En effet, la DML nécessite de partitionner sans remise l'échantillon conduisant ainsi à une estimation sur un très faible nombre d'individus. La TMLE est, quant à elle, particulièrement sensible aux violations aléatoires de la positivité qui surviennent couramment dans les échantillons de taille modérée [213].

7.4 Limitations et perspectives

Outre les limites inhérentes aux différents travaux présentés dans les chapitres 3 à 6, des limites plus globales méritent discussion.

La majorité des travaux présentés se basaient sur des études de simulations apportant un éclairage sur les seuls scénarios considérés [219]. Ces résultats ne sont pas forcément généralisables dans d'autres contextes, eu égard à la complexité du monde réel. Il peut également exister un biais, de non-neutralité, de la part des investigateurs envers la méthode proposée. En effet, la plupart des nouvelles méthodes statistiques ne sont testées que dans des situations qui leurs sont favorables [220, 221]. Le premier travail (chapitre 3) ne cherchait pas à démontrer la supériorité d'une méthode particulière sur les concurrentes mais à étudier leur robustesse dans diverses circonstances. Le plan de simulation a déjà été utilisé dans des études montrant la supériorité de l'IPW ou de l'appariement sur le SP sur d'autres compétiteurs [102, 176, 222, 223]. Or, dans nos simulations, ces méthodes se révèlent être les moins performantes. Le seul point questionnable est le choix d'une allocation déséquilibrée du traitement (un individu traité pour quatre non-traités) pour étudier l'ATT. Ce déséquilibre reflète au contraire un choix de favoriser la méthode la moins performante (appariement sur le SP [1]) en se plaçant dans un scénario où tous les traités pourront être appariés malgré le caliper. Les performances équivalentes, voire légèrement plus robustes, de la GC par rapport à la TMLE n'étaient pas attendues. Le second travail (chapitre 4) comparait cinq approches de ML associées avec la GC plus une sixième consistant en une combinaison des quatre premières au moyen du SL. Une meilleure performance du SL était attendue suite aux démonstrations théoriques [224]. Deux mécanismes de générations de données ont été considérés : un premier identique au précédent travail et un second plus réaliste, donc complexe (incluant notamment une interaction et des formes fonctionnelles complexes), afin de ne pas favoriser les méthodes paramétriques pénalisées (LASSO et elasticnet). Comme attendu, le SL conduisait aux meilleures performances excepté dans le scénario le plus simple où le *Support Vector Machine* était parfois supérieur. Les méthodes paramétriques conduisaient à des performances correctes lorsque la taille d'échantillon était supérieure ou égale à 500 individus. La dernière étude de simulation (chapitre 5) adaptait la génération des données du premier travail. Néanmoins, deux taux de censure (un réaliste et un extrême) étaient considérés afin de comparer la GC et l'IPW dans des scénarios respectivement équilibré et en faveur de l'IPW. Il était attendu que l'EDR étudié performe aussi bien que la plus performante des méthodes. De plus, un scénario omettant un facteur de confusion de l'ensemble d'ajustement a été investigué pour comparer la robustesse des trois approches à une violation de l'hypothèse d'échangeabilité conditionnelle. Ce scénario peut également être vu comme une tentative de comparaison des trois approches à la mauvaise spécification. Il est en effet difficile de comparer directement cette caractéristique puisque les modèles de travail sont différents (modèle de Cox versus modèle logistique, coefficients et nombre de variables différents). Avec du recul, il est possible que la GC ait été favorisée par rapport à l'IPW par un trop faible nombre de covariables utilisées lors de la modélisation en présence de forte censure à droite. En effet, des problèmes de convergence liés au faible nombre d'évènements aurait potentiellement été observé. Cependant, la SDR n'aurait vraisemblablement pas été affectée. De plus, toutes les méthodes n'ont pu être comparées dans ces études de simulation. Les recommandations récentes stipulent que les méthodes ayant les meilleures performances générales ainsi que celles étant couramment usitées devraient être étudiées [219, 225]. Ainsi, il aurait été intéressant d'étudier également le comportement de la DML et de la TMLE pour le troisième travail. Cependant, la SDR est plus simple à construire que la DML ou la TMLE dont une extension en temps continu de cette dernière vient seulement d'être construite sans être publiée [226]. Outre les compétiteurs à considérer, Boulesteix [225] recommande de reporter un ensemble de mesures de performances permettant d'apporter l'éclairage le plus large possible. Ceci inclut à la fois des

1. Notons que la première version du manuscrit présentait une comparaison de la GC, de l'IPW et de l'appariement sur le SP selon les plus proches voisins. Ces résultats n'ont pas été publiés suite aux commentaires des réviseurs de l'article et sont disponibles dans l'Annexe H.

estimations du biais, de la variance, du biais de la variance, de la couverture, de l'erreur de type I et de la puissance. Ces différentes mesures sont utiles seulement lorsqu'elles sont prises ensemble. En effet, une méthode peut être légèrement plus biaisée qu'une autre mais avoir une variance autrement plus faible, pouvant la rendre finalement plus performante. Enfin le taux de convergence et l'erreur de Monte-Carlo devraient être également rapportés dans les études de simulation [219, 227]. Comme pour un nombre insuffisant de mesures de performances, ne pas tenir compte des problèmes de convergence rencontrés peut conduire à une représentation erronée de la réalité.

L'intérêt des méthodes étudiées, et plus globalement de l'inférence causale, envers les facteurs non-modifiables est également questionnable. Ce sujet a fait, et fait toujours, débat dans la littérature et sort quelque peu du cadre de la présente thèse (voir [228, 229] pour un survol des débats les plus récents). Pour en toucher quelques mots, Holland [14] aussi bien que Rubin [230] disaient dès 1986 qu'une causalité ne pouvait être établie qu'en présence d'une manipulation. Glymour et Glymour [231] argumentaient, au contraire, que l'intervention n'était pas nécessaire. Par exemple, ce n'est pas parce que telle intervention est à ce jour impossible qu'elle ne le sera pour toujours. Naimi et Kaufman [232] démontraient l'intérêt de l'approche causale pour des facteurs non-modifiables, tels que l'ethnie ou le sexe, dans le cadre d'analyses de médiation. L'intérêt est de trouver un moyen d'action (*i.e.*, une variable) sur lequel intervenir pour provoquer un changement dans le facteur non-modifiable d'intérêt.

Hill argumentait dès 1962 de l'intérêt pour le biostatisticien de s'imprégner de l'environnement clinique dans lequel il évolue [233]. De façon analogue, un clinicien ne devrait-il pas s'intéresser aux méthodes d'analyses statistiques ? La méconnaissance des outils statistiques a conduit à de nombreuses publications problématiques [234]. Plug-Stat® vise à faire un pont entre le clinicien et une analyse correcte de ses données. Pour faciliter son utilisation, nous avons ici cherché à automatiser plusieurs étapes de l'estimation causale notamment la vérification des hypothèses, encore trop peu étudiées [89, 90, 235]. De plus, l'utilisation d'un logiciel vérifié et validé permettrait d'éviter des erreurs de codage potentiellement délétères [236] et réduirait le biais humain améliorant ainsi l'analyse et la transparence des résultats [103]. Similairement, un nombre croissant de revues demandent désormais la publication du code statistique et/ou les données utilisé pour la publication [237, 238]. L'export des données est aisé depuis Plug-Stat® sous réserve d'un accord réglementaire, tandis que celui du code R est théoriquement possible. Néanmoins, il est important de noter que la facilité d'analyse augmente le risque de *cherry picking* et de *p-hacking* [239, 240]. En effet, la pression autour de la publication [241] et une volonté de rentabiliser l'achat de Plug-Stat® pourraient mener à de nombreuses analyses afin d'en trouver des publiables, car positives [242], et ce à tort. Ce phénomène a déjà été observé pour les méta-analyses [243]. Enfin, Moodie et Stephens [103] notent que l'automatisation des analyses peut limiter la réflection clinique autour de la planification de l'étude en excluant les méthodologistes.

Une solution, dépendante de la politique commerciale de l'entreprise IDBC, serait de proposer des prestations d'analyses incluant la planification en utilisant Plug-Stat® en interne. Une telle utilisation de Plug-Stat® rendrait la réalisation des analyses statistiques plus rapide et plus sûre tout en facilitant le recrutement et la formation des statisticiens. Plug-Stat® gagnerait également a être étendu à des contextes plus variés. Citons comme exemples les traitements continus [205] et les analyses de médiation [29]. Notons que parmi les analyses proposées par Plug-Stat® figure la survie nette. Des approches de GC et de SDR ont récemment été proposées dans ce contexte [244]. L'extension aux traitements dépendant du temps [11] est également envisageable. Néanmoins, un phénomène connu avec la GC dans ce contexte est le *g-null paradox*. En présence de rétroaction entre le traitement et un facteur de confusion, l'hypothèse nulle aura tendance à être rejetée à tort même si l'estimand théorique est identifiable [245]. Enfin, l'estimation de stratégies thérapeutiques adaptatives est un domaine de recherche particulièrement croissant [246, 247]. Dans une optique conditionnelle plutôt que marginale, il serait pertinent de s'intéresser à ce type de méthodes.

Un problème majeur des études observationnelles est qu'il n'est pas possible de contrôler la confusion due à des variables non-mesurées. Une analyse de sensibilité permet d'avoir une idée sur l'importance du biais nécessaire pour changer les résultats. Si de nombreuses méthodes existent [70,248], l'utilisation des *E-values* [71] semble particulièrement pertinente et pourrait être implémentée Plug-Stat®. Contrairement aux autres possibilités, cette approche est facilement implémentable et ne repose pas sur des hypothèses fortes [249,250].

Une autre perspective de recherche serait l'utilisation de combinaisons d'approches de ML et d'estimation causale pour les ECR. L'ajustement sur des facteurs de risque pré-randomisation de l'évènement est recommandé par les régulateurs internationaux [251,252] afin d'améliorer la précision de l'estimation de l'effet causal [253,254]. De plus, lorsqu'une variable d'ajustement est utilisée dans le processus de randomisation (*e.g.*, randomisation stratifiée), cet ajustement permet de corriger l'erreur de type I [255]. Enfin, cet ajustement est également intéressant pour corriger une éventuelle confusion aléatoire, liée aux déséquilibres résiduels entre les variables, courantes dans les échantillons modérés [256]. L'ajustement est classiquement réalisé au moyen de régressions traditionnelles estimant un effet causal conditionnel au lieu de marginal [257]. Ce changement d'estimand théorique peut être néfaste selon l'interprétation souhaitée de l'ECR. Notons qu'une seule approche marginale est présentée dans le dernier rapport de la FDA [251], celle de Ge *et al.* [258] qui est similaire à la GC. Les méthodes d'estimation causale permettent de résoudre ce problème en ciblant un effet marginal et le ML peut être à nouveau intéressant afin d'obtenir une spécification correcte du modèle de travail. Moore et van der Laan [259] notaient que, pour un EDR, il serait également opportun d'estimer $g(\mathrm{L})$ avec du ML dans ce contexte. De plus, le ML peut être utilisé pour sélectionner l'ensemble d'ajustement [260].

Enfin, de nombreuses approches de ML tenant compte de la censure à droite sont développées [261,262]. L'incorporation de ces approches au sein de méthodes causales est un sujet actuellement étudié par notre équipe.

7.5 Implications pratiques pour Plug-Stat®

L'IPW implémentée dans Plug-Stat® pourraient être remplacée par la GC où $\mathrm{Q}(\mathrm{A},\mathrm{L})$ serait estimé par un SL. Les seuls réels avantages de l'IPW sont de pouvoir vérifier le respect de l'hypothèse de positivité et d'être intuitivement plus pertinente en présence d'un évènement rare. Cependant, l'algorithme développé dans le chapitre 6 permet de vérifier le respect de l'hypothèse quelle que soit la méthode employée. Le chapitre 5 montrait que la GC et la SDR possèdent des performances similaires, voire supérieures, à l'IPW en présence d'un fort taux de censure, donc d'un évènement rare. De plus, puisque la GC et les EDR ne nécessitent pas de vérifier l'équilibre entre les groupes du pseudo-échantillon pondéré, leur utilisation en est facilitée.

L'utilisation de ML en GC semble être une perspective intéressante. Néanmoins, cette combinaison n'est pas supportée par la théorie, pouvant compliquer la publication d'études l'employant. Les EDR peuvent être envisagés afin d'obtenir une spécification correcte dans ce contexte puisque leur association avec le ML est supportée par la théorie. Une critique supplémentaire de l'association entre la GC et le ML est que les facteurs de confusion faiblement liés à l'évènement seraient potentiellement exclus de $\mathrm{Q}(\mathrm{A},\mathrm{L})$ par certaines approches de ML qui sélectionnent les variables. Ce problème peut être surmonté en combinant des approches de ML faisant une sélection et d'autres n'en faisant pas dans le SL. Les EDR, au contraire de la GC, tiendraient compte de ces variables en les sélectionnant dans $g(\mathrm{L})$, évitant ainsi ce phénomène de confusion résiduelle. L'utilisation des courbes d'influence avec la TMLE est à la fois une force et une faiblesse. Elle permet une estimation de la variance supportée par la théorie et se révèle être plus rapide que le bootstrap. Néanmoins, l'estimation peut être biaisée dans des conditions réelles telles que de la non-positivité aléatoire ou une prévalence du traitement déséquilibré (cf Annexe C). Ceci est un problème majeur en pratique puisque la majorité des études sont sujettes à des violations aléatoires de positivité et peuvent comparer des groupes déséquilibrés. Notons qu'une extension de la

TMLE, dite TMLE collaborative, a été récemment proposée pour lutter contre ces problèmes [212]. L'utilisation du boostrap serait une solution pour estimer correctement la variance dans ces situations problématiques.

L'estimation classique de la variance, basée sur le test statistique final, ne tient compte que dudit test. Lorsque des étapes de modélisation, telle que la sélection des variables, sont nécessaires, cette estimation de la variance n'est plus correcte [263]. Ainsi, bootstrapper l'entièreté de la procédure permet également d'obtenir une estimation valide de la variance en tenant compte des différentes étapes de modélisation [264].

De fait, Plug-Stat®gagnerait à proposer la feuille de route suivante à ses utilisateurs :

1. Définition de l'évènement et des modalités de traitement
2. Définition de la population cible au moyen des critères d'éligibilité et définition de l'estimand théorique
3. Choix, sur expertise clinique, de l'ensemble d'ajustement L incluant les facteurs de confusion connus et des facteurs de risque
4. Application de PCART sur L afin de vérifier le respect de l'hypothèse de positivité
   (a) En cas de sous-groupes problématiques, juger sur expertise clinique si la violation est structurelle ou aléatoire
   (b) Si violation structurelle, redéfinir les critères d'éligibilité à l'étape 2
5. Application de la GC avec un SL pour estimer l'effet causal
6. Analyse de sensibilité pour une éventuelle confusion résiduelle

Ce processus se rapproche à la fois du *target trial* [10] et de la feuille de route causale [54]. Ces approches facilitent la réflexion autour de l'étude en amont, à la fois sur l'identifiabilité et sur l'intérêt en vie réelle [265–267]. L'occurrence d'un biais de temps immortel est déjà géré par Plug-Stat®lors de l'implémentation de la cohorte [268]. À ce jour, les deux premières étapes sont déjà implémentées dans Plug-Stat®et une version fonctionnelle de la quatrième étape a été récemment implémentée sur un serveur de développement. La troisième étape consiste actuellement à définir L d'après un seuil de significativité statistique déterminé par l'utilisateur. Cependant, une sélection basée sur expertise clinique est déjà possible. La cinquième étape est basée sur l'IPW ou sur des régressions traditionnelles (*e.g.*, modèles linéaires généralisés ou modèle de Cox) selon le type d'évènement étudié et l'estimand ciblé. L'implémentation d'une nouvelle méthode sera donc à prévoir. La dernière étape serait, quant à elle, nouvelle dans Plug-Stat®.

Dans le futur, le module prédictif présenté dans l'annexe A pourra être utilisé dans un objectif causal afin de prédire, par exemple, l'efficacité attendue de diverses stratégies thérapeutiques chez un individu donné. C'est-à-dire de la prédiction contrefactuelle [269].

Pour conclure, la GC semble être une méthode à considérer pour l'automatisation de Plug-Stat®. Elle ne nécessite pas d'hypothèse d'équilibre au contraire des méthodes basées sur le SP. L'utilisation de ML permet de construire un modèle de travail correctement spécifié et évite l'utilisation d'EDR qui sont moins accessibles pour le clinicien. Les bonnes performances de la GC ont été vérifiées empiriquement par des études de simulation. Enfin, la vérification de la positivité est également automatisée au moyen d'arbres de décision pour permettre à l'utilisateur de redéfinir sa population d'étude.

Bibliographie

[1] Le Borgne F, Fournier MC, Loncle C, Foucher Y. Plug-Stat® : un nouveau logiciel statistique sur mesure pour mieux valoriser les données de cohortes. Revue d'Épidémiologie et de Santé Publique. 2018 May;66 :S179–S180. 2

[2] Foucher Y, Le Borgne F, Legendre C, Morelon E, Buron F, Girerd S, et al. Lack of impact of pre-emptive deceased-donor kidney transplantation on graft outcomes : a propensity score-based study. Nephrology, Dialysis, Transplantation. 2019;34(5) :886–891. 2

[3] Masset C, Boucquemont J, Garandeau C, Buron F, Morelon E, Girerd S, et al. Induction Therapy in Elderly Kidney Transplant Recipients With Low Immunological Risk. Transplantation. 2020 Mar;104(3) :613–622. 2

[4] Ville S, Branchereau J, Cornuaud A, Dantal J, Legendre C, Buron F, et al. The weekend effect in kidney transplantation outcomes : a French cohort-based study. Transplant International. 2020 May;33(9) :1030-9. 2

[5] Foucher Y, Fournier MC, Legendre C, Morelon E, Buron F, Girerd S, et al. Comparison of machine perfusion versus cold storage in kidney transplant recipients from expanded criteria donors : a cohort-based study. Nephrology, Dialysis, Transplantation. 2020 Jun;35(6) :1043–1070. 2

[6] Foucher Y, Lorent M, Albano L, Roux S, Pernin V, Le Quintrec M, et al. Renal transplantation outcomes in obese patients : a French cohort-based study. BMC nephrology. 2021 Mar;22(1) :79. 2

[7] Westreich D. Epidemiology by design : a causal approach to the health sciences. Oxford University Press; 2020. 4, 7, 10

[8] Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Social Science & Medicine. 2018 Aug;210 :2–21. 4

[9] Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, et al. Observational Studies Analyzed Like Randomized Experiments : An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. Epidemiology. 2008 Nov;19(6) :766–779. 4

[10] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. American Journal of Epidemiology. 2016;183(8) :758–764. 4, 11, 90

[11] Hernán M, Robins JM. Causal Inference : What if? Chapman & Hall/CRC; 2020. 4, 7, 10, 15, 88

[12] Hill AB. The Environment and Disease : Association or Causation? Proceedings of the Royal Society of Medicine. 1965 May;58(5) :295–300. 4

[13] Höfler M. The Bradford Hill considerations on causality : a counterfactual perspective. Emerging Themes in Epidemiology. 2005 Nov;2(1) :11. 4

[14] Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986 Dec;81(396) :945–960. 4, 88

[15] Ioannidis JPA. Exposure-wide epidemiology : revisiting Bradford Hill. Statistics in Medicine. 2016;35(11) :1749–1762. 4

[16] Shimonovich M, Pearce A, Thomson H, Keyes K, Katikireddi SV. Assessing causality in epidemiology : revisiting Bradford Hill to incorporate developments in causal thinking. European Journal of Epidemiology. 2021 Sep ;36(9) :873-87. 5

[17] Ferguson KD, McCann M, Katikireddi SV, Thomson H, Green MJ, Smith DJ, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs) : a novel and systematic method for building directed acyclic graphs. International Journal of Epidemiology. 2020 Feb ;49(1) :322–329. 5, 8

[18] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999 Jan ;10(1) :37–48. 5

[19] Rohrer JM. Thinking Clearly About Correlations and Causation : Graphical Causal Models for Observational Data. Advances in Methods and Practices in Psychological Science. 2018 Mar ;1(1) :27–42. 5, 6, 8

[20] Pearl J. Causal Diagrams for Empirical Research. Biometrika. 1995 ;82(4) :669–688. 5

[21] Pearl J. Causality : Models, Reasoning and Inference. 2nd ed. USA : Cambridge University Press ; 2009. 5, 15

[22] Pearl J, Mackenzie D. The Book of Why : The New Science of Cause and Effect. 1st ed. USA : Basic Books, Inc. ; 2018. 5, 11

[23] Spirtes P, Glymour CN, Scheines R. Causation, prediction, and search. The MIT Press ; 2000. 5

[24] Suzuki E, Shinozaki T, Yamamoto E. Causal Diagrams : Pitfalls and Tips. Journal of Epidemiology. 2020 Apr ;30(4) :153–162. 5, 8

[25] Tennant PW, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research : review and recommendations. International Journal of Epidemiology. 2021 ;17(50) :620-32. 5, 7, 8

[26] Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. International Journal of Epidemiology. 2011 Jun ;40(3) :780–785. 6, 8, 11

[27] Simpson EH. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society Series B (Methodological). 1951 ;13(2) :238–241. 6

[28] Arnold KF, Davies V, de Kamps M, Tennant PWG, Mbotwa J, Gilthorpe MS. Reflections on modern methods : generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. International Journal of Epidemiology. 2020 Dec ;49(6) :2074-82. 6

[29] Vanderweele TJ. Explanation in causal inference : Methods for mediation and interaction. Oxford University Press ; 2015. 6, 88

[30] Geiger D, Verma T, Pearl J. Identifying independence in bayesian networks. Networks. 1990 ;20(5) :507–534. 7

[31] VanderWeele TJ, Shpitser I. On the definition of a confounder. Annals of statistics. 2013 Feb ;41(1) :196–220. 7

[32] Hernán MA, Hernández-Díaz S, Robins JM. A Structural Approach to Selection Bias :. Epidemiology. 2004 Sep ;15(5) :615–625. 7

[33] Luque-Fernandez MA, Schomaker M, Redondo-Sanchez D, Jose Sanchez Perez M, Vaidya A, Schnitzer ME. Educational Note : Paradoxical collider effect in the analysis of non-communicable disease epidemiological data : a reproducible illustration and web application. International Journal of Epidemiology. 2019 Apr ;48(2) :640–653. 7

[34] Nguyen TQ, Schmid I, Stuart EA. Clarifying causal mediation analysis for the applied researcher : Defining effects based on what we want to learn. Psychological Methods. 2021 ;26(2) :255–271. 7

[35] Banack HR, Kaufman JS. Does selection bias explain the obesity paradox among individuals with cardiovascular disease? Annals of Epidemiology. 2015 May;25(5) :342–349. 8

[36] Hernández-Díaz S, Wilcox AJ, Schisterman EF, Hernán MA. From causal diagrams to birth weight-specific curves of infant mortality. European journal of epidemiology. 2008;23(3) :163–166. 8

[37] Robins JM. Data, Design, and Background Knowledge in Etiologic Inference. Epidemiology. 2001 May;12(3) :313–320. 8

[38] Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation : an application to birth defects epidemiology. American Journal of Epidemiology. 2002 Jan;155(2) :176–184. 8

[39] Nilsson A, Bonander C, Strömberg U, Björk J. A directed acyclic graph for interactions. International Journal of Epidemiology. 2021;50(2) :613–619. 8

[40] van der Laan MJ, Rose S. Targeted learning : causal inference for observational and experimental data. Springer series in statistics. Springer; 2011. 8, 11, 17, 85

[41] Greenland S. For and Against Methodologies : Some Perspectives on Recent Causal and Statistical Inference Debates. European Journal of Epidemiology. 2017 Jan;32(1) :3–20. 8

[42] Suzuki E, Mitsuhashi T, Tsuda T, Yamamoto E. A typology of four notions of confounding in epidemiology. Journal of Epidemiology. 2017 Feb;27(2) :49–55. 8

[43] Glymour C, Zhang K, Spirtes P. Review of Causal Discovery Methods Based on Graphical Models. Frontiers in Genetics. 2019;10. 8

[44] VanderWeele TJ. Principles of confounder selection. European Journal of Epidemiology. 2019 Mar;34(3) :211–219. 8

[45] Ikram MA. The disjunctive cause criterion by VanderWeele : An easy solution to a complex problem? European Journal of Epidemiology. 2019;34(3) :223–224. 8

[46] Witte J, Didelez V. Covariate selection strategies for causal inference : Classification and comparison. Biometrical Journal. 2019;61(5) :1270–1289. 8

[47] Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009 Jul;20(4) :512–522. 8

[48] Schnitzer ME, Lok JJ, Gruber S. Variable selection for confounder control, flexible modeling and Collaborative Targeted Minimum Loss-based Estimation in causal inference. The International Journal of Biostatistics. 2016 May;12(1) :97–115. 8, 85

[49] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5) :688–701. 9

[50] Neyman J. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Statistical Science. 1990;5(4) :465–472. Traduction de la publication originale polonaise par Dabrowska, D. M. et Speed, T. P. 9

[51] 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 Jan;7(9) :1393–1512. 9, 15

[52] Hernán MA. A definition of causal effect for epidemiological research. Journal of Epidemiology & Community Health. 2004 Apr;58(4) :265–271. 9, 10, 11

[53] Lundberg I, Johnson R, Stewart BM. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. American Sociological Review. 2021 Jun;86(3) :532–565. 9, 10

[54] Petersen ML, van der Laan MJ. Causal Models and Learning from Data : Integrating Causal Modeling and Statistical Estimation. Epidemiology. 2014 May;25(3) :418–426. 9, 10, 90

[55] Díaz I. Machine learning in the estimation of causal effects : targeted minimum loss-based estimation and double/debiased machine learning. Biostatistics. 2020 Apr;21(2) :353–358. 9, 85

[56] Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right : A Classification of Data Science Tasks. Chance. 2019 Jan;32(1) :42–49. 9

[57] Schomaker M. Regression and Causality. arXiv :200611754. 2020 Jun. *Pre-print non publié à ce jour.* 9

[58] Mao H, Li L, Yang W, Shen Y. On the propensity score weighting analysis with survival outcome : Estimands, estimation, and inference. Statistics in Medicine. 2018;37(26) :3745–3763. 10

[59] Imbens GW. Nonparametric Estimation of Average Treatment Effects Under Exogeneity : A Review. The Review of Economics and Statistics. 2004 Feb;86(1) :4–29. 10, 13

[60] Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores : From naïve enthusiasm to intuitive understanding. Statistical Methods in Medical Research. 2012 Jun;21(3) :273–293. 10, 13, 19

[61] Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA. A causal framework for classical statistical estimands in failure-time settings with competing events. Statistics in Medicine. 2020 Apr;39(8) :1199–1236. 10

[62] Dawid AP. Conditional Independence in Statistical Theory. Journal of the Royal Statistical Society Series B (Methodological). 1979;41(1) :1–31. 10

[63] Hernán MA. Does water kill? A call for less casual causal inferences. Annals of Epidemiology. 2016 Oct;26(10) :674–680. 10

[64] Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. International Journal of Obesity. 2008 Aug;32(S3) :S8–S14. 10

[65] Hudgens MG, Halloran ME. Toward Causal Inference With Interference. Journal of the American Statistical Association. 2008 Jun;103(482) :832–842. 10

[66] Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. Statistical methods in medical research. 2012 Feb;21(1) :55–75. 10

[67] Breskin A, Edmonds A, Cole SR, Westreich D, Cocohoba J, Cohen MH, et al. G-computation for policy-relevant effects of interventions on time-to-event outcomes. International Journal of Epidemiology. 2020 Dec;49(6) :2021-9. 10, 19

[68] Moore KL, Neugebauer R, van der Laan MJ, Tager IB. Causal inference in epidemiological studies with strong confounding. Statistics in Medicine. 2012;31(13) :1380–1404. 10, 85

[69] 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 Feb;21(1) :31–54. 10, 18, 19, 84

[70] Zhang X, Stamey JD, Mathur MB. Assessing the impact of unmeasured confounders for credible and reliable real-world evidence. Pharmacoepidemiology and Drug Safety. 2020;29(10) :1219–1227. 10, 89

[71] VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research : Introducing the E-Value. Annals of Internal Medicine. 2017 Aug;167(4) :268-74. 10, 89

[72] Pirracchio R, Carone M, Rigon MR, Caruana E, Mebazaa A, Chevret S. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. Statistical Methods in Medical Research. 2016 Oct;25(5) :1968-54. 11

[73] Greifer N, Stuart EA. Choosing the Estimand When Matching or Weighting in Observational Studies. arXiv :210610577. 2021 Jun. *Pre-print non publié à ce jour.* 11

[74] Broadbent A. Causation and prediction in epidemiology : A guide to the "Methodological Revolution". Studies in History and Philosophy of Science Part C : Studies in History and Philosophy of Biological and Biomedical Sciences. 2015 Dec;54 :72–80. 11

[75] Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. Annual Review of Public Health. 2013;34(1) :61–75. 11

[76] Whitcomb BW, Naimi AI. Defining, Quantifying, and Interpreting "Noncollapsibility" in Epidemiologic Studies of Measures of "Effect". American Journal of Epidemiology. 2021 May;190(5) :697-700. 11

[77] Greenland S, Pearl J. Collapsibility Analysis using Graphical Models : Adjustments and their Consequences. International Statistical Review. 2011 Dec;79(3) :401–426. 11

[78] Daniel R, Zhang J, Farewell D. Making apples from oranges : Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biometrical Journal. 2021 Mar;63(3) :528-57. 11

[79] Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Analysis. 2013 Jul;19(3) :279–296. 11

[80] Sjölander A, Dahlqwist E, Zetterqvist J. A Note on the Noncollapsibility of Rate Differences and Rate Ratios. Epidemiology. 2016;27(3) :4. 11

[81] Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. American Journal of Epidemiology. 2017 Jan;185(1) :65–73. 11, 17, 18

[82] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983 Apr;70(1) :41–55. 11, 12, 13, 15

[83] Westreich D, Edwards JK, Cole SR, Platt RW, Mumford SL, Schisterman EF. Imputation approaches for potential outcomes in causal inference. International Journal of Epidemiology. 2015 Oct;44(5) :1731–1737. 11, 16

[84] Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly Robust Estimation of Causal Effects. American Journal of Epidemiology. 2011 Apr;173(7) :761–767. 11, 16, 17

[85] Hernán MA, Robins JM. Instruments for Causal Inference : An Epidemiologist's Dream? Epidemiology. 2006 Jul;17(4) :360–372. 11

[86] Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 1996;91(434) :444–455. 11

[87] Caniglia EC, Murray EJ. Difference-in-Difference in the Time of Cholera : a Gentle Introduction for Epidemiologists. Current Epidemiology Reports. 2020 Dec;7(4) :203–211. 11

[88] Dukes O, Vansteelandt S. A Note on G-Estimation of Causal Risk Ratios. American Journal of Epidemiology. 2018 May;187(5) :1079–1084. 11

[89] Grose E, Wilson S, Barkun J, Bertens K, Martel G, Balaa F, et al. Use of Propensity Score Methodology in Contemporary High-Impact Surgical Literature. Journal of the American College of Surgeons. 2020 Jan;230(1) :101-12.e2. 12, 13, 88

[90] Ali MS, Groenwold RHH, Belitser SV, Pestman WR, Hoes AW, Roes KCB, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal : a systematic review. Journal of Clinical Epidemiology. 2015 Feb;68(2) :112–121. 12, 13, 14, 88

[91] Webster-Clark M, Stürmer T, Wang T, Man K, Marinac-Dabic D, Rothman KJ, et al. Using propensity scores to estimate effects of treatment initiation decisions : State of the science. Statistics in Medicine. 2021;40(7) :1718–1735. 12, 13, 84

[92] Thoemmes FJ, Kim ES. A Systematic Review of Propensity Score Methods in the Social Sciences. Multivariate Behavioral Research. 2011 Feb;46(1) :90–118. 12

[93] King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Political Analysis. 2019;27(4) :434–454. 12, 14

[94] Joffe MM, Rosenbaum PR. Invited Commentary : Propensity Scores. American Journal of Epidemiology. 1999 Aug;150(4) :327–333. 12

[95] Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Statistics in Medicine. 1997 Feb;16(1–3) :285–319. 12

[96] Westreich D, Greenland S. The Table 2 Fallacy : Presenting and Interpreting Confounder and Modifier Coefficients. American Journal of Epidemiology. 2013 Feb;177(4) :292–298. 12

[97] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable Selection for Propensity Score Models. American Journal of Epidemiology. 2006 Jun;163(12) :1149–1156. 12

[98] De Luna X, Waernbaum I, Richardson TS. Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika. 2011 Dec;98(4) :861–875. 12

[99] Lefebvre G, Delaney JAC, 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. 12

[100] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates. American Journal of Epidemiology. 2011 Dec;174(11) :1213–1222. 12

[101] Pearl J. Invited Commentary : Understanding Bias Amplification. American Journal of Epidemiology. 2011 Dec;174(11) :1223–1227. 12

[102] Austin PC, Grootendorst P, Normand SLT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect : a Monte Carlo study. Statistics in Medicine. 2007 Feb;26(4) :754–768. 12, 13, 87

[103] Moodie EEM, Stephens DA. Treatment Prediction, Balance, and Propensity Score Adjustment. Epidemiology. 2017 Sep;28(5) :e51. 12, 88

[104] Pirracchio R, Carone M. The Balance Super Learner : A robust adaptation of the Super Learner to improve estimation of the average treatment effect in the treated based on propensity score matching. Statistical Methods in Medical Research. 2018 Aug;27(8) :2504–2518. 12

[105] Westreich D, Lessler J, Funk MJ. Propensity score estimation : neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. Journal of Clinical Epidemiology. 2010 Aug;63(8) :826–833. 13, 18

[106] Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. American Journal of Epidemiology. 2015 Jan;181(2) :108–119. 13, 18

[107] Fong C, Hazlett C, Imai K. Covariate balancing propensity score for a continuous treatment : Application to the efficacy of political advertisements. The Annals of Applied Statistics. 2018 Mar;12(1) :156–177. 13, 18

[108] Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000 Sep;11(5) :550–560. 13, 14, 15

[109] Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. Statistics in Medicine. 2014;33(23) :4053–4072. 13

[110] Austin PC. The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. Medical Decision Making. 2009 Nov;29(6) :661–677. 13

[111] Garrido MM. Covariate Adjustment and Propensity Scores. JAMA. 2016 Apr;315(14) :1521–1522. 13

[112] Lanza ST, Moore JE, Butera NM. Drawing Causal Inferences Using Propensity Scores : A Practical Guide for Community Psychologists. American Journal of Community Psychology. 2013 Dec;52(3–4) :380–392. 13

[113] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011 May;46(3) :399–424. 13, 14

[114] Jackson JW, Schmid I, Stuart EA. Propensity Scores in Pharmacoepidemiology : Beyond the Horizon. Current Epidemiology Reports. 2017 Dec;4(4) :271–280. 13, 84

[115] Stuart EA. Matching methods for causal inference : A review and a look forward. Statistical Science. 2010 Feb;25(1) :1–21. 13

[116] Shiba K, Kawahara T. Using propensity scores for causal inference : pitfalls and tips. Journal of Epidemiology. 2021;31(8) :457-63. 13, 14, 15

[117] Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Statistics. 2011 Apr;10(2) :150–161. 13

[118] Austin PC. A comparison of 12 algorithms for matching on the propensity score. Statistics in Medicine. 2014 Mar;33(6) :1057–1069. 14

[119] Austin PC, Stuart EA. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. Statistical Methods in Medical Research. 2017 Dec;26(6) :2505–2525. 14

[120] Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Statistics in Medicine. 2014;33(10) :1685–1699. 14

[121] Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society : Series A (Statistics in Society). 2008 Apr;171(2) :481–502. 14

[122] Johnson DH. The Insignificance of Statistical Significance Testing. The Journal of Wildlife Management. 1999;63(3) :763–772. 14

[123] Mansournia MA, Nazemipour M, Naimi AI, Collins GS, Campbell MJ. Reflections on modern methods : demystifying robust standard errors for epidemiologists. International Journal of Epidemiology. 2021 Feb;50(1) :346-51. 14

[124] Morgan SL, Todd JJ. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects. Sociological Methodology. 2008 Aug;38(1) :231–282. 14

[125] 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. 14

[126] Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score : a primer for practitioners. BMJ. 2019 Oct;367 :l5657. 14

[127] Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of Stabilized Inverse Propensity Scores as Weights to Directly Estimate Relative Risk and Its Confidence Intervals. Value in Health. 2010 Mar;13(2) :273–277. 14

[128] Sato T, Matsuyama Y. Marginal Structural Models as a Tool for Standardization :. Epidemiology. 2003 Nov;14(6) :680–686. 14

[129] Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine. 2015 Dec;34(28) :3661–3679. 14

[130] Hirano K, Imbens GW, Ridder G. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica. 2003;71(4) :1161–1189. 15

[131] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects : a comparative study. Statistics in Medicine. 2004 Oct;23(19) :2937–2960. 15, 17

[132] Joffe MM, Have TRT, Feldman HI, Kimmel SE. Model Selection, Confounder Control, and Marginal Structural Models : Review and New Applications. The American Statistician. 2004 ;58(4) :272–279. 15

[133] Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. Computer Methods and Programs in Biomedicine. 2004 Jul ;75(1) :45–49. 15

[134] Breskin A, Cole SR, Westreich D. Exploring the subtleties of inverse probability weighting and marginal structural models. Epidemiology. 2018 May ;29(3) :352–355. 15

[135] Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. Statistics in Medicine. 2016 Dec ;35(30) :5642–5655. 15

[136] Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JAC. Methods for dealing with time-dependent confounding. Statistics in Medicine. 2013 Apr ;32(9) :1584–1618. 15

[137] Naimi AI, Cole SR, Kennedy EH. An Introduction to G Methods. International Journal of Epidemiology. 2016 Dec ;46(2) :756–762. 15

[138] Neison FGP. On a Method Recently Proposed for Conducting Inquiries into the Comparative Sanatory Condition of Various Districts, with Illustrations, Derived from Numerous Places in Great Britain at the Period of the Last Census. Journal of the Statistical Society of London. 1844 ;7(1) :40–68. 15

[139] Keiding N, Clayton D. Standardization and Control for Confounding in Observational Studies : A Historical Perspective. Statistical Science. 2014 Nov ;29(4) :529–558. 15

[140] Zhang Z. Estimating a Marginal Causal Odds Ratio Subject to Confounding. Communications in Statistics - Theory and Methods. 2008 Dec ;38(3) :309–321. 15

[141] Hernán MA, Robins JM. Estimating causal effects from epidemiological data. Journal of Epidemiology and Community Health. 2006 Jul ;60(7) :578–586. 15, 16

[142] Vansteelandt S, Keiding N. Invited Commentary : G-Computation-Lost in Translation? American Journal of Epidemiology. 2011 Apr ;173(7) :739–742. 15, 17

[143] Sjölander A. Regression standardization with the R package stdReg. European Journal of Epidemiology. 2016 Jun ;31(6) :563–574. 15

[144] Snowden JM, Rose S, Mortimer KM. Implementation of G-Computation on a Simulated Data Set : Demonstration of a Causal Inference Technique. American Journal of Epidemiology. 2011 Apr ;173(7) :731–738. 15, 16

[145] Wang A, Nianogo RA, Arah OA. G-computation of average treatment effects on the treated and the untreated. BMC Medical Research Methodology. 2017 Dec ;17(1). 16

[146] Keil AP, Buckley JP, M OK, Ferguson KK, Zhao S, White AJ. A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures. Environmental Health Perspectives. 2020 Apr ;128(4) :047004. 16

[147] Keil AP, Daza EJ, Engel SM, Buckley JP, Edwards JK. A Bayesian approach to the g-formula. Statistical Methods in Medical Research. 2018 Oct ;27(10) :3183–3204. 16

[148] Josefsson M, Daniels MJ. Bayesian semi-parametric G-computation for causal inference in a cohort study with MNAR dropout and death. Journal of the Royal Statistical Society : Series C (Applied Statistics). 2021 Mar ;70(2) :398-414. 16

[149] Wang A, Arah OA. G-computation demonstration in causal mediation analysis. European Journal of Epidemiology. 2015 Oct ;30(10) :1119–1127. 16

[150] Pósch K. Testing Complex Social Theories With Causal Mediation Analysis and G-Computation : Toward a Better Way to Do Causal Structural Equation Modeling. Sociological Methods & Research. 2021 Aug ;50(3) :1376–1406. 16

[151] Tchetgen Tchetgen EJ, Fulcher IR, Shpitser I. Auto-G-Computation of Causal Effects on a Network. Journal of the American Statistical Association. 2021 Aug ;116(534) :833-44. 16

[152] Vock DM, Vock LFB. Estimating the effect of plate discipline using a causal inference framework : an application of the G-computation algorithm. Journal of Quantitative Analysis in Sports. 2018 Jun ;14(2) :37–56. 16

[153] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment : Performance of Double-Robust Estimators When “Inverse Probability” Weights Are Highly Variable. Statistical Science. 2007 Nov ;22(4) :544–559. 16, 17, 84

[154] Lendle SD, Fireman B, van der Laan MJ. Targeted maximum likelihood estimation in safety analysis. Journal of Clinical Epidemiology. 2013 Aug ;66(8) :S91–S98. 16, 18, 85

[155] Tan Z. Comment : Understanding OR, PS and DR. Statistical Science. 2007 Nov ;22(4) :560–568. 16, 85

[156] Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. Journal of the American Statistical Association. 1994 Sep ;89(427) :846–866. 16, 17

[157] Kreif N, DiazOrdaz K. Machine Learning in Policy Evaluation : New Tools for Causal Inference. Oxford Research Encyclopedia of Economics and Finance. 2019 Jul. 16, 17, 85, 86

[158] Glynn AN, Quinn KM. An Introduction to the Augmented Inverse Propensity Weighted Estimator. Political Analysis. 2010 ;18(1) :36–56. 16, 17

[159] Kang JDY, Schafer JL. Demystifying Double Robustness : A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science. 2007 Nov ;22(4) :523–539. 17, 85

[160] Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. Statistical Methods in Medical Research. 2012 Feb ;21(1) :7–30. 17

[161] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005 Dec ;61(4) :962–973. 17

[162] Hampel FR. The Influence Curve and Its Role in Robust Estimation. Journal of the American Statistical Association. 1974 Jun ;69(346) :383–393. 17

[163] van der Vaart A. Higher Order Tangent Spaces and Influence Functions. Statistical Science. 2014 Nov ;29(4) :679–686. 17

[164] Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S. Demystifying statistical learning based on efficient influence functions. arXiv :210700681. 2021 Jul. *Pré-print non publié à ce jour.* 17

[165] Fisher A, Kennedy EH. Visually Communicating and Teaching Intuition for Influence Functions. The American Statistician. 2021 Apr ;75(2) :162–172. 17

[166] Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment : A tutorial. Statistics in Medicine. 2018 Jul ;37(16) :2530–2546. 17

[167] van der Laan MJ, Rubin D. Targeted Maximum Likelihood Learning. The International Journal of Biostatistics. 2006 Dec ;2(1) :Article 11. 17

[168] Porter KE, Gruber S, van der Laan MJ, Sekhon JS. The Relative Performance of Targeted Maximum Likelihood Estimators. The International Journal of Biostatistics. 2011 Aug ;7(1) :Article 31. 17, 85

[169] Pang M, Schuster T, Filion KB, Eberg M, Platt RW. Targeted Maximum Likelihood Estimation for Pharmacoepidemiologic Research. Epidemiology. 2016 Jul ;27(4) :570–577. 17

[170] Kreif N, Gruber S, Radice R, Grieve R, Sekhon JS. Evaluating treatment effectiveness under model misspecification : A comparison of targeted maximum likelihood estimation with bias-corrected matching. Statistical Methods in Medical Research. 2016 Oct ;25(5) :2315–2336. 17

[171] Benkeser D, Carone M, van der Laan MJ, Gilbert PB. Doubly robust nonparametric inference on the average treatment effect. Biometrika. 2017 Dec ;104(4) :863–880. 17

[172] Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Statistics in Medicine. 2010 Sep ;29(20) :2137–2148. 18

[173] Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Statistics in Medicine. 2013 Jul ;32(16) :2837–2849. 18

[174] Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes : A simulation study. Statistical Methods in Medical Research. 2016 Oct ;25(5) :2214–2237. 18

[175] Le Borgne F, Giraudeau B, Querard AH, Giral M, Foucher Y. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors : practical illustrations in kidney transplantation. Statistics in Medicine. 2016 Mar ;35(7) :1103–1116. 18

[176] Gayat E, Resche-Rigon M, Mary JY, Porcher R. Propensity score applied to survival data analysis through proportional hazards models : a Monte Carlo study. Pharmaceutical Statistics. 2012 ;11(3) :222–229. 18, 87

[177] Hajage D, Tubach F, Steg PG, Bhatt DL, De Rycke Y. On the use of propensity scores in case of rare exposure. BMC Medical Research Methodology. 2016 Dec ;16(1) :38. 18

[178] Abdia Y, Kulasekera KB, Datta S, Boakye M, Kong M. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated : A comparative study. Biometrical Journal. 2017 Sep ;59(5) :967–985. 18

[179] Neugebauer R, van der Laan M. Why prefer double robust estimators in causal inference? Journal of Statistical Planning and Inference. 2005 Feb ;129(1–2) :405–426. 18, 85

[180] Colson KE, Rudolph KE, Zimmerman SC, Goin DE, Stuart EA, Laan Mvd, et al. Optimizing matching and analysis combinations for estimating causal effects. Scientific Reports. 2016 Mar ;6(1). 18, 85

[181] Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods : when worlds collide—prediction, machine learning and causal inference. International Journal of Epidemiology. 2021 Jan ;49(6) :2058-64. 18

[182] Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. American Journal of Epidemiology. 2019 Dec ;188(12) :2222-39. 18

[183] Alam S, Moodie EEM, Stephens DA. Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. Statistics in Medicine. 2019 Apr ;38(9) :1690–1702. 18

[184] Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statistics in medicine. 2010 Feb ;29(3) :337–346. 18

[185] Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation : a simulation study. Pharmacoepidemiology and drug safety. 2008 Jun ;17(6) :546–555. 18

[186] Neugebauer R, Fireman B, Roy JA, Raebel MA, Nichols GA, O'Connor PJ. Super learning to hedge against incorrect inference from arbitrary parametric assumptions in marginal structural modeling. Journal of Clinical Epidemiology. 2013 Aug ;66(8) :S99–S109. 18

[187] Gruber S, Logan RW, Jarrín I, Monge S, Hernán MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. Statistics in Medicine. 2015 Jan ;34(1) :106–117. 18

[188] Austin PC. Using Ensemble-Based Methods for Directly Estimating Causal Effects : An Investigation of Tree-Based G-Computation. Multivariate Behavioral Research. 2012 Feb ;47(1) :115–135. 18

[189] van der Laan M, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology. 2007 Sep ;6(1) :Article 25. 18

[190] Naimi AI, Balzer LB. Stacked generalization : an introduction to super learning. European Journal of Epidemiology. 2018 May;33(5) :459–464. 18

[191] Hernán MA. The Hazards of Hazard Ratios. Epidemiology. 2010 Jan;21(1) :13–15. 19

[192] Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Analysis. 2015 Oct;21(4) :579–593. 19

[193] Keil AP, Edwards JK, Richardson DR, Naimi AI, Cole SR. The parametric G-formula for time-to-event data : towards intuition with a worked example. Epidemiology (Cambridge, Mass). 2014 Nov;25(6) :889–897. 19

[194] Wen L, Young JG, Robins JM, Hernán MA. Parametric g-formula implementations for causal survival analyses. Biometrics. 2021 Jun;77(2) :740-53. 19

[195] Díaz I, Colantuoni E, Hanley DF, Rosenblum M. Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. Lifetime Data Analysis. 2019 Jul;25(3) :439–468. 19

[196] Benkeser D, Carone M, Gilbert PB. Improved estimation of the cumulative incidence of rare outcomes. Statistics in Medicine. 2018;37(2) :280–293. *Sous presse*. 19

[197] Gollob HF, Reichardt CS. Taking Account of Time Lags in Causal Models. Child Development. 1987 Feb;58(1) :80–92. 19

[198] Ferreira Guerra S, Schnitzer ME, Forget A, Blais L. Impact of discretization of the timeline for longitudinal causal inference methods. Statistics in Medicine. 2020 Nov;39(27) :4069–4085. 19

[199] Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology. 2008 Jul;168(6) :656–664. 19

[200] Westreich D, Cole SR. Invited Commentary : Positivity in Practice. American Journal of Epidemiology. 2010 Mar;171(6) :674–677. 19

[201] Bahamyirou A, Blais L, Forget A, Schnitzer ME. Understanding and diagnosing the potential for bias when using machine learning methods with doubly robust causal estimators. Statistical Methods in Medical Research. 2019 Jun;28(6) :1637–1650. 19, 84

[202] Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core Concepts in Pharmacoepidemiology : Violations of the Positivity Assumption in the Causal Analysis of Observational Data : Consequences and Statistical Approaches. Pharmacoepidemiology and Drug Safety. 2021 Nov;30(11) :1471–1485. 84

[203] Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan MJ. Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models. Journal of Causal Inference. 2014 Sep;2(2) :147–185. 85

[204] Keil AP, Mooney SJ, Jonsson Funk M, Cole SR, Edwards JK, Westreich D. Resolving an apparent paradox in doubly robust estimators. American Journal of Epidemiology. 2018 Apr;187(4) :891–892. 85

[205] Zhao S, van Dyk DA, Imai K. Propensity score-based methods for causal inference in observational studies with non-binary treatments. Statistical Methods in Medical Research. 2020 Mar;29(3) :709–727. 85, 88

[206] Naimi AI, Mishler AE, Kennedy EH. Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. American Journal of Epidemiology. 2021 Jul :kwab201. *Sous presse*. 85, 86

[207] Mullainathan S, Spiess J. Machine Learning : An Applied Econometric Approach. Journal of Economic Perspectives. 2017 May;31(2) :87–106. 85

[208] Balzer LB, Petersen ML. Machine Learning in Causal Inference : How do I love thee? Let me count the ways. American Journal of Epidemiology. 2021 Aug;8(8) :1483-7. 85, 86

[209] Bickel PJ, Götze F, van Zwet WR. Resampling Fewer Than n Observations : Gains, Losses, and Remedies for Losses. Statistica Sinica. 1997 ;7(1) :1–31. 85

[210] Zivich PN, Breskin A. Machine Learning for Causal Inference : On the Use of Cross-fit Estimators. Epidemiology. 2021 May ;32(3) :393–401. 85, 86

[211] Mooney SJ, Keil AP, Westreich DJ. 13 Questions About Using Machine Learning in Causal Research (You Won't Believe the Answer to Number 10 !). American Journal of Epidemiology. 2021 Aug ;190(8) :1476-82. 85

[212] Schnitzer ME. Comment : Increasing Real World Usage of Targeted Minimum Loss-Based Estimators. Statistical Science. 2020 ;35(3) :496–498. 85, 90

[213] Balzer LB, Westling T. Demystifying Statistical Inference When Using Machine Learning in Causal Research. American Journal of Epidemiology. 2021 Jul ;(kwab200). *Sous presse*. 85, 86

[214] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal. 2018 Feb ;21(1) :C1–C68. 85, 86

[215] Robinson PM. Root-N-Consistent Semiparametric Regression. Econometrica. 1988 Jul ;56(4) :931–954. 85

[216] Kennedy EH, Balakrishnan S. Discussion of "Data-driven confounder selection via Markov and Bayesian networks" by Jenny Häggström. Biometrics. 2018 Jun ;74(2) :399–402. 85

[217] Hünermund P, Louw B, Caspi I. Double Machine Learning and Bad Controls – A Cautionary Tale. arXiv :210811294. 2021 Aug. *Pré-print non publié à ce jour*. 86

[218] Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics. 2005 May ;21(9) :1979–1986. 86

[219] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019 May ;38(11) :2074–2102. 87, 88

[220] Buchka S, Hapfelmeier A, Gardner PP, Wilson R, Boulesteix AL. On the optimistic performance evaluation of newly introduced bioinformatic methods. Genome Biology. 2021 ;22(1). 87

[221] Boulesteix AL, Lauer S, Eugster MJA. A Plea for Neutral Comparison Studies in Computational Sciences. PLOS ONE. 2013 Apr ;8(4) :e61562. 87

[222] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Statistics in Medicine. 2007 Jul ;26(16) :3078–3094. 87

[223] Austin PC. The performance of different propensity-score methods for estimating relative risks. Journal of Clinical Epidemiology. 2008 Jun ;61(6) :537–545. 87

[224] Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Statistical Methodology. 2005 Jul ;2(2) :131–154. 87

[225] Boulesteix AL. Ten Simple Rules for Reducing Overoptimistic Reporting in Methodological Computational Research. PLOS Computational Biology. 2015 Apr ;11(4) :e1004191. 87

[226] Rytgaard HC, Gerds TA, van der Laan MJ. Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. arXiv :210502088. 2021 May. *Pré-print non publié à ce jour*. 87

[227] Boulesteix AL, Groenwold RH, Abrahamowicz M, Binder H, Briel M, Hornung R, et al. Introduction to statistical simulations in health research. BMJ Open. 2020 Dec ;10(12) :e039921. 88

[228] Robins JM, Weissman MB. Counterfactual causation and streetlamps : what is to be done ? International Journal of Epidemiology. 2016 Dec ;45(6) :1830–1835. 88

[229] Pearl J. Does Obesity Shorten Life ? Or is it the Soda ? On Non-manipulable Causes. Journal of Causal Inference. 2018 Sep ;6(2). 88

[230] Rubin DB. Comment : Which ifs have causal answers. Journal of the American Statistical Association. 1986 Dec;81(396) :961–962. 88

[231] Glymour C, Glymour MR. Commentary : Race and Sex Are Causes. Epidemiology. 2014;25(4) :488–490. 88

[232] Naimi AI, Kaufman JS. Counterfactual Theory in Social Epidemiology : Reconciling Analysis and Action for the Social Determinants of Health. Current Epidemiology Reports. 2015 Mar;2(1) :52–60. 88

[233] Hill AB. The Statistician in Medicine. Journal of the Institute of Actuaries. 1962 Sep;88(2) :178–191. 88

[234] Peng R. The reproducibility crisis in science : A statistical counterattack. Significance. 2015;12(3) :30–32. 88

[235] Lonjon G, Porcher R, Ergina P, Fouet M, Boutron I. Potential Pitfalls of Reporting and Bias in Observational Studies With Propensity Score Analysis Assessing a Surgical Procedure : A Methodological Systematic Review. Annals of Surgery. 2017 May;265(5) :901–909. 88

[236] Schwab S, Held L. Statistical programming : Small mistakes, big impacts. Significance. 2021 Jun;18(3) :6–7. 88

[237] Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible Research : Moving toward Research the Public Can Really Trust. Annals of Internal Medicine. 2007 Mar;146(6) :450–453. 88

[238] Hernán MA, Wilcox AJ. Epidemiology, Data Sharing, and the Challenge of Scientific Replication. Epidemiology. 2009 Mar;20(2) :167–168. 88

[239] Ioannidis JPA. Why Most Published Research Findings Are False. PLOS Medicine. 2005 Aug;2(8) :e124. 88

[240] Hofacker CF. Abuse of statistical packages : the case of the general linear model. American Journal of Physiology. 1983 Sep;245(3) :R299–R302. 88

[241] Neill US. Publish or perish, but at what cost? The Journal of Clinical Investigation. 2008 Jul;118(7) :2368–2368. 88

[242] Sterling TD. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance–Or Vice Versa. Journal of the American Statistical Association. 1959;54(285) :30–34. 88

[243] Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. The Milbank Quarterly. 2016 Sep;94(3) :485–514. 88

[244] Syriopoulou E, Rutherford MJ, Lambert PC. Inverse probability weighting and doubly robust standardization in the relative survival framework. Statistics in Medicine. 2021 Nov;40(27) :6069-92. 88

[245] McGrath S, Young JG, Hernán MA. Revisiting the g-null paradox. Epidemiology. 2021 Oct. *Sous presse*. 88

[246] Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007 Jun;63(2) :447–455. 88

[247] Zhao YQ, Laber EB. Estimation of optimal dynamic treatment regimes. Clinical Trials. 2014 Aug;11(4) :400–407. 88

[248] Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. International Journal of Epidemiology. 2014 Dec;43(6) :1969–1985. 89

[249] VanderWeele TJ, Ding P, Mathur M. Technical Considerations in the Use of the E-Value. Journal of Causal Inference. 2019 Sep;7(2). 89

[250] VanderWeele TJ, Mathur MB, Ding P. Correcting Misinterpretations of the E-Value. Annals of Internal Medicine. 2019 Jan;170(2) :131. 89

[251] FDA. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products : Guidance for Industry. FDA-2019-D-0934 ; 2021. 89

[252] EMA. Guideline on adjustment for baseline covariates in clinical trials. EMA/CHMP/295050/2013 ; 2015. 89

[253] Benkeser D, Díaz I, Luedtke A, Segal J, Scharfstein D, Rosenblum M. Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. Biometrics. 2020 ;n/a. 89

[254] Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials : an assessment of 12 outcomes from 8 studies. Trials. 2014 Apr ;15(1) :139. 89

[255] Kahan BC, Morris TP. Adjusting for multiple prognostic factors in the analysis of randomised trials. BMC Medical Research Methodology. 2013 Jul ;13(1) :99. 89

[256] Thompson DD, Lingsma HF, Whiteley WN, Murray GD, Steyerberg EW. Covariate adjustment had similar benefits in small and large randomized controlled trials. Journal of Clinical Epidemiology. 2015 Sep ;68(9) :1068–1075. 89

[257] Remiro-Azócar A, Heath A, Baio G. Conflating marginal and conditional treatment effects : Comments on "Assessing the performance of population adjustment methods for anchored indirect comparisons : A simulation study". Statistics in Medicine. 2021 ;40(11) :2753–2758. 89

[258] Ge M, Durham LK, Meyer RD, Xie W, Thomas N. Covariate-Adjusted Difference in Proportions from Clinical Trials Using Logistic Regression and Weighted Risk Differences. Drug Information Journal. 2011 Jul ;45(4) :481–493. 89

[259] Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes : Targeted maximum likelihood estimation. Statistics in medicine. 2009 Jan ;28(1) :39–64. 89

[260] Wager S, Du W, Taylor J, Tibshirani RJ. High-dimensional regression adjustments in randomized experiments. Proceedings of the National Academy of Sciences. 2016 Nov ;113(45) :12673–12678. 89

[261] Golmakani MK, Polley EC. Super Learner for Survival Data Prediction. The International Journal of Biostatistics. 2020 Feb ;16(2). 89

[262] Tanner KT, Sharples LD, Daniel RM, Keogh RH. Dynamic survival prediction combining landmarking with a machine learning ensemble : Methodology and empirical comparison. Journal of the Royal Statistical Society : Series A (Statistics in Society). 2021 Jan ;184(1) :3–30. 89

[263] Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. The Annals of Statistics. 2013 Apr ;41(2) :802–837. 90

[264] Efron B. Estimation and Accuracy after Model Selection. Journal of the American Statistical Association. 2014 Jul ;109(507) :991–1007. 90

[265] Ahern J. Start With the "C-Word," Follow the Roadmap for Causal Inference. American Journal of Public Health. 2018 May ;108(5) :621–621. 90

[266] Didelez V. Commentary : Should the analysis of observational data always be preceded by specifying a target experimental trial? International Journal of Epidemiology. 2016 Dec ;45(6) :2049–2051. 90

[267] García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence : an application to colorectal cancer screening. European Journal of Epidemiology. 2017 Jun ;32(6) :495–500. 90

[268] Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of clinical epidemiology. 2016 Nov ;79 :70–75. 90

[269] Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. European Journal of Epidemiology. 2020 Jul ;35(7) :615–617. 90

Annexe A

Développement d'un module prédictif dans Plug-Stat®

A.1 Brève introduction à la prédiction

Dans le domaine médical, la modélisation statistique sert souvent deux objectifs : expliquer ou prédire. De nombreux auteurs ont décrit les différences liées à ces approches [270–276]. En quelques mots, si la modélisation causale a pour but d'étudier et de quantifier d'éventuels liens causaux entre différentes variables, la modélisation prédictive réfère à l'utilisation des données disponibles afin de faire correspondre plusieurs variables entre elles. Toutes les variables sont ainsi sélectionnées dès lors qu'elles participent à améliorer les capacités prédictives, ce qui facilite l'automatisation.

Le développement du module d'analyse prédictive comprend le développement de deux sous-modules emboîtés. Le premier évalue les capacités prédictives d'une variable quand le second permet la construction du score prédictif. L'utilisateur peut utiliser le premier sous-module seul ou il peut se servir du second qui fera appel au premier pour évaluer les capacités prédictives du score nouvellement construit.

A.2 Évaluation des capacités prédictives d'une variable

La Figure A.1 illustre l'interface principale de Plug-Stat®. L'implémentation du nouveau module nécessite de proposer à l'utilisateur un nouveau type d'analyse (Figure A.1, point G). Notons que, pour les analyses prédictives, il n'est pas nécessaire de définir des groupes de traitement (Figure A.1, point E). Une fois cette analyse sélectionnée, l'utilisateur s'engagera dans une succession d'étapes pré-définies.

La première étape consiste à choisir la variable d'intérêt. Si la variable d'intérêt est qualitative, l'utilisateur définit les modalités d'intérêt et de référence. Ensuite, l'utilisateur se voit proposer les différents évènements à prédire d'après la liste des critères de jugement définis à l'installation de Plug-Stat®. Pour une variable quantitative, la troisième étape consiste à définir le seuil à partir duquel seront calculées les mesures prédictives suivantes : sensibilité, spécificité, valeurs prédictives positive et négative et aire sous la courbe ROC. La dichotomisation peut se faire automatiquement selon l'indice de Youden ou l'indice *upper-left* [277]. L'utilisateur peut également choisir un seuil différent manuellement. Afin de faciliter cette étape dépendante de l'utilisateur, nous avons créé une application RShiny interactive où l'utilisateur peut directement voir l'effet de ses choix (Figure A.2). En présence d'un évènement censuré à droite, l'utilisateur se voit également proposer cette interface avec un choix supplémentaire : le temps auquel la prédiction doit être faite. Les capacités prédictives sont ensuite automatiquement calculées à partir du tableau de contingence et

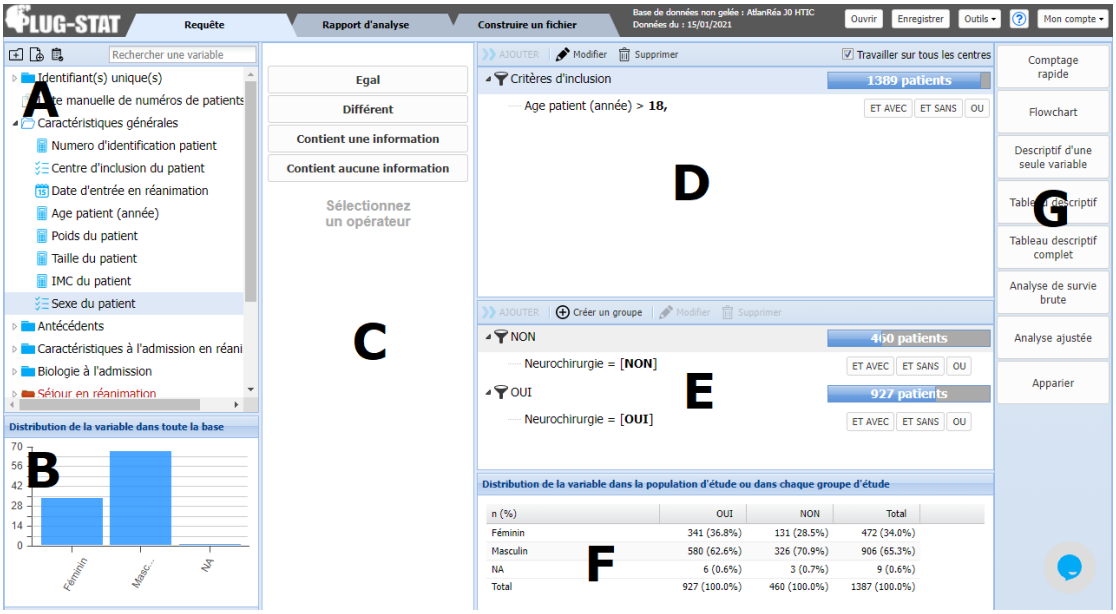


FIGURE A.1 – Interface principale de Plug-Stat®. Elle comprend des modules présentant les variables disponibles (A), des statistiques descriptives dynamiques (B et F), d'opérateurs (C), de définition des critères d'inclusion (D), de définition des groupes d'étude (E) et d'analyses (G)

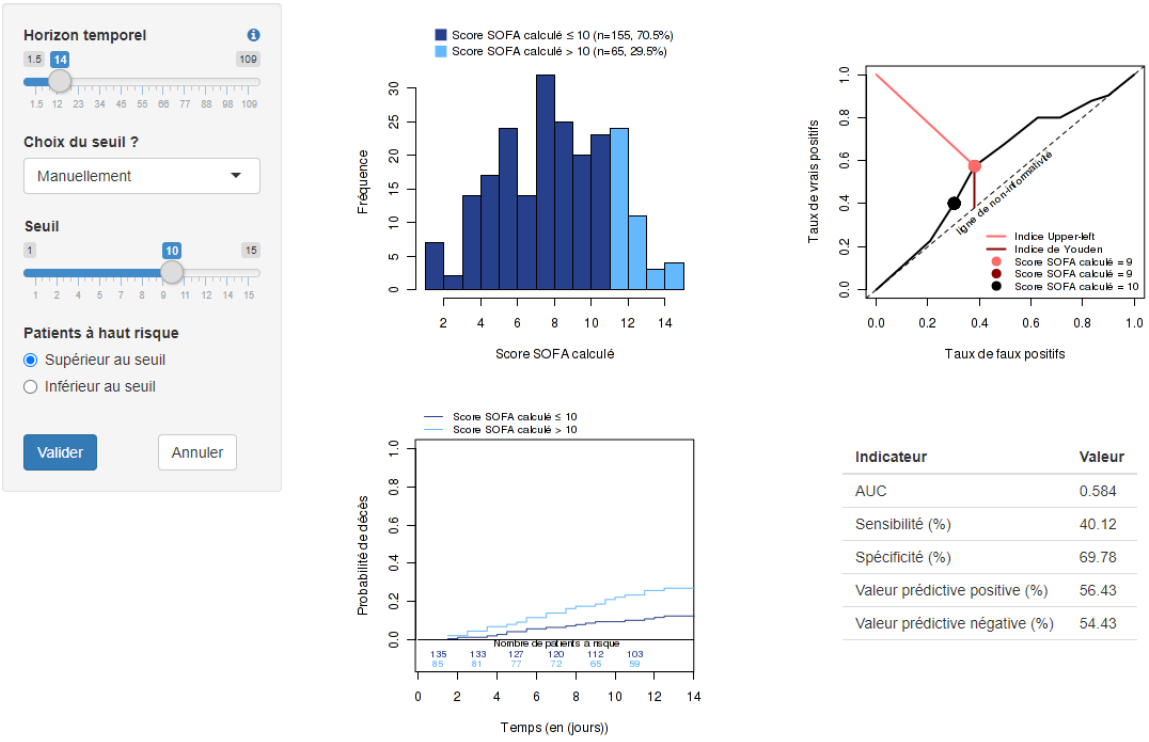


FIGURE A.2 – Application RShiny permettant à l'utilisateur de choisir le seuil prédictif et l'horizon temporel

de l'éventuelle courbe ROC. Notons que la censure à droite est prise en compte par pondération inverse [278]. Les intervalles de confiance de ces différentes mesures sont obtenus par bootstrap avec un nombre d'itérations défini par l'utilisateur (200, 500 ou 1000). À titre illustratif, les courbes de survies brutes et stratifiées sur la variable d'intérêt, la distribution de la variable d'intérêt et la courbe ROC sont fournies selon les scénarios. Notons que ces différentes figures sont personnalisables via des interfaces RShiny spécifiques. La dernière section de cette annexe présentera un exemple de rapport finalement fourni à l'utilisateur.

A.3 Construction et évaluation d'un score prédictif

Similairement au sous-module précédent, l'utilisateur pourra accéder à cette fonctionnalité via un bouton spécifique (Figure A.1, point G). Une fois cette analyse sélectionnée, l'utilisateur s'engagera dans une succession d'étapes pré-définies.

La première étape consiste à choisir le critère de jugement. Ensuite l'utilisateur pourra choisir les variables à inclure dans son score. L'utilisateur définira deux ensembles, les variables à tester (*i.e.*, une procédure de sélection sera appliquée) et les variables à forcer. Dès lors qu'au moins une variable est incluse dans l'ensemble à tester, une procédure de sélection sera effectuée par une pénalisation lasso [279, 280]. Le paramètre de pénalisation sera choisi par défaut par validation croisée (10 partitions) [281]. L'utilisateur pourra cependant changer la valeur de ce paramètre afin de modifier la complexité de son score. Une application RShiny a été développée afin d'aider l'utilisateur en montrant de manière interactive l'impact de la variation de la pénalisation (Figure A.3).

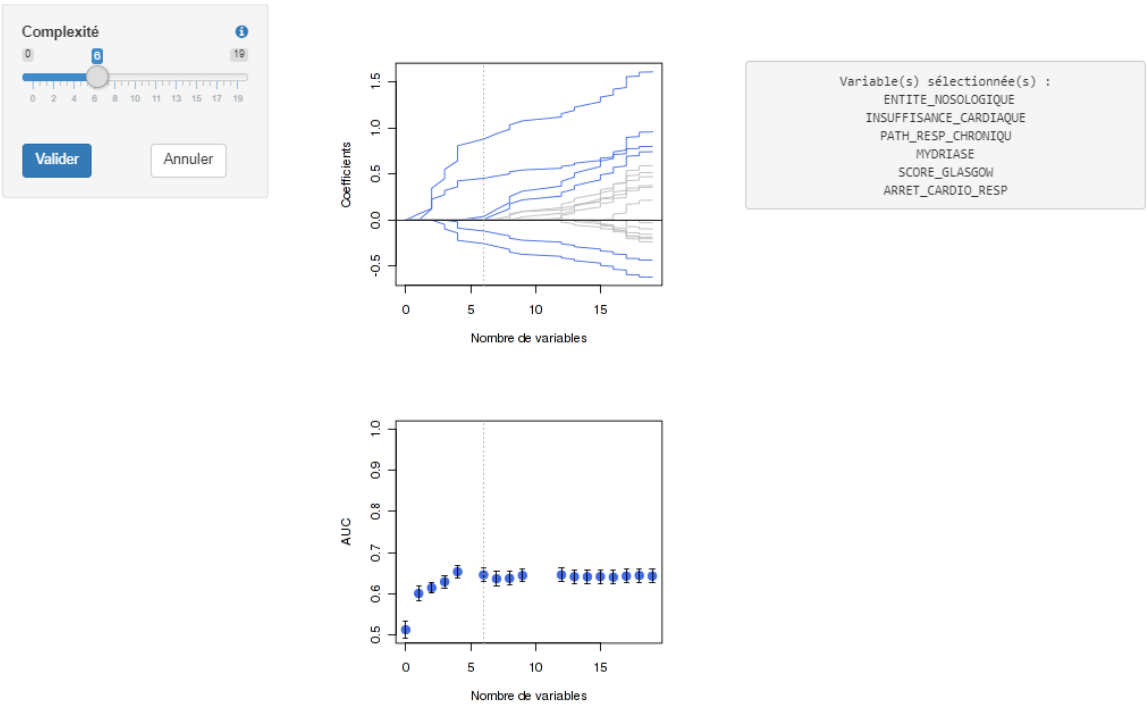


FIGURE A.3 – Application RShiny permettant à l'utilisateur de définir la complexité de son score prédictif

Si toutes les variables choisies par l'utilisateur sont forcées dans la modélisation, une régression logistique multiple ou un modèle de Cox multiple sera appliqué. Ensuite, l'utilisateur retrouvera la première application RShiny (Figure A.2) avec le score comme variable d'intérêt. Le rapport final est similaire à la section précédente, la principale différence étant que les estimations sont obtenues au moyen d'une validation-croisée par boostrap [282] afin d'éviter le sur-apprentissage. Les résultats sont présentés pour les échantillons apparant et de validation interne. La calibration du

score est vérifiée au moyen d'une courbe de calibration personnalisable [283]. Le score de Brier, intégré sur le temps en présence de censure à droite [284], et son intervalle de confiance sont également rapportés. Notons que le rapport est actuellement en phase de finalisation, sa forme définitive sera similaire à celle du rapport présenté dans la section suivante. Afin de limiter le temps de calcul, l'éventuel paramètre de pénalisation n'est pas réestimé à chaque itération. Foucher et Danger montraient que ce raccourci computationnel n'était pas associé à un sur-ajustement [285].

Bibliographie spécifique

[270] Breiman L. Statistical Modeling : The Two Cultures (with comments and a rejoinder by the author). Statistical Science. 2001 Aug;16(3) :199–231. I

[271] Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right : A Classification of Data Science Tasks. Chance. 2019 Jan;32(1) :42–49. I

[272] Shmueli G. To Explain or to Predict? Statistical Science. 2010 Aug;25(3) :289–310. I

[273] Arnold KF, Davies V, de Kamps M, Tennant PWG, Mbotwa J, Gilthorpe MS. Reflections on modern methods : generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. International Journal of Epidemiology. 2020 Dec;49(6) :2074-82. I

[274] Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence. 2020 Jul;2(7) :369–375. I

[275] Rose S. Intersections of machine learning and epidemiological methods for health services research. International Journal of Epidemiology. 2020 Dec;49(6) :1763–1770. I

[276] Efron B. Prediction, Estimation, and Attribution. Journal of the American Statistical Association. 2020 Apr;115(530) :636–655. I

[277] Perkins NJ, Schisterman EF. The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. American Journal of Epidemiology. 2006 Apr;163(7) :670–675. I

[278] Le Borgne F, Combescure C, Gillaizeau F, Giral M, Chapal M, Giraudeau B, et al. Standardized and weighted time-dependent receiver operating characteristic curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors. Statistical Methods in Medical Research. 2018;27(11) :3397–3410. III

[279] Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010 Feb;33 :1–22. III

[280] Simon N, Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software. 2011 Mar;39 :1–13. III

[281] de Rooij M, Weeda W. Cross-Validation : A Method Every Psychologist Should Know. Advances in Methods and Practices in Psychological Science. 2020 May :2515245919898466. III

[282] Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics. 2005 May;21(9) :1979–1986. III

[283] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration : the Achilles heel of predictive analytics. BMC Medicine. 2019 Dec;17(1) :230. IV

[284] Kvamme H, Borgan Ø, Scheel I. Time-to-Event Prediction with Neural Networks and Cox Regression. Journal of Machine Learning Research. 2019;20(129) :1–30. IV

[285] Foucher Y, Danger R. Time dependent ROC curves for the estimation of true prognostic capacity of microarray data. Statistical Applications in Genetics and Molecular Biology. 2012 Nov;11(6) :1. IV

A.4 Exemple de rapport d'analyse prédictive fourni par Plug-Stat®

Rapport statistique

5 avril 2019

Population étudiée. Les données sont issues de la cohorte française AtlanREA (www.atlanrea.org) constituée de patients traumatisés graves ou cérébrolésés hospitalisés. Pour cette étude seuls les patients avec une hypertension intracranienne ont été inclus. Les patients sont suivis de leur admission à leur sortie ou leur décès durant le séjour en réanimation. Une lettre d'information a été donnée après une explication orale. Tous les patients ont donné leur consentement éclairé. Seuls les patients intubés ont été considérés. Les patients qui étaient incapables de répondre aux questions ont été exclus.

Analyses statistiques. Le J0 a été défini comme le moment où une hypertension intracranienne est détectée (définie par le premier événement parmi une pression intracranienne supérieure à 20 cm d'eau ou le début d'un des traitements suivants : barbituriques, osmothérapie, craniectomie ou hypothermie). Les capacités de la variables d'intérêt à prédire la mortalité en réanimation ont été étudiées graphiquement par la courbe caractéristique de performances (ROC) dépendante du temps à un horizon temporel fixé à 14 jours. La censure à droite a été prise en compte par pondération inverse [1]. De potentiels seuils de discrimination entre les patients à haut et bas risques ont été obtenu par l'indice de Youden et l'indice upper-left. Le seuil de discrimination a finalement été fixé à 10 arbitrairement. Les caractéristiques des deux groupes à J0 ont été comparées avec un test du Chi2 ou un test exact de Fisher pour les variables catégorielles et par un test de Student pour les variables continues. Les courbes d'incidence cumulées brutes et stratifiées ont été obtenues par l'estimateur de Aalen-Johansen [2] et comparées par le test de Gray [3]. Les indicateurs dépendants du temps suivants sont également proposés en considérant à risque les patients supérieurs à 10 : sensibilité, spécificité, valeurs prédictives positive et négative, ainsi que les rapports de vraisemblance positif et négatifs [4]. Les intervalles de confiance à 95% (IC95%) ont été obtenus par bootstrap (200 itérations) [5]. Les analyses statistiques ont été réalisées avec le logiciel Plug-Stat® (www.labcom-risca.com) basé sur le logiciel R [6].

Résultats. Parmi les 293 patients inclus, 68 sont exclus pour cause de données manquantes sur la variable d'intérêt et 7 sont exclus pour cause de données manquantes sur la mortalité en réanimation dont 2 avec également la variable d'intérêt manquante. La courbe ROC et la distribution de la variable d'intérêt sont illustrées par les Figures 1 et 2, respectivement. Des seuils de 9 et 9 ont été obtenus en maximisant les indices de Youden (J=0.19) et Upper-left (UL=0.57), respectivement. Les caractéristiques initiales de l'échantillon et des groupes sont décrites dans la Table 1. On dénombre respectivement 65 patients à haut risque (29.5%) et 155 patients à bas risque (70.5%). La prévalence initiale de la mortalité en réanimation est de 26.8% dans l'échantillon et de 40.0% et 21.3% dans les groupes à haut et bas risque, respectivement. Les courbes d'incidence cumulée stratifiée par groupe et brute sont respectivement illustrées dans les Figures 3 et 4. La Table 2 présente les capacités prédictives de la variable d'intérêt. L'aire sous la courbe ROC (AUC) à 14 jours est de 0.584 (IC95% 0.499; 0.677). Le temps médian de survie est de 25 jours (min=2, max=109).

Références.

1. Le Borgne F, Combescure C, Gillaizeau F, et al. Standardized and weighted time-dependent receiver operating characteristic curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors. Stat Methods Med Res. 2018.
2. Aalen O, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics. 1978.
3. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. The annals of Statistics. 1988.

4. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. Biometrics. 2000.
5. Carpenter J, Bithell J. Boostrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine. 2000.
6. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017. URL https://www.R-project.org/.

Figure 1 : Courbe ROC à 14 jours

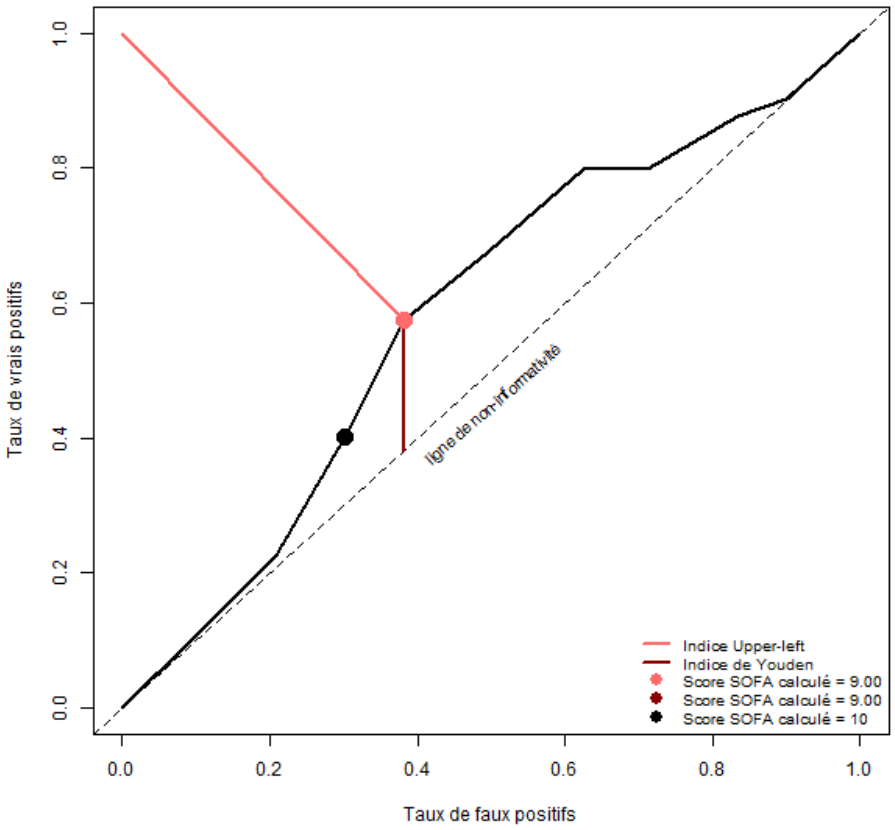


Figure 2 : Distribution de la variable Score SOFA calculé

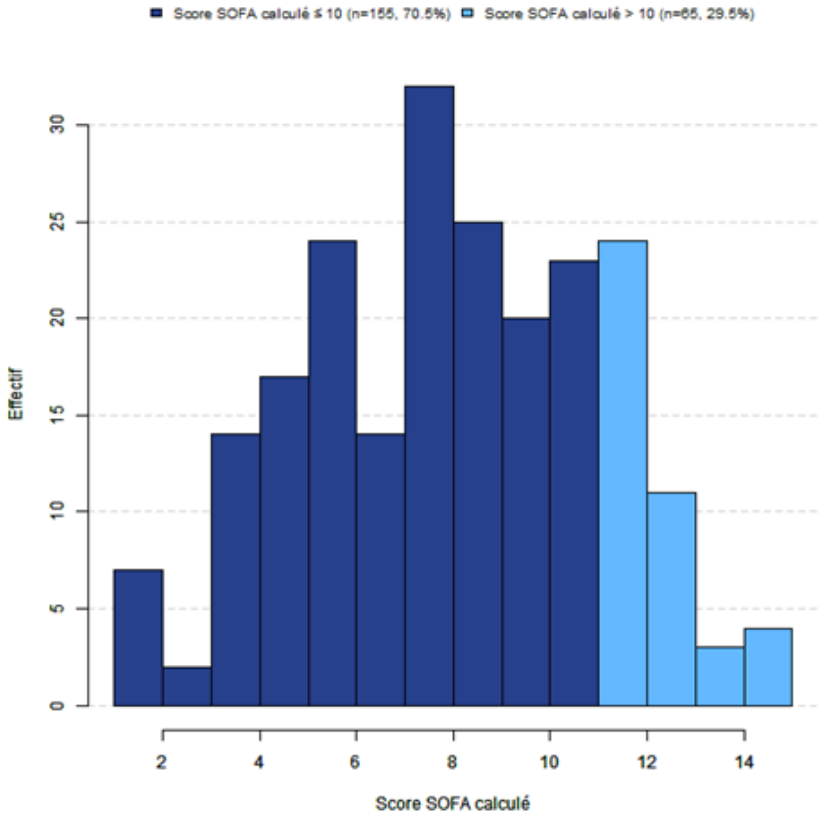


Figure 3 : Courbes de l'incidence cumulée du décès estimées par l'estimateur de Aalen-Johansen

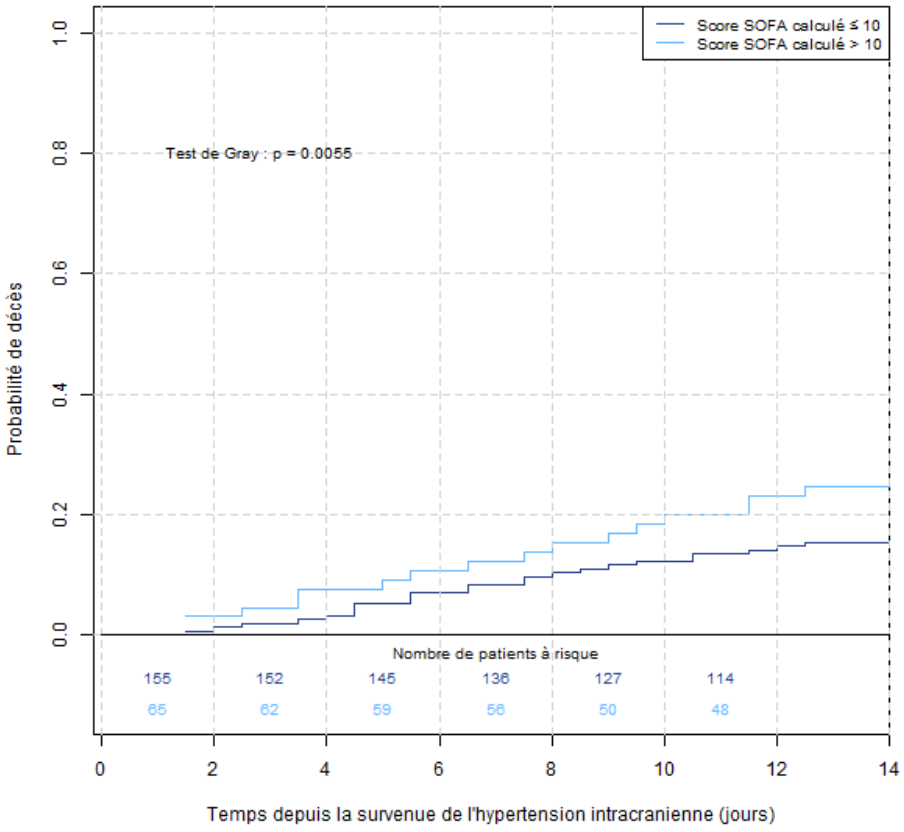
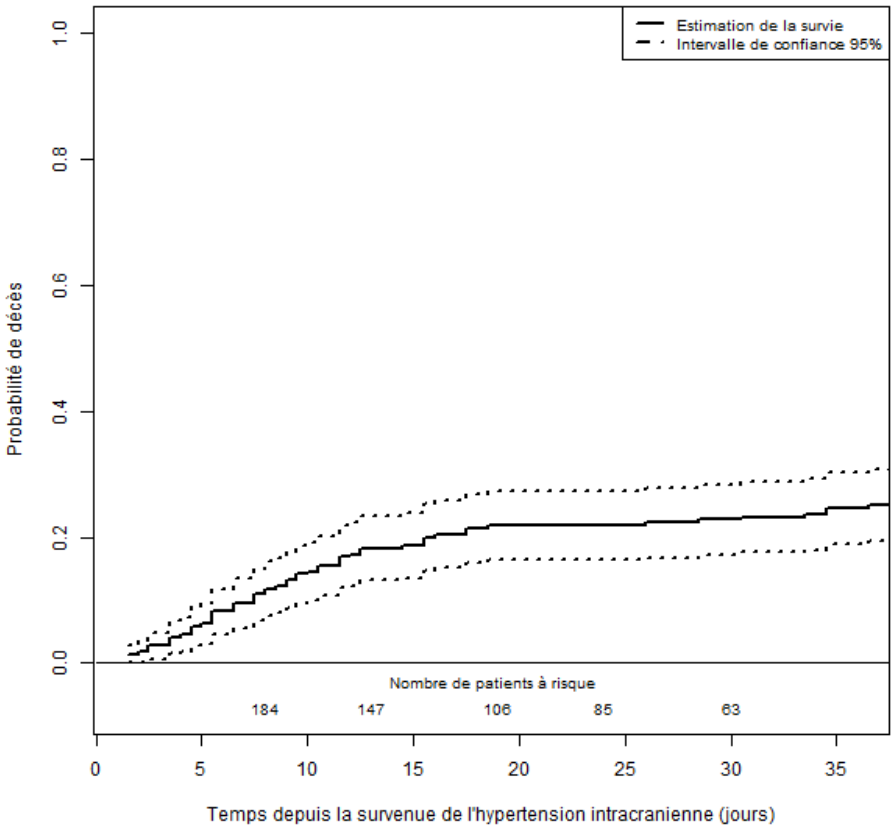


Figure 4 : Courbes de l'incidence cumulée du décès estimées par l'estimateur de Aalen-Johansen



Rapport statistique

5 avril 2019

Tableau 1 : Descriptif à l'origine selon les groupes d'étude (les p-values sont obtenues avec un test du Chi2 ou un test exact de Fisher (‡) pour les variables catégorielles et par un test de Student pour les variables continues).

	Global (n=220)			Inférieur ou égal au seuil (n=155)			Supérieur au seuil (n=65)			p-value
	NA	n	%	NA	n	%	NA	n	%	
Entité nosologique à l'admission en réanimation	0			0			0			0.6335‡
Cérébrolésés		5	2.3		3	1.9		2	3.1	
Traumatisés graves		215	97.7		152	98.1		63	96.9	
Patient masculin	0	180	81.8	0	127	81.9	0	53	81.5	0.9445
Antécédent d'insuffisance rénale	1	3	1.4	1	2	1.3	0	1	1.5	1.0000‡
Pathologie respiratoire chronique	4	6	2.8	3	3	2.0	1	3	4.7	0.3645‡
Antécédent de diabète	1	14	6.4	1	12	7.8	0	2	3.1	0.2402‡
Alcoolisme chronique	12	43	20.7	8	29	19.7	4	14	23.0	0.6013
Antécédent de cancer	4	3	1.4	2	1	0.7	2	2	3.2	0.2044‡
Tabagisme actif	27	53	27.5	16	38	27.3	11	15	27.8	0.9510
Mydriase à l'admission en réanimation	3	73	33.6	2	44	28.8	1	29	45.3	0.0186
Score de Glasgow ≥ 9	0	43	19.5	0	41	26.5	0	2	3.1	0.0001
PAS ≤ 90 mmHg avant l'admission en réanimation	5	72	33.5	2	42	27.5	3	30	48.4	0.0032
Arrêt cardio respiratoire avant l'admission en réanimation	0	8	3.6	0	3	1.9	0	5	7.7	0.0512‡
Evacuation d'un hématome avant l'admission en réanimation	0	44	20.0	0	32	20.6	0	12	18.5	0.7118
Pose d'une dérivation ventriculaire externe avant l'admission en réanimation	0	9	4.1	0	8	5.2	0	1	1.5	0.2873‡
Lobectomie avant l'admission en réanimation	0	9	4.1	0	7	4.5	0	2	3.1	1.0000‡
Craniectomie de décompression avant l'admission en réanimation	0	24	10.9	0	18	11.6	0	6	9.2	0.6051

Rapport statistique

5 avril 2019

	Global (n=220)			Inférieur ou égal au seuil (n=155)			Supérieur au seuil (n=65)			p-value
	NA	n	%	NA	n	%	NA	n	%	
Transfusion avant l'admission en réanimation	2	68	31.2	1	39	25.3	1	29	45.3	0.0037
Pneumopathie pré-HTIC	0	9	4.1	0	6	3.9	0	3	4.6	0.7257‡
	NA	**m**	**e-t**	**NA**	**m**	**e-t**	**NA**	**m**	**e-t**	
Age (en années)	0	39.7	19.2	0	39.8	19.7	0	39.6	18.2	0.9686
IMC (en kg.m-2)	12	25.0	4.9	4	24.8	4.7	8	25.4	5.3	0.4856
Pression intracrânienne à l'admission en réanimation (en cm d'eau)	27	22.0	14.7	19	21.3	13.4	8	23.8	17.4	0.3243
Hémoglobinémie (en g/dL)	0	11.0	2.3	0	11.1	2.2	0	10.8	2.5	0.3174
Leucocytes (en g/L)	0	17.4	6.6	0	17.6	6.5	0	16.8	6.9	0.4339
Prothrombine (en %)	1	71.3	18.8	1	73.3	18.8	0	66.7	18.0	0.0166
Plaquettes (en g/L)	0	177.7	71.6	0	188.6	71.4	0	151.9	65.6	0.0003
Fibrinogénémie (en g/L)	11	2.3	1.2	8	2.4	1.1	3	2.1	1.2	0.0781
Lactatémie (en mg/L)	21	2.4	1.4	19	2.2	1.2	2	2.8	1.8	0.0118
pH plasmatique	3	7.3	0.1	3	7.3	0.1	0	7.3	0.1	<0.0001
Bicarbonates plasmatiques (en mmol/L)	3	21.3	3.5	3	21.7	3.6	0	20.3	3.2	0.0055
PaO2 (en mmHg)	9	114.9	49.7	6	122.4	48.8	3	96.9	47.7	0.0006
FIO2	6	0.4	0.2	6	0.4	0.1	0	0.5	0.2	0.0003
Rapport PaO2/FiO2 à l'admission en réanimation	12	301.2	156.2	9	328.2	138.1	3	237.6	177.8	0.0005
Créatininémie (en mmol/L)	0	84.1	38.1	0	77.6	23.6	0	99.6	57.3	0.0038
Protidémie (en g/L)	9	56.2	10.7	6	57.7	10.5	3	52.4	10.3	0.0008
Glycémie (en g/L)	26	7.9	2.5	17	8.0	2.3	9	7.8	2.9	0.6687
Urémie (en g/L)	4	5.2	2.1	3	5.1	2.2	1	5.5	2.0	0.1830
Calcémie (en mmol/L)	8	2.0	0.2	6	2.0	0.2	2	1.9	0.2	0.0012
Score IGS II	7	45.8	12.3	6	42.4	10.7	1	53.7	12.1	<0.0001

FIO2, fraction inspirée d'oygène ; IGS, indice gravité simplifié ; IMC, indice de masse corporelle ; HTIC, hypertension intracrânienne ; ISS, injury severity score ; PaO2, Pression partielle en oxygène dans le sang artériel. m, moyenne; NA, not available (manquant); e-t, écart-type.

Tableau 2 : Capacités prédictives (n=220)

Indicateur	Valeur	95% CI
AUC	0.584	[0.499 ; 0.677]
Sensibilité (%)	40.12	[26.96 ; 53.14]
Spécificité (%)	69.78	[62.12 ; 76.76]
Valeur prédictive positive (%)	56.43	[41.33 ; 68.73]
Valeur prédictive négative (%)	54.43	[40.66 ; 66.18]
Rapport de vraisemblance positif	1.33	[0.87 ; 1.91]
Rapport de vraisemblance négatif	0.86	[0.68 ; 1.07]

IC, Intervalle de confiance.

Annexe B

Mesures de performance pour les études de simulation

TABLEAU B.1 – Résumé des principales mesures de performance des études de simulation, adapté de Morris, White et Crowther (2019)

Mesure de performances	Définition	Estimation	Erreur de Monte-Carlo
Biais absolu	$\mathrm{E}[\hat{\theta}] - \theta$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i - \theta$	$\sqrt{\frac{1}{n_{\text{sim}}(n_{\text{sin}^{-1}})} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \bar{\theta})^2}$
Erreur standard empirique (eSE)	$\sqrt{\mathrm{Var}(\hat{\theta})}$	$\sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \bar{\theta})^2}$	$\frac{\widehat{\text{eSE}}}{\sqrt{2(n_{\text{sim}}-1)}}$
Erreur standard asymptotique (aSE)[a]	$\sqrt{\mathrm{E}[\widehat{\mathrm{Var}}(\hat{\theta})]}$	$\sqrt{\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \widehat{\mathrm{Var}}(\hat{\theta}_i)]}$	$\sqrt{\frac{\widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}(\hat{\theta})]}{4n_{\text{sim}} \times \widehat{a\text{SE}}^2}}$
Biais d'estimation de la variance (%)[a]	$100\left(\frac{a\text{SE}}{e\text{SE}} - 1\right)$	$100\left(\frac{\widehat{a\text{SE}}}{\widehat{e\text{SE}}} - 1\right)$	$100\left(\frac{\widehat{a\text{SE}}}{\widehat{e\text{SE}}}\right) \sqrt{\frac{\widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}(\hat{\theta})]}{4n_{\text{sim}} \times \widehat{a\text{SE}}^4} + \frac{1}{2(n_{\text{sim}}-1}}$
Erreur quadratique moyenne (MSE)	$\mathrm{E}[(\hat{\theta} - \theta)^2]$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2$	$\sqrt{\frac{\sum_{i=1}^{n_{\text{sim}}} [(\hat{\theta}_i - \theta)^2 - \widehat{\text{MSE}}]^2}{n_{\text{sim}}(n_{\text{sin}^{-1}})}}$
Couverture (%)	$\mathrm{P}(\hat{\theta}_{\text{inf}} \leq \theta \leq \hat{\theta}_{\text{sup}})$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \mathbf{1}(\hat{\theta}_{\text{inf},i} \leq \theta \leq \hat{\theta}_{\text{sup},i})$	$\sqrt{\frac{\widehat{\text{Couv.}}(1-\widehat{\text{Couv.}})}{n_{\text{sim}}}}$
Rejet (%)[b]	$\mathrm{P}(p_{value} \leq \alpha)$	$\frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(p_{value,i} \leq \alpha)$	$\sqrt{\frac{\widehat{\text{Rejet}}(1-\widehat{\text{Rejet}})}{n_{\text{sim}}}}$

[a] $\widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}(\hat{\theta})] \approx \frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} \left\{ \widehat{\mathrm{Var}}(\hat{\theta}_i) - \frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \widehat{\mathrm{Var}}\left(\hat{\theta}_j\right) \right\}^2$

[b] Correspond respectivement à l'erreur de type I et à la puissance statistique sous les hypothèses nulle et alternative.

Annexe C

Étude de simulation de Léger *et al.* comparant différentes approches d'estimation causale en cas de violation aléatoire de positivité

Biometrical Journal **XX** (2020) XX, zzz–zzz / DOI: 10.1002/bimj.200100000

Causal inference in case of near-violation of positivity: comparison of methods

Maxime Léger [1,2], **Arthur Chatton** [1,3], **Florent Le Borgne** [1,3] , **Romain Pirrachio** [4], **Sigismond Lasocki** [2], and **Yohann Foucher*** [1,5]

[1] INSERM UMR 1246 - SPHERE, Université de Nantes, Université de Tours, Nantes, France

[2] Département d'Anesthésie-Réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France

[3] IDBC-A2COM, Nantes, France

[4] Department of Anesthesia and Perioperative Care, University of California, San Francisco, USA

[5] Centre Hospitalier Universitaire de Nantes, Nantes, France

Received zzz, revised zzz, accepted zzz

In causal studies, the near-violation of the positivity may occur by chance, because of sample-to-sample fluctuation despite the theoretical veracity of the positivity assumption in the population. It may mostly happen when the exposure prevalence is low or when the sample size is small. We aimed to compare the robustness of g-computation (GC), inverse probability weighting (IPW), truncated IPW, targeted maximum likelihood estimation (TMLE) and truncated TMLE in this situation, using simulations and one real application. We also tested different extrapolation situations for the sub-group with a positivity violation. The results illustrated that the near-violation of the positivity impacted all methods. We demonstrated the robustness of GC and TMLE-based methods. Truncation helped in limiting the bias in near-violation situations, but at the cost of bias in normal conditions. The application illustrated the variability of the results between the methods and the importance of choosing the most appropriate one. In conclusion, compared to propensity score-based methods, methods based on outcome regression should be preferred when suspecting near-violation of the positivity assumption.

Key words: Causal Inference; Doubly Robust Estimators; G-Computation; Positivity; Propensity Score; Real-world Evidence; Simulations

1 Introduction

There is growing interest in causal methods (Hernán and Robins, 2020), notably the propensity score (PS)-based methods (Austin, 2011; Williamson et al., 2012). The PS is related to the exposure prediction. One can distinguish four different approaches: matching, stratification, conditional adjustment, and inverse probability weighting (IPW) (Rosenbaum and Rubin, 1983; Robins et al., 2000). IPW and matching on PS estimate marginal effects, while stratification and conditioning estimate conditional effects. In the settings of non-linear link functions, marginal and conditional estimates may differ due to the non-collapsibility issues. IPW and matching emerge as preferable methods for estimating marginal effects (Austin, 2013). However both IPW and matching suffer efficiency limitations: IPW due to extreme weights and matching due to non-matching subjects resulting in loss of information. Despite the problems of the most extreme subjects, IPW emerges as a preferable options in terms of both bias and precision (Lendle et al., 2013; Le Borgne et al., 2016; Hajage et al., 2016; Abdia et al., 2017). An alternative is g-computation (GC), also known as parametric g-formula or (g-)standardisation (Robins, 1986; Vansteelandt and Keiding, 2011; Snowden et al., 2011). The latter estimator relies on an outcome model rather than an exposure model like for PS-based methods. Some estimators combine GC and PS to create doubly robust estimators (DREs)

*Corresponding author: e-mail: Yohann.Foucher@univ-nantes.fr

aiming to minimize the impact of model misspecification on consistency (Bang and Robins, 2005; Neugebauer and van der Laan, 2005). One of the most studied doubly-robust methods is the targeted maximum likelihood estimation (TMLE) (van der Laan and Rubin, 2006).

1.1 Positivity violation

Regardless of the type of estimator, in order to conclude causally, one has to make several assumptions: consistency, conditional exchangeability, and positivity. Positivity is met if, for any combination of the covariates, there is a non-null probability of being exposed or unexposed. Positivity violations can occur in two situations: i) theoretical violation: we know that there are patients with a null probability of being exposed or unexposed, e.g., if certain patients present a contraindication to receiving a treatment of interest; ii) near or practical violation, sampling variability may result in subjects having a null probability of being exposed or unexposed for certain combinations of covariate values. This may be particularly frequent for cases of low exposure prevalence or small sample sizes (Westreich and Cole, 2010).

The theoretical violation is a consequence of a conceptual problem in the study design and calls for restricting the studied population (Westreich and Cole, 2010; Petersen et al., 2012), i.e., excluding patients with a theoretical null probability of being exposed or unexposed (for instance, patients with a counter-indication for one of the studied treatments). In contrast, in case of near-violation, the target population is well defined. In this situation, the goal is to select an estimator that doesn't suffer from the near-violation. For IPW, one can empirically set threshold values for truncating (Cole and Hernán, 2008) or trimming the PS (Crump et al., 2009). These approaches aim to limit the maximum contribution of extreme observations. Truncation has the advantage of preserving clinical equipoise in the target population, whereas excluding certain subjects would result in a trimmed population that would change the estimand.

1.2 Extrapolation issue

For the methods based on the outcome regression, the problem is an extrapolation of the outcome prediction for the patients affected by the near-violation, rather than using actual observations in the data (van der Laan, 2003; Neugebauer and van der Laan, 2006).

Let (Y, A, Z) denote the binary outcome ($Y = 1$ for events and 0 otherwise), the binary exposure ($A = 1$ for exposed individuals and 0 otherwise), and the p baseline covariates $(Z_1, ..., Z_p)$. Let define $f(Z_1|A)$, the density function of the quantitative covariate Z_1 conditional to A, Z_1 being a true confounder which causes both the exposure status A and the outcome Y. As illustrated in Figure 1, consider a near-violation of the positivity for $Z_1 > \alpha$ and an effect of a theoretical increase in the conditional probability of the outcome under $A = 1$ for larger values of Z_1.

Because of the lack of information when $Z_1 > \alpha$ due to the near-violation of the positivity assumption, the estimation of the exposure effect relies on extrapolating the observed effect, i.e., when $Z_1 \leq \alpha$. Even when the outcome model is adequately specified in the region supported by data, the model may be inadequate for the region suffering from positivity near-violation.

The causal inferences will depend on the formulation of non-testable hypotheses.

One can note that this illustration (with Z_1 as a quantitative confounder can be extended for a binary confounder. Consider that Z_1 represents the gender. If there is no information regarding the outcomes among exposed women, one cannot properly infer the average exposure effect in the target population.

1.3 Framework

The literature does not provide a clear answer as to the most reliable method in cases of positivity near-violation. Indeed, even though several studies have compared the previous methods in the context of positivity violation (Lendle et al., 2013; Petersen et al., 2012; Moore et al., 2012), suggesting better stability and reduced bias for GC and DRE, they did not investigate the extrapolation issue.

In the situation of positivity near-violation, Petersen et al. (2012) introduced the problem of extrapolation. Nevertheless, they did not study its impact.

In this context, we performed a simulation-based study to evaluate the robustness of IPW, truncated IPW, GC, TMLE, and truncated TMLE in the situation of the extrapolation issue and positivity near-violation. We also evaluated one application from a real dataset. This study is structured as follows: section 2 outlines the methods used, section 3 presents the design and the results of the simulation study, in section 4, we apply the developed application to a real dataset, and finally we discuss the results and provide recommendations.

2 Methods

2.1 Setting and notations

Consider a resulting sample of size n in which one can observe the realizations of these random variables (y, a, z). Define $\pi_a = P(Y = 1|do(A = a))$ as the expected proportions of event if the entire population is exposed ($do(A = 1)$) or unexposed ($do(A = 0)$) (Pearl et al., 2016). The average exposure effect on the entire population is defined as $\Delta = \pi_1 - \pi_0$. The corresponding marginal causal odds ratio is expressed as $\text{OR} = (\pi_1/(1 - \pi_1))/(\pi_0/(1 - \pi_0))$.

2.2 Inverse Probability Weighting (IPW)

Formally, the PS for a subject i ($i = 1, ..., n$) is $p_i = P(A = 1|z_i)$, *i.e.*, the probability that a subject is exposed according to her/his observed characteristics z_i (Rosenbaum and Rubin, 1983). The PS is often estimated from logistic regression, but other models or algorithms can be used such as random forest, boosting, or super learner (Austin, 2012; Pirracchio and Carone, 2018). IPW results in weighting the contribution of each subject i by $\omega_i = A_i P(A_i = 1)/p_i + (1 - A_i)P(A_i = 0)/(1 - p_i)$, where $P(A_i = 1)$ and $P(A_i = 0)$ denote the marginal probability of exposure and its complementary. The use of such stabilized weights are preferred to optimize the variance estimation (Robins et al., 2000; Xu et al., 2010). Based on ω_i, the maximization of the weighted likelihood of the logistic regression with Y as the outcome and A as the unique explanatory variables allows us to obtain $\hat{\pi}_0^{IPW}$, $\hat{\pi}_1^{IPW}$, and $\widehat{\text{OR}}^{IPW}$.

2.3 Truncated IPW

The weights ω_i can largely inflate for a subject i concerned by positivity near-violation. The usual approach is to truncate the lowest and the highest p_i estimations by the 10$^{\text{th}}$ and 90$^{\text{th}}$ percentiles, respectively (Cole and Hernán, 2008). We also analyzed alternative thresholds, including the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentiles, as well as the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles of the estimated PS. We obtainded truncated stabilized weights, and the estimations $\hat{\pi}_0^{T-IPW}$, $\hat{\pi}_1^{T-IPW}$, and $\widehat{\text{OR}}^{T-IPW}$.

2.4 G-computation (GC)

GC is based on the outcome regression, frequently called the Q-model (Snowden et al., 2011). The logistic regression is often used when Y is binary. Other models or algorithms can constitute alternatives (Austin, 2012). Consider the following Q-model: $\text{logit}\{P(Y = 1|A, Z)\} = \gamma A + \beta Z$. Once fitted, one can compute for each subject i the two expected probabilities of events if she/he is exposed or unexposed, *i.e.*, $\hat{P}(Y_i = 1|do(A_i = 1), z_i)$ and $\hat{P}(Y_i = 1|do(A_i = 0), z_i)$, respectively (Snowden et al., 2011). One can then obtain $\hat{\pi}_a^{GC} = n^{-1}\sum_i \hat{P}(Y_i = 1|do(A_i = a), z_i)$ for $a = 0, 1$; $\hat{\Delta}^{GC} = \hat{\pi}_0^{GC} - \hat{\pi}_1^{GC}$ and $\widehat{\text{OR}}^{GC} = (\hat{\pi}_1^{GC}/(1 - \hat{\pi}_1^{GC}))/(\hat{\pi}_0^{GC}/(1 - \hat{\pi}_0^{GC}))$. This method is implemented in the `RISCA` package, in R (Foucher et al., 2019).

2.5 Targeted Maximum Likelihood Estimation (TMLE)

The first step is to fit the Q-model and estimate the two expected probabilities of events $\hat{\pi}_1^{GC}$ and $\hat{\pi}_0^{GC}$. The additional "targeting" step involves the estimation of p_i, which is then used to update the initial estimates obtained by the Q-model. This step aims to compute first: the clever covariates $H(1,Z) = A/(\text{expit}(\hat{p}_i))$ and $H(0,Z) = (1-A)/(1-\text{expit}(\hat{p}_i))$, where $\text{expit}(\dot{)}$ represents the inverse logit function ($\frac{\exp(\dot{)}}{1+\exp(\dot{)}}$), and second: a vector fluctuation parameter $\hat{\epsilon} = (\hat{\epsilon}_0, \hat{\epsilon}_1)$ estimated through a maximum likelihood procedure. The fluctuation parameter is computed using an outcome model where the logit of the initial prediction of the Q-model is an offset in an intercept-free logistic regression with the clever covariates as explanatory variables (Luque-Fernandez et al., 2018). Therefore, we can generate updated estimates of the set of potential outcomes (Y_1^* and Y_0^*) by incorporating information from the mechanisms to reduce potential biases. We generate $\text{logit}(\hat{Y_1^*}) = \text{logit}(\hat{Y_1}) + \hat{\epsilon} \times H_1$ and $\text{logit}(\hat{Y_0^*}) = \text{logit}(\hat{Y_0}) + \hat{\epsilon} \times H_0$ (Schuler and Rose, 2017). In the presence of residual confounders, the PS provides additional information to improve the initial estimates. It results in the estimations $\hat{\pi}_0^{TMLE}$ and $\hat{\pi}_1^{TMLE}$, *i.e.*, the updated values of $\hat{\pi}_0^{GC}$ and $\hat{\pi}_0^{GC}$, respectively. This method is implemented in the `tmle` package, in R (Gruber and van der Laan, 2012).

2.6 TMLE with truncated PS

As for IPW, the TMLE can use truncated PS in its second stage. The usual method is the truncation of the lowest and highest values of p_i by 0.1 and 0.9, respectively. We also analyzed other alternative truncation levels: 0.05/0.95 and 0.025/0.975. One can then obtain $\hat{\pi}_1^{T-TMLE}$, $\hat{\pi}_0^{T-TMLE}$ and $\widehat{\text{OR}}^{T-TMLE}$. We used the `gbounds` arguments in the `tmle` function of the `tmle` package in R (Gruber and van der Laan, 2012).

2.7 Variance estimators

For each method, was obtained from the usual and well-validated method. For IPW, we used a robust sandwich-type variance estimator (Robins et al., 2000), with the `sandwich` package in R (Zeileis, 2006). For GC, we generated 1,000 bootstrapped samples. This method is implemented in the `RISCA` package in R (Foucher et al., 2019). For TMLE, we used the efficient curve based variance estimator, implemented in the `tmle` package in R (Gruber and van der Laan, 2012).

To improve the comparability of the results, we additionally used the bootstrap for IPW and TMLE-based methods.

3 Simulation study

3.1 Data generation

Figure S1 (supplementary material) represents the directed acyclic graph of the simulations. We first independently generated covariates $Z = (Z_1, ..., Z_9)$: six binary covariates using Bernoulli distributions with different probabilities (0.1 for Z_1, 0.4 for Z_2, 0.7 for Z_4, 0.5 for Z_5, 0.3 for Z_7 and 0.8 for Z_8), and three continuous covariates using a Gaussian distribution with mean at 0 and standard deviation at 1. We generated the exposure A according to a Bernoulli distribution with probability obtained from a logistic model with the following linear predictor: $\alpha_0 + \alpha Z_1 + \alpha Z_2 + \alpha Z_4 + \alpha Z_6 + \alpha Z_7 + \alpha Z_8$, α being the regression coefficients associated with the covariates as detailed in Table S1, and α_0 was set to 1.05 or -0.45 to simulate a prevalence of exposed patients at 80% or 50%, respectively. This design allows us to expect situations of positivity near-violation (Figures S2 and S3), especially for Z_1 which was generated with a 10% prevalence. A prevalence of 50% improved the PS distribution overlap between exposed and

non-exposed subjects, and reducing the risk of positivity near-violation. Furthermore, because the near-violation is more susceptible for small samples, we studied several sample sizes: $n = 100, 200, 500$, and 1000.

We randomly generated the outcome from a Bernoulli distribution with probability obtained from a logistic model with the following linear predictor: $-0.8 + \beta_A A + \beta_Z Z_1 + \beta_Z Z_2 + \beta_Z Z_3 + \beta_Z Z_4 + \beta_Z Z_5 + \beta_Z Z_6 + \beta_{A,Z_1} A * Z_1$, where (β_A, β_Z) were the regression coefficients of A and Z, respectively. To create an extrapolation issue as illustrated in Figure 1, we considered an interaction between A and Z_1 in the outcome generating model to obtain a poorly calibrated model in the area where Z_1 violated the positivity assumption. The values of β_A and β_Z are presented in Table S1. The regression coefficient β_{A,Z_1} of the interaction ranged from $0.0 \cdot \beta_A$ to $2.0 \cdot \beta_A$, according to the intensity of the extrapolation issues : 0.0 for no issue, 0.3 for low issues, 0.9 for moderate issues and 2.0 for high issues.

For each of the 32 scenarios (4 sample sizes, 2 exposures, and 4 extrapolation scenarios), we generated 1,000 datasets. Among the generated datasets for a 50% exposure prevalence, the near-violation of the positivity assumption (no unexposed subjects with Z_1 = 1) concerned 0.0% of the datasets for n = 1000 or 500 subjects, 1.3% for n = 200 subjects, and 14.1% for n = 100 subjects. For an 80% exposure prevalence, this near-violation concerned 0.2% of the datasets for n = 1000, 7.2% for n = 500 subjects, 31.8% for n = 200 subjects, and 58.2% for n = 100 subjects.

3.2 Estimations

We used correctly specified exposure and outcome models to study the impact of positivity near-violation and the extrapolation issue. The interaction between Z_1 and A was introduced in both the models for data generation and the models estimated in each simulated dataset. Even if the outcome model was theoretically well specified, its estimation could result in poor calibrated predictions where there was no data support in the near-violation area.

The interaction between Z1 and A was introduced in both the models for data generation and the models estimated in each simulated dataset. Even if the outcome model is theoretically well specified, its estimation may result in poor calibrated predictions where there is no data support in the near-violation area.

We estimated the true values of π_1 and π_0 by averaging the values obtained from a univariate logistic model (the exposure as the only covariate), fitted from datasets generated as above, except that the exposure A was simulated independently of the covariates Z (Gayat et al., 2012).

To ensure comparability between methods, we decided to set the same strategy of variables' selection. Our set of covariates corresponded to all the outcome causes, theoretically defined by the simulation design (Figure S1), i.e., Z_1, Z_2, Z_3, Z_4, Z_5, and Z_6 (Chatton et al., 2020). We did not study data-adaptive methods to optimize our set of covariates (for instance, the collaborative targeted maximum likelihood estimation (van der Laan and Gruber, 2010)), or even a data-adaptive choice of the truncated PS threshold (Bembom and van der Laan, 2008).

The main estimand was the $\log(\text{OR})$. We reported several associated criteria: the mean absolute bias (MAB) ($MAB = E(\log(\widehat{\text{OR}})) - \log(\text{OR})$) , the variance estimation ratio (VER) by the ratio of estimated model standard deviation to empirical standard deviation ($VER = (\sqrt{E[\widehat{Var}(\log(\widehat{\text{OR}}))]}/\sqrt{Var(\log(\widehat{\text{OR}}))})$, the mean square error (MSE) ($MSE = E[(\log(\widehat{\text{OR}}) - \log(\text{OR}))^2]$) , the coverage rate of the 95% confidence interval (95%CI), and the statistical power. We also reported the mean bias of the probability of an event under the two counterfactual treatments as well as their difference (Δ). We computed the Monte Carlo standard errors for each metric (Morris et al., 2019). We performed all of the analyses using the `R` software package (R Core Team, 2014).

3.3 Results

The results are presented in Figures 2, 3, and 4 for an 80% exposure prevalence. For the methods with truncation, we report in this subsection the results obtained by using the 10^{th} and 90^{th} percentiles, which were associated with the lower MSE values. We also performed the analyses for the 5^{th} and 95^{th} percentiles, and the 2.5th and 97.5th percentiles. These additional results are detailed as supplementary information in Tables S2 and S3 for an exposure prevalence at 80%, and in Tables S6 and S7 for an exposure prevalence at 50% (with Figures S6-8). The bootstrap-based results for an 80% exposure prevalence are presented in Tables S4, S5, with Figures S4 and S5. The results under the null hypothesis are presented in Tables S8-9 with Figures S9-11 for a prevalence of 80%, and in Tables S10-11 with Figures S12-14 for a prevalence of 50%. The standard Monte Carlo errors were negligible and are not presented in the results.

3.3.1 Mean bias

The truncated IPW estimator was biased in almost every situation. For the other methods, the bias increased as the near-violation of positivity was accentuated, i.e., when the sample size decreased (Figure 2). This increase was more significant for IPW estimators. For instance, for the scenario without an extrapolation issue, the MAB was 0.065 for a sample size of 100 subjects, versus -0.002 for 1000 subjects.

The extrapolation issue increased the MAB for methods based on the outcome modeling (GC, TMLE and truncated TMLE), but only when the level was high. For instance, for 200 subjects without extrapolation issue, the MAB for GC and truncated TMLE were -0.006 and 0.000 respectively, versus -0.038 and -0.032 with high extrapolation issue. Even when its level was high, the extrapolation issue had minor consequences when the sample size was equal to or higher than 500 subjects. The TMLE seemed to be the most robust method across all scenarios, especially for small sample sizes ($n = 200$).

For a prevalence of exposure of 50% where the positivity near-violation was lower, the MABs were lower for all of the methods, with comparable results in terms of bias. Even in the most extreme situations (100 subjects with high extrapolation issue), the methods remained robust.

3.3.2 Variance

As illustrated in Figure 3, the decreases in the variance associated with the sample size was comparable across all methods. The extrapolation issue did not affect the variance estimation. However, GC was associated with larger variance when the sample size was smaller. The estimated standard deviation for GC was 1.167 for 100 subjects, 0.399 for 200 subjects, 0.244 for 500 subjects, and 0.177 for 1000 subjects. GC was the only method based on bootstrapping, which can explain this result. Therefore and for comparability sake, we subsequently used bootstrapping for the other methods (Tables S4, S5 and Figure S4). In this situation, variance was similar among all methods for 100 subjects.

Note that regardless of the method used for variance estimation, the standard deviations were similar when the prevalence was 50%. For 100 subjects without extrapolation issue, we estimated a standard deviation at 0.447 for GC, 0.458 for the IPW, 0.433 for the truncated IPW, 0.406 for the TMLE, and 0.405 for the truncated TMLE.

The VER was lower for the TMLE-based methods. This over-optimistic estimation of the variance was partially corrected for the largest sample sizes. More precisely, the VER for TMLE were 0.715 for 100 subjects, 0.798 for 200 subjects, 0.839 for 500 subjects, and 0.923 for 1000 subjects. The use of bootstrapping corrected this over-optimistic estimation (Figure S4). Note that truncated TMLE was associated with lower variances (Tables S2 and S3).

3.3.3 MSE, coverage and power

As illustrated in Figure 4, we observed an increase in the MSE values with the level of the positivity near-violation, in agreement with the previously reported increase in the MAB values. Nevertheless, the

MSE was not significantly affected by the problem of extrapolation. The MSE was lower for GC and truncated methods in the most extreme situation. For instance, for 100 subjects, MSE values were 0.331, 0.507, 0.388, 0.475 and 0.326 for GC, TMLE, truncated TMLE, IPW and truncated IPW, respectively. The lowest MSE was always obtained with the truncated IPW. The second method was GC. Truncated IPW and GC were the two methods with the best bias-variance tradeoff. Note that when the prevalence was 50% (Tables S6 and S7), the MSE for the different methods was similar. However, truncated IPW remained the method with the lowest MSE.

As presented in Figure 4, IPW-based methods and GC resulted in nominal coverage values regardless of the sample sizes. TMLE and truncated TMLE underestimated the variance, resulting in coverage issues. For TMLE-based methods, the underestimated variance results in anti-conservative confidence intervals. More precisely, for scenarios without extrapolation issues, the coverage value of TMLE was 84.6% for 100 subjects, 88.4% for 200 subjects, 88.6% for 500 subjects, and 91.4% for 1000 subjects. The use of bootstrapping allowed to correct this underestimation. However, as reported in Tables S4 and S5, we obtained values greater than 95%, regardless of the extrapolation issue: 97.1% for 100 subjects, 96.4% for 200 subjects, 95.7% for 500 subjects, and 94.7% for 1000 subjects. The previous results under the alternative hypothesis remained consistent under the null hypothesis. The type I error rate was close to the nominal 5% value at for all methods, except for the TMLE-based methods (variance estimation with efficient curves), with values close to 10% throughout the scenarios.

The progressive increase in the extrapolation issue allowed a slight increase in the statistical power for all methods, regardless of the magnitude of the positivity near-violation. In contrast, the statistical power was strongly impacted by the size of the population for all methods, with values around 20% per 100 subjects compared to values around 90% per 1000 subjects. The IPW presented the lowest values, while the truncated TMLE was the highest statistical power method. These results were in agreement with the over-optimistic estimation of the variance for TMLE-based methods. The power of GC was close to the truncated TMLE. The use of truncated methods improved the statistical power. For example, for 200 subjects without extrapolation issues, the powers were 36.0% for truncated TMLE, 35.9% for TMLE, 31.4% for GC, 25.1% for IPW and 23.5% for truncated IPW. For 1000 subjects, the powers were 87.1% for truncated TMLE, 81.0% for TMLE, 85.6% for GC, 76.9% for IPW and 86.3% for truncated IPW.

4 Application: effect of barbiturates in patients with intracranial hypertension

We compared the five methods on a real dataset, in situations We compared the five methods on a real dataset, in situations that could suggest a near-violation of the positivity assumption. We studied barbiturate prescription for the treatment for refractory intracranial hypertension during the first 24 hours post-admission, and its relationship to, in-hospital mortality.

4.1 Methods

We included 1,584 patients from the AtlanREA cohort (www.atlanrea.org, CNIL DR-2013-047). These patients were admitted to an intensive care unit (ICU) in France's western region between March 2013 and February 2018, and were monitored for intracranial pressure.

For covariates selection, to be consistent with the simulations, we selected the covariates causing the outcome (Chatton et al., 2020). For this purpose, as proposed by VanderWeele and Shpitser (2011), we asked experts which covariates caused the outcome (i.e., a history of head trauma, use of osmotherapy, type of brain injury, age, SAPS II score, signs of intracranial hypertension on admission, lactate and creatinine levels on admission). We did not test interactions. We applied B-spline transformations to continuous covariates when the log-linearity assumption did not hold. For IPW-based approaches, we additionally checked the balance between the two weighted groups with standardized differences. We performed complete case analyses.

4.2 Description of the cohort

Among the 1,584 patients, 1,119 had no missing data on the outcome or covariates. One hundred and twenty-seven (127) patients were in the treated group versus 992 control patients (no barbiturate during the first 24 hours post-admission).

We performed a comparison of analyzed patients versus patients excluded due to missing data and the results are shown in Table S12. Excluded patients were mainly less severe (higher Glasgow scores and lower SAPS II scores), with a higher proportion of women, and a different distribution of hospital care centers. Table S13 provides a comparison between the control and barbiturate-treated groups.

Sixty-six patients in the group administered barbiturates died in ICU compared to 256 in the control group. One can note that only six patients in the treated group (4.7%) were over 70 years old versus 126 (12.7%) in the control group (Figure S15). The age ranged was from 19 to 90 years old in the control group versus 19 to 76 years old in the treated group. One can explain a near-violation of the positivity because of two main reasons. Firstly, elderly patients have a lower probability of receiving last-line treatment for intracranial hypertension because of therapy limitations (Calland et al., 2012). Secondly, the treatment prevalence was small, resulting in only 127 patients with barbiturates and the possible sample-to-sample fluctuation.

4.3 Marginal effects estimates

In situations where the age-related near-violation of the positivity concerned 10% of the sample, we first performed an analyses of the overall sample. Next, we restricted the inclusion of patients to those younger than 70 years old. Figure S16 confirms that the patient age, for which we described the positivity violation, was associated with in-hospital mortality. The results are presented in Table 1 and plotted in Figure S17.

By observing the entire sample results, one can notice significant differences between the different methods. The most extreme effects were obtained with the truncated methods, while the previous simulation-based results highlighted their higher bias. More precisely, the truncated IPW (10^{th} and 90^{th} percentiles) had the highest OR (2.909, 95%CI from 1.990 to 4.254), while the truncated TMLE (bounds at 0.1 and 0.9) had the lowest OR (1.043, 95%CI from 0.814 to 1.338). The IPW and the TMLE were the two methods with the highest variance (0.362 and 0.299, respectively). The techniques with the lowest variances were the truncated approaches (0.127 for truncated TMLE and 0.194 for truncated IPW). Only the methods based on the TMLE have a 95% CI for the OR incorporating the value 1.

By comparing the results obtained from the entire sample with those reduced to patients under 70 years old, one can note relative stability in the estimates achieved by the five methods. Nevertheless, the estimations did not vary in the same direction: a slight increase between the estimations performed on the entire sample versus those in the subgroup for the GC, IPW and truncated IPW, and a modest decrease in values for the TMLE-based methods. The methods with the closest results between the entire cohort and the sub-sample were based on the outcome model (TMLE, truncated TMLE and GC). We reported a more considerable difference for IPW and truncated IPW. For instance, the OR obtained with truncated TMLE varied from 1.043 (95%CI from 0.814 to 1.338) to 1.047 (95%CI from 0.791 to 1.396), whereas the values obtained with IPW ranged from 2.158 (95%CI from 1.060 to 4.390) to 2.237 (95%CI from 1.082 to 4.624). Population restriction leads to an increase in variance, especially for TMLE-based methods (Figure S17).

The conclusions that can be drawn from the 95%CI did not change between the overall population and the restricted population. However, one can note that only the TMLE-based methods resulted in non-significant statistical effects, i.e., rendering the study statistically "inconclusive", in contrast to the results obtained by the other methods.

5 Discussion

The results of the simulations illustrated that the near-violation of the positivity assumption could impact the bias and precision of the five methods. In terms of MAB, one can conclude that methods based on

the outcome modeling showed the best results. The addition of an extrapolation issue altered the MAB for these methods, but in a magnitude similar to the one observed for the IPW-based approaches. Whilst the truncated methods introduced bias, they reduced the variance estimation, as previously described by Moore et al. (2012). Methods with the best balance between variance and bias were truncated IPW and GC. TMLE-based methods were associated with an over-optimistic estimate of the variance, resulting in lower coverage than the nominal value. We did not observe this issue when the prevalence of exposure was 50%, i.e., reducing the positivity near-violation. Although the TMLE is a doubly robust estimator consistent when at least one nuisance model is well-specified, the variance estimation can be challenging. Petersen et al. (2014) and Lendle et al. (2017) reported the potential inflation of the type I error and poor coverage in the presence of positivity near-violations. Our results confirm their findings and the potential of a bootstrap-based approach as an alternative. We performed additional simulations with the average exposure effect as the estimand (instead of the logOR), our results were consistent (data not shown).

Whilst the simulations illustrated important differences between each methods performance, the 'real-dataset' application emphasized the importance of the method chosen. Indeed, the clinical conclusion varied according to the specific method. In agreement with the simulations, the variances of truncated methods were smaller, but this benefit has to be counterbalanced with the risk of bias (Cole and Hernán, 2008; Ju et al., 2019). The main concern lies in their optimal cut-off choice, giving us the best bias-variance tradeoff. We have studied consensual thresholds, defined either by a bound value of PS (for TMLE) or by the value of a percentile of the weights (for IPW). An alternative would lie in establishing an algorithm seeking the best bias-variance tradeoff, which would be guided by the data. This solution has recently been studied to choose data-driven PS truncation thresholds adapted to IPW (Bembom and van der Laan, 2008) or to TMLE (Ju et al., 2019), with promising results for positivity violation situations. Another solution may also lie in the use of modern methods such as limited overlap, matching and entropy weightings to reduce the influence of the most extreme observations and focus on the data area with the most overlap, therefore capturing the processing effect for which we have the most information (Zhou et al., 2020). These techniques enable us to estimate an average treatment effect on the population overlap (Li et al., 2018).

Causal inference in observational studies relies directly on the assumption that all participants are eligible to be exposed (or unexposed). Our results confirmed the importance of this assumption since all the methods compared were affected in terms of bias and/or variance. This assumption's violation is more identifiable by using PS-based approaches since it consists of regressing the exposure probability. In contrast, GC involves outcome modeling, and this violation can remain unidentified (Kang and Schafer, 2007). For IPW, subjects who have a low likelihood of exposure but who are exposed, results in extreme weights with unstable estimations and high variances (Kang and Schafer, 2007). The inflated variance and the associated extreme weighting obtained in this way can alert investigators. Unfortunately, the situations at risk of extrapolation are not directly identifiable, and only the violations of positivity can be revealed.

The near-violation of the positivity represents an obstacle to causal inferences only when it concerns true confounders, i.e., those associated with both the exposure and the outcome (Westreich and Cole, 2010). In contrast, imbalance of variables was only associated with exposure, also called instrumental variables, and will have no impact on the bias.

Several authors have previously documented different techniques for detecting restrictions on the positivity assumption in the context of PS analysis (Cole and Hernán, 2008; Austin and Stuart, 2015). The first approach is to study the distribution of the exposure regimen for each covariate, but this can become tedious when dealing with many covariates. One can also use standardized differences (Austin and Stuart, 2015). Another possibility is to compare a groups weights distributions, or even to focus on the distribution of PS. Histogram of the PS distribution by exposure group is an example of interesting representation. In practice, one can assess the positivity assumption by searching for a lack of sufficient overlap of the PS distributions between the exposure groups. However, while useful to diagnose potential positivity violations, these techniques do not provide any quantitative estimate of the estimator bias due to positivity near-violation. Petersen et al. (2012) proposed a parametric bootstrap approach to provide an optimistic bias estimate specifically targeted for positivity violations and near-violations.

Our study has several limitations. Firstly, we only considered TMLE-based methods, while other DRE approaches exist, such as the augmented inverse probability of treatment weighting (A-IPTW) (Glynn and Quinn, 2010). We focused on TMLE because of its better stability compared to A-IPTW (Neugebauer and van der Laan, 2005; Porter et al., 2011; Luque-Fernandez et al., 2018). Secondly, we did not study the different methods for the construction of the model, as this would have multiplied the number of possible approaches to compare. For instance, an alternative to reduce the variance of TMLE is the collaborative TMLE (C-TMLE), which uses a sequential selection of covariates estimating PS (Porter et al., 2011; Lendle et al., 2013; Pirracchio et al., 2018). Machine learning techniques were also proposed for GC (Austin, 2012), or for PS-based methods (Pirracchio et al., 2015). The improvement of the methods we studied by machine learning techniques is an interesting perspective of our work, especially because it can help to reduce the problem of extrapolation. Thirdly, our simulation-based study was not associated with theoretical justification, and it does not demonstrate which method is the best in all situations. Even though our results are in agreement with the current literature, additional studies are required, such as incorporating the extrapolation issue for patients with a higher susceptibility of positivity near-violation.

To conclude, our study illustrates that all the causal methods were sensitive to the near-positivity violation. Nevertheless, we reported the methods' robustness based on the outcome model (GC and TMLE), even with an extrapolation issue. The truncated method, whilst attractive in terms of variance reduction, should be used with caution due to the associated risk of increased bias. G-computation appears to present the best compromise when considering its ability to reduce the bias and its statistical power.

Acknowledgements The authors would like to thank the members of AtlanREA Group for their involvement in the study, the physicians who helped recruit patients, and all patients who participated in this study. We also thank the clinical research associates who participated in the data collection. The analyses and interpretation of these data are the responsibility of the authors. This work was partially supported by a public grant overseen by the French National Research Agency (ANR) to create the Common Laboratory RISCA (www.labcom-risca.com, reference: ANR-16-LCV1-0003-01). The funder had no role in study design; analyses, and interpretation of data; writing the report; or the decision to submit the report for publication.

Conflict of Interest The authors have declared no conflict of interest.

References

Abdia, Y., Kulasekera, K. B., Datta, S., Boakye, M. and Kong, M. (2017) Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, **59**, 967–985.

Austin, P. C. (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*, **46**, 399–424.

— (2012) Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation. *Multivariate Behavioral Research*, **47**, 115–135.

— (2013) The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*, **32**, 2837–2849.

Austin, P. C. and Stuart, E. A. (2015) Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*, **34**, 3661–3679.

Bang, H. and Robins, J. M. (2005) Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, **61**, 962–973.

Bembom, O. and van der Laan, M. J. (2008) Data-adaptive selection of the truncation level for Inverse-Probability-of-Treatment-Weighted estimators. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Calland, J. F., Ingraham, A. M., Martin, N., Marshall, G. T., Schulman, C. I., Stapleton, T., Barraco, R. D. and Eastern Association for the Surgery of Trauma (2012) Evaluation and management of geriatric trauma: an Eastern Association for the Surgery of Trauma practice management guideline. *J Trauma Acute Care Surg*, **73**, S345–350.

Chatton, A., Le Borgne, F., Leyrat, C., Gillaizeau, F., Rousseau, C., Barbin, L., Laplaud, D., Léger, M., Giraudeau, B. and Foucher, Y. (2020) G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Sci Rep*, **10**.

Cole, S. R. and Hernán, M. A. (2008) Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.*, **168**, 656–664.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96**, 187–199.

Foucher, Y., Borgne, F. L., Dantan, E., Gillaizeau, F., Chatton, A. and Combescure, C. (2019) RISCA: Causal Inference and Prediction in Cohort-Based Analyses. URL: `https://CRAN.R-project.org/package=RISCA`.

Gayat, E., Resche-Rigon, M., Mary, J.-Y. and Porcher, R. (2012) Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm Stat*, **11**, 222–229.

Glynn, A. N. and Quinn, K. M. (2010) An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, **18**, 36–56.

Gruber, S. and van der Laan, M. J. (2012) tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, **51**, 1–35. URL: `http://www.jstatsoft.org/v51/i13/`. Doi:10.18637/jss.v051.i13.

Hajage, D., Tubach, F., Steg, P. G., Bhatt, D. L. and De Rycke, Y. (2016) On the use of propensity scores in case of rare exposure. *BMC Medical Research Methodology*, **16**.

Hernán, M. A. and Robins, J. M. (2020) *Causal Inference: What if?* Chapman & Hall/CRC. URL: `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`.

Ju, C., Schwab, J. and van der Laan, M. J. (2019) On adaptive propensity score truncation in causal inference. *Stat Methods Med Res*, **28**, 1741–1760.

Kang, J. D. Y. and Schafer, J. L. (2007) Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statist. Sci.*, **22**, 523–539.

Le Borgne, F., Giraudeau, B., Querard, A. H., Giral, M. and Foucher, Y. (2016) Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. *Stat Med*, **35**, 1103–1116.

Lendle, S. D., Fireman, B. and van der Laan, M. J. (2013) Targeted maximum likelihood estimation in safety analysis. *J Clin Epidemiol*, **66**, S91–98.

Lendle, S. D., Schwab, J., Petersen, M. L. and Laan, M. J. v. d. (2017) ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *Journal of Statistical Software*, **81**, 1–21. URL: `https://www.jstatsoft.org/index.php/jss/article/view/v081i01`. Number: 1.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2018) Balancing Covariates via Propensity Score Weighting. *J Am Stat Assoc*, **113**, 390–400.

Luque-Fernandez, M. A., Schomaker, M., Rachet, B. and Schnitzer, M. E. (2018) Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, **37**, 2530–2546.

Moore, K. L., Neugebauer, R., van der Laan, M. J. and Tager, I. B. (2012) Causal inference in epidemiological studies with strong confounding. *Stat Med*, **31**, 1380–1404.

Morris, T. P., White, I. R. and Crowther, M. J. (2019) Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**, 2074–2102.

Neugebauer, R. and van der Laan, M. J. (2005) Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, **129**, 405–426.

— (2006) G-computation estimation for causal inference with complex longitudinal data. *Comput. Stat. Data Anal.*, **51**, 1676–1697.

Pearl, J., Glymour, M. and Jewell, N. P. (2016) *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M. and van der Laan, M. (2014) Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models. *J Causal Inference*, **2**, 147–185.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. and van der Laan, M. J. (2012) Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*, **21**, 31–54.

Pirracchio, R. and Carone, M. (2018) The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research*, **27**, 2504–2518.

Pirracchio, R., Petersen, M. L. and van der Laan, M. (2015) Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, **181**, 108–119.

Pirracchio, R., Yue, J. K., Manley, G. T., van der Laan, M. J. and Hubbard, A. E. (2018) Collaborative targeted maximum likelihood estimation for variable importance measure: Illustration for functional outcome prediction in mild traumatic brain injuries. *Stat Methods Med Res*, **27**, 286–297.

Porter, K. E., Gruber, S., van der Laan, M. J. and Sekhon, J. S. (2011) The relative performance of targeted maximum likelihood estimators. *Int J Biostat*, **7**.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins, J. M. (1986) 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*, **7**, 1393–1512.

Robins, J. M., Hernán, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Schuler, M. S. and Rose, S. (2017) Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *Am. J. Epidemiol.*, **185**, 65–73.

Snowden, J. M., Rose, S. and Mortimer, K. M. (2011) Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am. J. Epidemiol.*, **173**, 731–738.

van der Laan, M. J. and Gruber, S. (2010) Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, **6**, 17–17.

van der Laan, M. J. and Rubin, D. B. (2006) Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, **2**.

van der Laan, Mark J., R. J. (2003) *Unified Methods for Censored Longitudinal Data and Causality*. Springer: New-York.

VanderWeele, T. J. and Shpitser, I. (2011) A new criterion for confounder selection. *Biometrics*, **67**, 1406–1413.

Vansteelandt, S. and Keiding, N. (2011) Invited Commentary: G-Computation–Lost in Translation? *Am J Epidemiol*, **173**, 739–742.

Westreich, D. and Cole, S. R. (2010) Invited commentary: positivity in practice. *Am. J. Epidemiol.*, **171**, 674–677; discussion 678–681.

Williamson, E. J., Morley, R., Lucas, A. and Carpenter, J. (2012) Propensity scores: From naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, **21**, 273–293.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C. and Smith, D. (2010) Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health*, **13**, 273–277.

Zeileis, A. (2006) Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, **16**, 1–16.

Zhou, Y., Matsouaka, R. A. and Thomas, L. (2020) Propensity score weighting under limited overlap and model misspecification. *Stat Methods Med Res*, Online ahead of print.

Table 1 Results obtained by using g-computation (GC), inverse probability weighting (IPW), truncated IPW, targeted maximum likelihood estimator (TMLE) and truncated TMLE for estimating barbiturates effects.

AtlanREA cohort: the barbiturates effect

	π_1	π_0	Δ	$\log(\text{OR})$	SD	OR	95%CI OR
Whole sample							
GC	0.396	0.272	0.124	0.560	0.194	1.750	1.207 - 2.524
IPW	0.447	0.273	0.174	0.769	0.362	2.158	1.060 - 4.390
Truncated IPW [10%-90%]	0.515	0.267	0.248	1.068	0.194	2.909	1.990 - 4.254
Truncated IPW [5%-95%]	0.470	0.271	0.199	0.872	0.236	2.391	1.506 - 3.795
Truncated IPW [2.5%-97.5%]	0.467	0.272	0.195	0.853	0.245	2.347	1.453 - 3.792
TMLE	0.320	0.289	0.031	0.146	0.299	1.158	0.645 - 2.079
Truncated TMLE [0.1-0.9]	0.298	0.288	0.010	0.043	0.127	1.043	0.814 - 1.338
Truncated TMLE [0.05-0.95]	0.311	0.289	0.022	0.107	0.156	1.112	0.820 - 1.509
Truncated TMLE [0.025-0.975]	0.311	0.289	0.022	0.108	0.219	1.114	0.725 - 1.711
Restricted sample							
GC	0.370	0.243	0.127	0.606	0.196	1.833	1.259 - 2.670
IPW	0.418	0.243	0.175	0.805	0.371	2.237	1.082 - 4.624
Truncated IPW [10%-90%]	0.499	0.238	0.261	1.160	0.203	3.188	2.142 - 4.746
Truncated IPW [5%-95%]	0.447	0.242	0.205	0.932	0.242	2.539	1.579 - 4.082
Truncated IPW [2.5%-97.5%]	0.447	0.243	0.204	0.902	0.250	2.464	1.509 - 4.025
TMLE	0.280	0.263	0.017	0.090	0.344	1.094	0.558 - 2.145
Truncated TMLE [0.1-0.9]	0.272	0.263	0.009	0.046	0.143	1.047	0.791 - 1.386
Truncated TMLE [0.05-0.95]	0.274	0.263	0.011	0.056	0.183	1.058	0.739 - 1.515
Truncated TMLE [0.025-0.975]	0.275	0.263	0.012	0.066	0.269	1.068	0.630 - 1.809

Abbreviations: π_1, the expected proportions of event if the entire population is exposed ; π_0, the expected proportions of event if the entire population is unexposed; Δ, the risk difference (π_1 - π_0) ; $\log(\text{OR})$, the logarithm of the odds ratio ; SD, the standard deviation for the logarithm of the odds ratio; OR, The corresponding odds-ratio ; 95%CI, 95% confidence interval of the odds-ratio.

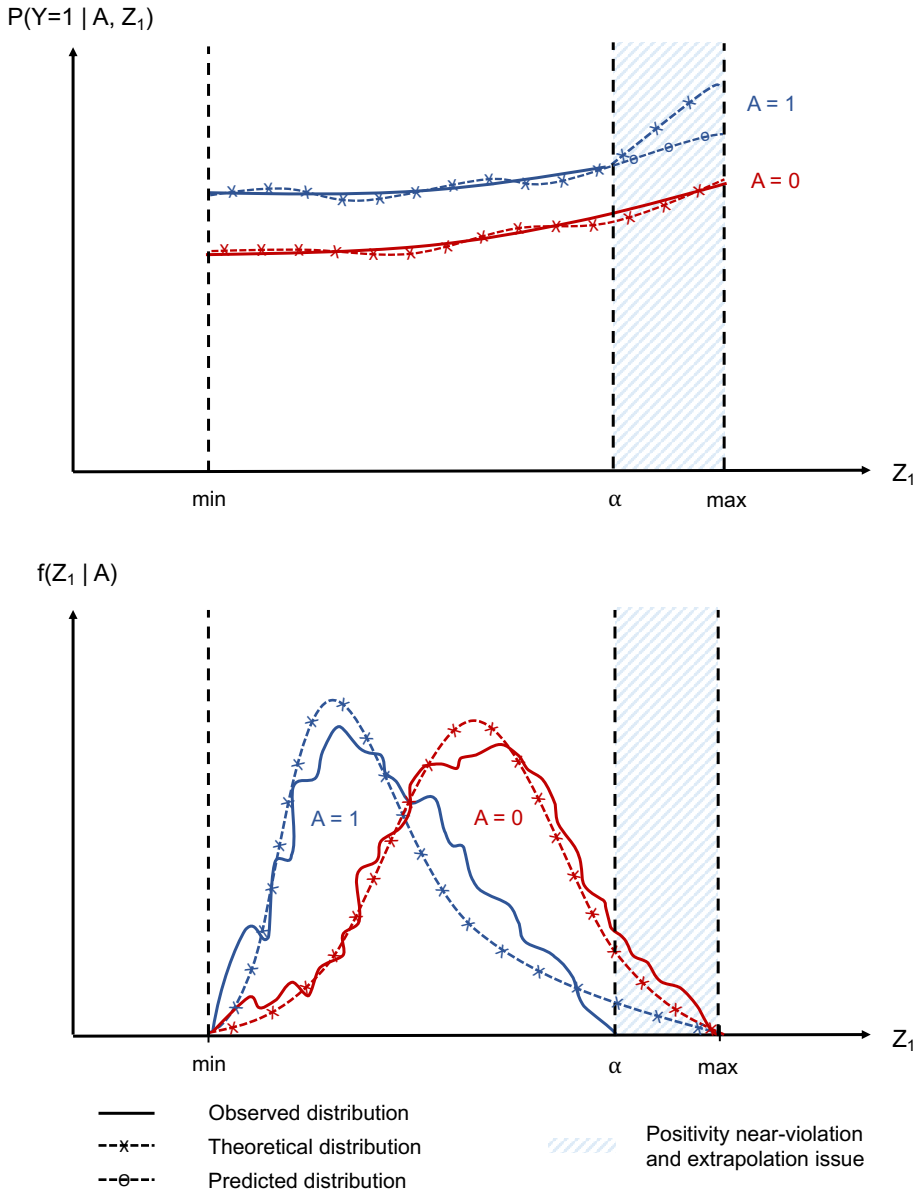


Figure 1 A representative illustration of the extrapolation issue occurring with a positivity near-violation. The left y-axis represents the conditional distribution function of the covariate Z_1 according to the exposure status. The right y-axis represents the conditional probability of the outcome.

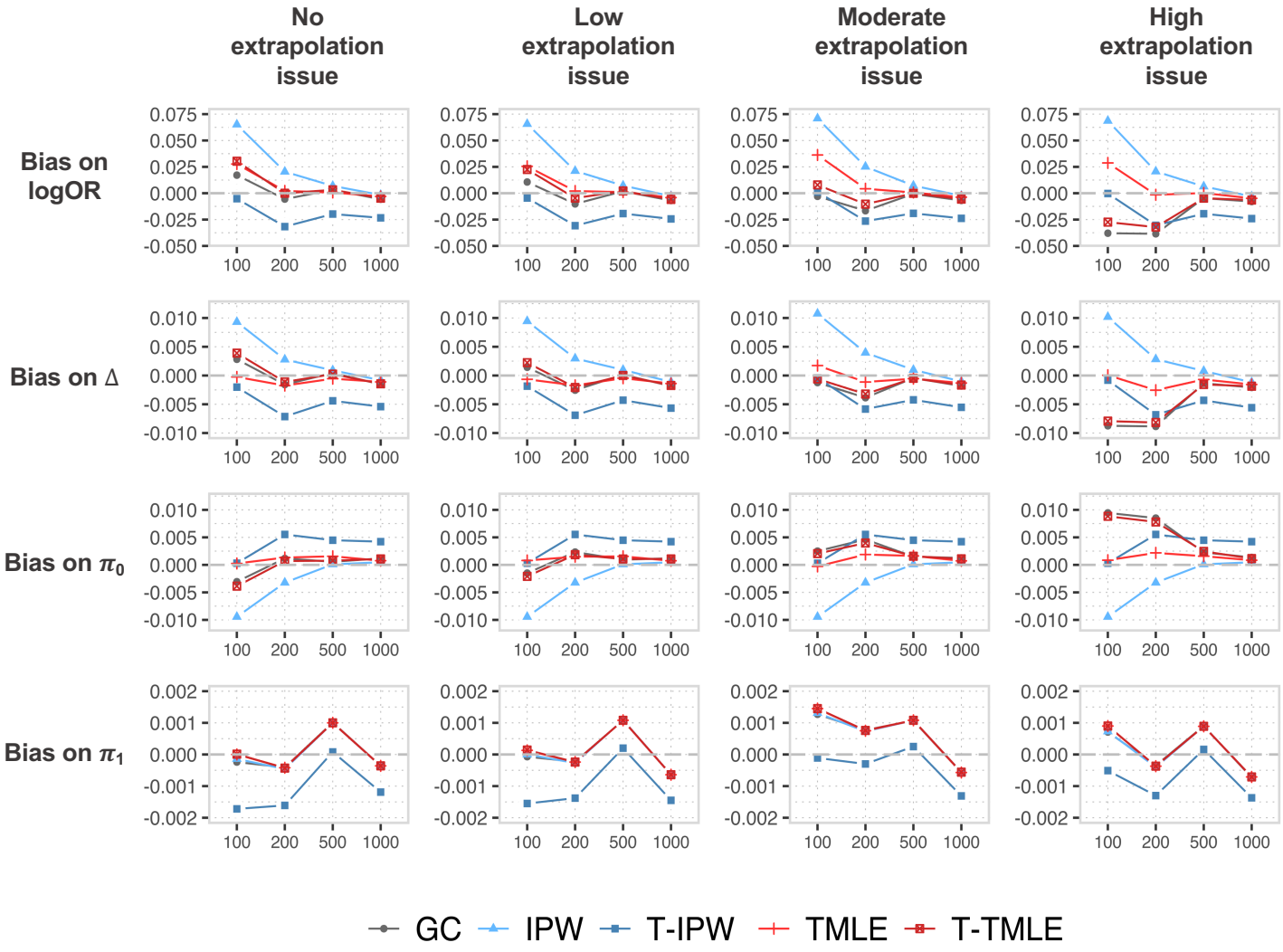


Figure 2 The mean absolute bias (y-axis) according to different sample size (from 100 to 1,000, x-axis) and extrapolation issue. Abbreviations: GC, g-computation; IPW, inverse probability weighting; T-IPW, truncated inverse probability weighting (thresholds: 10^{th} and 90^{th} percentiles) ; TMLE, targeting maximum likelihood estimator; T-TMLE, truncated targeting maximum likelihood estimator (thresholds: bounds at 0.1 and 0.9); π_1, the expected proportions of event if the entire population is exposed; π_0, the expected proportions of event if the entire population is unexposed; Δ, the corresponding difference (π_1 - π_0); OR, the corresponding odds-ratio.

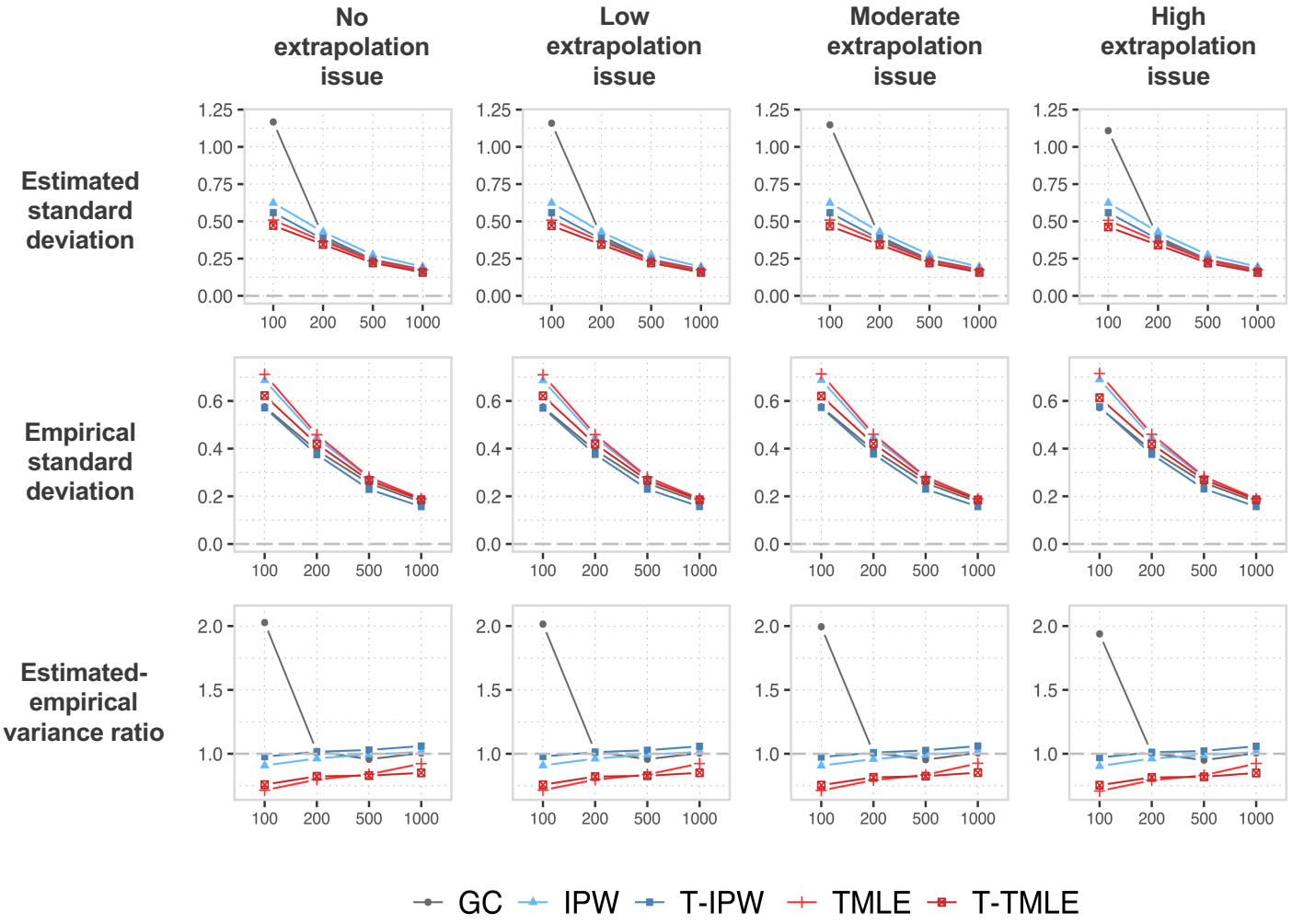


Figure 3 Graphical representation of the evolution of accuracy (empirical standard deviation, estimated standard deviation and, the variance estimation ratio) according to different sample size (from 100 to 1,000, x-axis) and extrapolation issue. The target parameter was $\log(\text{OR})$. Abbreviations: GC, g-computation; IPW, inverse probability weighting; T-IPW, truncated inverse probability weighting (thresholds: 10^{th} and 90^{th} percentiles) ; TMLE, targeting maximum likelihood estimator; T-TMLE, truncated targeting maximum likelihood estimator (thresholds: bounds at 0.1 and 0.9).

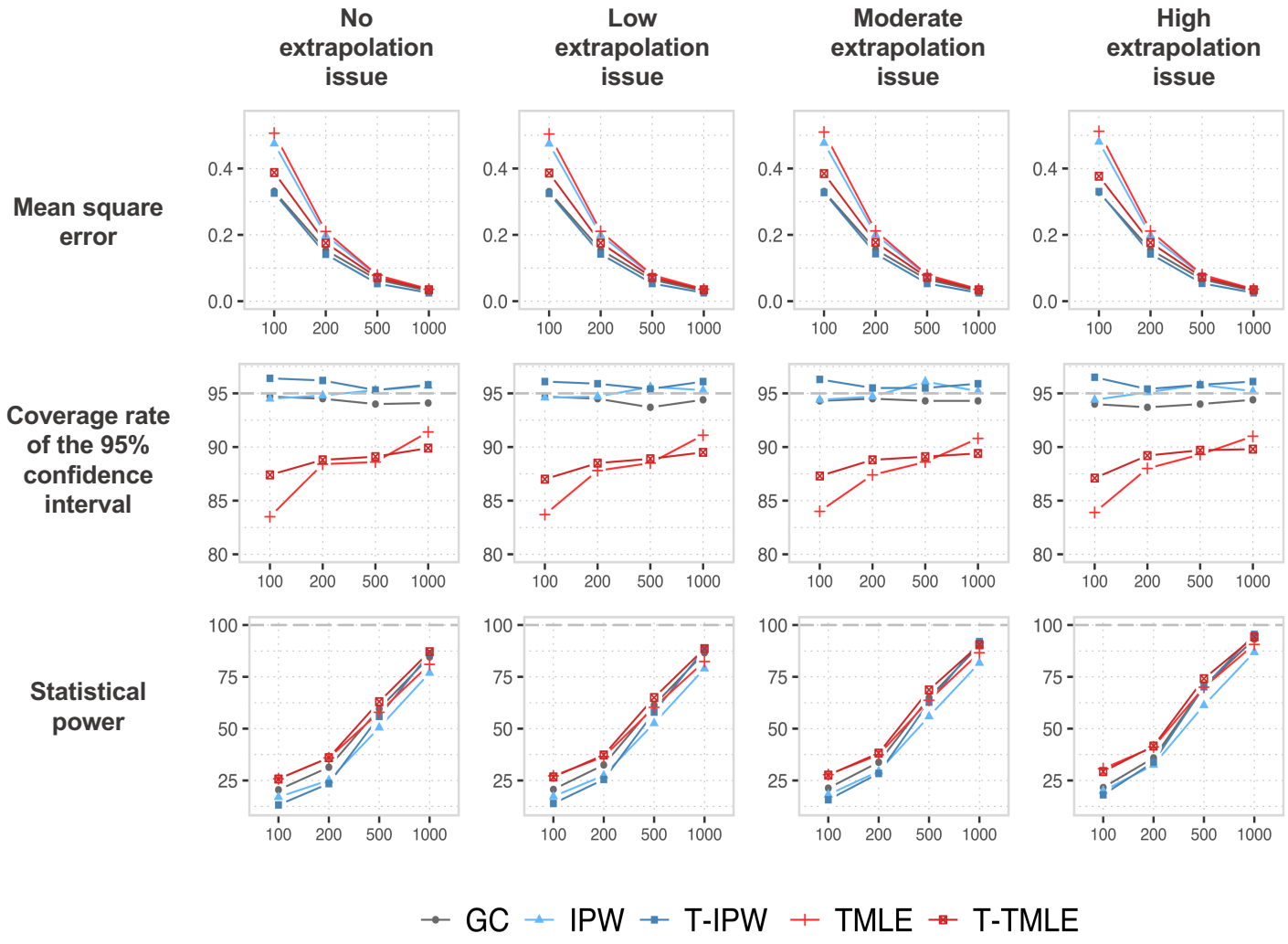


Figure 4 The mean square error, the coverage of the 95% confidence interval and the statistical power according to different sample size (from 100 to 1,000, x-axis), and extrapolation issue. The target parameter was $\log(\text{OR})$. Abbreviations: GC, g-computation; IPW, inverse probability weighting; T-IPW, truncated inverse probability weighting (thresholds: 10^{th} and 90^{th} percentiles) ; TMLE, targeting maximum likelihood estimator; T-TMLE, truncated targeting maximum likelihood estimator (thresholds: bounds at 0.1 and 0.9).

Annexe D

Éléments supplémentaires au chapitre 3

G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study

Arthur Chatton[1,2], Florent Le Borgne[1,2], Clémence Leyrat[1,3], Florence Gillaizeau[1,4], Chloé Rousseau[1,4,5], Laetitia Barbin[4], David Laplaud[4,6], Maxime Léger[1,7], Bruno Giraudeau[1,8], and Yohann Foucher[1,4,*]

[1]INSERM UMR 1246 - SPHERE, Université de Nantes, Université de Tours, Nantes, France.
[2]A2COM-IDBC, Pacé, France.
[3]Department of Medical Statistics & Cancer Survival Group, London School of Hygiene and Tropical Medicine, London, UK.
[4]Centre Hospitalier Universitaire de Nantes, Nantes, France.
[5]INSERM CIC1414, CHU Rennes, Rennes, France.
[6]Centre de Recherche en Transplantation et Immunologie INSERM UMR1064, Université de Nantes, Nantes, France.
[7]Département d'Anesthésie-Réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France.
[8]INSERM CIC1415, CHRU de Tours, Tours, France.
[*]Yohann.Foucher@univ-nantes.fr

ABSTRACT

Controlling for confounding bias is crucial in causal inference. Distinct methods are currently employed to mitigate the effects of confounding bias. Each requires the introduction of a set of covariates, which remains difficult to choose, especially regarding the different methods. We conduct a simulation study to compare the relative performance results obtained by using four different sets of covariates (those causing the outcome, those causing the treatment allocation, those causing both the outcome and the treatment allocation, and all the covariates) and four methods: g-computation, inverse probability of treatment weighting, full matching and targeted maximum likelihood estimator. Our simulations are in the context of a binary treatment, a binary outcome and baseline confounders. The simulations suggest that considering all the covariates causing the outcome led to the lowest bias and variance, particularly for g-computation. The consideration of all the covariates did not decrease the bias but significantly reduced the power. We apply these methods to two real-world examples that have clinical relevance, thereby illustrating the real-world importance of using these methods. We propose an R package *RISCA* to encourage the use of g-computation in causal inference.

Supplementary Materials

R code for RISCA use:

```
library(RISCA)

#data simulation
#treatment = 1 if the patients have been the treatment of interest and 0 otherwise
treatment <- rbinom(600, 1, prob=0.5)
covariate <- rnorm(600, 0, 1)
covariate[treatment==1] <- rnorm(sum(treatment==1), 0.3, 1)
outcome <- rbinom(600, 1, prob=1/(1+exp(-2-0.26*treatment-0.7*covariate)))
tab <- data.frame(outcome, treatment, covariate)

#Raw effect of the treatment
```

```
glm.raw <- glm(outcome ~ treatment, data=tab, family = binomial(link=logit))
summary(glm.raw)

#Conditional effect of the treatment
glm.multi <- glm(outcome ~ treatment + covariate, data=tab, family = binomial)
summary(glm.multi)

#Marginal effects of the treatment (ATE)
gc.ate <- GC.Logistic(glm.obj=glm.multi, data=tab, group="treatment", effect="ATE",
var.method="simulations", iterations=1000)

#Sum-up of the 3 ORs
data.frame( raw=exp(glm.raw$coefficients[2]),
conditional=exp(glm.multi$coefficients[2]),
marginal.ate=exp(gc.ate$logOR[,1]) )
```

n	method	set	mean bias				logOR			
			π_0	π_1	$\Delta\pi$	logOR	MSE	VEB	coverage	type I
100	GC	outcome	0.001	0.000	-0.001	-0.002	0.396	-7.1	93.1	5.6
		treatment	0.000	0.001	0.001	0.005	0.457	-7.7	92.8	5.7
		common	0.001	-0.000	-0.001	-0.003	0.419	-5.2	93.6	5.3
		entire	0.000	0.001	0.001	0.005	0.439	-10.9	91.7	6.3
	IPTW	outcome	-0.000	0.001	0.002	0.009	0.455	11.5	97.3	2.7
		treatment	-0.004	0.005	0.008	0.037	0.599	-4.9	94.7	5.3
		common	-0.000	0.001	0.001	0.006	0.462	7.6	96.6	3.5
		entire	-0.004	0.005	0.009	0.040	0.622	-5.6	95.1	5.0
	TMLE	outcome	0.001	0.000	-0.000	0.001	0.432	-14.7	89.8	10.2
		treatment	-0.003	0.005	0.008	0.037	0.549	-23.8	85.0	15.0
		common	0.001	-0.000	-0.001	-0.003	0.452	-9.8	91.4	8.6
		entire	-0.004	0.006	0.009	0.044	0.533	-30.6	81.3	18.7
	FM	outcome	-0.002	0.004	0.006	0.029	0.530	-21.9	88.1	11.9
		treatment	-0.006	0.007	0.013	0.062	0.648	-35.4	79.8	20.1
		common	-0.002	0.003	0.005	0.026	0.534	-22.6	87.7	12.3
		entire	-0.007	0.007	0.014	0.067	0.653	-35.7	79.7	20.2
300	GC	outcome	-0.001	0.000	0.001	0.005	0.217	-2.2	94.7	5.2
		treatment	-0.001	0.001	0.001	0.006	0.251	-2.0	94.5	5.2
		common	-0.001	0.000	0.001	0.004	0.234	-1.7	94.5	5.3
		entire	-0.001	0.001	0.001	0.006	0.235	-3.0	94.4	5.1
	IPTW	outcome	-0.001	0.001	0.002	0.008	0.236	19.3	98.1	1.9
		treatment	-0.002	0.001	0.003	0.012	0.319	5.8	96.2	3.8
		common	-0.001	0.001	0.002	0.008	0.249	12.0	97.0	3.0
		entire	-0.002	0.001	0.003	0.012	0.314	9.2	97.1	2.9
	TMLE	outcome	-0.001	0.000	0.001	0.004	0.230	-3.6	93.9	6.1
		treatment	-0.001	0.001	0.002	0.011	0.301	-10.3	91.3	8.7
		common	-0.001	0.000	0.001	0.004	0.246	-2.5	94.0	6.0
		entire	-0.001	0.001	0.003	0.011	0.281	-12.5	90.3	9.7
	FM	outcome	-0.001	0.001	0.002	0.010	0.285	-17.8	89.4	10.6
		treatment	-0.003	0.002	0.006	0.025	0.374	-37.0	78.5	21.5
		common	-0.001	0.000	0.001	0.007	0.305	-23.1	87.2	12.7
		entire	-0.004	0.003	0.006	0.028	0.359	-34.4	80.5	19.5
500	GC	outcome	-0.000	-0.000	-0.000	-0.000	0.169	-2.4	94.2	5.8
		treatment	-0.000	-0.000	0.000	0.002	0.196	-2.6	94.2	5.7
		common	-0.000	-0.000	0.000	0.000	0.182	-2.1	94.5	5.5
		entire	-0.000	-0.000	0.000	0.001	0.182	-3.0	93.9	5.8
	IPTW	outcome	-0.001	-0.000	0.001	0.003	0.181	18.6	98.0	2.0
		treatment	-0.001	0.000	0.001	0.005	0.244	6.7	96.5	3.5
		common	-0.001	0.000	0.001	0.003	0.192	11.3	97.1	2.9
		entire	-0.001	-0.000	0.001	0.003	0.237	11.3	97.0	3.0
	TMLE	outcome	-0.000	-0.000	-0.000	-0.000	0.178	-3.5	93.6	6.4
		treatment	-0.001	-0.000	0.000	0.002	0.232	-7.2	92.1	8.0
		common	-0.000	-0.000	0.000	0.000	0.191	-2.7	94.3	5.8
		entire	-0.001	-0.000	0.000	0.001	0.216	-8.7	91.5	8.5
	FM	outcome	-0.001	-0.000	0.000	0.002	0.216	-16.3	90.2	9.8
		treatment	-0.003	0.001	0.004	0.015	0.295	-38.5	77.5	22.6
		common	-0.000	-0.000	0.000	0.002	0.260	-30.4	83.2	16.9
		entire	-0.002	0.001	0.003	0.014	0.284	-36.1	79.2	20.8
2000	GC	outcome	-0.000	0.000	0.000	0.001	0.083	-1.0	94.7	5.5
		treatment	0.000	0.000	0.000	0.001	0.096	-0.7	94.8	5.2
		common	0.000	0.000	0.000	0.001	0.090	-0.7	94.9	5.3
		entire	-0.000	0.000	0.000	0.002	0.089	-0.8	94.8	5.4
	IPTW	outcome	-0.000	0.001	0.001	0.004	0.089	20.1	98.1	1.8
		treatment	-0.000	0.000	0.001	0.002	0.116	11.5	97.1	2.9
		common	-0.000	0.001	0.001	0.003	0.094	12.9	97.2	2.8
		entire	-0.000	0.000	0.001	0.003	0.111	16.7	97.7	2.2
	TMLE	outcome	0.000	0.000	0.000	0.001	0.087	-1.3	94.3	5.6
		treatment	0.000	0.000	0.000	0.001	0.111	-0.8	94.7	5.4
		common	0.000	0.000	0.000	0.001	0.093	-0.6	94.7	5.3
		entire	-0.000	0.000	0.001	0.002	0.104	-1.6	94.3	5.7
	FM	outcome	-0.000	0.000	0.000	0.002	0.126	-28.5	84.1	15.9
		treatment	-0.000	0.001	0.001	0.005	0.162	-44.4	72.3	27.7
		common	-0.000	0.000	0.000	0.002	0.197	-54.2	62.7	37.2
		entire	-0.000	0.001	0.001	0.004	0.137	-33.9	80.7	19.4

Simulation results comparing the ATE estimation under the null hypothesis.

n	method	set	mean bias				logOR			
			π_0	π_1	$\Delta\pi$	logOR	MSE	VEB	coverage	power
100	GC	outcome	0.002	-0.000	-0.003	0.000	0.526	-10.9	93.1	6.7
		treatment	0.002	-0.000	-0.003	-0.001	0.584	-10.4	93.1	6.3
		common	0.003	-0.000	-0.003	-0.002	0.555	-9.2	93.6	6.4
		entire	0.001	-0.000	-0.001	0.003	0.562	-13.4	91.8	7.5
	IPTW	outcome	0.003	-0.000	-0.004	-0.010	0.579	6.5	96.6	3.2
		treatment	0.002	-0.000	-0.002	-0.011	0.715	-4.3	94.4	5.5
		common	0.004	-0.000	-0.005	-0.011	0.593	1.6	95.9	4.0
		entire	-0.001	-0.000	0.001	-0.006	0.744	-4.8	94.5	5.5
	TMLE	outcome	0.003	-0.000	-0.004	-0.022	0.696	30.6	95.9	3.9
		treatment	0.003	-0.000	-0.004	-0.037	0.875	210.4	99.2	0.7
		common	0.004	-0.000	-0.004	-0.022	0.709	5.3	94.3	5.6
		entire	0.001	-0.000	-0.001	-0.028	0.884	498.8	99.5	0.5
	FM	outcome	-0.001	-0.000	0.001	0.005	0.662	-20.6	88.9	11.3
		treatment	-0.004	-0.000	0.004	0.010	0.817	-35.3	79.5	20.2
		common	0.000	-0.000	-0.000	0.002	0.658	-20.3	89.0	11.0
		entire	-0.005	-0.000	0.005	0.010	0.835	-36.6	78.5	21.4
300	GC	outcome	-0.000	0.000	0.000	0.004	0.268	-2.2	94.7	5.1
		treatment	-0.000	0.000	0.000	0.003	0.305	-3.0	94.4	5.3
		common	0.000	0.000	0.000	0.003	0.290	-2.5	94.7	5.3
		entire	-0.000	0.000	0.001	0.004	0.282	-3.0	94.7	5.1
	IPTW	outcome	0.000	0.000	-0.000	0.001	0.287	16.0	98.0	2.0
		treatment	-0.002	0.000	0.002	0.008	0.365	4.7	96.1	3.9
		common	0.000	0.000	-0.000	0.001	0.307	7.6	96.8	3.2
		entire	-0.002	0.000	0.003	0.008	0.353	10.4	97.1	2.9
	TMLE	outcome	-0.001	0.000	0.001	0.003	0.347	-11.2	92.1	7.9
		treatment	-0.001	0.000	0.001	-0.001	0.451	71.6	99.3	0.7
		common	-0.001	0.000	0.002	0.005	0.371	-14.7	90.9	9.1
		entire	-0.001	0.000	0.001	-0.003	0.427	106.5	99.6	0.4
	FM	outcome	-0.001	0.000	0.002	0.005	0.338	-13.8	90.9	9.1
		treatment	-0.004	0.000	0.004	0.011	0.444	-34.2	80.3	19.8
		common	-0.002	0.000	0.002	0.007	0.351	-17.0	89.9	10.1
		entire	-0.004	0.000	0.004	0.011	0.437	-33.2	81.8	18.2
500	GC	outcome	0.000	-0.001	-0.001	-0.003	0.203	-0.7	95.0	5.0
		treatment	0.001	-0.001	-0.002	-0.006	0.232	-1.6	94.8	5.1
		common	0.001	-0.001	-0.002	-0.005	0.221	-1.3	94.8	4.9
		entire	0.000	-0.001	-0.001	-0.005	0.215	-1.4	94.8	5.2
	IPTW	outcome	0.001	-0.001	-0.002	-0.008	0.219	16.6	97.9	2.1
		treatment	0.001	-0.001	-0.002	-0.008	0.276	6.2	96.2	3.8
		common	0.001	-0.001	-0.002	-0.009	0.233	8.8	96.7	3.2
		entire	0.000	-0.001	-0.001	-0.007	0.265	11.7	97.0	3.0
	TMLE	outcome	-0.000	-0.001	-0.001	-0.005	0.264	-15.8	90.0	10.0
		treatment	0.001	-0.001	-0.002	-0.012	0.341	51.5	99.0	1.0
		common	-0.000	-0.001	-0.001	-0.004	0.283	-17.2	89.4	10.6
		entire	0.001	-0.001	-0.002	-0.011	0.320	69.4	99.4	0.6
	FM	outcome	-0.000	-0.001	-0.001	-0.004	0.254	-11.7	92.1	7.9
		treatment	-0.001	-0.001	-0.000	-0.004	0.337	-33.2	80.9	19.1
		common	0.000	-0.001	-0.001	-0.005	0.276	-18.5	89.2	10.7
		entire	-0.000	-0.001	-0.001	-0.006	0.324	-30.6	82.7	17.3
2000	GC	outcome	0.000	0.000	0.000	0.000	0.102	-1.7	94.4	5.8
		treatment	0.000	0.000	-0.000	-0.000	0.115	-0.9	94.9	5.2
		common	0.000	0.000	-0.000	-0.000	0.110	-1.1	95.1	4.9
		entire	0.000	0.000	0.000	0.001	0.107	-1.6	94.3	5.8
	IPTW	outcome	0.001	0.000	-0.001	-0.005	0.108	15.9	97.6	2.4
		treatment	0.000	0.000	-0.000	-0.000	0.132	8.8	96.4	3.6
		common	0.001	0.000	-0.001	-0.005	0.115	9.0	96.6	3.4
		entire	0.000	0.000	0.000	0.000	0.127	13.9	97.3	2.7
	TMLE	outcome	0.000	0.000	-0.000	-0.001	0.129	-19.2	88.3	11.7
		treatment	0.000	0.000	0.000	-0.001	0.164	34.4	99.0	1.0
		common	0.000	0.000	-0.000	-0.000	0.138	-18.7	88.6	11.4
		entire	0.000	0.000	0.000	-0.000	0.154	37.2	99.1	0.9
	FM	outcome	0.000	0.000	0.000	0.000	0.129	-13.3	91.0	9.0
		treatment	0.000	0.000	0.000	-0.001	0.196	-43.1	73.4	26.7
		common	0.000	0.000	-0.000	-0.001	0.144	-22.6	87.5	12.6
		entire	0.000	0.000	0.000	-0.001	0.155	-28.1	84.0	16.0

Simulation results comparing the ATT estimation under the null hypothesis.

n	method	set	mean bias				logOR			
			π_0	π_1	$\Delta\pi$	logOR	MSE	VEB	coverage	type I
100	GC	outcome	-0.113	-0.000	0.113	0.471	0.686	-5.1	83.3	16.8
		treatment	-0.123	-0.000	0.123	0.512	0.754	-5.7	83.6	15.9
		common	-0.113	-0.000	0.112	0.468	0.702	-3.6	85.4	14.9
		entire	-0.124	-0.000	0.124	0.516	0.739	-7.9	81.1	18.0
	IPTW	outcome	-0.113	-0.000	0.113	0.471	0.699	9.5	90.7	9.5
		treatment	-0.126	-0.000	0.126	0.526	0.804	1.6	87.6	12.9
		common	-0.113	-0.000	0.112	0.467	0.710	3.9	89.3	10.9
		entire	-0.127	-0.000	0.127	0.531	0.813	4.3	89.0	11.4
	TMLE	outcome	-0.113	-0.000	0.113	0.473	0.771	1.2	87.2	13.1
		treatment	-0.129	-0.000	0.129	0.548	0.929	35.3	93.5	6.8
		common	-0.113	-0.000	0.113	0.471	0.766	4.1	87.9	12.4
		entire	-0.128	-0.000	0.128	0.545	0.933	119.1	96.2	3.9
	FM	outcome	-0.115	-0.000	0.115	0.480	0.764	-12.9	82.1	18.3
		treatment	-0.127	-0.000	0.127	0.536	0.886	-26.3	75.0	25.4
		common	-0.114	-0.000	0.113	0.474	0.765	-13.9	82.3	18.0
		entire	-0.129	-0.000	0.129	0.546	0.891	-26.1	74.0	26.4
300	GC	outcome	-0.115	-0.000	0.115	0.470	0.544	-2.5	57.7	42.2
		treatment	-0.126	-0.000	0.126	0.513	0.598	-3.2	59.3	40.9
		common	-0.115	-0.000	0.115	0.469	0.553	-2.6	62.8	37.8
		entire	-0.126	-0.000	0.125	0.513	0.588	-3.8	54.2	45.2
	IPTW	outcome	-0.115	-0.000	0.115	0.471	0.547	11.3	69.2	31.0
		treatment	-0.129	-0.000	0.129	0.527	0.621	3.0	66.4	33.7
		common	-0.115	-0.000	0.115	0.469	0.555	3.7	68.2	31.9
		entire	-0.129	-0.000	0.129	0.528	0.614	9.3	67.9	32.2
	TMLE	outcome	-0.116	-0.000	0.116	0.475	0.568	1.6	67.9	32.2
		treatment	-0.134	-0.000	0.134	0.551	0.678	-7.8	65.3	34.9
		common	-0.116	-0.000	0.116	0.474	0.576	4.2	72.0	28.1
		entire	-0.134	-0.000	0.134	0.549	0.666	1.8	68.4	31.8
	FM	outcome	-0.115	-0.000	0.115	0.471	0.568	-8.2	62.7	37.5
		treatment	-0.130	-0.000	0.129	0.530	0.652	-23.3	54.3	45.9
		common	-0.115	-0.000	0.115	0.471	0.580	-14.3	61.7	38.4
		entire	-0.131	-0.000	0.130	0.535	0.653	-22.4	53.7	46.4
500	GC	outcome	-0.116	-0.000	0.115	0.470	0.515	-1.9	37.9	61.7
		treatment	-0.127	-0.000	0.126	0.514	0.565	-1.8	39.0	60.4
		common	-0.116	-0.000	0.116	0.471	0.522	-1.7	42.6	57.5
		entire	-0.126	-0.000	0.126	0.513	0.558	-2.3	33.6	65.8
	IPTW	outcome	-0.116	-0.000	0.115	0.470	0.516	11.6	49.8	9.5
		treatment	-0.130	-0.000	0.130	0.530	0.585	4.1	46.8	52.8
		common	-0.116	-0.000	0.116	0.472	0.523	4.4	48.9	50.4
		entire	-0.130	-0.000	0.130	0.528	0.578	11.2	47.7	51.5
	TMLE	outcome	-0.116	-0.000	0.116	0.473	0.529	2.7	50.8	48.5
		treatment	-0.136	-0.000	0.136	0.555	0.631	-12.5	44.0	55.7
		common	-0.117	-0.000	0.117	0.476	0.538	5.2	55.9	43.4
		entire	-0.135	-0.000	0.135	0.551	0.619	-7.5	43.1	56.4
	FM	outcome	-0.116	-0.000	0.116	0.473	0.533	-8.7	44.4	55.1
		treatment	-0.131	-0.000	0.131	0.534	0.608	-23.3	36.8	62.7
		common	-0.117	-0.000	0.117	0.476	0.544	-15.2	44.2	55.4
		entire	-0.131	-0.000	0.130	0.532	0.603	-21.1	37.0	62.5
2000	GC	outcome	-0.115	0.000	0.115	0.465	0.477	-0.6	0.5	99.5
		treatment	-0.126	0.000	0.126	0.509	0.522	-0.7	0.7	99.3
		common	-0.115	0.000	0.115	0.466	0.479	-0.5	1.1	99.0
		entire	-0.126	0.000	0.126	0.508	0.520	-0.8	0.3	99.7
	IPTW	outcome	-0.115	0.000	0.115	0.465	0.477	11.7	1.1	98.9
		treatment	-0.129	0.000	0.129	0.522	0.536	4.5	1.1	98.9
		common	-0.115	0.000	0.115	0.466	0.479	4.8	1.6	98.4
		entire	-0.129	0.000	0.129	0.521	0.534	10.9	0.9	99.1
	TMLE	outcome	-0.116	0.000	0.116	0.468	0.482	4.8	2.4	97.6
		treatment	-0.135	0.000	0.135	0.546	0.565	-16.3	1.7	98.4
		common	-0.116	0.000	0.116	0.469	0.485	5.8	4.1	95.9
		entire	-0.135	0.000	0.135	0.545	0.563	-15.6	0.8	99.2
	FM	outcome	-0.115	0.000	0.115	0.465	0.481	-9.7	2.2	97.8
		treatment	-0.129	0.000	0.129	0.523	0.547	-29.2	2.2	97.8
		common	-0.115	0.000	0.115	0.466	0.484	-15.6	2.9	97.1
		entire	-0.130	0.000	0.130	0.524	0.542	-19.2	1.2	98.8

Simulation results comparing the ATT estimation under the null hypothesis in the presence of an unmeasured confounder.

n	method	set	mean bias				logOR			
			π_0	π_1	$\Delta\pi$	logOR	MSE	VEB	coverage	power
100	GC	outcome	-0.112	-0.002	0.111	0.493	0.716	-2.7	86.3	50.7
		treatment	-0.123	-0.002	0.122	0.538	0.786	-3.6	85.5	46.0
		common	-0.112	-0.002	0.111	0.492	0.735	-1.7	87.8	45.6
		entire	-0.123	-0.002	0.121	0.538	0.768	-5.5	83.5	50.7
	IPTW	outcome	-0.113	-0.002	0.111	0.494	0.727	11.7	92.7	36.6
		treatment	-0.126	-0.002	0.125	0.553	0.837	2.9	89.1	36.3
		common	-0.113	-0.002	0.111	0.494	0.743	5.6	90.8	38.4
		entire	-0.126	-0.002	0.124	0.553	0.838	6.0	90.4	33.9
	TMLE	outcome	-0.113	-0.002	0.111	0.498	0.794	16.1	94.3	26.3
		treatment	-0.129	-0.002	0.128	0.575	0.955	38.1	97.1	16.6
		common	-0.114	-0.002	0.112	0.499	0.794	19.6	93.9	24.7
		entire	-0.128	-0.002	0.126	0.570	0.953	110.0	98.4	9.7
	FM	outcome	-0.114	-0.002	0.113	0.503	0.787	-9.0	85.0	45.1
		treatment	-0.128	-0.002	0.127	0.567	0.911	-22.5	76.6	49.4
		common	-0.115	-0.002	0.113	0.504	0.795	-10.6	84.3	44.2
		entire	-0.128	-0.002	0.126	0.566	0.921	-23.8	76.2	49.6
300	GC	outcome	-0.118	0.001	0.119	0.476	0.557	-1.0	62.7	94.6
		treatment	-0.128	0.001	0.129	0.517	0.607	-1.2	63.1	91.0
		common	-0.117	0.001	0.118	0.474	0.563	0.1	67.1	91.5
		entire	-0.128	0.001	0.129	0.520	0.601	-2.6	58.9	94.5
	IPTW	outcome	-0.117	0.001	0.119	0.476	0.558	12.1	73.5	88.9
		treatment	-0.131	0.001	0.132	0.530	0.629	4.3	69.6	85.3
		common	-0.117	0.001	0.118	0.474	0.564	6.1	71.6	88.2
		entire	-0.131	0.001	0.132	0.532	0.624	10.0	71.2	85.5
	TMLE	outcome	-0.118	0.001	0.119	0.479	0.579	16.8	80.9	78.8
		treatment	-0.135	0.001	0.137	0.551	0.685	1.0	75.7	76.8
		common	-0.118	0.001	0.119	0.478	0.584	20.6	84.1	73.0
		entire	-0.136	0.001	0.137	0.552	0.676	9.7	78.6	74.6
	FM	outcome	-0.118	0.001	0.119	0.478	0.580	-6.1	66.5	88.5
		treatment	-0.131	0.001	0.132	0.532	0.658	-20.2	58.4	87.3
		common	-0.117	0.001	0.118	0.475	0.589	-11.3	65.2	86.8
		entire	-0.132	0.001	0.133	0.538	0.664	-20.6	57.7	87.7
500	GC	outcome	-0.115	-0.001	0.114	0.457	0.509	-1.5	45.4	99.4
		treatment	-0.125	-0.001	0.125	0.500	0.556	-0.5	46.3	99.0
		common	-0.115	-0.001	0.114	0.456	0.513	-0.1	50.8	98.9
		entire	-0.126	-0.001	0.125	0.501	0.552	-2.1	40.7	99.4
	IPTW	outcome	-0.115	-0.001	0.114	0.457	0.509	11.1	56.6	98.7
		treatment	-0.129	-0.001	0.128	0.513	0.574	5.0	53.3	97.7
		common	-0.115	-0.001	0.114	0.456	0.514	5.4	56.0	98.3
		entire	-0.129	-0.001	0.128	0.514	0.571	9.8	53.8	98.0
	TMLE	outcome	-0.115	-0.001	0.114	0.458	0.521	18.1	69.4	95.6
		treatment	-0.134	-0.001	0.133	0.536	0.618	-2.3	57.4	94.8
		common	-0.115	-0.001	0.114	0.459	0.527	20.5	73.6	93.2
		entire	-0.134	-0.001	0.133	0.535	0.611	1.4	57.2	95.5
	FM	outcome	-0.115	-0.001	0.114	0.458	0.525	-7.5	51.9	98.0
		treatment	-0.129	-0.001	0.129	0.517	0.597	-20.3	43.0	97.6
		common	-0.114	-0.001	0.113	0.455	0.530	-12.7	51.9	96.9
		entire	-0.129	-0.001	0.129	0.517	0.596	-19.7	43.1	97.6
2000	GC	outcome	-0.115	0.000	0.116	0.466	0.479	-1.0	0.7	100.0
		treatment	-0.126	0.000	0.126	0.510	0.524	-0.7	0.9	100.0
		common	-0.115	0.000	0.115	0.466	0.480	-0.5	1.6	100.0
		entire	-0.126	0.000	0.126	0.510	0.523	-0.7	0.3	100.0
	IPTW	outcome	-0.115	0.000	0.115	0.465	0.478	10.7	1.7	100.0
		treatment	-0.129	0.000	0.130	0.523	0.539	4.2	1.6	100.0
		common	-0.115	0.000	0.115	0.466	0.480	4.3	2.3	100.0
		entire	-0.129	0.000	0.130	0.523	0.537	10.6	1.2	100.0
	TMLE	outcome	-0.116	0.000	0.116	0.467	0.483	18.0	5.4	100.0
		treatment	-0.135	0.000	0.136	0.548	0.568	-6.6	2.6	100.0
		common	-0.116	0.000	0.116	0.469	0.486	18.6	8.3	100.0
		entire	-0.135	0.000	0.135	0.546	0.565	-4.6	1.7	100.0
	FM	outcome	-0.115	0.000	0.116	0.466	0.484	-8.4	2.9	100.0
		treatment	-0.130	0.000	0.130	0.524	0.548	-26.4	3.0	100.0
		common	-0.115	0.000	0.115	0.466	0.485	-13.9	3.7	100.0
		entire	-0.130	0.000	0.130	0.525	0.543	-16.3	1.4	100.0

Simulation results comparing the ATT estimation under the alternative hypothesis in the presence of an unmeasured confounder.

n	method	set	mean bias				logOR			
			π_0	π_1	$\Delta\pi$	logOR	MSE	VEB	coverage	type I
100	GC	outcome	-0.058	0.057	0.115	0.479	0.628	-6.4	74.9	23.4
		treatment	-0.065	0.065	0.131	0.546	0.716	-7.7	74.3	23.3
		common	-0.058	0.057	0.115	0.479	0.643	-4.9	78.0	21.3
		entire	-0.065	0.066	0.131	0.547	0.704	-10.0	70.8	26.0
	IPTW	outcome	-0.058	0.057	0.115	0.481	0.638	9.6	85.0	15.0
		treatment	-0.065	0.065	0.130	0.548	0.754	-1.6	80.4	19.6
		common	-0.058	0.057	0.115	0.480	0.650	3.1	83.1	16.9
		entire	-0.065	0.065	0.130	0.550	0.753	1.7	81.8	18.1
	TMLE	outcome	-0.058	0.057	0.115	0.479	0.635	-6.5	75.5	24.4
		treatment	-0.065	0.064	0.130	0.548	0.746	-11.8	74.2	25.8
		common	-0.058	0.057	0.115	0.480	0.649	-4.1	79.4	20.6
		entire	-0.065	0.065	0.130	0.549	0.735	-16.9	68.7	31.3
	FM	outcome	-0.058	0.056	0.115	0.483	0.687	-15.1	75.3	24.7
		treatment	-0.068	0.067	0.135	0.575	0.823	-28.8	65.1	34.9
		common	-0.060	0.057	0.116	0.491	0.707	-18.3	74.1	25.9
		entire	-0.068	0.067	0.135	0.574	0.818	-28.0	65.3	34.7
300	GC	outcome	-0.056	0.056	0.113	0.462	0.513	-1.4	45.0	55.1
		treatment	-0.064	0.064	0.128	0.528	0.586	-1.5	44.8	55.1
		common	-0.056	0.056	0.112	0.461	0.518	-0.6	50.8	49.7
		entire	-0.064	0.064	0.128	0.528	0.580	-2.6	38.4	61.1
	IPTW	outcome	-0.057	0.056	0.113	0.462	0.515	12.7	57.0	43.0
		treatment	-0.064	0.063	0.127	0.524	0.592	4.0	54.9	45.1
		common	-0.056	0.056	0.112	0.461	0.520	5.9	56.5	43.6
		entire	-0.064	0.063	0.127	0.525	0.587	9.6	56.0	44.1
	TMLE	outcome	-0.057	0.056	0.112	0.462	0.514	-1.2	45.9	54.2
		treatment	-0.064	0.063	0.127	0.522	0.589	-3.0	49.8	50.3
		common	-0.056	0.056	0.112	0.461	0.519	-0.2	51.8	48.3
		entire	-0.064	0.063	0.127	0.522	0.583	-4.6	44.1	56.0
	FM	outcome	-0.057	0.056	0.113	0.465	0.534	-10.3	50.2	49.9
		treatment	-0.064	0.064	0.129	0.532	0.624	-27.4	41.2	58.9
		common	-0.056	0.056	0.112	0.464	0.562	-25.4	49.8	50.3
		entire	-0.065	0.064	0.129	0.534	0.622	-25.6	41.2	58.9
500	GC	outcome	-0.057	0.056	0.113	0.462	0.494	-2.1	23.2	76.8
		treatment	-0.065	0.064	0.129	0.528	0.564	-2.0	22.6	77.3
		common	-0.057	0.056	0.113	0.462	0.498	-1.9	29.1	71.2
		entire	-0.065	0.064	0.129	0.527	0.559	-2.5	17.5	82.1
	IPTW	outcome	-0.057	0.056	0.113	0.462	0.494	11.3	33.3	66.6
		treatment	-0.065	0.063	0.128	0.524	0.565	3.2	32.4	67.4
		common	-0.057	0.056	0.113	0.462	0.498	4.3	33.8	65.9
		entire	-0.065	0.063	0.128	0.524	0.562	8.9	32.3	67.5
	TMLE	outcome	-0.057	0.056	0.113	0.462	0.494	-1.6	24.0	75.8
		treatment	-0.065	0.063	0.127	0.523	0.563	-2.6	27.9	71.9
		common	-0.057	0.056	0.113	0.462	0.498	-1.4	29.6	70.1
		entire	-0.064	0.063	0.127	0.522	0.559	-3.2	22.4	77.4
	FM	outcome	-0.057	0.056	0.113	0.461	0.507	-13.1	31.2	68.6
		treatment	-0.065	0.064	0.128	0.529	0.589	-29.7	25.2	74.7
		common	-0.057	0.056	0.113	0.463	0.540	-34.1	35.4	64.5
		entire	-0.065	0.063	0.128	0.528	0.582	-25.3	23.8	76.1
2000	GC	outcome	-0.057	0.056	0.113	0.460	0.468	-0.9	0.0	100.0
		treatment	-0.065	0.064	0.128	0.525	0.534	-0.7	0.0	100.0
		common	-0.057	0.056	0.113	0.460	0.469	-0.7	0.1	99.9
		entire	-0.064	0.064	0.128	0.525	0.533	-0.9	0.0	100.0
	IPTW	outcome	-0.057	0.056	0.113	0.459	0.468	12.1	0.1	99.9
		treatment	-0.064	0.063	0.127	0.520	0.530	4.6	0.2	99.8
		common	-0.057	0.056	0.113	0.459	0.468	5.0	0.2	99.8
		entire	-0.064	0.063	0.127	0.520	0.529	11.0	0.0	100.0
	TMLE	outcome	-0.057	0.056	0.113	0.459	0.467	-0.8	0.0	100.0
		treatment	-0.064	0.063	0.127	0.518	0.528	-1.0	0.1	99.9
		common	-0.057	0.056	0.113	0.459	0.468	-0.7	0.1	99.9
		entire	-0.064	0.063	0.127	0.518	0.527	-0.9	0.0	100.0
	FM	outcome	-0.057	0.056	0.112	0.459	0.479	-34.3	1.9	98.1
		treatment	-0.064	0.063	0.127	0.521	0.549	-48.2	2.5	97.5
		common	-0.056	0.056	0.113	0.462	0.515	-60.0	10.6	89.4
		entire	-0.064	0.063	0.127	0.518	0.532	-24.9	0.3	99.7

Simulation results comparing the ATE estimation under the null hypothesis in the presence of an unmeasured confounder.

n	method	set	mean bias				logOR			
			π_0	π_1	$\Delta\pi$	logOR	MSE	VEB	coverage	power
100	GC	outcome	-0.056	0.057	0.113	0.478	0.634	-6.9	76.2	67.5
		treatment	-0.065	0.065	0.129	0.551	0.727	-7.9	75.5	62.2
		common	-0.057	0.057	0.113	0.479	0.650	-5.2	79.2	62.8
		entire	-0.065	0.065	0.129	0.550	0.714	-10.6	71.6	67.0
	IPTW	outcome	-0.057	0.056	0.113	0.480	0.646	8.1	85.4	54.3
		treatment	-0.064	0.065	0.128	0.556	0.769	-2.4	80.9	51.8
		common	-0.057	0.056	0.113	0.482	0.657	2.9	84.2	55.5
		entire	-0.063	0.065	0.128	0.556	0.773	-0.7	82.0	50.1
	TMLE	outcome	-0.056	0.056	0.113	0.478	0.642	-7.0	77.3	66.5
		treatment	-0.063	0.064	0.128	0.552	0.757	-12.2	75.1	60.0
		common	-0.057	0.056	0.113	0.481	0.657	-4.1	80.6	60.8
		entire	-0.063	0.065	0.128	0.553	0.748	-17.6	69.6	66.6
	FM	outcome	-0.058	0.057	0.115	0.497	0.710	-16.7	75.2	62.1
		treatment	-0.065	0.066	0.130	0.575	0.832	-29.0	66.6	64.9
		common	-0.056	0.057	0.114	0.491	0.713	-18.5	75.0	61.4
		entire	-0.065	0.066	0.131	0.577	0.830	-28.4	66.2	65.3
300	GC	outcome	-0.057	0.058	0.115	0.480	0.532	-1.9	43.5	98.7
		treatment	-0.065	0.066	0.131	0.548	0.608	-2.8	42.9	97.7
		common	-0.057	0.058	0.115	0.480	0.539	-1.9	48.7	97.8
		entire	-0.065	0.066	0.131	0.547	0.600	-2.9	37.1	98.9
	IPTW	outcome	-0.057	0.058	0.115	0.480	0.533	12.4	55.3	97.6
		treatment	-0.064	0.066	0.130	0.546	0.615	3.0	53.1	94.2
		common	-0.057	0.058	0.115	0.481	0.541	4.7	54.4	96.6
		entire	-0.064	0.066	0.130	0.545	0.607	9.6	54.8	94.8
	TMLE	outcome	-0.057	0.058	0.115	0.480	0.532	-1.0	44.9	98.7
		treatment	-0.064	0.066	0.130	0.545	0.612	-3.3	48.0	96.2
		common	-0.057	0.058	0.115	0.481	0.540	-1.1	50.1	97.6
		entire	-0.064	0.066	0.130	0.543	0.603	-3.4	42.8	97.9
	FM	outcome	-0.057	0.058	0.115	0.481	0.552	-11.9	48.1	96.6
		treatment	-0.065	0.066	0.132	0.556	0.652	-29.0	40.2	95.3
		common	-0.057	0.058	0.115	0.486	0.588	-27.4	47.6	93.1
		entire	-0.065	0.067	0.132	0.557	0.644	-25.5	39.1	96.0
300	GC	outcome	-0.057	0.056	0.113	0.469	0.501	-1.0	23.0	100.0
		treatment	-0.065	0.064	0.129	0.536	0.573	-2.0	22.8	99.9
		common	-0.057	0.056	0.113	0.468	0.505	-1.0	29.0	99.9
		entire	-0.065	0.064	0.129	0.536	0.569	-1.9	17.5	100.0
	IPTW	outcome	-0.057	0.056	0.113	0.469	0.501	12.6	32.9	100.0
		treatment	-0.064	0.064	0.127	0.531	0.573	3.6	32.9	99.6
		common	-0.057	0.056	0.113	0.468	0.505	5.1	33.7	99.9
		entire	-0.064	0.064	0.127	0.532	0.569	10.2	32.6	99.6
	TMLE	outcome	-0.057	0.056	0.113	0.468	0.501	-0.4	23.7	100.0
		treatment	-0.064	0.063	0.127	0.530	0.571	-2.1	28.0	99.8
		common	-0.057	0.056	0.113	0.468	0.505	-0.5	29.3	99.9
		entire	-0.064	0.064	0.127	0.530	0.566	-1.9	22.9	99.9
	FM	outcome	-0.057	0.056	0.113	0.471	0.518	-14.2	30.3	99.8
		treatment	-0.064	0.064	0.128	0.537	0.598	-29.3	25.0	99.5
		common	-0.057	0.056	0.113	0.475	0.555	-35.3	34.9	98.4
		entire	-0.064	0.064	0.128	0.537	0.592	-25.2	23.4	99.6
300	GC	outcome	-0.058	0.057	0.115	0.474	0.482	-1.3	0.0	100.0
		treatment	-0.066	0.065	0.130	0.541	0.550	-1.3	0.0	100.0
		common	-0.058	0.057	0.115	0.474	0.483	-0.7	0.1	100.0
		entire	-0.066	0.065	0.131	0.541	0.550	-1.7	0.0	100.0
	IPTW	outcome	-0.058	0.057	0.115	0.474	0.482	11.3	0.1	100.0
		treatment	-0.065	0.064	0.129	0.536	0.547	3.7	0.1	100.0
		common	-0.058	0.057	0.114	0.473	0.483	4.9	0.1	100.0
		entire	-0.065	0.064	0.129	0.537	0.546	9.3	0.1	100.0
	TMLE	outcome	-0.058	0.057	0.115	0.474	0.482	-1.1	0.0	100.0
		treatment	-0.065	0.064	0.129	0.534	0.545	-1.6	0.1	100.0
		common	-0.058	0.057	0.114	0.473	0.483	-0.7	0.1	100.0
		entire	-0.065	0.064	0.129	0.535	0.544	-2.1	0.0	100.0
	FM	outcome	-0.058	0.057	0.115	0.475	0.497	-36.3	2.0	100.0
		treatment	-0.065	0.064	0.129	0.539	0.569	-49.2	2.2	100.0
		common	-0.058	0.057	0.115	0.479	0.534	-60.7	9.7	100.0
		entire	-0.065	0.064	0.129	0.535	0.549	-26.0	0.2	100.0

Simulation results comparing the ATE estimation under the alternative hypothesis in the presence of an unmeasured confounder.

	Relapse at 1 year	Fist-line treatment allocation*	Both	Nothing
Female patient				X
At least one relapse 1 year before treatment initiation			X	
Gd-enhancing lesion on MRI	X			
EDSS score >3			X	
Previous immunomodulatory treatment				X
Patient age	X			
Disease duration	X			

Classification of covariates in the multiple sclerosis application according to an expert knowledge. *: No covariate is only associated with the treatment allocation, treatment and common sets are the same.

	Favourable GOS at 3 months	Barbiturates allocation	Both	Nothing
Female patient				X
Diabetes			X	
Nosological entity: Severe trauma			X	
SAP $\leq$ 90 mmHg before admission*	X			
Evacuation of subdural or extradural hematoma		X		
External ventricular drain		X		
Evacuation of cerebral hematoma or lobectomy			X	
Decompressive craniectomy			X	
Blood transfusion before admission	X			
Pneumonia before increased ICP			X	
Osmotherapy	X			
GCS score $\geq$ 8*			X	
Patient age*			X	
Haemoglobin	X			
Platelets				X
Serum creatinine				X
Arterial pH			X	
Serum proteins				X
Serum urea*				X
PaO2/FiO2 ratio			X	
SAPS II score			X	

Classification of covariates in the intensive care unit application according to an expert knowledge. *: No include into a covariate set due to the association with the SAPS II Score.

Set	Characteristics	Overall		Fingolimod		Natalizumab		STD (%)
Outcome	At least one relapse (n, %)	526.9	83.8	273.4	83.9	253.5	83.7	0.5
	EDSS score $>$ 3 (n, %)	286.2	45.5	148.6	45.6	137.6	45.5	0.3
	Gd-enhancing lesion on MRI (n, %)	310.0	49.3	161.1	49.4	148.9	49.2	0.5
	Patient age, years (mean, sd)	37.1	9.6	37.2	9.9	37.1	9.3	1.2
	Disease duration, years (mean, sd)	8.5	6.4	8.6	6.2	8.5	6.5	0.2
Treatment*	At least one relapse (n, %)	526.2	83.7	272.7	83.7	253.5	83.6	0.1
	EDSS score $>$ 3 (n, %)	288.5	45.9	149.5	45.9	139.0	45.8	0.1
Common*	At least one relapse (n, %)	526.2	83.7	272.7	83.7	253.5	83.6	0.1
	EDSS score $>$ 3 (n, %)	288.5	45.9	149.5	45.9	139.0	45.8	0.1
Entire	Female patient (n, %)	481.4	76.6	250.2	76.8	231.2	76.4	0.9
	At least one relapse (n, %)	527.8	84.0	274.3	84.2	253.5	83.8	1.1
	Gd-enhancing lesion on MRI (n, %)	308.8	49.1	160.4	49.2	148.4	49.0	0.4
	EDSS score $>$ 3 (n, %)	285.2	45.4	148.1	45.5	137.1	45.3	0.3
	Previous IMT (n, %)	555.7	88.4	288.2	88.5	267.5	88.4	0.2
	Patient age, years (mean, sd)	37.1	9.7	37.2	10.0	37.1	9.3	1.3
	Disease duration, years (mean, sd)	8.6	6.4	8.6	6.4	8.6	6.4	0.6

The PS-adjusted samples for weighted analysis of the relapsing-remitting multiple sclerosis relapse in the year after the treatment initiation according to the covariate sets. Qualitative characteristics are presented by using the weighted effective (n) and the weighted percentage. Continuous characteristics are presented with weighted mean following by weighted standard deviation (sd). STD: Standardised differences in %, EDSS: Expanded Disability Status Scale, Gd: Gadolinium, MRI: Magnetic Resonance Imaging, and IMT: immunomodulatory treatment. *: No covariate is only associated with the treatment allocation, treatment and common sets are the same.

Set	Characteristics	Overall		Fingolimod		Natalizumab		STD (%)
Outcome	At least one relapse (n, %)	481.1	76.5	248.1	76.1	233.0	76.9	1.8
	EDSS score $>$ 3 (n, %)	236.2	37.6	114.2	35.0	122.0	40.3	10.8
	Gd-enhancing lesion on MRI (n, %)	271.6	43.2	145.6	44.7	126.0	41.6	6.2
	Patient age, years (mean, sd)	36.6	9.6	36.0	9.8	37.2	9.2	12.2
	Disease duration, years (mean, sd)	8.6	6.3	8.1	5.9	9.0	6.8	14.4
Treatment*	At least one relapse (n, %)	483.7	76.9	250.7	76.9	233.0	76.9	0.0
	EDSS score $>$ 3 (n, %)	253.3	40.3	131.3	40.3	122.0	40.3	0.0
Common*	At least one relapse (n, %)	483.7	76.9	250.7	76.9	233.0	76.9	0.0
	EDSS score $>$ 3 (n, %)	253.3	40.3	131.3	40.3	122.0	40.3	0.0
Entire	Female patient (n, %)	483.8	76.9	258.8	79.4	225.0	74.3	12.2
	At least one relapse (n, %)	475.9	75.7	242.9	74.5	233.0	76.9	5.5
	Gd-enhancing lesion on MRI (n, %)	266.5	42.4	140.5	43.1	126.0	41.6	3.1
	EDSS score $>$ 3 (n, %)	255.8	40.7	133.8	41.1	122.0	40.3	1.6
	Previous IMT (n, %)	545.6	86.7	282.6	86.7	263.0	86.8	0.3
	Patient age, years (mean, sd)	37.1	9.3	37.0	9.3	37.2	9.2	1.6
	Disease duration, years (mean, sd)	8.9	6.5	8.8	6.3	9.0	6.8	3.5

The PS-adjusted samples for matched (FM) analysis of the relapsing-remitting multiple sclerosis relapse in the year after the treatment initiation according to the covariate sets. Qualitative characteristics are presented by using the matched effective (n) and the matched percentage. Continuous characteristics are presented with matched mean following by matched standard deviation (sd). STD: Standardised differences in %, EDSS: Expanded Disability Status Scale, Gd: Gadolinium, MRI: Magnetic Resonance Imaging, and IMT: immunomodulatory treatment. *: No covariate is only associated with the treatment allocation, treatment and common sets are the same.

Set	Characteristics	Overall		Barbiturates		Control		STD (%)
Outcome	Diabetes (n, %)	4.1	2.7	2.1	2.7	2.0	2.7	0.2
	Evacuation of cerebral hematoma or lobectomy (n, %)	29.2	19.6	15.2	20.3	14.0	18.9	3.4
	Decompressive craniectomy (n, %)	26.1	17.5	14.1	18.8	12.0	16.2	6.9
	Blood transfusion before admission (n, %)	17.0	11.4	8.0	10.7	9.0	12.2	4.6
	Osmotherapy (n, %)	74.2	49.9	37.2	49.8	37.0	50.0	0.4
	Pneumonia before increased ICP (n, %)	26.3	17.7	13.3	17.8	13.0	17.6	0.7
	Nosological entity: Severe trauma (n, %)	56.5	38.0	27.5	36.8	29.0	39.2	4.9
	Haemoglobin, g/dL (mean, sd)	12.1	2.3	12.2	2.2	12.1	2.5	0.6
	Arterial pH (mean, sd)	7.3	0.1	7.3	0.1	7.3	0.1	2.2
	PaO2/FiO2 ratio (mean, sd)	322.3	193.0	317.9	172.2	326.6	212.9	4.5
	SAPS II score (mean, sd)	47.9	11.5	48.3	10.1	47.6	12.9	6.3
Treatment	Diabetes (n, %)	4.0	2.7	2.0	2.7	2.0	2.7	0.3
	Evacuation of subdural or extradural hematoma (n, %)	16.2	10.8	8.2	10.9	8.0	10.8	0.2
	External ventricular drain (n, %)	52.1	34.9	27.1	35.9	25.0	33.8	4.5
	Evacuation of cerebral hematoma or lobectomy (n, %)	28.8	19.3	14.8	19.6	14.0	18.9	1.8
	Decompressive craniectomy (n, %)	25.9	17.4	13.9	18.5	12.0	16.2	6.1
	Pneumonia before increased ICP (n, %)	27.0	18.1	14.0	18.5	13.0	17.6	2.5
	Nosological entity: Severe trauma (n, %)	57.1	38.2	28.1	37.3	29.0	39.2	3.9
	Arterial pH (mean, sd)	7.3	0.1	7.3	0.1	7.3	0.1	1.9
	PaO2/FiO2 ratio (mean, sd)	323.3	195.9	320.1	178.8	326.6	212.9	3.3
	SAPS II score (mean, sd)	47.8	11.5	48.0	10.0	47.6	12.9	4.2
Common	Diabetes (n, %)	4.0	2.7	2.0	2.7	2.0	2.7	0.1
	Evacuation of cerebral hematoma or lobectomy (n, %)	28.9	19.4	14.9	19.8	14.0	18.9	2.3
	Decompressive craniectomy (n, %)	25.3	17.0	13.3	17.8	12.0	16.2	4.1
	Pneumonia before increased ICP (n, %)	26.5	17.8	13.5	18.0	13.0	17.6	1.2
	Nosological entity: Severe trauma (n, %)	57.2	38.4	28.2	37.6	29.0	39.2	3.3
	Arterial pH (mean, sd)	7.3	0.1	7.3	0.1	7.3	0.1	3.6
	PaO2/FiO2 ratio (mean, sd)	323.8	192.8	321.1	172.0	326.6	212.9	2.9
	SAPS II score (mean, sd)	47.9	11.6	48.3	10.1	47.6	12.9	6.7
Entire	Female patients (n, %)	66.7	44.8	35.7	47.7	31.0	41.9	11.8
	Diabetes (n, %)	4.0	2.7	2.0	2.7	2.0	2.7	0.2
	Evacuation of subdural or extradural hematoma (n, %)	15.4	10.3	7.4	9.9	8.0	10.8	3.0
	External ventricular drain (n, %)	52.3	35.2	27.3	36.5	25.0	33.8	5.8
	Evacuation of cerebral hematoma or lobectomy (n, %)	30.6	20.5	16.6	22.2	14.0	18.9	8.0
	Decompressive craniectomy (n, %)	27.1	18.2	15.1	20.2	12.0	16.2	10.3
	Blood transfusion before admission (n, %)	16.6	11.1	7.6	10.1	9.0	12.2	6.4
	Osmotherapy (n, %)	75.3	50.6	38.3	51.2	37.0	50.0	2.4
	Pneumonia before increased ICP (n, %)	26.2	17.6	13.2	17.6	13.0	17.6	0.1
	Nosological entity: Severe trauma (n, %)	56.0	37.6	27.0	36.1	29.0	39.2	6.4
	Haemoglobin, g/dL (mean, sd)	12.2	2.3	12.2	2.2	12.1	2.5	4.5
	Platelets, counts/mm^3 (mean, sd)	208.8	82.7	212.5	90.7	205.1	74.2	9.0
	Serum creatinine, mmol/L (mean, sd)	70.6	30.9	70.2	28.6	71.1	33.3	2.7
	Arterial pH (mean, sd)	7.3	0.1	7.3	0.1	7.3	0.1	0.5
	Serum proteins, g/L (mean, sd)	59.4	9.5	59.2	9.4	59.6	9.7	4.2
	PaO2/FiO2 ratio (mean, sd)	317.9	196.8	309.3	180.5	326.6	212.9	8.8
	SAPS II score (mean, sd)	47.9	11.5	48.2	10.1	47.6	12.9	5.5

The PS-adjusted samples for weighted analysis of dichotomised Glasgow Outcome Scale score at 3 months according to the covariate sets. Qualitative characteristics are presented by using the weighted effective (n) and the weighted percentage. Continuous characteristics are presented with weighted mean following by weighted standard deviation (sd). STD: Standardised differences in %, SAPS: Simplified Acute Physiology Score, ICP: Intracranial Pressure, PaO2: arterial partial Pressure of Oxygen, and FiO2: Fraction of Inspired Oxygen.

Set	Characteristics	Overall		Barbiturates		Control		STD (%)
Outcome	Diabetes (n, %)	5.7	2.3	3.7	2.1	2.0	2.7	4.0
	Evacuation of cerebral hematoma or lobectomy (n, %)	42.2	16.7	28.2	15.8	14.0	18.9	8.1
	Decompressive craniectomy (n, %)	39.5	15.7	27.5	15.5	12.0	16.2	2.0
	Blood transfusion before admission (n, %)	28.2	11.2	19.2	10.8	9.0	12.2	4.3
	Osmotherapy (n, %)	121.8	48.3	84.8	47.7	37.0	50.0	4.7
	Pneumonia before increased ICP (n, %)	45.9	18.2	32.9	18.5	13.0	17.6	2.3
	Nosological entity: Severe trauma (n, %)	95.0	37.7	66.0	37.1	29.0	39.2	4.4
	Haemoglobin, g/dL (mean, sd)	12.1	2.3	12.1	2.3	12.1	2.5	2.1
	Arterial pH (mean, sd)	7.4	0.1	7.4	0.1	7.3	0.1	7.6
	PaO2/FiO2 ratio (mean, sd)	313.1	174.0	307.4	155.4	326.6	212.9	10.3
	SAPS II score (mean, sd)	48.4	11.1	48.7	10.3	47.6	12.9	9.8
Treatment	Diabetes (n, %)	5.1	2.0	3.1	1.7	2.0	2.7	6.6
	Evacuation of subdural or extradural hematoma (n, %)	28.2	11.2	20.2	11.4	8.0	10.8	1.7
	External ventricular drain (n, %)	74.0	29.4	49.0	27.5	25.0	33.8	13.6
	Evacuation of cerebral hematoma or lobectomy (n, %)	43.5	17.3	29.5	16.6	14.0	18.9	6.1
	Decompressive craniectomy (n, %)	45.6	18.1	33.6	18.9	12.0	16.2	6.9
	Pneumonia before increased ICP (n, %)	41.5	16.5	28.5	16.0	13.0	17.6	4.2
	Nosological entity: Severe trauma (n, %)	97.1	38.5	68.1	38.3	29.0	39.2	1.9
	Arterial pH (mean, sd)	7.4	0.1	7.4	0.1	7.3	0.1	12.4
	PaO2/FiO2 ratio (mean, sd)	318.2	180.0	314.7	165.0	326.6	212.9	6.3
	SAPS II score (mean, sd)	46.6	10.9	46.2	10.0	47.6	12.9	11.6
Common	Diabetes (n, %)	5.1	2.0	3.1	1.8	2.0	2.7	6.4
	Evacuation of cerebral hematoma or lobectomy (n, %)	51.6	20.5	37.6	21.1	14.0	18.9	5.5
	Decompressive craniectomy (n, %)	38.7	15.4	26.7	15.0	12.0	16.2	3.3
	Pneumonia before increased ICP (n, %)	42.7	16.9	29.7	16.7	13.0	17.6	2.4
	Nosological entity: Severe trauma (n, %)	57.2	38.4	28.2	37.6	29.0	39.2	3.3
	Arterial pH (mean, sd)	7.4	0.1	7.4	0.1	7.3	0.1	9.4
	PaO2/FiO2 ratio (mean, sd)	319.3	185.0	316.3	172.8	326.6	212.9	5.4
	SAPS II score (mean, sd)	48.2	10.8	48.4	9.9	47.6	12.9	7.4
Entire	Female patients (n, %)	113.6	45.1	82.6	46.4	31.0	41.9	9.1
	Diabetes (n, %)	8.0	3.2	6.0	3.4	2.0	2.7	3.8
	Evacuation of subdural or extradural hematoma (n, %)	23.5	9.3	15.5	8.7	8.0	10.8	7.1
	External ventricular drain (n, %)	104.0	41.3	79.0	44.4	25.0	33.8	21.9
	Evacuation of cerebral hematoma or lobectomy (n, %)	44.3	17.6	30.3	17.0	14.0	18.9	5.0
	Decompressive craniectomy (n, %)	46.4	18.4	34.4	19.3	12.0	16.2	8.2
	Blood transfusion before admission (n, %)	26.5	10.5	17.5	9.8	9.0	12.2	7.5
	Osmotherapy (n, %)	109.6	43.5	72.6	40.8	37.0	50.0	18.6
	Pneumonia before increased ICP (n, %)	45.5	18.1	32.5	18.3	13.0	17.6	1.8
	Nosological entity: Severe trauma (n, %)	77.1	30.6	48.1	27.0	29.0	39.2	26.0
	Haemoglobin, g/dL (mean, sd)	12.5	2.4	12.7	2.3	12.1	2.5	23.8
	Platelets, counts/mm^3 (mean, sd)	218.5	89.8	224.1	95.5	205.1	74.2	22.3
	Serum creatinine, mmol/L (mean, sd)	69.7	30.0	69.1	28.6	71.1	33.3	6.4
	Arterial pH (mean, sd)	7.3	0.1	7.3	0.1	7.3	0.1	4.7
	Serum proteins, g/L (mean, sd)	59.4	8.9	59.4	8.6	59.6	9.7	3.1
	PaO2/FiO2 ratio (mean, sd)	308.1	187.3	300.3	175.9	326.6	212.9	13.5
	SAPS II score (mean, sd)	48.3	11.1	48.7	10.2	47.6	12.9	9.7

The PS-adjusted samples for matched analysis (FM) of dichotomised Glasgow Outcome Scale score at 3 months according to the covariate sets. Qualitative characteristics are presented by using the matched effective (n) and the matched percentage. Continuous characteristics are presented with weighted mean following by matched standard deviation (sd). STD: Standardised differences in %, SAPS: Simplified Acute Physiology Score, ICP: Intracranial Pressure, PaO2: arterial partial Pressure of Oxygen, and FiO2: Fraction of Inspired Oxygen.

Annexe E

Éléments supplémentaires au chapitre 4

Supplementary materials of the manuscript by Le Borgne et al. entitled "G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes"

Florent Le Borgne[1,2], Arthur Chatton[1,2], Maxime Léger[1,3], Rémi Lenain[1,4], and Yohann Foucher[1,5]

[1] INSERM UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France.

[2] IDBC-A2COM, Pacé, France.

[3] Département d'Anesthésie Réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France.

[4] Lille University Hospital, Lille, France

[5] Nantes University Hospital, Nantes, France.

Table of contents:

Definition of the criteria reported in the simulations.

Let $\hat{\theta}_k$ and $\widehat{sd}(\hat{\theta}_k)$ the average causal effect (ACE) and its standard deviation estimated in the kth simulated data set and $\bar{\theta}$ the true ACE $\bar{\theta}$ ($k = 1, \ldots, 1000$). The criteria used in our simulation study are the following:

i) The mean bias (MB) : $\frac{1}{1000}\sum_{k=1}^{1000}(\hat{\theta}_k - \bar{\theta}) * 100$

ii) The root mean square error (RMSE) : $\sqrt{\frac{1}{1000}\sum_{k=1}^{1000}(\hat{\theta}_k - \bar{\theta})^2}$

iii) The empirical standard deviation (ESD): $\sqrt{\frac{1}{999}\sum_{k=1}^{1000}(\hat{\theta}_k - \bar{\theta})^2}$

iv) The asymptotic standard deviation (ASD): $\frac{1}{1000}\sum_{k=1}^{1000}\widehat{sd}(\hat{\theta}_k)$

v) The variance estimation bias: $100 * (ASD - ESD)/ESD$

vi) The empirical coverage rate of the nominal 95% confidence interval (95%CI): $\frac{1}{1000}\sum_{k=1}^{1000}(\mathrm{I}(95\%CI_{inf,k} \leq \theta \leq 95\%CI_{sup,k}))$, where $95\%CI_{inf,k}$ and $95\%CI_{sup,k}$ are the lower and upper bounds of the 95%CI estimated by bootstrap in the kth simulated data set, respectively.

vii) The statistical power: $\frac{1}{1000}\sum_{k=1}^{1000}(\mathrm{I}(95\%CI_{inf,k} > 0) + \mathrm{I}(95\%CI_{sup,k} < 0))$

Table S1. Models used for simulations in the realistic situation (Figure 1A in the main text).

Distribution	Linear predictor
X_1 ~ Gaussian	0
X_2 ~ Gaussian	$\beta_0 + \beta_1 * X_1$
X_3 ~ Gaussian	$\beta_0 - \beta_1 * X_1 - \beta_1 * X_2$
X_4 ~ Gaussian	$\beta_0 + \beta_1 * X_3$
X_5 ~ Gaussian	0
$\tilde{X}_6$ ~ Gaussian	0
X_6 ~ Bernouilli	1 if $\tilde{X}_6$ > 0.66 and 0 otherwise (prevalence ~ 25%)
$\tilde{X}_7$ ~ Gaussian	$\beta_0 - \beta_1 * X_5$
X_7 ~ Bernouilli	1 if $\tilde{X}_7$ > -0.40 and 0 otherwise (prevalence ~ 40%)
X_8 ~ Gaussian	$\beta_0 - \beta_1 * X_6$
$\tilde{X}_9$ ~ Gaussian	$\beta_0 + \beta_1 * X_7$
X_9 ~ Bernouilli	1 if $\tilde{X}_9$ > -0.80 and 0 otherwise (prevalence ~ 75%)
X_{10} ~ Gaussian	$\beta_0 + \beta_1 * X_8$
X_{11} ~ Gaussian	0
$\tilde{X}_{12}$ ~ Gaussian	$\beta_0 + \beta_1 * X_9$
X_{12} ~ Bernouilli	1 if $\tilde{X}_{12}$ > 0.84 and 0 otherwise (prevalence ~ 25%)
$\tilde{X}_{13}$ ~ Gaussian	$\beta_0 + \beta_1 * X_{10}$
X_{13} ~ Bernouilli	1 if $\tilde{X}_{13}$ > -0.09 and 0 otherwise (prevalence ~ 50%)
X_{14} ~ Gaussian	$\beta_0 - \beta_1 * X_{12} - \beta_1 * X_{11}$
X_{15} ~ Gaussian	$\beta_0 - \beta_1 * X_{12}$
$\tilde{X}_{16}$ ~ Gaussian	0
X_{16} ~ Bernouilli	1 if $\tilde{X}_{16}$ > -0.66 and 0 otherwise (prevalence ~ 75%)
$\tilde{X}_{17}$ ~ Gaussian	$\beta_0 - \beta_1 * X_{16}$
X_{17} ~ Bernouilli	1 if $\tilde{X}_{17}$ > -0.92 and 0 otherwise (prevalence ~ 50%)
X_{18} ~ Gaussian	0
$\tilde{X}_{19}$ ~ Gaussian	0
X_{19} ~ Bernouilli	1 if $\tilde{X}_{19}$ > 0.66 and 0 otherwise (prevalence ~ 25%)
$\tilde{X}_{20}$ ~ Gaussian	0
X_{20} ~ Bernouilli	1 if $\tilde{X}_{20}$ > 0.66 and 0 otherwise (prevalence ~ 25%)
X_{21} ~ Gaussian	0
$\tilde{X}_{22}$ ~ Gaussian	0
X_{22} ~ Bernouilli	1 if $\tilde{X}_{22}$ > 0.66 and 0 otherwise (prevalence ~ 25%)
Z ~ Bernouilli	$\beta_0 + \beta_1 * X_1 - \beta_1 * X_3 + \beta_1 * X_5 - \beta_1 * X_7 + \beta_1 * X_9 - \beta_1 * X_{11} + \beta_1 * X_{13} - \beta_1 * X_{15} - \beta_1 * X_{17} + \beta_1 * X_{19} - \beta_1 * X_{21}$
Y ~ Bernouilli	$-1.1 + \beta_1 * I(X_2 > -0.40) - \beta_1 * X_3 + (\beta_1/2) * X_3^2 + \beta_1 * X_6 + \beta_1 * X_7 + \beta_1 * X_{10} + \beta_1 * 0.5 * X_{11}^2 - \beta_1 * X_{14} - \beta_1 * I(X_{15} > -0.57) + \beta_1 * X_{18} + \beta_1 * X_{19} + \beta_1 * Z + \beta_1 * 0.5 * Z * X_{18}$

For the Gaussian distributions, the standard errors were 1 and the link function with the linear predictor was the identity function. For the Bernoulli distribution, the link function with the linear predictor was the logit function. $\beta_0 = -0.4$, $\beta_1 = log(2.00)$, $I(a) = 1$ if a is true and 0 otherwise.

Table S2. Models used for simulations in the simplistic situation (Figure 1B in the main text).

Distribution	Linear predictor
X_1 ~ Bernouilli	0
X_2 ~ Bernouilli	0
X_3 ~ Gaussian	0
X_4 ~ Bernouilli	0
X_5 ~ Bernouilli	0
X_6 ~ Gaussian	0
X_7 ~ Bernouilli	0
X_8 ~ Bernouilli	0
X_9 ~ Gaussian	0
Z ~ Bernouilli	$-0.8 + \beta_2 * X_1 + \beta_1 * X_2 - \beta_2 * X_4 - \beta_1 * X_5 + \beta_2 * X_7 + \beta_1 * X_8$
Y ~ Bernouilli	$-0.8 + \beta_Z * Z + \beta_2 * X_1 - \beta_2 * X_2 - \beta_2 * X_3 + \beta_1 * X_4 - \beta_1 * X_5 + \beta_1 * X_6$

For the Gaussian distributions, the standard errors were 1 and the link function with the linear predictor was the identity function. For the Bernoulli distribution, the link function with the linear predictor was the logit function. $\beta_1 = log(1.50)$, $\beta_2 = log(3.00)$, and $\beta_Z = log(1.75)$.

Table S3. Performances of G-computation in a realistic situation with the following Q-models: the theoretical logistic regression, elasticnet logistic regression, lasso logistic regression, neural network, support vector machine, boosted CART and super learner.

n (EPV)	Method	RMSE	MB (%)	ESD	ASD	VEB (%)	Cover (%)	Power (%)
100 (2.7)	Perfectly specified logistic regression	0.092	0.422	0.092	0.102	10.7	95.8	12.9
	Elasticnet logistic regression	0.117	5.751	0.102	0.096	-5.1	88.7	33.5
	Lasso logistic regression	0.113	5.051	0.101	0.101	0.4	91.3	28.6
	Neural network	0.069	-4.286	0.054	0.059	8.8	86.6	12.4
	Support vector machine	0.059	-0.648	0.059	0.055	-6.4	92.6	36.5
	Boosted CART	0.065	-5.114	0.039	0.017	-56.3	63.8	11.3
	Super learner	0.071	0.425	0.071	0.068	-3.7	93.1	30.8
500 (13.6)	Perfectly specified logistic regression	0.039	-0.036	0.039	0.038	-2.1	94.1	65.5
	Elasticnet logistic regression	0.047	1.700	0.044	0.043	-2.5	92.1	70.4
	Lasso logistic regression	0.045	1.398	0.043	0.043	-0.1	93.1	68.7
	Neural network	0.046	-2.075	0.041	0.048	19.0	96.1	29.5
	Support vector machine	0.043	1.365	0.041	0.041	-0.5	93.7	74.6
	Boosted CART	0.056	-4.896	0.028	0.026	-6.3	57.2	31.9
	Super learner	0.040	0.142	0.040	0.040	-1.6	95.2	65.0
1000 (27.3)	Perfectly specified logistic regression	0.027	-0.017	0.027	0.027	-0.9	94.9	93.6
	Elasticnet logistic regression	0.032	0.936	0.030	0.030	-0.7	92.2	92.4
	Lasso logistic regression	0.032	0.801	0.031	0.030	-1.3	92.9	91.5
	Neural network	0.036	-1.774	0.031	0.041	31.1	98.3	41.9
	Support vector machine	0.035	1.456	0.032	0.032	0.4	91.4	91.7
	Boosted CART	0.047	-4.017	0.241	0.023	-2.4	61.8	54.8
	Super learner	0.031	0.485	0.031	0.031	1.2	94.6	89.3

Abbreviations: MB = mean bias; RMSE = root mean square error; ESD = empirical standard deviation; ASD = asymptotic standard deviation; VEB = variance estimation bias; EPV = events per variable.

Table S4. Performances of G-computation in a simplistic situation with the following Q-models: the theoretical logistic regression, elasticnet logistic regression, lasso logistic regression, neural network, support vector machine, boosted CART and super learner.

n (EPV)	Method	RMSE	MB (%)	ESD	ASD	VEB (%)	Cover (%)	Power (%)
100 (4.1)	Perfectly specified logistic regression	0.093	-0.382	0.093	0.094	0.8	94.6	18.5
	Elasticnet logistic regression	0.096	0.293	0.096	0.097	1.0	93.9	20.2
	Lasso logistic regression	0.096	0.081	0.096	0.100	4.5	94.6	18.0
	Neural network	0.087	-1.607	0.086	0.103	19.7	96.6	10.9
	Support vector machine	0.078	-2.309	0.075	0.075	0.7	93.8	20.1
	Boosted CART	0.075	-3.874	0.065	0.061	-6.3	82.2	18.6
	Super learner	0.084	-1.255	0.083	0.083	-0.6	94.1	20.0
500 (20.4)	Perfectly specified logistic regression	0.041	-0.258	0.041	0.040	-1.8	93.6	66.6
	Elasticnet logistic regression	0.041	0.079	0.041	0.041	-1.2	94.5	67.5
	Lasso logistic regression	0.041	-0.052	0.041	0.041	0.1	94.8	65.1
	Neural network	0.043	-1.143	0.042	0.057	36.4	98.5	28.3
	Support vector machine	0.041	-0.253	0.041	0.042	3.4	94.3	64.1
	Boosted CART	0.055	-4.327	0.034	0.030	-11.0	69.6	36.5
	Super learner	0.040	-0.338	0.040	0.042	4.9	95.6	63.9
1000 (40.8)	Perfectly specified logistic regression	0.028	0.132	0.028	0.028	1.3	94.7	94.3
	Elasticnet logistic regression	0.028	0.243	0.028	0.029	1.8	95.1	93.8
	Lasso logistic regression	0.029	0.164	0.028	0.029	2.0	95.2	93.1
	Neural network	0.033	-1.184	0.031	0.038	23.2	97.4	62.8
	Support vector machine	0.030	0.233	0.030	0.031	4.0	95.8	89.9
	Boosted CART	0.043	-3.316	0.027	0.025	-8.2	73.1	70.7
	Super learner	0.029	-0.056	0.029	0.030	5.0	96.3	90.0

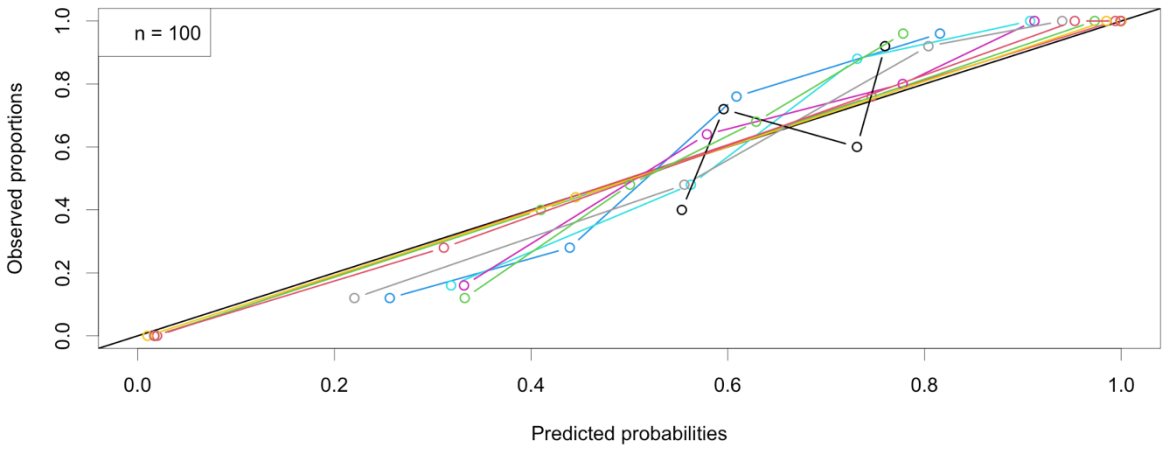
Abbreviations: MB = mean bias; RMSE = root mean square error; ESD = empirical standard deviation; ASD = asymptotic standard deviation; VEB = variance estimation bias; EPV = events per variable.

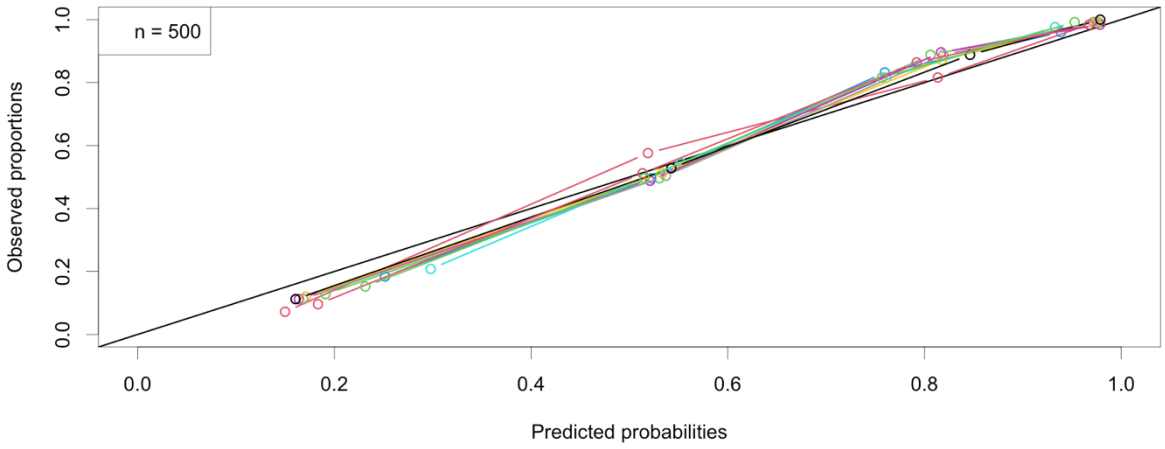
Table S5. List of the two sets of variables retained for the analysis of the case study. The set A contains all the available covariates that occurred prior to the exposure (i.e., prior to the first episode of intracranial hypertension). The set B is reduced to the variables that cause the outcome. These two sets were defined by M.L. based on his prior knowledge.

	Set A: variables included in the machine learning techniques	**Set B: variables included in the investigator-based logistic regression**
Female patient	X	X
Diabetes	X	X
Nosological entity: Severe trauma	X	X
SAP ≤ 90 mmHg before admission	X	X
Evacuation of subdural or extradural hematoma (*)	X	
External ventricular drain	X	
Evacuation of cerebral hematoma or lobectomy (*)	X	X
Decompressive craniectomy (*)	X	
Blood transfusion before admission	X	
Pneumonia before increased HICP	X	
Osmotherapy (*)	X	X
GCS score ≥ 8	X	X
Patient age	X	X
Hemoglobin	X	
Platelets	X	
Serum creatinine	X	
Arterial pH	X	
Serum proteins	X	
Serum urea	X	X
PaO_2/FiO_2 ratio	X	X
SAPS II score	X	

GOS score was dichotomised into favourable outcomes (good recovery or moderate disability) or unfavourable outcomes (severe disability, vegetative state or death). Abbreviations: GOS: Glasgow Outcome Scale, SAP: Systolic Arterial Pressure, HICP: High Intracranial Pressure, GCS: Glasgow Coma Scale, PaO_2: arterial partial Pressure of Oxygen, FiO_2: Fraction of Inspired Oxygen, and SAPS: Simplified Acute Physiology Score. (*) Before HICP

Figure S1. Calibration plots related to the predictions based on the lasso logistic regression in the realistic situation for 10 simulated datasets.





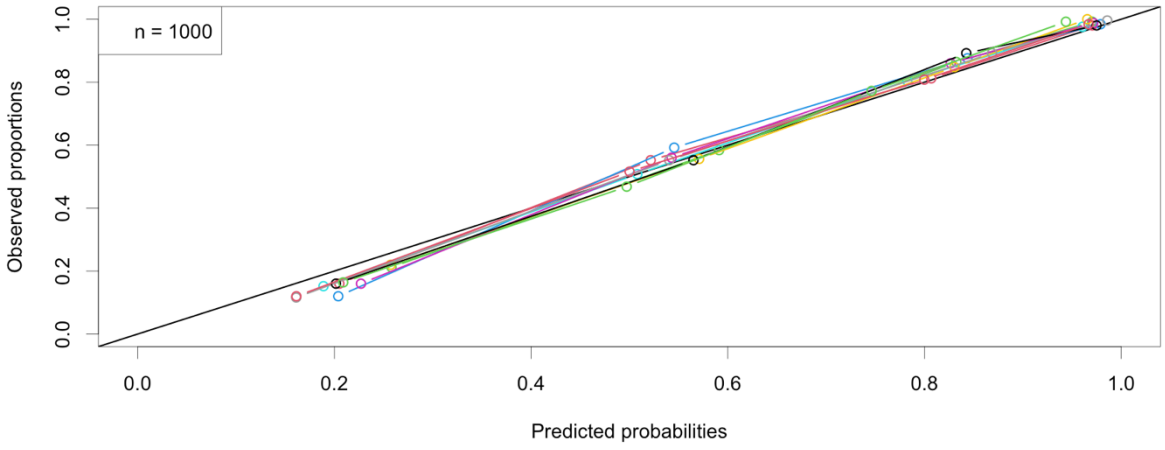
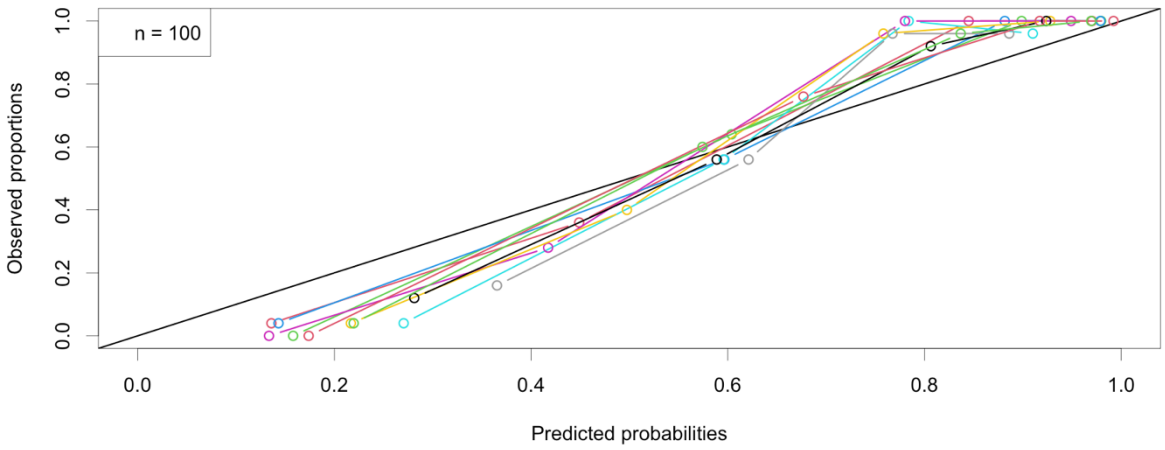
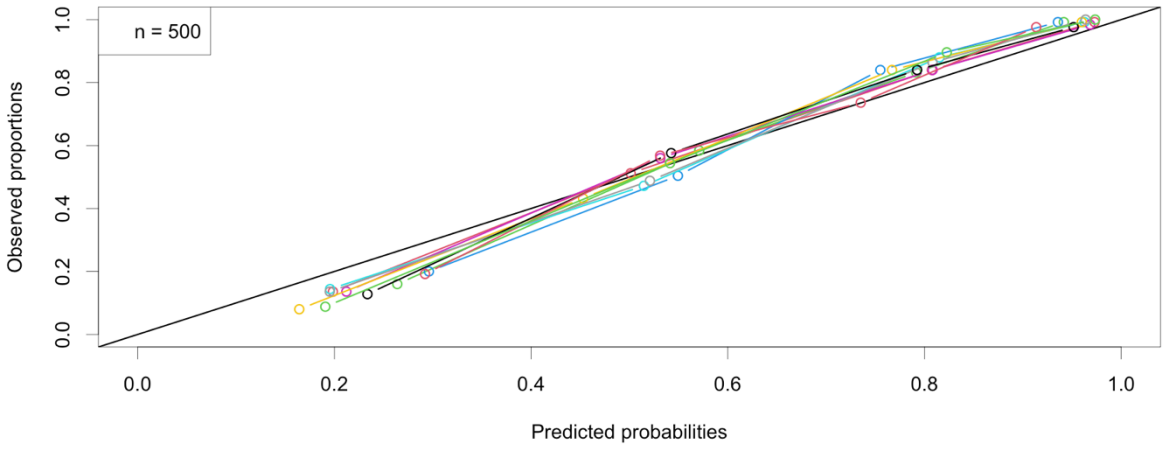


Figure S2. Calibration plots related to the predictions based on the elasticnet logistic regression in the realistic situation for 10 simulated datasets.





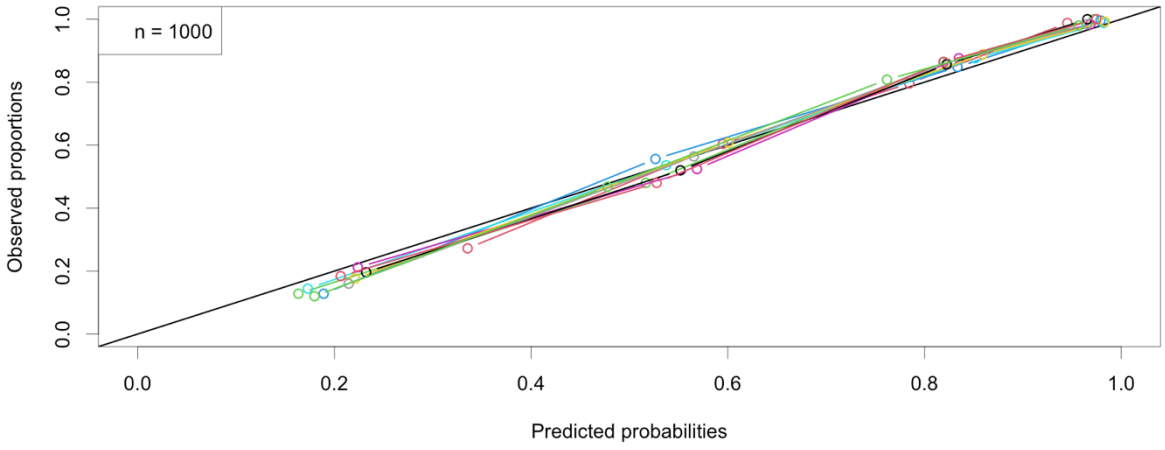
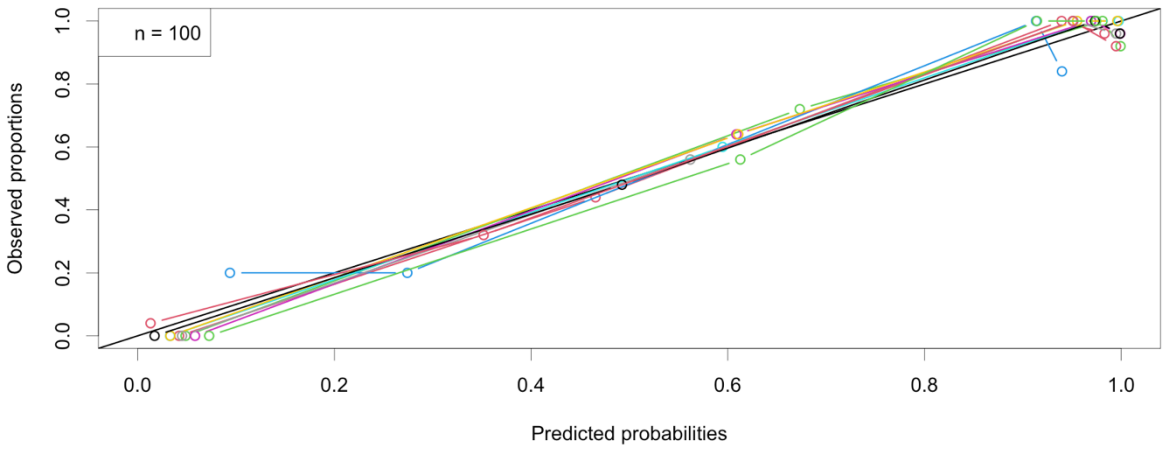
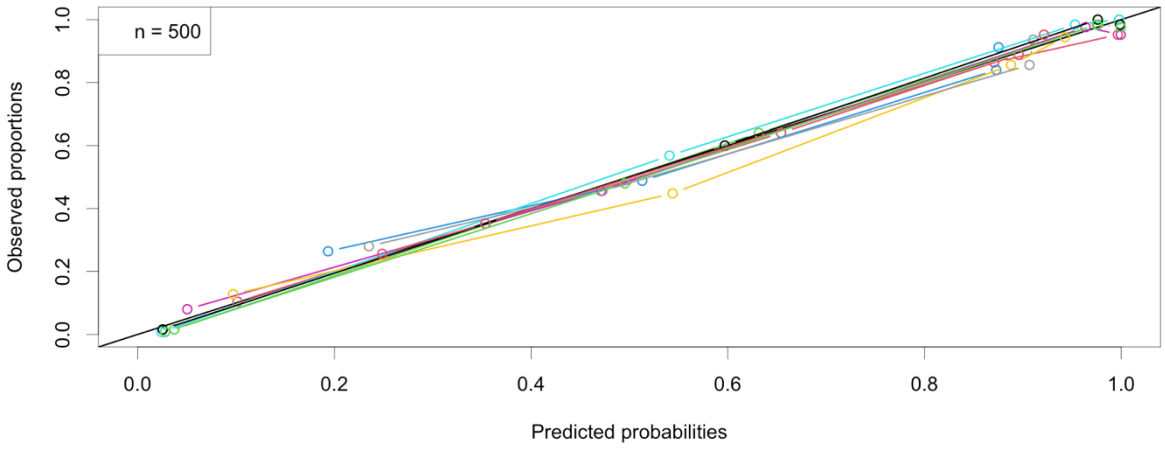


Figure S3. Calibration plots related to the predictions based on the neural network in the realistic situation for 10 simulated datasets.





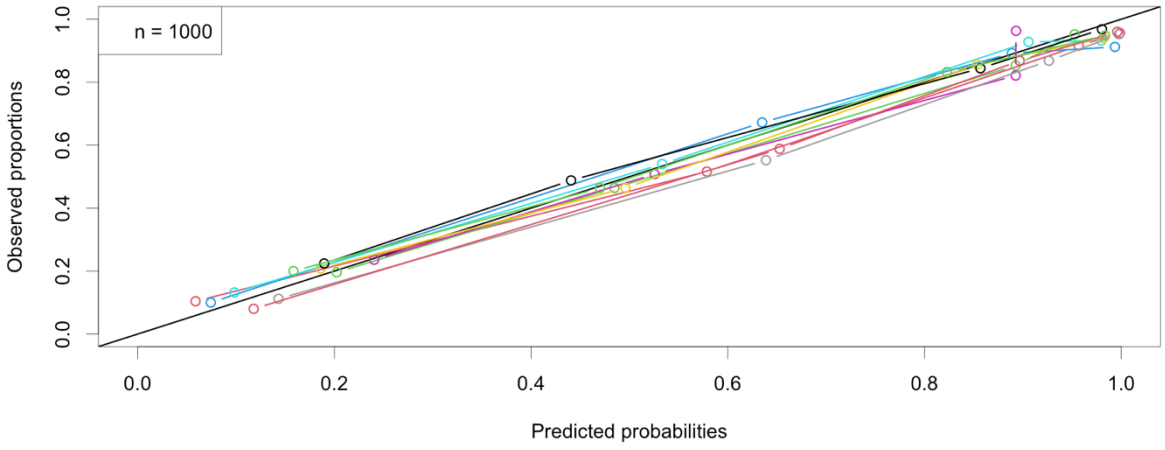


Figure S4. Calibration plots related to the predictions based on the support vector machine in the realistic situation for 10 simulated datasets.

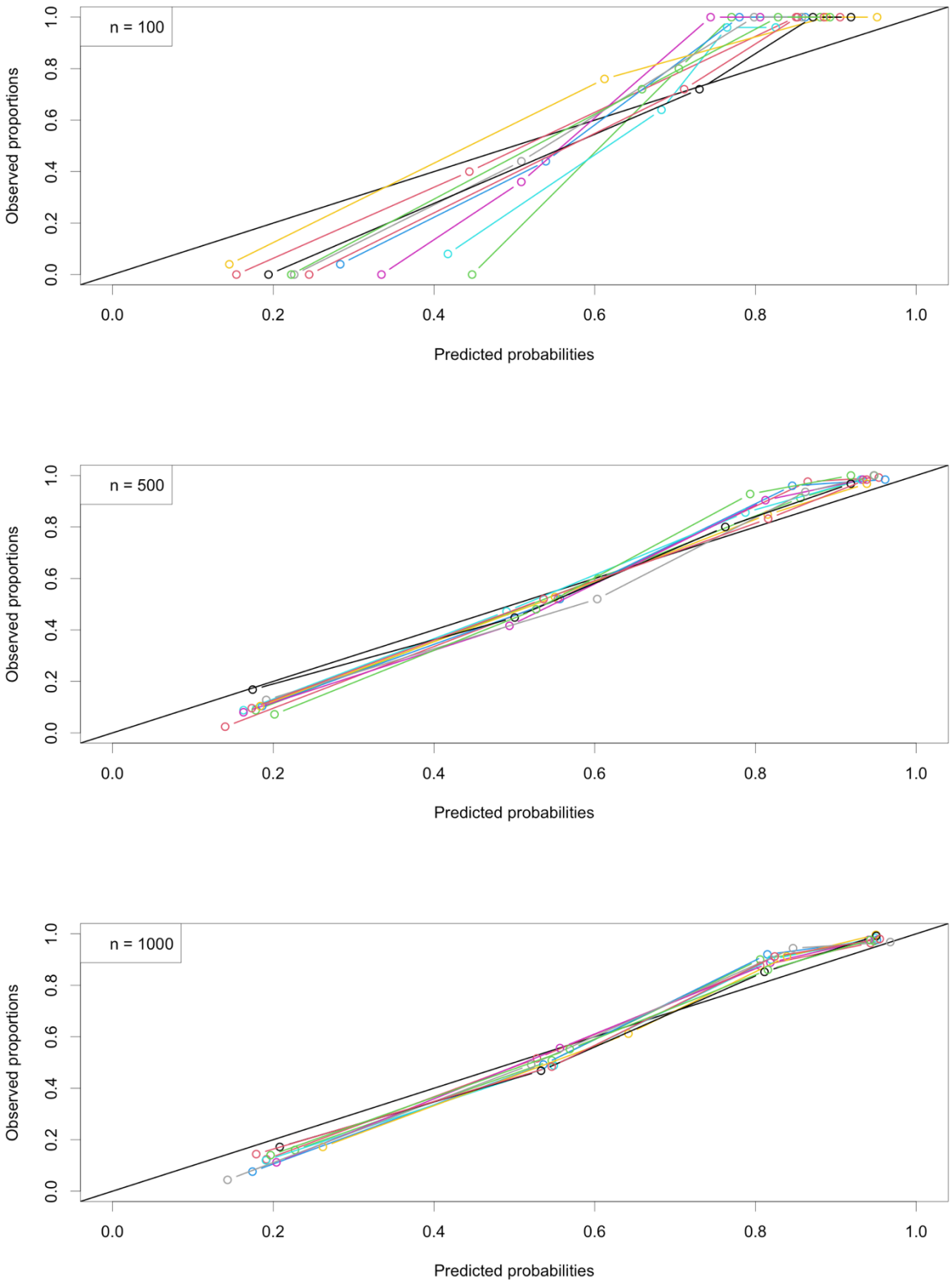
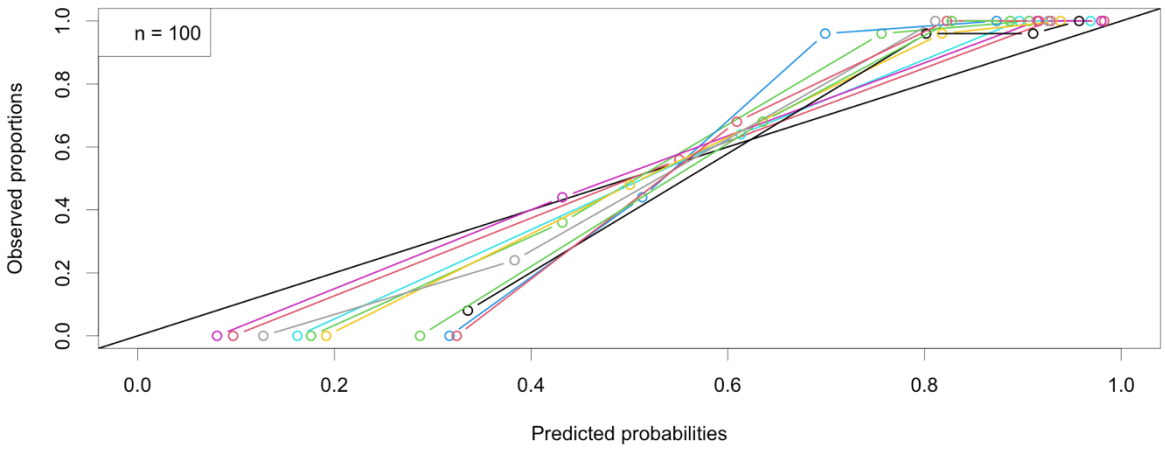
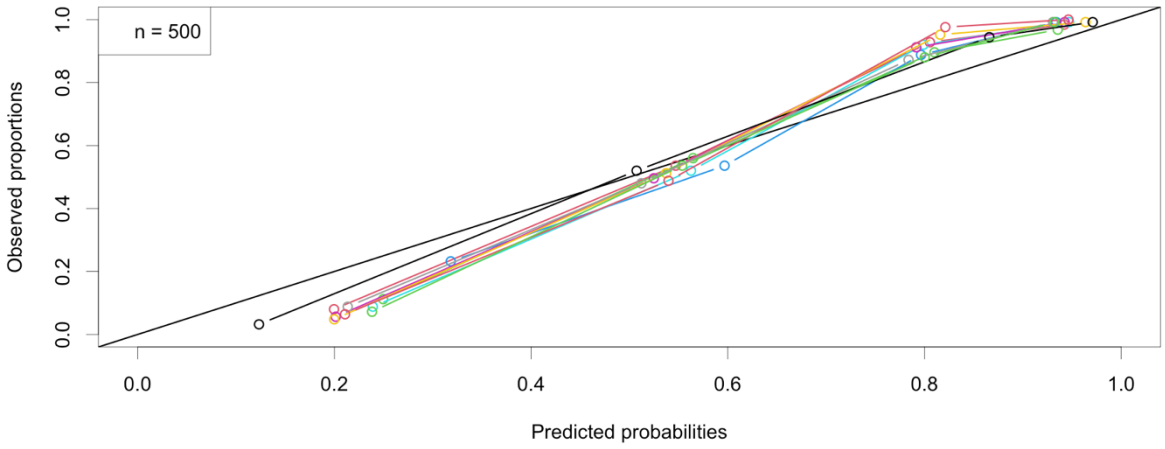


Figure S5. Calibration plots related to the predictions based on the super learner in the realistic situation for 10 simulated datasets.





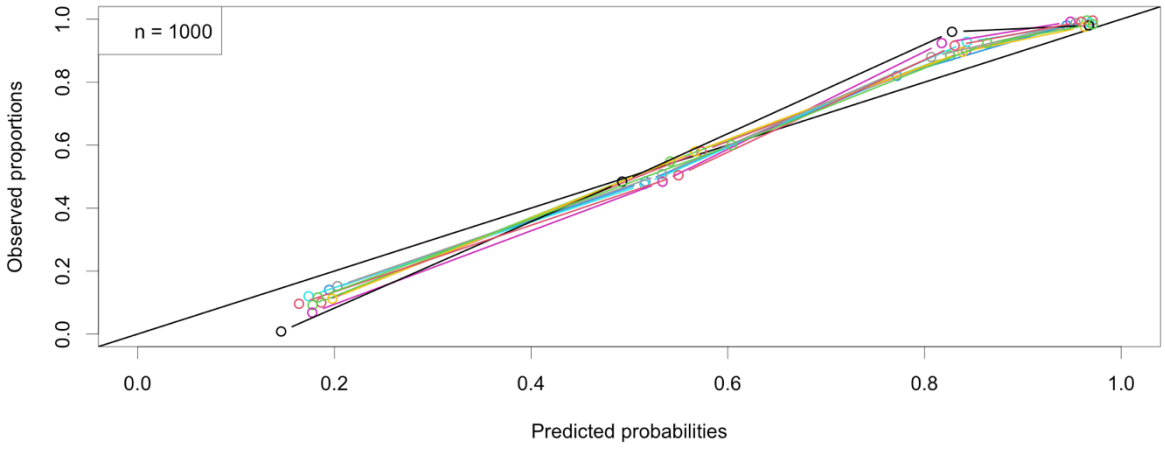
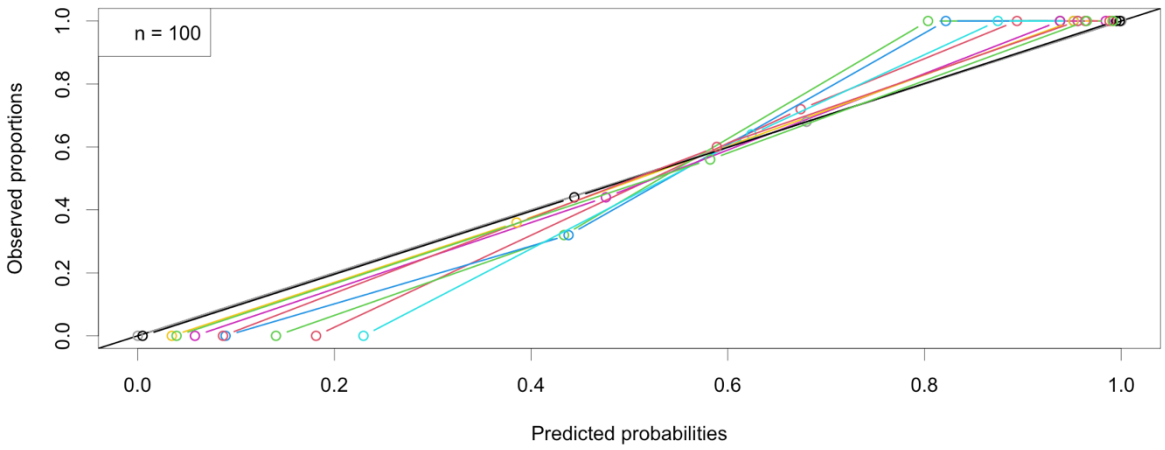
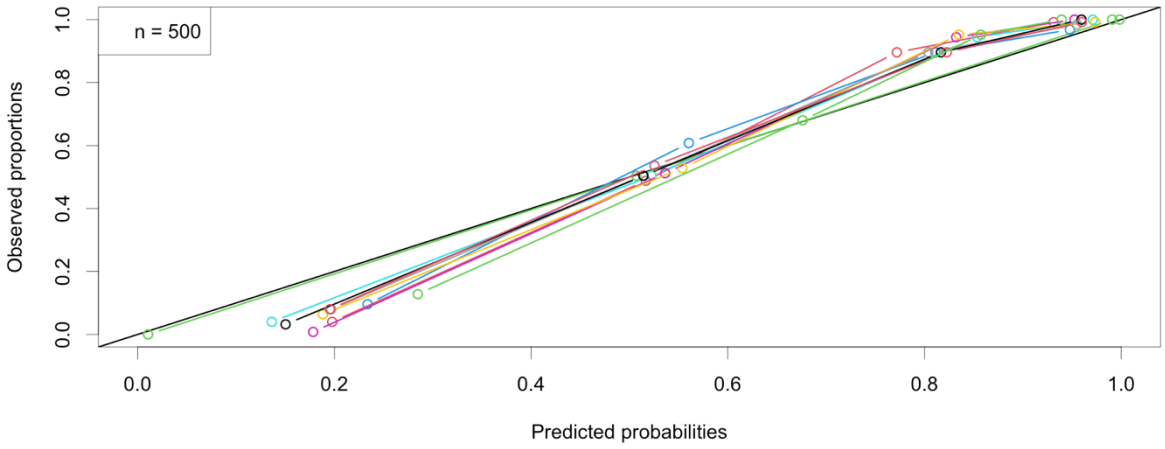


Figure S6. Calibration plots related to the predictions based on the boosted classification and regression trees in the realistic situation for 10 simulated datasets.





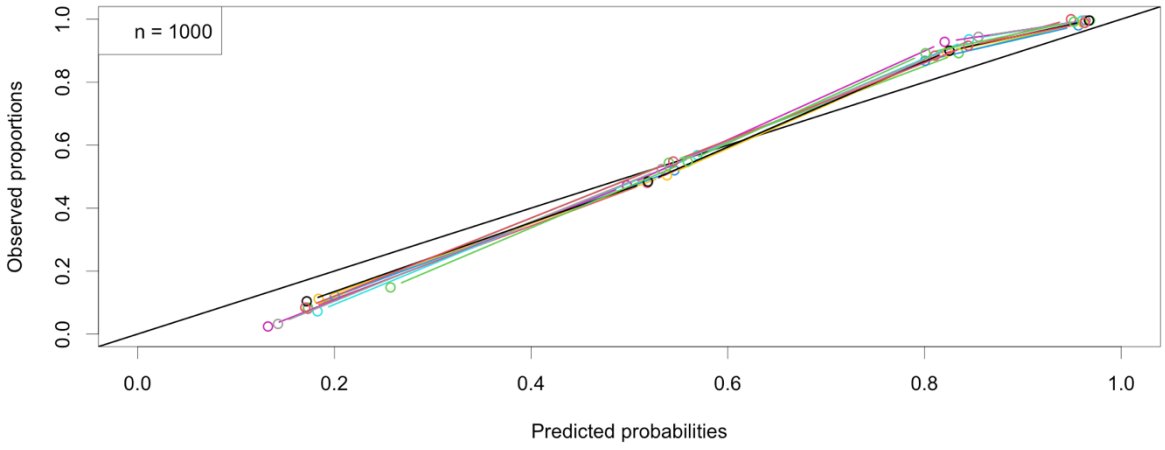
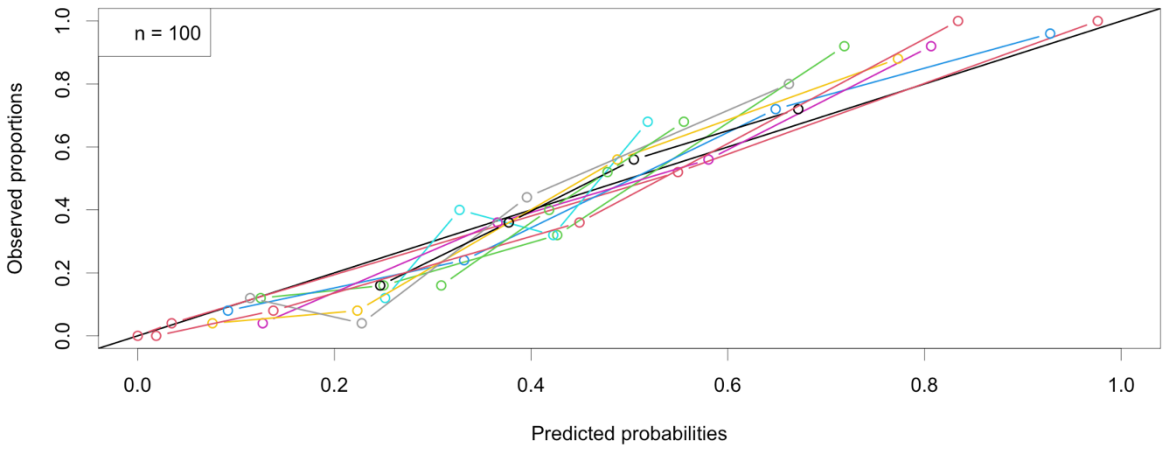
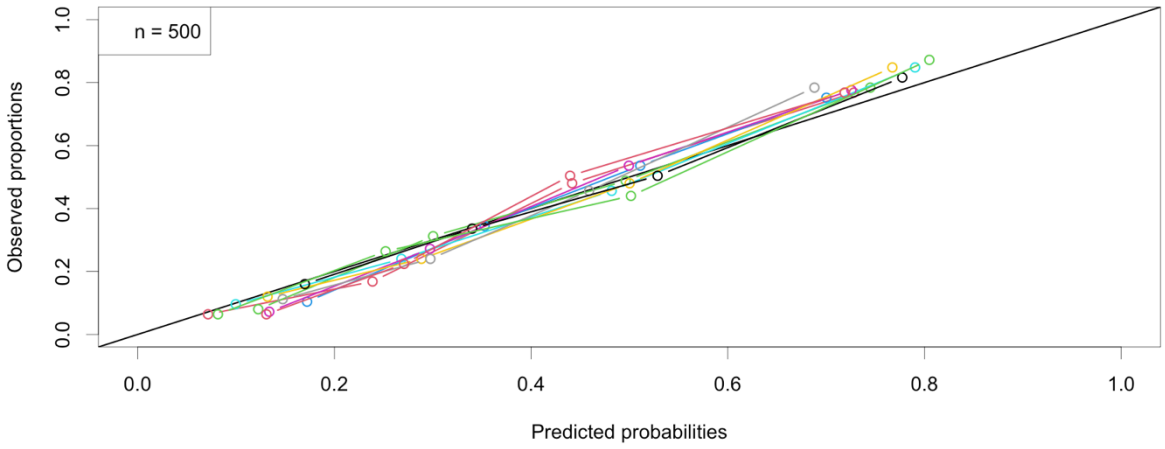


Figure S7. Calibration plots related to the predictions based on the lasso logistic regression in the simplistic situation for 10 simulated datasets.





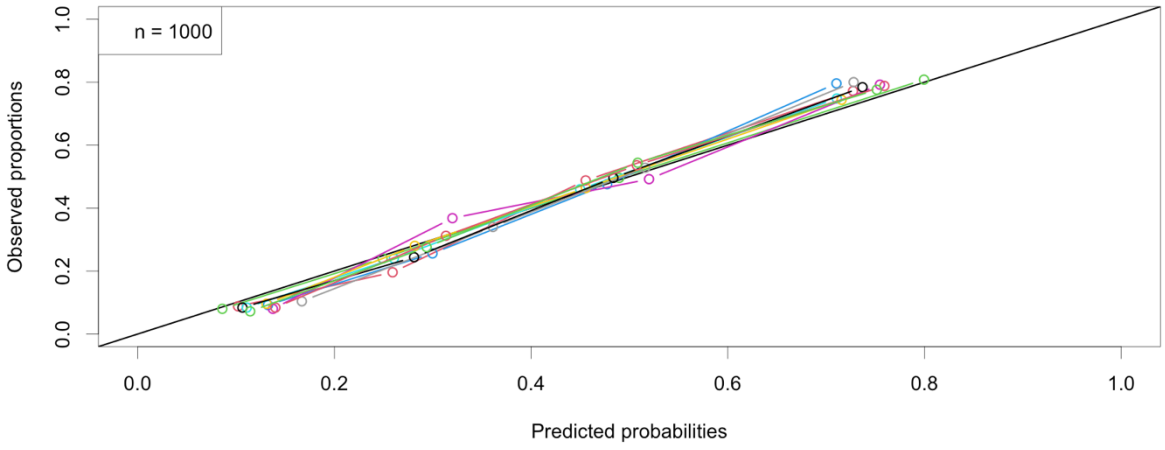
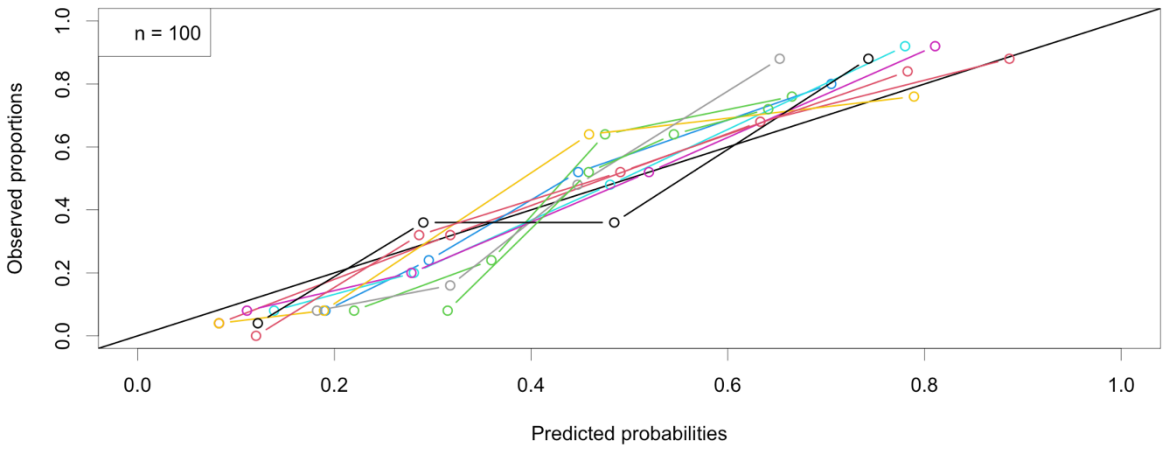
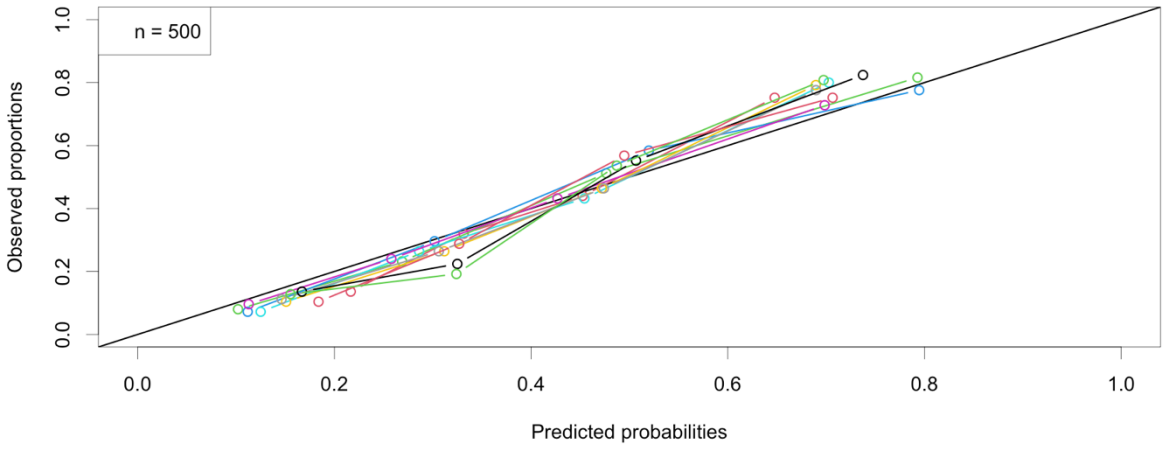


Figure S8. Calibration plots related to the predictions based on the elasticnet logistic regression in the simplistic situation for 10 simulated datasets.





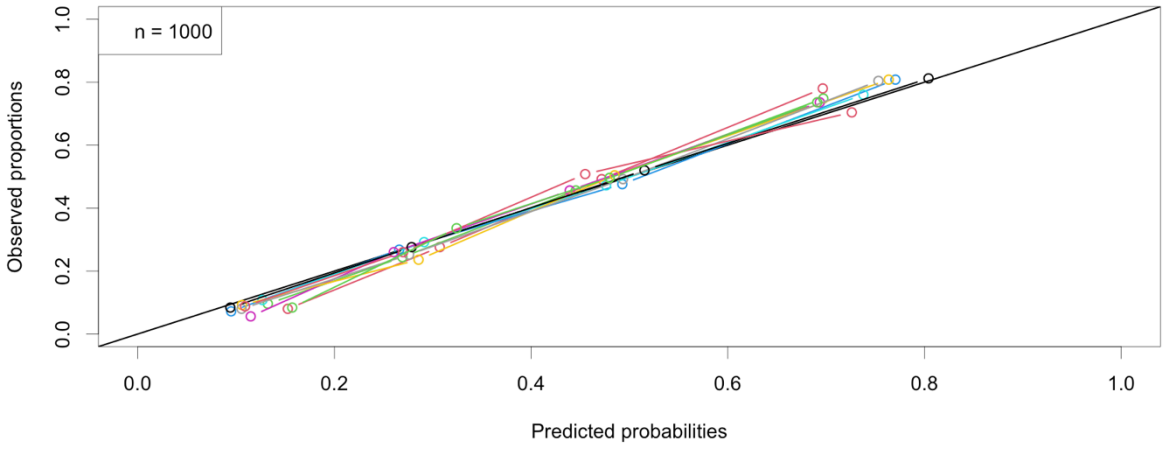
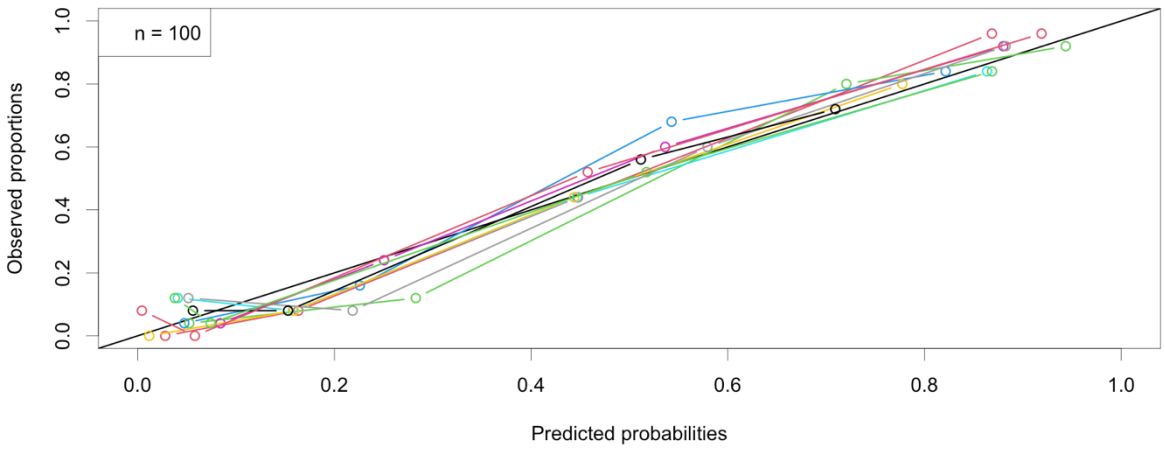
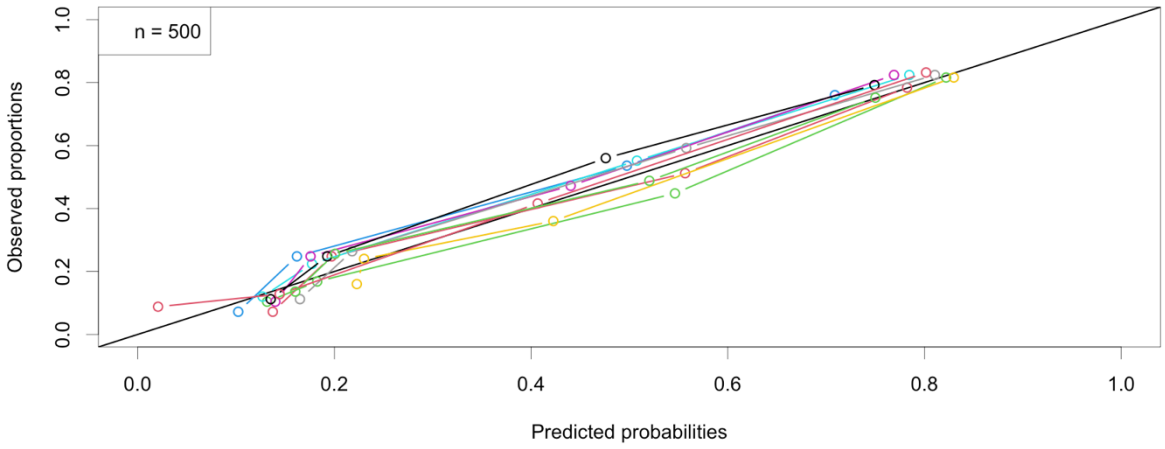


Figure S9. Calibration plots related to the predictions based on the neural network in the simplistic situation for 10 simulated datasets.





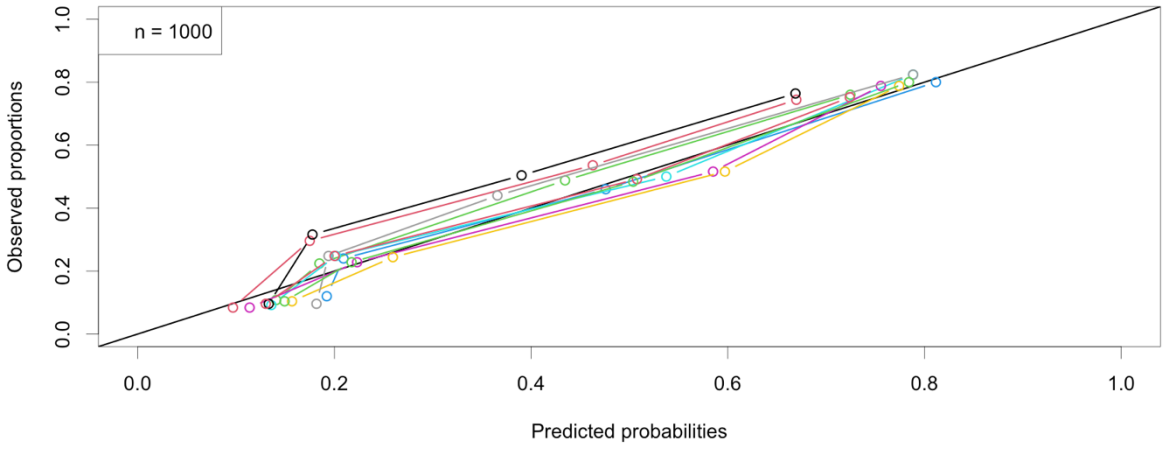
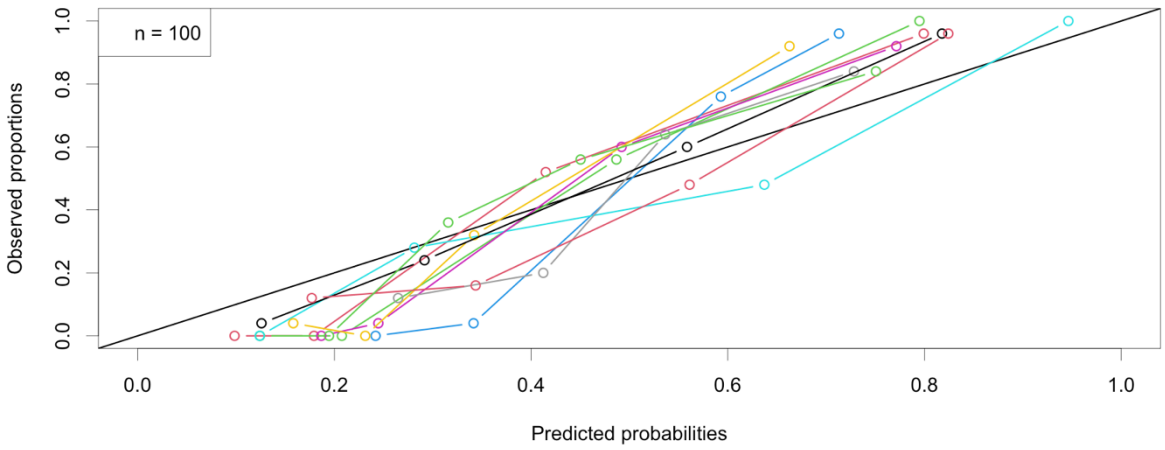
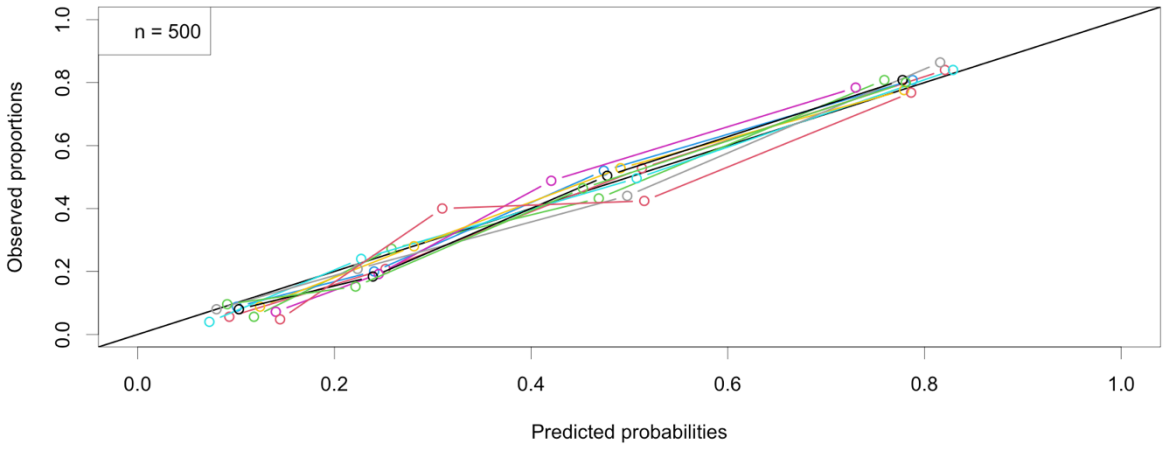


Figure S10. Calibration plots related to the predictions based on the support vector machine in the simplistic situation for 10 simulated datasets.





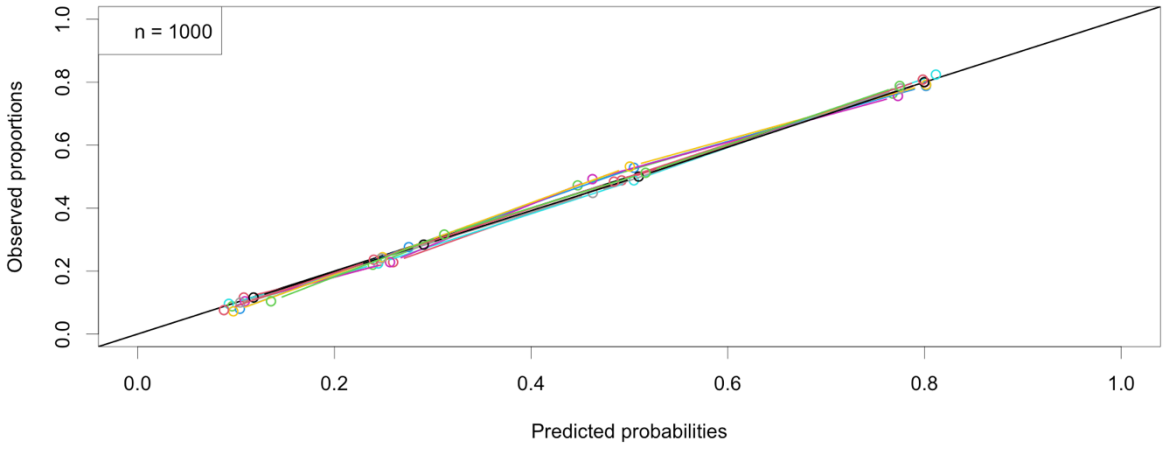
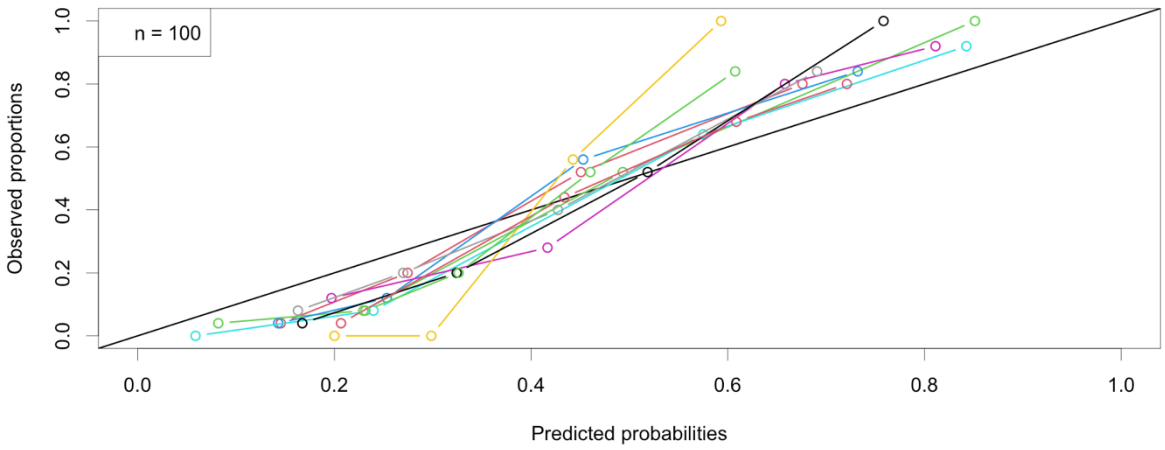
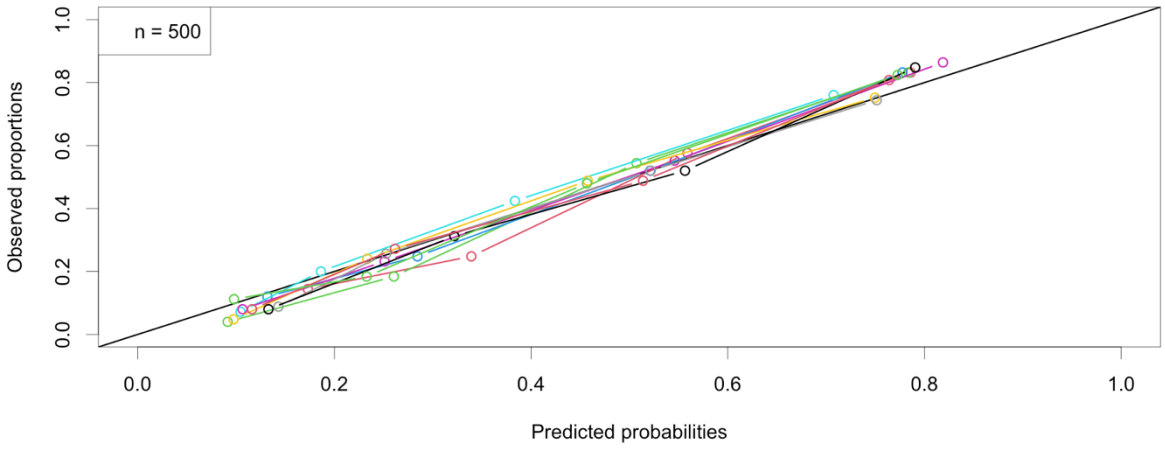


Figure S11. Calibration plots related to the predictions based on the super learner in the simplistic situation for 10 simulated datasets.





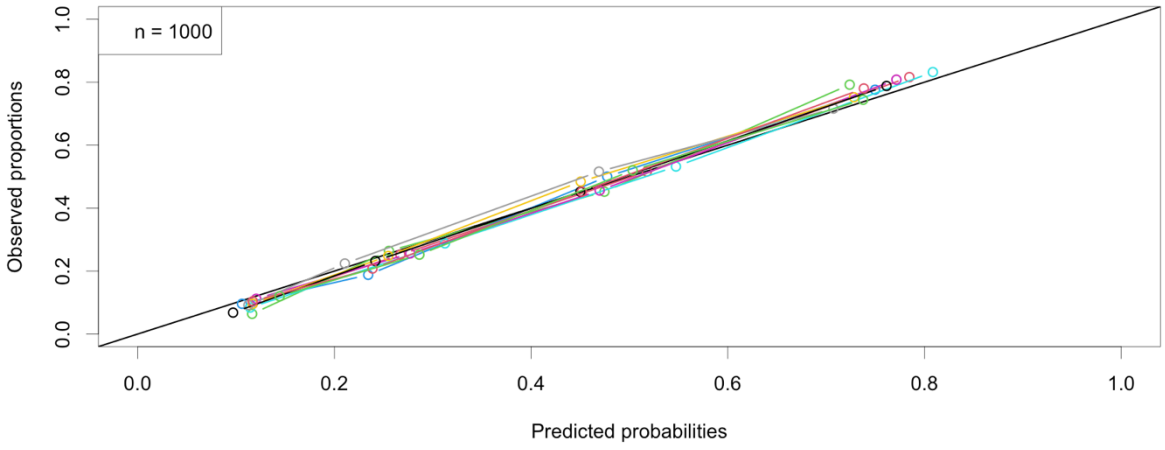
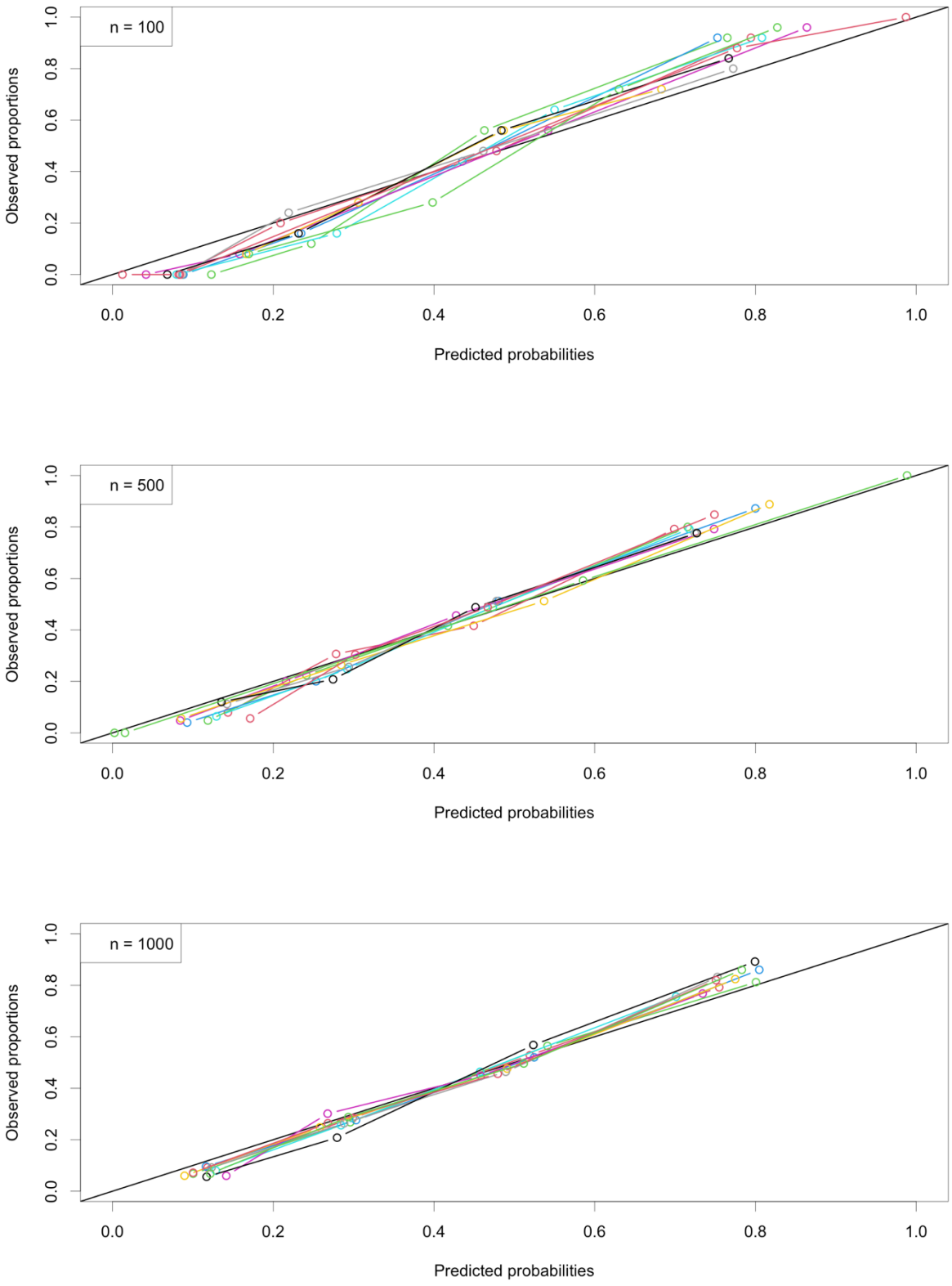


Figure S12. Calibration plots related to the predictions based on the boosted classification and regression trees in the simplistic situation for 10 simulated datasets.



Annexe F

Éléments supplémentaires au chapitre 5

G-computation for continuous-time data: a comparison with inverse probability weighting

Statistical Methods in Medical research

Supplementary materials

Arthur Chatton

INSERM UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France.
IDBC-A2COM, Pacé, France.

Florent Le Borgne

INSERM UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France.
IDBC-A2COM, Pacé, France.

Clémence Leyrat

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK.
Cancer Survival Group, London School of Hygiene and Tropical Medicine, London, UK.

Yohann Foucher

INSERM UMR 1246 - SPHERE, Nantes University, Tours University, Nantes, France.
Centre Hospitalier Universitaire de Nantes, Nantes, France.

E-mail: Yohann.Foucher@univ-nantes.fr

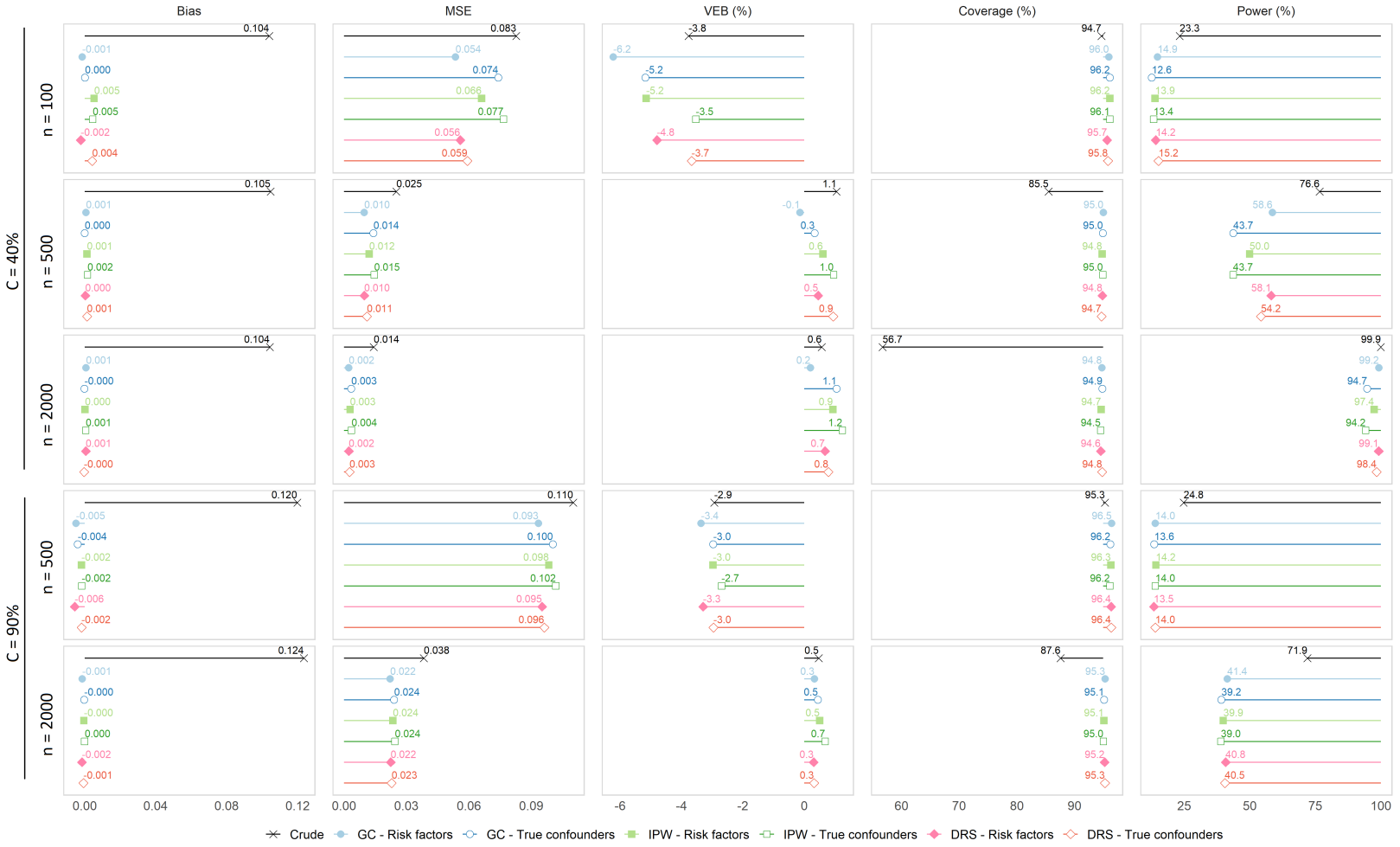


Fig. 1. Performances of the g-computation (GC), inverse probability weighting (IPW) and Doubly Robust Standardisation (DRS) under the alternative hypothesis to estimate the log average hazard ratio. Theoretical values of the log average hazard ratio equal to 0.210 and 0.256 for censoring rates of 40% and 90%, respectively. Abbreviations: C, censoring rate; n, sample size.

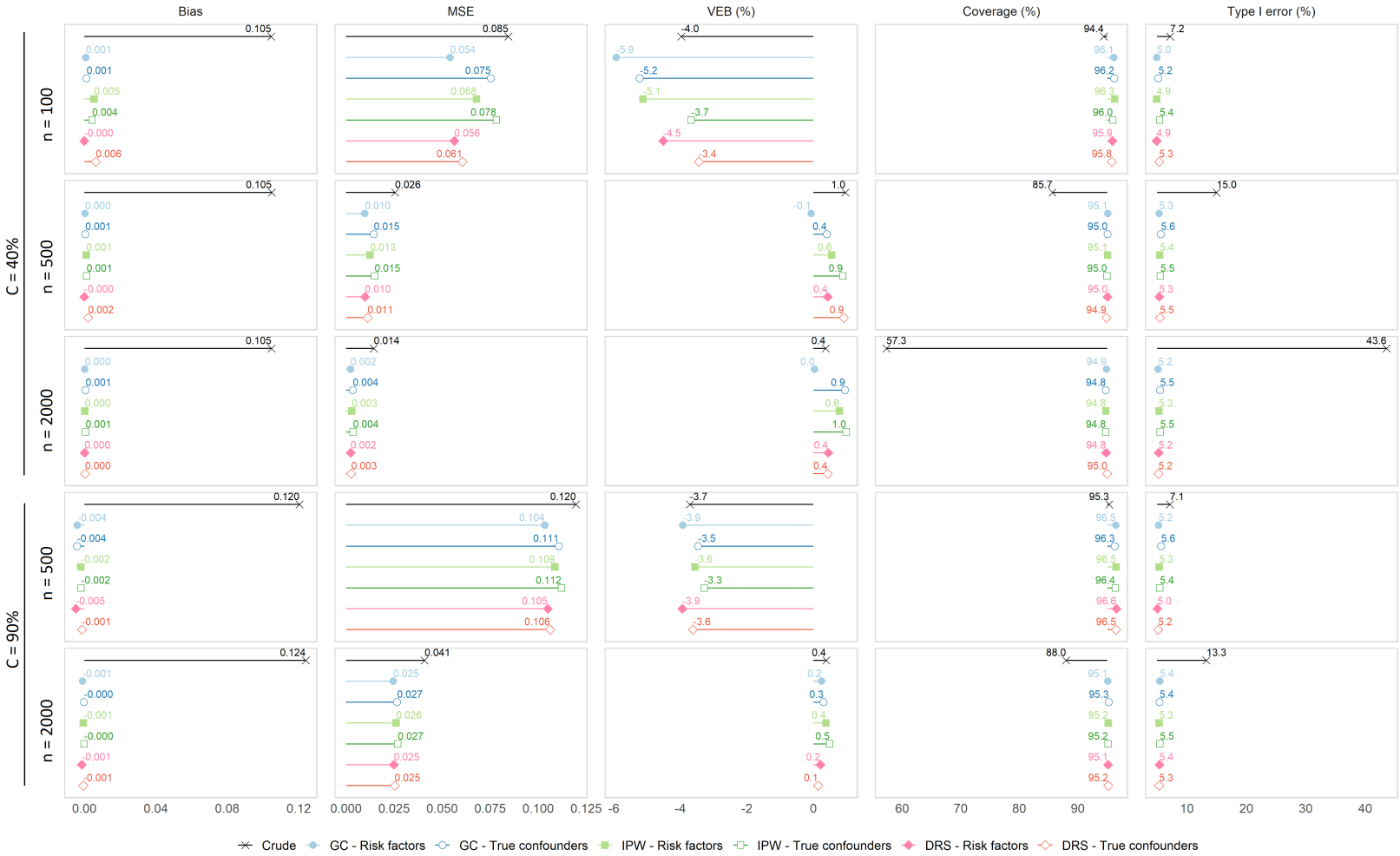


Fig. 2. Performances of g-computation (GC), inverse probability weighting (IPW) and Doubly Robust Standardisation (DRS) under the null hypothesis to estimate the log average hazard ratio. Theoretical values of the log average hazard ratio equals to 0.000. Abbreviations: C, censoring rate; n, sample size.

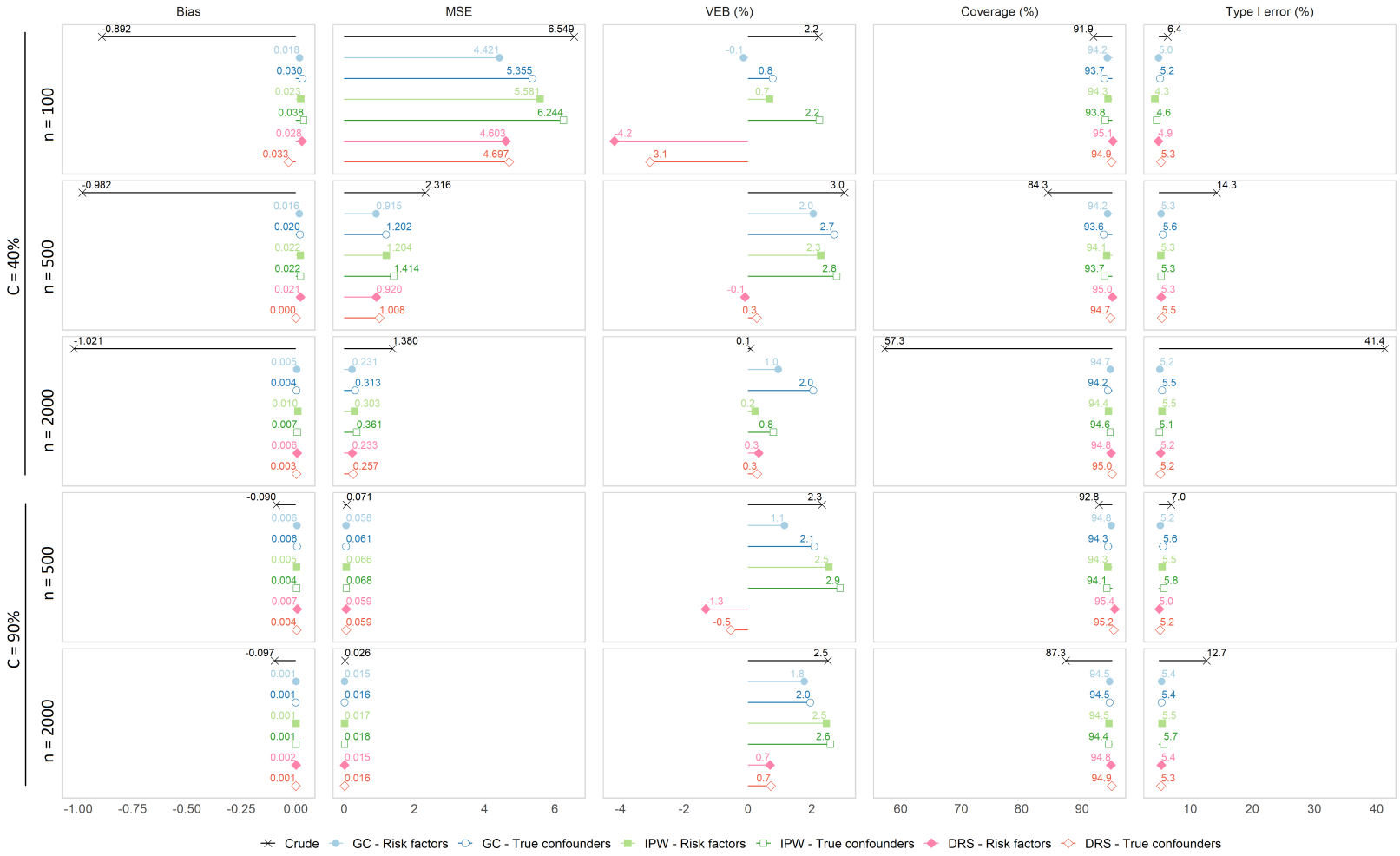


Fig. 3. Performances of g-computation (GC), inverse probability weighting (IPW) and Doubly Robust Standardisation (DRS) under the null hypothesis to estimate the Restricted Mean Survival Times difference at time τ. τ equals to 40.0 and 13.0 for censoring rates of 40% and 90%, respectively. Theoretical value of restricted mean survival times difference equals to 0.000. Abbreviations: C, censoring rate; n, sample size.

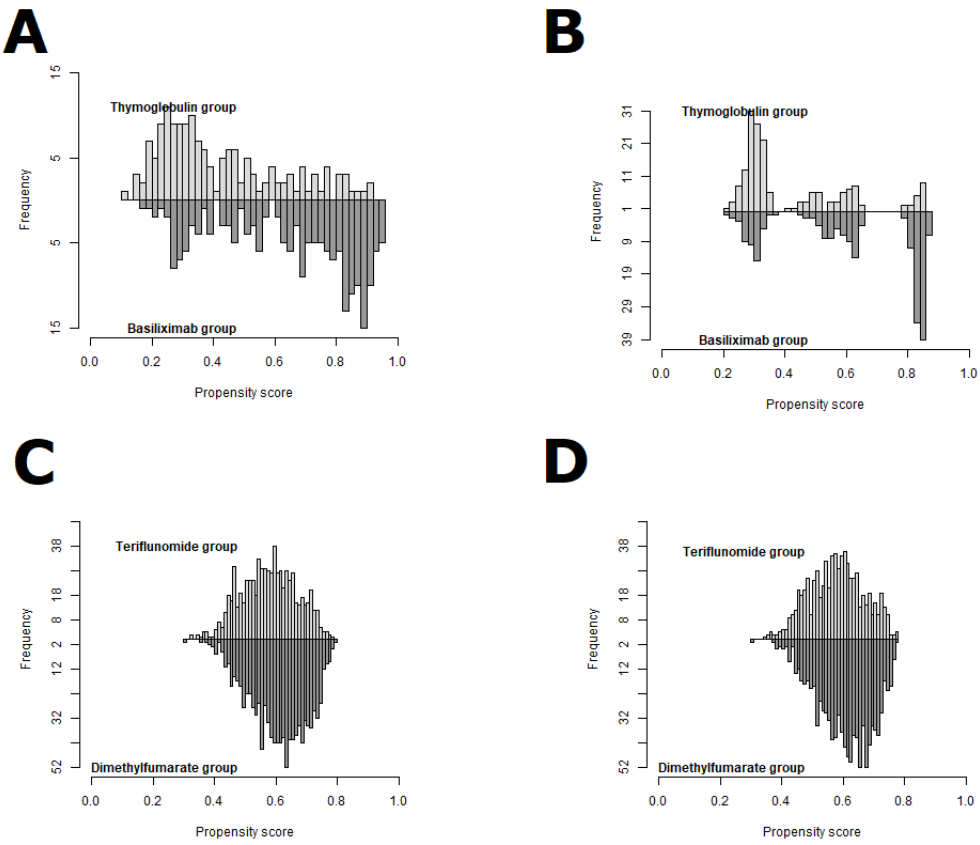


Fig. 4. Positivity plots in real-life applications. **A -** Kidney transplantation, risk factors (N=352.6, weighted sample); **B -** Kidney transplantation, true confounders (N=382.2, weighted sample); **C -** Multiple sclerosis, risk factors (N=1769.7, weighted sample); and **D -** Multiple sclerosis, true confounders (N=1770.4, weighted sample).

Table 1. The PS-adjusted samples of the time-to-first relapse of multiple sclerosis according to the covariate sets.

Set	Characteristics	Overall		TRF		DMF		STD (%)
Risk factors	Relapse within the previous year (n, %)	977.8	55.3	393.0	55.2	584.9	55.3	0.3
	Gado. Positive lesion on MRI at baseline (n, %)	605.2	34.2	244.2	34.3	361.0	34.1	0.3
	EDSS level (mean, sd)	1.7	1.2	1.7	1.3	1.7	1.2	1.2
	Disease duration (mean, sd)	7.6	7.4	7.6	7.5	7.6	7.3	0.1
	Age at initiation (mean, sd)	39.3	10.8	39.3	10.8	39.3	10.8	0.0
True confounders	Relapse within the previous year (n, %)	978.5	55.3	394.0	55.2	584.5	55.3	0.1
	Gado. Positive lesion on MRI at baseline (n, %)	606.0	34.2	245.3	34.4	360.8	34.1	0.5
	Disease duration, years (mean, sd)	7.6	7.4	7.6	7.5	7.6	7.3	0.3
	Age at initiation (mean, sd)	39.3	10.8	39.3	10.8	39.3	10.8	0.4

Qualitative characteristics are presented by using the weighted effective (n) and the weighted percentage. Continuous characteristics are presented with weighted mean following by weighted standard deviation (sd).
Abbreviations: DMF, Dimethylfurmarate; EDSS, Expanded Disability Status Scale; Gado, Gadolinium; MRI, Magnetic resonance imaging; MS, Multiple sclerosis; STD, Standardised differences in %; and TRF, Teriflunomid.

Table 2. The PS-adjusted samples of the kidney's post-transplant cardiovascular complications according to the covariate sets.

Set	Characteristics	Overall		ATG		BSX		STD (%)
Risk factors	Male Recipient (n, %)	259.2	73.5	114.5	73.4	144.7	73.6	0.5
	Recurrent causal nephropathy (n, %)	57.7	16.4	24.3	15.6	33.4	17.0	3.9
	History of diabetes (n, %)	112.1	31.8	49.1	31.5	63.0	32.1	1.3
	History of hypertension (n, %)	303.8	86.2	136.5	87.5	167.3	85.1	7.1
	History of cardiovascular disease (n, %)	198.2	56.2	90.5	58.0	107.7	54.8	6.5
	History of dyslipedemia (n, %)	203.6	57.7	91.8	58.9	111.7	56.8	4.2
	Donor hypertension (n, %)	210.0	59.6	89.9	57.7	120.1	61.1	6.9
	Vascular cause of donor death (n, %)	259.5	73.6	116.2	74.5	143.3	72.9	3.7
	Recipient BMI, kg.m^2 (mean, sd)	26.5	4.2	26.4	4.5	26.7	4.0	6.4
	Duration on waiting list, months (mean, sd)	17.5	20.9	17.9	19.8	17.1	21.7	3.9
	Recipient creatinemia, μmol/L (mean, sd)	80.0	37.7	79.9	39.2	80.0	36.6	0.2
	Recipient age (mean, sd)	70.9	4.8	71.0	5.1	70.9	4.6	1.9
	Donor age (mean, sd)	72.7	8.3	72.6	9.2	72.8	7.5	1.5
True confounders	Recipient age	71.0	4.9	71.0	5.1	71.1	4.8	0.8
	Donor age	72.9	8.2	72.9	9.0	72.9	7.5	1.1

Qualitative characteristics are presented by using the weighted effective (n) and the weighted percentage. Continuous characteristics are presented with weighted mean following by weighted standard deviation (sd).
Abbreviations: ATG, Thymoglobulin; BMI, Body mass index; BSX, Basiliximab; CMV, Cytomegalovirus; EBV, Epstein-Barr virus; ECD, Expanded criteria donor; HLA, Human leucocyte antigen; and STD, Standardised differences.

Annexe G

Éléments supplémentaires au chapitre 6

Supplementary materials for the manuscript: ''Identifying positivity violations with decision trees: introducing the PCART algorithm''.

Gabriel Danelian,[1,2] Yohann Foucher,[3,4] Maxime Léger,[3,5] Florent Le Borgne,[2,3] and Arthur Chatton[2,3]

[1] Université de Lille, Lille, France

[2] IDBC/A2COM, Pacé, France.

[3] UMR INSERM 1246 - SPHERE, Université de Nantes, Université de Tours, Nantes, France.

[4] Centre Hospitalier Universitaire de Nantes, Nantes, France.

[5] Département d'anesthésie-réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France.

List of contents:

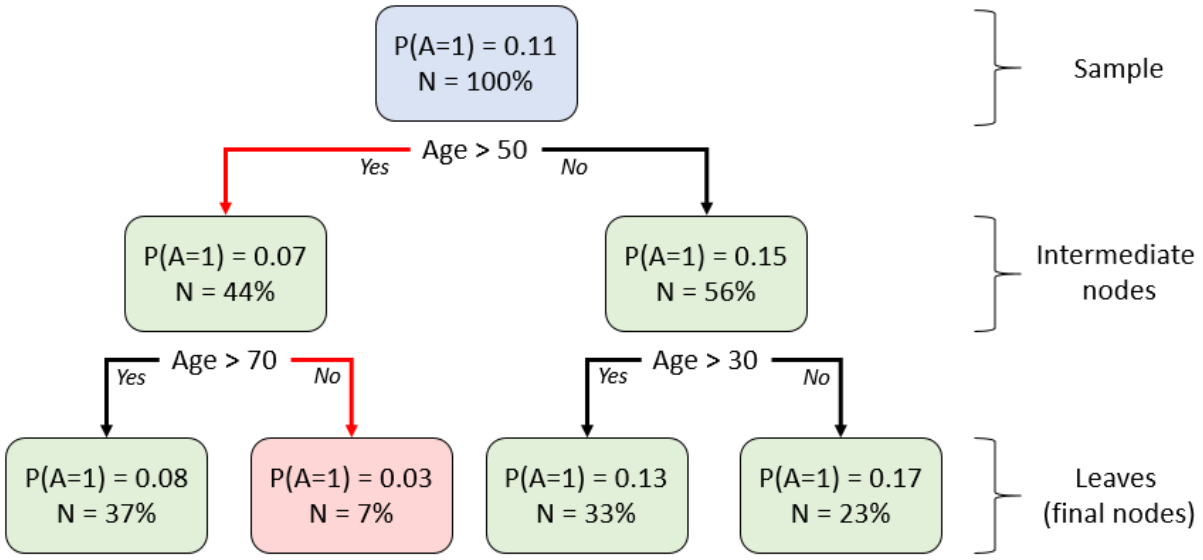


Figure 1: Example of decision tree identifying a subgroup responsible of non-positivity. In red, the path illustrating the problematic leaf (i.e., the subgroup 50 < Age ≤ 70), in green the non-problematic subgroups. A: Treatment, N: Percentage of sample contained in the subgroup.

Supplementary Table 1: Covariates sets used in the different studies.

Authors	Outcome	Adjustment set
Léger et al.	Mortality	History of head trauma Type of brain injury Creatinine levels at admission Osmotherapy Age SAPS II score Intracranial hypertension at admission Lactatemia at admission
Querard et al.	Patient and graft survival	Recipient age Year of transplantation
Foucher et al.	Patient and graft survival	Center Year of transplantation History of cardiovascular disease History of hypertension History of malignancy History of dyslipemia History of B or C hepatitis History of diabetes Recipient gender HLA-A-B-DR incompatibilities Donor hypertension Retransplantation Recipient age Duration on waiting list Donor age Cold ischemia time
Masset et al.	Patient and graft survival	Recipient gender History of cardiovascular disease Recipient age
	Time to first infection	Preemptive transplantation History of dyslipidemia Donor gender Donor hypertension Recipient age Cold ischemia time
	Time to first acute rejection	History of vascular disease Vascular cause of donor death Donor hypertension Cold ischemia time
	Time to malignancy	Recipient gender Preemptive transplantation History of diabetes History of cardiac disease History of malignancy Donor hypertension Donor blood group Recipient age
	Time to post-transplant diabete	History of cardiovascular disease Donor gender Recipient body mass index Donor age Cold ischemia time
	Time to cardiac complication	Recipient gender Recurrent causal nephropathy History of hypertension History of cardiovascular disease

		History of B or C hepatitis
		Donor hypertension
		Recipient body mass index
		Center
	Delayed graft function	History of vascular disease
		History of cardiac disease
		Donor gender
		Recipient body mass index
		Donor creatininemia
		Cold ischemia time

Supplementary Table 2: Results of PCART with α=5% and β=5%.

Authors	Problematic subgroup	n	Part of the total sample (%)	Probability of treatment (%)
Léger et al.	No osmotherapy at admission	732	67.3	2.9
	Age ≥ 75 years	73	6.7	2.7
	Lactatemia < 1 mmol/L	197	18.1	4.1
	No IH at admission nor history of head trauma	710	65.3	4.8
	No IH & severe trauma at admission	385	35.4	4.9
	No IH at admission & Creatinine < 150 mmol/L	740	68.0	4.9
	No IH at admission & SAPS II score < 55	532	48.9	4.1
	40 ≤ SAPS II score < 45 & No severe trauma	69	6.3	4.3
	25 ≤ SAPS II score < 55 & 50 ≤ Creatinine < 60	135	12.4	4.4
Querard et al.	Recipient age < 30 years	391	8.1	3.1
Foucher et al.	Transplant before 2015 & A - D centres[a]	376	19.0	2.9
	Transplant before 2013 & CIT ≥ 20	101	5.1	5.0
Masset et al.	None detected	-	-	-

Abbreviations: CIT, Cold Ischemia Time; IH, Intracranial Hypertension; and n, sample size.
[a] The centres were anonymised.

Supplementary Table 3: Results of PCART with α=5% and β=1%.

Authors	Problematic subgroup	n	Part of the total sample (%)	Probability of treatment (%)	New[a]
Léger et al.	Lactatemia < 1 mmol/L & Creatinine < 50	61	5.6	0.0	Yes
	No osmotherapy & Age ≥ 75 years	60	5.5	0.0	Yes
	No osmotherapy & 55 ≤ Age < 60 years	77	7.1	0.0	Yes
	No osmotherapy & 30 ≤ SAPS II score < 35 years	68	6.2	0.0	Yes
	No osmotherapy & 40 ≤ SAPS II score < 50 years	222	20.4	0.9	Yes
	No osmotherapy & Creatinine < 50	144	13.2	0.7	Yes
	No osmotherapy & Lactatemia < 4	74	6.8	0.0	Yes
	Age ≥ 70 years & 1 ≤ Lactatemia < 2	62	5.7	0.0	Yes
	No IH at admission & 55 ≤ Age < 60	77	7.1	0.0	Yes
	No IH at admission & Age ≥ 70 years	116	10.7	0.9	Yes
Querard et al.	None detected	-	-	-	-
Foucher et al.	None detected	118	6.2	0.0	Yes
Masset et al.	None detected	-	-	-	-

Abbreviations: IH, Intracranial Hypertension; and n, sample size.
[a] New subgroup with respect to the Supplementary Table 2.

Supplementary Table 4: Results of PCART with α=5% and β=10%.

Authors	Problematic subgroup	n	Part of the total sample (%)	Probability of treatment (%)	New[a]
Léger et al.	No osmotherapy at admission	732	67.3	2.9	No
	Age ≥ 55 years	479	44.0	7.3	Yes
	40 ≤ Age < 45	85	7.8	9.4	Yes
	Lactatemia < 1 mmol/L	61	5.6	0.0	No
	No IH at admission	754	69.3	5.1	Yes
	SAPS II score < 55	767	70.5	9.8	Yes
	Creatinine < 60 mmol/L	384	35.3	8.3	Yes
	110 ≤ Creatinine < 140	85	7.8	9.4	Yes
Querard et al.	Recipient age < 40 years	938	19.4	5.8	Yes
	Recipient age ≥ 65 years	817	16.9	9.0	Yes
Foucher et al.	Transplant before 2013 & Recipient age < 60	103	5.2	6.8	Yes
	Transplant after 2018 & CIT < 20	113	5.7	8.8	Yes
	Transplant before 2013 & CIT ≥ 20	101	5.1	5.0	No
	Duration on waiting list ≥ 10m & centre A[b]	128	6.5	5.5	Yes
	Transplant after 2016 & centre E[b]	101	5.1	6.9	Yes
	Transplant before 2015 & A - D centres[b]	376	19.0	2.9	No
	Transplant before 2013 & Female recipient	116	5.9	8.6	Yes
	Transplant before 2013 & HLA-A-B-DR incompatibilities < 5	243	12.3	9.9	Yes
	Transplant before 2013 & No donor hypertension	130	6.6	10.0	Yes
	Transplant before 2013 & History of dyslipidemia	151	7.6	9.9	Yes
Masset et al.	No history of CV disease & A centre[b,c]	47	12.7	6.4	Yes

Abbreviations: CIT, Cold Ischemia Time; CV, Cardiovascular; IH, Intracranial Hypertension; and n, sample size.

[a] New subgroup with respect to the Supplementary Table 2, [b] The centres were anonymised, [c] For time to cardiac complication.

Supplementary Table 5: Results of PCART with α=1% and β=5%.

Authors	Problematic subgroup	n	Part of the total sample (%)	Probability of treatment (%)	New[a]
Léger et al.	No osmotherapy at admission	732	67.3	2.9	No
	Age ≥ 75 years	73	6.7	2.7	No
	70 ≤ SAPS II score < 75	20	1.8	0.0	Yes
	Lactatemia < 1 mmol/L	197	18.1	4.1	No
	6mmol/L ≤ Lactatemia < 9mmol/L	33	3.0	3.0	Yes
	No IH at admission nor history of head trauma	710	65.3	4.8	No
	No IH & severe trauma at admission	385	35.4	4.9	No
	No IH at admission & Creatinine < 150 mmol/L	740	68.0	4.9	No
	History of head trauma & No severe trauma	11	1.0	0.0	Yes
	Creatinine ≥ 100 mmol/L & No severe trauma	40	3.7	2.5	Yes
	40 ≤ Creatinine < 50 & Severe trauma	40	3.7	5.0	Yes
	120 ≤ Creatinine < 140 & Severe trauma	22	2.0	4.5	Yes
Querard et al.	Recipient age < 30 years	391	8.1	3.1	No
Foucher et al.	Transplant before 2015 & A - D centres[b]	376	19.0	2.9	No
	Transplant before 2013 & Recipient age < 50	25	1.3	0.0	Yes
	Transplant in 2013 or 2014 & CIT > 30	21	1.1	4.8	Yes
	Transplant before 2013 & CIT ≥ 20	101	5.1	5.0	No
	Duration on waiting list ≥ 40m & centre A[b]	31	1.6	3.2	Yes
	Duration on waiting list ≥ 60m & Transplant after 2018	31	1.6	3.2	Yes
	30m ≤ Duration on waiting list < 40m & Transplant after 2018	21	1.1	4.8	Yes
	Transplant after 2018 & Recipient age ≥ 70	66	3.3	1.6	Yes
	Transplant in 2011 or 2012 & 75 ≤ Recipient age < 80	30	1.5	3.3	Yes
	Transplant after 2018 & Centre F[b]	25	1.3	4.0	Yes
	Transplant in 2014 & Centre B[b]	32	1.6	0.0	Yes
	Transplant before 2013 & Centre E[b]	89	4.5	3.4	Yes
	HLA-A-B-DR incompatibilities ≥ 5 & Centre A[b]	21	1.1	4.8	Yes
	Transplant in 2010 & History of malignancy	22	1.1	4.5	Yes
	Transplant before 2013 & HLA-A-B-DR incompatibilities < 2	28	1.4	3.6	Yes
	Transplant in 2010 & Retransplant	21	1.1	4.8	Yes
Masset et al.	No history of donor hypertension & A centre[b,c]	11	3.0	0.0	Yes
	No hemodialysis & Donor blood AB group[d]	6	1.7	0.0	Yes
	Donor creatinine < 50 mmol/L & CIT ≥ 20h[e]	8	2.1	0.0	Yes
	15 ≤ CIT < 20 & 55 ≤ Donor age < 65[f]	9	2.4	0.0	Yes

Abbreviations: CIT, Cold Ischemia Time; IH, Intracranial Hypertension; and n, sample size.

[a] New subgroup with respect to the Supplementary Table 2, [b] The centres were anonymised, [c] For time to cardiac complication, [d] For time to malignancy, [e] For delayed graft function occurrence, [f] For time to post-transplant diabetes.

Supplementary Table 6: Results of PCART with α=10% and β=5%.

Authors	Problematic subgroup	n	Part of the total sample (%)	Probability of treatment (%)	New[a]
Léger et al.	No osmotherapy at admission	732	67.3	2.9	No
	Age ≥ 55 years & SAPS II score < 45	150	13.8	2.7	Yes
	Age ≥ 65 years & 45 ≤ SAPS II score < 60	116	10.7	4.3	Yes
	50 ≤ Creatinine < 60 & 25 ≤ SAPS II score < 55	135	12.4	4.4	Yes
	Lactatemia < 1 mmol/L	197	18.1	4.1	No
	No IH at admission nor history of head trauma	710	65.3	4.8	No
	No IH and severe trauma at admission	385	35.4	4.9	No
	No IH at admission & Creatinine < 150 mmol/L	740	68.0	4.9	No
	No IH at admission & SAPS II score < 55	532	48.9	4.1	No
	No IH at admission & Age ≥ 50 years	434	39.9	3.0	Yes
Querard et al.	None detected	-	-	-	-
Foucher et al.	Transplant before 2015 & A - D centres[b]	376	19.0	2.9	No
Masset et al.	None detected	-	-	-	-

Abbreviations: IH, Intracranial Hypertension; and n, sample size.

[a] New subgroup with respect to the Supplementary Table 2, [b] The centres were anonymised.

Annexe H

Résultats non-publiés du chapitre 3

Table 1 : Performances of g-computation (GC) and Inverse Probability of Treatment Weighting (IPTW) approaches under the alternative hypothesis to estimate the ATE effect.

n	selection strategy	method	mean bias				logOR				
			π_0	π_1	$\Delta\pi$	logOR	eSE	RMSE	aSE	coverage	power
100	outcome	GC	-0.0	0.1	0.1	0.006	0.479	0.479	0.432	94.1	16.1
		IPTW	-0.1	0.1	0.1	0.023	0.549	0.549	0.573	96.0	9.5
	treatment	GC	0.1	0.1	0.0	0.006	0.550	0.550	0.490	94.1	13.5
		IPTW	-0.3	0.2	0.5	0.055	0.692	0.694	0.630	93.5	13.5
	outcome and treatment	GC	0.0	-0.0	-0.0	0.003	0.504	0.504	0.464	94.6	14.5
		IPTW	-0.1	0.0	0.1	0.021	0.556	0.557	0.564	95.6	10.4
	all	GC	0.0	0.2	0.1	0.012	0.532	0.532	0.456	93.5	15.2
		IPTW	-0.3	0.3	0.7	0.063	0.712	0.715	0.644	93.7	13.3
300	outcome	GC	0.1	0.1	0.0	0.000	0.256	0.256	0.248	94.3	36.9
		IPTW	0.0	0.1	0.1	0.005	0.277	0.277	0.319	97.5	19.7
	treatment	GC	0.1	0.1	0.0	0.002	0.297	0.297	0.286	93.9	28.8
		IPTW	-0.1	0.1	0.1	0.015	0.369	0.370	0.376	95.3	20.0
	outcome and treatment	GC	0.0	0.0	-0.0	0.001	0.274	0.274	0.266	94.3	32.6
		IPTW	0.0	0.1	0.1	0.006	0.289	0.289	0.317	96.9	21.2
	all	GC	0.1	0.1	0.0	0.002	0.278	0.278	0.265	93.9	32.5
		IPTW	-0.1	0.1	0.1	0.014	0.365	0.366	0.382	96.1	19.9
500	outcome	GC	-0.0	0.0	0.1	0.003	0.196	0.196	0.192	94.2	56.1
		IPTW	-0.1	0.1	0.1	0.008	0.213	0.213	0.245	97.6	35.5
	treatment	GC	-0.0	0.0	0.1	0.004	0.228	0.228	0.222	94.2	45.0
		IPTW	-0.1	0.1	0.2	0.013	0.283	0.284	0.294	95.5	28.5
	outcome and treatment	GC	-0.0	-0.0	0.0	0.003	0.211	0.211	0.206	94.5	51.3
		IPTW	-0.0	0.1	0.1	0.007	0.224	0.224	0.244	96.9	36.4
	all	GC	-0.0	0.1	0.1	0.004	0.213	0.213	0.206	94.0	49.9
		IPTW	-0.1	0.1	0.2	0.014	0.277	0.277	0.297	96.3	28.2
2000	outcome	GC	-0.0	0.0	0.0	0.002	0.095	0.095	0.096	95.4	98.8
		IPTW	0.0	0.1	0.1	0.005	0.102	0.102	0.121	98.1	94.9
	treatment	GC	-0.0	0.0	0.1	0.003	0.111	0.111	0.111	95.3	95.7
		IPTW	0.0	0.1	0.1	0.004	0.134	0.134	0.147	96.9	78.9
	outcome and treatment	GC	-0.0	0.0	0.0	0.002	0.102	0.102	0.103	95.3	97.8
		IPTW	0.0	0.1	0.1	0.005	0.107	0.108	0.121	97.5	94.0
	all	GC	-0.0	0.1	0.1	0.003	0.103	0.103	0.103	95.2	97.7
		IPTW	0.0	0.1	0.1	0.004	0.129	0.129	0.148	97.5	79.8

Ten thousand simulations were performed. Theoretical values : $\pi_0 = 23.6\%, \pi_1 = 31.5\%$ and logOR = 0.390. ATE : average treatment effect in the entire population; n : sample size; π_0 : percentage of event in untreated patients; π_1 : percentage of event for treated patients; $\Delta\pi : \pi_1 - \pi_0$; OR : Odds-ratio; eSE : empirical standard error; RMSE : root mean square error; aSE : asymptotic standard error; coverage : the percentage of 95% confidence intervals including the theoretical value; and power : the percentage of rejection of the null hypothesis.

Table II : Performances of g-computation (GC), Inverse Probability of Treatment Weighting (IPTW) and Propensity Score Matching (PSM) approaches under the alternative hypothesis to estimate the ATT effect.

n	selection strategy	method	mean bias				logOR				
			π_0	π_1	$\Delta\pi$	logOR	eSE	RMSE	aSE	coverage	power
100	outcome	GC	0.3	-0.8	-1.0	-0.071	0.471	0.476	0.481	96.3	9.9
		IPTW	0.6	-0.8	-1.3	-0.076	0.554	0.560	0.637	97.6	5.6
		PSM (mp=80)	-2.2	-3.6	-1.4	-0.022	0.714	0.714	0.840	99.4	1.0
	treatment	GC	0.5	-0.8	-1.3	-0.078	0.537	0.543	0.541	96.3	8.2
		IPTW	0.2	-0.8	-1.0	-0.033	0.735	0.736	0.709	94.5	10.0
		PSM (mp=69)	-2.5	-4.4	-1.9	-0.069	0.741	0.744	0.887	99.5	0.6
	outcome and treatment	GC	0.3	-0.8	-1.1	-0.078	0.498	0.504	0.516	96.5	8.7
		IPTW	0.6	-0.8	-1.3	-0.079	0.558	0.564	0.621	97.0	6.1
		PSM (mp=86)	-1.5	-2.9	-1.4	-0.018	0.727	0.727	0.814	99.1	1.5
	all	GC	0.4	-0.8	-1.2	-0.071	0.516	0.521	0.507	96.0	9.2
		IPTW	0.2	-0.8	-1.0	-0.020	0.768	0.769	0.732	94.5	9.9
		PSM (mp=64)	-2.8	-4.5	-1.7	-0.069	0.717	0.720	0.908	99.7	0.4
300	outcome	GC	-0.2	0.2	0.4	0.005	0.284	0.284	0.276	94.4	31.9
		IPTW	-0.1	0.2	0.3	0.002	0.313	0.313	0.353	97.4	17.3
		PSM (mp=90)	-2.2	-1.5	0.7	0.098	0.441	0.452	0.440	96.2	17.8
	treatment	GC	-0.2	0.2	0.3	0.003	0.328	0.328	0.315	94.6	25.7
		IPTW	-0.3	0.2	0.4	0.017	0.410	0.410	0.412	95.0	17.0
		PSM (mp=76)	-3.6	-2.8	0.7	0.085	0.485	0.492	0.474	96.4	15.2
	outcome and treatment	GC	-0.2	0.2	0.3	0.002	0.306	0.306	0.298	94.8	28.5
		IPTW	-0.1	0.2	0.3	0.001	0.328	0.328	0.349	96.4	18.7
		PSM (mp=92)	-1.9	-1.3	0.5	0.089	0.450	0.459	0.436	95.7	18.7
	all	GC	-0.2	0.2	0.4	0.006	0.306	0.306	0.291	94.0	28.9
		IPTW	-0.3	0.2	0.5	0.021	0.404	0.405	0.419	95.5	16.5
		PSM (mp=76)	-3.6	-2.8	0.8	0.080	0.451	0.458	0.469	97.3	13.5
500	outcome	GC	0.0	-0.1	-0.1	-0.011	0.216	0.216	0.212	94.7	47.3
		IPTW	0.1	-0.1	-0.2	-0.015	0.237	0.237	0.269	97.4	28.2
		PSM (mp=93)	-1.8	-1.3	0.5	0.083	0.330	0.340	0.329	95.4	30.7
	treatment	GC	0.0	-0.1	-0.1	-0.011	0.248	0.249	0.244	94.6	37.9
		IPTW	-0.1	-0.1	-0.0	-0.001	0.305	0.305	0.316	95.8	24.6
		PSM (mp=77)	-3.3	-2.8	0.5	0.070	0.359	0.366	0.355	95.7	25.4
	outcome and treatment	GC	0.0	-0.1	-0.1	-0.013	0.233	0.233	0.230	94.8	42.0
		IPTW	0.1	-0.1	-0.2	-0.015	0.249	0.250	0.268	96.8	29.9
		PSM (mp=93)	-1.5	-1.4	0.2	0.068	0.332	0.339	0.328	95.4	29.5
	all	GC	-0.0	-0.1	-0.1	-0.010	0.231	0.232	0.225	94.4	43.1
		IPTW	-0.1	-0.1	-0.0	-0.002	0.296	0.296	0.320	96.2	22.8
		PSM (mp=78)	-3.3	-2.8	0.5	0.065	0.336	0.342	0.351	96.7	24.3
2000	outcome	GC	-0.0	0.0	0.0	-0.001	0.106	0.106	0.106	95.0	96.6
		IPTW	0.1	0.0	-0.1	-0.005	0.115	0.115	0.133	97.8	88.0
		PSM (mp=96)	-1.7	-0.7	1.0	0.105	0.172	0.202	0.159	88.8	88.0
	treatment	GC	0.0	0.0	-0.0	-0.002	0.122	0.122	0.122	95.1	90.8
		IPTW	-0.0	0.0	0.0	0.002	0.148	0.148	0.156	96.2	73.5
		PSM (mp=78)	-3.4	-2.8	0.6	0.068	0.185	0.198	0.173	92.5	76.6
	outcome and treatment	GC	0.0	0.0	-0.0	-0.002	0.115	0.115	0.115	95.1	93.7
		IPTW	0.1	0.0	-0.1	-0.005	0.123	0.123	0.132	96.6	86.4
		PSM (mp=96)	-1.4	-1.1	0.3	0.072	0.158	0.173	0.159	93.5	85.6
	all	GC	-0.0	0.0	0.0	-0.000	0.113	0.113	0.113	95.0	94.6
		IPTW	-0.0	0.0	0.1	0.002	0.141	0.141	0.156	97.2	74.7
		PSM (mp=79)	-3.4	-2.2	1.2	0.094	0.175	0.199	0.172	92.1	82.4

Ten thousand simulations were performed. Theoretical values : π_0 = 33.4%, π_1 = 42.7% and logOR = 0.400. ATT : average treatment effect for the treated ; n : sample size ; mp : percentage of matched treated ; π_0 : percentage of event in untreated patients ; π_1 : percentage of event for treated patients ; $\Delta\pi$: $\pi_1 - \pi_0$; OR : Odds-ratio ; eSE : empirical standard error ; RMSE : root mean square error ; aSE : asymptotic standard error ; coverage : the percentage of 95% confidence intervals including the theoretical value ; and power : the percentage of rejection of the null hypothesis.

Supplementary Table II : Performances of g-computation (GC) and Inverse Probability of Treatment Weighting (IPTW) approaches under the null hypothesis to estimate the ATE effect.

n	selection strategy	method	mean bias				logOR				
			π_0	π_1	$\Delta\pi$	logOR	eSE	RMSE	aSE	coverage	type I
100	outcome	GC	0.1	0.1	-0.1	0.004	0.498	0.498	0.447	94.3	5.7
		IPTW	-0.1	0.1	0.2	0.028	0.572	0.572	0.584	96.2	3.8
	treatment	GC	0.1	0.2	0.1	0.014	0.575	0.575	0.505	93.8	6.2
		IPTW	-0.4	0.4	0.7	0.068	0.720	0.723	0.639	92.7	7.3
	outcome and treatment	GC	0.1	0.1	-0.0	0.008	0.524	0.524	0.479	94.3	5.7
		IPTW	-0.1	0.1	0.2	0.029	0.580	0.580	0.576	95.3	4.7
	all	GC	0.2	0.2	0.0	0.011	0.555	0.555	0.472	93.6	6.4
		IPTW	-0.5	0.4	0.9	0.075	0.732	0.735	0.653	93.3	6.7
300	outcome	GC	0.0	0.1	0.1	0.005	0.263	0.263	0.256	94.7	5.3
		IPTW	-0.0	0.1	0.1	0.010	0.288	0.288	0.326	97.4	2.6
	treatment	GC	0.1	0.1	0.0	0.004	0.305	0.305	0.294	94.6	5.4
		IPTW	-0.1	0.0	0.2	0.016	0.376	0.376	0.381	95.4	4.6
	outcome and treatment	GC	0.0	0.1	0.0	0.004	0.281	0.281	0.275	95.0	5.0
		IPTW	-0.0	0.1	0.1	0.008	0.299	0.299	0.324	96.6	3.4
	all	GC	0.0	0.1	0.1	0.005	0.286	0.286	0.274	94.6	5.4
		IPTW	-0.1	0.0	0.2	0.016	0.372	0.372	0.386	96.1	3.9
500	outcome	GC	0.0	0.0	-0.0	0.000	0.204	0.204	0.199	94.4	5.6
		IPTW	0.0	0.0	0.0	0.005	0.221	0.221	0.250	97.5	2.5
	treatment	GC	0.0	0.0	-0.0	0.001	0.236	0.236	0.228	94.3	5.7
		IPTW	0.0	0.1	0.0	0.006	0.289	0.289	0.299	95.7	4.3
	outcome and treatment	GC	0.0	0.0	-0.0	0.001	0.219	0.219	0.213	94.3	5.7
		IPTW	-0.0	0.1	0.1	0.005	0.230	0.231	0.249	96.6	3.4
	all	GC	0.0	0.0	-0.0	0.001	0.221	0.221	0.213	94.4	5.6
		IPTW	0.0	0.0	0.0	0.005	0.284	0.284	0.301	96.1	3.9
2000	outcome	GC	-0.0	0.0	0.0	0.002	0.100	0.100	0.099	95.0	5.0
		IPTW	-0.0	0.0	0.1	0.004	0.106	0.106	0.124	97.8	2.2
	treatment	GC	-0.0	0.0	0.0	0.002	0.114	0.114	0.114	94.9	5.1
		IPTW	-0.0	0.0	0.0	0.002	0.137	0.137	0.150	96.9	3.1
	outcome and treatment	GC	-0.0	0.0	0.0	0.002	0.106	0.106	0.106	94.9	5.1
		IPTW	-0.0	0.1	0.1	0.004	0.111	0.111	0.123	97.2	2.8
	all	GC	-0.0	0.0	0.0	0.002	0.107	0.107	0.107	94.9	5.1
		IPTW	-0.0	0.0	0.0	0.0	0.132	0.132	0.150	97.3	2.7

Ten thousand simulations were performed. Theoretical values : $\pi_0 = \pi_1 = 23.6\%$ and logOR = 0.000. ATE : average treatment effect in the entire population; n : sample size; π_0 : percentage of event in untreated patients; π_1 : percentage of event for treated patients; $\Delta\pi : \pi_1 - \pi_0$; OR : Odds-ratio; eSE : empirical standard error; RMSE : root mean square error; aSE : asymptotic standard error; coverage : the percentage of 95% confidence intervals including the theoretical value; and type I : the percentage of rejection of the null hypothesis.

Supplementary Table III : Performances of g-computation (GC), Inverse Probability of Treatment Weighting (IPTW) and Propensity Score Matching (PSM) approaches under the null hypothesis to estimate the ATT effect.

n	selection strategy	method	mean bias				logOR				
			π_0	π_1	$\Delta\pi$	logOR	eSE	RMSE	aSE	coverage	type I
100	outcome	GC	-0.1	-0.3	-0.2	-0.010	0.488	0.488	0.499	96.6	3.4
		IPTW	0.1	-0.3	-0.4	-0.012	0.564	0.564	0.654	98.0	2.0
		PSM (mp=79)	-2.8	-2.7	0.1	0.014	0.725	0.725	0.858	99.8	0.2
	treatment	GC	-0.1	-0.3	-0.2	-0.006	0.549	0.549	0.562	96.8	3.2
		IPTW	-0.2	-0.3	-0.1	0.030	0.739	0.739	0.728	95.1	4.9
		PSM (mp=68)	-3.5	-3.2	0.3	0.023	0.753	0.753	0.912	99.7	0.3
	outcome and treatment	GC	-0.1	-0.3	-0.2	-0.015	0.511	0.511	0.537	97.0	3.0
		IPTW	0.1	-0.3	-0.4	-0.017	0.565	0.565	0.639	97.8	2.2
		PSM (mp=86)	-2.2	-2.1	0.1	0.013	0.748	0.748	0.829	99.5	0.5
	all	GC	-0.2	-0.3	-0.2	0.000	0.537	0.537	0.523	96.3	3.7
		IPTW	-0.2	-0.3	-0.1	0.042	0.773	0.774	0.749	94.8	5.2
		PSM (mp=64)	-3.7	-3.4	0.3	0.027	0.737	0.737	0.934	99.9	0.1
300	outcome	GC	0.0	0.1	0.1	0.001	0.294	0.294	0.287	94.6	5.4
		IPTW	0.1	0.1	0.0	-0.002	0.323	0.323	0.363	97.6	2.4
		PSM (mp=90)	-2.0	-1.4	0.6	0.034	0.440	0.442	0.444	96.6	3.4
	treatment	GC	0.1	0.1	0.0	-0.003	0.339	0.339	0.327	95.0	5.0
		IPTW	-0.1	0.1	0.2	0.017	0.413	0.413	0.421	95.3	4.7
		PSM (mp=76)	-3.4	-2.6	0.8	0.042	0.489	0.490	0.480	96.3	3.7
	outcome and treatment	GC	0.1	0.1	0.1	-0.003	0.318	0.318	0.310	94.9	5.1
		IPTW	0.1	0.1	-0.0	-0.004	0.338	0.338	0.359	96.5	3.5
		PSM (mp=92)	-1.7	-1.3	0.4	0.023	0.454	0.455	0.440	95.8	4.2
	all	GC	0.0	0.1	0.1	0.002	0.315	0.315	0.302	94.4	5.6
		IPTW	-0.2	0.1	0.3	0.022	0.405	0.406	0.428	96.0	4.0
		PSM (mp=76)	-3.5	-2.6	0.9	0.047	0.459	0.461	0.476	97.0	3.0
500	outcome	GC	-0.0	0.0	0.0	-0.001	0.225	0.225	0.221	94.8	5.2
		IPTW	0.1	0.0	-0.1	-0.004	0.246	0.246	0.277	97.4	2.6
		PSM (mp=92)	-1.9	-1.1	0.8	0.042	0.330	0.333	0.332	95.8	4.2
	treatment	GC	0.1	0.0	-0.1	-0.004	0.258	0.258	0.252	94.5	5.5
		IPTW	-0.0	0.0	0.0	0.004	0.312	0.312	0.323	95.4	4.6
		PSM (mp=77)	-3.3	-2.5	0.8	0.042	0.366	0.368	0.360	95.4	4.6
	outcome and treatment	GC	0.1	0.0	-0.0	-0.005	0.242	0.243	0.239	94.8	5.2
		IPTW	0.1	0.0	-0.1	-0.006	0.258	0.258	0.275	96.6	3.4
		PSM (mp=93)	-1.6	-1.2	0.4	0.023	0.334	0.335	0.331	95.5	4.5
	all	GC	-0.0	0.0	0.0	-0.001	0.240	0.240	0.233	94.5	5.5
		IPTW	-0.0	0.0	0.0	0.006	0.303	0.303	0.326	96.3	3.7
		PSM (mp=78)	-3.3	-2.4	0.9	0.046	0.343	0.346	0.356	96.6	3.4
2000	outcome	GC	0.0	0.0	0.0	-0.000	0.110	0.110	0.110	95.0	5.0
		IPTW	0.2	0.0	-0.1	-0.006	0.120	0.120	0.136	97.7	2.3
		PSM (mp=96)	-1.7	-0.7	1.0	0.051	0.172	0.180	0.160	92.0	8.0
	treatment	GC	0.1	0.0	-0.1	-0.003	0.127	0.127	0.126	94.8	5.2
		IPTW	0.0	0.0	0.0	0.001	0.150	0.150	0.159	96.3	3.7
		PSM (mp=78)	-3.4	-2.5	0.9	0.045	0.189	0.194	0.175	92.5	7.5
	outcome and treatment	GC	0.1	0.0	-0.0	-0.003	0.119	0.119	0.119	94.9	5.1
		IPTW	0.2	0.0	-0.1	-0.006	0.127	0.127	0.136	96.5	3.5
		PSM (mp=96)	-1.3	-1.0	0.3	0.017	0.160	0.161	0.160	95.0	5.0
	all	GC	0.0	0.0	0.0	-0.000	0.118	0.118	0.116	95.1	4.9
		IPTW	0.0	0.0	0.0	0.001	0.144	0.144	0.159	97.4	2.6
		PSM (mp=79)	-3.4	-1.9	1.5	0.075	0.177	0.192	0.174	93.0	7.0

Ten thousand simulations were performed. Theoretical values : $\pi_0 = \pi_1 = 33.4\%$ and logOR = 0.000. ATE : average treatment effect for the treated ; n : sample size ; mp : percentage of matched treated ; π_0 : percentage of event in untreated patients ; π_1 : percentage of event for treated patients ; $\Delta\pi : \pi_1 - \pi_0$; OR : Odds-ratio ; eSE : empirical standard error ; RMSE : root mean square error ; aSE : asymptotic standard error ; coverage : the percentage of 95% confidence intervals including the theoretical value ; and type I : the percentage of rejection of the null hypothesis.

DOCTORAT BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE NANTES

Titre : Prédiction contrefactuelle pour l'estimation causale à partir de données de vie réelle

Mots clefs : Biais de confusion, Etudes de simulation, G-computation, Inférence causale, Machine learning, Score de propension.

Résumé : L'absence de randomisation pour les données de vie réelle complique l'analyse statistique et nécessite l'utilisation de méthodes complexes. L'application web Plug-Stat® facilite ce type d'analyse en proposant des interfaces intuitives pour les non-spécialistes. La présente thèse cherche à optimiser Plug-Stat® en automatisant le plus possible l'étape d'estimation causale. Trois travaux étudiant le comportement de ces méthodes et un quatrième proposant un outil d'aide à la décision sont présentés. Le premier travail compare par simulations les méthodes d'inférence causale les plus courantes selon différents ensembles d'ajustement. Le second travail compare différentes approches de *machine learning* en combinaison avec la *g-computation* pour éviter les biais liés à une mauvaise spécification du modèle. Le troisième travail présente le développement d'un estimateur de *g-computation* en présence de censure à droite. Ce nouvel estimateur est également combiné avec un score de propension pour former un estimateur doublement robuste. Ces trois méthodes sont ensuite comparées par simulations. Le dernier travail propose un algorithme d'aide à la vérification de l'hypothèse de positivité. Au final, la *g-computation* semble être une méthode à considérer pour l'automatisation de Plug-Stat®. Elle ne nécessite pas d'hypothèse d'équilibre et le *machine learning* évite les problèmes de spécification. Enfin, la vérification de la positivité est automatisée au moyen d'arbres de décision pour permettre à l'investigateur de redéfinir sa population d'étude.

Title: Counterfactual prediction in causal estimation from real-life data

Keywords: Causal inference, Confounding bias, G-computation, Machine learning, Propensity score, Simulation studies.

Abstract: The lack of randomisation in observational studies makes statistical analysis harder and requires complex methods. Plug-Stat®, a web application, facilitates such analyses by proposing intuitive interfaces. This thesis searches to optimise Plug-Stat® by maximally automating the causal estimation step. This thesis presents three works investigating the behaviour of these specific methods, and a fourth one showing a decision-making tool for causal studies. The first study compares by simulations the most common causal inference methods across several adjustment sets. The second work compares several machine learning approaches associated with the g-computation to avoid model misspecifications. The third work presents the development of a g-computation estimator in the presence of right censoring. This novel estimator is also associated with a propensity score to form a doubly robust estimator. The three methods are then compared through simulations. The last work proposes an algorithm able to check potential positivity violations. The g-computation should be considered for the automation of Plug-Stat®. It is free of the balancing assumption, and machine learning avoids misspecifications. Positivity checking is also automated through decision trees for helping the investigator to redefine his study population.