Université Sorbonne
Paris Cité

Thèse de doctorat de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole Doctorale Pierre Louis de Santé Publique :

Epidémiologie et Sciences de l'Information Biomédicale (ED 393)

Laboratoire – Equipe de recherche :

INSERM UMR 1123, ECEVE

Epidémiologie clinique et évaluation économique appliquée aux populations vulnérables

Utilisation du score de propension et du score pronostique en pharmacoépidémiologie

Par David Hajage

Thèse de doctorat de Biostatistique

Dirigée par le Professeur Florence Tubach (Directeur de Thèse)

et Yann De Rycke (Co-Directeur de Thèse)

Présentée et soutenue publiquement à Paris le 2 février 2017

Président du jury	Matthieu Resche-Rigon	PU-PH
Rapporteur	Stefan Michiels	Directeur de recherche
Rapporteur	Bruno Falissard	PU-PH
Examinateur	Pascale Tubert-Bitter	Directeur de recherche
Examinateur	Laurence Meyer	PU-PH
Directeur de Thèse	Florence Tubach	PU-PH
Co-Directeur de Thèse	Yann De Rycke	Ingénieur de recherche

"There is a natural connection between causes and strategies that should be maintained : if one wants to obtain a goal, it is good (in the pre-utility sence of good) strategy to introduce a cause for that goal."

Cartwright, Nancy.
"Causal Laws and Effective Strategies."
Noûs 13, no. 4 (1979) : 419–37.

Remerciements

En tout premier lieu, je remercie mes directeurs de thèse, Florence Tubach et Yann De Rycke.

Florence, ce coup de fil me proposant de rejoindre ton équipe a probablement été l'étape la plus importante de mon cursus. Je profite de ton enseignement, de ta rigueur scientifique, de ton soutien sans faille et de ton humanité depuis maintenant cinq ans. Avoir ta confiance malgré mes doutes permanents ne cesse de m'étonner, et m'honore tout à la fois.

Yann, les poules qui ont la chance d'avoir croisé ton chemin le savent : ton intelligence et ta créativité n'ont d'égal que ta gentillesse et ton humour. Tu es pour moi un ami, un modèle et un moteur. Je souhaite que la fin de ce travail ne marque pas seulement la fin d'une belle aventure, mais qu'elle marque également le début de nombreuses autres joyeuses découvertes (et soirées au Waïkiki) en ta compagnie.

Avoir un mentor est important, en avoir deux aussi complémentaires et bienveillants à mon égard est une chance rare dans ce milieu. Merci à vous deux d'avoir cru en moi.

Je remercie Stefan Michiels, Bruno Falissard, Pascale Tubert-Bitter, Matthieu Resche-Rigon et Laurence Meyer de m'avoir fait l'honneur de s'intéresser à mon travail et de participer à ce jury afin de me faire part de leurs remarques constructives.

Je remercie également Guillaume Chauvet, pour sa gentillesse et sa disponibilité. Avoir le quart de ton intelligence et de tes connaissances réglerait pas mal de mes problèmes. En attendant, merci de les avoir utilisées pour faire avancer ce travail.

Candice, chère collègue de bureau, merci pour ton soutien moral quotidien, pour tes conseils méthodo et de bricolage, pour le café, pour les blagues et les fous rires, bref merci pour ta disponibilité dans tous les domaines et de tous les instants.

Mes remerciements vont aussi à France Mentré, pour avoir grandement facilité la réalisation de cette thèse en me permettant d'utiliser son serveur de calcul, pour sa compréhension lors de mes demandes de réinscription, et sa patience pour répondre à mes nombreuses questions concernant l'organisation de ma soutenance.

Merci à l'équipe du DEBRC ex DEPIREC ex DEBRC pour cette ambiance de travail agréable et solidaire, et tous les bons moments passés. Vous allez me manquer.

Merci Débo, d'être auprès de nous, toujours, et d'avoir mis à profit ton trouble de l'*in*attention pour corriger ce manuscrit.

A mes parents et mes frères.

A Maud, mon épouse, Adam et Juliette, mes enfants. Pour votre amour inconditionnel et votre compréhension durant ces années difficiles par bien des aspects. Sans vous rien n'a de sens.

Résumé

Résumé français

Les études observationnelles en pharmacoépidémiologie sont souvent mises en place pour évaluer un médicament mis sur le marché récemment ou concurrencé par de nombreuses alternatives thérapeutiques. Cette situation conduit à devoir évaluer l'effet d'un médicament dans une cohorte comprenant peu de sujets traités, c'est à dire une population où l'exposition d'intérêt est rare. Afin de prendre en compte les facteurs de confusion dans cette situation, certains auteurs déconseillent l'utilisation du score de propension au profit du score pronostique, mais cette recommandation ne s'appuie sur aucune étude évaluant spécifiquement les faibles prévalences de l'exposition, et ignore le type d'estimation, conditionnelle ou marginale, fournie par chaque méthode d'utilisation du score pronostique.

La première partie de ce travail évalue les méthodes basées sur le score de propension pour l'estimation d'un effet marginal en situation d'exposition rare. La deuxième partie évalue les performances des méthodes basées sur le score pronostique rapportées dans la littérature, introduit de nouvelles méthodes basées sur le score pronostique adaptées à l'estimation d'effets conditionnels ou marginaux, et les compare aux performances des méthodes basées sur le score de propension. La dernière partie traite des estimateurs de la variance des effets du traitement. Nous présentons les conséquences liées à la non prise en compte de l'étape d'estimation du score de propension et du score pronostique dans le calcul de la variance. Nous proposons et évaluons de nouveaux estimateurs tenant compte de cette étape.

Résumé anglais

Pharmacoepidemiologic observational studies are often conducted to evaluate newly marketed drugs or drugs in competition with many alternatives. In such cohort studies, the exposure of interest is rare. To take into account confounding factors in such settings, some authors advise against the use of the propensity score in favor of the prognostic score, but this recommendation is not supported by any study especially focused on infrequent exposures and ignores the type of estimation provided by each prognostic score-based method.

The first part of this work evaluates the use of propensity score-based methods to estimate the marginal effect of a rare exposure. The second part evaluates the performance of the prognostic score based methods already reported in the literature, compares them with the propensity score based methods, and introduces some new prognostic score-based methods intended to estimate conditional or marginal effects. The last part deals with variance estimators of the treatment effect. We present the opposite consequences of ignoring the estimation step of the propensity score and the prognostic score. We show some new variance estimators accounting for this step.

Production scientifique dans le cadre de la thèse

Articles

1) Hajage D, Tubach F, Steg PG, Bhatt DL, De Rycke Y. On the use of propensity scores in case of rare exposure. *BMC Med Res Methodol*. 2016 Mar 31 ;16 :38.

2) Hajage D, De Rycke Y, Chauvet G, Tubach F. Estimation of conditional and marginal odds ratios using the prognostic score. *Stat Med*. 2016 Nov 17.

3) Hajage D, Chauvet G, Tubach F, De Rycke Y. Variance estimation for weighted propensity score estimators. *Soumis, under review*.

Présentations orales

1) Utilisation des scores de propension en cas d'exposition rare. *EpiClin 8/21ème Journées des Statisticiens des Centres de Lutte Contre le Cancer*, Bordeaux. 2014.

2) Les méthodes de prise en compte des scores pronostiques : une étude de simulation. *EpiClin 10/23ème Journées des Statisticiens des Centres de Lutte Contre le Cancer*, Strasbourg. 2016.

3) More efficient methods to take into account prognostic scores in observational studies. *37th annual conference of the International Society for Clinical Biostatistics*, Birmingham. 2016.

Présentations affichées

1) On the use of propensity score method in case of rare exposure. *30th International Conference on Pharmacoepidemiology & Therapeutic Risk Management*, Taïwan. 2014.

2) Better alternatives to existing methods to take into account prognostic scores in observational studies. *32th International Conference on Pharmacoepidemiology & Therapeutic Risk Management*, Dublin. 2016.

Liste des abréviations

AMM Autorisation de mise sur le marché

ATE Average treatment effect on the whole population (en français : effet marginal du traitement dans l'ensemble de la population)

ATT Average treatment effect on the treated population (en français : effet marginal du traitement chez les traités)

CTE Conditional treatment effect (en français : effet conditionnel du traitement)

DR Différence de risques

HR Hazard ratio

ITE Individual treatment effect (en français : effet individuel du traitement)

MTE Marginal treatment effect (en français : effet marginal du traitement)

OR Odds ratio

RR Risque relatif

SNIIRAM Système National d'Information Interrégimes de l'Assurance Maladie

SPN Score pronostique

SPP Score de propension

Table des matières

Remerciements **2**

Résumé **4**

Résumé français 4

Résumé anglais 5

Production scientifique dans le cadre de la thèse **6**

Articles 6

Présentations orales 7

Présentations affichées 7

Liste des abréviations **8**

1 Introduction **12**

1.1 Pharmacoépidémiologie 12

1.1.1 Définition et objectifs 12

1.1.2 Sources de données en pharmacoépidémiologie 13

1.1.3 Plans d'études pharmacoépidémiologiques 13

1.1.4 Types d'études pharmacoépidémiologiques 14

1.2 Causalité 15

1.2.1 Définition 15

1.2.2 Facteurs de confusion et biais d'indication 17

1.2.3 Différentes mesures de l'effet d'une exposition 18

1.2.3.1 Selon le paramètre de mesure utilisé 18

1.2.3.2 Selon la population d'intérêt 19
1.2.4 Inférence causale en situation expérimentale : les essais randomisés . 20
1.3 Méthodes d'inférence causale pour l'analyse des études observationnelles . . 21
1.4 Spécificités de l'utilisation de ces méthodes en pharmacoépidémiologie . . . 22
1.5 Organisation du document 23

2 Score de propension **24**
2.1 Définition et hypothèses 24
2.2 Estimation et utilisation 25
2.2.1 Estimation du score de propension 25
2.2.2 Utilisation du score de propension 26
2.3 Brève revue de la littérature 28
2.4 Utilisation du score de propension quand la prévalence de l'exposition est faible 29
Article 1 30

3 Score pronostique **49**
3.1 Définition et hypothèses 49
3.2 Estimation et utilisation 50
3.2.1 Estimation du score pronostique 50
3.2.2 Utilisation du score pronostique 51
3.2.3 Brève revue de la littérature 52
3.3 Evaluation des méthodes d'utilisation existantes et développement de nouvelles méthodes d'utilisation 53
Article 2 56

4 Estimateurs de la variance de l'effet **87**
4.1 Problèmes liés à la non la prise en compte de l'étape d'estimation du score . 87
4.1.1 Conséquence avec le score pronostique 87
4.1.2 Conséquence avec le score de propension 88

4.2 Méthode de linéarisation pour l'estimation de la variance d'un estimateur complexe 89
4.3 Application au score pronostique 90
4.4 Application à la pondération sur le score de propension 92
4.4.1 Estimateurs de l'effet du traitement 94
4.4.2 Estimateurs de variance de Lunceford & Davidian (2004) et de Williamson et al. (2014) 94
4.4.3 Estimateurs obtenus par linéarisation 96
4.4.4 Etude de simulation 97
Article 3 97

5 Conclusion **132**
5.1 Résumé de la thèse 132
5.2 Perspectives 135

Bibliographie **138**

1 | Introduction

1.1 Pharmacoépidémiologie

1.1.1 Définition et objectifs

Strom et al. (2012) définissent la pharmacoépidémiologie comme l'étude de l'utilisation et des effets (en termes de bénéfice ou de risque) des médicaments chez un grand nombre d'individus, en conditions réelles. Une étude pharmacoépidémiologique est menée après l'autorisation de mise sur le marché (post-AMM) d'un médicament. Une grande taille d'échantillon issue d'une population non sélectionnée permet d'apporter des informations plus précises ou nouvelles par rapports aux données accumulées au cours des études pré-AMM :

— quantification plus précise de l'effet d'un médicament et de l'incidence des évènements indésirables (Heerdink et al. 2002) ;
— information concernant des patients généralement exclus des études pré-AMM (notamment les personnes âgées, les enfants, les femmes enceintes ou les sujets présentant des pathologies associées) (McKenzie et al. 1976) ;
— description des indications et des modalités de prescription ou d'utilisation, et quantification du mésusage (Lunde & Baksaas 1988) ;
— découverte d'effets (bénéfiques ou délétères) rares ou retardés, des conséquences du mésusage, d'interactions médicamenteuses inattendues (Herbst et al. 1971) ;

— comparaison à d'autres traitements existants et prescrits dans la même indication mais n'ayant pas fait l'objet d'études expérimentales (Cameron et al. 2015) ;
— évaluation médico-économique (Reed et al. 2008).

Si ces informations sont mesurées à partir d'une étude menée selon un plan d'étude et d'analyse adéquat (c'est-à-dire en assurant la validité interne en limitant les biais systématiques [1]) et incluant un échantillon de sujets représentatif de la population cible (pour assurer la validité externe), celles-ci peuvent être extrapolées à la population cible et orienter les décisions de soins et de politiques de santé publique.

1.1.2 Sources de données en pharmacoépidémiologie

Les sources de données pour les études en pharmacoépidémiologie sont multiples, mais peuvent être divisées en deux catégories :

— les données issues d'études ad-hoc, conçues spécifiquement pour répondre à une question de recherche ;
— les bases de données déjà existantes, non spécifiques mais comportant des informations sur les médicaments ; il peut alors s'agir de données issues d'une étude initialement conçue pour répondre à un autre objectif de recherche (analyse post-hoc), ou de données médico-administratives récoltées dans un objectif hors recherche, comme les données de remboursement de l'assurance maladie (Système National d'Information Interrégimes de l'Assurance Maladie ou SNIIRAM) ou les données d'hospitalisation (Programme de Médicalisation des Systèmes d'Information ou PMSI).

1.1.3 Plans d'études pharmacoépidémiologiques

La pharmacoépidémiologie utilise les méthodes de l'épidémiologie clinique pour l'étude de l'utilisation et de l'effet des médicaments. Les études observationnelles longitudinales de

1. comme les biais de sélection, de suivi, d'attrition, d'évaluation ou d'interprétation.

type **cohorte**, menées sur de larges populations non sélectionnées, y jouent un rôle majeur. D'autres plans d'étude non-expérimentaux, comme les études cas-témoins, les études « case-only » ou les séries temporelles (dont le niveau de preuve est plus faible et qui ne seront pas abordées dans ce document), ainsi que des essais randomisés pragmatiques sont aussi utilisés (Tubach et al. 2011).

1.1.4 Types d'études pharmacoépidémiologiques

De manière simple, on peut distinguer deux types d'étude pharmacoépidémiologiques :

- les **études descriptives**, qui décrivent par exemple les caractéristiques des personnes utilisant un médicament d'intérêt (prévalence du mésusage en particulier), la fréquence de survenue de certains évènements comme les effets secondaires chez des patients traités par un médicament d'intérêt, ou la durée de maintien des traitements chroniques ;
- les **études étiologiques** (ou analytiques), qui étudient l'association entre l'exposition à un médicament et la survenue d'un évènement d'intérêt (critère d'efficacité ou de sécurité).

Ces études ne nécessitent pas obligatoirement l'utilisation de méthodologies d'analyse particulières (autres que la prise en compte du plan d'échantillonnage de la population d'étude). Quantifier la fréquence du mésusage parmi les utilisateurs d'un médicament est, par exemple, une information descriptive directement interprétable. En revanche, l'estimation d'une association n'est bien souvent interprétable que si elle a tenu compte des **facteurs de confusion**, c'est-à-dire des variables qui, non prises en compte, peuvent créer une association apparente ou, au contraire, masquer une association réelle, c'est-à-dire **causale**.

1.2 Causalité

1.2.1 Définition

Des critères empiriques de causalité d'une association entre une exposition (par exemple un médicament) et un évènement (par exemple un décès) ont été listés et discutés par Austin Bradford Hill en 1965. Aucun critère n'est nécessaire ni suffisant pour conclure à la causalité, mais les chances que l'association observée soit liée au hasard ou à un biais systématique sont plus faibles en présence du critère qu'en son absence. Abondamment discutés et critiqués, ils sous-tendent depuis 50 ans le raisonnement (ou *inférence*) causal dans le domaine de l'épidémiologie (et donc, par extension, dans celui de la pharmacoépidémiologie). Nous les listons ci-après, en les définissant brièvement.

Tableau 1.1 : Critères de causalité (Hill 1965).

Critère	Définition
La force de l'association	La taille de l'association observée, mesurée par exemple par la valeur d'une différence ou d'un rapport de deux probabilités d'évènement, est importante.
La reproductibilité	L'association observée a été répliquée indépendamment par différentes études, à différents endroits et à différents moments.
La spécificité	L'effet observé n'est dû qu'à une seule et même exposition de nature causale.
La temporalité	L'exposition causale précède son effet.
Le gradient biologique	Il existe une relation, par exemple monotone, entre le niveau d'exposition (ou « dose ») et le risque d'évènement.

Critère	Définition
La plausibilité biologique	Il existe un mécanisme biologique expliquant l'association observée.
La cohérence biologique	L'association observée est en accord avec les connaissances déjà accumulées sur la maladie considérée.
L'expérimentation	L'association a été observée au moyen d'un plan d'étude expérimental.
L'analogie	L'association observée ressemble à d'autres relations causales déjà décrites, ou à leurs mécanismes.

Aucun de ces critères n'est jugé nécessaire (dans le sens « obligatoirement documenté ») ni suffisant par leur auteur, et aucun ne permet donc d'affirmer ou d'infirmer la causalité d'une association. Six de ces critères sont communément considérés comme forts : la force, la reproductibilité, la temporalité, le gradient, la plausibilité et la cohérence biologique. Ces critères ont récemment été revisités par John P. A. Ioannidis à l'occasion de leur cinquantième anniversaire (Ioannidis 2016). Seuls deux critères « survivent » à cet examen (la reproductibilité et l'expérimentation), les autres étant, selon le point de vue de l'auteur, au mieux rarement documentables et au pire source d'erreurs. Ainsi, toujours selon cet auteur, une forte association reflète souvent une étude présentant des biais importants, et est rarement reproductible. Les associations causales sont rarement spécifiques, les maladies étant souvent multifactorielles. La temporalité est souvent difficile à documenter (biais protopathique). L'observation d'un gradient biologique masque souvent les dizaines de tentatives de modélisation qui ne la retrouvaient pas. La plausibilité ne reflète que l'état des connaissances actuelles. La cohérence est en pratique difficile à définir, et subjective. Enfin l'analogie peut ne conduire qu'à de fausses interprétations ou à la perpétuation des erreurs précédentes.

Si nous admettons, à la suite de Ioannidis, que l'expérimentation et la reproductibilité sont les deux piliers les plus importants sous-tendant l'interprétation causale d'une association, l'objectif étiologique de la pharmacoépidémiologie peut être reformulé comme suit : tenter de reproduire, dans un contexte non-expérimental et en conditions réelles d'utilisation des médicaments, les résultats observés dans les essais contrôlés randomisés réalisés en pré-AMM, plan expérimental qui permet de s'affranchir des facteurs de confusion. En post-AMM, des méthodes ont été développées pour s'en approcher à partir de données observationnelles de type cohorte, soumises aux facteurs de confusion et en particulier au **biais d'indication**.

1.2.2 Facteurs de confusion et biais d'indication

Un facteur de confusion F est défini comme un facteur associé à la fois à une exposition d'intérêt T (mais sans en être la conséquence) et un critère de jugement d'intérêt Y. En d'autres termes, si le niveau d'exposition T dépend d'un facteur F qui influence également le devenir Y des sujets, distinguer l'effet de l'exposition de l'effet de ce facteur devient difficile. Si l'on prend l'exemple d'une exposition binaire (exposé « oui » ou « non »), la comparaison directe, ou naïve, du devenir des exposés et des non exposés fournit une estimation biaisée de l'effet de l'exposition en présence d'un (ou plusieurs) facteur(s) de confusion.

Un exemple typique d'étude présentant des facteurs de confusion est une étude observationnelle dans laquelle un médicament est prescrit préférentiellement à des sujets ayant un risque plus élevé (ou plus faible) de présenter l'évènement d'intérêt (dans le cas d'un critère de jugement binaire) : il s'agit de la définition habituelle donnée au **biais d'indication**.

1.2.3 Différentes mesures de l'effet d'une exposition

La suite de cette introduction considère le cas particulier d'une exposition binaire ($T = 1$ en cas d'exposition, et $T = 0$ en l'absence d'exposition). Ce type d'exposition est très répandu dans la littérature médicale, puisqu'elle est applicable à toutes les situations où la question scientifique est de comparer deux alternatives thérapeutiques. Cette perte de généralisabilité est également compensée par des explications rendues plus simples et didactiques.

1.2.3.1 Selon le paramètre de mesure utilisé

Différentes mesures d'association entre une exposition T et un critère de jugement d'intérêt Y peuvent être calculées, selon la nature du critère de jugement Y. Certaines de ces estimations fréquemment utilisées sont listées ci-après.

Si le critère de jugement Y est quantitatif, la mesure d'association la plus utilisée est la différence de moyenne entre les sujets exposés et les sujets non exposés : $E(Y|T = 1) - E(Y|T = 0)$. D'autres mesures pourraient être utilisées (comme l'aire sous la courbe ROC), mais le sont en pratique très rarement dans le contexte d'évaluation de l'effet d'un traitement.

Si le critère de jugement Y est binaire ($Y = 1$ en cas de survenue de l'événement au cours du suivi, $Y = 0$ en l'absence de survenue de l'événement au cours du même suivi), plusieurs types de mesures sont utilisés, comme la différence de risques ($DR = E(Y|T = 1) - E(Y|T = 0)$), le risque relatif ($RR = E(Y|T = 1)/E(Y|T = 0)$), ou l'odds ratio ($OR = \frac{E(Y|T=1)}{1-E(Y|T=1)} / \frac{E(Y|T=0)}{1-E(Y|T=0)}$).

Si le critère de jugement est censuré (délai de survenue d'un évènement Y possiblement censuré à droite), plusieurs types de mesures absolues (comme la différence des survies sans évènement moyennes ou médianes) ou relatives (comme le hazard ratio) sont utilisés.

1.2.3.2 Selon la population d'intérêt

La mesure d'une association dépend également des sujets auxquels elle se rapporte [2]. On peut, entre autres, définir :

L'effet individuel ou « individual treatment effect » (ITE). Pour un sujet i, l'effet individuel d'un traitement est la différence entre la mesure du critère de jugement Y quand le sujet est exposé, et la mesure du critère de jugement Y quand le sujet est non exposé. Si on note $Y_{i,1}$ et $Y_{i,0}$ ces deux valeurs, l'effet individuel est simplement égal à $Y_{i,1} - Y_{i,0}$. Comme il est impossible d'observer $Y_{i,1}$ et $Y_{i,0}$ en même temps (l'une de ces valeurs est dite *contrefactuelle*), l'effet individuel n'est pas observable.

L'effet conditionnel ou « conditional treatment effect » (CTE). L'effet conditionnel d'un traitement est l'effet du traitement chez un certain profil d'individus, défini par les caractéristiques mesurées et prises en compte dans cette estimation. L'effet conditionnel peut être identique pour tous les profils d'individus, ou différent s'il existe une *hétérogénéité de l'effet*, par exemple une interaction entre le traitement et une ou plusieurs des caractéristiques définissant les profils d'individus. Il mesure l'effet du passage, pour un certain profil d'individus, du statut de « non traité » à celui de « traité » et est donc souvent interprété comme un effet sujet-spécifique.

L'effet marginal ou « marginal treatment effect » (MTE). L'effet marginal d'un traitement est l'effet du traitement dans un groupe d'individus, c'est-à-dire au niveau d'une population. La valeur de l'effet marginal dépend donc du groupe d'individus dans lequel il est mesuré. Deux groupes d'individus ont reçu une attention particulière dans la littérature : l'ensemble de la population d'intérêt, et l'ensemble de la population des sujets effectivement traités. Le premier permet de mesurer l'effet du passage de l'ensemble de la population du statut de « non traité »

2. 10 Types of Treatment Effect You Should Know About. *Evidence in Governance and Politics (EGAP)*. http://egap.org/methods-guides/10-types-treatment-effect-you-should-know-about

à celui de « traité ». On le retrouve dans la littérature sous la dénomination de « average treatment effect on the whole population » (ATE). Le second permet de mesurer l'effet du passage de l'ensemble de la population effectivement traitée du statut de « non traité » à celui de « traité ». On le retrouve sous la dénomination de « average treatment effect on the treated » (ATT).

Même en l'absence d'hétérogénéité de l'effet du traitement, les estimations conditionnelles et marginales ne coïncident pas toujours en valeur. En effet, selon la nature du paramètre estimé, les effets conditionnels et marginaux pourront avoir des valeurs théoriques différentes. On parle de paramètres « non-collapsibles » c'est-à-dire que CTE, ATE et ATT ne coïncident pas, en dehors du cas particulier d'absence totale d'effet du traitement (Greenland et al. 1999). La différence de moyennes, la différence de risques, le risque relatif[3], la différence de survies moyennes ou médianes, précédemment définis, sont des exemples de mesures d'association « collapsibles ». L'odds ratio et le hazard ratio sont, en revanche, « non-collapsibles ». Quand ces types de mesure sont utilisés, il convient alors de réfléchir avec soin à la population d'intérêt répondant aux objectifs de l'étude.

1.2.4 Inférence causale en situation expérimentale : les essais randomisés

Un essai contrôlé randomisé est une étude comparant deux stratégies thérapeutiques sur un critère de jugement Y quelconque. Chaque individu reçoit l'une ou l'autre des alternatives thérapeutiques comparées, avec une probabilité identique pour tous les individus inclus et égale à 0.5 dans la plupart des essais (situation de l'essai à deux bras avec un ratio de randomisation 1 : 1). Ainsi, quelle que soit l'exposition finalement reçue, la probabilité d'être exposé ($T = 1$) ou non ($T = 0$) est la même pour tous les individus au moment du tirage au sort. Les deux groupes de sujets ainsi constitués devraient être alors *en tous points comparables*, c'est-à-dire que la distribution des caractéristiques initiales des

3. Le risque relatif n'est en fait collapsible que si l'effet du traitement est le même pour tous les profils d'individus (Greenland et al. 1999)

individus, mesurées ou non dans l'étude, devraient être identiques [4], sauf pour l'exposition attribuée par le tirage au sort. On parle alors d'**échangeabilité** entre les groupes (Greenland & Robins 2009) : le devenir des deux groupes (dans le cas d'un critère de jugement binaire, la probabilité d'évènement Y) serait en moyenne identique s'ils recevaient tous les deux le même traitement. En conséquence, le devenir du groupe effectivement non exposé ($T = 0$) peut être considéré comme une réalisation de ce que serait devenu le groupe effectivement exposé ($T = 1$) en l'absence d'exposition, et inversement. Ainsi, une différence observée entre le devenir des deux groupes peut être considérée comme uniquement liée au traitement initialement attribué. Une mesure d'association quantifiant cette différence est donc une mesure causale de l'effet du traitement, exempt de tout biais d'indication. Il s'agit de plus d'une mesure marginale, populationnelle, de l'effet du traitement.

1.3 Méthodes d'inférence causale pour l'analyse des études observationnelles

Le mécanisme brièvement décrit au paragraphe précédent permet de tirer une conclusion causale des résultats d'un essai randomisé dans la situation idéale d'absence de biais potentiellement induit par un défaut ou un non-respect du tirage au sort, des exclusions *a posteriori* ou des perdus de vue, ou des différences systématiques en termes de réalisation de l'intervention, de suivi ou d'évaluation du critère de jugement. Ce mécanisme est également à la base de deux méthodes statistiques dont l'objectif est de tirer une conclusion causale des résultats d'une étude observationnelle, c'est-à-dire d'une étude où l'exposition d'intérêt n'est pas tirée au sort, et où la comparaison naïve des groupes d'exposition est sujette au biais d'indication. La première méthode, basée sur le **score de propension**, introduite par Rosenbaum & Rubin (1983), cherche à reproduire des conditions de *quasi-randomisation* en équilibrant les caractéristiques initiales observées entre les deux groupes de traitement. La seconde méthode, introduite par Ben B. Hansen (2008) et basée sur le

4. Des déséquilibres liés au hasard sont évidemment possibles entre les deux groupes.

score pronostique (parfois retrouvée sous la dénomination de *disease risk score* dans la littérature), cherche à estimer, à partir des caractéristiques initiales observées, le devenir des sujets traités en l'absence d'exposition, et inversement.

Ces deux méthodes sont limitées à la prise en compte des facteurs confondants observés, ou plus précisément *mesurés*. D'autres méthodes, notamment les méthodes reposant sur les **variables instrumentales**[5] (Angrist et al. 1996 ; Brookhart et al. 2006 ; Brookhart & Schneeweiss 2007) ou sur l'ajustement sur le **prior-event rate ratio** (Yu et al. 2012 ; Uddin et al. 2015), permettent la prise en compte les facteurs de confusion *mesurés et non mesurés*. Leur utilisation est cependant limitée respectivement par la disponibilité d'un facteur répondant à la définition d'une variable instrumentale, et par la possibilité d'estimer le taux d'évènements dans la population de l'étude avant le début de l'exposition au traitement d'intérêt. Elles ne seront pas abordées dans ce travail.

1.4 Spécificités de l'utilisation de ces méthodes en pharmacoépidémiologie

Les études observationnelles en pharmacoépidémiologie sont souvent mises en place pour évaluer un médicament mis sur le marché récemment ou concurrencé par de nombreuses alternatives thérapeutiques. Cette situation conduit à devoir évaluer l'effet d'un médicament dans une population comprenant peu de sujets traités, c'est-à-dire une population où **l'exposition d'intérêt est rare**.

L'Effective Health Care Program (de l'Agency for Healthcare Research and Quality) recommande, dans son guide méthodologique dédié aux protocoles d'études observationnelles (Velentgas et al. 2013), l'utilisation du score pronostique au lieu du score de propension dans les études cherchant à évaluer l'association entre un évènement fréquent et une expo-

5. Variables mesurées qui, comme l'allocation aléatoire d'une randomisation, sont fortement liées à l'attribution du traitement, mais pas aux facteurs de confusion et au critère de jugement d'intérêt (Greenland 2000).

sition rare. Cependant, ce guide ne définit pas à partir de quel seuil une exposition devrait être considérée comme rare. De plus, aucune étude n'a cherché à évaluer spécifiquement le score de propension ou à le comparer au score pronostique dans cette situation.

1.5 Organisation du document

La suite de ce document est organisée comme suit. Le **chapitre 2** traite de l'utilisation des méthodes basées sur le score de propension en situation d'exposition rare. Le **chapitre 3** évalue les performances des méthodes basées sur le score pronostique rapportées dans la littérature, introduit de nouvelles méthodes basées sur le score pronostique, et les compare aux performances des méthodes basées sur le score de propension. Le **chapitre 4** traite des estimateurs de la variance des effets du traitement estimés à l'aide du score de propension et du score pronostique. Enfin, une **conclusion** résume l'apport de ce travail et expose les grandes lignes de futures recherches.

2 | Score de propension

2.1 Définition et hypothèses

Le score de propension est défini par la probabilité d'être exposé au traitement conditionnellement aux caractéristiques observées du sujet avant l'exposition (Rosenbaum & Rubin 1983). Conditionnellement au score de propension, la distribution de ces caractéristiques observées est indépendante de l'exposition au traitement (Rosenbaum & Rubin 1983) (comme elle l'est aussi dans les essais contrôlés randomisés pour les caractéristiques observées ou non). A condition que certaines hypothèses soient satisfaites (listées dans le Tableau 2.1 ci-dessous) et d'une spécification correcte du modèle utilisé pour estimer le score de propension (Austin & Stuart 2015a), l'utilisation du score de propension permet de palier au problème du biais d'indication dans les études observationnelles, en induisant l'équilibre des caractéristiques initiales observées entre les groupes de sujets traités et non traités (Austin 2009b).

Tableau 2.1 : Hypothèses sur lesquelles repose une analyse basée sur le score de propension.

Hypothèse	Définition
Cohérence (consistency)	Le devenir potentiel d'un sujet particulier s'il recevait l'exposition effectivement observée est égal au devenir observé.
Echangeabilité (exchangeability)	Il n'existe aucun facteur de confusion non mesuré.

Hypothèse	Définition
Positivité (positivity)	Chaque sujet a une probabilité non-nulle d'être exposé et non exposé.

Les scores de propension ont été développés dans l'objectif d'estimer des effets marginaux (Austin et al. 2007). En fonction des méthodes d'utilisation du score de propension, l'ATE ou l'ATT peuvent être estimés (Resche-Rigon et al. 2012).

2.2 Estimation et utilisation

Dans une étude expérimentale randomisée, le score de propension est fixé par le plan expérimental (ou plus précisément par le ratio d'allocation [1]) et est donc connu pour tous les sujets de l'étude. Dans une étude observationnelle, le score de propension est *a priori* inconnu, et doit être estimé à partir des données observées.

2.2.1 Estimation du score de propension

L'analyse par score de propension fonctionne en deux étapes successives (Austin 2011a), dont la première correspond à l'estimation de p_T, le vecteur des probabilités individuelles d'être exposé au traitement conditionnellement aux caractéristiques initiales observées : $p_T = P(T = 1|X)$, où $X = (X_1, \ldots, X_K)$ est un vecteur de K variables observées. Plusieurs stratégies ont été proposées pour estimer ces probabilités individuelles (Westreich et al. 2010), mais la méthode la plus courante consiste à utiliser un modèle de régression logistique avec l'exposition observée T en variable binaire à expliquer et le vecteur de covariables explicatives X (Austin 2011a). Ensuite, les coefficients estimés par ce modèle sont utilisés pour dériver $\hat{p}_{T,i}$, le score de propension estimé pour chaque sujet i. En pratique,

1. Le ratio d'allocation est souvent de 1 :1, le score de propension est alors de 0.5 pour tous les sujets de l'étude.

il est recommandé d'estimer ce modèle en incluant toutes les variables associées au critère de jugement Y (c'est-à-dire les facteurs de confusion et les variables pronostiques) et aucune variable instrumentale (c'est-à-dire les variables associées uniquement à l'exposition et pas au critère de jugement) (Austin 2008b ; Pirracchio et al. 2012).

2.2.2 Utilisation du score de propension

Une fois que le score de propension $\hat{p}_T$ est estimé, la seconde étape consiste à l'utiliser dans l'estimation de l'effet de l'exposition T sur Y.

Quatre méthodes d'utilisation du score de propension estimé ont été décrites dans la littérature : l'ajustement sur le score de propension (Austin et al. 2007), la stratification sur le score de propension (Rosenbaum & Rubin 1984 ; Lunceford & Davidian 2004), l'appariement sur le score de propension (Austin 2009b ; Rubin & Thomas 1996 ; Abadie & Imbens 2009), et la pondération sur le score de propension (Rosenbaum 1987 ; Austin 2010b).

Ajustement sur le score de propension. Le score de propension estimé $\hat{p}_T$ est directement inclus dans un modèle multivarié, comportant donc deux variables explicatives du critère de jugement Y : l'exposition T et le score de propension $\hat{p}_T$. Certains auteurs considèrent que l'ajustement sur le score de propension fournit une estimation de l'effet conditionnel du traitement (Forbes & Shortreed 2008).

Stratification sur le score de propension. La population est divisée en sous-groupes selon les quantiles du score de propension estimé. L'effet du traitement est ensuite estimé au sein de chaque sous-groupe. Enfin, ces estimations sont rassemblées (moyenne pondérée) pour obtenir une estimation globale de l'effet du traitement [2]. Si l'on cherche à estimer l'effet marginal du traitement dans l'ensemble de la population (ATE), chaque estimation est pondérée de manière identique (puisque chaque sous-groupe comporte un nombre approximativement identique

2. Selon la nature du critère de jugement, il est parfois possible d'utiliser un modèle qui moyenne directement l'effet du traitement, comme un modèle de Cox stratifié (Austin 2013).

de sujets). Si l'on cherche à estimer l'effet marginal chez les sujets effectivement traités (ATT), chaque estimation est pondérée proportionnellement au nombre de sujets traités dans le sous-groupe correspondant (Austin & Schuster 2014). En pratique, la stratification sur les quintiles du score de propension est largement utilisée dans la littérature, car l'utilisation des quintiles permettrait de réduire de 90% le biais dû aux facteurs de confusion mesurés (Cochran 1968).

Appariement sur le score de propension. Les sujets exposés au traitement sont appariés avec des sujets non exposés ayant une valeur proche du score de propension. De nombreux algorithmes d'appariement ont été étudiés (Abadie & Imbens 2009 ; Austin 2014a), mais l'approche la plus simple et la plus utilisée est l'appariement 1 pour 1, « gourmand [3] », sans remplacement, avec le plus proche voisin en utilisant un caliper de taille prédéfinie [4] (c'est-à-dire que les sujets appariés ne peuvent différer de plus de la taille du caliper) (Donald B. Rubin 1985 ; Austin 2008a ; Ali et al. 2015). Une taille de caliper égale à 20% de l'écart-type de logit$(\hat{p}_T)$ (le logit du score de propension) semble avoir des performances correctes dans de nombreuses configurations (Austin 2011a ; Austin 2011c). Une fois l'appariement effectué, les exposés et les non exposés sont comparés au sein de la population appariée afin d'estimer l'effet marginal du traitement chez les sujets effectivement traités (ATT).

Pondération sur le score de propension. La pondération consiste à attribuer à chaque sujet un poids calculé en fonction de la valeur du score de propension, de manière à obtenir une *pseudo-population* dans laquelle les caractéristiques initiales des sujets exposés et non exposés au traitement ont tendance à être équilibrées (Resche-Rigon et al. 2012 ; Austin & Stuart 2015a). Dans cette *pseudo-population*, les exposés et les non exposés peuvent être directement comparés afin d'estimer l'effet marginal du traitement. Plusieurs types de pondération W peuvent être utilisés en

3. traduction libre de « greedy » c'est-à-dire qu'un sujet traité est apparié avec le sujet non traité ayant le score de propension le plus proche, même si ce dernier aurait pu constituer une meilleure possibilité d'appariement avec un autre sujet traité.

4. « greedy nearest-neighbour 1 : 1 matching without replacement within specified caliper widths » en anglais.

fonction de l'estimation d'intérêt (Austin & Schuster 2014). Les poids permettant l'estimation de l'ATE sont calculés comme suit : $W_{ATE} = \frac{T}{\hat{p}_T} + \frac{1-T}{1-\hat{p}_T}$. Multiplier ce poids par la probabilité estimée d'être exposé conduit à l'estimation de l'ATT : $W_{ATT} = \hat{p}_T \times W_{ATE} = T + \frac{\hat{p}_T}{1-\hat{p}_T}(1-T)$.

2.3 Brève revue de la littérature

Plusieurs auteurs ont démontré grâce à des études de simulation que l'ajustement et la stratification sur le score de propension ont des performances médiocres pour l'estimation des effets marginaux (Austin 2007 ; Forbes & Shortreed 2008 ; Graf & Schumacher 2008) comme des effets conditionnels (Austin et al. 2007). L'appariement et la pondération sur le score de propension ont par contre de bonnes performances pour estimer les effets marginaux (Austin 2010b ; Austin 2013 ; Austin & Schuster 2014), du fait d'une réduction plus efficace du déséquilibre des distributions des caractéristiques observées entre les sujets exposés et non exposés au traitement (Austin 2009c). Malgré leur infériorité en termes de performances, l'ajustement et la stratification ont longtemps été les méthodes les plus utilisées (Shah et al. 2005 ; Dahabreh et al. 2012), mais des revues de la littérature récentes indiquent que l'appariement pourrait être la méthode majoritaire aujourd'hui, la pondération restant sous exploitée (Gayat et al. 2010 ; Thoemmes & Kim 2011 ; Ali et al. 2015) malgré ses avantages en termes de performances, de flexibilité (elle permet d'estimer soit l'ATE, soit l'ATT), et de présentation (les analyses et les résultats peuvent être rapportés de manière proche de ceux d'un essai randomisé) (Deb et al. 2016).

Plusieurs auteurs ont discuté et évalué les performances des méthodes basées sur le score de propension dans des conditions extrêmes telles que les petits effectifs (Pirracchio et al. 2012), ou les évènements rares (Cepeda et al. 2003 ; Patorno et al. 2014 ; Leyrat et al. 2014), et ont démontré que ces méthodes conservent des performances acceptables dans ces situations. Mais l'utilisation du score de propension peut devenir difficile en cas de faible prévalence de l'exposition, situation fréquemment rencontrées dans les études

observationnelles de pharmacoépidémiologie, car le schéma d'étude impose rarement une prévalence élevée de l'exposition (Rassen & Schneeweiss 2012).

2.4 Utilisation du score de propension quand la prévalence de l'exposition est faible

Dans une situation d'exposition rare, la première étape de l'analyse par score de propension peut être contrariée par des problèmes de séparation dans le modèle de régression logistique utilisé pour estimer les probabilités conditionnelles d'être exposé. Cette situation expose la méthodologie du score de propension à ses limites d'applicabilité : problème liée à la positivité (présence de profils de sujets chez qui le traitement n'a encore jamais été prescrit) et à la spécification du modèle utilisé pour estimer le score de propension (difficulté de prendre en compte toutes les variables nécessaires du fait d'un nombre insuffisant de sujets exposés). Même si des recommandations encouragent l'utilisation de méthodes alternatives dans ce contexte (Arbogast et al. 2012 ; Velentgas et al. 2013), à notre connaissance, aucune étude n'a spécifiquement évalué les conséquences de la rareté d'une exposition sur les performances d'une analyse utilisant le score de propension.

Nous nous sommes donc intéressés à cette problématique dans un article publié dans BMC Medical Research Methodology (Hajage, Florence Tubach, et al. 2016). Ce travail de simulation était illustré par une application sur une étude observationnelle déjà publiée évaluant l'effet du thiazolidinedione sur le risque cardiovasculaire (Roussel et al. 2013).

L'étude était focalisée sur les deux méthodes d'utilisation du score de propension les plus performantes (l'appariement et la pondération sur le score de propension) et sur l'estimation d'un hazard ratio marginal (ATE ou ATT). La distribution des temps jusqu'à l'évènement était générée selon une loi exponentielle, selon un processus dérivé de celui d'Havercroft & Didelez (2012). Cet algorithme a l'avantage de générer une cohorte de sujets fictifs dans laquelle le hazard ratio marginal est directement fixé à la valeur théorique

désirée. Les scenarios évalués étaient définis selon le nombre de sujets, la prévalence de l'exposition, le nombre de covariables, le niveau de l'association entre les covariables et le risque d'évènement, le niveau de l'association entre les covariables et la probabilité d'être exposé, le niveau de l'association entre l'exposition d'intérêt et le risque d'évènement, et le taux de censures.

La conclusion de cette étude était que l'appariement et la pondération sur le score de propension peuvent être sévèrement biaisés en situation d'exposition rare, à moins qu'un nombre important de sujets soit disponible pour l'analyse. Globalement, les plus mauvaises performances étaient obtenues avec la pondération sur le score de propension quand celle-ci cherchait à estimer l'ATE, et les meilleures avec la pondération sur le score de propension quand celle-ci cherchait à estimer l'ATT. Cette dernière méthode est donc à privilégier en situation d'exposition rare, si l'estimation de l'ATT répond aux objectifs de la recherche.

Article 1

Hajage *et al. BMC Medical Research Methodology* (2016) 16:38
DOI 10.1186/s12874-016-0135-1

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

On the use of propensity scores in case of rare exposure

David Hajage[1,3,4,5,6*], Florence Tubach[2,3,4,5,6], Philippe Gabriel Steg[7,8,9], Deepak L. Bhatt[10] and Yann De Rycke[2,3,4,5,6]

Abstract

Background: Observational post-marketing assessment studies often involve evaluating the effect of a rare treatment on a time-to-event outcome, through the estimation of a marginal hazard ratio. Propensity score (PS) methods are the most used methods to estimate marginal effect of an exposure in observational studies. However there is paucity of data concerning their performance in a context of low prevalence of exposure.

Methods: We conducted an extensive series of Monte Carlo simulations to examine the performance of the two preferred PS methods, known as PS-matching and PS-weighting to estimate marginal hazard ratios, through various scenarios.

Results: We found that both PS-weighting and PS-matching could be biased when estimating the marginal effect of rare exposure. The less biased results were obtained with estimators of average treatment effect in the treated population (ATT), in comparison with estimators of average treatment effect in the overall population (ATE). Among ATT estimators, PS-weighting using ATT weights outperformed PS-matching. These results are illustrated using a real observational study.

Conclusions: When clinical objectives are focused on the treated population, applied researchers are encouraged to estimate ATT with PS-weighting for studying the relative effect of a rare treatment on time-to-event outcomes.

Keywords: Propensity scores, Observational studies, Pharmacoepidemiology, Rare exposure, Hazard ratio, Monte Carlo simulations

Background

Post-marketing assessment of the risk and the benefit of a drug in real-world setting frequently relies on observational studies (such as prospective cohorts), comparing treated and untreated subjects on a time-to-event outcome. Effect of the drug exposure is then evaluated through the estimation of a hazard ratio [1–4].

By nature, observational studies may end up with an imbalance of baseline characteristics between exposed and unexposed subjects. If some of these characteristics are also associated with the outcome of interest, we are confronted with confounding factors, and the crude analysis of the treatment effect will be biased [5, 6].

Among the methods used to account for confounding factors in observational studies, propensity score (PS) analysis has been increasingly used [7]. PS analysis was developed to take into account the problem of confounding in observational studies [8], inducing baseline balance of measured confounding factors between groups of exposed and unexposed subjects. PS analysis works with two successive steps [9, 10]. The first step corresponds to the estimation of the probability of exposure conditional on baseline confounding factors. In the second step, these conditional probability estimates are used for the estimation of treatment effect. Several methods have previously been described and extensively compared [11–16]: adjustment on PS [8, 12], stratification on PS [11, 17], matching on PS [8, 14, 18], and PS-weighting estimation [15, 19]. Using empirical case studies and Monte

*Correspondence: david.hajage@aphp.fr
[1] APHP, Hôpital Louis Mourier, Département d'Epidémiologie et Recherche Clinique, 178 Rue des Renouillers, 92700 Colombes, France
[3] APHP, Hôpital Bichat, Centre de Pharmacoépidémiologie (Cephepi), 46 Rue Henri Huchard, F-75018 Paris, France
Full list of author information is available at the end of the article

Carlo simulations, several authors recently showed that PS-matching and PS-weighting more effectively reduced the imbalance between exposed and unexposed subjects in baseline covariates than the two other methods [11, 20], and should be the two preferred methods for the estimation of a marginal hazard ratio [16].

Unlike traditional regression analysis (i.e. incorporating exposure and confounding factors in the same regression model) which provides conditional estimation of the treatment effect, PS-weighting and PS-matching provide marginal estimation. While conditional effects denote an average effect for a specific strata defined by the vector of covariates included in the model, marginal effects denote an effect at the population level. The marginal estimation is similar to the causal estimation provided by a proper randomized clinical trial [10]. Furthermore, PS analysis outperforms conditional analysis when many confounding factors are taken into account: in this situation, conditional analysis may encounter convergence problems [21], particularly when the number of events of interest is small.

Several authors have discussed the use of PS analysis in some extreme situations such as small sample size [22] or rare outcome of interest [23–25]. But the use of PS analysis is also challenging in the case of rare exposure. This situation could frequently be encountered in pharmacoepidemiologic observational studies, particularly when study design does not require a high prevalence of exposure (for example, studies performed on electronic healthcare data, databases constituted with a nonspecific objective or analyzed for a different purpose than initially defined, evaluation of newly marketed drugs [26]). In this setting, the first step of PS analysis (i.e. conditional probability of treatment estimation) can be problematic, due to separation issues with the logistic model used for PS estimation, unless a large sample size is available. Although some recommendations encourage the use of alternative methods like disease risk score (DRS) in this setting [27, 28], to our knowledge, no study specifically assessed the effect of infrequent exposure on PS analysis. Even among the recent literature comparing DRS and PS based methods [29, 30], no article has explored the infrequent exposure setting.

Therefore, our objective was to evaluate the performance of PS-matching and PS-weighting to estimate the marginal hazard ratio associated with a rare exposure in the context of an observational study. An illustration is also provided from a real observational dataset, assessing the association between thiazolidinedione use and major cardiovascular outcomes.

Methods

A Monte Carlo simulation study

We used Monte Carlo simulations to examine the ability of some PS methods to estimate the marginal hazard ratio (HR) associated with a binary treatment in the context of rare exposure. They consisted in:

1. randomly generating independent datasets with several settings defined by exposure prevalence, covariates effect on exposure allocation and on outcome of interest, number of covariates, censoring rate, and exposure effect on outcome of interest (section 'Data-generating process');
2. applying each analytical method to analyze representative samples of each data set (section 'Statistical analyses in simulated data sets');
3. computing several criteria to evaluate and comparing the performance of each method (section 'Performance criteria').

Definitions

In a cohort of N subjects, let E be an indicator variable denoting exposure status ($E = 1$ for exposed subjects, $E = 0$ otherwise), Y be an indicator variable of the event of interest ($Y = 1$ if subject has experimented the event, $Y = 0$ otherwise), and t the observed follow-up time. Let B and C be two baseline covariates, the first one being binary and the second one continuous. Finally, let U represent an unmeasured latent general health baseline variable.

Data-generating process

We used a data-generating process derived from Havercroft et al., who provide an algorithm to generate data from a desired marginal structural model for survival outcome with time-dependent confounding on exposure causal effect [31]. In our simulation process, only baseline confounding was generated.

The key aspect of the algorithm proposed by Havercroft et al. is the use of an unmeasured uniformly distributed variable $U \sim \mathcal{U}(0,1)$ which represents a latent 'general health' process. A value of U close to 0 indicates poor health, and U close to 1 indicates good health.

First, for each subject, we randomly generate three normally distributed covariates (X_B, X_C, and X_U) from the following multivariate normal distribution:

$$X = [X_B, X_C, X_U] \sim \mathcal{N}(0, \Sigma)$$

Variables B, C and U are then computed by applying the following transformations to X_B, X_C and X_U:

$$B = \begin{cases} 1 \text{ if } X_B > 0 \\ 0 \text{ if } X_B \leq 0 \end{cases},$$

$C = X_C$, and

$U = P(X_U < x)$ (the cumulative distribution function of X_U).

By construction, B follows a Bernoulli distribution $\mathcal{B}(0.5)$, C follows a normal distribution $\mathcal{N}(0, \sigma_C)$, and U follows a uniform distribution $\mathcal{U}(0,1)$. B, C, U are related

to each other through covariance parameters $\sigma_{U,B}$, $\sigma_{U,C}$ and $\sigma_{B,C}$.

The exposure allocation E is drawn from a Bernoulli distribution $E \sim \mathcal{B}(p_z)$, where

$$p_z = \text{logit}^{-1}\left(\delta_0 + \delta_B B + \delta_C C\right). \quad (1)$$

δ_0 is the intercept, selected so that the prevalence of exposed subjects in the simulated sample is fixed at a desired proportion p, and δ_B and δ_C are the regression coefficients of this exposure allocation logistic model. For each targeted prevalence, we used an iterative process to determine the value of δ_0 that induced the desired prevalence p:

1. We simulated 100,000 subjects, and computed the individual probabilities of exposure with Eq. 1. The average of these individual probabilities is the theoretical prevalence of exposure, $p^\star$, in the sample.
2. Minimizing $(p^\star - p)^2$ (with the R function `optim`) allows us to obtain the parameter δ_0 that induced desired prevalence of exposure p.
3. This process was repeated 1,000 times and values of δ_0 were averaged to increase precision of the estimation.

An event time T with exponential distribution is generated from U as follows:

$$T = \frac{-log(U)}{\lambda \exp(\gamma E)}, \quad (2)$$

where λ is a constant baseline hazard function, and γ is the marginal effect of E on event time (i.e. $\gamma = log(HR)$). Censoring time T_c is drawn from a uniform distribution $\mathcal{U}(0, c)$ where c is chosen to achieve a desired censoring rate r_c in the simulated sample. Finally, the observed time-to-event outcome is obtained with the following decision rule:

$$Y = 1, t = T \text{ if } T \leq T_c$$

$$Y = 0, t = T_c \text{ if } T > T_c$$

Applied for N subjects, this algorithm generates a sample corresponding to the directed acyclic graph represented on Fig. 1. The key mechanism by which the algorithm generates confounding in the estimation of the marginal exposure effect is the way in which the exposure E and the time t to event outcome Y depends (directly or undirectly) both on U. The relationship between U and Y is straightforward, as U is used to generate event times T (Eq. 2). The relationship between U and E is mediated by the two other covariates B and C, which are 'natively' correlated with U (through parameters $\sigma_{U,B}$ and $\sigma_{U,C}$), and then used to calculate the probability of exposure allocation (Eq. 1). There is confounding due to U being a common ancestor of E and Y. B and C are sufficient to adjust for confounding, because E is independent of U given B and C.

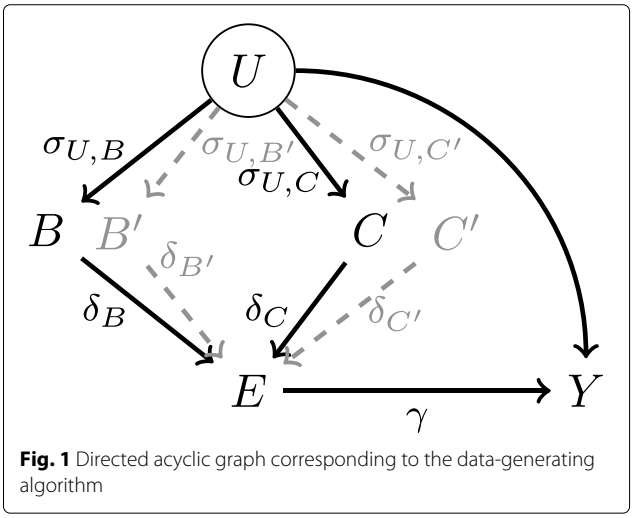


Fig. 1 Directed acyclic graph corresponding to the data-generating algorithm

In all simulations, the following parameters were fixed:

- $N = 10,000$
- $\lambda = 0.1$
- $\sigma_U^2 = \sigma_B^2 = \sigma_C^2 = 1$

We allowed the following parameters to vary across simulations:

- the prevalence of exposure: $p \in \{1\,\%, 2\,\%, 5\,\%, 10\,\%\}$;
- the strength of the correlation between covariates B and C: $\sigma_{B,C} \in \{0, 0.1, 0.3, 0.5\}$ (no, weak, moderate, or strong correlation);
- the strength of the association between covariates and U: $\sigma_{U,B} = \sigma_{U,C} \in \{0, 0.1, 0.3, 0.5\}$ (no, weak, moderate, or strong association);
- the strength of the association between covariates and exposure allocation: $\exp(\delta_B) = \exp(\delta_C) \in \{1, 1.2, 1.5, 2\}$ (no, weak, moderate, or strong association);
- the strength of the marginal association between exposure and outcome: $HR = \exp(\gamma) \in \{1, 1.2, 1.5, 2\}$ (no, weak, moderate, or strong association);
- the censoring rate: $rc \in \{20\,\%, 50\,\%, 80\,\%\}$;

For the intelligibility of the description of the data-generating process, only two covariates (B and C) were previously described. In order to study the impact of the number of confounding factors, two additional covariates, B' and C', were generated in some scenarios, according to the same process. In these scenarios, B' is binary, C' is continuous, and B, B', C, C', and U are related to each other through covariance parameters $\sigma_{U,B} = \sigma_{U,B'}$, $\sigma_{U,C} = \sigma_{U,C'}$, $\sigma_{B,C} = \sigma_{B',C'}$ and $\sigma_{B,B'} = \sigma_{C,C'} = 0$. These two additional covariates are represented in gray on Fig. 1. A

detailed document that encapsulates the data-generating process and all of the simulation scenarios in one place is included in the supplemental material (Additional file 1).

Statistical analyses in simulated data sets

First, in each simulated cohort, random representative samples of increasing size were selected. When studying a rare exposure and limited sample sizes, it is not uncommon to have no event Y in the exposed group. These samples could not be analysed. Dropping all samples with no events in the exposed group would lead to over-represent samples with enough events, and would therefore break the simulation settings when studying small sample sizes. To prevent this situation, samples were not selected according to fixed sample sizes, but according to fixed numbers of events y in the exposed group. More precisely, in each simulated cohort, we selected the first set of subjects among which there were y events in the exposed group, with y varying from 2 to 200, with increment of 2. This allows having enough events in all analysed samples, while ensuring the selection of representative samples of the underlying cohort.

Then, each representative sample was analyzed with the following statistical methods.

Propensity score (PS) analysis with PS-weighting

First, individual PS (i.e. individual probability of being exposed given baseline covariates) was estimated with the following logistic model:

$$PS = \text{logit}^{-1}\left(\hat{\delta}_0 + \hat{\delta}_B B + \hat{\delta}_C C\right) \quad (3)$$

The propensity score of each patient was estimated from the predicted probability of treatment given his(her) covariates.

Then, we applied the Cox proportional hazards model given by the following equation:

$$\lambda(t) = \lambda_0(t)\exp(\hat{\gamma}E) \quad (4)$$

with each subject weighted using the propensity score, and robust standard error estimator [32].

The PS related literature differentiates between the average treatment effect in the entire eligible population (ATE) and the average treatment effect in treated subjects (ATT) [33]. Indeed, two types of weights could be used depending on the desired estimate, as follow:

$$W_{ATE} = \frac{E}{PS} + \frac{1-E}{1-PS}$$

$$W_{ATT} = E + \frac{PS(1-E)}{1-PS}$$

With ATE weights, we considered stabilized weights [34, 35] by multiplying previous (un-stabilized) weights by $E\bar{p} + (1-E)(1-\bar{p})$ (where $\bar{p}$ is the overall probability of being exposed, i.e. the prevalence of exposure estimated in the selected sample).

Propensity score (PS) analysis with PS-matching

First, individual PS were estimated with Eq. 3. Then, we used greedy nearest-neighbour 1:1 matching within specified caliper widths to form pairs of exposed and unexposed subjects matched on the logit of the propensity score, without replacement. We used calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score as this caliper width has been found to perform well in a wide variety of settings [36].

Once matching was completed, we used an univariate Cox proportional hazards regression model with exposure as the only variable to estimate ATT. We used robust estimate of the standard error of the regression coefficient that accounted for the clustering within matched sets [32].

Performance criteria

We performed 5000 simulations per scenario. Results were assessed in terms of the following:

- Bias of the exposure effect estimation: $E(\hat{\gamma} - \gamma)$.
- Root mean squared error (RMSE) of the exposure effect estimation, defined as: $\sqrt{E((\hat{\gamma} - \gamma)^2)}$.
- Variability ratio of the exposure effect, defined as: $\frac{\frac{1}{5000}\sum_{i=1}^{5000}\hat{SE}(\hat{\gamma}_i)}{\sqrt{\frac{1}{4999}\sum_{i=1}^{5000}\left(\hat{\gamma}_i - \bar{\hat{\gamma}}\right)^2}}$, where $\hat{SE}(\hat{\gamma}_i)$ is the estimated standard error of exposure effect $\hat{\gamma}$ in the simulation i. A ratio > 1 (or < 1) suggests that standard errors overestimate (or underestimate) the variability of the estimate of exposure effect [25, 37].
- Coverage: proportion of times γ is enclosed in the 95 % confidence interval of γ estimated from the model.

The mean sample size n were also computed for each scenario.

The data-generating algorithm used in this simulation study allows to generate data with a desired level of ATE. But PS-matching and PS-weighting using ATT weights methods do not provide the same type of estimation (ATT). For these two methods in each evaluated scenario, performance metrics were estimated relative to the corresponding theoretical ATT hazard ratios.

In case of null treatment effect, the true marginal effect is null and do not vary over the sample. Theoretical ATE and ATT are equal: $HR = \exp(\gamma) = 1$. In case of non-null treatment effect, theoretical ATT were computed as followed:

- Using the parameters of the select scenario, we simulated a cohort of 100,000 subjects. Whatever the 'real' exposure status simulated, we generated two potential event times for each subject: first assuming

that the subject was unexposed and then assuming that the subject was exposed to the treatment.
- In the sample regrouping each subject twice (once with the outcome under treatment, and once with the outcome with no treatment), we fitted a Cox model using only subjects who were "really" exposed. The obtained coefficient corresponded to the ATT of the population.
- We repeated this process 1,000 times and averaged the values to increase the precision of this estimation.

Software

All simulations and analyses were performed using `R` software version 3.1.1 (R Foundation for Statistical Computing, Vienna, Austria). Critical parts (in terms of performances, mostly data sets generation procedure) of the simulation program were coded using `C++`, and integrated into `R` code with the help of `Rcpp` package [38].

Results

Results were displayed using a reference configuration: prevalence of exposure $p = 5$ %, moderate association between confounding factors and outcome ($\sigma_{U,B} = \sigma_{U,C} = 0.3$), moderate association between confounding factors and exposure ($\exp(\delta_B) = \exp(\delta_C) = 1.5$), no marginal association between exposure and outcome ($\exp(\gamma) = HR = 1$), two independant confounding factors (one binary, one continuous, $\sigma_{B,C} = 0$), and a censorting rate r_c of 50 %. Then, the effects of change of each of the simulation parameters (compared to the value used in the reference configuration) were reported. More precisely, when the value of a parameter is changed, all other parameters are fixed to the value used in the reference configuration.

The strength of confounding was defined in four classes:

- No confounding: $\sigma_{U,B} = \sigma_{U,C} = 0$ and $\exp(\delta_B) = \exp(\delta_C) = 1$
- Weak confounding: $\sigma_{U,B} = \sigma_{U,C} = 0.1$ and $\exp(\delta_B) = \exp(\delta_C) = 1.2$
- Moderate confounding: $\sigma_{U,B} = \sigma_{U,C} = 0.3$ and $\exp(\delta_B) = \exp(\delta_C) = 1.5$
- Strong confounding: $\sigma_{U,B} = \sigma_{U,C} = 0.5$ and $\exp(\delta_B) = \exp(\delta_C) = 2$

To make the comparison across the different scenarios possible, table and figures of this section report the mean sample size n.

Results for the reference configuration

Results for the reference configuration previously defined are presented in Table 1.

When $y = 20$ (20 events in the exposed group, approximatively 700 analyzed subjects overall), PS-weighting using ATE weights (PSW-ATE) and PS-matching were the most biased methods, followed by PS-weighting using ATT weights (PSW-ATT), and the latter was the only method having coverage below the nominal level. Bias and coverage deteriorated when sample size decreased ($y = 10$, approximatively 350 analyzed subjects overall), particularly for PSW-ATE. When sample size increased ($y = 30$, approximatively 1100 subjects overall), PSW-ATE and PS-matching showed very similar results, and PSW-ATT was still the best method according to bias and coverage performance parameters.

Variability ratios suggested that standard errors underestimate the variability of the exposure effect estimate for methods PSW-ATE and PS-matching when the sample size was low. Variability ratios increased with the sample size, and became clearly larger than 1 for PSW-ATT method (meaning that standard errors tend to be overestimated). The lowest RMSE were observed with the PSW-ATT method.

Table 2 reports the distribution of ATE and ATT weights according to exposure status. Despite the use of stabilized

Table 1 Results for the reference configuration

Method	y	n	Bias	V ratio	RMSE	1-coverage	% match
PSW-ATE	10	364	0.056	0.914	0.406	0.091	
	20	728	0.028	0.982	0.271	0.065	
	30	1092	0.018	1.009	0.216	0.057	
PSW-ATT	10	364	0.026	0.983	0.321	0.060	
	20	728	0.013	1.019	0.222	0.047	
	30	1092	0.008	1.031	0.180	0.046	
PS-matching	10	364	0.051	0.925	0.473	0.062	99.0
	20	728	0.026	0.964	0.316	0.056	99.5
	30	1092	0.017	0.990	0.250	0.053	99.7

Bias, variability ratio, RMSE, and 1-coverage according to analytical method, number of events in the exposed group y, and mean analyzed sample size n, for one scenario ($p = 5$ %, $\sigma_{U,B} = \sigma_{U,C} = 0.3$, $\sigma_{B,C} = 0$, $\exp(\delta_B) = \exp(\delta_C) = 1.5$, $HR = 1$, 2 confounding factors, censoring rate $r_c = 50$ %). The mean percentage of matched exposed subjects is reported for the PS-matching method

Table 2 Distribution of ATE and ATT weights for the reference configuration

		ATE				ATT			
		Weights				Weights			
y	E	Mean	Var	Min	Max	Mean	Var	Min	Max
10	0	1.000	0.001	0.887	3.596	0.052	0.001	0.000	2.940
	1	0.995	0.383	0.064	17.072	1.000	0.000	1.000	1.000
20	0	1.000	0.001	0.922	2.305	0.052	0.001	0.000	1.436
	1	0.999	0.296	0.064	10.461	1.000	0.000	1.000	1.000
30	0	1.000	0.001	0.932	1.727	0.052	0.001	0.001	0.848
	1	0.999	0.265	0.109	10.465	1.000	0.000	1.000	1.000

Mean, variance, minimum and maximum ATE and ATT weights according to type of weights, number of events in the exposed group *y* and exposure status *E* for one scenario ($p = 5$ %, $\sigma_{U,B} = \sigma_{U,C} = 0.3$, $\sigma_{B,C} = 0$, $\exp(\delta_B) = \exp(\delta_C) = 1.5$, $HR = 1$, 2 confounding factors, censoring rate $r_c = 50$ %)

weights, ATE (but not ATT) weights could reach extreme values in the exposed population.

Effect of the prevalence of exposure

Figure 2 show that bias decreased when sample size and/or prevalence increased. Bias decreased more slowly for PSW-ATE than for PSW-ATT. At lower prevalences of exposure (1 and 2 %), PS-matching encountered severe convergence issues, which explained the appearance of the corresponding bias curve. At this level of prevalence, neither PSW-ATE nor PSW-ATT had satisfactory coverage properties unless a

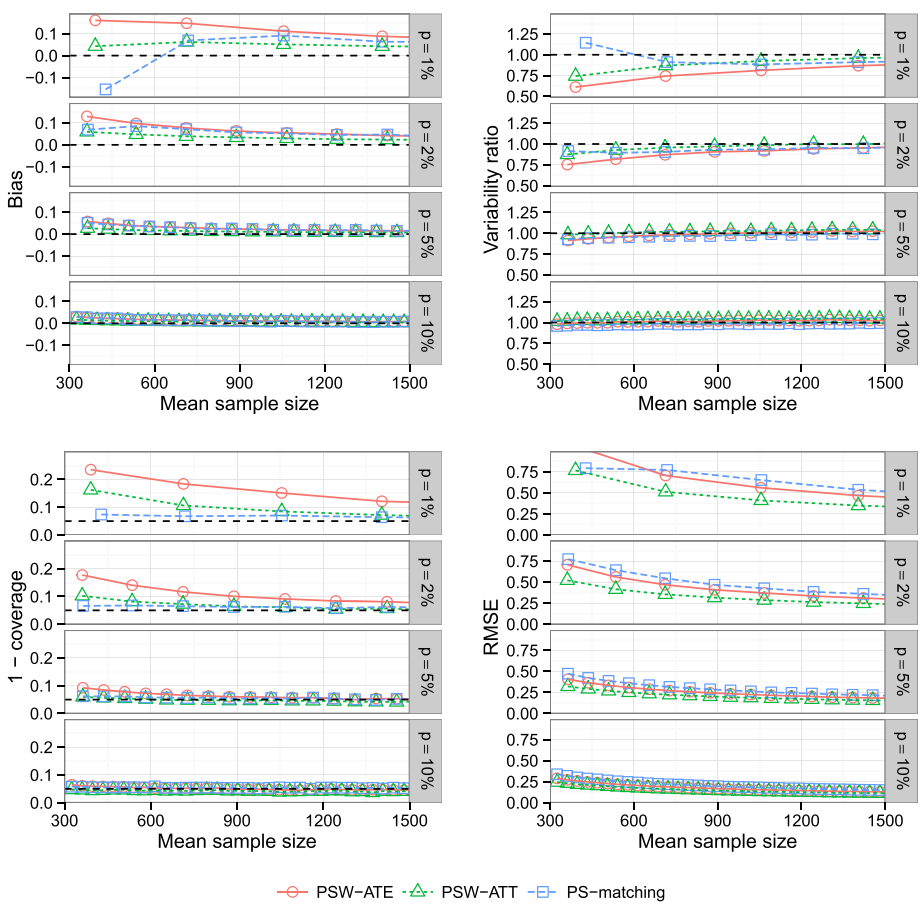


Fig. 2 Effect of exposure prevalence. Bias of exposure effect, variability ratio, 1 - coverage and RMSE according to *prevalence p* of exposure and mean sample size, for one continuous and one dichotomous confounder, $\sigma_{U,B} = \sigma_{U,C} = 0.3$, $\sigma_{B,C} = 0$, $\exp(\delta_B) = \exp(\delta_C) = 1.5$, $r_c = 50$ % and $HR = 1$, with weighting by inverse of PS using ATE and ATT weights and PS-matching

large sample size was analyzed (Fig. 2), the worst method being the use of ATE weights. Standard errors were underestimated at lower levels of prevalence and/or sample size, and became slightly overestimated for PSW-ATT method when prevalence and sample size increased. PSW-ATT method had the lowest RMSE levels. When prevalence was 10 %, bias, coverage and variability ratio were satisfactory for all methods.

Effect of the marginal effect of exposure on outcome event

Influence of theoretical HR is illustrated on Fig. 3. In these scenarios, theoretical values of ATT hazard ratio (used to evaluate the performance of PS-matching and PSW-ATT methods) were 1, 1.471 and 1.935, for theoretical values of ATE hazard ratio of 1, 1.5 and 2 respectively.

All results were mostly unchanged with varying effect of exposure. PSW-ATT was both the less biased method and had the lowest RMSE levels.

Effect of the strength of confounding

Results are illustrated on Fig. 4. In terms of bias, increasing the strength of confounding had a favorable impact on PSW-ATT and PS-matching methods. In contrast, with PSW-ATE method, bias increased with the strength of confounding.

At strong level of confounding, standard errors were overestimated when using PSW-ATT. Consequently, coverage probabilities were greater than the nominal coverage probability, but PSW-ATT remained the most performant method in terms of RMSE.

Effect of the number of confounding factors

Results are illustrated on Fig. 5. The number of confounding factors had a important impact on the bias found with PSW-ATE method, in contrast to the one found with methods estimating ATT. Increasing the number of confounders increased the variability ratio of PSW-ATT method, which consequently seemed too conservative. Conversely, coverage properties of PSW-ATE method

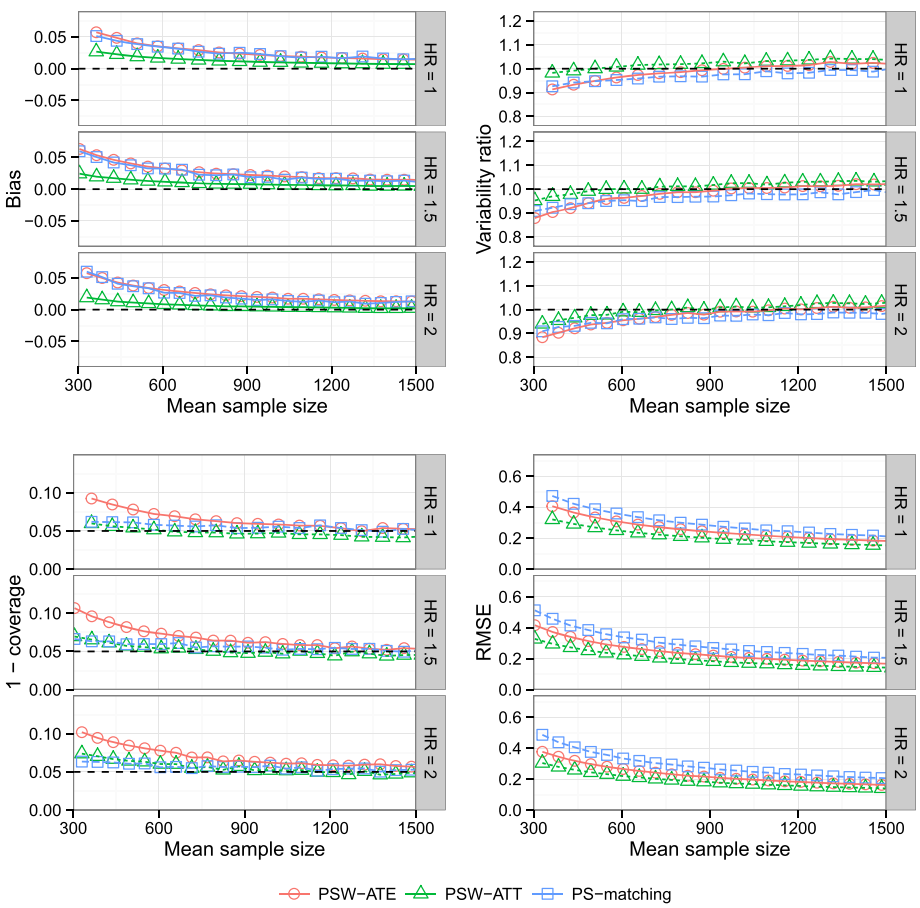


Fig. 3 Effect of theoretical hazard ratio. Bias of exposure effect, variability ratio, 1 - coverage and RMSEw according to *theoretical exposure effect* (HR) and mean sample size, for one continuous and one dichotomous confounder, $\sigma_{U,B} = \sigma_{U,C} = 0.3$, $\sigma_{B,C} = 0$, $\exp(\delta_B) = \exp(\delta_C) = 1.5$, $r_c = 50$ % and $p = 5$ %, with weighting by inverse of PS using ATE and ATT weights and PS-matching

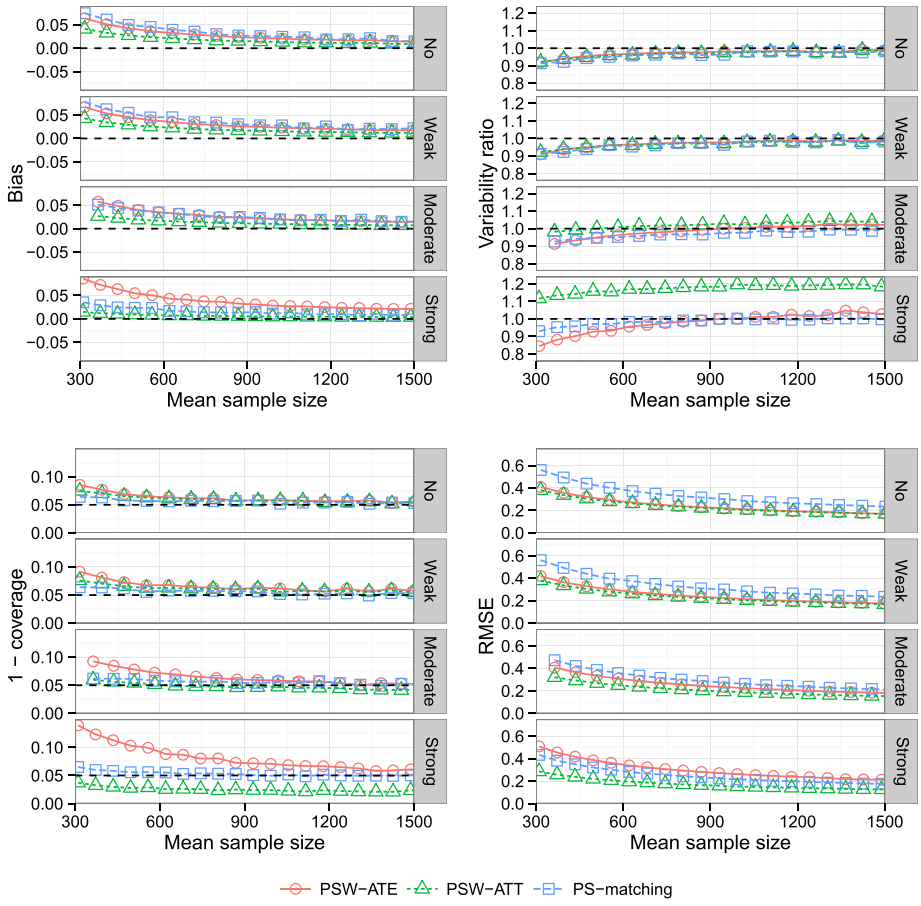


Fig. 4 Effect of strength of confounding. Bias of exposure effect, variability ratio, 1 - coverage and RMSE according to *strength of confounding* and mean sample size, for one continuous and one dichotomous confounder, $\sigma_{B,C} = 0$, $HR = 1$, $r_c = 50$ % and $p = 5$ %, with weighting by inverse of PS using ATE and ATT weights and PS-matching

deteriorated with the transition from two to four confounders. Again, the method with the lowest RMSE values was PSW-ATT, whatever the number of confounding factors.

Effect of the censoring rate

Results are illustrated on Fig. 6. Bias increased with increasing censoring rate for all methods. At the lower level of censoring ($r_c = 20$ %), PSW-matching method was less biased than PSW-ATE method. The opposite was observed at the highest level of censoring. Bias found with PSW-ATT method never exceeded the bias found with PSW-ATE method.

Again, coverage properties and RMSE levels were more satisfactory with PSW-ATT than with PSW-ATE method.

Effect of the correlation between covariates *B* and *C*

Results are illustrated on Fig. 7. Whatever the method, the overall effect of the correlation level between confounding factors was modest.

Real observational dataset illustration

To illustrate these results, we applied the PS methods described above in an already published real observational study [39]. The objective of this study was to compare the occurrence of death, non-fatal myocardial infaction, and congestive heart failure in patients with diabetes, according to the use of thiazolidinedione (TZD), in the REACH (REduction of Atherothrombosis for Continued Health) Registry, an international prospective cohort of patients with either established atherosclerotic arterial disease or at risk for atherothrombosis [40–43]. Patients were enrolled in 44 countries between December 2003 and December 2004. In each country, the protocol was submitted to the institutional review boards according to local requirements, and signed informed consent was obtained for all patients.

From the REACH Registry, we selected 28,332 patients with type 2 diabetes and available data on TZD use. This population (mean age 68 years, standard deviation 9.6 years, 61 % of male) has been previously described, and

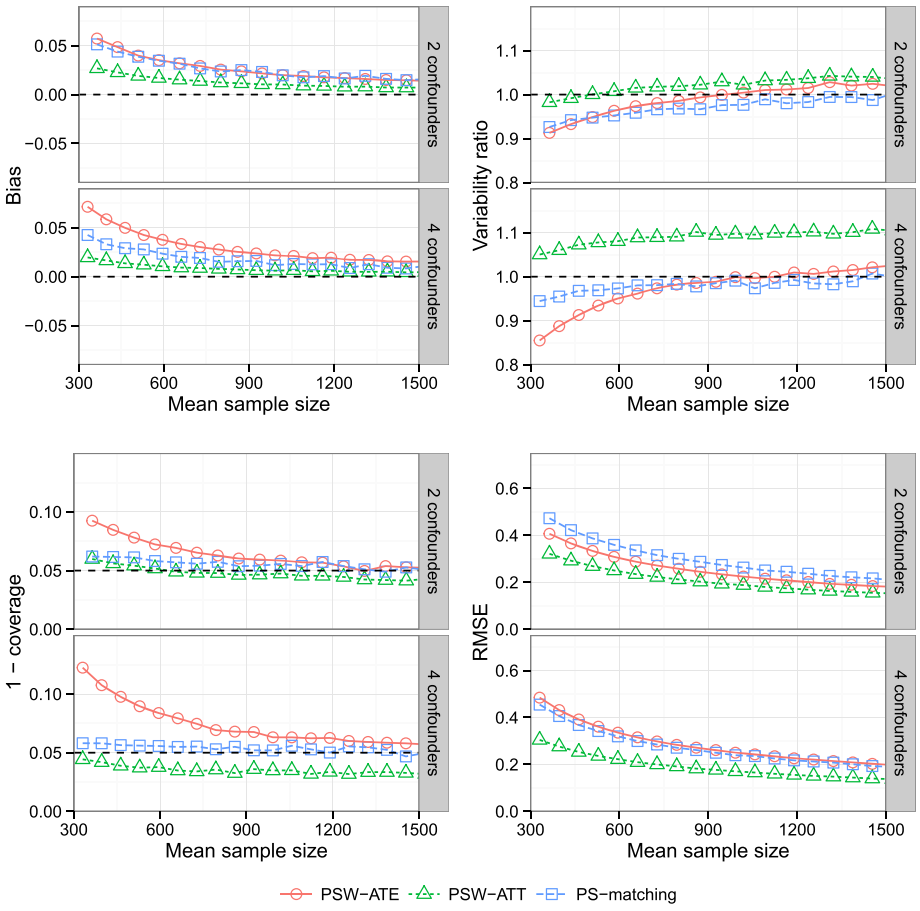


Fig. 5 Effect of the number of confounders. Bias of exposure effect, variability ratio, 1 - coverage and RMSE according to *number of confounders* (2 or 4 confounders) and mean sample size, for $\sigma_{U,B} = \sigma_{U,C} = 0.3$, $\sigma_{B,C} = 0$, $\exp(\delta_B) = \exp(\delta_C) = 1.5$, $HR = 1$, $r_c = 50$ % and $p = 5$ %, with weighting by inverse of PS using ATE and ATT weights and PS-matching

is composed of 4997 TZD users at baseline (prevalence of exposure 17 %).

The list of co-variables used to calculate the propensity score was the same as in the original publication, and included age, geographic region of enrolment, height, body mass index, smoking status, atrial fibrillation/flutter, history of congestive heart failure, treated hypertension, use of lipid-lowering agents, anti-platelet agents, oral anti-coagulants, non-steroidal anti-inflammatory agents, diuretics, cardiovascular agents, peripheral arterial claudication medications, insulin, and use of other anti-diabetic agents. Before the use of PS methods, some known risk factors of cardiovascular events were imbalanced between TZD users and non-users, according to their absolute standardized differences (ASD) (Fig. 8). Compared to the ASD observed in the previous simulations (data not shown), some variables had ASD comparable to the 'weak' confounding condition (like continuous 'age' or binary 'Atrial fibrillation' variables), but also comparable to the 'moderate' (like continuous 'BMI' or binary 'Insulin' variables), or 'strong' confounding condition (like the multimodal 'region' variable). After application of the estimated propensity score to the entire dataset, all variables including those not used in the PS estimation (like formal education and employment) were correctly balanced between TZD users and non-users.

In this application, all event types where regrouped into the same composite outcome (time to the occurrence of the first event). An event occurred in 12 % of subjects. TZD effect was estimated with PS-matching and PS weighting approaches. None of these methods found a significant effect of TZD. No treatment effect heterogeneity was detected (test for homogeneity of the TZD effect across deciles of the PS, p-value = 0.5425).

We then 1) randomly dropped some TZD users to create a new dataset with a pre-specified lower prevalence of exposure 2) applied the three PS-based methods to a representative sample of this new dataset. This two-step process was repeated 2,000 times for prevalences ranging from 17 % (real) down to 5 % and increasing sample

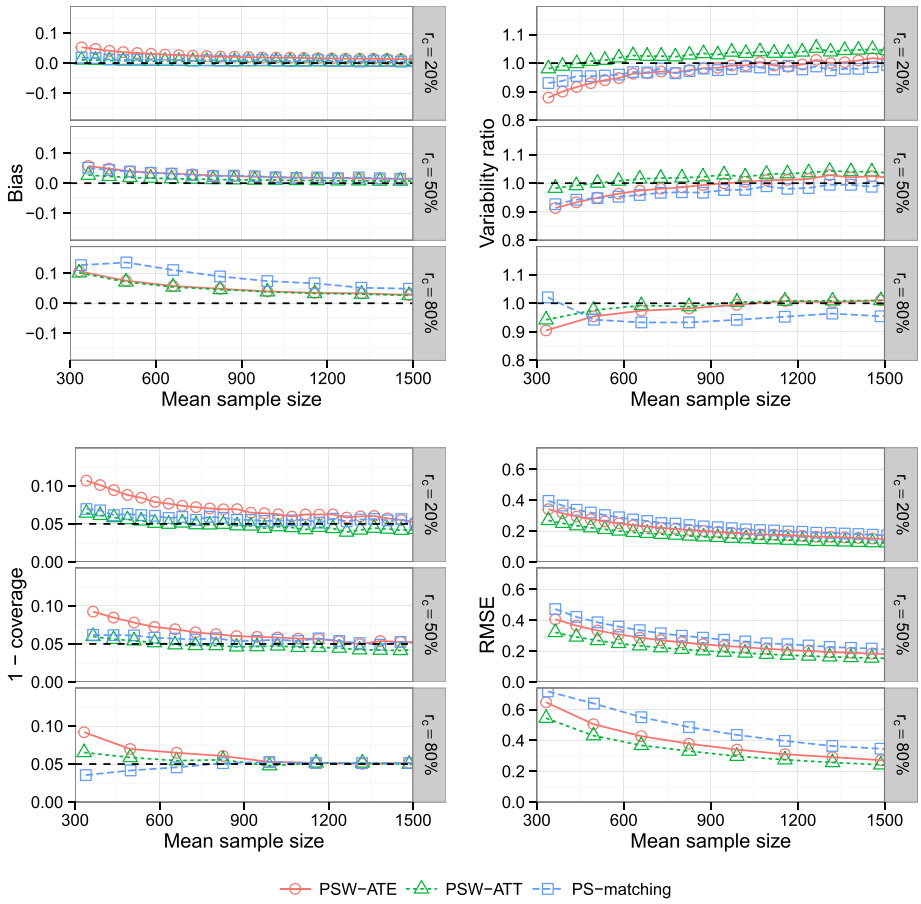


Fig. 6 Effect of censoring rate. Bias of exposure effect, variability ratio, 1 - coverage and RMSE according to *censoring rate* (r_C) and mean sample size, for one continuous and one dichotomous confounder, $\sigma_{U,B} = \sigma_{U,C} = 0.3$, $\sigma_{B,C} = 0$, $\exp(\delta_B) = \exp(\delta_C) = 1.5$, $HR = 1$ and $p = 5$ %, with weighting by inverse of PS using ATE and ATT weights and PS-matching

sizes (selected according to the number of events in the exposed group, like in our simulations). We chose to limit the exploration of the real observational dataset to prevalence of exposure higher than 5 %, because event rate was only 12 % in the REACH cohort, and the number of events in the exposed group is then limited. Bias (relatively to the TZD effect estimated by each method applied in the entire cohort) was averaged and drawn on Fig. 9.

As demonstrated in the simulation study, we observed that ATE estimations were severely biased compared to TZD effect estimated in the full dataset, particularly for the smallest prevalences, even if a large sample size was analyzed. In contrast, ATT estimations through PS-weighting using ATT weights were uniformly less biased, whatever the prevalence and the sample size used. In this application, results observed with PS-matching and PSW-ATT methods seemed superimposed, but this is due to the extremely poor performances of PSW-ATE method, and bias was actually higher with PS-matching than with PSW-ATT.

Discussion

The present simulation study shows that in case of rare exposure, PS-weighting or PS-matching can be biased for estimating the marginal hazard ratio of an exposure. This result was particularly clearcut with PS-weighting analysis using ATE weights, even if stabilized weights were used across all analyses. All methods were converging to their theoretical value with increasing sample size and/or prevalence, but the use of ATE weights and PS-matching needed more subjects than the use of ATT weights. This result leads to limiting the use of PS analysis in case of rare exposure if a sufficient number of subjects is not available, and to favour PS-weighting method using ATT weights when the number of subjects is limited.

Nevertheless, ATT estimation is not consistent with the study objectives in all cases. Small prevalence of exposure could be encountered in two main situations. First, a drug on the market for a long time, and actually little prescribed: in this situation, estimating ATE may not be of great interest, and estimating ATT makes more

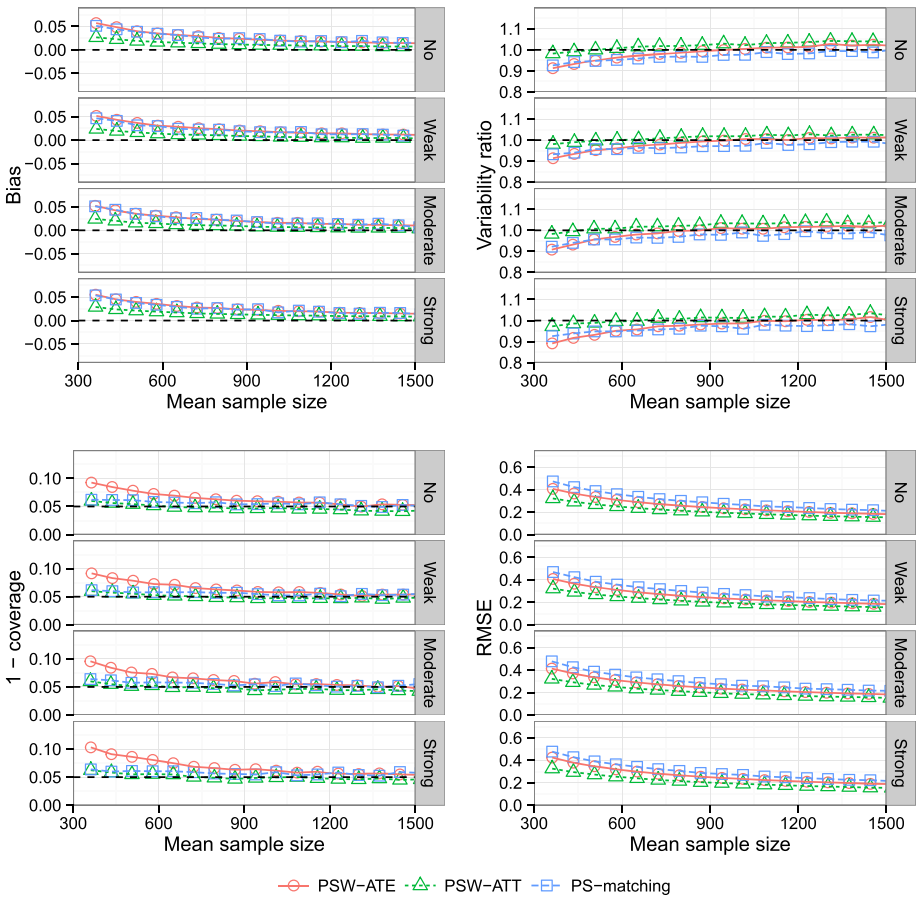


Fig. 7 Effect of correlation between covariates. Bias of exposure effect, variability ratio, 1 - coverage and RMSE according to *correlation between covariates B and C* ($\sigma_{B,C}$) and mean sample size, for one continuous and one dichotomous confounder, $\sigma_{U,B} = \sigma_{U,C} = 0.3$, $\exp(\delta_B) = \exp(\delta_C) = 1.5$, $HR = 1$ and $p = 5$ %, with weighting by inverse of PS using ATE and ATT weights and PS-matching

clinical sense. Second, a newly marketed drug, that is not intended to remain uncommon: this situation is a subject of special attention from the health authorities, and early assessment of the drug effect if the entire population was exposed would be of great interest to public health policy. Our simulation results stress the importance of looking for methods less influenced by exposure prevalence.

The concerns with ATE estimation in case of rare exposure were sustained by our real dataset illustration. The number of potential confounders taken into account were high, and some variables had absolute standardized differences comparable to the 'moderate' and 'strong' confounding conditions of the simulations. We assumed from the former simulation results that the high degree of bias observed with PSW-ATE method in the REACH study is due to the strength of confounding and the number of confounders present in the database, which had a large impact on ATE estimates. Hence, results observed in the REACH study were consistent with the simulation results.

Pirracchio et al. [22] concluded from their simulation study that 'even in case of small study samples or low prevalence of treatment, both propensity score matching and inverse probability of treatment weighting can yield unbiased estimations of treatment effect'. However this study explored more specifically the context of small sample size (ranging from 1000 down to 40) rather than low prevalence of exposure (ranging from 50 % down to 20 %). While some conventions exist on the definition of a rare disease [44], there is, to our knowledge, no such definition of a rare exposure. Nevertheless, we felt that a 1:4 exposure ratio represented a quite common exposure, and more extreme situations could be encountered in observational studies, for example those focusing on a newly marketed medications or when many therapeutic strategies are available. To the best of our knowledge, the present study is the first to focus on the performance of PS-based methods in the context of a rare exposure (10 % down to 1 %) and small sample sizes. This explains that,

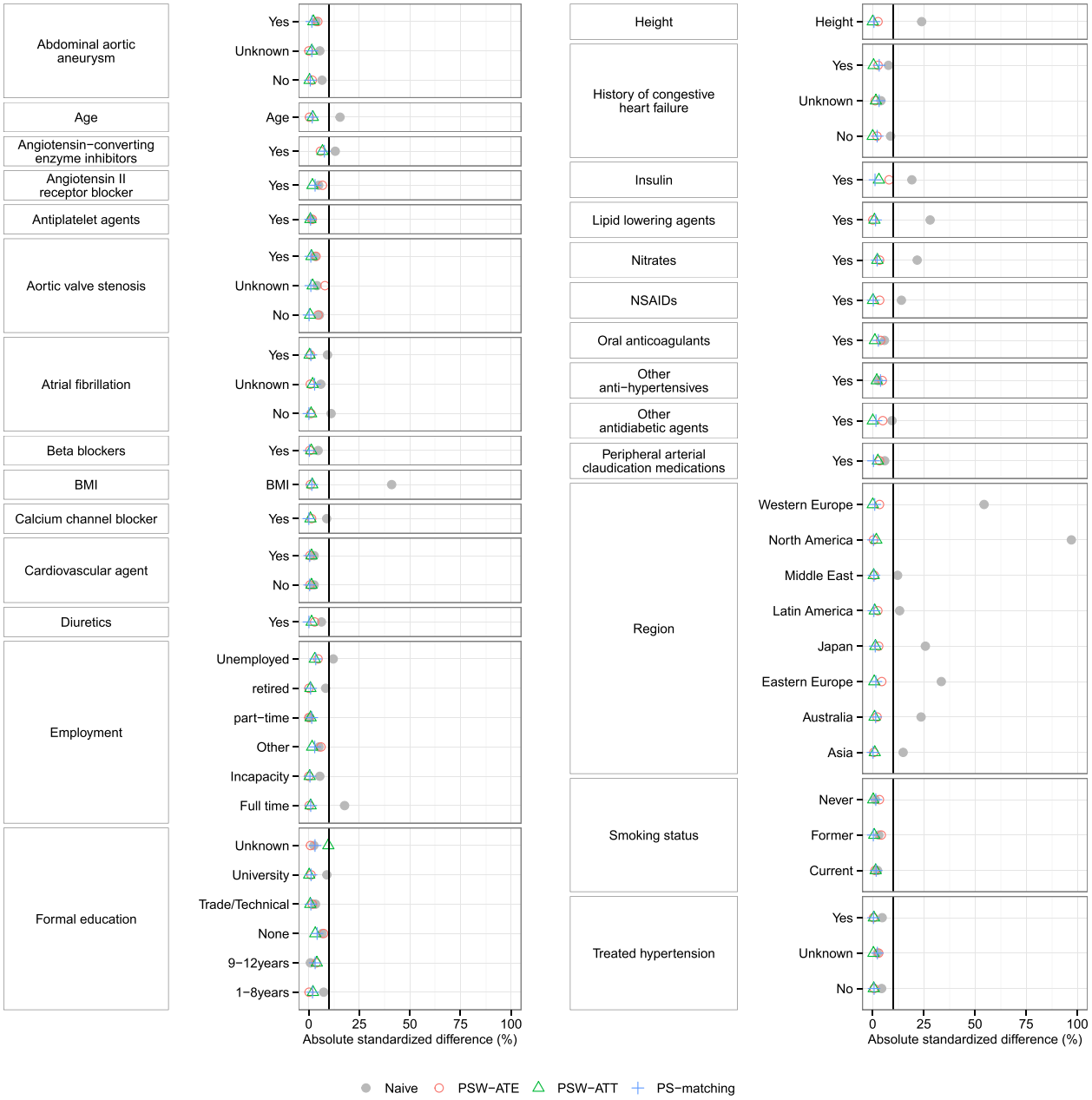


Fig. 8 Imbalances in the REACH cohort, defined as the standardized means differences of covariate values between the two treatment groups. Solid black line represents an absolute standardized difference of 10 %

unlike Pirracchio et al., we conclude that PS-based methods could lead to rather biased estimates when prevalence is low, particularly when estimating average treatment effect in the whole population.

Without focusing specifically on rare exposure issue, Austin et al. have compared the performance of different propensity score methods for estimating absolute effects [45] and relative effects [16] of treatments on survival outcomes. In these two simulation studies, low prevalences of exposure were also simulated. The authors did not observe any major performance issue using PS-weighting or PS-matching when proportion of treated subjects was fixed to 10 % or 5 %. For the estimation of absolute effects, they reported that PS-matching tended to decrease bias compared with PS-weighting approaches. However, all methods compared in this article were applied on simulated cohorts of 10,000 subjects. With fewer subjects, we observed that 1) all methods could be biased, 2) PS-weighting using ATT weights outperformed PS-matching for the estimation of ATT, and 3) PS-weighting using ATE

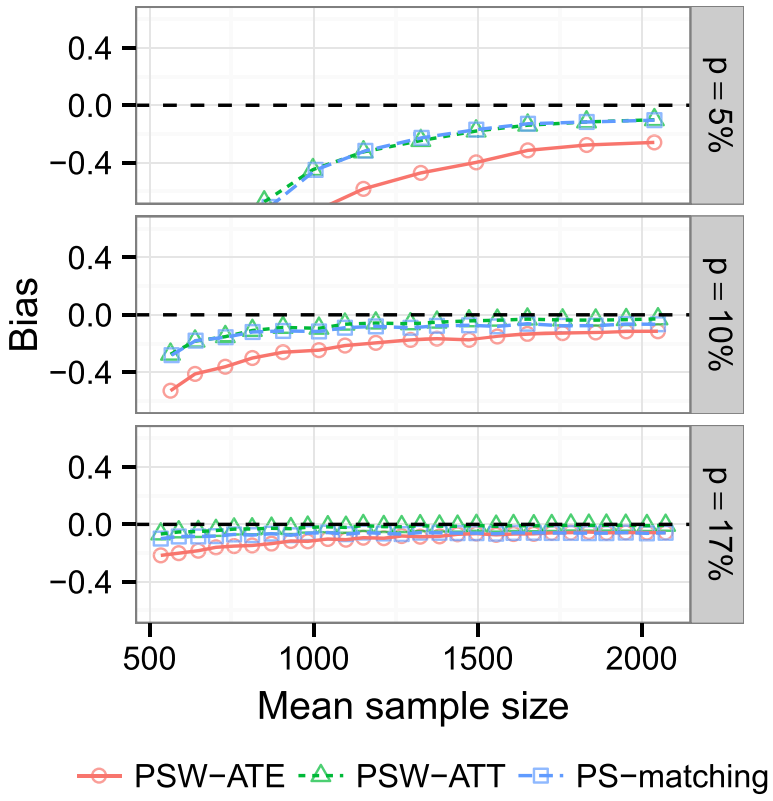


Fig. 9 Real observational dataset illustration. Bias of TZD effect estimation in the REACH cohort, using PS-matching and PS-weighting approaches, according to prevalence *p* and mean sample size

weights was the method which performance deteriorates most with the decrease of exposure prevalence.

The context of rare exposure is also addressed by authors interested in 'the prognostic analogue of the propensity score', a.k.a. disease risk score (DRS) [29, 46, 47]. Actually, Effective Health Care Program recommends the use of disease risk score instead of propensity score when the exposure is infrequent [27, 28], but without defining when an exposure should be considered as infrequent. No study has compared propensity and disease risk score methods for the estimation of an exposure effect in the context of rare exposure. Arbogast et al. [29] compared the performance of disease risk score, propensity score and traditional multivariable regression to evaluate a treatment effect on a Poisson outcome, but prevalence of exposure was fixed to 10 %, and computations were based on the analysis of samples consisting of 10,000 subjects. The authors concluded that all methods performed well when there was an adequate number of events per covariates. Our simulation results also suggest that all PS-based methods are unbiased at this level of prevalence when a large sample size is analyzed. Wyss and colleagues [30] compared PS and DRS matching, and concluded that the use of DRS yielded to match more exposed subjects than the use of PS, and this improved the precision of the effect estimate. However, the prevalence of exposure was fixed to 30 % in all the scenarios considered. Intuitively, this advantage of DRS should be less apparent in case of lower prevalence of exposure. Among the scenarios and sample sizes explored in the present article, the percentages of matched exposed subjects were high (Q25 = 99.7 %, Q50 = 99.8 %, Q75 = 99.9 %). Thus, further investigation is needed to assess if DRS really performs better than PS in the context of rare exposure, especially as the relative performance of the different DRS-based methods for estimating ATE and ATT are today a research area [27].

In the setting of rare exposure, we found that application of PS-based methods could provide biased estimates unless a large sample size was available. PS method being a two-step estimator, the appropriateness of the estimation in the second step relies on correct modelling of the probability of exposure during the first step, which could be problematic in case of infrequent exposure, due to separation issues. Of note, alternative strategies than logistic regression have been proposed to estimate individual probability of exposure [48], but we found no information about how they would be affected by a rare exposure issue.

All PS methods rely on the validity of estimates of individual exposure probability, and thus on the validity of the logistic regression fitted for these estimations. A classical rule when fitting a logistic model is to have an adequate number of outcomes per predictor (at least five or ten outcomes per predictor [49, 50]). This explains why we chose to limit the number of confounding factors in our simulations: in case of small prevalence of exposure, the number of exposed subjects, and therefore the number of variables that could be included in the logistic model, is limited. The bias observed in some of our simulations could not be explained by an inadequate number of exposed subjects per co-variables in all cases: even with only two confounding factors, bias was still present with a sample size of 500 subjects and an exposure prevalence of 5 % (and thus 25 exposed subjects on average) or 10 % (50 exposed subjects on average). Therefore, the previously mentioned 'rule of thumb' fails to provide sufficiently accurate estimates of individual exposure probability, particularly when estimating ATE with PS-weighting method.

Other reasons might explain that the ATT estimates were more reliable that ATE estimates in the context of rare exposure. First, ATT estimates apply to a much more homogeneous population, so less confounding might be involved. Another reason might be that strong confounding and limited overlap between treatment groups leads to a violation of the positivity assumption. We observed that ATE (but not ATT) weighting can yield extreme weights in the exposed population, as well as biased and highly variable estimates.

One of the strengths of this study is the use of an algorithm which directly generates data with desired marginal HR and confounding on exposure causal effect. Indeed, several simulations studies evaluating the performance of PS methods to estimate marginal HR used a conditional model to link the outcome with the exposure and (time-dependent or not) confounding factors, even though the measures used to estimate exposure effect on outcome are sometimes non-collapsible [51, 52] (i.e. conditional and marginal treatment effects will not coincide). Two more approximate strategies are typically used to deal with this issue: the use of a high number of simulations to determine the value of the conditional hazard ratio that induced the desired marginal hazard ratio [16]; or the *post-hoc* verification that conditional and marginal treatment effects are in the same range [53]. Another solution is to use a collapsible estimate of exposure effect, like risk differences [15], but this type of estimator is less used to report the effect of an exposure in real studies. Nevertheless, even if we did not use a conditional model to generate simulated datasets, a rather similar issue remains in this article: our algorithm simulates a desired hazard ratio in the entire cohort (ATE), but not a desired hazard ratio in the treated population (ATT). Thus, a possible explanation for the discrepancies between methods estimating ATE and ATT is that they are compared to different theoretical values of the treatment effect. However, this issue was minimized in this study 1) by choosing a null treatment effect in the majority of the reported scenarios (in this case, ATE and ATT are both null), and 2) by estimating the theoretical ATT as precisely as possible with a large number of simulations of potential outcomes in other cases. Moreover, if this estimation of theoretical ATT was not sufficiently accurate, this would probably disadvantage methods estimating ATT, which reinforce the findings of this study.

Conclusions

In conclusion, this simulation study showed that in case of rare exposure, marginal treatment effect estimation through propensity score analysis can be severely biased, in particular when focusing on average treatment effect in the entire eligible population (ATE). When clinical objectives are focused on the treated population, PS-weighting using ATT weights should be the preferred estimator of the treatment effect. Further work in this area is needed to provide improved analytical strategies for the estimation of the marginal treatment effect in the context of an observational study with a rare exposure.

Availability of data and materials

The `R` code corresponding to the data-generating process and the statistical methods used in this article can be obtained on request to David Hajage (david.hajage@aphp.fr).

Real dataset supporting the findings (REACH Registry) can be obtained on request to Philippe Gabriel Steg (gabriel.steg@aphp.fr).

Additional file

Additional file 1: Data-generation process and simulated scenarios. (DOCX 119 kb)

Abbreviations

ATE: average treatment effect; ATT: average treatment effect in the treated; DRS: disease risk score; HR: hazard ratio; PS: propensity score; PSW: propensity score weighting; REACH: REduction of Atherothrombosis for Continued Health; TZD: thiazolidinedione.

Competing interests

This article reports the results of a simulation study, and the following disclosures are therefore unrelated to the submitted work.
PG Steg discloses the following relationships (unrelated with submitted work):

- Research grants (to INSERM U1148): Servier, Sanofi
- Speaker or consultant (including steering committee, DMC and CEC memberships) : Amarin, AstraZeneca, Bayer, Boehringer-Ingelheim, BristolMyersSquibb, Daiichi-Sankyo-Lilly, GlaxoSmithKline, Medtronic, Merck-Sharpe Dohme, Novartis, Otsuka, Pfizer, Regado, Sanofi, Servier, The Medicines Company, Vivus
- Stockholder: Aterovax

DL Bhatt discloses the following relationships (unrelated with submitted work):

- Advisory Board: Cardax, Elsevier Practice Update Cardiology, Medscape Cardiology, Regado Biosciences
- Board of Directors: Boston VA Research Institute, Society of Cardiovascular Patient Care
- Chair: American Heart Association Get With The Guidelines Steering Committee
- Data Monitoring Committees: Duke Clinical Research Institute, Harvard Clinical Research Institute, Mayo Clinic, Population Health Research Institute
- Honoraria: American College of Cardiology (Senior Associate Editor, Clinical Trials and News, ACC.org), Belvoir Publications (Editor in Chief, Harvard Heart Letter), Duke Clinical Research Institute (clinical trial steering committees), Harvard Clinical Research Institute (clinical trial steering committee), HMP Communications (Editor in Chief, Journal of Invasive Cardiology), Journal of the American College of Cardiology (Associate Editor; Section Editor, Pharmacology), Population Health Research Institute (clinical trial steering committee), Slack Publications (Chief Medical Editor, Cardiology Today's Intervention), WebMD (CME steering committees); Other: Clinical Cardiology (Deputy Editor)
- Research Funding: Amarin, AstraZeneca, Bristol-Myers Squibb, Eisai, Ethicon, Forest Laboratories, Ischemix, Medtronic, Pfizer, Roche, Sanofi Aventis, The Medicines Company
- Unfunded Research: FlowCo, PLx Pharma, Takeda

Authors' contributions

DH and YDR carried out all Monte simulations and statistical analysis, and drafted the manuscript. FT supervised the project and the elaboration of the manuscript. DLB and PGS provided real dataset, helped the interpretation of the results, and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The REACH Registry was sponsored by Sanofi-Aventis, Bristol-Myers Squibb and the Waksman Foundation (Tokyo, Japan), and is endorsed by the World Heart Federation. This specific analysis did not receive any funding.

Author details

[1]APHP, Hôpital Louis Mourier, Département d'Epidémiologie et Recherche Clinique, 178 Rue des Renouillers, 92700 Colombes, France. [2]APHP, Hôpital Bichat, Département d'Epidémiologie et Recherche Clinique, 46 Rue Henri Huchard, F-75018 Paris, France. [3]APHP, Hôpital Bichat, Centre de Pharmacoépidémiologie (Cephepi), 46 Rue Henri Huchard, F-75018 Paris, France. [4]Univ Paris Diderot, Sorbonne Paris Cité, UMR 1123 ECEVE, F-75018 Paris, France. [5]INSERM, UMR 1123 ECEVE, F-75018 Paris, France. [6]INSERM, CIE-1425, F-75018 Paris, France. [7]FACT, DHU FIRE, Univ Paris-Diderot, Sorbonne Paris-Cité, F-75018 Paris, France. [8]LVTS, INSERM U-1148, Hôpital Bichat, HUPNVS, AP-HP, F-75018 Paris, France. [9]NHLI, Imperial College, Royal Brompton Hospital, London, UK. [10]Brigham and Women's Hospital Heart & Vascular Center and Harvard Medical School, Boston, Massachusetts, USA.

Received: 28 May 2015 Accepted: 15 March 2016
Published online: 31 March 2016

References

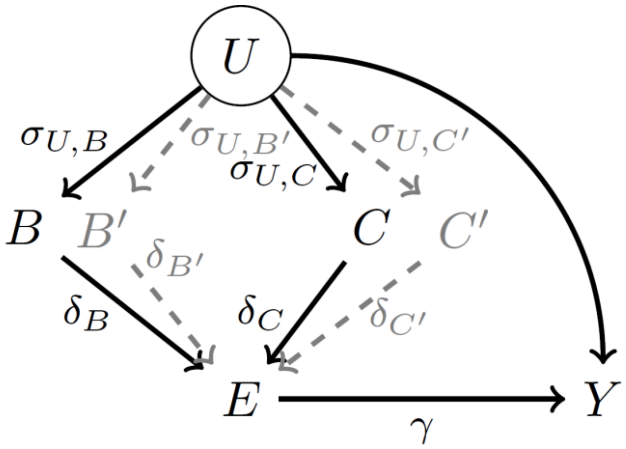
1. Rafaniello C, Lombardo F, Ferrajolo C, Sportiello L, Parretta E, Formica R, Potenza S, Rinaldi B, Irpino A, Raschetti R, Vanacore N, Rossi F, Capuano A. Predictors of mortality in atypical antipsychotic-treated community-dwelling elderly patients with behavioural and psychological symptoms of dementia: a prospective population-based cohort study from Italy. Eur J Clin Pharmacol. 2014;70(2):187–95. doi:10.1007/s00228-013-1588-3.
2. Weinhandl ED, Gilbertson DT, Collins AJ, Foley RN. Relative safety of peginesatide and epoetin alfa. Pharmacoepidemiol Drug Saf. 2014;23(10):1003–11. doi:10.1002/pds.3655.
3. Eftekhari K, Ghodasra DH, Haynes K, Chen J, Kempen JH, VanderBeek BL. Risk of retinal tear or detachment with oral fluoroquinolone use: a cohort study. Pharmacoepidemiol Drug Saf. 2014;23(7):745–52. doi:10.1002/pds.3623.
4. Beigel F, Steinborn A, Schnitzler F, Tillack C, Breiteneicher S, John JM, Van Steen K, Laubender RP, Göke B, Seiderer J, Brand S, Ochsenkühn T. Risk of malignancies in patients with inflammatory bowel disease treated with thiopurines or anti-TNF alpha antibodies. Pharmacoepidemiology and Drug Safety. 2014;23(7):735–44. doi:10.1002/pds.3621.
5. Kestenbaum B. Methods to Control for Confounding. In: Epidemiology and Biostatistics. New York: Springer; 2009. p. 101–11.
6. Rothman KJ, Greenland S, Lash TL, (eds). Modern Epidemiology. 530 Walnut Street, Philadelphia, PA 19106 USA: Lippincott Williams & Wilkins; 2008.
7. Glynn RJ, Schneeweiss S, Sturmer T. Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006;98(3):253–9. doi:10.1111/j.1742-7843.2006.pto_293.x.
8. Rosenbaum PR, 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.
9. Austin PC. A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. Multivar Behav Res. 2011;46(1):119–51. doi:10.1080/00273171.2011.540480.
10. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. Stat Med. 2014;33(7):1242–58. doi:10.1002/sim.5984.
11. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937–960. doi:10.1002/sim.1903.
12. Austin PC, Grootendorst P, Normand S-LT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. Stat Med. 2007;26(4):754–68. doi:10.1002/sim.2618.
13. Austin PC. The performance of different propensity-score methods for estimating relative risks. J Clin Epidemiol. 2008;61(6):537–45. doi:10.1016/j.jclinepi.2007.07.011.
14. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biom J Biom Z. 2009;51(1):171–84. doi:10.1002/bimj.200810488.
15. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Stat Med. 2010;29(20):2137–148. doi:10.1002/sim.3854.
16. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med. 2013;32(16):2837–849. doi:10.1002/sim.5705.
17. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. J. Am Stat Assoc. 1984;79(387):516. doi:10.2307/2288398.
18. Rubin DB, Thomas N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. Biometrics. 1996;52(1):249. doi:10.2307/2533160.
19. Rosenbaum PR. Model-Based Direct Adjustment. J Am Stat Assoc. 1987;82(398):387. doi:10.2307/2289440.
20. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Dec Mak An Int J Soc Med Dec Mak. 2009;29(6):661–77. doi:10.1177/0272989X09341755.
21. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006;59(5):437–47. doi:10.1016/j.jclinepi.2005.07.004.
22. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the Propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol. 2012;12(1):70. doi:10.1186/1471-2288-12-70.
23. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. Am J Epidemiol. 2003;158(3):280–7. doi:10.1093/aje/kwg115.
24. Patorno E, Glynn RJ, Hernández-Díaz S, Liu J, Schneeweiss S. Studies with many covariates and few outcomes: selecting covariates and

implementing propensity-score-based confounding adjustments. Epidemiol (Cambridge, Mass). 2014;25(2):268–78. doi:10.1097/EDE.0000000000000069.
25. Leyrat C, Caille A, Donner A, Giraudeau B. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. Stat Med. 2014. doi:10.1002/sim.6185.
26. Rassen JA, Schneeweiss S. Newly marketed medications present unique challenges for nonrandomized comparative effectiveness analyses. J Comp Eff Res. 2012;1(2):109–11. doi:10.2217/cer.12.12.
27. Arbogast PG, Seeger JD, DEcIDE Methods Center Summary Variable Working Group. Summary Variables in Observational Research: Propensity Scores and Disease Risk Scores. Effective Health Care Program Research Report No. 33. (Prepared by DEcIDE Methods Center under Contract No. HHSA 290-2005-0016-I, Task Order 10.) AHRQ Publication No. 11(12)-EHC055-EF. Rockville, MD: Agency for Healthcare Research and Quality. May 2012. http://effectivehealthcare.ahrq.gov/reports/final.cfm.
28. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, (eds). Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. AHRQ Methods for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013.
29. Arbogast PG, Ray WA. Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders. Am J Epidemiol. 2011;143. doi:10.1093/aje/kwr143.
30. Wyss R, Ellis AR, Brookhart MA, Jonsson Funk M, Girman CJ, Simpson RJ, Stürmer T. Matching on the disease risk score in comparative effectiveness research of new treatments. Pharmacoepidemiol Drug Saf. 2015;24(9):951–61. doi:10.1002/pds.3810.
31. Havercroft WG, Didelez V. Simulating from marginal structural models with time-dependent confounding. Stat Med. 2012;31(30):4190–206. doi:10.1002/sim.5472.
32. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008;27(12):2037–049. doi:10.1002/sim.3150.
33. Imbens G. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. Rev Econ Stat. 2004.
34. Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550–60. doi:10.2307/3703997.
35. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168(6):656–64. doi:10.1093/aje/kwn164.
36. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150–61. doi:10.1002/pst.433.
37. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2011;25(24):4279–292. doi:10.1002/sim.2673.
38. Eddelbuettel D, Francois R. Rcpp: Seamless R and C++ Integration. J Stat Softw. 2011;40(8):1–18.
39. Roussel R, Hadjadj S, Pasquet B, Wilson PW, Smith SC Jr, Goto S, Tubach F, Marre M, Porath A, Krempf M, Bhatt DL, Steg PG. Thiazolidinedione use is not associated with worse cardiovascular outcomes: a study in 28,332 high risk patients with diabetes in routine clinical practice: brief title: thiazolidinedione use and mortality. Int J Cardiol. 2013;167(4):1380–4. doi:10.1016/j.ijcard.2012.04.019.
40. Bhatt DL, Eagle KA, Ohman EM, Hirsch AT, Goto S, Mahoney EM, Wilson PWF, Alberts MJ, D'Agostino R, Liau C-S, Mas J-L, Röther J, Smith SC, Salette G, Contant CF, Massaro JM, Steg PG, REACH Registry Investigators. Comparative determinants of 4-year cardiovascular event rates in stable outpatients at risk of or with atherothrombosis. JAMA. 2010;304(12):1350–1357. doi:10.1001/jama.2010.1322.
41. Steg PG, Bhatt DL, Wilson PWF, D'Agostino R, Ohman EM, Röther J, Liau C-S, Hirsch AT, Mas J-L, Ikeda Y, Pencina MJ, Goto S, REACH Registry Investigators. One-year cardiovascular event rates in outpatients with atherothrombosis. JAMA. 2007;297(11):1197–1206. doi:10.1001/jama.297.11.1197.
42. Ohman EM, Bhatt DL, Steg PG, Goto S, Hirsch AT, Liau C-S, Mas J-L, Richard A-J, Röther J, Wilson PWF, REACH Registry Investigators. The REduction of Atherothrombosis for Continued Health (REACH) Registry: an international, prospective, observational investigation in subjects at risk for atherothrombotic events-study design. Am Heart J. 2006;151(4):786–110. doi:10.1016/j.ahj.2005.11.004.
43. Bhatt DL, Steg PG, Ohman EM, Hirsch AT, Ikeda Y, Mas J-L, Goto S, Liau C-S, Richard AJ, Röther J, Wilson PWF, REACH Registry Investigators. International prevalence, recognition, and treatment of cardiovascular risk factors in outpatients with atherothrombosis. JAMA. 2006;295(2):180–9. doi:10.1001/jama.295.2.180.
44. Lavandeira A. Orphan drugs: legal aspects, current situation. Haemophilia: The Official J World Fed Hemophilia. 2002;8(3):194–8.
45. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. Stat Methods Med Res. 2014. 0962280213519716, doi:10.1177/0962280213519716.
46. Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008;95(2):481–8. doi:10.1093/biomet/asn004.
47. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. Pharmacoepidemiol Drug Saf. 2012;21:138–47. doi:10.1002/pds.3231.
48. 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. J Clin Epidemiol. 2010;63(8):826–33. doi:10.1016/j.jclinepi.2009.11.020.
49. Vittinghoff E, McCulloch CE. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. Am J Epidemiol. 2007;165(6):710–8. doi:10.1093/aje/kwk052.
50. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–1379.
51. Greenland S. Interpretation and Choice of Effect Measures in Epidemiologic Analyses. Am J Epidemiol. 1987;125(5):761–8.
52. Gail MH, Wieand S, Piantadosi S. Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. Biometrika. 1984;71(3):431. doi:10.2307/2336553.
53. Xiao Y, Abrahamowicz M, Moodie EEM. Accuracy of Conventional and Marginal Structural Cox Model Estimators: A Simulation Study. Int J Biostat. 2010;6(2):Article 13.

1. Definitions

- E: indicator variable denoting exposure status (E = 1 for exposed subjects, E = 0 otherwise)
- Y: indicator variable of the event of interest (Y = 1 if subject has experimented the event, Y = 0 otherwise)
- t: observed follow-up time.
- B, B', C and C': two baseline covariates (confounding factors)
- U: unmeasured latent general health baseline variable

2. Relationships between covariates



3. Simulation parameters

- the prevalence of exposure: $p \in \{1\%,2\%,5\%,10\%\}$;
- the number of confounding factors: one binary (B) and one continuous (C), or two binaries (B and B') and two continuous (C and C')
- the strength of the correlation between covariates B and C, and between covariates B' and C': $\sigma_{B,C} = \sigma_{B',C'} \in \{0,0.1,0.3,0.5\}$ (no, weak, moderate, or strong correlation);
- the strength of the association between covariates and U: $\sigma_{U,B} = \sigma_{U,C} = \sigma_{U,B'} = \sigma_{U,C'} \in \{0,0.1,0.3,0.5\}$ (no, weak, moderate, or strong association);
- the strength of the association between covariates and exposure allocation: $\exp(\delta_B) = \exp(\delta_C) = \exp(\delta_{B'}) = \exp(\delta_{C'}) \in \{1,1.2,1.5,2\}$ (no, weak, moderate, or strong association);
- the strength of the marginal association between exposure and outcome: $HR = \exp(\gamma) \in \{1,1.2,1.5,2\}$ (no, weak, moderate, or strong association);
- the censoring rate: $r_c \in \{20\%,50\%,80\%\}$.

4. Data generating process

1) randomly generate of five normally distributed covariates (sample size N = 10000):

$X = [X_B, X_C, X_{B'}, X_{C'}, X_U] \sim N(0,\Sigma)$ with correlation matrix $\Sigma = \begin{bmatrix} 1 & \sigma_{B,C} & 0 & 0 & \sigma_{U,B} \\ \sigma_{B,C} & 1 & 0 & 0 & \sigma_{U,C} \\ 0 & 0 & 1 & \sigma_{B',C'} & \sigma_{U,B'} \\ 0 & 0 & \sigma_{B',C'} & 1 & \sigma_{U,C'} \\ \sigma_{U,B} & \sigma_{U,C} & \sigma_{U,B'} & \sigma_{U,C'} & 1 \end{bmatrix}$

2) transform each variable as follows:
- $B = 1$ if $X_B > 0$, $B = 0$ if $X_B \leq 0$
- $B' = 1$ if $X_{B'} > 0$, $B' = 0$ if $X_{B'} \leq 0$
- $C = X_C$,
- $C' = X_{C'}$,
- $U = P(X_U < x)$

3) draw the exposure allocation E from a Bernoulli distribution $E \sim B(p_z)$, where :
- $p_z = \text{logit}^{-1}(\delta_0 + \delta_B B + \delta_C C)$ when scenario implies variables B and C only, or
- $p_z = \text{logit}^{-1}(\delta_0 + \delta_B B + \delta_C C + \delta_{B'} B + \delta_{C'} C)$ when scenario implies variables B, B', C and C',

δ_0 is chosen using an iterative process so that exposure prevalence in the simulated sample is fixed at a desired proportion p.

4) generate the event time T with exponential distribution as follows:
$T = -Log(U)/(\lambda \exp(\gamma E))$ with $\lambda = 0.1$

5) generate the censoring time T_c drawn from a uniform distribution U(0,c), where c is chosen using an iterative process to achieve a desired censoring rate r_c in the simulated sample.

6) obtain the observed time-to-event outcome with the following decision rule:
- $Y = 1$, $t = T$ if $T <= T_c$
- $Y = 0$, $t = T_c$ if $T > T_c$

5. Reported scenarios

All simulation parameters were crossed using a factorial design, resulting in 6144 different scenarios.
For the sake of concision, only a limited number of scenarios were reported in the 'Results' section. We first chose a reference configuration, and then reported the effects of change of each of the simulation parameters on the results.
The above table lists all the reported scenarios (**the reference configuration is highlighted in bold font** on each line).

	Exposure prevalence	Confounding factors B, B', C and C'	Correlation between covariates B/C and B'/C'	Correlation between covariates and U	Association between covariates and exposure allocation	Marginal association between exposure and outcome (HR)	Censoring rate
Reference configuration	**5%**	**B and C only**	**No**	**Moderate**		**1**	**50%**
Scenarios with a varying prevalence	1%-2%-**5%**-10%	**B and C only**	**No**	**Moderate**		**1**	**50%**
Scenarios with a varying HR	**5%**	**B and C only**	**No**	**Moderate**		**1**-1.2-1.5-2	**50%**
Scenarios with a varying strength of confounding	**5%**	**B and C only**	**No**	No-Weak-**Moderate**-Strong		**1**	**50%**
Scenario with a varying number of confounding factors	**5%**	**B and C only**-B, B', C and C'	**No**	**Moderate**		**1**	**50%**
Scenario with a varying censoring rate	**5%**	**B and C only**	**No**	**Moderate**		**1**	20%-**50%**-80%
Scenario with a varying correlation between covariates	**5%**	**B and C only**	**No**-Weak-Moderate-Strong	**Moderate**		**1**	**50%**

3 | Score pronostique

3.1 Définition et hypothèses

Le score pronostique a été défini par Ben B. Hansen (2008) comme une fonction d'une ou plusieurs covariables permettant d'induire une indépendance conditionnelle entre le devenir potentiel d'un sujet en l'absence d'exposition au traitement et les covariables considérées séparément (Ben B. Hansen 2008 ; Leacy & Stuart 2014). En cas d'exposition Y binaire, certains auteurs l'appellent *disease risk score* (Arbogast & Ray 2009 ; Arbogast & Ray 2011 ; Arbogast et al. 2012 ; Tadrous et al. 2013 ; Wyss et al. 2015 ; Schmidt et al. 2016 ; Connolly & Gagne 2016), le définissant alors comme un score de risque de l'évènement d'intérêt en l'absence d'exposition.

Donc, comme le score de propension, le score pronostique est une mesure résumant plusieurs covariables en une seule dimension. Sa valeur chez un sujet est corrélée au devenir du sujet en l'absence d'exposition au traitement (quelle que soit l'exposition réellement observée dans l'étude). L'analyse par score pronostique nécessite la satisfaction des mêmes hypothèses que celles déjà décrites pour l'analyse par score de propension (Tableau 2.1 page 24), à une différence près concernant la positivité : l'analyse par score pronostique nécessite qu'il n'existe *aucune valeur du score pronostique* pour laquelle le fait d'être traité (ou non traité) est certain, alors que l'analyse par score de propension nécessite qu'il n'existe *aucun profil de sujets*, défini par les covariables considérées séparément, pour lequel le fait d'être traité (ou non traité) est certain (Ben B. Hansen 2008). L'hypothèse de positivité est donc

moins stricte (et plus simple à satisfaire) dans l'analyse par score pronostique.

3.2 Estimation et utilisation

3.2.1 Estimation du score pronostique

Comme le score de propension, le score pronostique réel est inconnu et doit être estimé à partir des données disponibles pour l'analyse. Le score pronostique Ψ_0 est la valeur attendue du critère en l'absence d'exposition au traitement, conditionnellement aux caractéristiques initiales (Leacy & Stuart 2014). Le plus souvent, son estimation utilise des méthodes de régression standards (Arbogast & Ray 2009) adaptées à la nature du critère de jugement Y. Si le critère de jugement est continu, binaire, catégoriel ou ordinal, la méthode la plus simple pour estimer Ψ_0 est d'utiliser un modèle de régression linéaire généralisée estimé au sein du sous-groupe de sujets non exposés au traitement : $f(E(Y|T=0,X)) = X\beta$ où f est la fonction de lien et β le vecteur des coefficients associés aux covariables X. Pour cette étape, plusieurs variantes ont été décrites :

- l'utilisation d'un échantillon de sujets non exposés issu d'une population indépendante (par exemple une population sélectionnée avant la mise sur le marché d'un médicament) (Glynn et al. 2012 ; Schmidt et al. 2016) ;
- l'utilisation de l'ensemble de la population analysée (au lieu du sous-groupe de sujets non exposés seulement), en rajoutant l'exposition T en variable explicative : $f(E(Y|T,X)) = X\beta + T\beta_T$ (en fait, cette dernière méthode conduit à l'estimation du score de Miettinen (Miettinen 1976), qui pourait être moins robuste à une mauvaise spécification du modèle pronostique (Ben B. Hansen 2008 ; Leacy & Stuart 2014 ; Pike et al. 1979)).

Quelle que soit la méthode utilisée pour estimer ce modèle pronostique, les coefficients estimés ($\hat{\beta}$) sont ensuite utilisés pour dériver $\widehat{\Psi}_{0,i}$ (le pronostic en l'absence d'exposition

au traitement) pour chaque sujet i de l'étude, quelle que soit son exposition réellement observée.

3.2.2 Utilisation du score pronostique

Une fois que le score pronostique $\widehat{\Psi}_{0,i}$ est estimé pour chaque sujet i, la seconde étape consiste à l'utiliser dans l'estimation de l'effet de l'exposition au traitement T sur le critère de jugement Y. Trois méthodes d'utilisation du score pronostique, calquées sur trois des méthodes d'utilisation du score de propension vues au chapitre précédent, ont été décrites dans la littérature (Arbogast et al. 2012) : l'ajustement sur le score pronostique, la stratification sur le score pronostique, et l'appariement sur le score pronostique.

Ajustement sur le score pronostique. Le score pronostique estimé $\widehat{\Psi}_0$ est directement inclus dans un modèle multivarié, comportant donc deux variables explicatives du critère de jugement Y : l'exposition T et le score pronostique $\widehat{\Psi}_0$. $\widehat{\Psi}_0$ peut être utilisé tel quel ou divisé en quantiles (par exemple, en quintiles ou déciles) (Arbogast & Ray 2011).

Stratification sur le score pronostique. L'effet du traitement est estimé au sein de sous-groupes définis par le score pronostique estimé. L'approche la plus répandue consiste à utiliser des strates de taille approximativement égales définies sur les quantiles (quintiles ou déciles) du score pronostique. L'estimation de l'effet du traitement au sein de chaque sous-groupe permet d'explorer facilement l'existence d'un effet du traitement différent selon le pronostic (Arbogast et al. 2012) (c'est-à-dire une hétérogénéité de l'effet). La moyenne pondérée de ces estimations est ensuite calculée pour obtenir une estimation de l'effet du traitement dans l'ensemble de la population.

Appariement sur le score pronostique. Les sujets exposés au traitement sont appariés avec des sujets non exposés ayant une valeur proche du score pronostique. Les mêmes procédures d'appariement que pour le score de propension peuvent être

utilisées (Leacy & Stuart 2014 ; Connolly & Gagne 2016). L'effet du traitement est ensuite estimé dans la population appariée.

3.2.3 Brève revue de la littérature

La littérature concernant le score pronostique est beaucoup plus limitée que celle concernant le score de propension. Si nous nous limitons au score pronostique tel que formalisés par Ben B. Hansen (2008), seules cinq études de simulation ont cherché à évaluer les performances des trois méthodes existantes basées sur le score pronostique (Arbogast & Ray 2011 ; Leacy & Stuart 2014 ; Wyss et al. 2015 ; Pfeiffer & Riedl 2015 ; Schmidt et al. 2016). Globalement, ces études ne permettent pas de dégager des recommandations générales d'utilisation des méthodes basées sur le score pronostique. Premièrement, même si certains auteurs conseillent l'utilisation du score pronostique à la place du score de propension quand l'exposition est rare (Arbogast et al. 2012 ; Velentgas et al. 2013), cette recommandation d'experts ne repose sur aucun travail de simulation ayant comparé ces deux approches dans cette situation particulière. Deuxièmement, aucune de ces cinq études n'a comparé les trois méthodes d'utilisation du score pronostique en même temps : Arbogast & Ray (2011) ainsi que Schmidt et al. (2016) ont évalué l'ajustement uniquement ; Leacy & Stuart (2014) ainsi que Wyss et al. (2015) ont évalué l'appariement uniquement ; enfin Pfeiffer & Riedl (2015) ont évalué l'ajustement et l'appariement uniquement. Troisièmement, le type d'effet du traitement (CTE, ATE ou ATT) estimé par chacune de ces méthodes n'est pas clair : la question était soit en partie éludée par le choix d'une mesure d'association collapsible (Arbogast & Ray 2011 ; Leacy & Stuart 2014), soit répondue partiellement par l'évaluation d'un seul type d'effet à la fois (CTE (Pfeiffer & Riedl 2015 ; Schmidt et al. 2016) ou ATT (Leacy & Stuart 2014 ; Wyss et al. 2015), l'ATE n'ayant jamais été évalué). Enfin, seules deux études incluaient des scénarios impliquant une hétérogénéité de l'effet du traitement (Wyss et al. 2015 ; Schmidt et al. 2016).

3.3 Evaluation des méthodes d'utilisation existantes et développement de nouvelles méthodes d'utilisation

Toutes ces questions non répondues à ce jour ont été abordées dans un article accepté dans Statistics in Medicine en 2016 (Hajage, De Rycke, et al. 2016). Ce travail de simulation était illustré par une application sur les données du SNIIRAM cherchant à évaluer l'effet du ratio traitement de fond/traitement total de l'asthme (rapport entre le nombre de délivrances de corticostéroïdes inhalés sur le nombre total de délivrances de médicaments antiasthmatiques (Laforest et al. 2014)) sur la survenue d'exacerbations de l'asthme à 1 an.

Le critère de jugement Y ainsi que l'exposition T (simulés pour $n = 5000$ sujets) étaient binaires. Le modèle pronostique était estimé à l'aide d'une régression logistique au sein du sous-groupe non exposé :

$$\text{logit}(P(Y = 1|T = 0, X)) = X\beta_0,$$

où X était un vecteur de 9 covariables distribuées normalement. Une fois estimés, les coefficients de ce modèle permettaient d'estimer $\widehat{\Psi}_0 = \text{logit}(\hat{p}_0)$, le vecteur des probabilités individuelles d'évènement estimées en l'absence d'exposition et exprimées en logit.

L'étude comparait les trois méthodes d'utilisation existantes du score pronostique ainsi que les quatre méthodes d'utilisation du score de propension décrites au précédent chapitre. La mesure d'association estimée était l'OR (mesure d'association non-collapsible) conditionnel (CTE) ou marginal (ATE ou ATT).

Nous avons également évalué quatre nouvelles méthodes d'utilisation du score pronostique que nous avons développé, chacune cherchant à estimer un type d'effet spécifique (soit le CTE, soit l'ATE, soit l'ATT). Deux de ces méthodes utilisent, en plus de $\hat{\Psi}_0$, un score pronostique dérivé d'un modèle estimé au sein du sous-groupe *exposé* au traitement, appelé

$\hat{\Psi}_1 = \text{logit}(\hat{p}_1)$.

— La première de ces nouvelles méthodes (SPN-CTE1), développée pour estimer le CTE, inclut $\widehat{\Psi}_0$ comme un terme *offset* au sein d'un modèle logistique estimé au sein du sous-groupe exposé au traitement :

$$\text{logit}(P(Y = 1|T = 1)) = \widehat{\Gamma}_1 + \widehat{\Psi}_0.$$

L'effet du traitement $\widehat{\Gamma_1}$ est donc estimé en comparant le risque individuel d'évènement observé sous traitement (terme à gauche du signe égal) au risque individuel d'évènement prédit en l'absence de traitement (c'est-à-dire $\widehat{\Psi}_0$).

— La deuxième méthode (SPN-ATT), développée pour estimer l'ATT, peut être décrite de manière similaire à la première méthode, c'est-à-dire en incluant un terme *offset* dans un modèle logistique estimé au sein du sous-groupe exposé au traitement :

$$\text{logit}(P(Y = 1|T = 1)) = \widehat{\Gamma}_2 + \text{logit}(\widehat{P}_0^1),$$

où $\widehat{P}_0^1$ est le taux d'évènements estimé dans le sous-groupe exposé au traitement en l'absence d'exposition : $\widehat{P}_0^1 = \frac{\sum_{i=1}^n T_i \hat{p}_{0,i}}{\sum_{i=1}^n T_i}$. Dans ce dernier modèle, le terme *offset* a la même valeur pour tous les sujets. Par conséquent, le terme à gauche peut être remplacé par son estimation empirique $\widehat{O^1} = \frac{\sum_{i=1}^n T_i Y_i}{\sum_{i=1}^n T_i}$, c'est-à-dire le taux d'évènements observés dans le sous-groupe exposé au traitement. L'effet du traitement est donc, en fait, simplement évalué par

$$\widehat{\Gamma}_2 = \text{logit}(\widehat{O^1}) - \text{logit}(\widehat{P}_0^1).$$

Cette méthode compare donc le taux d'évènements observés chez les sujets traités au taux d'évènement prédits chez ces mêmes sujets mais en l'absence d'exposition au traitement.

— La troisième méthode (SPN-CTE2), développée pour estimer le CTE, utilise à la fois

$\widehat{\Psi}_0$ et $\widehat{\Psi}_1$. L'effet du traitement est en effet estimé, dans l'ensemble de la population, par :

$$\widehat{\Gamma}_3 = \widehat{L}_1 - \widehat{L}_0,$$

où $\widehat{L}_1 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Psi}_{1,i}$ et $\widehat{L}_0 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Psi}_{0,i}$, c'est-à-dire la moyenne des différences entre les valeurs individuelles du score pronostique en présence de l'exposition et les valeurs individuelles du score pronostique en l'absence d'exposition.

— Enfin, la quatrième méthode (SPN-ATE), développée pour estimer l'ATE, utilise également $\widehat{\Psi}_0$ et $\widehat{\Psi}_1$:

$$\widehat{\Gamma}_4 = \text{logit}(\widehat{P}_1) - \text{logit}(\widehat{P}_0),$$

où $\widehat{P}_1 = \frac{\sum_{i=1}^{n} \hat{p}_{1,i}}{n}$ et $\widehat{P}_0 = \frac{\sum_{i=1}^{n} \hat{p}_{0,i}}{n}$. L'effet du traitement est donc estimé, dans l'ensemble de la population, par la différence entre le taux d'évènements prédits en présence de l'exposition et le taux d'évènements prédits en l'absence d'exposition.

Les différents scénarios évalués étaient définis par la prévalence de l'exposition T, le taux d'évènements Y, l'effet de l'exposition sur le risque d'évènement, et la présence d'une hétérogénéité de l'effet du traitement.

La conclusion de cette étude était qu'aucune des trois méthodes d'utilisation existantes du score pronostique ne permettait d'estimer l'ATE : l'ajustement sur le score pronostique fournissait une estimation non biaisée du CTE, la stratification une estimation biaisée du CTE, et l'appariement une estimation non biaisée de l'ATT. Cette dernière méthode était donc, jusqu'à notre travail, la seule méthode d'utilisation du score pronostique (« équivalent pronostique des scores de propension » selon Ben B. Hansen (2008)) permettant d'estimer un effet marginal. De plus, ces trois méthodes sous-estimaient systématiquement la variabilité de l'effet du traitement, particulièrement en cas de forte prévalence de l'exposition, avec pour conséquence des taux de recouvrement parfois très éloignés du taux nominal de 95%. Nos nouvelles méthodes d'utilisation du score pronostique estimaient le type d'effet pour lequel elles ont chacune été développées, et avaient de bonnes performances quelle que soit la prévalence de l'exposition. Ces bonnes performances s'expliquaient aussi par les es-

timateurs de variance utilisés (qui seront détaillés dans le prochain chapitre de cette thèse). Enfin, la nouvelle méthode d'estimation de l'ATE basée sur le score pronostique (SPN-ATE) avait des performances supérieures à la pondération sur le score de propension en cas d'exposition peu fréquente. Les nouvelles méthodes d'utilisation du score pronostique développées dans ce travail constituent donc une alternative intéressante aux méthodes basées sur le score de propension, particulièrement en cas d'exposition rare.

Article 2

Research Article

Statistics in Medicine

Received 17 March 2016, Accepted 21 October 2016 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7170

Estimation of conditional and marginal odds ratios using the prognostic score

David Hajage,[a,b,c,e*†] Yann De Rycke,[a,b,c,e] Guillaume Chauvet[f,g] and Florence Tubach[a,b,d,e]

Introduced by Hansen in 2008, the prognostic score (PGS) has been presented as 'the prognostic analogue of the propensity score' (PPS). PPS-based methods are intended to estimate marginal effects. Most previous studies evaluated the performance of existing PGS-based methods (adjustment, stratification and matching using the PGS) in situations in which the theoretical conditional and marginal effects are equal (i.e., collapsible situations). To support the use of PGS framework as an alternative to the PPS framework, applied researchers must have reliable information about the type of treatment effect estimated by each method. We propose four new PGS-based methods, each developed to estimate a specific type of treatment effect. We evaluated the ability of existing and new PGS-based methods to estimate the conditional treatment effect (CTE), the (marginal) average treatment effect on the whole population (ATE), and the (marginal) average treatment effect on the treated population (ATT), when the odds ratio (a non-collapsible estimator) is the measure of interest. The performance of PGS-based methods was assessed by Monte Carlo simulations and compared with PPS-based methods and multivariate regression analysis. Existing PGS-based methods did not allow for estimating the ATE and showed unacceptable performance when the proportion of exposed subjects was large. When estimating marginal effects, PPS-based methods were too conservative, whereas the new PGS-based methods performed better with low prevalence of exposure, and had coverages closer to the nominal value. When estimating CTE, the new PGS-based methods performed as well as traditional multivariate regression. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: causal inference; confounding; observational study; prognostic score; propensity score

1. Introduction

As part of observational prospective studies comparing treated versus untreated subjects on a binary outcome, the approaches used to estimate the treatment effect depend mainly on the clinical objective [e.g., whether an applied researcher is interested in the treatment effect at the subject level (conditional treatment effect) or at the population level (marginal treatment effect)]. Conditional treatment effect evaluates the effect of switching the treatment status of a particular individual profile and therefore, may be more relevant from a clinical perspective. Marginal treatment effect is an averaged estimation useful for evaluating policy changes or public health interventions over a relevant target population and may mimic the treatment effect measured in a randomized clinical trial. Conditional and marginal treatment effect measures are equal when risk differences or relative risks are used to quantify the association between the exposure and the binary outcome but not when a non-collapsible estimator of the treatment effect is used, such as the odds ratio, even in the absence of treatment effect heterogeneity [1, 2]. Therefore, estimating the MTE is especially important in observational studies when the objective is to verify that the MTE

[a]*APHP, Hôpital Pitié-Salpêtrière, Département de Biostatistiques, Santé publique et Information médicale, F-75013, Paris, France*
[b]*APHP, Centre de Pharmacoépidémiologie (Cephepi), F-75013, Paris, France*
[c]*Univ Paris Diderot, Sorbonne Paris Cité, UMR 1123 ECEVE, F-75010, Paris, France*
[d]*Université Pierre et Marie Curie – Paris 6, Sorbonne Universités, Paris, France*
[e]*INSERM, UMR 1123 ECEVE, F-75018, Paris, France*
[f]*Ecole Nationale de la Statistique et de lAnalyse de l'Information (ENSAI), F-35170, Bruz, France*
[g]*IRMAR, UMR CNRS 6625, Rennes, France*
[*]*Correspondence to: D. Hajage, APHP, Hôpital Pitié-Salpêtrière, Département de Biostatistiques, Santé publique et Information médicale, F-75013, Paris, France.*
[†]*E-mail: david.hajage@aphp.fr*

estimated in a randomized clinical trial can be reproduced in the 'real-world setting': when the OR (or other non-collapsible measures) are compared across different studies, the same type of estimate must be used to interpret any difference [3].

Among the methods used to account for confounding factors in observational studies, multivariate regression analysis and propensity score (PPS) analysis are the most common. Multivariate regression analysis allows for estimating the CTE. The PPS framework was developed to induce balance of observed confounding factors between groups of treated and untreated subjects [4] and is designed to estimate the MTE [5].

In 2008, Hansen proposed an alternative to the PPS framework, the prognostic score (PGS), presented as 'the prognostic analogue of the propensity score' and intended to estimate MTE as well [6]. Indeed, the scores share many similarities: both are summary scores, both reduce observed confounding, and both imply a two-step analysis [7]. However, Hansen discussed the theoretical properties of PGS framework, not the different methods that account for the estimated PGS. Different methods could lead to different types of treatment effect estimates.

Prognostic score methods are increasingly used particularly in pharmacoepidemiology [8, 9], but to the best of our knowledge, no study has focused on the type of estimate provided by the three reported PGS-based methods – covariate adjustment on the PGS [7, 10], stratification on the PGS [7, 9], and matching on the PGS [11, 12] – in the context of a non-collapsible measure of treatment effect. To support the use of the PGS framework as an alternative to the PPS framework in some settings, applied researchers must have reliable information about the type of treatment effect estimated by each method.

In this study, we focused on the treatment effect estimation in non-randomized observational studies comparing treated versus untreated subjects on a binary outcome, with the OR used to measure the treatment effect. To this end, we used Monte Carlo simulations to examine the ability of multivariate logistic regression, PPS-based and PGS-based methods (the three previously cited ones, termed *existing methods* as well as newly developed ones, termed *new methods*) to estimate the conditional and marginal OR associated with a binary outcome. Using a non-collapsible measure such as the OR allows for more easily identifying the type of effect estimated by each method, even if the use of collapsible estimators such as relative risks or risk differences may be preferable where applicable [13].

The remainder of this article is organized as follows. Section 2 briefly reviews the definition of several types of treatment effects. Section 3 provides a brief explanation of PPS and PGS concepts. Section 4 details the methods evaluated in this study, and Section 5 describes the data-generating process and the performance criteria used to evaluate and compare these methods. Section 6 provides the simulation results, and Section 7 provides a real case study aimed at estimating the effect of high inhaled corticosteroids to total asthma drug ratio on the occurrence of asthma-related exacerbations. Section 8 discusses study results and provides areas for future research.

2. Different types of measures of treatment effect

We first review some definitions of treatment effects commonly used in observational studies.

2.1. Definitions

Let T be an indicator variable denoting treatment status ($T = 1$ for treated subjects, $T = 0$ otherwise) and Y be an indicator variable of the binary event of interest ($Y = 1$ if the subject has experienced the event, $Y = 0$ otherwise). Furthermore, let Y_1 and Y_0 denote the counterfactual outcomes with or without exposure to treatment, whatever the real exposure status.

2.2. Individual treatment effect

For a particular subject i, the individual treatment effect is simply $Y_{1,i} - Y_{0,i}$. Because we cannot observe $Y_{0,i}$ and $Y_{1,i}$ simultaneously, the individual treatment effect is not observable.

2.3. Conditional treatment effect

The CTE is the average effect of treatment for a particular set of covariate values (i.e., an estimation of the treatment effect at the subject level). The regression coefficient associated with the treatment from a multivariate logistic model including the treatment status and other covariates is an estimate of the CTE (logit of OR). The CTE is constant for subjects with the same set of covariates values, and constant

over all subjects if there is no effect modification (e.g., no interaction between the treatment status and other covariates).

2.4. Marginal treatment effect

The MTE is the average effect of the treatment on a particular set of subjects (i.e., at the population level). The most popular methods to estimate MTE in observational studies are based on PPS. When the treatment effect is measured with an OR, conditional and marginal treatment effects are not collapsible, and estimates of conditional and marginal treatment effects will not coincide (except if the true effect is null). The MTE depends on the set of subjects for whom the average effect is computed. In this simulation study, we focused on the two most studied treatment effects derived from two particular sets of subjects: the average treatment effect on the entire population under study (ATE) and the average treatment effect on the subjects who received the treatment (ATT).

3. Propensity and prognostic scores overview

3.1. Propensity score

3.1.1. Theory. The PPS is defined as the probability of treatment exposure conditionally on the observed baseline covariates [14]. Conditionally on the PPS, the distribution of observed baseline covariates is independent of exposure to treatment. If the assumptions of consistency (subjects' potential outcome under their observed treatment status is equal to their observed outcome), exchangeability (which implies no unmeasured confounding), positivity (every subject has a non-zero probability of both treatment and non-treatment) and no misspecification of the propensity score model [15] are satisfied, the use of PPS overcomes the problem of selection bias in observational studies, thereby inducing balance of observed characteristics between groups of exposed and unexposed subjects [4]. In experimental designs (e.g., a randomized clinical trial), the true PPS is known for all subjects under study. In observational designs (e.g., a cohort study), the true PPS is unknown and must be estimated from the available observed data. PPS-based methods can provide marginal estimates of the treatment effect [5]. Depending on the method used, ATE or ATT is estimated [16].

3.1.2. Methods. Propensity score analysis works with two successive steps [17]. The first step corresponds to the estimation of Φ, that is, the (logit of) vector of individual probabilities of being exposed conditionally to observed baseline covariates ($\Phi = \text{logit}(p_T) = \text{logit}(p(T = 1|X))$, where $X = (X_1, \ldots, X_K)$ is a vector of K baseline measured covariates). Several strategies have been proposed to estimate individual probabilities of exposure [18], but this estimation usually involves a logistic regression model, regressing observed treatment assignment on baseline covariates [17]. Then, regression coefficients are used to derive $\hat{\Phi}_i$ for each subject i. In practice, the PPS should be estimated by including all variables related to the outcome (i.e., true confounders and prognostic variables) in the PPS estimator model [19, 20].

For the second step, four methods have been described to account for the estimated PPS in estimating the treatment effect: adjustment on the PPS [5], stratification on the PPS [21, 22], matching on the PPS [4, 23, 24], and weighting using the PPS [25, 26]. Several authors have used Monte Carlo simulations to demonstrate that adjustment and stratification poorly estimate marginal ORs as compared with matching and weighting [2, 27, 28]. Both methods also poorly estimate conditional ORs [5]. Moreover, PPS-matching and PPS-weighting can more effectively reduce the imbalance between exposed and unexposed subjects [29].

For the adjustment on the PPS method, the individual estimated PPS $\hat{p}_T$ is directly included as a covariate in the logistic regression model evaluating the treatment effect.

Stratification on the PPS consists of stratifying the sample according to quantiles of the PPS. The treatment effect is then estimated within each stratum and pooled (by using a weighted mean) to estimate the overall effect. Stratification on quintiles of the PPS is widely used in practice, because it may remove 90% of the bias because of measured confounders [30]. This approach was retained in the current study.

Matching on the PPS consists of matching exposed and unexposed subjects with similar values of PPS. Different approaches could be used to match [31], and the most simple and common approach is greedy nearest-neighbor 1:1 matching without replacement within specified caliper widths [32] (the approach used in the simulation study to follow). Calipers of width equal to 0.2 of the standard deviation

of $\hat{\Phi}$ perform well in a wide variety of settings [17]. Once matching is completed, a univariate logistic regression model with exposure as the only covariate could be used to estimate ATT [33].

The weighting method consists of applying a logistic regression model to the outcome with exposure as the only covariate, each subject being weighted according to its PPS value, and using a robust estimator of the standard error [15, 16]. In the resulting weighted pseudo-population, exposed and unexposed groups will tend to have balanced characteristics. Several types of weights (W) could be used depending on the population of interest [34]. Weights to estimate ATE are calculated as follows: $W_{ATE} = \frac{T}{\hat{p}_T} + \frac{1-T}{1-\hat{p}_T}$. Additionally, weighting each subject by the probability of being treated allows for estimating ATT: $W_{ATT} = \hat{p}_T \times W_{ATE} = T + \frac{\hat{p}_T}{1-\hat{p}_T}(1-T)$. To decrease the variability of ATE estimation, stabilized weights [15] could also be used by multiplying previous (un-stabilized) weights W_{ATE} by $T\bar{p}_T + (1-T)(1-\bar{p}_T)$ (where $\bar{p}_T$ is the overall probability of being treated estimated in the sample).

3.1.3. Some strengths and caveats. The use of PPS has been recommended in studies assessing the association between a common binary exposure and rare event(s) [8]. Indeed, PPS analysis can outperform multivariate conditional analysis when accounting for many confounding factors. In this situation, multivariate analysis may encounter convergence problems, particularly when the number of events of interest is small [35, 36]. The PPS, by reducing the dimensionality problem of controlling for multiple confounders into one-dimensional summary measure, is less sensitive to this issue. Nevertheless, all PPS-based methods rely on the validity of estimates of individual exposure probability and thus on the validity of the logistic regression fitted for these estimations. In cases of rare exposure, another dimensionality issue could arise, and a recent work has shown that PPS-based methods could provide biased estimates of the treatment effect in this setting, particularly when focusing on ATE [37].

Although generalizations of PPS methods have been proposed in cases of multiple [38] or non-binary exposures [39], these approaches are complex, and most observational studies involving PPS-based methods assess binary exposures [35].

3.2. Prognostic score

3.2.1. Theory. Hansen's prognostic score [6] (also called 'disease risk score' when the outcome of interest is binary) is defined as 'any scalar or vector-valued function of the covariates that when conditioned on induces independence between the potential outcome under the control condition and the unreduced covariate' (i.e., covariates considered separately) [11]. Thus, like the PPS, it is a summary measure of the covariates. Its value is related to the outcome of the corresponding subject not exposed to treatment (possibly contrary to fact). Similar to the true PPS, the true PGS is in practice unknown and can be estimated by using standard regression modeling techniques [10]. If the outcome of interest is continuous, binary, categorical, or ordinal, one possible PGS is given by the conditional expectation of the outcome under the unexposed condition given the observed covariates [11].

3.2.2. Methods. Like the PPS framework, PGS analysis works with two successive steps [40]. The first step corresponds to the estimation of Ψ_0, the vector of individual expected outcomes under the unexposed status conditional on the observed baseline covariates. If the outcome of interest is binary, Ψ_0 corresponds to the vector of the logit of expected probabilities of an event ($\Psi_0 = \text{logit}(p_E) = \text{logit}(p(Y=1|T=0,X))$, where $X = (X_1, \ldots, X_K)$ is a vector of K baseline observed covariates).

The simplest method to estimate Ψ_0 is to use a logistic regression model fitted in the unexposed population, regressing the observed outcome on baseline covariates. The use of an independent set of unexposed subjects (as opposed to the 'same-sample' estimation) has also been described for this step [7, 41]. The use of the full sample (instead of the unexposed subgroup only) leads to the estimation of Miettinen's multivariate confounder score [42], which may be less robust to model misspecification than Hansen's PGS [6, 11, 43]. Whatever the sample used to fit the prognostic model, the estimated regression coefficients are then used to derive $\hat{\Psi}_{0,i}$ for each subject i under study, whatever the real exposure status.

Three methods have been reported to account for PGS for estimating the treatment effect [9]: adjustment on the PGS, stratification on the PGS and matching on the PGS. Thus, the three methods are similar to the three PPS-based methods presented earlier. These three PGS-based methods act as if the PGS were a known quantity rather than an estimate of an unknown quantity.

Adjustment on the PGS simply consists of including $\hat{\Psi}_0$ in addition to T in the logistic regression model evaluating the treatment effect. $\hat{\Psi}_0$ could be expressed as a linear term or as score quantiles (e.g., quintiles or deciles) [40].

Stratification on the PGS involves comparing outcomes between treated and untreated subjects within strata defined by the PGS. A common approach is to use approximately equal sized strata defined by the quantiles (quintiles or deciles) of the PGS. Then, the effect of treatment on outcome is estimated within each stratum, so this strategy provides a natural way to examine the presence of an effect modification [9]. The weighted mean of the within-stratum treatment effects is calculated to obtain the overall (pooled) treatment effect.

Matching on the PGS consists of matching exposed and unexposed subjects with similar values of the PGS. The same matching procedure as for PPS-matching could be used, with the same width of caliper (0.2 of the standard deviation of $\hat{\Psi}_0$)[11,44]. When the outcome is binary, matching cases (subjects with $Y = 1$) and controls (subjects with $Y = 0$) on the PGS has been reported [45].

3.2.3. Brief literature overview. If we restrict our discussion to Hansen's definition of the PGS, to our knowledge, five simulation studies have evaluated the performance of PGS-based methods: Arbogast *et al.* [40], Leacy and Stuart [11], Wyss *et al.* [12], Pfeiffer and Riedl [45] and Schmidt *et al.* [41]. The first two focused on a collapsible measure of treatment effect. Arbogast *et al.* simulated a Poisson outcome with a binary exposure (exposure prevalence was 10%) and compared the performance of adjustment on the PGS, adjustment on the PPS, and multivariate regression modeling. The authors found that PGS analysis performed well unless covariates used to derive the PGS were strongly associated with the exposure and weakly with the outcome or unless the number of events per confounder was small (< 5 events per covariate). They found no setting for which the use of PGS performed better than multivariate regression or (adjustment on) PPS, except when the traditional model was misspecified because of exclusion of covariates associated with the outcome. Leacy and Stuart considered a continuous outcome and a binary exposure (exposure prevalence 20%) and compared different matching methods, accounting for PGS, PPS, and a combination of both, to estimate the average treatment effect on treated subjects. They concluded that 'methods combining the estimated propensity and prognostic scores should be preferred to methods utilizing the propensity score alone, but that full matching on a prognostic score may be preferred if the researcher is confident that the prognostic score model has been correctly specified'. Wyss *et al.* considered a binary outcome and a binary exposure (exposure prevalence 30%) and compared the performance of matching on the PGS and matching on the PPS to estimate the ATT with the OR. By using simulations and an empirical example, the overlap between treatment groups was often greater with the PGS than the PPS, so the use of PGS may lead to matching a larger proportion of the treated population. Pfeiffer and Riedl [45] compared PPS and PGS in the context of a binary outcome (using Poisson and logistic regression models) with estimate the CTE. The authors found no bias in estimation when the model was adjusted for the PGS on a log/logit scale. Matching cases and controls on the PGS and analyzing them by using conditional logistic regression also yielded unbiased estimates of conditional effect. Finally, Schmidt *et al.* compared multivariate logistic regression, adjustment and weighting on the PPS, and adjustment on the PGS to evaluate a conditional OR with a low number of events and a low number of exposed subjects per covariates. PGS models were estimated in large independent training samples ($n = 5000$) and used in small test samples ($n = 400$). The exposure prevalence was 50% in all simulated datasets. At the highest number of events and exposed subjects per covariate ($\geqslant 2.5$ per covariate), PPS models performed better than multivariate logistic regression and the PGS model. At the lowest number (0.5 per covariate), levels of bias and coverage were unacceptable for all methods, even if the PGS model was the less biased method.

Among these five simulations studies, only two included one or more scenarios with the presence of an effect modification: Wyss *et al.* [12], and Schmidt *et al.* [41]. In Schmidt *et al.*, this effect modification was present only in the training datasets, to assess the effect of a misspecification of the PGS model.

3.2.4. Some strengths and caveats. The PGS has been recommended in studies investigating the association between a common outcome and rare exposure(s) [8,46]. Indeed, the Effective Health Care Program (which aims to produce effectiveness and comparative effectiveness research for clinicians, consumers, and policymakers) recommends the use of the PGS instead of the PPS when the exposure is infrequent (without defining which exposure should be considered infrequent) [8]. Because the existing methods rely on a PGS derived by using the unexposed population only, PGS-based methods seem attractive in cases of rare exposure, that is, situations with a high number of unexposed subjects. This scenario could be particularly true when the estimator of interest is the ATT but may not apply when the estimator of interest is the ATE, because Hansen demonstrated that identifying the ATE requires additional conditioning on any effect modifiers (e.g., but not limited to, cases of interaction between the exposure and some

of the other covariates). To estimate the ATE, conditioning on both a PGS developed within unexposed subjects and a PGS developed within exposed subjects seems necessary according to Leacy and Stuart [11] (citing an unpublished work by Waernbaum). All existing PGS-based methods use a PGS developed within unexposed subjects only, whereas a new PGS method to estimate the ATE by using both scores may encounter an issue when exposure is rare (because of the score developed with a low number of exposed subjects), which may limit the theoretical advantage of the PGS in this setting. Furthermore, we found no study evaluating the effect of exposure prevalence on the performance of PPS and PGS methods [37].

With a new treatment, it might be difficult to identify confounders that predict treatment, so there might be a reason to choose the PGS framework in that setting [12]. Also, the PGS is easier to use than the PPS in cases of multiple exposures or multiple exposure levels because of no need to model a complex exposure in the first step [8, 46]. Nevertheless, the use of the PGS when the outcome does not follow a generalized linear model (e.g., a time-to-event outcome) does not seem straightforward and has not been reported, except in one empirical example (with matching on the PGS method) [12].

Finally, the use of the PGS is much more limited than the PPS in the biomedical and statistical literature. The relative merits of the different PGS-based methods, the type of estimated treatment effect provided by each method, the type of covariates that need to be included in the PGS model, their advantages and disadvantages compared with the PPS, the development of use and reporting guidelines need further investigation.

3.3. Summary

The literature is less abundant regarding the PGS than the PPS. Outside the 'common outcome, rare exposure' situation, for which PGS-based methods could be preferred to PPS-based methods (regardless of the type of treatment effect), the two frameworks seem in a position of equipoise when choosing between the two methods to evaluate a binary exposure. The PGS-matching method leading to a greater number of matches as compared with PPS-matching is an important advantage but is not sufficient to recommend the wide use of the PGS: the performance of PGS-matching and all other PGS and PPS-based methods still needs to be compared. Finally, the type of treatment effect (CTE, ATE, or ATT) estimated by each PGS-based method is unclear: the few authors who paid attention to this issue overcame the problem by using a collapsible measure [11, 40] or evaluated only one effect type at a time (CTE [41, 45] or ATT [11, 12]; ATE has never been explored).

4. Statistical methods

All statistical methods in this study are described hereafter (and in Table I). To simplify our notations, the estimators $\hat{\Gamma}$ are numbered for only the four new PGS-based methods that we propose in Section 4.4 ($\hat{\Gamma}_1$ to $\hat{\Gamma}_4$).

4.1. Multivariate logistic regression model)

A standard logistic regression model was fitted: $\text{logit}(P(Y = 1|T, \boldsymbol{X})) = \hat{\beta}_0 + \hat{\Gamma}T + \sum_{k=1}^{K} \hat{\beta}_k X_k$. The coefficient $\hat{\Gamma}$ is an estimate of the CTE.

4.2. Propensity score-based methods

First, the following propensity model was estimated in the full sample: $P(T = 1|\boldsymbol{X}) = \text{logit}^{-1}(\hat{\alpha}_0 + \sum_{k=1}^{K} \hat{\alpha}_k X_k)$. We estimated $\hat{\Phi} = \text{logit}(\hat{p}_T)$ from the predicted probability of treatment given subjects' baseline covariates.

Then, PPS adjustment (PPS-ADJ), PPS stratification (PPS-STRAT), PPS matching (PPS-MATCH) and PPS weighting (with W_{ATE} (PPS-WATE) and W_{ATT} (PPS-WATT)) were used as described in Section 3.1.2.

4.2.1. Propensity score adjustment. The individual estimated PPS was included as a covariate in a logistic regression model:

$$\text{logit}(P(Y = 1|T, \hat{p}_T)) = \hat{\beta}_0 + \hat{\Gamma}T + \hat{\beta}\hat{p}_T.$$

4.2.2. Propensity score stratification. For PPS stratification, the sample was divided according to quintiles of PPS distribution. Then, the following model was fitted in each stratum:

$$\text{logit}(P(Y = 1|T, S = s)) = \hat{\beta}_{0,s} + \hat{\Gamma}_s T$$

where $s = 1, \ldots, 5$ refers to the stratum.

Table I. Evaluated statistical methods.

Method	Score used		
	Φ	Ψ_0	Ψ_1
Traditional multivariate regression			
MULTI			
PPS-based methods			
PPS-ADJ	✓		
PPS-STRAT	✓		
PPS-MATCH	✓		
PPS-WATE	✓		
PPS-WATT	✓		
Existing PGS-based methods			
PGS-ADJ		✓	
PGS-STRAT		✓	
PGS-MATCH		✓	
New PGS-based methods			
PGS-CTE1		✓	
PGS-ATT		✓	
PGS-CTE2		✓	✓
PGS-ATE		✓	✓

ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment effect on the treated population; CTE, conditional treatment effect; MATCH, matching; MULTI, Multivariate logistic regression model; PGS, prognostic score; PPS, propensity score; STRAT, stratification.

Finally, the results in each stratum were pooled:

$$\hat{\Gamma} = \sum_{s=1}^{5} \frac{n_s}{n} \hat{\Gamma}_s \text{ and } \hat{V}(\hat{\Gamma}) = \sum_{s=1}^{5} \left(\frac{n_s}{n}\right)^2 \hat{V}(\hat{\Gamma}_s)$$

with n_s the sample size in stratum s, n the total sample size, and $\hat{V}(\hat{\Gamma}_s)$ the model-based estimators of the variance. The use of PPS quintiles implied that $n_s \approx \frac{n}{5}$.

4.2.3. Propensity score matching. We used greedy nearest-neighbor 1:1 matching without replacement: each treated subject was randomly selected and matched to the nearest untreated subject based on a caliper width of 0.2 of the standard deviation of Φ. Then, a logistic regression model was fitted in the matched sample:

$$\text{logit}(P(Y = 1|T)) = \hat{\beta}_0 + \hat{\Gamma} T.$$

We used a robust estimator of the standard error of the regression coefficient that accounted for the clustering within matched sets [33].

4.2.4. Propensity score weighting. A logistic regression model was fitted in the entire population, with the treatment as the only covariate, and each subject was weighted in the likelihood by using the (unstabilized) weights W_{ATE} or W_{ATT} defined in Section 3.1.2. Robust estimators of the standard errors were used [15].

4.3. Existing prognostic score-based methods

Two prognostic models were fitted – one in the subgroup of unexposed subjects, the other in the subgroup of exposed subjects – as follows:

$$\text{logit}(P(Y = 1|T = 0, \boldsymbol{X})) = \hat{\beta}_{00} + \sum_{k=1}^{K} \hat{\beta}_{0k} X_k \quad (1)$$

$$\text{logit}(P(Y=1|T=1,\boldsymbol{X})) = \hat{\beta}_{10} + \sum_{k=1}^{K} \hat{\beta}_{1k} X_k \tag{2}$$

From Equation (1), we derived $\hat{\Psi}_0 = \text{logit}(\hat{p}_0)$, the vector of estimated individual prognostics as if all subjects were unexposed; from Equation (2), we derived $\hat{\Psi}_1 = \text{logit}(\hat{p}_1)$, the vector of estimated individual prognostics as if all patients were exposed to treatment.

Then, PGS adjustment (PGS-ADJ), PGS stratification (PGS-STRAT) and PGS matching (PGS-MATCH) were used as described in Section 3.2.2. For these three methods (referred to as *existing PGS-based methods*), only $\hat{\Psi}_0$ estimated from the unexposed prognostic model (1) was used for treatment effect estimation.

Matching cases and controls on the PGS is reported in only one simulation study (comparing its performance against matching cases and controls on the PPS [45]). This method was not evaluated in this article.

4.3.1. Prognostic score adjustment. PGS adjustment is similar to PPS adjustment and consists in modeling the probability of an event by a logistic regression model in the entire population, by using the individual prognostic score $\hat{\Psi}_0$ as a covariate, in addition to the exposure status T:

$$\text{logit}\left(P\left(Y=1|T,\hat{\Psi}_0\right)\right) = \hat{\beta}_0 + \hat{\Gamma}T + \hat{\beta}\hat{\Psi}_0.$$

4.3.2. Prognostic score stratification. For PGS stratification, the sample was divided by quintiles of the Ψ_0 distribution. Then, an approach similar to PPS stratification (Section 4.2.2) was used to estimate the overall treatment effect and its variance.

4.3.3. Prognostic score matching. PGS matching was performed similar to PPS matching, based on a caliper width of 0.2 of the standard deviation of Ψ_0, and robust estimate of the standard error.

4.4. New Prognostic score-based methods

Four new methods were evaluated. These methods involve only $\hat{\Psi}_0$ or both $\hat{\Psi}_0$ and $\hat{\Psi}_1$. Each method was developed with a specific type of treatment effect in mind (CTE, ATE or ATT). Furthermore, unlike the three existing methods, they involve a variance estimation method that accounts for the fact that the true PGSs (Ψ_0 and/or Ψ_1) are unknown and have been estimated (Section 4.4.3).

4.4.1. Methods using only the unexposed estimated prognostic score $\hat{\Psi}_0$. The first method (PGS-CTE1) was designed to estimate the CTE. It consists of modeling the probability of event by a logistic regression model in the exposed sample by using the individual prognostic score $\hat{\Psi}_0$ as an offset term (i.e., with a parameter estimate constrained to 1):

$$\text{logit}(P(Y=1|T=1)) = \hat{\Gamma}_1 + \hat{\Psi}_0. \tag{3}$$

The second method (PGS-ATT) was designed to estimate the ATT. It is similar to the previous method but uses the expit of the prognostic score ($\text{logit}^{-1}(\hat{\Psi}_0) = \hat{p}_0$) averaged over the population of exposed subjects:

$$\text{logit}(P(Y=1|T=1)) = \hat{\Gamma}_2 + \text{logit}\left(P_0^1\right), \tag{4}$$

where P_0^1 is the estimated probability of an event in the exposed population under the unexposed condition: $P_0^1 = \frac{\sum_{i=1}^n T_i \hat{p}_{0,i}}{\sum_{i=1}^n T_i}$. In the model (4), the offset term has the same value for all exposed subjects, and therefore the left term could be replaced by its empirical estimation $O^1 = \frac{\sum_{i=1}^n T_i Y_i}{\sum_{i=1}^n T_i}$, that is, the observed probability of an event in the exposed population. Thus, the treatment effect is estimated as follows:

$$\hat{\Gamma}_2 = \text{logit}(O^1) - \text{logit}\left(P_0^1\right), \tag{5}$$

(i.e., the difference between the [logit of] observed probability of an event in the exposed population and the [logit of] estimated probability of an event of the same population under the unexposed condition).

4.4.2. Methods using both unexposed and exposed estimated prognostic scores $\hat{\Psi}_0$ *and* $\hat{\Psi}_1$. The third method (PGS-CTE2) was designed to estimate the CTE by using $\hat{\Psi}_1$ in addition to $\hat{\Psi}_0$:

$$\hat{\Gamma}_3 = L_1 - L_0, \tag{6}$$

with $L_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}_{1,i}$ and $L_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{\Psi}_{0,i}$ (i.e., the mean of the difference between individual PGS of the overall population under the exposed condition and the individual PGS of the overall population under the unexposed condition).

Finally, the fourth method (PGS-ATE) was designed to estimate the ATE:

$$\hat{\Gamma}_4 = \text{logit}(P_1) - \text{logit}(P_0), \tag{7}$$

with $P_1 = \frac{\sum_{i=1}^{n} \hat{p}_{1,i}}{n}$ and $P_0 = \frac{\sum_{i=1}^{n} \hat{p}_{0,i}}{n}$ (i.e., the difference between the [logit of] estimated probability of an event in the overall population under the exposed condition and the [logit of] estimated probability of an event of the overall population under the unexposed condition).

4.4.3. Variance of $\hat{\Gamma}$ *estimators.* An approximately unbiased variance estimator of $\hat{\Gamma}$ for the four new PGS-based methods was obtained by using the influence function linearization technique developed by Deville [47]. Briefly, for a parameter Γ and an estimator $\hat{\Gamma}$, the linearization involves finding a variable U such that

$$E\left(\frac{s_U^2}{n}\right) \simeq V(\hat{\Gamma})$$

where

$$s_U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \bar{U})^2 \text{ and } \bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i.$$

The variable U usually depends on unknown parameters, which can be estimated from the sample to obtain an estimated linearized variable $\hat{U}$. This leads to the variance estimator:

$$\hat{V}(\hat{\Gamma}) = \frac{s_{\hat{U}}^2}{n}. \tag{8}$$

In this section, we report only the final expressions of the estimated linearized variables for the four estimators considered. Detailed calculation rules can be found in Deville [47], and the derivation for each linearized variable estimator is described in detail in the Appendix.

For the estimator $\hat{\Gamma}_1$ (method PGS-CTE1):

$$\hat{U}_i(\hat{\Gamma}_1) = \frac{T_i(Y_i - \hat{p}_{Yi|1}) - (1-T_i)(Y_i - \hat{p}_{0,i})Ax_i}{n^{-1}\sum_{i=1}^{n} T_i\hat{p}_{Yi|1}(1-\hat{p}_{Yi|1})}$$

where $\hat{p}_{Yi|1} = \text{logit}^{-1}(\hat{\Gamma}_1 + \hat{\Psi}_0)$ is the estimated probability of event for subject i obtained from model 3, $x_i = (X_{1,i}, \ldots, X_{K,i})^\intercal$ is the vector of covariates values for subject i, and

$$A = \left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{Yi|1}(1-\hat{p}_{Yi|1})x_i\right]^\intercal \left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i x_i^\intercal\right]^{-1}.$$

For estimator $\hat{\Gamma}_2$ (method PGS-ATT):

$$\hat{U}_i(\hat{\Gamma}_2) = \frac{T_i(Y_i - O^1)}{O^1(1-O^1)\bar{T}} - \frac{T_i\left(\hat{p}_{0,i} - P_0^1\right) + (1-T_i)(Y_i - \hat{p}_{0,i})Bx_i}{P_0^1(1-P_0^1)\bar{T}}$$

where:

$$B = \left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i x_i^{\intercal}\right]^{-1}.$$

For estimator $\hat{\Gamma}_3$ (method PGS-CTE2):

$$\hat{U}_i(\hat{\Gamma}_3) = T_i(Y_i - \hat{p}_{1,i})C_1 x_i - (1-T_i)(Y_i - \hat{p}_{0,i})C_0 x_i$$

where:

$$C_1 = \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{1,i}(1-\hat{p}_{1,i})x_i x_i^{\intercal}\right]^{-1},$$

$$C_0 = \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i x_i^{\intercal}\right]^{-1}.$$

For estimator $\hat{\Gamma}_4$ (method PGS-ATE):

$$\hat{U}_i(\hat{\Gamma}_4) = \frac{\hat{p}_{1,i} + T_i(Y_i - \hat{p}_{1,i})D_1 x_i}{P_1(1-P_1)} - \frac{\hat{p}_{0,i} + (1-T_i)(Y_i - \hat{p}_{0,i})D_0 x_i}{P_0(1-P_0)}$$

where:

$$D_1 = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{p}_{1,i}(1-\hat{p}_{1,i})x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{1,i}(1-\hat{p}_{1,i})x_i x_i^{\intercal}\right]^{-1},$$

$$D_0 = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{p}_{0,i}(1-\hat{p}_{0,i})x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i x_i^{\intercal}\right]^{-1}.$$

5. Simulation setup

5.1. Data-generating process

We used a data-generating process similar to that used in the Austin *et al.* studies to examine different aspects of PPS analysis [2, 34, 48, 49].

We aimed to generate datasets in which the exposure allocation T is drawn from a Bernoulli distribution $T \sim B(p_T)$, with

$$p_T = \text{logit}^{-1}\left(\alpha_0 + \sum_{k=1}^{K}\alpha_k X_k\right), \tag{9}$$

and the binary event E is drawn from a Bernoulli distribution, with $E \sim B(p_E)$, with

$$p_E = \text{logit}^{-1}\left(\beta_0 + \gamma T + \sum_{k=1}^{K}\beta_k X_k + \delta X_2 T\right), \tag{10}$$

where $X_k, k = 1, \ldots, K$, are $K = 9$ normally distributed independent covariates ($X_k \sim N(0;1)$, $k \in 1, \ldots, K$).

5.1.1. Simulation parameters. The true regression coefficients α_k and β_k were set to values presented in Table II, so that variables X_1 to X_3 were associated with both exposure and outcome (true confounders), X_4 to X_6 with exposure only (instrumental variables) and X_7 to X_9 with outcome only (predictors). An interaction term, δ, between exposure status and a true confounder was also included to induce an effect modification (so that the treatment effect depended on the values of X_2 when $\delta \neq 0$).

Coefficients α_0 (true intercept of the exposure model (9)), β_0 (true intercept of the outcome model (10)), and γ (true conditional regression coefficient associated with treatment in model (10)) were selected so that the prevalence of exposed subjects π_T, the event rate π_E and the treatment effect of interest Γ (whether

Table II. Regression coefficients used in the data-generating process. α_k: coefficients of the exposure model, β_k: coefficients of the outcome model.

Parameter	Value	Parameter	Value	Resulting variable
α_1	log(1.5)	β_1	log(1.5)	Weak confounder
α_2	log(2)	β_2	log(2)	Moderate confounder
α_3	log(2.5)	β_3	log(2.5)	Strong confounder
α_4	log(1.5)	β_4	log(1)	Weak instrumental variable
α_5	log(2)	β_5	log(1)	Moderate instrumental variable
α_6	log(2.5)	β_6	log(1)	Strong instrumental variable
α_7	log(1)	β_7	log(1.5)	Weak predictor
α_8	log(1)	β_8	log(2)	Moderate predictor
α_9	log(1)	β_9	log(2.5)	Strong predictor

conditional Γ_{CTE}, marginal ATE Γ_{ATE}, or marginal ATT Γ_{ATT}) were fixed at the desired values in the simulated samples. These three simulation parameters are mutually dependent, and we used an iterative process to determine the values of α_0, β_0, and γ that induce the desired π_T, π_E and Γ. This process is described below.

First, for each $N = 10,000$ subjects, we generated covariates X_1 to X_9, and computed the individual probability of being exposed ($\tilde{p}_{T,i}$) by using Equation (9). The average of these individual probabilities is the expected exposure prevalence $\tilde{\pi}_T = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{T,i}$ in the simulated sample. Similarly, we computed the individual probability of an event ($\tilde{p}_{E,i}$) by using Equation (10), and the corresponding average, the expected event rate $\tilde{\pi}_E = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{E,i}$ in the sample. We also computed the average probability of an event first assuming that all subjects were untreated ($\tilde{\pi}_{E,0} = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{E_0,i}$) and then assuming that all subjects were treated ($\tilde{\pi}_{E,1} = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{E_1,i}$). The difference between the logit of these two average probabilities is the expected ATE, $\tilde{\Gamma}_{ATE} = \text{logit}(\tilde{\pi}_{E,1}) - \text{logit}(\tilde{\pi}_{E,0})$, in the sample. Finally, we computed the same average probabilities weighted by individual probabilities of being exposed ($\tilde{\pi}'_{E,0} = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{T,i}\tilde{p}_{E_0,i}$ and $\tilde{\pi}'_{E,1} = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{T,i}\tilde{p}_{E_1,i}$). The difference between the logit of these two weighted average probabilities is the expected ATT, $\tilde{\Gamma}_{ATT} = \text{logit}(\tilde{\pi}'_{E,1}) - \text{logit}(\tilde{\pi}'_{E,0})$, in the sample.

Using an iterative process, one could successively modify α_0, β_0 and γ until the expected treatment prevalence, the expected event rate and the expected treatment effect are arbitrarily close to the desired value in the simulated cohort. This process was performed by minimizing:

- $(\pi_T - \tilde{\pi}_T)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{ATE} - \tilde{\Gamma}_{ATE})^2$ to obtain the parameters α_0, β_0 and γ that induced the desired exposure prevalence, event rate, and ATE;
- $(\pi_T - \tilde{\pi}_T)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{ATT} - \tilde{\Gamma}_{ATT})^2$ to obtain the parameters α_0, β_0 and γ that induced the desired exposure prevalence, event rate, and ATT;
- $(\pi_T - \tilde{\pi}_T)^2 + (\pi_E - \tilde{\pi}_E)^2$ to obtain the parameters α_0 and β_0 that induced the desired exposure prevalence and event rate for a theoretical CTE $\Gamma_{CTE} = \gamma$ (indeed, the parameter γ used in the outcome model (10) induced the desired conditional treatment effect).

We obtained three separate sets of $(\alpha_0, \beta_0, \gamma)$ parameters (one for the ATE, one for the ATT, and one for the CTE). To increase precision, this minimization process was repeated in 1000 simulated samples of size N to obtain 1000 sets of parameter values, which were averaged to obtain the final parameter values used in the simulation study. Thus, these final parameters were obtained with 10,000,000 simulated subjects for each scenario explored in this study (scenarios are defined in the next section).

With Equation (9), the probability of being treated depends on only subject characteristics. Thus, for a given desired treatment prevalence, all the parameters α_0 obtained with the previously described minimization process were approximatively equal, whatever the event rate or treatment effect.

5.1.2. Simulated datasets generation. All simulated datasets included n=5000 subjects and were generated according to the exposure model (9) and outcome model (10) by using the parameters presented in Table II and derived from the algorithm described in Section 5.1.1. Several scenarios were explored, defined by the following:

- the exposure prevalence: $\pi_T \in \{0.10, 0.20, 0.50\}$;
- the event rate: $\pi_E \in \{0.20, 0.50\}$;

- the treatment effect: $\exp(\Gamma_{CTE}) = \exp(\Gamma_{ATE}) = \exp(\Gamma_{ATT}) \in \{1/1.50, 1/1.25, 1, 1.25, 1.50, 1.75, 2, 2.25, 2.5\}$;
- the effect modification: $\exp(\delta) \in \{1, 2\}$.

For each dataset, three outcome variables were generated, one inducing the desired CTE, one inducing the desired ATE, and one inducing the desired ATT. A total of B=10,000 datasets were generated for each scenario.

5.2. *Evaluation of the treatment effect*

All statistical methods (described in Section 4 and Table I) were applied to each outcome of each simulated dataset, with the exception of scenarios involving an effect modification.

Indeed, in case of non-null interactions, CTE is not equal for all individuals. The CTE depends on the value of the covariates, and because all covariates X_k are continuous, the CTE may differ for each individual. Thus, in cases of non-null interactions, the correct estimation of the CTE for each subject profile needs to account for the interaction terms between exposure status and other covariates (in our simulations, the interaction between X_2 and T). None of the evaluated methods, particularly those intended to estimate the CTE, such as Multivariate logistic regression model (MULTI), PGS-CTE1, and PGS-CTE2, explicitly accounts for any potential interaction between the treatment and other covariates. Therefore, we did not estimate the CTE in scenarios involving a non-null interaction term.

5.3. *Covariates used in each statistical method*

Each statistical method was applied considering four different sets of covariates (V_1 to V_4):

- all variables (V_1): X_1 to X_9;
- only true confounders (V_2): X_1 to X_3;
- all variables associated with the outcome (V_3): X_1 to X_3 (true confounders) and X_7 to X_9 (predictors);
- all variables associated with the exposure (V_4): X_1 to X_6 (true confounders and instrumental variables).

5.4. *Performance criteria*

Results were assessed in terms of the following criteria:

- Bias of the exposure effect estimation: $E(\hat{\Gamma} - \Gamma)$;
- Variability ratio of the exposure effect, defined as follows: $\frac{\frac{1}{B}\sum_{b=1}^{B} \hat{SE}(\hat{\Gamma}_b)}{\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\Gamma}_b - \bar{\hat{\Gamma}})^2}}$, where $\hat{SE}(\hat{\Gamma})$ is the estimated standard error of exposure effect $\hat{\Gamma}$;
- Root mean square error (RMSE): $\sqrt{E(\hat{\Gamma} - \Gamma)^2}$;
- Coverage: proportion of times Γ is included in the 95% confidence interval of Γ estimated from the model.

5.5. *Software*

All simulations and analyses involved use of `R` 3.1.1 (R Foundation for Statistical Computing, Vienna, Austria). Minimization (Section 5.1.1) involved the function `optim` in `R`. Matching procedures involved the `Match` function from the `Matching` package [50]. Robust standard errors were computed with the `svyglm` function from the `survey` package [51].

6. Simulation results

6.1. *Bias in the estimation of the treatment effect*

We first comment on the bias of each analytical method according to each type of treatment effect (CTE, ATE or ATT). For conciseness, we comment on only the results extracted from scenarios with parameter values $\pi_T = 0.20$, $\pi_E = 0.50$ and no effect modification, but results from some other scenarios are described in the remaining sections. According to our simulation parameters, this is a medium scenario, chosen to not favor any particular method. All variables were used to control for confounding. Biases according to the CTE are presented in Figure 1, according to the ATE in Figure 2, and according to the ATT in Figure 3.

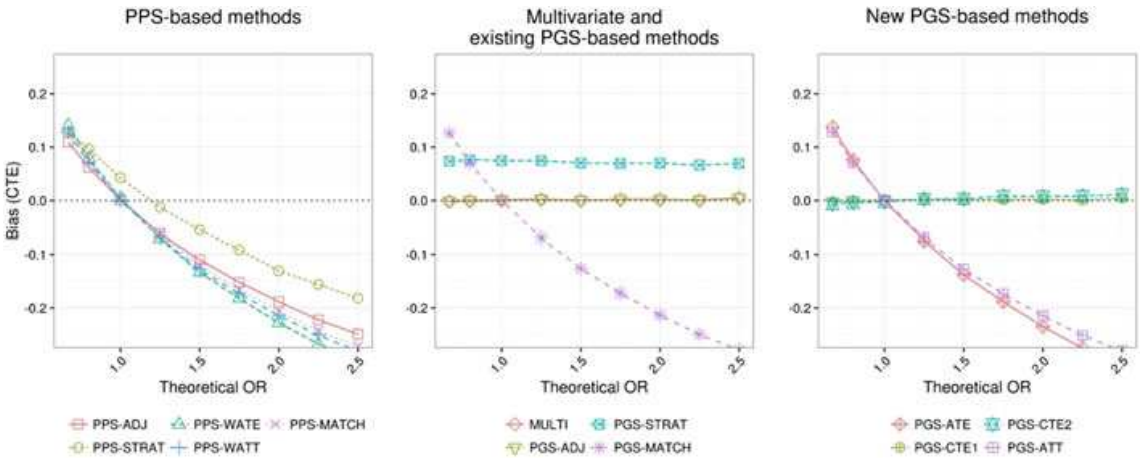


Figure 1. Bias in conditional treatment effect. $\pi_T = 0.20$, $\pi_E = 0.50$, no effect modification ($\exp(\delta) = 1$). All variables (X_1 to X_9) were used to control for confounding. ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment; CTE, conditional treatment effect; MATCH, matching; OR, odd ratio; PGS, prognostic score; PPS, propensity score; RMSE, Root mean square; STRAT, stratification. ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment effect on the treated population; CTE, conditional treatment effect; MATCH, matching; MULTI, Multivariate logistic regression model; PGS, prognostic score; PPS, propensity score; STRAT, stratification.

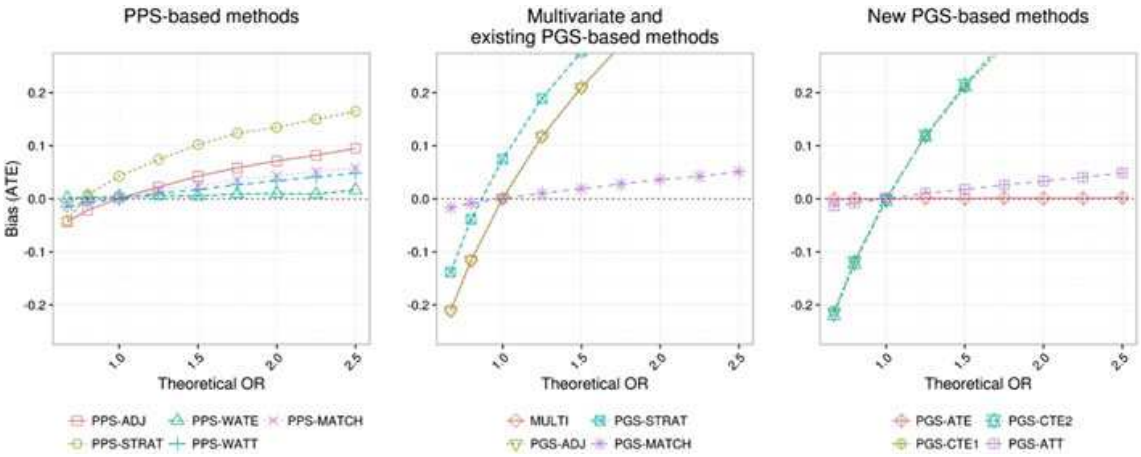


Figure 2. Bias in average treatment effect. $\pi_T = 0.20$, $\pi_E = 0.50$, no effect modification ($\exp(\delta) = 1$). All variables (X_1 to X_9) were used to control for confounding. ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment effect on the treated population; CTE, conditional treatment effect; MATCH, matching; MULTI, Multivariate logistic regression model; PGS, prognostic score; PPS, propensity score; STRAT, stratification.

Multivariate logistic regression model provided a nearly perfect estimation of the CTE, whatever the true underlying OR (Figure 1). All PPS-based methods were biased when CTE was the targeted effect of interest unless there was no treatment effect at all (i.e., true OR was 1), except for PPS-STRAT, which was biased even in this case. PGS-ADJ provided treatment effect estimates nearly identical to multivariate regression. PGS-STRAT estimations of the CTE were biased, but the bias level was constant whatever the true OR. PGS-MATCH provided estimations close to those provided by PPS-based methods. Finally, among the new PGS-based methods, PGS-CTE1 and PGS-CTE2 provided a nearly perfect estimation of CTE, and PGS-ATT and PGS-ATE did not estimate CTE, except when the true OR was 1.

If ATE was the targeted effect (Figure 2), PPS-WATE was the least biased method among PPS-based methods. All existing PGS-based methods provided a biased assessment of the ATE, but PGS-MATCH yielded very similar results to those provided by PPS-MATCH and PPS-WATT. Among the new PGS-based methods evaluated in this study, only PGS-ATE provided unbiased estimates of the ATE whatever

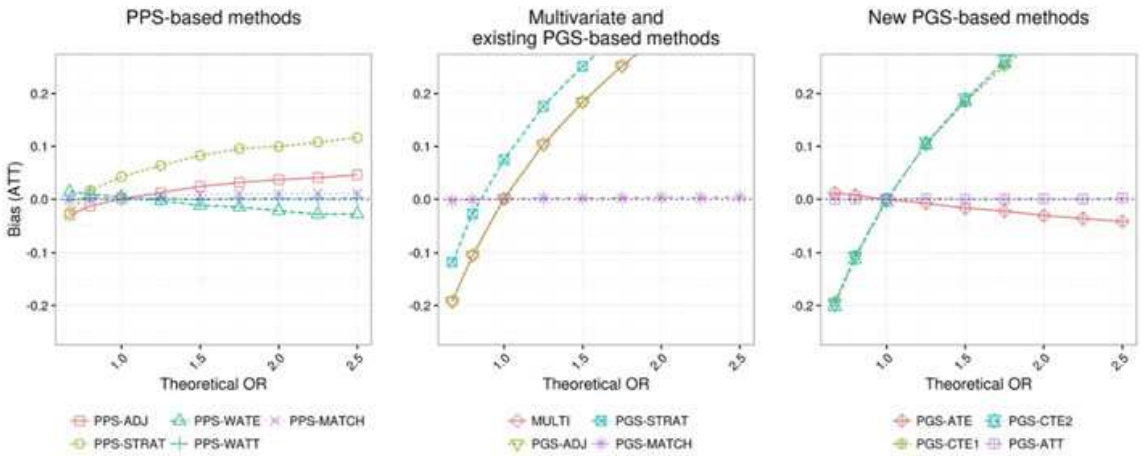


Figure 3. Bias in average treatment effect on the treated. $\pi_T = 0.20$, $\pi_E = 0.50$, no effect modification ($\exp p(\delta) = 1$). All variables (X_1 to X_9) were used to control for confounding. ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment effect on the treated population; CTE, conditional treatment effect; MATCH, matching; MULTI, Multivariate logistic regression model; PGS, prognostic score; PPS, propensity score; STRAT, stratification.

the theoretical OR. PGS-ATT produced estimates close to those of PGS-MATCH, PPS-MATCH and PPS-WATT.

Finally when targeting ATT (Figure 3), PPS-WATT provided a nearly perfect estimation of the ATT, and PPS-MATCH gave a slight bias that increased with the theoretical OR. Among PGS-based methods, only PGS-MATCH and PGS-ATT provided unbiased estimates of the treatment effect. The proportion of the exposed population that could be matched was approximatively the same whatever the theoretical OR, and was 80% for PPS-MATCH and 99% for PGS-MATCH. Thus, the slight bias observed with PPS-MATCH was possibly related to the number of exposed subjects being discarded.

In summary, all the new PGS-based methods provided unbiased estimates of the type of treatment effect for which they were developed. Among the existing PGS methods, PGS-ADJ estimated the CTE, PGS-MATCH estimated the ATT, and PGS-STRAT yielded biased estimates of the CTE. None of the existing PGS methods could estimate the ATE. As was already known, PPS-WATE estimated the ATE and PPS-WATT and PPS-MATCH estimated the ATT. PPS-ADJ and PPS-STRAT did not estimate any of the treatment effects of interest in this study. Because of their weak performance previously reported elsewhere [2, 5] for estimating both conditional and marginal effects, PPS-ADJ and PPS-STRAT were excluded from the remaining results section.

6.2. Effect of the exposure prevalence π_T

Simulation results for varying exposure prevalence are presented in Table III. For clarity, each statistical method is described with the effect type for which they were the most efficient as revealed in Section 6.1: CTE for MULTI, PGS-ADJ, PGS-STRAT, PGS-CTE1 and PGS-CTE2; ATE for PPS-WATE and PGS-ATE; and ATT for PPS-MATCH, PPS-WATT, PGS-MATCH and PGS-ATT. All scenarios presented in this section had an event rate $\pi_E = 0.5$ and no effect modification ($\delta = 0$). As in the previous section, all variables were used to control for confounding.

The upper part of Table III presents the results for a null theoretical treatment effect. When CTE is the targeted treatment effect, MULTI provided good estimates of the treatment effect (bias close to 0) and its variance (variability ratio close to 1) and had a nearly perfect coverage whatever the exposure prevalence. PGS-STRAT resulted in severe convergence issues (because of no event in some strata) when exposure prevalence was low. PGS-ADJ led to results very similar to those with MULTI when $\pi_T = 10\%$, but its variability ratio and coverage deteriorated with increasing exposure prevalence. In contrast, the PGS-CTE1 and PGS-CTE2 performance parameters were little affected by the exposure prevalence and were comparable to that with MULTI. PGS-CTE1 bias and coverage were good whatever the exposure prevalence, whereas PGS-CTE2 seemed less efficient than PGS-CTE1 for the lowest exposure prevalence. However, with $\pi_T = 50\%$, PGS-CTE2 was one of the methods associated with the smallest RMSE (with MULTI method).

Table III. Simulation results for the estimation of CTE, ATE and ATT according to theoretical treatment effect and prevalence of exposure.

	$\pi_T = 10\%$				$\pi_T = 20\%$				$\pi_T = 50\%$			
	Bias	VR	RMSE	Coverage	Bias	VR	RMSE	Coverage	Bias	VR	RMSE	Coverage
$\Gamma = \log(\text{OR}) = \log(1)$												
CTE estimation												
MULTI	0.000	1.003	0.129	0.952	0.000	1.007	0.100	0.954	−0.001	0.993	0.084	0.950
PGS-ADJ	0.000	0.933	0.130	0.934	0.001	0.886	0.102	0.918	0.002	0.750	0.095	0.859
PGS-STRAT	0.004	6.198	0.487	0.899	0.074	0.951	0.144	0.853	0.073	0.813	0.123	0.788
PGS-CTE1	0.000	1.002	0.130	0.950	0.001	1.009	0.102	0.951	−0.002	0.988	0.097	0.946
PGS-CTE2	−0.005	0.993	0.204	0.948	−0.002	0.994	0.130	0.951	−0.002	0.992	0.084	0.949
ATE estimation												
PPS-WATE	0.015	0.909	0.261	0.939	0.004	0.982	0.157	0.955	0.001	1.115	0.082	0.974
PGS-ATE	−0.001	0.996	0.130	0.946	0.000	0.996	0.085	0.949	−0.001	0.993	0.055	0.950
ATT estimation												
PPS-WATT	0.000	1.148	0.108	0.977	0.000	1.131	0.092	0.974	0.000	1.037	0.106	0.964
PPS-MATCH	0.001	1.093	0.125	0.970	0.004	1.110	0.090	0.972	0.009	1.123	0.071	0.968
PGS-ATT	0.000	1.002	0.089	0.950	0.001	1.009	0.069	0.951	−0.001	0.989	0.065	0.944
PGS-MATCH	−0.001	0.964	0.116	0.943	0.001	0.921	0.084	0.929	0.009	0.776	0.071	0.867
$\Gamma = \log(\text{OR}) = \log(2)$												
CTE estimation												
MULTI	0.006	0.998	0.138	0.950	0.002	0.990	0.105	0.948	0.001	0.994	0.085	0.951
PGS-ADJ	0.006	0.937	0.139	0.935	0.002	0.876	0.106	0.915	0.004	0.749	0.095	0.859
PGS-STRAT	0.073	1.807	0.227	0.904	0.070	0.898	0.136	0.866	0.064	0.803	0.117	0.808
PGS-CTE1	0.006	0.999	0.139	0.951	0.002	0.991	0.107	0.948	0.003	0.993	0.096	0.948
PGS-CTE2	0.021	0.977	0.219	0.947	0.007	0.983	0.136	0.948	0.003	0.987	0.086	0.949
ATE estimation												
PPS-WATE	0.029	0.893	0.304	0.922	0.010	0.947	0.183	0.941	0.001	1.101	0.089	0.972
PGS-ATE	0.005	0.968	0.158	0.940	0.001	0.979	0.098	0.946	0.000	0.989	0.059	0.944
ATT estimation												
PPS-WATT	0.005	1.135	0.120	0.975	0.001	1.102	0.099	0.974	0.001	1.028	0.111	0.964
PPS-MATCH	0.010	1.083	0.136	0.968	0.010	1.091	0.097	0.966	0.023	1.120	0.075	0.964
PGS-ATT	0.004	0.998	0.104	0.950	0.001	0.984	0.078	0.948	0.001	0.993	0.068	0.948
PGS-MATCH	0.004	0.966	0.127	0.944	0.004	0.918	0.090	0.928	0.055	0.795	0.090	0.771

All scenarios had an event rate $\pi_E = 0.5$ and no effect modification ($\delta = 0$). All variables were used to control for confounding.

ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment effect on the treated population; CTE, conditional treatment effect; MATCH, matching; MULTI, Multivariate logistic regression model; PGS, prognostic score; PPS, propensity score; RMSE, Root mean square error; STRAT, stratification; VR, Variability ratio.

When focusing on ATE estimation, the PPS-WATE method was biased with $\pi_T = 10\%$. Bias decreased with increasing exposure prevalence, but this method became too conservative with $\pi_T = 50\%$. In contrast, the PGS-ATE method was mostly unbiased whatever the exposure prevalence. The performance parameters were overall better for the PGS-ATE method than the PPS-WATE method (variability ratios closer to 1, smaller RMSE, coverages closer to the nominal value).

For ATT estimation, the two matching methods (PPS-MATCH and PGS-MATCH) had acceptable performance with $\pi_T = 10\%$, but PPS-MATCH was too conservative. If bias increased slightly for these two methods with increasing exposure prevalence, the performance of PGS-MATCH decreased with $\pi_T = 50\%$, with a coverage far from its nominal value. The proportion of the exposed population that could be matched was 90% for PPS-MATCH and 100% for PGS-MATCH with $\pi_T = 10\%$. It was 80% for PPS-MATCH and 99% for PGS-MATCH with $\pi_T = 20\%$ and was 50% for PPS-MATCH and 77% for PGS-MATCH with $\pi_T = 50\%$.

PPS-WATT and PGS-ATT were both mostly unbiased. PPS-WATT tended to overestimate the variance of the treatment effect (and thus was too conservative), whereas PGS-ATT performed well and had coverage closer to the nominal value.

Table IV. Simulation results for the estimation of ATE and ATT according to real treatment effect and the presence of effect modification.

	$\delta = 0$				$\delta = log(2)$			
	Bias	VR	RMSE	Coverage	Bias	VR	RMSE	Coverage
Γ = log(OR) = log(1)								
ATE estimation								
PPS-WATE	0.004	0.982	0.157	0.955	0.005	0.990	0.155	0.958
PGS-ATE	0.000	0.996	0.085	0.949	0.000	0.993	0.081	0.948
ATT estimation								
PPS-WATT	0.000	1.131	0.092	0.974	0.002	1.114	0.093	0.974
PPS-MATCH	0.004	1.110	0.090	0.972	−0.022	1.107	0.093	0.967
PGS-ATT	0.001	1.009	0.069	0.951	0.001	0.999	0.070	0.950
PGS-MATCH	0.001	0.921	0.084	0.929	0.001	0.907	0.085	0.926
Γ = log(OR) = log(2)								
ATE estimation								
PPS-WATE	0.010	0.947	0.183	0.941	0.011	0.963	0.180	0.947
PGS-ATE	0.001	0.979	0.098	0.946	0.001	0.985	0.094	0.948
ATT estimation								
PPS-WATT	0.001	1.102	0.099	0.974	0.001	1.123	0.097	0.975
PPS-MATCH	0.010	1.091	0.097	0.966	−0.016	1.092	0.097	0.964
PGS-ATT	0.001	0.984	0.078	0.948	0.001	0.990	0.077	0.946
PGS-MATCH	0.004	0.918	0.090	0.928	0.002	0.919	0.090	0.927

All scenarios had an exposure prevalence of $\pi_T = 20\%$ and event rate of $\pi_E = 50\%$. All variables were used to control for confounding.
ATE, average treatment effect on the whole population; ATT, average treatment effect on the treated population; MATCH, matching; RMSE, Root mean square.

These results were very similar when considering a non-null theoretical treatment effect (lower part of Table III), with the exception of the two matching methods with $\pi_T = 50\%$: the bias of the two methods increased, particularly when considering PGS-MATCH.

6.3. Effect of the effect modification

Results according to the presence ($\delta = \log(2)$) or absence ($\delta = 0$) of an effect modification are presented in Table IV. In these scenarios, exposure prevalence was $\pi_T = 20\%$, and event rate was $\pi_E = 50\%$. All variables were used to control for confounding. Therefore, results presented in the left part of the table (no effect modification) are the same as those presented in Section (6.2) and are repeated for readability. As stated in Section 5.2, the CTE was not evaluated because it may differ for each subject with $\delta \neq 0$.

For all methods estimating ATE or ATT, no significant change was observed with the introduction of an effect modification.

6.4. Effect of the event rate p_E

Results according to the event rate π_E are presented in Table V. Exposure prevalence was $\pi_T = 20\%$, there was no effect modification ($\delta = 0$), and all variables were used to control for confounding. Results presented in the right part of the table ($\pi_E = 50\%$) are the same as those presented in the two previous sections.

For the PGS-STRAT method, the performance deteriorated the most with event rate decreased to $\pi_E = 20\%$. The method showed severe convergence issues because of the absence of events in some PGS strata.

To a lesser extent, performance was also altered for PGS-CTE2, with an increase in bias and RMSE (but little impact on coverage). This finding was also true for the two matching methods (PPS-MATCH and PGS-MATCH) but mainly with a non-null treatment effect. All other methods were little affected by the decrease in event rate to $\pi_E = 20\%$.

6.5. Effect of the set of covariates used to control for confounding

Results are presented in Table VI for scenarios with exposure prevalence $\pi_T = 20\%$, event rate $\pi_E = 50\%$, and no effect modification ($\delta = 0$).

Table V. Simulation results for the estimation of CTE, ATE and ATT according to real treatment effect and the event rate.

	$\pi_E = 20\%$				$\pi_E = 50\%$			
	Bias	VR	RMSE	Coverage	Bias	VR	RMSE	Coverage
$\Gamma = \log(OR) = \log(1)$								
CTE estimation								
MULTI	0.000	0.997	0.112	0.949	0.000	1.007	0.100	0.954
PGS-ADJ	0.001	0.849	0.115	0.901	0.001	0.886	0.102	0.918
PGS-STRAT	−0.673	46.284	1.543	0.920	0.074	0.951	0.144	0.853
PGS-CTE1	0.000	1.000	0.116	0.950	0.001	1.009	0.102	0.951
PGS-CTE2	−0.020	0.986	0.179	0.948	−0.002	0.994	0.130	0.951
ATE estimation								
PPS-WATE	0.000	1.005	0.150	0.963	0.004	0.982	0.157	0.955
PGS-ATE	0.000	0.988	0.093	0.948	0.000	0.996	0.085	0.949
ATT estimation								
PPS-WATT	0.003	1.059	0.116	0.965	0.000	1.131	0.092	0.974
PPS-MATCH	0.005	1.092	0.102	0.969	0.004	1.110	0.090	0.972
PGS-ATT	0.000	1.000	0.080	0.949	0.001	1.009	0.069	0.951
PGS-MATCH	0.002	0.888	0.093	0.917	0.001	0.921	0.084	0.929
$\Gamma = \log(OR) = \log(2)$								
CTE estimation								
MULTI	0.002	1.000	0.109	0.949	0.002	0.990	0.105	0.948
PGS-ADJ	0.003	0.829	0.114	0.895	0.002	0.876	0.106	0.915
PGS-STRAT	−0.250	27.643	1.029	0.918	0.070	0.898	0.136	0.866
PGS-CTE1	0.003	1.000	0.114	0.949	0.002	0.991	0.107	0.948
PGS-CTE2	−0.012	0.995	0.162	0.950	0.007	0.983	0.136	0.948
ATE estimation								
PPS-WATE	0.000	1.012	0.147	0.965	0.010	0.947	0.183	0.941
PGS-ATE	−0.001	1.002	0.087	0.946	0.001	0.979	0.098	0.946
ATT estimation								
PPS-WATT	0.004	1.057	0.119	0.963	0.001	1.102	0.099	0.974
PPS-MATCH	0.024	1.110	0.101	0.965	0.010	1.091	0.097	0.966
PGS-ATT	0.001	1.002	0.080	0.950	0.001	0.984	0.078	0.948
PGS-MATCH	0.010	0.886	0.093	0.916	0.004	0.918	0.090	0.928

All scenarios had an exposure prevalence of $\pi_T = 20\%$ and no effect modification ($\delta = 0$). All variables were used to control for confounding.

ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment; CTE, conditional treatment effect; MATCH, matching; MULTI, Multivariate logistic regression model; PGS, prognostic score; PPS, propensity score; RMSE, Root mean square; STRAT, stratification.

With a null treatment effect (upper part of the table) and when the targeted treatment effect was CTE (i.e., for MULTI, PGS-ADJ, PGS-STRAT, PGS-CTE1, and PGS-CTE2), the best set of covariates in terms of coverage and RMSE was that including only the true confounding factors. When targeting ATE or ATT, the best RMSE was obtained when controlling for all variables associated with the outcome and no instrumental variables, but coverage was closer to the nominal value when accounting for only true confounders.

In contrast, with a non-null treatment effect (lower part of the table), accounting for only true confounders to estimate CTE led to very poor performance. Overall, the set of all variables associated with the outcome was the best choice, in terms of RMSE, for all methods. With this set of covariates, the lowest RMSE was obtained with MULTI, PGS-ADJ and PGS-CTE1 for the CTE estimation (with suboptimal coverage for PGS-ADJ), PGS-ATE for the ATE estimation, and PGS-ATT and PPS-WATT for the ATT estimation (with suboptimal coverage for PPS-WATT).

7. Case study

7.1. Data source

Data were obtained from a (yet-unpublished) study evaluating the efficacy of personalized support of asthmatic patients, organized by the French National Health Insurance Fund (Caisse Nationale de l'Assurance

Table VI. Simulation results for the estimation of CTE, ATE and ATT according to real treatment effect and the set of covariates used to adjust for confounding: all variables (V_1), true confounders (V_2), variables associated with outcome (V_3) or variables associated with exposure (V_4).

	V_1		V_2		V_3		V_4	
	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
$\Gamma = \log(OR) = \log(1)$								
CTE estimation								
MULTI	0.100	0.954	0.081	0.952	0.092	0.953	0.088	0.951
PGS-ADJ	0.102	0.918	0.081	0.953	0.093	0.943	0.089	0.928
PGS-STRAT	0.144	0.853	0.129	0.876	0.139	0.869	0.138	0.852
PGS-CTE1	0.102	0.951	0.082	0.953	0.093	0.955	0.090	0.953
PGS-CTE2	0.130	0.951	0.092	0.952	0.105	0.949	0.115	0.951
ATE estimation								
PPS-WATE	0.157	0.955	0.088	0.961	0.082	0.973	0.159	0.949
PGS-ATE	0.085	0.949	0.076	0.951	0.068	0.949	0.095	0.949
ATT estimation								
PPS-WATT	0.092	0.974	0.073	0.969	0.067	0.982	0.098	0.964
PPS-MATCH	0.090	0.972	0.087	0.954	0.082	0.966	0.098	0.956
PGS-ATT	0.069	0.951	0.071	0.954	0.064	0.954	0.077	0.953
PGS-MATCH	0.084	0.929	0.087	0.955	0.080	0.943	0.093	0.936
$\Gamma = \log(OR) = \log(2)$								
CTE estimation								
MULTI	0.105	0.948	0.168	0.604	0.097	0.948	0.170	0.655
PGS-ADJ	0.106	0.915	0.167	0.606	0.098	0.938	0.170	0.602
PGS-STRAT	0.136	0.866	0.121	0.896	0.130	0.886	0.125	0.888
PGS-CTE1	0.107	0.948	0.167	0.613	0.098	0.949	0.170	0.671
PGS-CTE2	0.136	0.948	0.175	0.632	0.107	0.950	0.189	0.752
ATE estimation								
PPS-WATE	0.183	0.941	0.101	0.954	0.096	0.962	0.185	0.939
PGS-ATE	0.098	0.946	0.086	0.949	0.077	0.948	0.110	0.947
ATT estimation								
PPS-WATT	0.099	0.974	0.082	0.962	0.075	0.978	0.106	0.962
PPS-MATCH	0.097	0.966	0.094	0.951	0.089	0.963	0.103	0.953
PGS-ATT	0.078	0.948	0.080	0.949	0.073	0.950	0.086	0.947
PGS-MATCH	0.090	0.928	0.094	0.951	0.086	0.940	0.100	0.932

All scenarios had an exposure prevalence of $\pi_T = 20\%$, an event rate of $\pi_E = 50\%$ and no effect modification ($\delta = 0$). ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment; CTE, conditional treatment effect; MATCH, matching; MULTI, Multivariate logistic regression model; PGS, prognostic score; PPS, propensity score; RMSE, Root mean square; STRAT, stratification.

Maladie des Travailleurs Salariés, CNAMTS). We analyzed subjects from the control group only, who did not receive the personalized intervention. Data were extracted from the French health insurance database (SNIIRAM), which is the national claims database, linked to the French hospital discharge database (PMSI) containing individual records of all hospital stays.

Eligible subjects were 18–39 years old defined as having asthma because they filled at least four prescriptions for asthma-related medications during the year before inclusion. Demographics, reimbursements for health care expenditure, costly long-term disease status, complementary universal health insurance (CMUc) status (indicates low socioeconomic level), and the general practitioner's city zip code were available for 31,332 subjects.

In this case study, the objective was to assess the impact of the asthma drug ratio (exposure of interest) on the risk of asthma exacerbation. Asthma drug ratio is the proportion of reimbursed units of inhaled corticosteroids (ICSs) to overall number of reimbursed respiratory medication units [52]. ICSs are the cornerstone therapy in persistent asthma to prevent exacerbations. Previous studies of persistent asthma patients showed significantly fewer asthma outcomes among patients with high ratios.

Two groups were defined by ICS ratio: $< 70\%$ (low ICS ratio group) and $\geqslant 70\%$ (high ICS ratio group). The outcome was the occurrence of asthma exacerbation within 1 year, defined as a filled prescription for oral corticosteroids within 7 days after a medical consultation (with a general practitioner or a pneumonologist) or a hospitalization for asthma exacerbation.

Table VII. Characteristics of patients in the case study.

	ICS ratio < 70 $N = 18850$	*ICS ratio* $\geqslant 70$ $N = 12482$	Overall $N = 31332$
Age	32 [27–36]	32 [27–36]	32 [27–36]
Male	8369 (44)	5019 (40)	13,388 (43)
Long term disease status for asthma			
	2155 (11)	800 (6)	2955 (9)
Complementary Universal Health Insurance status			
	4602 (24)	1849 (15)	6451 (21)
Pneumonologist consultation in the past year			
	2860 (15)	2394 (19)	5254 (17)
Hospitalization for asthma in the past year			
	208 (1)	21 (0)	229 (1)
Number of asthma exacerbation in the past year			
0	9020 (48)	7139 (57)	16,159 (52)
1	4382 (23)	2735 (22)	7117 (23)
2–3	3677 (20)	1954 (16)	5631 (18)
$\geqslant 4$	1771 (9)	654 (5)	2425 (8)
Quintiles of the social deprivation index			
first	2270 (12)	1776 (14)	4046 (13)
second	3307 (18)	2337 (19)	5644 (18)
third	3764 (20)	2561 (20)	6325 (20)
fourth	3552 (19)	2266 (18)	5818 (19)
fifth	5957 (32)	3542 (28)	9499 (30)
Physician's type of area			
Rural	1991 (11)	1430 (11)	3421 (11)
Urban	16,859 (89)	11,052 (89)	27,911 (89)
Medical density in the municipality			
	206 [129–296]	199 [126–296]	204 [127–296]
Asthma exacerbation within one year (primary outcome)			
	8615 (46)	5005 (40)	13620 (43)

Data are median [Q25-Q75] or *N* (%).
ICSs, inhaled corticosteroids.

Table VIII. Estimated ORs in the case study.

Method	log(OR)	Standard error
Unadjusted	−0.2291	0.0234
CTE estimation		
MULTI	−0.0972	0.0252
PGS-ADJ	−0.0968	0.0248
PGS-STRAT	−0.1201	0.0248
PGS-CTE1	−0.0986	0.0253
PGS-CTE2	−0.1046	0.0255
ATE estimation		
PPS-WATE	−0.0877	0.0242
PGS-ATE	−0.0884	0.0226
ATT estimation		
PPS-WATT	−0.0905	0.0241
PPS-MATCH	−0.0840	0.0260
PGS-ATT	−0.0896	0.0230
PGS-MATCH	−0.0758	0.0244

ADJ, adjustment; ATE, average treatment effect on the whole population; ATT, average treatment; CTE, conditional treatment effect; MATCH, matching; OR, odd ratio; PGS, prognostic score; PPS, propensity score; RMSE, Root mean square; STRAT, stratification.

7.2. Results

All the statistical methods described in Section 4 were applied. Approximatively 40% of the subjects were part of the high ICS ratio group. Variables accounted for in the multivariate regression, PPS or PGS analysis are described in Table VII. The estimated log ORs are reported in Table VIII. ORs had qualitatively similar values, which indicates the protective effect of a high ICS ratio on asthma exacerbation.

As in the simulation study, standard errors of the treatment effects estimated with MULTI, PGS-CTE1, and PGS-CTE2 were slightly higher that those estimated by PGS-ADJ and PGS-STRAT, so the latter may overestimate the statistical significance of the association. According to simulation results, the same was expected when comparing PGS-ATT and PGS-MATCH methods, but here, PGS-ATT had a lower standard error. The standard error was lower for PGS-ATE than PPS-WATE and was lower for PGS-ATT than PPS-WATT. These results were consistent with the overestimation of the treatment effect variability for these PPS-based methods observed in the simulation study.

8. Discussion

Since their introduction by Hansen in 2008, PGS methods have been a subject of growing interest in the medical and statistical literature, but their properties and performance have been much less studied than those of PPS methods. Moreover, the PGS is presented as analogous to the PPS, but this assertion relies mostly on simulation studies focusing on a collapsible measure of treatment effect [11, 40] or the estimation of a conditional treatment effect [41, 45]. We found only one simulation study assessing PGS methods for evaluating the marginal effect with a non-collapsible measure (the OR) [12]. In the latter, only matching on the PGS was evaluated for estimating the average treatment effect on the treated population, and the theoretical treatment effect was null in most reported scenarios (in this setting, the marginal and conditional ORs coincide [2]). Overall, the type of treatment effect estimated by the existing PGS-based methods was poorly discussed in the literature. In the present article, we have demonstrated that (i) none of the three existing PGS-based methods (adjustment, stratification and matching on the PGS) allows for assessing the ATE; (ii) matching on the PGS is the only existing method providing a marginal estimate of the treatment effect (ATT); (iii) adjustment and stratification on the PGS led to a conditional estimate of the treatment effect (which was biased for stratification); and (iv) the three existing methods have unacceptable coverage when the proportion of exposed subjects in the study sample is large. Conversely, the PPS allows for estimating marginal effects (ATE or ATT depending on the method used) with acceptable performance, and is unable to estimate the CTE even with adjustment or stratification on the PPS methods (as was observed in Austin *et al.* [5]). Therefore, defining the PGS as analogous to the PPS is misleading for applied researchers, because the only treatment effect that could be estimated by both frameworks with the available methods is the ATT. In the present article, we proposed four new PGS-based methods (and their variance estimators) especially developed for estimating CTE, ATE or ATT and explored their performance for estimating the OR associated with a binary exposure in various settings.

We studied three types of treatment effect: CTE, ATE and ATT. Concerning the CTE, stratification on the PGS was biased, but this finding may be due to an insufficient number of strata, because we used only five strata to control for confounding. We chose this number of strata by analogy with stratification on the PPS: Cochran [30] showed that stratifying on quintiles of the PPS eliminates approximately 90% of the bias because of measured confounders. This situation may not be the case for stratification on the PGS and needs further investigation. However, adjustment on the PGS led to less biased estimations, although coverage of both stratification and adjustment on the PGS was highly affected by exposure prevalence. We believe that this issue is due to the fact that these methods consider the PGS as a known quantity and not an estimation of an unknown quantity. When exposure prevalence is low, the variance of $\hat{\Psi}_0$ (i.e., the predictions from a model fitted within unexposed subjects only) is negligible as compared with the 'model-based' variance of the treatment effect, and considering the PGS as a known quantity could be a valid approximation. However, this is no longer true when the exposure prevalence is higher. PGS variability is no longer negligible, which leads to an underestimation of the treatment effect variability and decreased coverage. Arbogast *et al.* did not find a coverage issue with adjustment on a PGS estimated in unexposed subjects, except when covariates were strongly associated with exposure and moderately with the outcome. However, all scenarios evaluated in Arbogast *et al.* implied a quite small proportion of exposed subjects (10%), and the coverage issue was very modest in our study at this level of exposure prevalence. Schmidt *et al.* [41] evaluated adjustment on PGS in simulated samples with 50% exposure,

but the PGS was derived on a separate training data set of 5000 subjects, whereas treatment effect estimation involved a test sample of 400 subjects. In this situation, the variance of the PGS should also be negligible. To confirm this assumption, we reran some simulations to evaluate the PGS-ADJ method when the exposure was fixed to 50% but deriving the PGS model in a independent training sample. Use of an independent sample of size $n = 5000$ (i.e., the same size as the sample with which we estimated the effect of the treatment in our study) led to a coverage level below the nominal level, as in Table III. As expected, increasing the sample size of the independent sample to 50,000 led to a correct coverage level.

We proposed two new approaches to estimate the CTE, with a variance estimator accounting for the estimated PGS variability. These methods had acceptable coverage whatever the exposure prevalence. The first method used the PGS derived from the unexposed population only and had better performance than the second method (which used both PGSs derived from the unexposed and exposed populations) with exposure at 10%. In our simulations, these methods were more efficient than existing PGS-based methods and often as efficient as traditional multivariate regression. Nevertheless, our study was limited to quite common exposure prevalence (10%–50%) and event rates (20% and 50%). The comparison of the two new PGS methods evaluating CTE needs further evaluation in other (more extreme) situations. Also, one can wonder why any method other than standard multivariate regression would be used to estimate the CTE if the performance is the same. In fact, we believe that if CTE is measure of interest, multivariate regression should be considered the gold standard. Nevertheless, three arguments in favor of PGS-based methods may still be considered in some situations. First, using a two-step procedure for estimating the treatment effect allows for separating the consideration of confounding factors from the estimation of the treatment effect on the outcome. A similar argument was also reported by Austin concerning PPS-based methods [53]. Second, a unique PGS model could be used to study several different exposures [9], whereas multivariate regression needs to refit the entire model for each exposure. Third, Glynn *et al.* [7] proposed the use of an independent set of unexposed subjects to estimate the prognostic score by using data from a period before the study period or from a separate population. The rational is that (i) unexposed subjects may be more representative (the population before the study period, e.g., before the marketing of a new drug, is more likely to include untreated subjects representative of the future treated subjects, that is those in the study sample after the drug marketing) and (ii) if the external sample is large, estimation of the prognostic model may be more reliable in settings with relatively few outcomes. Of course, the second and third arguments are relevant for PGS-based methods that use Ψ_0 only, such as PGS-CTE1.

Concerning the ATE, we proposed a new PGS-based method involving the PGS derived from the exposed population in addition to the unexposed population. As stated previously, existing PGS-based methods do not allow for assessing the ATE. When comparing the performance of our new method with the PPS-WATE method, we observed that the PPS-based method was biased for the lower exposure ($\pi_T = 10\%$). This result is consistent with previous findings from Hajage *et al.* [37], who evaluated the use of some PPS-based methods in cases of rare exposure for estimating a marginal hazard ratio. When exposure was frequent, bias was acceptable, but the variance of the estimation was overestimated and coverage became too high. This situation was also observed in some other simulation studies related to PPS [20, 26]. Austin *et al.* suggested that the situation could be due to the analysis being performed as though the PPS was a known quantity rather than an estimate of an unknown quantity [49]. This is the same issue that we described for existing PGS-based methods but with opposite consequences on the coverage. The variance estimator of our new PGS method (PGS-ATE) accounts for the variability associated with the PGS estimation and led to coverage rates closer to the nominal value.

Our method for estimating ATE by using the PGS framework is close to that proposed by Austin *et al.* [54] and Localio *et al.* [55]. They both proposed to compute the marginal odds ratio by averaging the predicted probability given that studied subjects were treated, and the same subjects were untreated. The model used to derive the predicted probabilities was a logistic regression model fitted in the entire cohort including the treatment as a covariate, whereas our method follows Hansen's recommendation to fit two independent models (one in the exposed group, one in the unexposed group) [6]. This method accounts for all effect modifiers [6] by using both a prognostic score for the potential outcome under treatment and a prognostic score for the potential outcome under control [11], without explicitly including interaction terms. Our method is also close to the 'model standardization approach' [56, 57]. This method was developed by Rosenblum and van der Laan in the context of a randomized trial to obtain more precise estimates of marginal treatment effects, with information for variables collected at baseline. As in our method, the authors used the predicted probabilities derived from two independent MULTIs, one fitted in the experimental arm and one fitted in the control arm. They derived a variance estimator that used the fact that the probability of being treated is known and equals 1/2 for all subjects (by randomization). Our

study showed that this approach can also be used in the context of an observational (non-randomized) study and proposed a variance estimator that does not require the probability of being treated to be fixed by the study design.

Finally concerning the ATT, our results showed that matching on the PPS can be more efficient than matching on the PGS. When the exposure was low ($\pi_T = 10\%$), bias was limited for the two methods, and PGS-matching had a coverage rate close to 95%, whereas PPS-matching was too conservative. However, when the treatment effect was non-null, bias increased more for PGS-matching than PPS-matching with increasing exposure. Moreover, the PGS-matching coverage became unacceptable with $\pi_T = 50\%$, even if the proportion of matched exposed subjects was always larger with PGS-matching than PPS-matching. This issue was not reported by Leacy and Stuart [11], who evaluated (among other matching methods) the performance of matching on the estimated PGS. However, Leacy and Stuart did not explore the effect of the exposure prevalence (fixed to 20% in all evaluated scenarios) and assessed the effect of a treatment on a continuous outcome, whereas our study focused on a binary outcome. The Wyss *et al.* study [12] focused on estimating the OR associated with a binary exposure on a binary outcome and did not report any similar performance issue with PGS-matching. The exposure prevalence was fixed to 30% in all scenarios, which is greater than the prevalence for which we started to observe a coverage rate below the nominal value (20%). The Wyss *et al.* study concluded that PGS-matching can improve precision of the effect estimates by allowing to match a larger proportion of the treated population. However, most of the scenarios reported in the Wyss *et al.* study had a null treatment effect, and we also reported an acceptable bias level of the PGS-matching method in this situation. Moreover, the authors did not report the variability ratios and type I error rates observed in their simulations. In our simulation results, even if we also noticed that a larger number of treated subjects could be matched by using the PGS than PPS, PGS-matching tended to underestimate the variance of the treatment effect with exposure $\geqslant 20\%$ (thus precision appeared 'improved') and to provide more biased estimates than PPS-matching when exposure prevalence was 50%.

To our knowledge, PGS-matching was the only PGS-based method reported in the literature for estimating the ATT. The performance of our new method (PGS-ATT) and its associated variance estimator were a little affected by the theoretical treatment effect and the treatment prevalence. Compared with PGS-matching, PPS-matching, and PPS-weighting using ATT weights, our method had the lowest RMSE and coverage rates closer to the nominal value.

Our study confirmed the results of Arbogast *et al.* [40], who stated that PGS analysis is less efficient when using covariates associated with exposure and not with the outcome. Also, our article is not the first to report a coverage issue with some PGS-based methods. A recent simulation study by Xu *et al.* [58] compared matching and stratification (but not adjustment) on the PPS or the PGS and multivariate logistic regression evaluating the safety of emerging treatment (measured with an OR). Analysis of the simulated samples was sequential with the Lan-DeMets approach [59] to mimic the post-market surveillance of a new drug. Like us, the authors observed that PGS-based methods had suboptimal type I error rates when the treatment was common. Nevertheless, the Xu *et al.* study is not directly comparable with our study. First, the PGS model was developed in the entire population (and not only the unexposed subgroup), including treatment and other covariates in the prognostic model and then taking as PGS the part of the linear predictor that is free of the treatment variable. Thus, the authors estimated the Miettinen's score, which could exaggerate the statistical significance of treatment effect estimates [43, 60]. Furthermore, they compared their estimations with a theoretical conditional treatment effect, so the results concerning the matching methods (which estimates ATT) are difficult to interpret, because the OR is not collapsible. In some sensitivity analyses, they observed that increasing the size of the entire sample, leaving the number of exposed subjects unchanged (i.e., reducing the prevalence of exposed subjects in the sample), tends to reduce the type I error rates associated with PGS-based methods. Their interpretation was that 'the inflation of type I error rate was partially due to the unavailability of optimal comparators for matching or stratification'. Because the percentage of matched exposed subjects was not reported, we cannot formally contest this interpretation, but we disagree with the authors for the following reasons. First, both our simulations and Wyss *et al.* [12] showed that matching on the PGS allows for matching a larger proportion of the treated population as compared with matching on the PPS. Second, Xu *et al.* did not report the variance of their estimations. Here, we showed that using a variance estimator that accounts for the variability of PGS estimation led to acceptable type I error rates even when treatment is common in the population. Thus, we believe that Xu *et al.* would have observed an underestimation of the treatment effect variance in their study if it was evaluated.

The simulation results we report are limited to a data-generating process focused on only a binary outcome and the estimation of the OR with a limited number of methods in a limited number of scenarios, which are the directions for future research. First, the PGS can be estimated for some other types of outcome variables (e.g., continuous, categorical, or ordinal) by using generalized linear regression modeling techniques, and the performance of the existing and new approaches should be evaluated and compared in each situation. Even with a binary outcome, the performance of PGS-based methods has never been explored for estimating differences in proportions or relative risks as was done for PPS-based methods [19, 26]. These measures of the treatment effect have the advantage of being collapsible measures (unlike the OR). The use of PGS to evaluate a time-to-event outcome has never been reported except in one empirical example [12], and this avenue warrants a specific exploration. Second, the performance of the new PGS-based methods should be evaluated in more extreme situations such as very low number of events. In our study, the impact of decreasing the event rate from 50% to 20% was limited for all methods except for stratification on the PGS. However, with an even lower event rate, the advantages of PGS-based methods over PPS-based methods we describe could disappear or be reversed. Third, the use of a 'same-sample' estimation of PGS may be vulnerable to overfitting [6] because only a subgroup of the subjects contributes to the estimation. This issue is even more acute for PGS-based methods using both unexposed and exposed estimated prognostic scores: in real-life settings, the number of exposed and unexposed subjects is likely unbalanced, and one of the PGS models is then evaluated with a smaller number of subjects. The use of regression modeling techniques, which are less subject to overfitting, and the generalization of our new PGS-based methods (in particular their variance estimators) to the use of an independent (larger) sample for fitting prognostic models needs further investigation. Fourth, none of the existing or new PGS-based methods estimating CTE are designed to account for a treatment effect heterogeneity. Developing new methods that could estimate the CTE for each subject profile in case of heterogeneity could be interesting. Fifth, we used unstabilized weights for ATE estimation with the PPS-weighting method, which could increase the variability of the estimated treatment effect because of excessive weight values. The performance of the PPS-WATE method could have been improved by the use of stabilized and/or trimmed weights [15]. Finally, a limitation of our study is the variance estimator used with PPS-weighting methods. Lunceford and Davidian developed a variance formula [61] that is suitable for only estimating ATE by using a risk difference (not the OR). Williamson *et al.* [62] developed a variance formula suitable for estimating the ATE by using the OR but not for the ATT. Moreover, to our knowledge, the latter has never been evaluated in the context of observational studies (but in the context of randomized trials). For these reasons and to use the same method for ATE and ATT estimations, we chose to use the variance estimation taken from a weighted regression model of the outcome on treatment with a robust variance estimator (as in Forbes and Shortreed [27] or Pirracchio *et al.* [20]). This method does not account for the fact that PPS is estimated from the data, which may lead to too-conservative confidence intervals [63]. This limitation also applies to PPS-adjustment [64], PPS-stratification [63] and PPS-matching [24] methods for which variance formulae have been recently developed. Bootstrap estimation of the variance has also been proposed [15, 65]. Thus, further evaluations are needed on this topic.

In conclusion, when evaluating a treatment effect with an observational study, an applied researcher should carefully choose the type of estimation that answers the clinical objectives. The PGS framework provides a good alternative to the PPS framework. The process involves modeling the prognostics for subjects, which can rely on a richer literature than modeling the propensity of being treated (particularly for a recently marketed drug, for which identifying confounders that predict treatment might be difficult) and can benefit more directly from the predictive scoring systems developed and available for many diseases. Nevertheless, the different methods accounting for the estimated PGS have received little attention. We showed that existing PGS-based methods do not allow for estimating the ATE and feature unacceptable performance issues when the proportion of exposed subjects is large. We propose new PGS-based methods estimating conditional or marginal treatment effects (ATE or ATT), with good performance in different scenarios defined by the exposure prevalence, theoretical treatment effect, presence of effect modification, and event rate. When estimating CTE, the new PGS-based methods often performed as well as multivariate logistic regression. When estimating marginal effects, the PPS-based methods were too conservative, whereas the new PGS-based methods performed better particularly with low exposure and had coverage closer to the nominal value. Therefore, these new methods may be recommended to account for PGS in observational studies when estimating the OR associated with a binary exposure and should be extended to and tested in other settings.

Appendix A Computation of the estimated linearized variables $\hat{U}$

A. Preliminary results

The first step of all PGS-based methods consists of estimating the prognostic of each subject under the unexposed and/or the exposed condition. For the subjects with $T_i = 0$, a logistic regression model leads to:

$$\hat{p}_{0,i} = \frac{\exp\left(x_i^\intercal \hat{\beta}_0\right)}{1 + \exp\left(x_i^\intercal \hat{\beta}_0\right)} \text{ which estimates } P(Y_i = 1 | T_i = 0, X_i) = p_{0,i},$$

with $\hat{\beta}_0 = (\hat{\beta}_{0,0}, \ldots, \hat{\beta}_{0,K})^\intercal$. We also note $\beta = (\beta_0, \ldots, \beta_K)^\intercal$. Using a Taylor expansion, we obtain:

$$\begin{aligned}
\hat{p}_{0,i} &= \left[1 + \exp\left(-x_i^\intercal \hat{\beta}_0\right)\right]^{-1} \\
&= \left[1 + \exp\left(-x_i^\intercal \beta\right) \exp\left\{-x_i^\intercal \left(\hat{\beta}_0 - \beta\right)\right\}\right]^{-1} \\
&\approx \left[1 + \exp\left(-x_i^\intercal \beta\right) \left\{1 - x_i^\intercal \left(\hat{\beta}_0 - \beta\right)\right\}\right]^{-1} \\
&= \left[1 + \exp\left(-x_i^\intercal \beta\right)\right]^{-1} \left[1 - \frac{x_i^\intercal \exp\left(-x_i^\intercal \beta\right) \left(\hat{\beta}_0 - \beta\right)}{1 + \exp\left(-x_i^\intercal \beta\right)}\right]^{-1} \\
&\approx p_{0,i} \left[1 + \frac{x_i^\intercal \exp\left(-x_i^\intercal \beta\right) \left(\hat{\beta}_0 - \beta\right)}{1 + \exp\left(-x_i^\intercal \beta\right)}\right] \\
&= p_{0,i} \left[1 + (1 - p_{0,i}) x_i^\intercal (\hat{\beta}_0 - \beta)\right] \\
&\Rightarrow \hat{p}_{0,i} - p_{0,i} \approx \hat{p}_{0,i}(1 - \hat{p}_{0,i}) x_i^\intercal (\hat{\beta}_0 - \beta)
\end{aligned} \tag{A.1}$$

The normal equation for the logistic regression, $\sum_{i=1}^n (1 - T_i)(\hat{p}_{0,i} - Y_i) x_i = 0$, can be written as follows:

$$\sum_{i=1}^n (1 - T_i)(\hat{p}_{0,i} - p_{0,i}) x_i = \sum_{i=1}^n (1 - T_i)(Y_i - p_{0,i}) x_i \tag{A.2}$$

By plugging (A.1) into (A.2), we obtain

$$\hat{\beta}_0 - \beta \approx \left[\sum_{i=1}^n (1 - T_i) \hat{p}_{0,i}(1 - \hat{p}_{0,i}) x_i x_i^\intercal\right]^{-1} \sum_{i=1}^n (1 - T_i)(Y_i - p_{0,i}) x_i. \tag{A.3}$$

By plugging (A.3) into (A.1), we obtain the approximation:

$$\boxed{\hat{p}_{0,i} - p_{0,i} \approx \left[\hat{p}_{0,i}(1 - \hat{p}_{0,i}) x_i^\intercal\right] \left[\sum_{j=1}^n (1 - T_j) \hat{p}_{0,j}(1 - \hat{p}_{0,j}) x_j x_j^\intercal\right]^{-1} \sum_{j=1}^n (1 - T_j)(Y_j - p_{0,j}) x_j.} \tag{A.4}$$

For the subjects with $T_i = 1$, a logistic regression model leads to

$$\hat{p}_{1,i} = \frac{\exp\left(x_i^\intercal \hat{\beta}_1\right)}{1 + \exp\left(x_i^\intercal \hat{\beta}_1\right)} \text{ which estimates } P(Y_i = 1 | T_i = 1, X_i) = p_{1,i},$$

where we note $\hat{\beta}_1 = (\hat{\beta}_{1,0}, \ldots, \hat{\beta}_{1,K})^\intercal$. By following a similar reasoning, we obtain the approximation:

$$\boxed{\hat{p}_{1,i} - p_{1,i} \approx \left[\hat{p}_{1,i}(1 - \hat{p}_{1,i}) x_i^\intercal\right] \left[\sum_{j=1}^n T_j \hat{p}_{1,j}(1 - \hat{p}_{1,j}) x_j x_j^\intercal\right]^{-1} \sum_{j=1}^n T_j (Y_j - p_{j,i}) x_j.} \tag{A.5}$$

B. Derivation of $\hat{U}_i(\hat{\Gamma}_1)$

The first estimator is obtained by fitting a logistic regression model on the subjects with $T_i = 1$, with $x_i^\intercal \hat{\beta}_0$ as an offset term, and the intercept as the only explicative variable.

The following estimated probabilities can be obtained from this logistic model:

$$\hat{p}_{Y_i|1} = \frac{\exp\left(x_i^\intercal \hat{\beta}_0 + \hat{\Gamma}_1\right)}{1 + \exp\left(x_i^\intercal \hat{\beta}_0 + \hat{\Gamma}_1\right)}$$

We also note

$$\tilde{p}_{Y_i|1} = \frac{\exp\left(x_i^\intercal \hat{\beta}_0 + \Gamma\right)}{1 + \exp\left(x_i^\intercal \hat{\beta}_0 + \Gamma\right)}$$

The normal equation for the logistic regression, $\sum_{i=1}^n T_i(\hat{p}_{Y_i|1} - Y_i) = 0$, can be written as follows:

$$\sum_{i=1}^n T_i(\hat{p}_{Y_i|1} - \tilde{p}_{Y_i|1}) = \sum_{i=1}^n T_i(Y_i - \tilde{p}_{Y_i|1}). \tag{A.6}$$

By applying a Taylor expansion similar to that leading to (A.1), we obtain:

$$\hat{p}_{Y_i|1} - \tilde{p}_{Y_i|1} \approx \hat{p}_{Y_i|1}(1 - \hat{p}_{Y_i|1})(\hat{\Gamma}_1 - \Gamma). \tag{A.7}$$

By plugging (A.7) into (A.6), we obtain:

$$\hat{\Gamma}_1 - \Gamma \approx \frac{\sum_{i=1}^n T_i(Y_i - \tilde{p}_{Y_i|1})}{\sum_{i=1}^n T_i\hat{p}_{Y_i|1}(1 - \hat{p}_{Y_i|1})} = Z_1 - Z_2 \tag{A.8}$$

with

$$Z_1 = \frac{\sum_{i=1}^n T_i(Y_i - p_{1,i})}{\sum_{i=1}^n T_i\hat{p}_{Y_i|1}(1 - \hat{p}_{Y_i|1})} \text{ and } Z_2 = \frac{\sum_{i=1}^n T_i(\tilde{p}_{Y_i|1} - p_{1,i})}{\sum_{i=1}^n T_i\hat{p}_{Y_i|1}(1 - \hat{p}_{Y_i|1})}.$$

Following Deville's [47] rules, the estimated linearized variable of Z_1 is

$$\hat{U}_i(Z_1) = \frac{T_i(Y_i - \hat{p}_{Y_i|1})}{n^{-1}\sum_{j=1}^n T_j\hat{p}_{Y_j|1}(1 - \hat{p}_{Y_j|1})}.$$

We now consider Z_2. With a derivation similar to that leading to (A.4), we obtain

$$\tilde{p}_{Y_i|1} - p_{1,i} \approx \left[\hat{p}_{Y_i|1}(1 - \hat{p}_{Y_i|1})x_i^\intercal\right]\left[\sum_{j=1}^n (1 - T_j)\hat{p}_{0,j}(1 - \hat{p}_{0,j})x_j x_j^\intercal\right]^{-1}\sum_{j=1}^n (1 - T_j)(Y_j - p_{0,j})x_j. \tag{A.9}$$

By plugging (A.9) into Z_2, we obtain

$$Z_2 \approx \frac{\sum_{i=1}^n (1 - T_i)(Y_i - p_{0,i})Ax_i}{\sum_{i=1}^n T_i\hat{p}_{Y_i|1}(1 - \hat{p}_{Y_i|1})}$$

with

$$A = \left[\frac{1}{n}\sum_{i=1}^n T_i\hat{p}_{Yi|1}(1 - \hat{p}_{Yi|1})x_i\right]^\intercal\left[\frac{1}{n}\sum_{i=1}^n (1 - T_i)\hat{p}_{0,i}(1 - \hat{p}_{0,i})x_i x_i^\intercal\right]^{-1}$$

The estimated linearized variable of Z_2 is

$$\hat{U}_i(Z_2) = \frac{(1-T_i)(Y_i - \hat{p}_{0,i})Ax_i}{n^{-1}\sum_{i=1}^{n} T_i \hat{p}_{Y_i|1}(1-\hat{p}_{Y_i|1})}$$

Overall, the estimated linearized variable of $\hat{\Gamma}_1$ is thus:

$$\boxed{\hat{U}_i(\hat{\Gamma}_1) = \frac{T_i(Y_i - \hat{p}_{Yi|1}) - (1-T_i)(Y_i - \hat{p}_{0,i})Ax_i}{n^{-1}\sum_{i=1}^{n} T_i \hat{p}_{Yi|1}(1-\hat{p}_{Yi|1})}}$$

C. Derivation of $\hat{U}_i(\hat{\Gamma}_2)$
The second estimator is

$$\hat{\Gamma}_2 = \text{logit}(O^1) - \text{logit}\left(P_0^1\right)$$

with

$$O^1 = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} \text{ and } P_0^1 = \frac{\sum_{i=1}^{n} T_i \hat{p}_{0,i}}{\sum_{i=1}^{n} T_i}.$$

First of all, by noting $f(x) = \text{logit}(x)$, we have

$$\begin{aligned}\hat{U}_i(\hat{\Gamma}_2) &= f'(O^1)\hat{U}_i(O^1) - f'\left(P_0^1\right)\hat{U}_i\left(P_0^1\right) \\ &= \frac{\hat{U}_i(O^1)}{O^1(1-O^1)} - \frac{\hat{U}_i\left(P_0^1\right)}{P_0^1\left(1-P_0^1\right)}.\end{aligned} \tag{A.10}$$

Following Deville's rules for a ratio parameter, the estimated linearized variable for O^1 is

$$\hat{U}_i(O^1) = \frac{T_i(Y_i - O^1)}{\bar{T}} \text{ with } \bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i.$$

We now consider P_0^1, which can be written as $P_0^1 = P_a + P_b$, with

$$P_a = \frac{\sum_{i=1}^{n} T_i p_{0,i}}{\sum_{i=1}^{n} T_i} \text{ and } P_b = \frac{\sum_{i=1}^{n} T_i(\hat{p}_{0,i} - p_{0,i})}{\sum_{i=1}^{n} T_i}.$$

We have $\hat{U}(P_0^1) = \hat{U}(P_a) + \hat{U}(P_b)$, with

$$\hat{U}(P_a) = \frac{T_i\left(\hat{p}_{0,i} - P_0^1\right)}{\bar{T}}. \tag{A.11}$$

We now consider P_b. Using (A.4), we obtain the following approximation:

$$P_b \simeq \frac{\sum_{i=1}^{n}(1-T_i)(Y_i - p_{0,i})Bx_i}{\sum_{i=1}^{n} T_i}$$

with

$$B = \left[\frac{1}{n}\sum_{i=1}^{n} T_i \hat{p}_{0,i}(1-\hat{p}_{0,i})x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i x_i^{\intercal}\right]^{-1}.$$

The estimated linearized variable of P_b is thus

$$\hat{U}_i(P_b) = \frac{(1-T_i)(Y_i - \hat{p}_{0,i})Bx_i}{\bar{T}}. \tag{A.12}$$

By plugging (A.11) and (A.12) into (A.10), we obtain

$$\boxed{\hat{U}_i(\hat{\Gamma}_2) = \frac{T_i(Y_i - O^1)}{O^1(1-O^1)\bar{T}} - \frac{T_i\left(\hat{p}_{0,i} - P_0^1\right) + (1-T_i)\left(Y_i - \hat{p}_{0,i}\right)Bx_i}{P_0^1\left(1-P_0^1\right)\bar{T}}}$$

D. Derivation of $\hat{U}_i(\hat{\Gamma}_3)$

The third estimator could be written as

$$\hat{\Gamma}_3 = L_1 - L_0$$

with:

$$L_1 = \frac{1}{n}\sum_{i=1}^{n} \text{logit}(\hat{p}_{1,i}) \text{ and } L_0 = \frac{1}{n}\sum_{i=1}^{n} \text{logit}(\hat{p}_{0,i})$$

This estimator can be rewritten as

$$\begin{aligned}
\hat{\Gamma}_3 &= \frac{1}{n}\sum_{i=1}^{n} x_i^\intercal(\hat{\beta}_1 - \hat{\beta}_0) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^\intercal\{(\hat{\beta}_1 - \tilde{\beta}) + (\tilde{\beta} - \beta) - (\hat{\beta}_0 - \beta)\} \\
&= \Gamma + \frac{1}{n}\sum_{i=1}^{n} x_i^\intercal\{(\hat{\beta}_1 - \tilde{\beta}) - (\hat{\beta}_0 - \beta)\} \\
\Rightarrow \hat{\Gamma}_3 - \Gamma &= \frac{1}{n}\sum_{i=1}^{n} x_i^\intercal\{(\hat{\beta}_1 - \tilde{\beta}) - (\hat{\beta}_0 - \beta)\}
\end{aligned} \tag{A.13}$$

with $\tilde{\beta} = \beta + (\Gamma, 0, \ldots, 0)^\intercal$. An approximation for $(\hat{\beta}_0 - \beta)$ is given by Equation (A.3). Similarly, we have

$$\hat{\beta}_1 - \tilde{\beta} \simeq \left[\sum_{i=1}^{n} T_i\hat{p}_{1,i}(1-\hat{p}_{1,i})x_i x_i^\intercal\right]^{-1} \sum_{i=1}^{n} T_i(Y_i - p_{1,i})x_i \tag{A.14}$$

We obtain from (A.13), (A.3), and (A.14) an estimated linearized variable of $\hat{\Gamma}_3$:

$$\boxed{\hat{U}_i(\hat{\Gamma}_3) = T_i(Y_i - \hat{p}_{1,i})C_1 x_i - (1-T_i)(Y_i - \hat{p}_{0,i})C_0 x_i}$$

with:

$$\begin{aligned}
C_1 &= \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^\intercal \left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{1,i}(1-\hat{p}_{1,i})x_i x_i^\intercal\right]^{-1}, \\
C_0 &= \left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^\intercal \left[\frac{1}{n}\sum_{i=1}^{n} (1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i x_i^\intercal\right]^{-1}.
\end{aligned}$$

E. Derivation of $\hat{U}_i(\hat{\Gamma}_4)$

The fourth estimator is given by

$$\hat{\Gamma}_4 = \text{logit}(P_1) - \text{logit}(P_0)$$

with $P_1 = \frac{\sum_{i=1}^{n}\hat{p}_{1,i}}{n}$ and $P_0 = \frac{\sum_{i=1}^{n}\hat{p}_{0,i}}{n}$.

First, by noting again $f(x) = \text{logit}(x)$, we have

$$\hat{U}_i(\hat{\Gamma}_4) = \frac{\hat{U}_i(P_1)}{P_1(1-P_1)} - \frac{\hat{U}_i(P_0)}{P_0(1-P_0)}. \tag{A.15}$$

We can write P_0 as

$$P_0 = \frac{1}{n}\sum_{i=1}^{n} p_{0,i} + \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_{0,i} - p_{0,i}). \tag{A.16}$$

By plugging Equation (A.4) into (A.16), we obtain

$$P_0 \simeq \frac{1}{n}\sum_{i=1}^{n} p_{0,i} + \frac{1}{n}\sum_{i=1}^{n}(1-T_i)(Y_i - p_{0,i})D_0 x_i$$

with

$$D_0 = \left[\frac{1}{n}\sum_{i=1}^{n}\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})x_i x_i^{\intercal}\right]^{-1}.$$

The estimated linearized variable of P_0 is

$$\hat{U}_i(P_0) = \hat{p}_{0,i} + (1-T_i)(Y_i - \hat{p}_{0,i})D_0 x_i$$

By symmetry, the estimated linearized variable of P_1 is

$$\hat{U}_i(P_1) = \hat{p}_{1,i} + T_i(Y_i - \hat{p}_{1,i})D_1 x_i$$

with

$$D_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\hat{p}_{1,i}(1-\hat{p}_{1,i})x_i\right]^{\intercal}\left[\frac{1}{n}\sum_{i=1}^{n}T_i\hat{p}_{1,i}(1-\hat{p}_{1,i})x_i x_i^{\intercal}\right]^{-1}.$$

Overall, the estimated linearized variable of $\hat{\Gamma}_4$ is

$$\boxed{\hat{U}_i(\hat{\Gamma}_4) = \frac{\hat{p}_{1,i} + T_i(Y_i - \hat{p}_{1,i})D_1 x_i}{P_1(1-P_1)} - \frac{\hat{p}_{0,i} + (1-T_i)(Y_i - \hat{p}_{0,i})D_0 x_i}{P_0(1-P_0)}}.$$

Acknowledgements

The simulation study was sponsored by the Agence nationale de sécurité du médicament et des produits de santé (ANSM) (no. AAP-2015-051). The case study corresponds to the investigation of a methodological point within the framework of a study funded by the French National Health Insurance Fund (Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés, CNAMTS). The authors thank Alexandre Lafourcade for work on the case study.

References

1. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**(5):761–768.
2. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2007; **26**(16):3078–3094.
3. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine* 2014; **33**(7):1242–1258.
4. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal. Biometrische Zeitschrift* 2009; **51**(1):171–184.
5. 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; **26**(4):754–768.
6. Hansen BB. The prognostic analogue of the propensity score. *Biometrika* 2008; **95**(2):481–488.
7. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology and Drug Safety* 2012; **21**:138–147.

8. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM (eds). *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*, AHRQ Methods for Effective Health Care. Agency for Healthcare Research and Quality (US): Rockville (MD), 2013.
9. Arbogast PG, Seeger JD, DEcIDE Methods Center Summary VWG. Summary variables in observational research: propensity scores and disease risk scores. *Effective Health Care Program Research Report* 2012; **33**.
10. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research* 2009; **18**(1):67–80.
11. Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine* 2014; **33**(20):3488–3508.
12. Wyss R, Ellis AR, Brookhart MA, Jonsson FM, Girman CJ, Simpson RJ, Stürmer T. Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and Drug Safety* 2015; **24**(9):951–961.
13. Miettinen OS, Cook EF. Confounding: essence and detection. *American Journal of Epidemiology* 1981; **114**(4): 593–603.
14. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55.
15. 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; **34**(28): 3661–3679.
16. Resche-Rigon M, Pirracchio R, Robin M, Latour RPD, Sibon D, Ades L, Ribaud P, Fermand JP, Thieblemont C, Socié G, Chevret S. Estimating the treatment effect from non-randomized studies: the example of reduced intensity conditioning allogeneic stem cell transplantation in hematological diseases. *BMC Hematology* 2012; **12**(1):1–10.
17. Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research* 2011; **46**(1):119–151.
18. 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; **63**(8):826–833.
19. Austin PC. The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology* 2008; **61**(6):537–545.
20. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. *BMC Medical Research Methodology* 2012; **12**(1):1–10.
21. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**(387):516–524.
22. Lunceford JK, 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.
23. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 1996; **52**(1): 249–264.
24. Abadie A, Imbens GW. Matching on the estimated propensity score, Technical Report 15301, 2009.
25. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987; **82**(398):387–394.
26. 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* 2010; **29**(20):2137–2148.
27. Forbes A, Shortreed S. Inverse probability weighted estimation of the marginal odds ratio: correspondence regarding 'The performance of different propensity score methods for estimating marginal odds ratios' by P. Austin, *Statictics in Medicine*, 2007; **26**:3078–3094. *Statistics in Medicine* 2008; **27**(26):5556–5559.
28. Graf E, Schumacher M. Comments on 'The performance of different propensity score methods for estimating marginal odds ratios' by Peter C. Austin, *Statistics in Medicine* 2007; **26**(16):3078–3094. *Statistics in Medicine* 2008; **27**(19): 3915–3917.
29. 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: An International Journal of the Society for Medical Decision Making* 2009; **29**(6):661–677.
30. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**(2):295–313.
31. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine* 2014; **33**(6): 1057–1069.
32. Donald B, Rubin PRR. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; **39**(1):33–38.
33. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 2008; **27**(12):2037–2049.
34. 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; **25**(5):2214–2237.
35. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology* 2006; **59**(5):437–447.
36. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology* 2003; **158**(3):280–287.
37. 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; **16**:1–16.
38. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* 2013; **32**(19):3388–3414.
39. Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* 2004; **99**:854–866.

40. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *American Journal of Epidemiology* 2011; **174**(5):613–620.
41. Schmidt AF, Klungel OH, Groenwold RHH, GetReal C. Adjusting for confounding in early postlaunch settings: going beyond logistic regression models. *Epidemiology (Cambridge, Mass.)* 2016; **27**(1):133–142.
42. Miettinen OS. Stratification by a multivariate confounder score. *American Journal of Epidemiology* 1976; **104**(6):609–620.
43. Pike MC, Anderson J, Day N. Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. *Epidemiology and Community Health* 1979; **33**(1):104–106.
44. Connolly JG, Gagne JJ. Comparison of calipers for matching on the disease risk score. *American Journal of Epidemiology* 2016; **183**(10):937–948.
45. Pfeiffer RM, Riedl R. On the use and misuse of scalar scores of confounders in design and analysis of observational studies. *Statistics in Medicine* 2015; **34**(18):2618–2635.
46. Tadrous M, Gagne JJ, Stürmer T, Cadarette SM. Disease risk score (DRS) as a confounder summary method: systematic review and recommendations. *Pharmacoepidemiology and Drug Safety* 2013; **22**(2):122–129.
47. Deville JC. Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* 1999; **25**(2):193–204.
48. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 2011; **10**(2):150–161.
49. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine* 2013; **32**(16):2837–2849.
50. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software* 2011; **42**(7):1–52.
51. Lumley T. Analysis of complex survey samples. *Journal of Statistical Software* 2004; **9**(1):1–19.
52. Laforest L, Licaj I, Devouassoux G, Chatte G, Martin J, Van Ganse Eric. Asthma drug ratios and exacerbations: claims data from universal health coverage systems. *The European Respiratory Journal* 2014; **43**(5):1378–1386.
53. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011; **46**(3):399–424.
54. Austin PC, Laupacis A. A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: a review. *The International Journal of Biostatistics* 2011; **7**(1).
55. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology* 2007; **60**(9):874–882.
56. Colantuoni E, Rosenblum M. Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in Medicine* 2015; **34**(18):2602–2617.
57. Rosenblum M, van der Laan MJ. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics* 2010; **6**(1).
58. Xu S, Shetterly S, Cook AJ, Raebel MA, Goonesekera S, Shoaibi A, Roy J, Fireman B. Evaluation of propensity scores, disease risk scores, and regression in confounder adjustment for the safety of emerging treatment with group sequential monitoring. *Pharmacoepidemiology and Drug Safety* 2016; **25**(4):453–461.
59. Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**(3):659–663.
60. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *Journal of Clinical Epidemiology* 1989; **42**(4):317–324.
61. Lunceford JK, 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.
62. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* 2014; **33**(5):721–737.
63. Williamson EJ, Morley R, Lucas A, Carpenter JR. Variance estimation for stratified propensity score estimators. *Statistics in Medicine* 2012; **31**(15):1617–1632.
64. Zou B, Zou F, Shuster JJ, Tighe PJ, Koch GG, Zhou H. On variance estimate for covariate adjustment by propensity score analysis. *Statistics in Medicine* 2016.
65. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine* 2014; **33**(24):4306–4319.

4 | Estimateurs de la variance de l'effet

4.1 Problèmes liés à la non la prise en compte de l'étape d'estimation du score

Un point commun important entre les méthodes basées sur le score de propension et les méthodes basées sur le score pronostique est qu'elles nécessitent toutes une première étape d'estimation du score qui sera ensuite utilisé pour corriger le biais d'indication. En effet, le score de propension et le score pronostique ne sont pas des valeurs théoriques, mais des valeurs estimées à partir des données observées. Ils sont donc en eux-mêmes des sources de variabilité qui doivent être prises en compte dans les estimateurs de la variance de l'effet du traitement.

4.1.1 Conséquence avec le score pronostique

La conséquence de la non prise en compte de l'étape d'estimation du score pronostique est intuitivement évidente. Prenons l'exemple la méthode SPN-ATT que nous avons proposée au chapitre précédent pour estimer l'ATT :

$$\hat{\Gamma}_2 = \text{logit}(\widehat{O^1}) - \text{logit}(\widehat{P_0^1}).$$

Ignorer la variabilité issue de l'estimation de $\widehat{P_0^1}$ conduit nécessairement à la sous-estimation de la variance de l'effet du traitement. Un raisonnement du même type peut être fait pour toutes les méthodes basées sur le score pronostique. De fait, cette sous-estimation de la variance a bien été mise en évidence dans nos simulations pour les méthodes d'utilisation existantes du score pronostique vues au chapitre 3 (Section 3.2.2 page 51), qui utilisent des estimateurs de variance ignorant cette source de variabilité. Le lien entre l'importance de la sous-estimation de la variance et la prévalence de l'exposition s'explique simplement : plus la prévalence de l'exposition est élevée, plus le nombre de sujets non exposés est faible, et plus la variabilité liée à l'étape d'estimation du score pronostique $\widehat{\Psi}_0$ est élevée (puisque les coefficients du modèle pronostique sont estimés dans le sous-groupe des non exposés uniquement, et donc un petit nombre de sujets). Ainsi, ignorer l'étape d'estimation du score pronostique a des conséquences plus graves en cas de prévalence élevée de l'exposition.

4.1.2 Conséquence avec le score de propension

A l'inverse, et de manière moins intuitive, ignorer l'étape d'estimation du score de propension a une conséquence diamétralement opposée à celle décrite pour le score pronostique : considérer le score de propension estimé comme une valeur théorique conduit à une surestimation de la variance de l'effet du traitement. Dans un article focalisé sur la méthode de pondération sur le score de propension, Williamson et al. (2014) ont fourni une explication intuitive en partant d'une situation où le score de propension théorique est connu : l'essai randomisé. Cette explication étant généralisable à la situation observationnelle (ainsi qu'aux autres méthodes d'utilisation du score de propension), nous la paraphrasons ci-après.

> *L'étape d'estimation du score de propension n'est pas une source de variabilité dans son sens usuel. Dans un essai randomisé, le score de propension théorique est de 0.5 pour tous les sujets de l'étude, et l'estimation non ajustée de l'effet du traitement est alors égale à l'estimation par pondération sur ce score de pro-*

pension théorique. La pondération sur le score de propension théorique ne tient alors pas compte des déséquilibres initiaux, liés au hasard, des facteurs pronostiques entre les deux groupes comparés. La variabilité du score de propension estimé reflète simplement ces déséquilibres des facteurs pronostiques inclus dans le modèle utilisé pour estimer le score de propension. L'estimation de l'effet du traitement par pondération sur le score de propension estimé entraîne une amélioration de l'équilibre des facteurs pronostiques (contrairement à l'estimation non ajustée), et donc une amélioration de la précision de l'estimation de l'effet du traitement. L'absence de prise en compte de l'étape d'estimation du score de propension ignore cette amélioration de la précision obtenue grâce au rééquilibrage des caractéristiques initiales. La précision apparaît alors faussement diminuée.

4.2 Méthode de linéarisation pour l'estimation de la variance d'un estimateur complexe

Dans le cas d'estimateurs complexes comme le sont les estimateurs de l'effet du traitement par les méthodes basées sur le score de propension ou le score pronostique précédemment décrites, il est possible d'obtenir des estimateurs approximativement sans biais de la variance en utilisant la méthode de linéarisation décrite par Deville (1999). Brièvement, pour un paramètre théorique Γ et un estimateur $\widehat{\Gamma}$, la linéarisation consiste à trouver une variable U telle que

$$E(\frac{s_U^2}{n}) \simeq \mathrm{Var}(\widehat{\Gamma})$$

où

$$s_U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \bar{U})^2 \text{ et } \bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i.$$

Habituellement, la variable U dépend de paramètres inconnus qui peuvent être estimés à

partir des données observées, pour obtenir une variable linéarisée estimée $\widehat{U}$. L'estimateur de la variance devient alors :

$$\widehat{\mathrm{Var}}(\hat{\Gamma}) = \frac{s^2_{\widehat{U}}}{n}$$

où

$$s^2_{\widehat{U}} = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{U}_i - \bar{\widehat{U}})^2 \text{ et } \bar{\widehat{U}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{U}_i.$$

Des règles de calculs détaillées et des exemples d'application à des estimateurs simples et complexes sont décrits par Deville (1999) dans un article particulièrement didactique. Nous reprenons ici un exemple simple mais instructif pour décrire le processus de linéarisation : l'estimateur de variance d'un ratio de deux moyennes. Dans un échantillon constitué de n sujets i indépendants, soit X et Y deux variables d'espérance μ_X et μ_Y respectivement. On estime

$$R = \frac{\mu_X}{\mu_Y} \text{ par } \hat{R} = \frac{\bar{Y}}{\bar{X}}.$$

On peut montrer que la variable linéarisée d'un ratio $\hat{R}$ est :

$$U_i(R) = \frac{1}{\mu_X}(Y_i - RX_i).$$

Les paramètres R et μ_X étant inconnus, on les remplace par leur estimateurs $\hat{R}$ et $\bar{X}$ pour obtenir la variable linéarisée estimée de $\widehat{R}$:

$$\widehat{U}_i(R) = \frac{1}{\bar{X}}(Y_i - \hat{R}X_i)$$

que l'on peut calculer à partir des données observées.

4.3 Application au score pronostique

Nous avons appliqué la méthode de linéarisation pour développer un estimateur de variance tenant compte de l'étape d'estimation du score pronostique pour chacune des nouvelles

méthodes d'utilisation du score pronostique présentées au chapitre 3 (les démonstrations étant fournies en annexe de l'article 2) :

Pour $\widehat{\Gamma}_1$ (SPN-CTE1, page 54)

$$\widehat{U_i}(\widehat{\Gamma}_1) = \frac{T_i(Y_i - \hat{p}_{Yi|1}) - (1-T_i)(Y_i - \hat{p}_{0,i})AX_i}{n^{-1}\sum_{i=1}^{n} T_i\hat{p}_{Yi|1}(1-\hat{p}_{Yi|1})}$$

où $\hat{p}_{Yi|1} = \text{logit}^{-1}(\widehat{\Gamma}_1 + \hat{\Psi}_{0,i})$ est une probabilité d'évènement estimée pour le sujet i, $X_i = (X_{1,i}, \ldots, X_{K,i})^\top$ est le vecteur des caractéristiques du sujet i et

$$A = \left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{Yi|1}(1-\hat{p}_{Yi|1})X_i\right]^\top \left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})X_iX_i^\top\right]^{-1}.$$

Pour $\widehat{\Gamma}_2$ (SPN-ATT, page 54)

$$\widehat{U_i}(\widehat{\Gamma}_2) = \frac{T_i(Y_i - \widehat{O^1})}{\widehat{O^1}(1-\widehat{O^1})\bar{T}} - \frac{T_i(\hat{p}_{0,i} - \widehat{P_0^1}) + (1-T_i)(Y_i - \hat{p}_{0,i})BX_i}{\widehat{P_0^1}(1-\widehat{P_0^1})\bar{T}}$$

où

$$B = \left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{0,i}(1-\hat{p}_{0,i})X_i\right]^\top \left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})X_iX_i^\top\right]^{-1}.$$

Pour $\widehat{\Gamma}_3$ (SPN-CTE2, page 54)

$$\widehat{U_i}(\widehat{\Gamma}_3) = T_i(Y_i - \hat{p}_{1,i})C_1X_i - (1-T_i)(Y_i - \hat{p}_{0,i})C_0X_i$$

où

$$C_1 = \left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]^\top \left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{1,i}(1-\hat{p}_{1,i})X_iX_i^\top\right]^{-1},$$

et

$$C_0 = \left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]^\top \left[\frac{1}{n}\sum_{i=1}^{n}(1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})X_iX_i^\top\right]^{-1}.$$

Pour $\widehat{\Gamma}_4$ (SPN-ATE, page 55)

$$\widehat{U_i}(\widehat{\Gamma}_4) = \frac{\hat{p}_{1,i} + T_i(Y_i - \hat{p}_{1,i})D_1 X_i}{\widehat{P}_1(1-\widehat{P}_1)} - \frac{\hat{p}_{0,i} + (1-T_i)(Y_i - \hat{p}_{0,i})D_0 X_i}{\widehat{P}_0(1-\widehat{P}_0)}$$

où

$$D_1 = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{p}_{1,i}(1-\hat{p}_{1,i})X_i\right]^\top \left[\frac{1}{n}\sum_{i=1}^{n} T_i\hat{p}_{1,i}(1-\hat{p}_{1,i})X_i X_i^\top\right]^{-1},$$

et

$$D_0 = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{p}_{0,i}(1-\hat{p}_{0,i})X_i\right]^\top \left[\frac{1}{n}\sum_{i=1}^{n} (1-T_i)\hat{p}_{0,i}(1-\hat{p}_{0,i})X_i X_i^\top\right]^{-1}.$$

Dans l'article 2, l'évaluation des performances des estimateurs de variance reposait sur le ratio entre la moyenne des écart-types estimés et l'écart-type empirique (écart-type des estimations de l'effet du traitement calculé sur l'ensemble des simulations d'un scenario donné). Ce critère de performance devrait être proche de la valeur 1 en cas d'estimation non biaisée de la variance. Un ratio supérieur (ou inférieur) à 1 suggère qu'en moyenne, les variances estimées surestiment (ou sous-estiment) la variabilité de l'effet du traitement.

Contrairement aux estimateurs de variance *model-based* utilisés avec les méthodes d'utilisation existantes du score pronostique (qui ne tiennent pas compte de l'étape d'estimation du score pronostique et sous-estimaient systématiquement la variance de l'effet du traitement), les estimateurs de variance développés pour les nouvelles méthodes d'utilisation du score pronostique estimaient convenablement la variance de l'effet du traitement (ratio des variabilités proche de 1) dans l'ensemble des scénarios considérés.

4.4 Application à la pondération sur le score de propension

Contrairement au cas du score pronostique, de nombreux estimateurs de variance ont déjà été proposés dans la littérature pour les méthodes basées sur le score de propension. [1]

1. Voir Zou et al. (2016) pour l'ajustement sur le score de propension, Williamson et al. (2012) pour la stratification sur le score de propension, et Abadie & Imbens (2009) pour l'appariement sur le score de propension.

Concernant la méthode de pondération sur le score de propension et dans le cas d'un critère de jugement binaire, trois estimateurs de variance ont été proposés :

— l'estimateur proposé par Robins et al. (2000) (estimateur *sandwich* issu d'un modèle de régression pondéré) adapté à l'estimation de l'ATE et de l'ATT par une différence de risque (DR), un risque relatif (RR) ou un odds-ratio (OR) ;
— l'estimateur proposé par Lunceford & Davidian (2004) adapté à l'estimation de l'ATE par une DR ;
— et l'estimateur proposé par Williamson et al. (2014) adapté à l'estimation de l'ATE par une DR, un RR ou un OR.

L'estimateur *sandwich* ne tient pas compte du fait que le score de propension théorique est inconnu mais est estimé à partir des données, et conduit donc à une surestimation de la variance et des intervalles de confiance trop conservateurs (Williamson et al. 2012). Les deux autres estimateurs ont quant à eux été développés en tenant compte de cette étape d'estimation du score de propension.

Mais même si l'estimateur de Williamson et al. (2014) et l'estimateur de Lunceford & Davidian (2004) présentent de nombreuses ressemblances, le premier semble estimer convenablement la variance de l'ATE, tandis que le second semble la surestimer et donc conduire à des intervalles de confiance trop conservateurs (Austin 2010b). De plus, comme nous l'avons vu au chapitre 2, la méthode de pondération sur le score de propension permet également d'estimer l'ATT, mais aucun de ces auteurs n'en fournit un estimateur de variance adapté.

Dans un article soumis en Novembre 2016, nous avons cherché à mettre en évidence les raisons expliquant la différence de performance constatée dans la littérature entre l'estimateur de Lunceford & Davidian (2004) et l'estimateur de Williamson et al. (2014). Nous avons également proposé une approche unifiée, basée sur la méthode de linéarisation, pour dériver des estimateurs de variance adaptés à l'estimation de l'ATE ou de l'ATT par une DR, un RR ou un OR. Dans la suite de cet exposé et par soucis de concision, nous n'exposerons

que les estimateurs adaptés à une DR, les autres estimateurs se déduisant facilement.

4.4.1 Estimateurs de l'effet du traitement

L'estimateur pondéré du taux d'évènements chez les sujets exposés au traitement s'écrit (Imbens 2004) :

$$\widehat{P}_{11} = \frac{\sum_{i=1}^{n} \frac{T_i}{\hat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} \frac{T_i}{\hat{p}_{Ti}}}.$$

De même, chez les non exposés :

$$\widehat{P}_{10} = \frac{\sum_{i=1}^{n} \frac{1-T_i}{1-\hat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} \frac{1-T_i}{1-\hat{p}_{Ti}}}.$$

Une estimation sans biais de l'ATE peut être obtenue en calculant $\widehat{\Gamma}_{DR1} = \widehat{P}_{11} - \widehat{P}_{10}$.

Pour l'ATT, nous avons calculé (Imbens 2004) :

$$\widehat{P}_{21} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i}$$

et

$$\widehat{P}_{20} = \frac{\sum_{i=1}^{n} (1-T_i) \frac{\hat{p}_{Ti}}{1-\hat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} (1-T_i) \frac{\hat{p}_{Ti}}{1-\hat{p}_{Ti}}}.$$

Une estimation sans biais de l'ATT peut être obtenue en calculant $\widehat{\Gamma}_{DR2} = \widehat{P}_{21} - \widehat{P}_{20}$.

4.4.2 Estimateurs de variance de Lunceford & Davidian (2004) et de Williamson et al. (2014)

Estimateur de Lunceford & Davidian (2004). En reprenant les notations de Lunceford & Davidian (2004), l'estimateur de variance de $\widehat{\Gamma}_{DR1}$ est obtenu par

$\widehat{\text{Var}}(\widehat{\Gamma}_{DR1}) = n^{-2}\sum_{i=1}^{n}\widehat{I}_i^2$ avec :

$$\widehat{I}_i = \frac{T_i}{\hat{p}_{Ti}}\left(Y_i - \widehat{P}_{11}\right) - \frac{1-T_i}{1-\hat{p}_{Ti}}(Y_i - \widehat{P}_{10}) - (T_i - \hat{p}_{Ti})\widehat{H}^\top \widehat{E}^{-1}X_i,$$

où

$$\widehat{H} = n^{-1}\sum_{j=1}^{n}\left\{T_j(Y_j - \widehat{P}_{11})\frac{1-\hat{p}_{Tj}}{\hat{p}_{Tj}} + (1-T_j)(Y_j - \widehat{P}_{10})\frac{\hat{p}_{Tj}}{1-\hat{p}_{Tj}}\right\}X_j$$

et

$$\widehat{E}^{-1} = n^{-1}\sum_{j=1}^{n}\hat{p}_{Tj}(1-\hat{p}_{Tj})X_jX_j^\top.$$

Estimateur de Williamson et al. (2014). En reprenant les notations de Williamson et al. (2014), l'estimateur de variance de $\widehat{\Gamma}_{DR1}$ est obtenu par :

$$n\widehat{\text{Var}}(\widehat{\Gamma}_{DR1}) = \widehat{V}_{un} - \hat{\mathbf{v}}^\top(2\widehat{M}_1 - \widehat{M}_2)\hat{\mathbf{v}}.$$

En notant que $\widehat{W}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{p}_{Ti}}$, et $\widehat{W}_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\hat{p}_{Ti}}$, nous avons

$$\widehat{V}_{un} = \frac{1}{\widehat{W}_1^2}\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{P}_{11})^2T_i}{\hat{p}_{Ti}^2} + \frac{1}{\widehat{W}_0^2}\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{P}_{10})^2(1-T_i)}{(1-\hat{p}_{Ti})^2},$$

$$\hat{\mathbf{v}} = \frac{1}{\widehat{W}_1}\frac{1}{n}\sum_{i=1}^{n}\frac{X_i(Y_i - \widehat{P}_{11})T_i(1-\hat{p}_{Ti})}{\hat{p}_{Ti}} + \frac{1}{\widehat{W}_0}\frac{1}{n}\sum_{i=1}^{n}\frac{X_i(Y_i - \widehat{P}_{10})(1-T_i)\hat{p}_{Ti}}{(1-\hat{p}_{Ti})},$$

$$\widehat{M}_1 = \left(n^{-1}\sum_{j=1}^{n}\hat{p}_{Tj}(1-\hat{p}_{Tj})X_jX_j^\top\right)^{-1},$$

et

$$\widehat{M}_2 = \widehat{M}_1\left(n^{-1}\sum_{j=1}^{n}X_jX_j^\top(T_j - \hat{p}_{Tj})^2\right)\widehat{M}_1.$$

Nous pouvons observer que $\widehat{M}_1 = \widehat{E}$. De même, $\hat{\mathbf{v}} = \widehat{H}$ si nous remplaçons $\widehat{W}_0$ et $\widehat{W}_1$ par leur approximation égale à 1. En fait, à cette dernière approximation près, le développement de $n^{-2}\sum_{i=1}^{n}\widehat{I}_i^2$ conduit à l'estimateur de Williamson et al. (2014), à l'exception notable que $\widehat{E}$ est utilisé dans l'estimateur de Williamson et

al. (2014), et $\widehat{E}^{-1}$ est utilisé dans l'estimateur de Lunceford & Davidian (2004), ce qui explique la différence entre les estimations fournies par ces deux estimateurs.

4.4.3 Estimateurs obtenus par linéarisation

Estimateur de variance de l'ATE. En utilisant les mêmes notations que celles utilisées dans Lunceford & Davidian (2004), la technique de linéarisation permet d'obtenir la variable linéarisée estimée suivante :

$$\widehat{U}_i(\widehat{\Gamma}_{DR1}) = \frac{T_i}{\hat{p}_{Ti}}\left(Y_i - \widehat{P}_{11}\right) - \frac{1-T_i}{1-\hat{p}_{Ti}}(Y_i - \widehat{P}_{10}) - (T_i - \hat{p}_{Ti})\widehat{H}^{\top}\widehat{E}X_i.$$

Nous pouvons constater que les expressions de $\widehat{U}_i(\widehat{\Gamma}_{DR1})$ dans l'équation précédente et de $\widehat{I}_i$ dans l'estimateur de Lunceford & Davidian (2004) sont proches, à l'exception de $\widehat{E}$ remplaçant $\widehat{E}^{-1}$, comme c'était aussi le cas dans l'estimateur de Williamson et al. (2014) ; ce dernier est par conséquent quasi équivalent à notre estimateur obtenu par linéarisation.

Estimateur de variance de l'ATT. Nous avons également développé une variable linéarisée estimée de $\widehat{\Gamma}_{DR2}$:

$$\widehat{U}_i(\widehat{\Gamma}_{DR2}) = \widehat{U}_i(\widehat{P}_{21}) - \widehat{U}_i(\widehat{P}_{20})$$

où

$$\widehat{U}_i(\widehat{P}_{21}) = \bar{T}^{-1}T_i(Y_i - \widehat{P}_{21})$$

et

$$\widehat{U}_i(\widehat{P}_{20}) = \bar{n}_T^{-1}\left\{(1-T_i)\frac{\hat{p}_{Ti}}{1-\hat{p}_{Ti}}(Y_i - \widehat{P}_{20}) + (T_i - \hat{p}_{Ti})\hat{\gamma}_{20}^{\top}X_i\right\},$$

avec

$$\hat{\gamma}_{20} = \left\{\sum_{j=1}^{n}\hat{p}_{Tj}(1-\hat{p}_{Tj})X_jX_j^{\top}\right\}^{-1}\sum_{j=1}^{n}(1-T_j)\frac{\hat{p}_{Tj}}{1-\hat{p}_{Tj}}X_j(Y_j - \widehat{P}_{20})$$

et

$$\bar{n}_T = n^{-1} \sum_{i=1}^{n} (1 - T_i) \frac{\hat{p}_{Ti}}{1 - \hat{p}_{Ti}}.$$

Toutes les démonstrations sont fournies dans l'annexe de l'article 3.

4.4.4 Etude de simulation

Les estimateurs de variance de Lunceford & Davidian (2004), de Williamson et al. (2014) et les estimateurs obtenus par linéarisation ont été évalués dans une étude de simulation, et comparés aux estimateurs *sandwich* ne tenant pas compte de l'étape d'estimation du score de propension (Robins et al. 2000). Ce travail était illustré par la même application sur les données du SNIIRAM que celle utilisée dans l'article 2. Cette étude retrouvait une surestimation de la variance de l'ATE avec l'estimateur de Lunceford & Davidian (2004) (ratio des variabilités supérieur à 1) comparable à celle retrouvée avec l'estimateur *sandwich*, tandis que l'estimateur de Williamson et al. (2014) et l'estimateur obtenu par linéarisation fournissaient des estimations correctes (et superposables) de la variance de l'ATE. De plus, les performances de l'estimateur de Lunceford & Davidian (2004) se détérioraient très fortement si les facteurs de confusion pris en compte ne suivaient pas des lois normales centrées et réduites. Enfin, l'estimateur de la variance de l'ATT obtenu par linéarisation fournissait des estimations correctes de la variance, conduisant à des intervalles de confiance plus étroits que ceux basés sur l'estimateur *sandwich*.

Article 3

Variance estimation for weighted propensity score estimators

David Hajage[a,b,c,e†], Guillaume Chauvet[f,g†], Florence Tubach[a,b,d,e], and Yann De Rycke[a,b,c,e]

[a]APHP, Hôpital Pitié-Salpêtrière, Département de Biostatistiques, Santé publique et Information médicale, F-75013, Paris, France

[b]APHP, Centre de Pharmacoépidémiologie (Cephepi), F-75013, Paris, France

[c]Univ Paris Diderot, Sorbonne Paris Cité, UMR 1123 ECEVE, F-75010, Paris, France

[d]Université Pierre et Marie Curie – Paris 6, Sorbonne Universités, Paris, France

[e]INSERM, UMR 1123 ECEVE, F-75018, Paris, France

[f]Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAI), F-35170, Bruz, France

[g]IRMAR, UMR CNRS 6625, Rennes, France

[†]These authors contributed equally to this work.

Abstract

Weighting using the estimated propensity score (PS) is widely used to estimate the marginal effect of a treatment on a binary outcome in observational studies. Lunceford and Davidian (2003) and Williamson *et al.* (2014) provided treatment effect variance estimators suitable when the average treatment effect on the overall population (ATE) is estimated. However, the estimation of the average treatment effect on the treated population (ATT) was not addressed by these authors. In this case, variance estimation taken from a weighted (using the PS) regression model of the outcome on treatment with robust variance estimator could be used, but could lead to too conservative confidence intervals. In this article, we propose a unified approach to derive variance estimators of the average treatment effect on the overall or on the treated population using risk difference, relative risk or odds ratio, and demonstrate the performance of the proposed estimators through a simulation study. The resulting variance estimators for the ATE were very close to those proposed by Williamson *et al.* and resulted in correct estimates of standard errors. The variance estimators for the ATT also led to correct estimates of standard errors, resulting in narrower confidence intervals than those based on robust estimator.

1 Introduction

Randomized controlled trials are the gold standard to assess the effect of a treatment. However, this assessment in real life setting, i.e. after the market access and in large and unselected populations, mainly relies on observational studies. In such situation, crude treatment effect estimate is likely to be biased by indication, and adjustment for baseline confounding factors is needed. Among other relevant statistical tools, the propensity score

(PS) framework is very popular in the statistical and applied literature to account for observed baseline confounders [1].

The PS is defined as the probability of treatment conditionally to these confounders. It is estimated from the available data, as the true PS is unlikely to be known in the context of observational studies. PS framework was developed to induce balance of observed confounding factors between groups of treated and untreated subjects [2]. It is designed to estimate marginal treatment effects (MTE) [3] (as opposed to conditional), i.e. the average effect of the treatment at the population level. The MTE depends on the set of individuals in whom the average effect is computed, and two particular average effects are commonly estimated: the average treatment effect on the entire population under study (ATE), and the average treatment effect on the subjects who are treated (ATT). The choice between estimating ATE or ATT depends mainly on the study objective [4].

Four methods have been proposed account for the estimated PS [1, 5]: adjustment on the PS, stratification on the PS, matching on the PS, and weighting using the PS. Several authors have demonstrated using simulation studies that adjustment and stratification on the PS perform poorly for estimating marginal effects [6, 7, 8]. Matching and weighting methods have better performance [9, 10, 11] because of a more effective reduction of the imbalance in the distribution of observed characteristics between the exposed and unexposed subjects [12]. Despite their weakness, adjustment and stratification methods have been the most used methods [13, 14]. But recent literature reviews indicated that the matching method could be the most commonly used method today. The weighting method remains underexploited [15, 16, 17] despite its advantages in terms of performance, flexibility (it allows estimating the ATE or the ATT), and reporting (presentation of analyses and results is very similar to randomized trial [18]).

The present study is focused on the use of the PS weighting method to evaluate the marginal effect of a treatment on a binary outcome. Binary outcomes are common in the clinical reasearch area, and the treatment effect could be measured using the risk difference (RD), the relative risk (RR) or the odds ratio (OR). Once the treatment effect has been estimated, a particular care needs to be taken with the method of variance estimation, which should account for the weighting scheme as well as the fact that the PS is estimated, rather than known with certainty.

Several authors have focused on the variance estimator of the PS weighting estimator. Lunceford and Davidian [19] developed a large-sample marginal variance estimator usable for the estimation of the difference in means of a continuous response. This variance estimator has also been used in the case of a binary outcome when the treatment effect was measured with RD [9]. Neither RR nor OR have been considered. Williamson *et al.* [20] were also interested in the variance estimator of the PS weighting estimator, and have addressed the evaluation of RD, RR and OR.

Lunceford and Davidian and Williamson *et al.* variance formulas have many similarities [20]. But Austin found that confidence intervals based on the Lunceford and Davidian variance estimator were too conservative [9]. Moreover, the performance of the Williamson *et al.* approach was only evaluated in the case of individually randomized controlled trials, and not observational studies [20]. Finally, only ATE was addressed, the evaluation of ATT was not in the scope of any of these authors.

Variance estimation taken from a weighted regression model of the outcome on treatment with sandwich-type (a.k.a. robust) variance estimator [21, 22] has also been proposed for ATE or ATT estimation. To the best of our knowledge, no other method has been published for the variance estimation of the ATT. However, this variance estimation method does not account for the fact that the true PS is unknown and has been estimated, and may lead to too conservative confidence intervals [23].

In this study, we have focused on the estimation of the treatment effect in observational (and thus not randomized) studies that compare treated versus untreated subjects on a binary outcome, using RD, RR, or OR as measures of the treatment effect. We have described the previously published variance estimators and explained the key differences between them. We have proposed a unified approach to derive variance estimators suitable for ATE and ATT estimation, using the influence function linearization technique developed by Deville [24]. We have compared their performance with Lunceford and Davidian, Williamson *et al.* and robust variance estimators through an extensive simulation study. We have applied all these methods to a real case study aiming at estimating the effect of high inhaled corticosteroids to total asthma drug ratio on the occurence of asthma-related exacerbations.

2 Weighting using the propensity score to estimate the treatment effect on a binary outcome

Let T_i be an indicator variable denoting treatment status for a subject i ($T_i = 1$ for a treated subject, $T_i = 0$ otherwise), Y_i be an indicator variable of the binary event of interest ($Y_i = 1$ if subject has experienced the event, $Y_i = 0$ otherwise), and X_i be some vector of baseline observed covariates.

The PS is the probability of treatment given the observed covariates $Pr(T_i = 1|X_i) \equiv p_{Ti}$, and is commonly estimated from a logistic regression model: $\text{logit}(\hat{p}_{Ti}) = X_i^\top \widehat{\alpha}$.

We considered the inverse probability weighted (IPW) estimator [25]:

$$\widehat{P}_{11} = \frac{\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}}}.$$

Similarly, we considered the IPW estimator:

$$\widehat{P}_{10} = \frac{\sum_{i=1}^{n} \frac{1-T_i}{1-\widehat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} \frac{1-T_i}{1-\widehat{p}_{Ti}}}.$$

Unbiased estimates of the ATE using RD, RR or OR could be obtained with:

$$\begin{aligned}
\widehat{\Gamma}_{RD1} &= \widehat{RD}_1 = \widehat{P}_{11} - \widehat{P}_{10}, \\
\widehat{\Gamma}_{RR1} &= \log(\widehat{RR}_1) = \log(\widehat{P}_{11}) - \log(\widehat{P}_{10}), \\
\widehat{\Gamma}_{OR1} &= \log(\widehat{OR}_1) = \text{logit}(\widehat{P}_{11}) - \text{logit}(\widehat{P}_{10}).
\end{aligned}$$

To estimate ATT, we considered [25]:

$$\widehat{P}_{21} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i}$$

and

$$\widehat{P}_{20} = \frac{\sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}}.$$

Unbiased estimates of the ATT using RD, RR or OR could be obtained with:

$$\begin{aligned}
\widehat{\Gamma}_{RD2} &= \widehat{RD}_2 = \widehat{P}_{21} - \widehat{P}_{20}, \\
\widehat{\Gamma}_{RR2} &= \log(\widehat{RR}_2) = \log(\widehat{P}_{21}) - \log(\widehat{P}_{20}), \\
\widehat{\Gamma}_{OR2} &= \log(\widehat{OR}_2) = \text{logit}(\widehat{P}_{21}) - \text{logit}(\widehat{P}_{20}).
\end{aligned}$$

3 Treatment effect variance estimators

3.1 Model-based estimators (sandwich estimators)

Robins *et al.* [21] suggested to use the sandwich (a.k.a. robust) variance estimator obtained from a weighted regression model.

To retrieve these estimations, we used the `svyglm` function from the `R` package `survey`, and fitted three weighted generalized linear models of the outcome on treatment, using a *identity* link function (to estimate the RD), a log link function (to estimate the log(RR)), or a logit link function (to estimate the log(OR)). Two types of weight were used, to estimate the ATE (with $\widehat{w}_1$) or the ATT (with $\widehat{w}_2$), like below [25]:

$$\begin{aligned}
\widehat{w}_{1i} &= \frac{T_i}{\widehat{p}_{Ti}} + \frac{1-T_i}{1-\widehat{p}_{Ti}}, \\
\widehat{w}_{2i} &= \widehat{p}_{Ti} \times \widehat{w}_{1i} = T_i + (1-T_i)\frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}.
\end{aligned}$$

3.2 Lunceford and Davidian estimator

Following Lunceford and Davidian [19] notations, the variance estimator for $\widehat{\Gamma}_{RD1}$ is computed as $n^{-2}\sum_{i=1}^{n}\widehat{I}_i^2$, with:

$$\widehat{I}_i = \frac{T_i}{\widehat{p}_{Ti}}\left(Y_i - \widehat{P}_{11}\right) - \frac{1-T_i}{1-\widehat{p}_{Ti}}(Y_i - \widehat{P}_{10}) - (T_i - \widehat{p}_{Ti})\widehat{H}^\top \widehat{E}^{-1} X_i, \tag{1}$$

where

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i(Y_i - \widehat{P}_{11})T_i(1-\widehat{p}_{Ti})}{\widehat{p}_{Ti}}\right\} + \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i(Y_i - \widehat{P}_{10})(1-T_i)\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}\right\} \tag{2}$$

and

$$\widehat{E}^{-1} = n^{-1}\sum_{i=1}^{n}\widehat{p}_{Ti}(1-\widehat{p}_{Ti})X_i X_i^\top. \tag{3}$$

3.3 Williamson et al. estimators

Following Williamson *et al.* notations [20], variance estimators for $\widehat{\Gamma}_{RD1}$, $\widehat{\Gamma}_{RR1}$ and $\widehat{\Gamma}_{OR1}$ are computed as

$$n\widehat{\mathrm{Var}}(\widehat{\Gamma}_{k1}) = \widehat{V}_{un} - \widehat{\mathbf{v}}^\top(2\widehat{M}_1 - \widehat{M}_2)\widehat{\mathbf{v}}. \tag{4}$$

In the previous equation, k stands for RD, RR or OR. Letting $\widehat{W}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{p}_{Ti}}$, $\widehat{W}_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\widehat{p}_{Ti}}$, $\widehat{K}_{RD1} = \widehat{K}_{RD0} = 1$, $\widehat{K}_{RR1} = \widehat{P}_{11}^{-1}$, $\widehat{K}_{RR0} = \widehat{P}_{10}^{-1}$, $\widehat{K}_{OR1} = \{\widehat{P}_{11}(1-\widehat{P}_{11})\}^{-1}$, and $\widehat{K}_{OR0} = \{\widehat{P}_{10}(1-\widehat{P}_{10})\}^{-1}$, we have

$$\widehat{V}_{un} = \frac{\widehat{K}_{k1}^2}{\widehat{W}_1^{\,2}}\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{P}_{11})^2 T_i}{\widehat{p}_{Ti}^2} + \frac{\widehat{K}_{k0}^2}{\widehat{W}_0^{\,2}}\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \widehat{P}_{10})^2(1-T_i)}{(1-\widehat{p}_{Ti})^2}, \tag{5}$$

$$\widehat{\mathbf{v}} = \frac{\widehat{K}_{k1}}{\widehat{W}_1}\frac{1}{n}\sum_{i=1}^{n}\frac{X_i(Y_i - \widehat{P}_{11})T_i(1-\widehat{p}_{Ti})}{\widehat{p}_{Ti}} + \frac{\widehat{K}_{k0}}{\widehat{W}_0}\frac{1}{n}\sum_{i=1}^{n}\frac{X_i(Y_i - \widehat{P}_{10})(1-T_i)\widehat{p}_{Ti}}{(1-\widehat{p}_{Ti})}, \tag{6}$$

$$\widehat{M}_1 = \left(n^{-1}\sum_{i=1}^{n}\widehat{p}_{Ti}(1-\widehat{p}_{Ti})X_i X_i^\top\right)^{-1}, \tag{7}$$

and

$$\widehat{M}_2 = \widehat{M}_1\left(n^{-1}\sum_{i=1}^{n}X_i X_i^\top (T_i - \widehat{p}_{Tj})^2\right)\widehat{M}_1. \tag{8}$$

We can observe that $\widehat{M}_1 = \widehat{E}$ (Equation 3). Moreover, in the case of the risk difference, $\widehat{\mathbf{v}} = \widehat{H}$ (Equation 2) if $\widehat{W}_0$ and $\widehat{W}_1$ are replaced by their approximated expectation of 1 [20]. In fact, apart from that approximation, the development of the Lunceford and Davidian estimator $n^{-2}\sum_{i=1}^{n}\widehat{I}_i^2$ leads to the Williamson *et al.* estimator, except that $\widehat{E}$ is used in the Williamson *et al.* estimator, and $\widehat{E}^{-1}$ is used in the Lunceford and Davidian estimator.

3.4 Estimators based on estimated linearized variables

We have developed variance estimators, approximately unbiased, using the influence function linearization technique described by Deville [24]. Briefly, for an estimator $\widehat{\gamma}$, the linearization involves to find a variable U such that $E(n^{-1}s_U^2) \simeq V(\widehat{\gamma})$, where $s_U^2 = (n-1)^{-1}\sum_{i=1}^n (U_i - \bar{U})^2$ and $\bar{U} = n^{-1}\sum_{i=1}^n U_i$. The variable U usually depends on unknown parameters, which can be estimated from the sample to obtain an estimated linearized variable $\widehat{U}$. This leads to the variance estimator $\widehat{V}(\widehat{\gamma}) = n^{-1}s_{\widehat{U}}^2$.

We only report in this section the final expressions of the estimated linearized variables. Detailed calculation rules can be found in Deville [24], and the derivation for each linearized variable estimator is described in detail in the Web Appendix A.

An estimated linearized variables of $\widehat{\Gamma}_{RD1}$ is:

$$\begin{aligned}
\widehat{U}_i(\widehat{\Gamma}_{RD1}) &= \widehat{U}_i(\widehat{P}_{11}) - \widehat{U}_i(\widehat{P}_{10}) \\
&= \left\{ \frac{T_i}{\widehat{p}_{Ti}}(Y_i - \widehat{P}_{11}) - (T_i - \widehat{p}_{Ti})\widehat{\gamma}_{11}^\top X_i \right\} \\
&\quad - \left\{ \frac{1-T_i}{1-\widehat{p}_{Ti}}(Y_i - \widehat{P}_{10}) + (T_i - \widehat{p}_{Ti})\widehat{\gamma}_{10}^\top X_i \right\} \qquad (9)
\end{aligned}$$

with

$$\widehat{\gamma}_{11} = \left\{ \sum_{j=1}^n \widehat{p}_{Tj}(1-\widehat{p}_{Tj}) X_j X_j^\top \right\}^{-1} \sum_{j=1}^n T_j \frac{1-\widehat{p}_{Tj}}{\widehat{p}_{Tj}} X_j (Y_j - \widehat{P}_{11}), \qquad (10)$$

$$\widehat{\gamma}_{10} = \left\{ \sum_{j=1}^n \widehat{p}_{Tj}(1-\widehat{p}_{Tj}) X_j X_j^\top \right\}^{-1} \sum_{j=1}^n (1-T_j) \frac{\widehat{p}_{Tj}}{1-\widehat{p}_{Tj}} X_j (Y_j - \widehat{P}_{10}). \qquad (11)$$

Following the same notations used in Lunceford and Davidian [19], $\widehat{U}_i(\widehat{\Gamma}_{RD1})$ can be reformulated as:

$$\widehat{U}_i(\widehat{\Gamma}_{RD1}) = \frac{T_i}{\widehat{p}_{Ti}}\left(Y_i - \widehat{P}_{11}\right) - \frac{1-T_i}{1-\widehat{p}_{Ti}}(Y_i - \widehat{P}_{10}) - (T_i - \widehat{p}_{Ti})\widehat{H}^\top \widehat{E} X_i. \qquad (12)$$

The expressions of $\widehat{I}_i$ in Equation 1 and $\widehat{U}_i(\widehat{\Gamma}_{RD1})$ in Equation 12 are nearly identical, except for $\widehat{E}$ replacing $\widehat{E}^{-1}$. The latter was also the key difference between the Lunceford and Davidian and the Williamson *et al.* estimators. Thus, the linearized estimator $n^{-1}s_{\widehat{U}}^2$ should be approximately equal to the Williamson *et al.* estimator $\widehat{\mathrm{Var}}(\widehat{\Gamma}_{RD1})$ (Equation 4) if, again, we consider that $\widehat{W}_0$ and $\widehat{W}_1$ are equal to 1.

We also developed an estimated linearized variable for $\widehat{\Gamma}_{RD2}$:

$$\begin{aligned}\widehat{U}_i(\widehat{\Gamma}_{RD2}) &= \widehat{U}_i(\widehat{P}_{21}) - \widehat{U}_i(\widehat{P}_{20}) \\ &= \bar{T}^{-1}T_i(Y_i - \widehat{P}_{21}) \\ &\quad -\bar{n}_T^{-1}\left\{(1-T_i)\frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}(Y_i - \widehat{P}_{20}) + (T_i - \widehat{p}_{Ti})\widehat{\gamma}_{20}^{\top}X_i\frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}\right\}, \end{aligned} \quad (13)$$

with

$$\widehat{\gamma}_{20} = \left\{\sum_{j=1}^{n}\widehat{p}_{Tj}(1-\widehat{p}_{Tj})X_jX_j^{\top}\right\}^{-1}\sum_{j=1}^{n}(1-T_j)\frac{\widehat{p}_{Tj}}{1-\widehat{p}_{Tj}}X_j(Y_j - \widehat{P}_{20}) \quad (14)$$

and

$$\bar{n}_T = n^{-1}\sum_{i=1}^{n}(1-T_i)\frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}. \quad (15)$$

Finally, by noting $f(x) = \log(x)$ and $g(x) = \text{logit}(x)$, we obtained the expression of all other estimated linearized variables:

$$\begin{aligned}\widehat{U}_i(\widehat{\Gamma}_{RRm}) &= \widehat{U}_i(f(\widehat{P}_{m1})) - \widehat{U}_i(f(\widehat{P}_{m0})) \\ &= f'(\widehat{P}_{m1})\widehat{U}_i(\widehat{P}_{m1}) - f'(\widehat{P}_{m0})\widehat{U}_i(\widehat{P}_{m0}) \\ &= \frac{\widehat{U}_i(\widehat{P}_{m1})}{\widehat{P}_{m1}} - \frac{\widehat{U}_i(\widehat{P}_{m0})}{\widehat{P}_{m0}}, \end{aligned} \quad (16)$$

and

$$\begin{aligned}\widehat{U}_i(\widehat{\Gamma}_{ORm}) &= \widehat{U}_i(g(\widehat{P}_{m1})) - \widehat{U}_i(g(\widehat{P}_{m0})) \\ &= g'(\widehat{P}_{m1})\widehat{U}_i(\widehat{P}_{m1}) - g'(\widehat{P}_{m0})\widehat{U}_i(\widehat{P}_{m0}) \\ &= \frac{\widehat{U}_i(\widehat{P}_{m1})}{\widehat{P}_{m1}(1-\widehat{P}_{m1})} - \frac{\widehat{U}_i(\widehat{P}_{m0})}{\widehat{P}_{m0}(1-\widehat{P}_{m0})} \end{aligned} \quad (17)$$

for $m \in \{1, 2\}$.

4 Monte-Carlo simulations

4.1 Methods

We conducted Monte-Carlo simulations to evaluate the performance of the variance estimators described in the previous section. Our simulation framework and parameters were deliberately close to those used in Austin's studies [9, 26] which examined different aspects of PPS analysis.

First, we randomly generated 10 independent normally distributed ($N(0, 1)$) variables $X_1 \ldots X_{10}$ for n=10,000 subjects. The exposure allocation T was drawn from a Bernoulli

distribution $T \sim B(p_T)$, with

$$\begin{aligned} p_T &= \text{logit}^{-1}(\alpha_{0,T} \\ &\quad +\alpha_L X_1 + \alpha_L X_2 + \alpha_L X_3 \\ &\quad +\alpha_M X_4 + \alpha_M X_5 + \alpha_M X_6 \\ &\quad +\alpha_H X_7 + \alpha_H X_8 + \alpha_H X_9 + \alpha_{VH} X_{10}). \end{aligned}$$

A binary event was also generated for each subject, with a probability p_Y equal to

$$\begin{aligned} p_Y &= \text{logit}^{-1}(\alpha_{0,Y} + \gamma T \\ &\quad +\alpha_L X_1 + \alpha_L X_2 + \alpha_L X_3 \\ &\quad +\alpha_M X_4 + \alpha_M X_5 + \alpha_M X_6 \\ &\quad +\alpha_H X_7 + \alpha_H X_8 + \alpha_H X_9 + \alpha_{VH} X_{10}). \end{aligned}$$

In the previous equations, γ denotes the conditional log(OR) relating the treatment T to the outcome Y. The other regression coefficients were set as follows to reflect low, medium, high and very high effects: $\alpha_L = \log(1.1)$, $\alpha_M = \log(1.25)$, $\alpha_H = \log(1.5)$ and $\alpha_{VH} = \log(2)$. $\alpha_{0,T}$, $\alpha_{0,Y}$ and γ were set to values that induce the desired treatment prevalence π_T, event rate π_Y and marginal effect Γ (RD, RR, or OR, ATE or ATT) in the simulated sample. The process allowing to find these parameter values used a minimization approach and is described in detail in the Web Appendix B.

We allowed the following parameters to vary across simulations:

- the treatment prevalence: $\pi_T \in \{0.25, 0.50\}$;
- the event rate: $\pi_Y \in \{0.25, 0.50\}$;
- the marginal treatment effect (ATE or ATT). Eight increasing treatment effects were evaluated for each measurement:
  - risk difference : $\Gamma_{RD} \in \{-0.20, -0.15, -0.10, -0.05, -0.02, 0, 0.02, 0.05\}$;
  - relative risk: $\Gamma_{RR} \in \{\log(1/1.60), \log(1/1.3), \log(1/1.2), \log(1/1.1), \log(1/1.05), \log(1), \log(1.05), \log(1.1)\}$;
  - odds ratio: $\Gamma_{OR} \in \{\log(1/2.20), \log(1/1.8), \log(1/1.5), \log(1/1.25), \log(1/1.1), \log(1), \log(1.1), \log(1.25)\}$;
- the sample size: $n \in \{500, 1{,}000, 2{,}000, \ldots, 10{,}000\}$.

Compared to Austin [9, 26], the main differences for the choice of these parameters relied on the treatment prevalence (which was fixed to 25 per cent in Austin's studies), the event rate (which was fixed to 29 per cent if all subjects in the population were not exposed), the

treatment effects (which were only focused on risk difference, and fixed to 0, -0.02, -0.05, -0.10, and -0.15) and the sample size (which was fixed to $10,000$).

For each scenario, we used B=10,000 replicates to calculate the following performance criteria:

- Bias: $\frac{1}{B}\sum_{b=1}^{B}(\widehat{\Gamma}_b - \Gamma)$;
- Root mean square error (RMSE): $\sqrt{\frac{1}{B}\sum_{b=1}^{B}(\widehat{\Gamma}_b - \Gamma)^2}$;
- Variability ratio (VR): $\frac{\frac{1}{B}\sum_{b=1}^{B}\widehat{SE}(\widehat{\Gamma}_b)}{\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\Gamma}_b - \bar{\widehat{\gamma}})^2}}$, where $\widehat{SE}(\widehat{\Gamma}) = \sqrt{\widehat{V}(\widehat{\Gamma})}$ is the estimated standard error of exposure effect $\widehat{\Gamma}$;
- Coverage: proportion of times Γ is enclosed in the 95% confidence interval ;
- Power: proportion of times 0 is not enclosed in the 95% confidence interval.

4.2 Results

Table 1 and Figure 1 show the simulation results according to the theoretical treatment effect, with $\pi_T = 0.25$ and $\pi_Y = 0.50$.

Bias for the estimation of the treatment effect (Table 1) was limited in all scenarios. RMSE values were the largest when estimating ATE with OR, and the smallest when estimating ATT with RD.

In practice, Williamson *et al.* variance estimators of the ATE and the corresponding estimators based on estimated linearized variables gave almost equal estimates. For clarity, we only drew the linearized estimators on Figure 1 because the points corresponding to the Williamson *et al.* estimators would be perfectly superimposed.

Overall, the Lunceford and Davidian and the sandwich variance estimators overestimated the variability of the ATE and have suboptimal coverages (Figure 1). The Williamson *et al.* and the linearized variance estimators provided the best estimations of the ATE variability (variability ratios close to one) and coverages were close to the nominal value (Figure 1). Linearized variance estimators of the ATT had better performance than sandwich estimators and also provided good coverages. For conciseness, only the results extracted from a limited number of scenarios were described, but results for other combinations of treatment prevalence and event rate are available in the Web Figure 1. Overall, they were analogous.

Figure 2 compares the coverage and the power of all methods for increasing sample size, with $\Gamma_{RD} = -0.05$, $\Gamma_{RR} = \log(1/1.10)$ and $\Gamma_{OR} = \log(1/1.25)$. Again, the confidence intervals based on the Lunceford and Davidian and sandwich variance estimators were too conservative. The Williamson *et al.* and the linearized variance estimators were the most efficient, and provided very close results when focusing on ATE. The gain in power with these estimators compared to the sandwich estimators was systematic and could reach 9%.

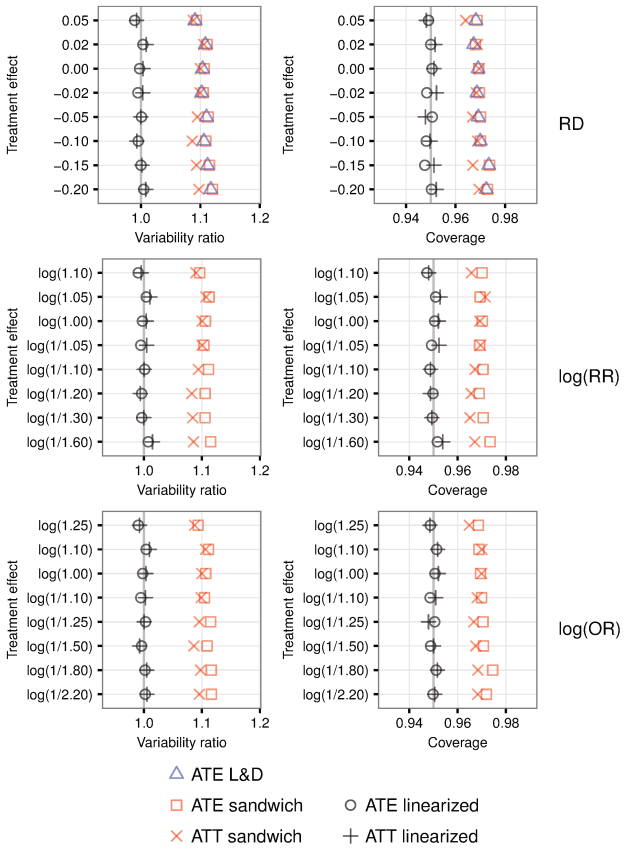


Figure 1: Variability ratio and coverage according to variance estimator and true risk difference, true (log) odds ratio and true (log) relative risk, with $\pi_T = 0.25$ and $\pi_Y = 0.50$. ATE L&D: estimation of the ATE variance using the Lunceford and Davidian estimator (section 3.2). ATE/ATT sandwich: estimation of the ATE/ATT variance using the sandwich estimators (section 3.1). ATE/ATT linearized: estimation of the ATE/ATT variance using the linearized estimators (section 3.4).

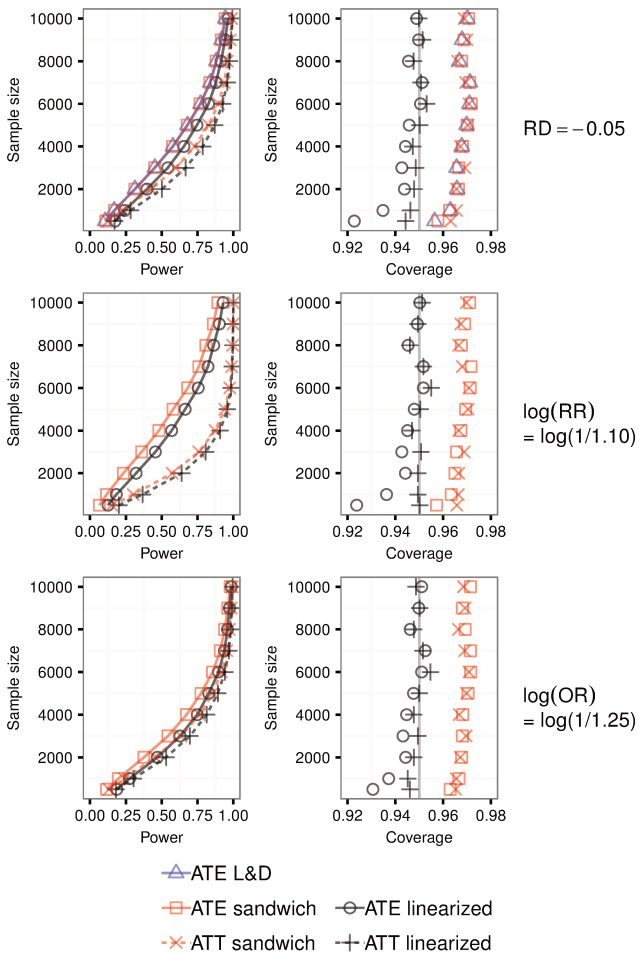


Figure 2: Coverage and power according to variance estimators, with $\pi_T = 0.25$ and $\pi_Y = 0.50$. ATE L&D: estimation of the ATE variance using Lunceford and davidian estimator (section 3.2). ATE/ATT sandwich: estimation of the ATE/ATT variance using sandwich estimators (section 3.1). ATE/ATT linearized: estimation of the ATE/ATT variance using linearized estimators (section 3.4).

A gain in power was observed for other combinations of treatment prevalence and event rate (Web Figures 2 to 5). Nevertheless, the linearized variance estimators of the ATE produced slightly unconservative confidence intervals for the smallest sample sizes, particularly with $n < 2{,}000$.

Linearized variance estimators of the ATT also provided a gain in power compared to sandwich estimators (Figure 2). They were less affected by the small sample size issue observed with the linearized variance estimators of the ATE, except when the treatment prevalence π_T was set at 0.50 (Web Figures 4 and 5).

5 Case study

5.1 Data source

Data were obtained from a (yet-unpublished) study evaluating the efficacy of personalized support of asthmatic patients, organized by the French public health assurance office (Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés, CNAMTS). We analyzed subjects from the control group only, who did not receive the personalized intervention. Data were extracted from the French health insurance database (SNIIRAM), which is the national claims database, linked to the French hospital discharge database (PMSI) containing individual records of all hospital stays.

Eligible subjects were 18 to 39 years old defined as having asthma because they filled at least 4 prescriptions for asthma-related medications during the year before inclusion. Demographics, reimbursements for health care expenditure, costly long-term disease (LTD) status, complementary universal health insurance (CMUc) status (indicates low socioeconomic level), and the general practitioner's city zip code were available for 31,332 subjects.

In this case study, the objective was to assess the impact of the asthma drug ratio (exposure of interest) on the risk of asthma exacerbation. Asthma drug ratio is the proportion of reimbursed units of inhaled corticosteroids (ICSs) to overall number of reimbursed respiratory medication units [27]. ICSs are the cornerstone therapy in persistent asthma to prevent exacerbations. Previous studies of persistent asthma patients showed significantly fewer asthma outcomes among patients with high ratios.

Two groups were defined by ICS ratio: $< 70\%$ (low ICS ratio group) and $\geq 70\%$ (high ICS ratio group). The outcome was the occurrence of asthma exacerbation within 1 year, defined as a filled prescription for oral corticosteroids within 7 days after a medical consultation (with a general practitioner or a pneumonologist) or a hospitalization for asthma exacerbation.

5.2 Results

All the statistical methods described in Sections 2 and 3 were applied. Approximatively 40% of the subjects were part of the high ICS-ratio group. Variables accounted for in the propensity analysis are described in Table 2. The estimated log ORs are reported in table 3. ORs had qualitatively similar values, which indicates the protective effect of a high ICS-ratio on asthma exacerbation.

Table 3 also reports the estimated standard errors. In accordance with our simulation results, the Williamson *et al.* and the linearized variance estimators provided similar standard errors that were smaller than those provided by sandwich estimators.

The surprisingly high value of the standard error based on the Lunceford and Davidian estimator was never observed in the simulation study. We hypothesized that this unusual

value was caused by the fact that no variable was standardized in this case study whereas all variables had a mean of zero and a standard error of one in the simulation study. To confirm this assumption, we reran some simulations in which all X_k variables had a mean of 10, and/or a standard error of 10 (Table 4). As expected, Lunceford and Davidian estimator was the only method highly influenced by the mean and the standard error of the covariates. Highest values were obtained when both the mean and the standard error of the covariates were increased.

6 Discussion

When evaluating a treatment effect in an observational study, applied researcher should carefully choose the type of estimation that answers the clinical objectives, in particular when treatment effect is estimated using a non-collapsible measure [28]. PS framework is popular for estimating marginal (as opposed to conditional) treatment effects. For this purpose, adjustment and stratification on the PS has been found less efficient than matching and weighting methods [29, 10]. Valid variance estimators are available when adjusting [30], stratifying [23] or matching [31] on the estimated PS are used. Weighting using the PS allows to estimate the ATE or the ATT, and is easily implementable using standard statistical softwares. But some of the previously published variance estimators of the weighted PS estimators appear suboptimal or limited to the ATE estimation.

In this study, we used the linearization technique [24] to develop new variance estimators suitable for evaluating ATE and ATT on a binary outcome using PS-weighting estimation of RD, RR or OR, and reported their performance in various settings. Unlike sandwich variance estimators, they were developped by accounting for the fact that the PS was estimated from the available data. Consequently, they provided more accurate estimations of the treatment effect variability, narrower confidence intervals and coverage rates closer to the nominal value. In the case of ATE estimation, the expression and the performance of the Williamson *et al.* estimators and the linearized estimators were almost identical. Finally, this study is the first providing the expression of a valid variance estimators of the ATT.

Lunceford and Davidian's derivation of their variance estimator was also attentive to the PS estimation step [19] but was only suitable for the evaluation of the ATE using RD. This estimator also appeared not fully efficient (as previously reported [9]) resulting in wider confidence intervals very similar to the ones based on sandwich variance estimators. Moreover, its performance became unacceptable when unstandardized variables were used to adjust for confounding, both in the simulations and in the case study. To our knowledge, this was never reported elsewhere.

Overall, the linearized variance estimators (which are almost equivalent to the Williamson *et al.* estimators when focusing only on the ATE) seem the best choice when the PS weighting

method is used to estimate the treatment effect on a binary outcome. Nevertheless, they tended to underestimate the variability of the treatment effect when the sample size was limited (in practice $n < 2,000$), particularly in the case of ATE estimation, and they should be used with caution in this setting or replaced by the sandwich variance estimators.

We have focused on binary outcomes which are frequent in the clinical research area. Time-to-event outcomes are also widely used. In this context, the treatment effect is commonly measured using the hazard ratio, but existing (sandwich) variance estimators appeared also too conservative in this context [10]. The use of a bootstrap-based estimator has been recently proposed and seems to accurately estimate the sampling variability [32]. But the extension of the linearized variance estimators proposed in the current study for time-to-event outcomes would be interesting and warrants a specific exploration.

In conclusion, we proposed a unified approach to derive new variance estimators of the treatment effect (on the overall or the treated population) on a binary outcome, measured using the RD, the RR or the OR, and estimated using PS-weighting approach. Provided that the number of subjects is large enough, they resulted in correct estimates of standard errors in different scenarios defined by the exposure prevalence, the event rate and the theoretical treatment effect. In the case of ATE estimation, their expression and their performance was similar to those of Williamson *et al.* estimators. They had better performance than sandwich estimators for the estimation of the ATT variability. These results allow valid large sample inference for estimators that use weighting on the estimated propensity score to estimate the treatment effect on the overall or the treated population.

Table 1: Bias and RMSE according to type and value of the theoretical treatment effect, with exposure prevalence $\pi_T = 0.25$ and event rate $\pi_Y = 0.50$.

True RD	Bias ×100	RMSE ×100	True log(RR)	Bias ×100	RMSE ×100	True log(OR)	Bias ×100	RMSE ×100
ATE			ATE			ATE		
-0.20	0.01	1.19	log(1/1.60)	-0.02	3.22	log(1/2.20)	0.00	5.16
-0.15	0.01	1.24	log(1/1.30)	0.01	3.02	log(1/1.80)	0.02	5.19
-0.10	0.04	1.29	log(1/1.20)	0.04	2.94	log(1/1.50)	0.13	5.28
-0.05	0.01	1.33	log(1/1.10)	0.00	2.84	log(1/1.25)	0.05	5.33
-0.02	0.02	1.36	log(1/1.05)	0.00	2.81	log(1/1.10)	0.06	5.46
0.00	0.01	1.38	log(1.00)	0.00	2.76	log(1.00)	0.06	5.51
0.02	0.00	1.38	log(1.05)	-0.02	2.69	log(1.10)	0.01	5.55
0.05	-0.01	1.43	log(1.10)	-0.04	2.69	log(1.25)	-0.03	5.75
ATT			ATT			ATT		
-0.20	0.02	1.13	log(1/1.60)	0.01	2.28	log(1/2.20)	0.04	4.83
-0.15	-0.02	1.14	log(1/1.30)	-0.05	2.03	log(1/1.80)	-0.12	4.83
-0.10	0.03	1.15	log(1/1.20)	0.03	1.94	log(1/1.50)	0.10	4.90
-0.05	0.02	1.14	log(1/1.10)	0.02	1.84	log(1/1.25)	0.07	4.90
-0.02	0.01	1.13	log(1/1.05)	0.02	1.79	log(1/1.10)	0.05	4.92
0.00	0.01	1.13	log(1.00)	0.01	1.74	log(1.00)	0.05	4.94
0.02	-0.01	1.12	log(1.05)	-0.02	1.70	log(1.10)	-0.05	4.95
0.05	-0.01	1.13	log(1.10)	-0.02	1.69	log(1.25)	-0.03	5.09

Table 2: Characteristics of patients in the case study.

	Median [Q25-Q75] or N (%)		
	ICS-ratio > 70	*ICS-ratio ≥ 70*	Overall
	N = 18850	N = 12482	N = 31332
Age	32 [27-36]	32 [27-36]	32 [27-36]
Male	8369 (44)	5019 (40)	13388 (43)
Long term disease status for asthma			
	2155 (11)	800 (6)	2955 (9)
Complementary Universal Health Insurance status			
	4602 (24)	1849 (15)	6451 (21)
Pneumonologist consultation in the past year			
	2860 (15)	2394 (19)	5254 (17)
Hospitalization for asthma in the past year			
	208 (1)	21 (0)	229 (1)
Number of asthma exacerbation in the past year			
0	9020 (48)	7139 (57)	16159 (52)
1	4382 (23)	2735 (22)	7117 (23)
2-3	3677 (20)	1954 (16)	5631 (18)
≥ 4	1771 (9)	654 (5)	2425 (8)
Quintiles of the social deprivation index			
1^{st}	2270 (12)	1776 (14)	4046 (13)
2^{nd}	3307 (18)	2337 (19)	5644 (18)
3^{rd}	3764 (20)	2561 (20)	6325 (20)
4^{th}	3552 (19)	2266 (18)	5818 (19)
5^{th}	5957 (32)	3542 (28)	9499 (30)
Physician's type of area			
Rural	1991 (11)	1430 (11)	3421 (11)
Urban	16859 (89)	11052 (89)	27911 (89)
Medical density in the municipality			
	206 [129-296]	199 [126-296]	204 [127-296]
Asthma exacerbation within one year (primary outcome)			
	8615 (46)	5005 (40)	13620 (43)

Table 3: Estimated treatment effects in the case study.

	RD	log(RR)	log(OR)
ATE			
Effect	-0,0215	-0,0498	-0,0877
Standard error			
L&D	5169,4586	–	–
Williamson	0,0056	0,0130	0,0228
Sandwich	0,0059	0,0138	0,0242
Linearized	0,0056	0,0130	0,0228
ATT			
Effect	-0,0219	-0,0532	-0,0905
Standard error			
L&D	–	–	–
Williamson	–	–	–
Sandwich	0,0058	0,0142	0,0241
Linearized	0,0056	0,0136	0,0230

Table 4: Standard error of the treatment effect (ATE using RD) according to covariates distribution, with $\pi_T = 0.25$, $\pi_Y = 0.50$ and $\Gamma_{RD} = -0.05$.

	$N(0,1)$	$N(10,1)$	$N(0,10)$	$N(10,10)$
Empirical	0,0134	0,0134	0,0134	0,0134
L&D	0,0148	3,4262	1,0834	53,9133
Williamson	0,0133	0,0133	0,0133	0,0133
Sandwich	0,0148	0,0148	0,0148	0,0148
Linearized	0,0133	0,0133	0,0133	0,0133

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* Jan 1983; **70**(1):41–55, doi: 10.1093/biomet/70.1.41.

[2] Austin PC. Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal. Biometrische Zeitschrift* Feb 2009; **51**(1):171–184, doi: 10.1002/bimj.200810488.

[3] 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* Feb 2007; **26**(4):754–768, doi: 10.1002/sim.2618.

[4] Austin PC. Different measures of treatment effect for different research questions. *Journal of Clinical Epidemiology* Jan 2010; **63**(1):9–10, doi:10.1016/j.jclinepi.2009.07.006.

[5] Rosenbaum PR. Model-Based Direct Adjustment. *Journal of the American Statistical Association* 1987; **82**(398):387–394, doi:10.2307/2289440.

[6] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* Jul 2007; **26**(16):3078–3094, doi: 10.1002/sim.2781.

[7] Forbes A, Shortreed S. Inverse probability weighted estimation of the marginal odds ratio: Correspondence regarding 'The performance of different propensity score methods for estimating marginal odds ratios' by P. Austin, Statictics in Medicine, 2007; 26:3078–3094. *Statistics in Medicine* Nov 2008; **27**(26):5556–5559, doi: 10.1002/sim.3362.

[8] Graf E, Schumacher M. Comments on 'The performance of different propensity score methods for estimating marginal odds ratios' by Peter C. Austin, Statistics in Medicine 2007; 26(16):3078–3094. *Statistics in Medicine* Aug 2008; **27**(19):3915–3917, doi: 10.1002/sim.3271.

[9] 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* Sep 2010; **29**(20):2137–2148, doi:10.1002/sim.3854.

[10] Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine* Jul 2013; **32**(16):2837–2849, doi: 10.1002/sim.5705.

[11] 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* Jan 2014; :0962280213519716doi: 10.1177/0962280213519716.

[12] 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: An International Journal of the Society for Medical Decision Making* 2009; **29**(6):661–677, doi:10.1177/0272989X09341755.

[13] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *Journal of Clinical Epidemiology* Jun 2005; **58**(6):550–559, doi: 10.1016/j.jclinepi.2004.10.016.

[14] Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, Rassen JA, Trikalinos TA, Kitsios GD. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *European Heart Journal* Aug 2012; **33**(15):1893–1901, doi: 10.1093/eurheartj/ehs114.

[15] Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary JY, Porcher R. Propensity scores in intensive care and anaesthesiology literature: A systematic review. *Intensive Care Medicine* Dec 2010; **36**(12):1993–2003, doi:10.1007/s00134-010-1991-5.

[16] Thoemmes FJ, Kim ES. A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research* Feb 2011; **46**(1):90–118, doi: 10.1080/00273171.2011.540475.

[17] Ali MS, Groenwold RHH, Belitser SV, Pestman WR, Hoes AW, Roes KCB, de Boer A, Klungel OH. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: A systematic review. *Journal of Clinical Epidemiology* Feb 2015; **68**(2):112–121, doi:10.1016/j.jclinepi.2014.08.011.

[18] Deb S, Austin PC, Tu JV, Ko DT, Mazer CD, Kiss A, Fremes SE. A Review of Propensity-Score Methods and Their Use in Cardiovascular Research. *Canadian Journal of Cardiology* Feb 2016; **32**(2):259–265, doi:10.1016/j.cjca.2015.05.015.

[19] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* Oct 2004; **23**(19):2937–2960, doi:10.1002/sim.1903.

[20] Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* Feb 2014; **33**(5):721–737, doi:10.1002/sim.5991. Bibtex:.

[21] Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* Sep 2000; **11**(5):550–560, doi:10.2307/3703997. ArticleType: research-article / Full publication date: Sep., 2000 / Copyright © 2000 Lippincott Williams & Wilkins.

[22] Resche-Rigon M, Pirracchio R, Robin M, Latour RPD, Sibon D, Ades L, Ribaud P, Fermand JP, Thieblemont C, Socié G, *et al.*. Estimating the treatment effect from non-randomized studies: The example of reduced intensity conditioning allogeneic stem cell transplantation in hematological diseases. *BMC Hematology* Aug 2012; **12**(1):10, doi:10.1186/1471-2326-12-10.

[23] Williamson EJ, Morley R, Lucas A, Carpenter JR. Variance estimation for stratified propensity score estimators. *Statistics in Medicine* Jul 2012; **31**(15):1617–1632, doi:10.1002/sim.4504.

[24] Deville JC. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey methodology* 1999; **25**(2):193–204.

[25] Imbens G. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics* 2004; .

[26] Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* Mar 2011; **10**(2):150–161, doi:10.1002/pst.433.

[27] Laforest L, Licaj I, Devouassoux G, Chatte G, Martin J, Van Ganse E. Asthma drug ratios and exacerbations: Claims data from universal health coverage systems. *The European Respiratory Journal* May 2014; **43**(5):1378–1386, doi:10.1183/09031936.00100113.

[28] Greenland S. Interpretation and Choice of Effect Measures in Epidemiologic Analyses. *American Journal of Epidemiology* Jan 1987; **125**(5):761–768.

[29] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research* May 2011; **46**(3):399–424, doi:10.1080/00273171.2011.568786.

[30] Zou B, Zou F, Shuster JJ, Tighe PJ, Koch GG, Zhou H. On variance estimate for covariate adjustment by propensity score analysis. *Statistics in Medicine* Jan 2016; :n/a–n/adoi:10.1002/sim.6943.

[31] Abadie A, Imbens GW. Matching on the Estimated Propensity Score. *National Bureau of Economic Research Working Paper Series* 2009; **No. 15301**, doi:10.3386/w15301.

[32] Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine* Jan 2016; :n/a–n/adoi: 10.1002/sim.7084.

Web-Based Supplementary Materials for "Variance estimation for weighted propensity score estimators"

David Hajage, Guillaume Chauvet, Florence Tubach, Yann De Rycke

Web Appendix A Derivation of the new variance estimators

A.1 Preliminary results

The probability p_{Ti} is estimated by fitting a logistic regression model. We note $\widehat{\alpha}$ the estimator of α (vector of the logistic model coefficients), and $\widehat{p}_{Ti}$ the estimated probability of treatment, estimated with:

$$\widehat{p}_{Ti} = \frac{\exp(X_i^\intercal \widehat{\alpha})}{1 + \exp(X_i^\intercal \widehat{\alpha})}.$$

Using a Taylor expansion, we obtain:

$$\begin{aligned}
\widehat{p}_{Ti} &= \{1 + \exp(-X_i^\intercal \widehat{\alpha}\}^{-1} \\
&= \{1 + \exp(-X_i^\intercal \alpha)\exp\{-X_i^\intercal(\widehat{\alpha} - \alpha)\}\}^{-1} \\
&\approx \{1 + \exp(-X_i^\intercal \alpha)\{1 - X_i^\intercal(\widehat{\alpha} - \alpha)\}\}^{-1} \\
&= \{1 + \exp(-X_i^\intercal \alpha)\}^{-1} \left\{1 - \frac{X_i^\intercal \exp(-X_i^\intercal \alpha)(\widehat{\alpha} - \alpha)}{1 + \exp(-X_i^\intercal \alpha)}\right\}^{-1} \\
&\approx p_{Ti}\left\{1 + \frac{X_i^\intercal \exp(-X_i^\intercal \alpha)(\widehat{\alpha} - \alpha)}{1 + \exp(-X_i^\intercal \alpha)}\right\} \\
&= p_{Ti}\left\{1 + (1 - p_{Ti})X_i^\intercal(\widehat{\alpha} - \alpha)\right\} \\
\Rightarrow \widehat{p}_{Ti} - p_{Ti} &\approx \widehat{p}_{Ti}(1 - \widehat{p}_{Ti})X_i^\intercal(\widehat{\alpha} - \alpha). \qquad (1)
\end{aligned}$$

The normal equation for the logistic regression, $\sum_{j=1}^{n}(\widehat{p}_{Tj} - T_j)X_j = 0$, can be written as:

$$\sum_{j=1}^{n}(\widehat{p}_{Tj} - p_{Tj})X_j = \sum_{j=1}^{n}(T_j - p_{Tj})X_j. \qquad (2)$$

By plugging (1) into (2), we obtain:

$$
\begin{aligned}
\sum_{j=1}^{n} \widehat{p}_{Tj}(1-\widehat{p}_{Tj}) X_j^\intercal (\widehat{\alpha}-\alpha) X_j &\approx \sum_{j=1}^{n} (T_j - p_{Tj}) X_j \\
\left\{ \sum_{j=1}^{n} \widehat{p}_{Tj}(1-\widehat{p}_{Tj}) X_j X_j^\intercal \right\} (\widehat{\alpha}-\alpha) &\approx \sum_{j=1}^{n} (T_j - p_{Tj}) X_j \\
\widehat{\alpha}-\alpha &\approx \left\{ \sum_{j=1}^{n} \widehat{p}_{Tj}(1-\widehat{p}_{Tj}) X_j X_j^\intercal \right\}^{-1} \sum_{j=1}^{n} (T_j - p_{Tj}) X_j. \quad (3)
\end{aligned}
$$

We obtain the approximation

$$
\widehat{\alpha}-\alpha \simeq A^{-1} \sum_{j=1}^{n} (T_j - p_{Tj}) X_j \quad \text{where} \quad A = \sum_{j=1}^{n} p_{Tj}(1-p_{Tj}) X_j X_j^\top. \tag{4}
$$

From (4), we obtain successively

$$
\begin{aligned}
\frac{1}{\widehat{p}_{Ti}} - \frac{1}{p_{Ti}} &\simeq -\left(\frac{1}{p_{Ti}} - 1 \right) X_i^\top A^{-1} \sum_{j=1}^{n} (T_j - p_{Tj}) X_j, \\
\frac{1}{1-\widehat{p}_{Ti}} - \frac{1}{1-p_{Ti}} &\simeq \left(\frac{p_{Ti}}{1-p_{Ti}} \right) X_i^\top A^{-1} \sum_{j=1}^{n} (T_j - p_{Tj}) X_j, \quad (5) \\
\frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}} - \frac{p_{Ti}}{1-p_{Ti}} &\simeq \left(\frac{p_{Ti}}{1-p_{Ti}} \right) X_i^\top A^{-1} \sum_{j=1}^{n} (T_j - p_{Tj}) X_j.
\end{aligned}
$$

A.2 First estimator: $\widehat{\Gamma}_{RD1}$

The marginal risk difference (RD) in the overall population is estimated by:

$$
\widehat{\Gamma}_{RD1} = RD_1 = \widehat{P}_{11} - \widehat{P}_{10} \quad \text{for} \quad \Gamma = \mathbb{P}_{11} - \mathbb{P}_{10}. \tag{6}
$$

where

$$
\begin{aligned}
\widehat{P}_{11} &= \frac{\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}}}, \quad (7) \\
\widehat{P}_{10} &= \frac{\sum_{i=1}^{n} \frac{1-T_i}{1-\widehat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} \frac{1-T_i}{1-\widehat{p}_{Ti}}}. \quad (8)
\end{aligned}
$$

We can write

$$
\widehat{\Gamma}_{RD1} - \Gamma = \left\{ \widehat{P}_{11} - \mathbb{P}_{11} \right\} - \left\{ \widehat{P}_{10} - \mathbb{P}_{10} \right\}. \tag{9}
$$

Since $\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}} \simeq n$, we have

$$
\widehat{P}_{11} - \mathbb{P}_{11} = \frac{\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}} (Y_i - \mathbb{P}_{11})}{\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}}} \simeq \frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}} (Y_i - \mathbb{P}_{11}). \tag{10}
$$

By using equation (5), we obtain

$$\begin{aligned}
\sum_{i=1}^{n} \frac{T_i}{\widehat{p}_{Ti}}(Y_i - \mathbb{P}_{11}) &\simeq \sum_{i=1}^{n} \frac{T_i}{p_{Ti}}(Y_i - \mathbb{P}_{11}) + \sum_{i=1}^{n} T_i \left(\frac{1}{\widehat{p}_{Ti}} - \frac{1}{p_{Ti}}\right)(Y_i - \mathbb{P}_{11}) \\
&\simeq \sum_{i=1}^{n} \frac{T_i}{p_{Ti}}(Y_i - \mathbb{P}_{11}) - \left\{\sum_{i=1}^{n} T_i \left(\frac{1}{p_{Ti}} - 1\right) X_i^\top (Y_i - \mathbb{P}_{11})\right\} A^{-1} \sum_{j=1}^{n} (T_j - p_{Tj}) X_j \\
&\simeq \sum_{i=1}^{n} \left(\frac{T_i}{p_{Ti}}(Y_i - \mathbb{P}_{11}) - (T_i - p_{Ti})\gamma_{11}^\top X_i\right) \qquad (11)
\end{aligned}$$

with

$$\gamma_{11} = \left\{\sum_{j=1}^{n} p_{Tj}(1 - p_{Tj}) X_j X_j^\top\right\}^{-1} \sum_{j=1}^{n} T_j \left(\frac{1}{p_{Tj}} - 1\right) X_j (Y_j - \mathbb{P}_{11}). \qquad (12)$$

We obtain a linearized variable

$$U_i(\widehat{P}_{11}) = \frac{T_i}{p_{Ti}}(Y_i - \mathbb{P}_{11}) - (T_i - p_{Ti})\gamma_{11}^\top X_i. \qquad (13)$$

An estimated linearized variable is

$$\widehat{U}_i(\widehat{P}_{11}) = \frac{T_i}{\widehat{p}_{Ti}}(Y_i - \widehat{P}_{11}) - (T_i - \widehat{p}_{Ti})\widehat{\gamma}_{11}^\top X_i \qquad (14)$$

with

$$\widehat{\gamma}_{11} = \left\{\sum_{j=1}^{n} \widehat{p}_{Tj}(1 - \widehat{p}_{Tj}) X_j X_j^\top\right\}^{-1} \sum_{j=1}^{n} T_j \left(\frac{1}{\widehat{p}_{Tj}} - 1\right) X_j (Y_j - \widehat{P}_{11}). \qquad (15)$$

By symmetry, an estimated linearized variable of $\widehat{P}_{10}$ is

$$\widehat{U}(\widehat{P}_{10}) = \frac{1 - T_i}{1 - \widehat{p}_{Ti}}(Y_i - \widehat{P}_{10}) + (T_i - \widehat{p}_{Ti})\widehat{\gamma}_{10}^\top X_i \qquad (16)$$

with

$$\widehat{\gamma}_{10} = \left\{\sum_{j=1}^{n} \widehat{p}_{Tj}(1 - \widehat{p}_{Tj}) X_j X_j^\top\right\}^{-1} \sum_{j=1}^{n} (1 - T_j) \frac{\widehat{p}_{Tj}}{1 - \widehat{p}_{Tj}} X_j (Y_j - \widehat{P}_{10}). \qquad (17)$$

Thus, an estimated linearized variable for $\widehat{\Gamma}_{RD1}$ is

$$\begin{aligned}
\widehat{U}_i(\widehat{\Gamma}_{RD1}) &= \widehat{U}_i(\widehat{P}_{11}) - \widehat{U}_i(\widehat{P}_{10}) \qquad (18) \\
&= \left\{\frac{T_i}{\widehat{p}_{Ti}}(Y_i - \widehat{P}_{11}) - (T_i - \widehat{p}_{Ti})\widehat{\gamma}_{11}^\top X_i\right\} \\
&- \left\{\frac{1 - T_i}{1 - \widehat{p}_{Ti}}(Y_i - \widehat{P}_{10}) + (T_i - \widehat{p}_{Ti})\widehat{\gamma}_{10}^\top X_i\right\}. \qquad (19)
\end{aligned}$$

A.3 Second estimator: $\widehat{\gamma}_{RD2}$

We now consider the estimator

$$\widehat{\Gamma}_{RD2} = \widehat{P}_{21} - \widehat{P}_{20} \quad \text{for} \quad \Gamma = \mathbb{P}_{21} - \mathbb{P}_{20}, \tag{20}$$

with

$$\widehat{P}_{21} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} \tag{21}$$

and

$$\widehat{P}_{20} = \frac{\sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}} Y_i}{\sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}}. \tag{22}$$

We can write

$$\widehat{\Gamma}_{RD2} - \Gamma \quad = \quad \left\{ \widehat{P}_{21} - \mathbb{P}_{21} \right\} - \left\{ \widehat{P}_{20} - \mathbb{P}_{20} \right\}. \tag{23}$$

We have

$$\widehat{P}_{21} - \mathbb{P}_{21} = \frac{\sum_{i=1}^{n} T_i (Y_i - \mathbb{P}_{21})}{\sum_{i=1}^{n} T_i} = \frac{1}{n\bar{T}} \sum_{i=1}^{n} T_i (Y_i - \mathbb{P}_{21}), \tag{24}$$

with $\bar{T} = n^{-1} \sum_{i=1}^{n} T_i$.
A linearized variable of $\widehat{P}_{21}$ is

$$U(\widehat{P}_{21}) = \frac{T_i (Y_i - \mathbb{P}_{21})}{\bar{T}}. \tag{25}$$

An estimated linearized variable of $\widehat{P}_{21}$ is

$$\widehat{U}(\widehat{P}_{21}) = \frac{T_i (Y_i - \widehat{P}_{21})}{\bar{T}}. \tag{26}$$

With the same reasoning as in Section A.2, we obtain successively

$$\widehat{P}_{20} - \mathbb{P}_{20} \quad \simeq \quad \frac{1}{n \times \bar{n}_T} \sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}} (Y_i - \mathbb{P}_{20}), \tag{27}$$

$$\sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}} (Y_i - \mathbb{P}_{20}) \quad \simeq \quad \sum_{i=1}^{n} \left\{ (1-T_i) \frac{p_{Ti}}{1-p_{Ti}} (Y_i - \mathbb{P}_{20}) + (T_i - p_{Ti}) \gamma_{20}^{\top} X_i \right\} \tag{28}$$

with

$$\bar{n}_T \quad = \quad \frac{1}{n} \sum_{i=1}^{n} (1-T_i) \frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}, \tag{29}$$

$$\gamma_{20} \quad = \quad \left\{ \sum_{j=1}^{n} p_{Tj} (1-p_{Tj}) X_j X_j^{\top} \right\}^{-1} \sum_{j=1}^{n} (1-T_j) \frac{p_{Tj}}{1-p_{Tj}} X_j (Y_j - \mathbb{P}_{20}). \tag{30}$$

This leads to

$$\widehat{P}_{20} - \mathbb{P}_{20} \simeq \frac{1}{n\ \bar{n}_T} \sum_{i=1}^{n} \left\{ (1-T_i) \frac{p_{Ti}}{1-p_{Ti}} (Y_i - \mathbb{P}_{20}) + (T_i - p_{Ti}) \gamma_{20}^{\top} X_i \right\}. \tag{31}$$

We obtain that a linearized variable for $\widehat{P}_{20}$ is

$$U_i(\widehat{P}_{20}) = \frac{1}{\bar{n}_T}\left\{(1-T_i)\frac{p_{Ti}}{1-p_{Ti}}(Y_i-\mathbb{P}_{20})+(T_i-p_{Ti})\gamma_{20}^{\top}X_i\right\}, \quad (32)$$

and an estimated linearized variable is

$$\widehat{U}_i(\widehat{P}_{20}) = \frac{1}{\bar{n}_T}\left\{(1-T_i)\frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}(Y_i-\widehat{P}_{20})+(T_i-\widehat{p}_{Ti})\widehat{\gamma}_{20}^{\top}X_i\right\}, \quad (33)$$

with

$$\widehat{\gamma}_{20} = \left\{\sum_{j=1}^{n}\widehat{p}_{Tj}(1-\widehat{p}_{Tj})X_jX_j^{\top}\right\}^{-1}\sum_{j=1}^{n}(1-T_j)\frac{\widehat{p}_{Tj}}{1-\widehat{p}_{Tj}}X_j(Y_j-\widehat{P}_{20}). \quad (34)$$

Thus, an estimated linearized variable for $\widehat{\Gamma}_{RD2}$ is

$$\begin{aligned}\widehat{U}_i(\widehat{\Gamma}_{RD2}) &= \widehat{U}_i(\widehat{P}_{21})-\widehat{U}_i(\widehat{P}_{20}) \quad (35)\\ &= \left\{\frac{T_i(Y_i-\widehat{P}_{21})}{\bar{T}}\right\}\\ &- \left\{\frac{1}{\bar{n}_T}\left((1-T_i)\frac{\widehat{p}_{Ti}}{1-\widehat{p}_{Ti}}(Y_i-\widehat{P}_{20})+(T_i-\widehat{p}_{Ti})\widehat{\gamma}_{20}^{\top}X_i\right)\right\}. \quad (36)\end{aligned}$$

Web Appendix B Data generating process

B.1 Simulation parameters

We used a data-generating process similar to the one found in Austin *et al.* studies to examine different aspects of propensity score analysis [1, 2].

First, we randomly generated 10 independent normally distributed ($N(0,1)$) variables $X_1 \dots X_{10}$ for n=10,000 subjects. The exposure allocation T was drawn from a Bernoulli distribution $T \sim B(p_T)$, with

$$\begin{aligned}p_T = \text{logit}^{-1}(&\alpha_{0,T} \quad (37)\\ &+\alpha_L X_1+\alpha_L X_2+\alpha_L X_3\\ &+\alpha_M X_4+\alpha_M X_5+\alpha_M X_6\\ &+\alpha_H X_7+\alpha_H X_8+\alpha_H X_9+\alpha_{VH}X_{10}).\end{aligned}$$

A binary event was also generated for each subject, with a probability p_Y equal to

$$\begin{aligned}p_Y = \text{logit}^{-1}(&\alpha_{0,Y}+\gamma T \quad (38)\\ &+\alpha_L X_1+\alpha_L X_2+\alpha_L X_3\\ &+\alpha_M X_4+\alpha_M X_5+\alpha_M X_6\\ &+\alpha_H X_7+\alpha_H X_8+\alpha_H X_9+\alpha_{VH}X_{10}).\end{aligned}$$

In the previous equation, γ denotes the conditional log odds ratio relating the treatment T

to the outcome Y. The other regression coefficients were set as follows to reflect low, medium, high and very high effects: $\alpha_L = \log(1.1)$, $\alpha_M = \log(1.25)$, $\alpha_H = \log(1.5)$ and $\alpha_{VH} = \log(2)$.

$\alpha_{0,T}$, $\alpha_{0,Y}$ and γ were set to values that induce the desired treatment prevalence π_T, event rate π_Y and marginal effect Γ (RD, RR, or OR, ATE or ATT) in the simulated sample. These three parameters are mutually dependent, and we used an iterative process to determine the values of $\alpha_{0,T}$, $\alpha_{0,Y}$, and γ that induce desired π_T, π_Y and Γ. First, we simulated n=10,000 subjects, and computed the individual probabilities of being exposed ($\tilde{p}_{T,i}$) with equation (37). The average of these individual probabilities is the expected exposure prevalence $\tilde{\pi}_T = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{T,i}$ in the simulated sample. Similarly, we computed the individual probabilities of event ($\tilde{p}_{Y,i}$) with equation (38), and the corresponding average, which is the expected event rate $\tilde{\pi}_Y = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Y,i}$ in the sample.

We also computed the average probability of event first assuming that all subjects were untreated ($\tilde{\pi}_{Y,0} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Y_0,i}$) and then assuming that all subjects were treated ($\tilde{\pi}_{Y,1} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{Y_1,i}$). The difference between these two average probabilities is the expected risk difference in the overall population, $\tilde{\Gamma}_{1,RD} = \tilde{\pi}_{Y,1} - \tilde{\pi}_{Y,0}$, in the sample. We computed the same average probabilities weighted by individual probabilities of being exposed ($\tilde{\pi}'_{Y,0} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{T,i}\tilde{p}_{Y_0,i}$ and $\tilde{\pi}'_{Y,1} = \frac{1}{n}\sum_{i=1}^{n}\tilde{p}_{T,i}\tilde{p}_{Y_1,i}$). The difference between these two weighted average probabilities is the expected risk difference in the treated population, $\tilde{\Gamma}_{2,RD} = \tilde{\pi}'_{Y,1} - \tilde{\pi}'_{Y,0}$, in the sample.

Using an iterative process, one could successively modify $\alpha_{0,T}$, $\alpha_{0,Y}$ and γ until the expected treatment prevalence, the expected event rate and the expected marginal risk difference are arbitrarily close to the desired value in the simulated cohort. This process was performed by minimizing:

- the quantity $(\pi_T - \tilde{\pi}_T)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{1,RD} - \tilde{\Gamma}_{1,RD})^2$ to obtain the parameters $\alpha_{0,T}$, $\alpha_{0,Y}$ and γ that induced the desired exposure prevalence, event rate, and risk difference in the overall population (ATE);
- and the quantity $(\pi_T - \tilde{\pi}_T)^2 + (\pi_E - \tilde{\pi}_E)^2 + (\Gamma_{2,RD} - \tilde{\Gamma}_{2,RD})^2$ to obtain the parameters $\alpha_{0,T}$, $\alpha_{0,Y}$ and γ that induced the desired exposure prevalence, event rate, and risk difference in the treated population (ATT).

To increase precision, this minimization process was repeated in 1,000 simulated samples, to obtain 1000 sets of parameters $\alpha_{0,T}$, $\alpha_{0,Y}$ and γ for ATE and ATT risk differences. These 1000 estimations were averaged to obtain the final parameters used in the simulation study.

The parameters suitable for relative risks and odds ratios were obtained using a similar approach, replacing $\tilde{\Gamma}_{1,RR}$ and $\tilde{\Gamma}_{2,RR}$ by $\tilde{\Gamma}_{1,RR} = \log(\tilde{\pi}_{Y,1}) - \log(\tilde{\pi}_{Y,0})$ and $\tilde{\Gamma}_{2,RR} = \log(\tilde{\pi}'_{Y,1}) - \log(\tilde{\pi}'_{Y,0})$, or by $\tilde{\Gamma}_{1,OR} = \text{logit}(\tilde{\pi}_{Y,1}) - \text{logit}(\tilde{\pi}_{Y,0})$ and $\tilde{\Gamma}_{2,OR} = \text{logit}(\tilde{\pi}'_{Y,1}) - \text{logit}(\tilde{\pi}'_{Y,0})$.

One can notice (with equation 38) that the probability of treatment depends on only the subjects characteristics. Thus, for a given desired treatment prevalence, all the parameters $\alpha_{0,T}$ obtained with the previously described minimization process were approximatively equal, whatever the desired event rate and treatment effect. Consequently, $\alpha_{0,T}$ was considered unique for a given treatment prevalence (i.e. the values obtained for the same value of treatment prevalence were averaged).

B.2 Datasets generation

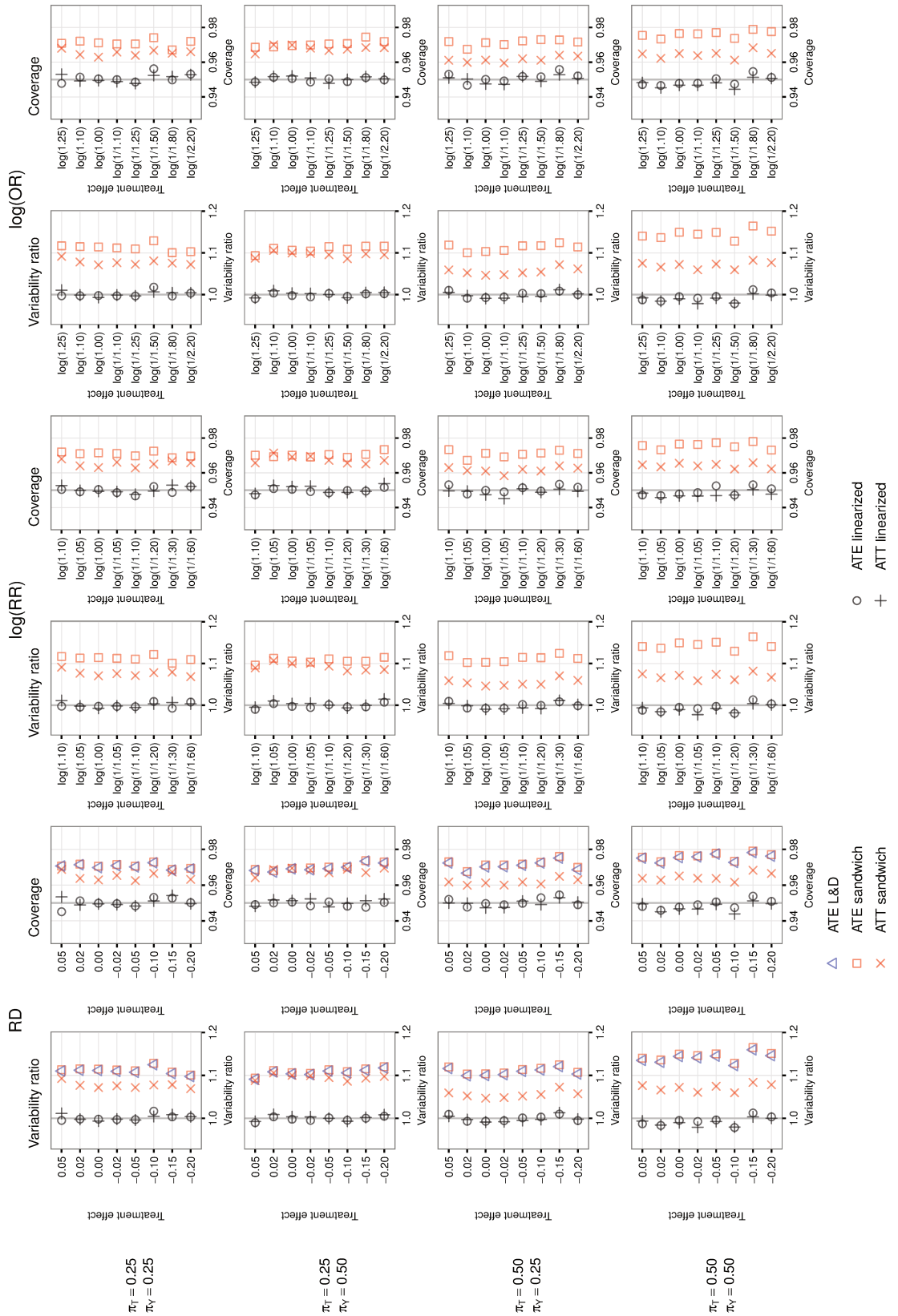
Several scenarios were explored, defined by:

- the treatment prevalence: $\pi_T \in \{0.25, 0.50\}$;
- the event rate: $\pi_Y \in \{0.25, 0.50\}$;
- the marginal treatment effect. Eight increasing levels of treatment effects were evaluated for each measurement, their value depending on the type of measurement:
  - risk difference : $\Gamma_{RD} \in$ {-0.20,-0.15,-0.10,-0.05,-0.02,0,0.02,0.05};
  - relative risk: $\Gamma_{RR} \in$ {log(1/1.60), log(1/1.3), log(1/1.2), log(1/1.1), log(1/1.05), log(1), log(1.05), log(1.1)};
  - odds ratio: $\Gamma_{OR} \in$ {log(1/2.20), log(1/1.8), log(1/1.5), log(1/1.25), log(1/1.1), log(1), log(1.1), log(1.25)}.
- the sample size: $n \in$ {500, 1,000, 2,000, ..., 10,000}.

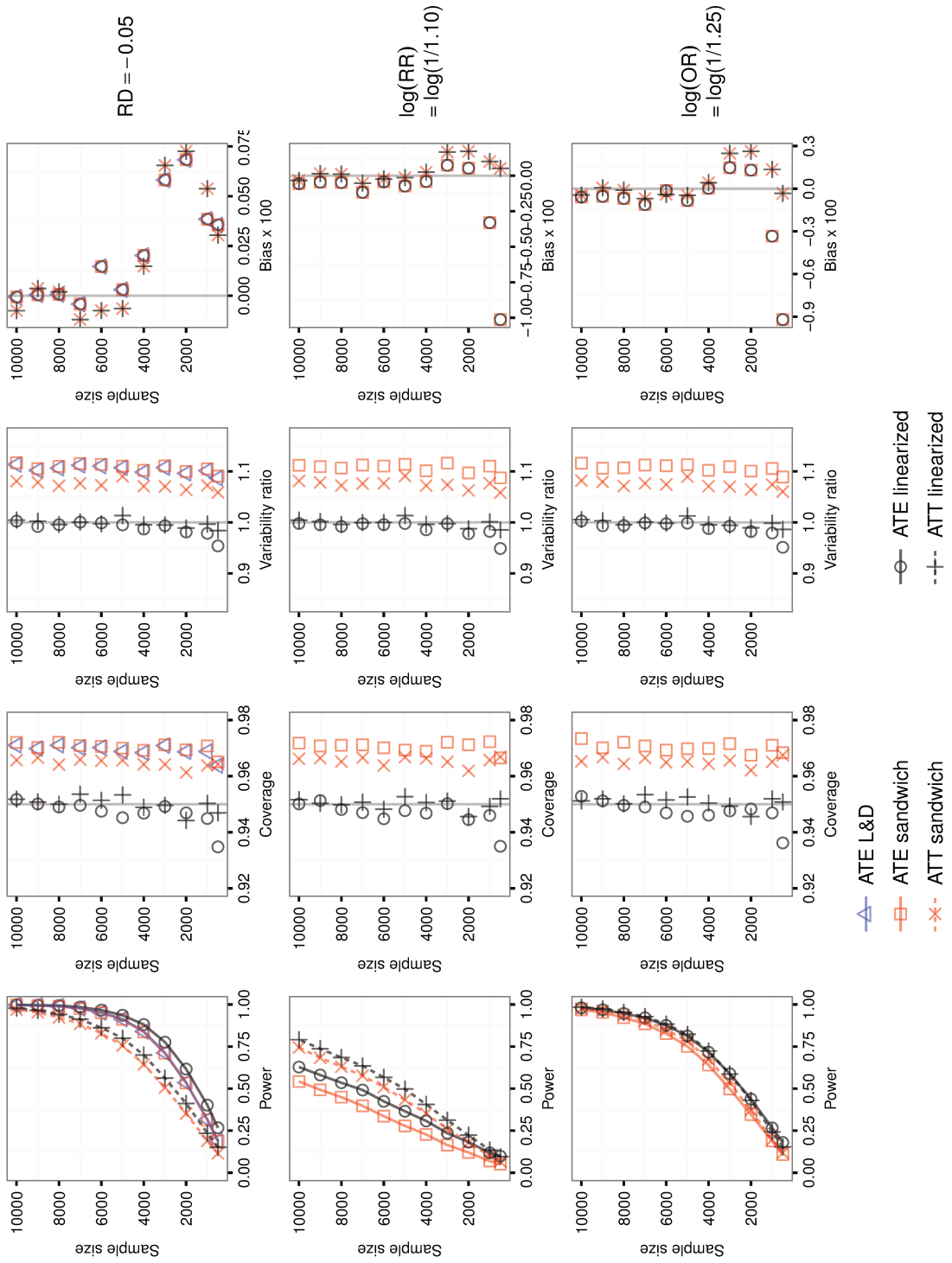
All simulated datasets included n=10,000 subjects. In each simulated dataset, six outcomes variables were generated, one for each evaluated treatment effect: ATE or ATT, using RD, RR or OR.

A total of B=10,000 datasets were generated for each scenario.

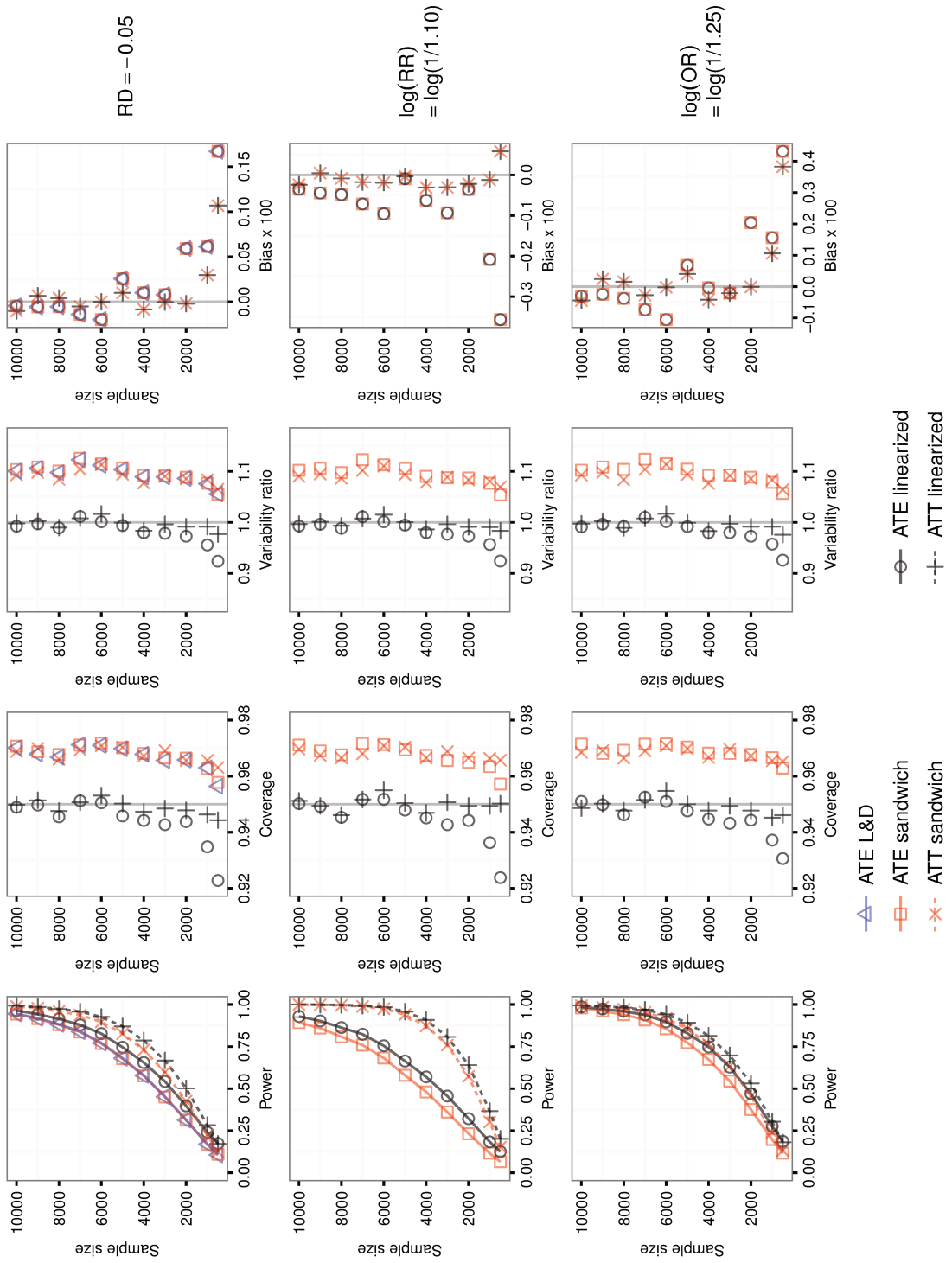
Web Appendix C Supplementary results



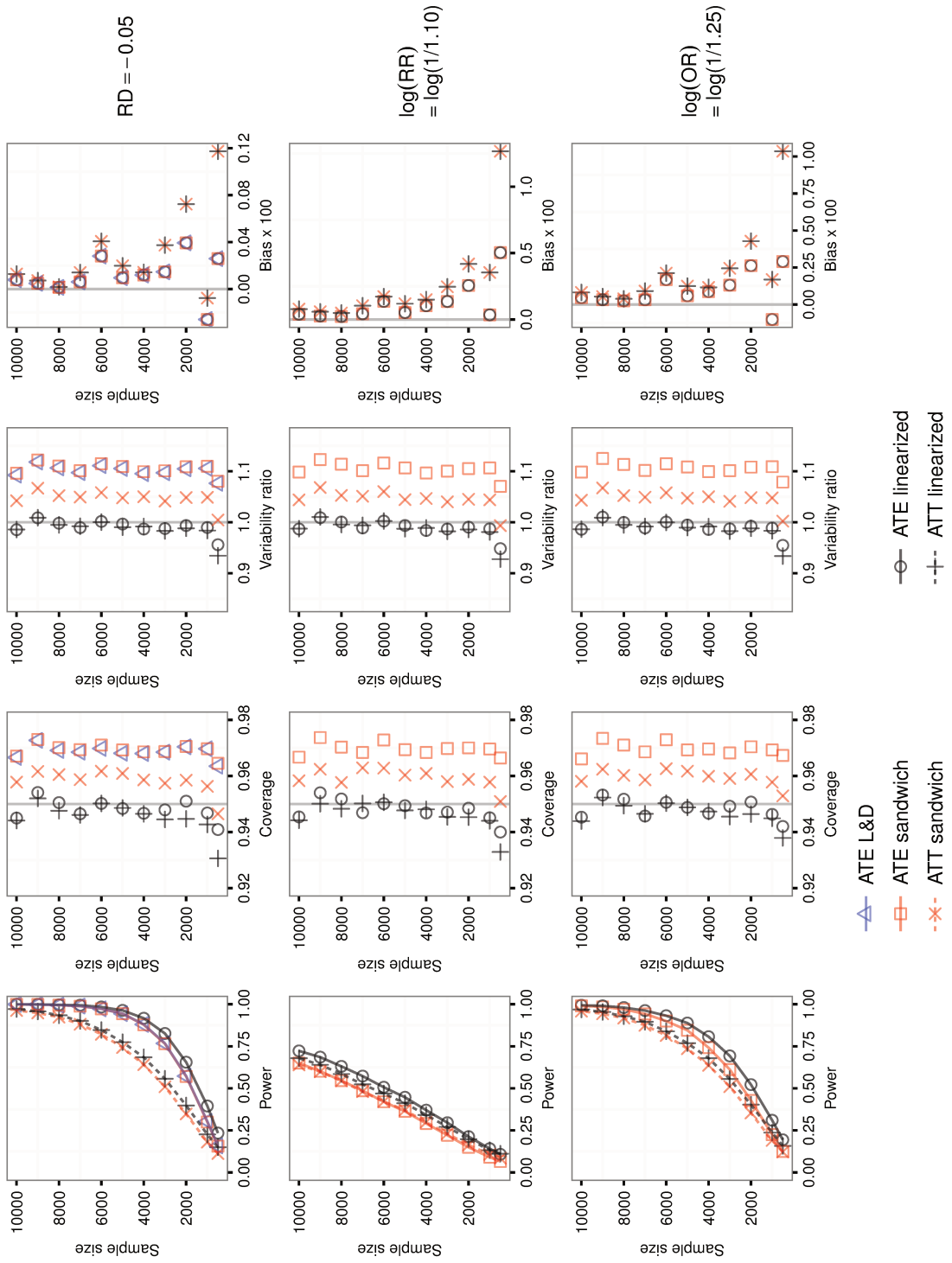
Web Figure 1: Variability ratio and coverage according to the variance estimator, true risk difference, true (log) odds ratio and true (log) relative risk, exposure prevalence π_T and event rate π_Y. ATE L&D: estimation of the ATE variance using the Lunceford and Davidian's estimator. ATE/ATT sandwich: estimation of the ATE/ATT variance using the sandwich estimators. ATE/ATT linearized: estimation of the ATE/ATT variance using the linearized estimators.



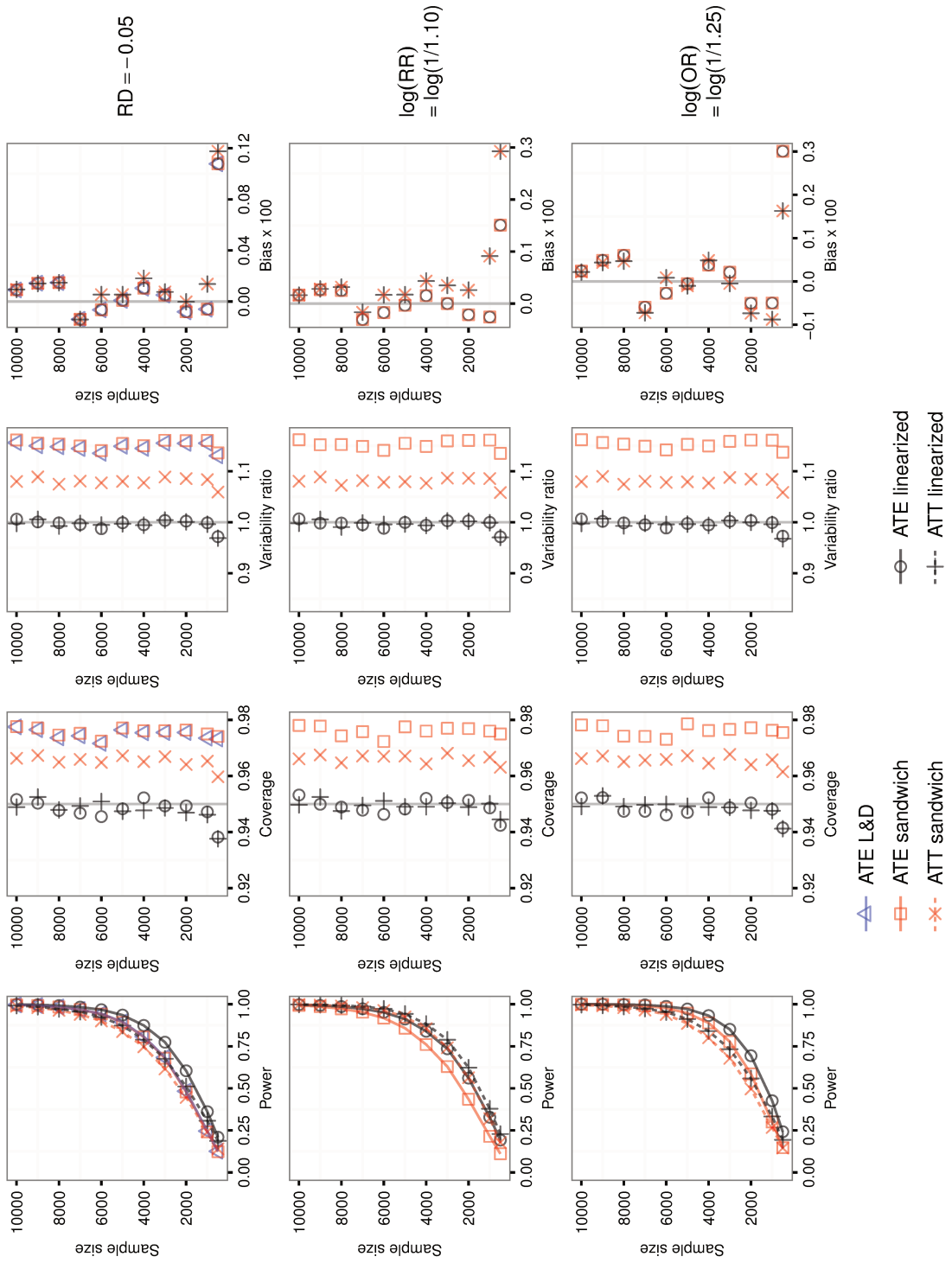
Web Figure 2: Performances according to variance estimator and sample size, with exposure prevalence $\pi_T = 0.25$ and event rate $\pi_Y = 0.25$. ATE L&D: estimation of the ATE variance using the Lunceford and Davidian's estimator. ATE/ATT sandwich: estimation of the ATE/ATT variance using the sandwich estimators. ATE/ATT linearized: estimation of the ATE/ATT variance using the linearized estimators.



Web Figure 3: Performances according to variance estimator and sample size, with exposure prevalence $\pi_T = 0.25$ and event rate $\pi_Y = 0.50$. ATE L&D: estimation of the ATE variance using the Lunceford and Davidian's estimator. ATE/ATT sandwich: estimation of the ATE/ATT variance using the sandwich estimators. ATE/ATT linearized: estimation of the ATE/ATT variance using the linearized estimators.



Web Figure 4: Performances according to variance estimator and sample size, with exposure prevalence $\pi_T = 0.50$ and event rate $\pi_Y = 0.25$. ATE L&D: estimation of the ATE variance using the Lunceford and Davidian's estimator. ATE/ATT sandwich: estimation of the ATE/ATT variance using the sandwich estimators. ATE/ATT linearized: estimation of the ATE/ATT variance using the linearized estimators.



Web Figure 5: Performances according to variance estimator and sample size, with exposure prevalence $\pi_T = 0.50$ and event rate $\pi_Y = 0.50$. ATE L&D: estimation of the ATE variance using the Lunceford and Davidian's estimator. ATE/ATT sandwich: estimation of the ATE/ATT variance using the sandwich estimators. ATE/ATT linearized: estimation of the ATE/ATT variance using the linearized estimators.

References

[1] Peter C. Austin. 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(20):2137–2148, September 2010.

[2] Peter C. Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161, March 2011.

5 | Conclusion

5.1 Résumé de la thèse

Nos travaux ont pour origine le constat suivant :

- les études de pharmacoépidémiologie à visée étiologique ont souvent pour objectif d'évaluer l'effet d'un médicament à partir de cohortes observationnelles reflétant la « vie réelle », mais sujettes au biais d'indication ;
- quand un médicament a été récemment mis sur le marché ou dispose de nombreuses alternatives thérapeutiques, le nombre de sujets exposés est faible par rapport au nombre de sujets non exposés.

Les méthodes basées sur le score de propension sont très populaires pour l'analyse d'études observationnelles dont l'objectif est d'évaluer l'effet d'un médicament en vie réelle. Cette popularité s'explique, entre autres, par leur relative facilité d'utilisation, et par la possibilité d'obtenir, sous certaines hypothèses, des estimations marginales, c'est-à-dire du même type que celles obtenues dans les essais randomisés (Deb et al. 2016), ces derniers constituant encore aujourd'hui le « gold standard » de l'évaluation thérapeutique. Une autre caractéristique intéressante est que la prise en compte des facteurs de confusion est réalisée via une modélisation de l'allocation du traitement, rendant cette technique d'analyse particulièrement attrayante dans les études où le nombre d'évènements est faible mais le nombre de sujets exposés au traitement est important (Cepeda et al. 2003 ; Patorno et al. 2014 ; Leyrat et al. 2014).

La situation réciproque (étude où le nombre d'évènements est important, mais le nombre de sujets exposés est faible) a, jusqu'à notre travail, été peu étudiée dans la littérature concernant le score de propension. Pirracchio et al. (2012) concluaient de leur étude de simulation que

> *même en cas de faible nombre de sujets ou de faible prévalence du traitement, l'appariement et la pondération sur le score de propension peuvent fournir des estimations non biaisées de l'effet du traitement.*

Cependant, malgré cette conclusion très générale, cette étude était principalement focalisée sur les situations de faible effectif, et non sur les situations de faible prévalence de l'exposition, puisque la plus faible prévalence étudiée était de 20%. En explorant des scénarios plus extrêmes (de 1% à 10% de prévalence), nous avons montré que ces deux méthodes d'utilisation du score de propension pouvaient conduire à des estimations biaisées de l'effet marginal de l'exposition, et ce problème était particulièrement net pour l'estimation de l'ATE. Ce résultat nous a conduits à recommander de privilégier l'estimation de l'ATT en situation de faible prévalence de l'exposition (si cela est cohérent avec la question scientifique), la pondération sur le score de propension ayant alors de meilleures performances que l'appariement.

Mais l'estimation de l'ATT ne correspond pas toujours à l'objectif de l'étude. En effet, les faibles prévalences de l'exposition se rencontrent dans deux situations principales :

— l'évaluation d'un médicament sur le marché depuis longtemps, mais peu prescrit : dans cette situation, l'estimation de l'ATT, c'est-à-dire l'effet du traitement chez les sujets l'ayant effectivement reçu, a un intérêt clinique certain ;

— l'évaluation d'un médicament récemment mis sur le marché, mais n'ayant pas vocation à rester peu prescrit : dans cette situation, objet d'une attention particulière de la part des autorités de santé, l'évaluation précoce de l'ATE, c'est-à-dire de l'effet du traitement si l'ensemble de la population cible était exposée, serait d'un intérêt majeur pour orienter des décisions de santé publique.

Nos résultats de simulation soulignaient donc le besoin de rechercher une alternative au score de propension, utilisable pour estimer l'ATE en situation de faible prévalence de l'exposition, et plus généralement moins influencée par la prévalence de l'exposition. Ceci nous a conduit à l'étude des propriétés des méthodes basées sur le score pronostique, du fait de leur popularité récente dans le domaine de la pharmacoépidémiologie (Arbogast & Ray 2009) et de leur recommandation en situation d'exposition rare (Arbogast et al. 2012). Présenté comme « l'analogue pronostique du score de propension » (Ben B. Hansen 2008), le score pronostique cherche à prendre en compte les facteurs de confusion via la modélisation du critère de jugement. Même si Ben B. Hansen (2008) destine cette méthode à l'estimation des effets marginaux, peu d'études se sont réellement intéressées au type d'estimation fournie par chaque méthode d'utilisation du score pronostique, ni à évaluer spécifiquement leurs performances en fonction de la prévalence de l'exposition.

Notre deuxième travail a démontré que les termes d'« analogue pronostique du score de propension » étaient, jusqu'à aujourd'hui, abusifs, puisqu'une seule des trois méthodes d'utilisation existantes du score pronostique permettait d'estimer un effet marginal (l'appariement sur le score pronostique, l'ajustement et la stratification estimant l'effet conditionnel), et qu'aucune d'entre elles ne permettait d'estimer l'ATE. Plus problématique encore, cette étude de simulation a montré que ces trois méthodes sous-estimaient systématiquement la variance de l'effet du traitement, en particulier si le nombre de sujets exposés n'était pas négligeable par rapport au nombre de sujets non exposés. Le développement de nouvelles méthodes d'utilisation du score pronostique (chacune adaptée à un seul type d'estimation : CTE, ATT ou ATE) ainsi que des estimateurs de variance correspondants nous a permis de rendre l'analyse par score pronostique plus flexible : elle permet de répondre à différents objectifs de recherche (alors que l'analyse par score de propension n'est réellement adaptée qu'à l'estimation des effets marginaux (Austin et al. 2007)) quel que soit le niveau de prévalence de l'exposition.

Nous avons mis en évidence que la sous-estimation de la variance des méthodes existantes basées sur le score pronostique était liée à la non prise en compte de l'étape d'estimation du

score pronostique. Contrairement à la conséquence d'une non prise en compte de l'étape d'estimation du score de propension (qui engendre une surestimation de la variance, et préserve le risque de première espèce en dessous du risque nominal fixé), ce problème, jusqu'alors non décrit, remet potentiellement en cause les conclusions des études de cohorte précédemment publiées utilisant une méthode basée sur le score pronostique, particulièrement si elles rapportent des résultats positifs (augmentation de la probabilité d'erreur de type I du fait d'intervalles de confiance sous-estimant le taux nominal de recouvrement).

Enfin, nous avons appliqué la technique de linéarisation, utilisée pour le développement des estimateurs de variance des nouvelles méthodes d'utilisation du score pronostique, pour développer des estimateurs de variance des effets du traitement estimés par pondération sur le score de propension. Contrairement aux autres méthodes d'utilisation du score de propension pour lesquels des estimateurs de variance valides existent déjà (Zou et al. 2016 ; Williamson et al. 2012 ; Abadie & Imbens 2009), nous avons mis en évidence une incohérence entre deux estimateurs de variance de l'ATE (Lunceford & Davidian 2004 ; Williamson et al. 2014), présentés comme similaires dans la littérature mais aux performances différentes (Austin 2010b ; Williamson et al. 2014). De plus, aucun estimateur de variance adapté à l'estimation de l'ATT par pondération sur le score de propension n'a été publié jusqu'ici. Dans notre troisième travail soumis pour publication, nous avons proposé une approche unifiée pour le développement d'estimateurs de variance de l'ATE et de l'ATT dans le cadre d'un critère de jugement binaire, et évalué leurs performances par rapport aux estimateurs existants prenant ou non en compte l'étape d'estimation du score de propension.

5.2 Perspectives

Le travail effectué dans le cadre de cette thèse ouvre de nombreuses perspectives de recherche.

Un prochain travail cherchera à évaluer les performances des méthodes d'utilisation du score pronostique (et plus particulièrement les nouvelles méthodes d'utilisation) pour l'estimation d'autres mesures d'association que l'odds-ratio : différence de moyennes pour les critères de jugement continus, différence de risques et risque relatif pour les critères binaires. En effet, l'extension des estimateurs de l'effet du traitement (et des estimateurs de variance correspondant) à ces autres mesures d'association ne pose aucune difficulté particulière. L'extension de ces méthodes pour l'évaluation de critères de jugement censurés serait également souhaitable ; cela nécessitera des développements plus poussés, notamment pour les méthodes estimant des effets marginaux.

Notre travail a permis de mettre en évidence la nécessité de prendre en compte l'étape d'estimation du score pronostique pour calculer la variance de l'effet du traitement. Des estimateurs de variance adaptés aux nouvelles méthodes d'utilisation du score pronostique ont été développés, mais il serait également intéressant d'en développer pour les trois méthodes d'utilisation du score pronostique précédemment décrites dans la littérature (ajustement, stratification et appariement). Toutes les méthodes d'utilisation du score pronostique pourraient ensuite être comparées sur un « pied d'égalité » concernant l'estimation de la variance.

L'évaluation des performances des différentes méthodes d'utilisation du score pronostique doit se poursuivre, notamment dans des situations proches de leur limite d'utilisation théorique, comme les évènements rares. En effet, nous n'avons pas exploré de scénario dont le taux d'évènements est inférieur à 20%. Or, l'étude d'évènements rares est fréquente en pharmacoépidémiologie, par exemple lors de l'évaluation de certains effets indésirables graves des médicaments. Le taux (ou le nombre) d'évènements au-dessous duquel une analyse par score pronostique est inenvisageable doit encore être déterminé.

Pour conclure, la situation de faible prévalence de l'exposition correspond à une situation fréquemment rencontrée en pharmacoépidémiologie, et l'utilisation du score pronostique se développe dans ce domaine. Mais il faut souligner que les méthodes étudiées ou déve-

loppées dans le cadre de cette thèse ne sont pas spécifiques à la pharmacoépidémiologie ou aux situations d'exposition rare. En particulier, les nouvelles méthodes basées sur le score pronostique semblent avoir des performances satisfaisantes pour prendre en compte les facteurs de confusion quelle que soit la prévalence de l'exposition. Ces méthodes constituent donc une alternative intéressante aux méthodes basées sur le score de propension dans tous les domaines où ces dernières sont déjà utilisées. Nous espérons donc que les applications qui utiliseront les résultats présentés dans cette thèse dépasseront le cadre de la pharmacoépidémiologie pour rejoindre le cadre plus général de l'épidémiologie clinique.

Bibliographie

Abadie, A. & Imbens, G.W., 2009. Matching on the Estimated Propensity Score. *National Bureau of Economic Research Working Paper Series*, No. 15301.

Aldridge, T.D. et al., 2014. First-trimester antihistamine exposure and risk of spontaneous abortion or preterm birth. *Pharmacoepidemiology and Drug Safety*, 23(10), pp.1043–1050.

Ali, M.S. et al., 2015. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal : A systematic review. *Journal of Clinical Epidemiology*, 68(2), pp.112–121.

Allison, P.D., 2009. *Fixed Effects Regression Models* 1 edition., Los Angeles : SAGE Publications, Inc.

Angrist, J.D., Imbens, G.W. & Rubin, D.B., 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434), pp.444–455.

Arbogast, P.G. & Ray, W.A., 2011. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *American Journal of Epidemiology*, 174(5), pp.613–620.

Arbogast, P.G. & Ray, W.A., 2009. Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research*, 18(1), pp.67–80.

Arbogast, P.G., Seeger, J.D. & DEcIDE Methods Center Summary Variable Working Group, 2012. Summary Variables in Observational Research : Propensity Scores and Disease Risk Scores. *Effective Health Care Program Research Report*, 33.

Austin, P.C., 2014a. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6), pp.1057–1069.

Austin, P.C., 2008a. A critical appraisal of propensity-score matching in the medical literature between

1996 and 2003. *Statistics in Medicine*, 27(12), pp.2037–2049.

Austin, P.C., 2011a. A Tutorial and Case Study in Propensity Score Analysis : An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. *Multivariate Behavioral Research*, 46(1), pp.119–151.

Austin, P.C., 2011b. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3), pp.399–424.

Austin, P.C., 2009a. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25), pp.3083–3107.

Austin, P.C., 2010a. Different measures of treatment effect for different research questions. *Journal of Clinical Epidemiology*, 63(1), pp.9–10.

Austin, P.C., 2011c. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2), pp.150–161.

Austin, P.C., 2009b. Some methods of propensity-score matching had superior performance to others : Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal. Biometrische Zeitschrift*, 51(1), pp.171–184.

Austin, P.C., 2013. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine*, 32(16), pp.2837–2849.

Austin, P.C., 2007. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, 26(16), pp.3078–3094.

Austin, P.C., 2010b. 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(20), pp.2137–2148.

Austin, P.C., 2008b. The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology*, 61(6), pp.537–545.

Austin, P.C., 2009c. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, 29(6), pp.661–677.

Austin, P.C., 2014b. The use of propensity score methods with survival or time-to-event outcomes : Reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 33(7),

pp.1242–1258.

Austin, P.C., 2016. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*.

Austin, P.C. & Laupacis, A., 2011. A Tutorial on Methods to Estimating Clinically and Policy-Meaningful Measures of Treatment Effects in Prospective Observational Studies : A Review. *The International Journal of Biostatistics*, 7(1).

Austin, P.C. & Schuster, T., 2014. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes : A simulation study. *Statistical Methods in Medical Research*, p.0962280213519716.

Austin, P.C. & Small, D.S., 2014. The use of bootstrapping when using propensity-score matching without replacement : A simulation study. *Statistics in Medicine*, 33(24), pp.4306–4319.

Austin, P.C. & Stuart, E.A., 2015a. 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*, 34(28), pp.3661–3679.

Austin, P.C. & Stuart, E.A., 2015b. Optimal full matching for survival outcomes : A method that merits more widespread use. *Statistics in Medicine*, 34(30), pp.3949–3967.

Austin, P.C. & Stuart, E.A., 2015c. 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*, p.0962280215584401.

Austin, P.C. et al., 2007. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect : A Monte Carlo study. *Statistics in Medicine*, 26(4), pp.754–768.

Austin, P.C. et al., 2010. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials : A review of trials published in leading medical journals. *Journal of Clinical Epidemiology*, 63(2), pp.142–153.

Beigel, F. et al., 2014. Risk of malignancies in patients with inflammatory bowel disease treated with thiopurines or anti-TNF alpha antibodies. *Pharmacoepidemiology and Drug Safety*, 23(7), pp.735–744.

Berglind, I.A. et al., 2014. Hormone therapy and risk of cardiovascular outcomes and mortality in women treated with statins. *Menopause (New York, N.Y.)*.

Berkowitz, S.A. et al., 2014. Initial Choice of Oral Glucose-Lowering Medication for Diabetes Mellitus : A

Patient-Centered Comparative Effectiveness Study. *JAMA internal medicine*.

Bhatt, D.L. et al., 2010. Comparative determinants of 4-year cardiovascular event rates in stable outpatients at risk of or with atherothrombosis. *JAMA*, 304(12), pp.1350–1357.

Brookhart, M.A. & Schneeweiss, S., 2007. Preference-Based Instrumental Variable Methods for the Estimation of Treatment Effects : Assessing Validity and Interpreting Results. *The International Journal of Biostatistics*, 3(1).

Brookhart, M.A. et al., 2006. Evaluating Short-Term Drug Effects Using a Physician-Specific Prescribing Preference as an Instrumental Variable : *Epidemiology*, 17(3), pp.268–275.

Burton, A. et al., 2006. The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24), pp.4279–4292.

Cameron, C. et al., 2015. Network meta-analysis incorporating randomized controlled trials and non-randomized comparative cohort studies for assessing the safety and effectiveness of medical treatments : Challenges and opportunities. *Systematic Reviews*, 4, p.147.

Cartwright, N., 1979. Causal Laws and Effective Strategies. *Noûs*, 13(4), pp.419–437.

Cepeda, M.S. et al., 2003. Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. *American Journal of Epidemiology*, 158(3), pp.280–287.

Chen, Y. & Briesacher, B.A., 2011. Use of Instrumental Variable in Prescription Drug Research with Observational Data : A Systematic Review. *Journal of clinical epidemiology*, 64(6), pp.687–700.

Cochran, W.G., 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2), pp.295–313.

Colantuoni, E. & Rosenblum, M., 2015. Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in Medicine*, 34(18), pp.2602–2617.

Cole, S.R. & Hernán, M.A., 2008. Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, 168(6), pp.656–664.

Connolly, J.G. & Gagne, J.J., 2016. Comparison of Calipers for Matching on the Disease Risk Score. *American Journal of Epidemiology*, 183(10), pp.937–948.

Cook, E.F. & Goldman, L., 1989. Performance of tests of significance based on stratification by a multiva-

riate confounder score or by a propensity score. *Journal of Clinical Epidemiology*, 42(4), pp.317–324.

Dahabreh, I.J. et al., 2012. Do observational studies using propensity score methods agree with randomized trials ? A systematic comparison of studies on acute coronary syndromes. *European Heart Journal*, 33(15), pp.1893–1901.

Deb, S. et al., 2016. A Review of Propensity-Score Methods and Their Use in Cardiovascular Research. *Canadian Journal of Cardiology*, 32(2), pp.259–265.

Deville, J.C., 1999. Variance estimation for complex statistics and estimators : Linearization and residual techniques. *Survey methodology*, 25(2), pp.193–204.

Donald B. Rubin, P.R.R., 1985. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician 39, 33-38. *The American Statistician*, 39(1).

Eddelbuettel, D. & Francois, R., 2011. Rcpp : Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8), pp.1–18.

Eftekhari, K. et al., 2014. Risk of retinal tear or detachment with oral fluoroquinolone use : A cohort study. *Pharmacoepidemiology and Drug Safety*, 23(7), pp.745–752.

EGAP (Evidence in Governance and Politics), 2016. 10 Types of Treatment Effect You Should Know About.

ENCePP, 2016. ENCePP Guide on Methodological Standards in Pharmacoepidemiology.

Forbes, A. & Shortreed, S., 2008. Inverse probability weighted estimation of the marginal odds ratio : Correspondence regarding “The performance of different propensity score methods for estimating marginal odds ratios” by P. Austin, Statictics in Medicine, 2007 ; 26 :30783094. *Statistics in Medicine*, 27(26), pp.5556–5559.

Frölich, M., 2004. Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators. *Review of Economics and Statistics*, 86(1), pp.77–90.

Gail, M.H., Wieand, S. & Piantadosi, S., 1984. Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates. *Biometrika*, 71(3), p.431.

Gayat, E. et al., 2010. Propensity scores in intensive care and anaesthesiology literature : A systematic review. *Intensive Care Medicine*, 36(12), pp.1993–2003.

Gayat, E. et al., 2012. Propensity score applied to survival data analysis through proportional hazards

models : A Monte Carlo study. *Pharmaceutical Statistics*, 11(3), pp.222–229.

Glynn, R.J., Gagne, J.J. & Schneeweiss, S., 2012. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology and Drug Safety*, 21, pp.138–147.

Glynn, R.J., Schneeweiss, S. & Sturmer, T., 2006. Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology. *Basic & clinical pharmacology & toxicology*, 98(3), pp.253–259.

Graf, E. & Schumacher, M., 2008. Comments on "The performance of different propensity score methods for estimating marginal odds ratios" by Peter C. Austin, Statistics in Medicine 2007 ; 26(16) :30783094. *Statistics in Medicine*, 27(19), pp.3915–3917.

Greenland, S., 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4), pp.722–729.

Greenland, S., 1987. Interpretation and Choice of Effect Measures in Epidemiologic Analyses. *American Journal of Epidemiology*, 125(5), pp.761–768.

Greenland, S. & Robins, J.M., 2009. Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, 6, p.4.

Greenland, S., Robins, J.M. & Pearl, J., 1999. Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1), pp.29–46.

Hajage, D. et al., 2016. Estimation of conditional and marginal odds ratios using the prognostic score. *Statistics in Medicine*, pp.n/a–n/a.

Hajage, D. et al., 2016. On the use of propensity scores in case of rare exposure. *BMC medical research methodology*, 16, p.38.

Hansen, B.B., 2008. The essential role of balance tests in propensity-matched observational studies : Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, 27(12), pp.2050–2054 ; discussion 2066–2069.

Hansen, B.B., 2008. The prognostic analogue of the propensity score. *Biometrika*, 95(2), pp.481–488.

Havercroft, W.G. & Didelez, V., 2012. Simulating from marginal structural models with time-dependent confounding. *Statistics in medicine*, 31(30), pp.4190–4206.

Heerdink, E.R., Urquhart, J. & Leufkens, H.G., 2002. Changes in prescribed drug doses after market

introduction. *Pharmacoepidemiology and Drug Safety*, 11(6), pp.447–453.

Herbst, A.L., Ulfelder, H. & Poskanzer, D.C., 1971. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *The New England Journal of Medicine*, 284(15), pp.878–881.

Hill, A.B., 1965. The Environment and Disease : Association or Causation ? *Proceedings of the Royal Society of Medicine*, 58(5), pp.295–300.

Hill, J., 2008. Discussion of research using propensity-score matching : Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, 27(12), pp.2055–2061 ; discussion 2066–2069.

Imai, K. & van Dyk, D.A., 2004. Causal Inference With General Treatment Regimes : Generalizing the Propensity Score. *Journal of the American Statistical Association*, 99, pp.854–866.

Imbens, G., 2004. Nonparametric Estimation of Average Treatment Effects Under Exogeneity : A Review. *The Review of Economics and Statistics*, 86(1), pp.4–29.

Ioannidis, J.P.A., 2016. Exposure-wide epidemiology : Revisiting Bradford Hill. *Statistics in Medicine*, 35(11), pp.1749–1762.

Kestenbaum, B., 2009. Methods to Control for Confounding. In *Epidemiology and Biostatistics*. Springer New York, pp. 101–111.

Kurth, T. et al., 2006. Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, and Propensity-based Weighting under Conditions of Nonuniform Effect. *American Journal of Epidemiology*, 163(3), pp.262–270.

Laforest, L. et al., 2014. Asthma drug ratios and exacerbations : Claims data from universal health coverage systems. *The European Respiratory Journal*, 43(5), pp.1378–1386.

Lan, K.K.G. & Demets, D.L., 1983. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3), pp.659–663.

Lavandeira, A., 2002. Orphan drugs : Legal aspects, current situation. *Haemophilia : The Official Journal of the World Federation of Hemophilia*, 8(3), pp.194–198.

Leacy, F.P. & Stuart, E.A., 2014. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated : A simulation study. *Statistics in Medicine*, 33(20), pp.3488–3508.

Lewis, J.D. et al., 2014. Proteinuria testing among patients with diabetes mellitus is associated with

bladder cancer diagnosis : Potential for unmeasured confounding in studies of pioglitazone and bladder cancer. *Pharmacoepidemiology and Drug Safety*, 23(6), pp.636–645.

Leyrat, C. et al., 2014. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Statistics in Medicine*, 33(20), pp.3556–3575.

Localio, A.R., Margolis, D.J. & Berlin, J.A., 2007. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology*, 60(9), pp.874–882.

Lu, C.Y., 2009. Observational studies : A review of study designs, challenges and strategies to reduce confounding. *International Journal of Clinical Practice*, 63(5), pp.691–697.

Lumley, T., 2004. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), pp.1–19.

Lunceford, J.K. & Davidian, M., 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects : A comparative study. *Statistics in Medicine*, 23(19), pp.2937–2960.

Lunde, P.K. & Baksaas, I., 1988. Epidemiology of drug utilization–Basic concepts and methodology. *Acta Medica Scandinavica. Supplementum*, 721, pp.7–11.

Månsson, R. et al., 2007. On the Estimation and Use of Propensity Scores in Case-Control and Case-Cohort Studies. *American Journal of Epidemiology*, 166(3), pp.332–339.

McCaffrey, D.F. et al., 2013. A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. *Statistics in medicine*, 32(19), pp.3388–3414.

McKenzie, M.W. et al., 1976. Adverse drug reactions leading to hospitalization in children. *The Journal of Pediatrics*, 89(3), pp.487–490.

Miettinen, O., 1974. Confounding and effect-modification. *American Journal of Epidemiology*, 100(5), pp.350–353.

Miettinen, O.S., 1976. Stratification by a multivariate confounder score. *American Journal of Epidemiology*, 104(6), pp.609–620.

Miettinen, O.S. & Cook, E.F., 1981. Confounding : Essence and detection. *American Journal of Epidemiology*, 114(4), pp.593–603.

Morgan, S.L. & Winship, C., 2007. *Counterfactuals and Causal Inference : Methods and Principles for*

Social Research, Cambridge University Press.

Normand, S.-L.T. et al., 2005. Readers guide to critical appraisal of cohort studies : 3. Analytical strategies to reduce confounding. *BMJ*, 330(7498), pp.1021–1023.

Ohman, E.M. et al., 2006. The REduction of Atherothrombosis for Continued Health (REACH) Registry : An international, prospective, observational investigation in subjects at risk for atherothrombotic events-study design. *American Heart Journal*, 151(4), pp.786.e1–10.

Patorno, E. et al., 2014. Studies with many covariates and few outcomes : Selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology (Cambridge, Mass.)*, 25(2), pp.268–278.

Pavlou, M. et al., 2016. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, 35(7), pp.1159–1177.

Peduzzi, P. et al., 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), pp.1373–1379.

Pfeiffer, R.M. & Riedl, R., 2015. On the use and misuse of scalar scores of confounders in design and analysis of observational studies. *Statistics in Medicine*, 34(18), pp.2618–2635.

Pike, M.C., Anderson, J. & Day, N., 1979. Some insights into Miettinen's multivariate confounder score approach to case-control study analysis. *Epidemiology and Community Health*, 33(1), pp.104–106.

Pirracchio, R. et al., 2013. 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*, p.0962280213507034.

Pirracchio, R., Resche-Rigon, M. & Chevret, S., 2012. Evaluation of the Propensity score methods for estimating marginal odds ratios in case of small sample size. *BMC Medical Research Methodology*, 12(1), p.70.

Pirracchio, R. et al., 2011. Utility of time-dependent inverse-probability-of-treatment weights to analyze observational cohorts in the intensive care unit. *Journal of clinical epidemiology*, 64(12), pp.1373–1382.

Rafaniello, C. et al., 2014. Predictors of mortality in atypical antipsychotic-treated community-dwelling elderly patients with behavioural and psychological symptoms of dementia : A prospective population-based cohort study from Italy. *European Journal of Clinical Pharmacology*, 70(2), pp.187–195.

Rassen, J.A. & Schneeweiss, S., 2012. Newly marketed medications present unique challenges for nonrando-

mized comparative effectiveness analyses. *Journal of Comparative Effectiveness Research*, 1(2), pp.109–111.

Rassen, J.A., Avorn, J. & Schneeweiss, S., 2010. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. *Pharmacoepidemiology and drug safety*, 19(8), pp.848–857.

Reed, S.D. et al., 2008. Updated estimates of survival and cost effectiveness for imatinib versus interferon-alpha plus low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukaemia. *PharmacoEconomics*, 26(5), pp.435–446.

Resche-Rigon, M. et al., 2012. Estimating the treatment effect from non-randomized studies : The example of reduced intensity conditioning allogeneic stem cell transplantation in hematological diseases. *BMC Hematology*, 12(1), p.10.

Robert J Glynn, Mark Lunt & Til Stümer, 2014. Trimming on Propensity or Disease Risk Scores to Enhance Validity in the Design of Comparative Effectiveness Studies, presented as the 30th International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Taipei, October 26th 2014.

Robins, J.M., Hernán, M.Á. & Brumback, B., 2000. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5), pp.550–560.

Rosenbaum, P.R., 1987. Model-Based Direct Adjustment. *Journal of the American Statistical Association*, 82(398), pp.387–394.

Rosenbaum, P.R. & Rubin, D.B., 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387), p.516.

Rosenbaum, P.R. & Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), pp.41–55.

Rosenblum, M. & van der Laan, M.J., 2010. Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables. *The International Journal of Biostatistics*, 6(1).

Rothman, K.J., Greenland, S. & Lash, T.L., 2008. *Modern Epidemiology*, Lippincott Williams & Wilkins.

Roussel, R. et al., 2013. Thiazolidinedione use is not associated with worse cardiovascular outcomes : A study in 28,332 high risk patients with diabetes in routine clinical practice : Brief title : Thiazolidinedione use and mortality. *Int J Cardiol*, 167(4), pp.1380–4.

Rubin, D.B. & Thomas, N., 1996. Matching Using Estimated Propensity Scores : Relating Theory to

Practice. *Biometrics*, 52(1), p.249.

Schemper, M., Wakounig, S. & Heinze, G., 2009. The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*, 28(19), pp.2473–2489.

Schmidt, A.F. et al., 2016. Adjusting for Confounding in Early Postlaunch Settings : Going Beyond Logistic Regression Models. *Epidemiology (Cambridge, Mass.)*, 27(1), pp.133–142.

Sekhon, J.S., 2011. Multivariate and Propensity Score Matching Software with Automated Balance Optimization : The Matching package for R. *Journal of Statistical Software*, 42(7), p.52.

Senn, S., Graf, E. & Caputo, A., 2007. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine*, 26(30), pp.5529–5544.

Shah, B.R. et al., 2005. Propensity score methods gave similar results to traditional regression modeling in observational studies : A systematic review. *Journal of Clinical Epidemiology*, 58(6), pp.550–559.

Sikirica, S. et al., 2014. Risk of death associated with the use of conventional vs. atypical antipsychotic medications : Evaluating the use of the Emilia-Romagna Region database for pharmacoepidemiological studies. *Journal of Clinical Pharmacy and Therapeutics*, 39(1), pp.38–44.

Steg, P.G. et al., 2007. One-year cardiovascular event rates in outpatients with atherothrombosis. *JAMA*, 297(11), pp.1197–1206.

Strom, B.L., Kimmel, S.E. & Hennessy, S., 2012. *Pharmacoepidemiology* 5th Edition., Chichester, West Sussex, UK : Wiley-Blackwell.

Stuart, E.A., 2008. Developing practical recommendations for the use of propensity scores : Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, 27(12), pp.2062–2065 ; discussion 2066–2069.

Stuart, E.A., Lee, B.K. & Leacy, F.P., 2013. Prognostic scorebased balance measures for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8 0), pp.S84–S90.e1.

Sturmer, T. et al., 2007. Adjustments for Unmeasured Confounders in Pharmacoepidemiologic Database Studies Using External Information. *Medical care*, 45(10 SUPL), pp.S158–S165.

Sturmer, T. et al., 2006. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology*, 59(5), pp.437–447.

Tadrous, M. et al., 2013. Disease Risk Score (DRS) as a Confounder Summary Method : Systematic Review

and Recommendations. *Pharmacoepidemiology and drug safety*, 22(2), pp.122–129.

Tadrous, M. et al., 2015. Performance of the disease risk score in a cohort study with policy-induced selection bias. *Journal of Comparative Effectiveness Research*, 4(6), pp.607–614.

Thoemmes, F.J. & Kim, E.S., 2011. A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, 46(1), pp.90–118.

Tubach, F. et al., 2011. Role of the post-marketing authorisation studies in drug risk surveillance : specifications and methodologies. *Thérapie*, 66(4), pp.355–362, 347–354.

Uddin, M.J. et al., 2015. Performance of prior event rate ratio adjustment method in pharmacoepidemiology : A simulation study. *Pharmacoepidemiology and Drug Safety*, 24(5), pp.468–477.

Velentgas, P. et al., 2013. *Developing a Protocol for Observational Comparative Effectiveness Research : A User's Guide*, Rockville (MD) : Agency for Healthcare Research and Quality (US).

Vittinghoff, E. & McCulloch, C.E., 2007. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, 165(6), pp.710–718.

Waernbaum, I., 2012. Model misspecification and robustness in causal inference : Comparing matching with doubly robust estimation. *Statistics in Medicine*, 31(15), pp.1572–1581.

Weinhandl, E.D. et al., 2014. Relative safety of peginesatide and epoetin alfa. *Pharmacoepidemiology and Drug Safety*, 23(10), pp.1003–1011.

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

Williamson, E.J., Forbes, A. & White, I.R., 2014. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, 33(5), pp.721–737.

Williamson, E.J. et al., 2012. Variance estimation for stratified propensity score estimators. *Statistics in Medicine*, 31(15), pp.1617–1632.

Wyss, R. et al., 2015. Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiology and Drug Safety*, 24(9), pp.951–961.

Wyss, R. et al., 2014. Reducing Bias Amplification in the Presence of Unmeasured Confounding Through Out-of-Sample Estimation Strategies for the Disease Risk Score. *Journal of Causal Inference*, 2(2), pp.131–

146.

Xiao, Y., Abrahamowicz, M. & Moodie, E.E.M., 2010. Accuracy of Conventional and Marginal Structural Cox Model Estimators : A Simulation Study. *The International Journal of Biostatistics*, 6(2).

Xu, S. et al., 2016. Evaluation of propensity scores, disease risk scores, and regression in confounder adjustment for the safety of emerging treatment with group sequential monitoring. *Pharmacoepidemiology and Drug Safety*, 25(4), pp.453–461.

Yu, M. et al., 2012. Prior event rate ratio adjustment : Numerical studies of a statistical method to address unrecognized confounding in observational studies. *Pharmacoepidemiology and Drug Safety*, 21 Suppl 2, pp.60–68.

Zou, B. et al., 2016. On variance estimate for covariate adjustment by propensity score analysis. *Statistics in Medicine*, 35(20), pp.3537–3548.