Université de Montréal

Sur les estimateurs doublement robustes avec sélection de modèles et de variables pour les données administratives

par

Asma Bahamyirou

Axe médicament et santé des populations

Faculté de pharmacie

Thèse présentée en vue de l'obtention du grade de

Philosophiæ Doctor (Ph.D.)

en Sciences pharmaceutiques, option Médicaments et santé des populations

December 20, 2021

Université de Montréal

Faculté de pharmacie

Cette thèse intitulée

Sur les estimateurs doublement robustes avec sélection de modèles et de variables pour les données administratives

présentée par

Asma Bahamyirou

a été évaluée par un jury composé des personnes suivantes :

Fahima Nekka, Ph.D
(président-rapporteur)

Mireille E. Schnitzer, Ph.D
(directeur de recherche)

David Haziza, Ph.D
(membre du jury)

Antoine Chanbaz, Ph.D
(examinateur externe)

Sébastien Lemieux, Ph.D
(représentant du doyen de la FESP)

Dédicace

À mes parents, Makpanawè Bahamyirou et Fidèle Acakpo-Addra pour tous les sacrifices consentis pour mon éducation. Cette thèse est pour vous. *Nlabalè, Akpé.*

Résumé

Les essais cliniques randomisés (ECRs) constituent la meilleure solution pour obtenir des effets causaux et évaluer l'efficacité des médicaments. Toutefois, vu qu'ils ne sont pas toujours réalisables, les bases de données administratives servent de solution de remplacement. Le sujet principal de cette thèse peut être divisée en deux parties, le tout, repartie en trois articles. La première partie de cette thèse traite de l'utilisation des estimateurs doublement robustes en inférence causale sur des bases de données administratives avec intégration des méthodes d'apprentissage automatique. Nous pouvons citer, par exemple, l'estimateur par maximum de vraisemblance ciblé (TMLE; [**73**]) et l'estimateur par augmentation de l'inverse de la probabilité de traitement (AIPTW; [**51**]). Ces méthodes sont de plus en plus utilisées [**65, 102, 86, 7, 37**] en pharmacoépidémiologie pour l'estimation des paramètres causaux, comme l'effet moyen du traitement. Dans la deuxième partie de cette thèse, nous développons un estimateur doublement robuste pour les données administratives et nous étendons une méthode existante [**121**] pour l'ajustement du biais de sélection utilisant un échantillon probabiliste de référence.

Le premier manuscrit de cette thèse présente un outil de diagnostic pour les analystes lors de l'utilisation des méthodes doublement robustes. Ce manuscrit démontre à l'aide d'une étude de simulation l'impact de l'estimation du score de propension par des méthodes flexibles sur l'effet moyen du traitement, et ce, en absence de positivité pratique. L'article propose un outil capable de diagnostiquer l'instabilité de l'estimateur en absence de positivité pratique et présente une application sur les médicaments contre l'asthme durant la grossesse.

Le deuxième manuscrit présente une procédure de sélection de modificateurs d'effet et d'estimation de l'effet conditionnel. En effet, cet article utilise une procédure de régularisation en deux étapes et peut être appliqué sur plusieurs logiciels standards. Finalement, il présente une application sur les médicaments contre l'asthme durant la grossesse.

Le dernier manuscrit développe une méthodologie pour ajuster un biais de sélection dans une base de données administratives dans le but d'estimer une moyenne d'une population, et ce, en présence d'un échantillon probabiliste provenant de la même population avec des covariables communes. En utilisant une méthode de régularisation, il montre qu'il est possible de sélectionner statistiquement les bonnes variables à ajuster dans le but de réduire l'erreur quadratique moyenne et la variance. Cet article décrit ensuite une application sur l'impact de la COVID-19 sur les Canadiens.

Mots-clés: inférence causale, doublement robuste, données administratives, score de propension, apprentissage automatique.

Abstract

Randomized clinical trials (RCTs) are the gold standard for establishing causal effects and evaluating drug efficacy. However, RCTs are not always feasible and the usage of administrative data for the estimation of a causal parameter is an alternative solution. The main subject of this thesis can be divided into two parts, the whole comprised of three articles. The first part studies the usage of doubly robust estimators in causal inference using administrative data and machine learning. Examples of doubly robust estimators are Targeted Maximum Likelihood Estimation (TMLE; [**73**]) and Augmented Inverse Probability of Treatment Weighting (AIPTW; [**51**]). These methods are more and more present in pharmacoepidemiology [**65, 102, 86, 7, 37**]. In the second part of this thesis, we develop a doubly robust estimator and extend an existing one [**121**] for the setting of administrative data with a supplemental probability sample. The first paper of this thesis proposes a diagnostic tool that uses re-sampling methods to identify instability in doubly robust estimators when using data-adaptive methods in the presence of near practical positivity violations. It demonstrates the impact of machine learning methods for propensity score estimation when near practical positivity violations are induced. It then describes an analysis of asthma medication during pregnancy. The second manuscript develops a methodology to statistically select effect modifying variables using a two stage procedure in the context of a single time point exposure. It then describes an analysis of asthma medication during pregnancy. The third manuscript describes the development of a variable selection procedure using penalization for combining a nonprobability and probability sample in order to adjust for selection bias. It shows that we can statistically select the right subset of the variables when the true

propensity score model is sparse. It demonstrates the benefit in terms of mean squared error and presents an application of the impact of COVID-19 on Canadians.
Keywords: causal inference, doubly robust, administrative data, propensity score, machine learning.

Table des matières

Dédicace 5

Résumé 7

Abstract 9

Liste des tableaux 17

Liste des figures 21

Liste des sigles et des abréviations 23

Remerciements 27

Chapitre 1. Introduction 29

Chapitre 2. Revue de la littérature 33

2.1. Échantillon non-probabiliste ou données administratives 33

2.2. Inférence causale 34

2.2.1. Hypothèses d'identifiabilité 36

2.2.2. Modèles structurels marginaux 38

2.2.3. Modificateur d'effet 39

2.2.3.1. Détection d'un modificateur d'effet et estimation du CATE 40

2.3. Méthode d'inférence 41

2.3.1. Méthode de pondération inverse 41

2.3.2. Méthode d'augmentation de l'inverse de la probabilité de traitement 43

2.3.2.1. Estimation des effects causaux par la régression 43
2.3.2.2. Méthode d'augmentation de l'inverse de la probabilité de traitement .. 44
2.3.2.3. Note sur la linéarité asymptotique de l'AIPTW 45
2.3.3. Méthode d'estimation par maximum de vraisemblance ciblée 46
2.3.3.1. Cas de l'effet moyen du traitement 47
2.3.4. Inférence en présence de positivité pratique 49
2.3.4.1. Estimateur TMLE collaboratif 50
2.3.4.2. Outils de diagnostic paramétrique 52
2.3.5. Inférence en présence de bias de sélection 52

2.4. Méthodes d'apprentissage automatique 54
2.4.1. Super Learner 55
2.4.2. LASSO 56
2.4.2.1. Définitions et notations 56
2.4.2.2. Propriété oracle et le LASSO adaptatif 57
2.4.2.3. LASSO adaptatif basé sur l'issue 58
2.4.2.4. LASSO hautement adaptatif (Highly Adaptive LASSO; HAL) 59
2.4.2.5. Inférence sélective 60

Chapitre 3. Objectifs 63

Chapitre 4. Understanding and Diagnosing the Potential for Bias when using Machine Learning Methods with Doubly Robust Causal Estimators 65

4.1. Introduction 67

4.2. Estimators 69
4.2.1. Targeted estimation 69

4.2.2. Inverse Probability of Treatment Weighting (IPTW) 69
4.2.3. Targeted Minimum Loss-Based Estimation 70
4.2.4. Super Learner 71
4.2.5. Collaborative Targeted Minimum Loss-Based Estimation 72

4.3. Simulation Scenario 73
4.3.1. Simulated data 73
4.3.2. Estimation 74

4.4. Bootstrap algorithm 77
4.4.1. BDT for a single data set 80

4.5. Data analysis: Asthma medication during pregnancy 82
4.5.1. Data description 82
4.5.2. Results of the Analysis 83
4.5.3. Bootstrap diagnostic test 85

4.6. Discussion 86

Chapitre 5. Doubly Robust Adaptive LASSO for Effect Modifier Discovery 89

5.1. Introduction 91

5.2. Methods 93
5.2.1. The framework 93
5.2.2. Adaptive LASSO 94
5.2.3. Highly Adaptive LASSO (HAL) 95
5.2.4. Selective inference 95
5.2.5. The model 96
5.2.5.1. Model definition 96
5.2.5.2. Estimation 98

5.3. Simulation study . . . 100
5.3.1. Data generation and parameter estimation . . . 100
5.3.2. Simulation results . . . 102
5.4. Data analysis: Asthma medication during pregnancy . . . 107
5.4.1. Data . . . 107
5.4.2. Results . . . 108
5.5. Discussion . . . 109
5.6. Appendix . . . 110
5.7. Addenda . . . 120
Chapitre 6. Data Integration through outcome adaptive LASSO and a collaborative propensity score approach . . . 123
6.1. Introduction . . . 126
6.2. Methods . . . 129
6.2.1. The framework . . . 129
6.2.2. Estimation of the propensity score . . . 130
6.2.2.1. Variable selection for propensity score . . . 131
6.2.2.2. Implementation of OALASSO . . . 132
6.2.2.3. SCAD variable selection for propensity score . . . 133
6.2.3. Inverse weighted estimators . . . 133
6.2.4. Augmented Inverse Probability Weighting . . . 134
6.3. Simulation study . . . 135
6.3.1. Data generation and parameter estimation . . . 135
6.3.2. Results . . . 137

6.4. Data Analysis 141
6.5. Discussion 143
6.6. Appendix 144
Chapitre 7. Conclusion générale et discussion 147
7.1. Résumé des résultats principaux 147
7.2. Limites et perspectives 150
Références bibliographiques 153

Liste des tableaux

1 Median and mean bias, median squared error and coverage for different bounds of g_n. Estimates taken over 500 generated datasets for different sample sizes, n..... 76

2 Results from one data set (estimates of the average treatment effect and standard error)..... 80

3 Results from the BDT and alternative method used on a single simulated data set investigating the absolute average bias, mean squared error and coverage for IPTW and TMLE for different bounds of g_n..... 81

4 Estimates of the effect of exposure to ICS on birth weight ($n = 4791$)..... 84

5 BDT results investigating the absolute average bias, root mean squared error and the percent coverage for TMLE, C-TMLE and IPTW. 85

1 Simulation results (Data generating scenario 1). Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: average estimated value of the coefficients of the MSM, $\%Cov$: percent coverage of the selective confidence interval $\times 100$ (Standard CI for the linear model case), $\%sel$: percent selection of variables $\times 100$, FCR: False coverage rate $\times 100$, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$ 112

2 Simulation results (Data generating scenario 2). Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval $\times 100$, %sel: percent selection of variables $\times 100$, FCR:

False coverage rate ×100, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$ 113

3 Simulation results (Data generating scenario 3). Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval ×100, %sel: percent selection of variables ×100, FCR: False coverage rate ×100, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$ 114

4 Simulation results for smaller sample size ($n = 100$). Estimates taken over 500 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval ×100, %sel: percent selection of variables ×100, FCR: False coverage rate ×100, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$ 115

5 Simulation results (Data generating scenario 1 with 50 noise covariates). Estimates taken over 100 generated datasets. $\widehat{\boldsymbol{\beta}}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval, %sel: percent selection of variables, FCR: False coverage rate, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0, ..., 0)$ 116

6 Baseline Characteristics of mothers in the cohort extraction ($N = 4{,}707$) 117

7 Estimates of the coefficients associated with interaction terms using naive linear model ($n = 4707$)..... 118

8 Estimates of the selected MSM coefficients using adaptive lasso ($n = 4707$) with 95% Post selection interval for the selected variables. *: means significant variables 118

9 Résultat avec un MSM non linéaire. GAM (Generalized Additive Model/Modèle additif généralisé)..... 121

1 Observed data structure 130

2 Scenario 1: Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; IPW-LASSO: IPW with LASSO regression for propensity score; IPW-OALASSO: IPW with OALASSO regression for propensity score; AIPW-Logistic: AIPW with logistic regression for propensity score; AIPW-LASSO: AIPW with LASSO regression for propensity score; AIPW-OALASSO: AIPW with OALASSO regression for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score; AIPW-Yang: Yang's proposed AIPW. . . 139

3 Scenario 2: Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; IPW-LASSO: IPW with LASSO regression for propensity score; IPW-OALASSO: IPW with OALASSO regression for propensity score; AIPW-Logistic: AIPW with logistic regression for propensity score; AIPW-LASSO: AIPW with LASSO regression for propensity score; AIPW-OALASSO: AIPW with OALASSO regression for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score; AIPW-Yang: Yang's proposed AIPW. . . 139

4 Scenario 3: Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; IPW-LASSO: IPW with LASSO regression for propensity score; IPW-OALASSO: IPW with OALASSO regression for propensity score; AIPW-Logistic: AIPW with logistic regression for propensity score; AIPW-LASSO: AIPW with LASSO regression for propensity score; AIPW-OALASSO:

AIPW with OALASSO regression for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score; AIPW-Yang: Yang's proposed AIPW. . . 140

5 Scenario 4 (non-linear model setting): Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; AIPW-Logistic (1): AIPW with logistic regression for propensity score and a misspecified model for the outcome; AIPW-Logistic (2): AIPW with logistic regression for propensity score and HAL for the outcome model; AIPW-Benkeser: AIPW with the collaborative propensity score. 140

6 Point estimate, standard error and 95% Wald confidence interval. IPW-Logistic (Grp LASSO/OA Grp LASSO): IPW with logistic regression (Group LASSO/outcome adaptive Group LASSO) for propensity score; AIPW-Logistic (Grp LASSO/OA Grp LASSO): AIPW with logistic regression (Group LASSO/outcome adaptive Group LASSO for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score. 143

7 **Distributions of common covariates from the two samples.** 144

8 Distributions of common covariates from the two samples. 145

Liste des figures

1 Boxplots of the ATE with different bounds on g_n for IPTW, TMLE and C-TMLE. $n = 1000$. 77

2 Density plots of the log of the true and estimated weights for: (a) treatment A=1 in subset of patients with A=1, (b) treatment A=0 in subset of patients with A=0, (c) treatment A=1 for all subjects and (d) treatment A=0 for all subjects. 78

1 Simulation results illustrations (Data generating scenario 1). Box plots of 1000 MSM coefficients estimates for the true EMs ($V^{(1)}$,$V^{3)}$). The true values of the coefficients are $(0.5, 1)$. Notation: Qcgc: models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL, NLin: Naive linear model, CLin: Correct linear model. $\%sel$: percent selection of a covariate $\times 100$, $\%cov$: coverage rate of the confidence interval of a coefficient estimate $\times 100$, FCR: False coverage rate of the model $\times 100$. 104

2 Simulation results illustrations (Data generating scenario 1). Box plots of 1000 MSM coefficients estimates for the non-EMs ($V^{(2)}$,$V^{4)}$). The true values of the coefficients are $(0, 0)$. Notation: Qcgc: models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL, NLin: Naive linear model, CLin: Correct linear model. $\%sel$: percent selection of a covariate $\times 100$, $\%cov$: coverage rate of the confidence interval of a coefficient estimate $\times 100$, FCR: False coverage rate of the model $\times 100$. 105

3 Simulation results for high-dimensional setting (Data generating scenario 1). Box plots of MSM coefficients estimates over 100 simulations for the potentials EMs $\boldsymbol{V} = (V^{(1)},V^{(2)},V^{(3)},V^{(4)})$. The true values of the coefficients are $(0.5,0,1,0)$. Notations: HAL: $\bar{Q}$ and g are estimated with HAL, $\%sel$: percent selection $\times 100$, $\%cov$: coverage rate $\times 100$, FCR: False coverage rate $\times 100$. 106

4 Illustrations for high-dimensional setting. Box plots of the MSM coefficients estimates over 100 simulations for the 50 noise covariates (for both $n = 1000$ and $n = 10000$). The true values of the coefficients are (0,...,0). 106

5 Percent coverage of the selective confidence interval associated to V_1 and V_3 for different sample size. Notation: Qcgc: models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL. 111

1 Population and observed samples. 129

2 Percent selection of each variable into the propensity score model over 1000 simulations under scenarios 1-3. 141

Liste des sigles et des abréviations

AIPTW	Estimateur par augmentation de l'inverse de la probabilité de traitement, de l'anglais *Augmented Inverse Probability of Treatment Weighting*
ATE	EffSergio90@et moyen du traitement, de l'anglais *Average treatment effect*
BDT	Bootstrap Dagnostic Tool
CATE	Effet moyen du traitement conditionnel, de l'anglais *Conditional average treatment effect*
C-TMLE	Estimateur par maximum de vraisemblance ciblé collaboratif, de l'anglais *Collaborative Targeted Maximum Likelihood Estimation*
ECRs	Essais cliniques randomisés, de l'anglais *Randomized Clinical Trials (RCTs)*

EM	Modificateur d'effet, de l'anglais *Effect modifier*
FCR	Taux de faux recouvrement, de l'anglais *False Coverage Rate*
Grp LASSO	Group LASSO
HAL	Highly Adaptive LASSO
IPTW	Estimateur par l'inverse de la probabilité de traitement, de l'anglais *Inverse Probability of Treatment Weighting*
LASSO	Least Absolute Shrinkage and Selection Operator
LFS	Labor Force Survey
MSM	Modèles structurels marginaux, de l'anglais *Marginal structural model*
MSE	Erreur quadratique moyenne, de l'anglais *Mean Square Error*

OALASSO	Outcome Adaptive LASSO
OLS	Moindres carrés ordinaires, de l'anglais *Ordinary Least Square*
SE	Erreur type, de l'anglais *Standard Error*
SL	Super Learner
STD	Écart type, de l'anglais *Standard Deviation*
SUTVA	Valeur de traitement des unités stable, de l'anglais *Stable unit treatment value assumption*
TIO	Trust In Others
TMLE	Estimateur par maximum de vraisemblance ciblé, de l'anglais *Targeted Maximum Likelihood Estimation*

Remerciements

Je tiens d'abord à remercier ma directrice de recherche, Dr Mireille E. Schnitzer de m'avoir accepté dans son équipe pour entreprendre mes études doctorales. Ton humilité et ta présence m'ont aider à mener ce voyage à bon port. J'ai beaucoup appris dans le domaine de l'inférence causale et de la biostatistique et je t'en suis infiniment reconnaissant.

Mes remerciements vont aussi à mon comité consultatif (Dr Lucie Blais et Dr Geneviève Lefebvre) pour toutes les suggestions et apports dans le cadre de cette thèse.

Je remercie également mes amis et collègues de la Faculté avec qui j'ai eu la chance de collaborer et qui ont contribué d'une façon ou d'une autre à l'aboutissement de ce travail: Dr Flory Tsobo Muanda, Dr Maëlle Dandjinou, Steve Ferreira.

Je tiens à remercier la Faculté de pharmacie de l'Université de Montréal, le conseil de recherches en sciences naturelles et en génie du Canada, l'institut de recherche en santé du Canada pour leur soutien financier.

Je souhaite exprimer mon infinie gratitude à ma famille. À mes parents qui se sont beaucoup sacrifiés pour moi et qui ont toujours été là pour m'encourager à viser plus haut. Aussi, à mes soeurs Sylvana et Massama, mon frère Elom, merci.

Enfin, à ma princesse Cathou, avec qui j'ai débuté ce voyage, pour tout ton soutien, ton amour. À Caleb et Elza qui nous ont rejoint en cours de route. Je vous aimes !

Je ne saurai terminer sans rendre grâce à mon Dieu très haut. À toi la Gloire !

Chapitre 1

Introduction

Les essais cliniques randomisés (ECRs) constituent le devis par excellence pour évaluer l'efficacité d'un médicament ou d'une intervention. Toutefois, pour examiner l'innocuité d'un médicament, le recours aux études observationnelles s'avère indispensable. Pour des raisons éthiques, économiques ou de faisabilité, les ECRs ne sont toujours pas réalisables notamment au sein des populations vulnérables. Notons par exemple les personnes âgées ou les femmes enceintes, qui sont généralement exclues de ces études. Ainsi pour obtenir des données probantes, les chercheurs réalisent des études observationnelles, en utilisant des banques de données médico-administratives qui ont de grandes tailles d'échantillons, permettant ainsi d'étudier des issues adverses rares et l'efficacité des médicaments dans les conditions réelles d'utilisation. C'est aussi le cas des statisticiens d'enquête qui souhaitent ainsi obtenir des statistiques officielles actuelles et à moindre coût. Cependant, les bases de données administratives ne viennent toutefois pas sans défis. Une des plus grandes limitations des études observationnelles est le manque de randomisation qui empêche les facteurs de risque mesurés et non mesurés d'être similaires entre les groupes de comparaisons. Ceci pourrait conduire à des estimations biaisées, qui auront des conséquences fâcheuses sur la santé des patients. On pourrait, par exemple, conclure à tort qu'un médicament n'est pas néfaste, alors qu'il l'est, en réalité. De plus, les bases de données administratives étant conçues pour d'autres objectifs, elles ne contiennent pas toujours les informations sur les habitudes de vie (tabac,

alcool) et les données ne sont pas toujours collectées avec exactitude (les données sur l'exposition, l'issue ou les facteurs de confusion). Par ailleurs, elles ne couvrent généralement qu'une portion de la population. Par conséquent, la présence d'un biais de confusion et de sélection ne doit pas être exclue lorsqu'on utilise ces sources de données, ce qui rend difficile l'obtention d'estimation sans biais. Par exemple, on peut citer la protection apparente observée dans les études sur l'effet des statines sur la maladie de Parkinson dû au fait que l'indication des statines n'ait pas été ajustée [**57**]. Supposons qu'on désire estimer l'effet du tabac pendant la grossesse sur le poids du nouveau-né. En utilisant un fichier administratif ne contenant que de naissances vivantes, on pourrait conclure à tort un effet protecteur du tabac sur le poids du nouveau-né [**109**].

Plusieurs méthodes statistiques avancées ont été développées ces dernières années pour estimer des paramètres d'intérêt sans bias à l'instar des essais cliniques. Parmi ces méthodes, nous pouvons citer la méthode de pondération inverse (inverse probability of treatment weigthing; IPTW [**107**]) qui est très utilisée dans les études observationnelles [**46, 91, 12, 105**] et le calcul g [**48**]. En effet, l'IPTW utilise le score de propension qui est la probabilité de recevoir le traitement reçu conditionnellement aux caractéristiques du patient. Le calcul-g quant à lui se base sur le modèle de l'issue, soit l'espérance de l'issue conditionnellement aux caractéristiques du patient. Ces méthodes permettent de rendre comparable les caractéristiques mesurées entre le groupe d'exposition et le groupe contrôle. Toutefois, l'estimation du score de propension et de l'espérance conditionnelle de l'issue se font par une modélisation et une mauvaise spécification de ces dernières aura pour conséquence la production d'un estimateur biaisé. Pour pallier ce problème, des estimateurs semi-paramétriques, asymptotiquement linéaires et doublement robustes ont été proposés. Le terme *doublement robuste* signifie qu'on est en mesure d'obtenir un estimateur consistant si le modèle du traitement ou de l'issue est bien spécifié. Parmi les estimateurs semi-paramétriques, nous citons, la version augmentée de IPTW (augmented inverse probability of treatment; AIPTW [**51**]) et l'estimateur par maximum de vraisemblance ciblée (targeted maximum likelihood estimation; TMLE [**73**]). L'AIPTW est une extension du IPTW qui prend en compte le modèle

de l'issue et est basée sur les équations d'estimations. Par contre, le TMLE est un cadre permettant de développer des estimateurs par maximum de vraisemblance. L'AIPTW et le TMLE sont des estimateurs semi-paramétriques, doublement robustes et localement efficaces, pour divers paramètres causaux. Ils permettent également l'utilisation de méthodes non paramétriques ou d'apprentissage automatique, lors de l'estimation des paramètres de nuisance. Cependant, ces méthodes peuvent être instables lorsqu'il y a peu de données dans une région de l'espace des co-variables. Dans la littérature, très peu d'outils [**78**] existent pour diagnostiquer une mauvaise estimation d'un paramètre causal dans ce contexte.

L'effet causal moyen d'une exposition à un traitement sur une issue est un paramètre important souvent estimé dans plusieurs problèmes de causalité. Toutefois, connaître l'effet dans les sous-groupes de la population permet d'identifier les parties de la population qui en bénéficieront le plus. De plus, il permet aussi de cibler, de manière optimale, les traitements à donner selon les groupes d'appartenance. Néanmoins, très peu de développements intègrent les méthodes de régularisation pour la sélection d'un modificateur d'effet [**33, 114**] en pharmacoépidémiologie.

Comme discuté plus haut, le biais de sélection résulte de l'utilisation de données administratives qui ne couvrent pas toute la population à l'étude. Ainsi, il y a un défi de taille à relever lors de l'estimation. Récemment, une méthode [**121**] fut proposée intégrant un échantillon probabiliste de référence. Cependant, cette méthode présuppose une bonne connaissance des co-variables à ajuster dans leur modèle. Dans la littérature, peu de travaux de recherche utilisent les méthodes de régularisation pour la sélection de variables lors de l'intégration d'un échantillon probabiliste [**101**] avec une source administrative.

Dans cette thèse, nous nous intéressons à l'instabilité des estimateurs doublement robustes quand on est proche de violer la positivité en pratique, la sélection de modificateurs d'effet et la combinaison de données administratives et de données d'enquêtes pour l'estimation d'un paramètre de la population. Dans le chapitre 2, nous présentons très brièvement les défis reliés à l'utilisation de données administratives, une revue de la littérature sur l'inférence causale dans un contexte d'une exposition qui ne varie pas dans le temps et quelques

méthodes d'inférence. Au chapitre 3, nous introduisons les objectifs de ce mémoire. Dans chapitre 4 qui est le manuscrit publié dans la revue *Statistical Methods in Medical Research*, nous démontrons comment le TMLE et l'IPTW peuvent être déstabilisés par l'inverse du score de propension lorsqu'on utilise des méthodes d'apprentissage automatique pour l'estimation de cette dernière. Dans le chapitre 5, nous proposons une procédure doublement robuste pour la sélection de modificateurs d'effet et l'estimation d'un effet conditionnel. Ce manuscrit est en révision pour la revue *International Journal of Biostatistics*. Le chapitre 6 décrit une méthode novatrice de combinaison de données administratives et de données d'enquêtes pour l'ajustement de biais de sélection. Ce manuscrit sera soumis dans la revue *Journal of Survey Statistics and Methodology*. Finalement, le chapitre 7 consiste en une discussion des résultats obtenus.

Chapitre 2

Revue de la littérature

La revue de la littérature est divisée en trois sections principales. La première section porte sur les données non probabilistes et les défis qui y sont reliés. La deuxième section porte sur l'inférence causale et les hypothèses sous-jacentes pour l'identification et l'estimation d'un paramètre causal. La troisième section décrit les méthodes d'inférence causale existantes et utilisées dans le cadre de ce mémoire.

2.1. Échantillon non-probabiliste ou données administratives

Depuis quelques années, les données provenant d'enquêtes non probabilistes sont de plus en plus utilisées en pratique. On parle par exemple de données administratives ou observationnelles provenant des organisations gouvernementales, des entreprises et d'autres organismes dans plusieurs domaines tels l'impôt, la santé ou l'éducation, et recueillies pour d'autres fins. En santé par exemple, ces données proviennent des registres du ministère de la Santé et des bases de données d'assureurs ou de compagnies privées de services en santé. Les estimations produites avec des sources de données non probabilistes sont sujettes à des défis [**64**]. Pour l'estimation d'effets relatifs, nous pouvons citer comme défi le manque de randomisation qui empêche les facteurs de risque mesurés et non mesurés d'être similaires

entre les groupes de comparaison (confusion). Ceci conduit inévitablement à des estimations biasées. De plus, les bases de données administratives étant conçues pour d'autres objectifs, elles ne couvrent pas toute la population à l'étude (erreur de couverture) et ne contiennent pas de poids d'échantillonnage. Par conséquent, il en résulte une présence de biais de sélection. L'utilisation de données sujettes à un biais de sélection peut entrainer une distorsion (sur/sous-estimation) du paramètre d'intérêt. Une façon de prévenir ce biais *a priori* est de recourir à des échantillons probabilistes, c'est à dire, tirer les sujets entrant dans notre étude de manière aléatoire. Vu le coût associé à cette démarche et l'abondance des données administratives de nos jours, d'autres approches sont de plus en plus considérées en pratique. Un autre défi de taille lors de l'utilisation de données administratives est la présence d'erreurs de mesure [**77**]. Dans ce contexte, il est clair que l'utilisation des données administratives pour produire des estimations nécessite un cadre d'inférence statistique valide.

Dans la suite de ce chapitre, nous discuterons de l'inférence causale dans un contexte de traitement binaire et des méthodes d'inférences permettant d'estimer des paramètres causaux avec des données administratives. Dans nos études, nous avons simplifié la réalité en supposant l'absence d'erreur de mesure. En d'autres termes, nous supposons que les données se trouvant dans nos fichiers administratifs sont de bonne qualité.

2.2. Inférence causale

L'inférence causale est un domaine dans lequel on cherche à estimer la relation de cause à effet entre deux variables, si causalité il y a. Pour y arriver, il est important de se questionner sur l'issue qui aurait été obtenue en cas d'exposition à un traitement et dans le cas contraire. Par exemple, si une personne malade se sent mieux après avoir pris un médicament (donc exposé à ce médicament), est-ce vraiment ce médicament qui l'a aidée ? Que lui serait-il arrivé si elle n'avait pas pris le médicament ? Le modèle de Rubin-Neyman [**21**] propose un cadre permettant de décrire un effet causal de manière intuitive.

Supposons que l'on s'intéresse à l'effet causal d'une exposition ponctuelle A sur une issue Y. Dans la suite du mémoire, supposons que A est binaire. Pour un individu i, nous observons

$A_i = 1$ si l'individu i est exposé (ou traité) et $A_i = 0$ sinon. Ainsi, pour l'individu i, dénotons par Y_i^0, respectivement Y_i^1, l'issue potentielle sous l'exposition $a = 0$, respectivement $a = 1$. De manière générale, les variables Y^a dénotent les issues potentielles ou contrefactuelles obtenues par l'individu dans le cadre d'une intervention hypothétique $A = a$. Elles sont dites potentielles, car elles décrivent une valeur qu'on aurait observée si l'exposition A avait été a, ce qui n'est pas forcément observé en réalité. Pour un individu i, l'effet causal individuel est la différence entre les deux issues potentielles $Y_i^1 - Y_i^0$. Remarquons que cet effet n'est pas directement observable en pratique. En effet, à un temps t et pour un traitement binaire, un individu ne peut qu'être soit traité ou non traité, mais pas les deux à la fois. Ainsi, nous ne pouvons observer que Y_i^1 ou Y_i^0. Pour une population d'intérêt, on peut ainsi définir par l'espérance de l'issue contrefactuelle $E(Y^a)$, l'issue moyenne observée si tous les individus de la population d'intérêt avaient reçu le traitement a. L'effet moyen du traitement sur une issue aléatoire Y est défini par $E(Y^1 - Y^0) = E(Y^1) - E(Y^0)$. Toutefois, on ne peut pas calculer les deux espérances pour la même population.

Les études cliniques randomisées sont généralement considérées comme le devis par excellence pour des questions de causalité. En effet, de manière aléatoire, on attribue le traitement $A = 1$ à une partie de la population. La randomisation faite *a priori* rend les groupes comparables en termes de leurs caractéristiques. Ainsi, sous certaines conditions, l'effet moyen du traitement peut être directement estimé sans biais par la différence des issues moyennes dans les deux groupes. Les études cliniques randomisées sont souvent irréalisables et souffrent généralement de problème d'adhérence au traitement. Dans ces situations, les données observationnelles ou administratives dans lesquelles le traitement n'est pas randomisé servent à estimer les paramètres causaux. La différence des issues moyennes dans les deux groupes dans ce contexte est biaisée pour l'effet moyen du traitement. Cependant, il est possible d'obtenir une estimation sans biais sous certaines présuppositions causales et en utilisant des méthodes de modélisation.

2.2.1. Hypothèses d'identifiabilité

Dans le but d'estimer des paramètres causaux en utilisant des données observationnelles, plusieurs présuppositions liées aux données sont requises. Supposons que nos données sont représentées par $O = (\boldsymbol{W}, A, Y)$ et distribuées suivant la loi de probabilité P_0 , c'est à dire $O \sim P_0$, avec $\boldsymbol{W}$ le vecteur de co-variables, A l'indicateur du traitement qui est 1 si le patient est traité et 0 sinon, Y la variable réponse. Nous dénotons par $O_i = (\boldsymbol{W}_i, A_i, Y_i)$, la i-ième observation. Une réalisation des variables aléatoires sera notée en minuscule. Pour l'effet moyen du traitement, les présuppositions sont: 1) la non-interférence, 2) la cohérence, 3) la positivité et 4) l'ignorabilité ou l'interchangeabilité conditionnelle [**69**].

La non-interférence signifie que l'issue potentielle Y_i^a d'un individu i est indépendante de l'exposition d'autres individus $j \neq i$ [**20**]. La cohérence est l'hypothèse qui stipule que l'issue observée Y d'un individu i sous l'exposition $A = a$ est identique à l'issue potentielle Y_i^a [**106**]. Ainsi, pour $A_i = a$, $Y_i^a = Y_i$. Cette hypothèse suppose également qu'il n'y a qu'une seule version d'un traitement a ou que les versions n'ont pas d'importance pour l'issue potentielle. Dans le cas d'une exposition binaire, elle est formulée par $Y = AY^1 + (1-A)Y^0$. Par conséquent, l'issue contrefactuelle Y^a, sous l'exposition $A = a$ est bien égale à l'issue observée. La non-interférence et la cohérence ensemble sont nommées l'hypothèse de "valeur de traitement des unités stable" (SUTVA) [**20**]. La positivité ou "l'hypothèse de traitement expérimental" exige que la probabilité de recevoir tout niveau de traitement conditionnellement aux co-variables soit positive pour chaque individu de la population. Soit $\boldsymbol{w}$ une observation du vecteur aléatoire $\boldsymbol{W}$. La positivité théorique stipule que $Pr(A = a|\boldsymbol{W} = \boldsymbol{w}) > 0$, $\forall A = a$ et $f_{\boldsymbol{W}}(\boldsymbol{w}) > 0$. La positivité peut être valide théoriquement, mais violée en pratique. Cette violation se produit en pratique lorsque certains individus ont une probabilité estimée de recevoir l'un ou l'autre des traitements conditionnellement à $\boldsymbol{W}$ proche de zéro. Ceci peut être dû à un nombre insuffisant d'individus dans certains groupes de la population d'intérêt. La positivité théorique est nécessaire pour l'identification ou la définition d'un paramètre causal. L'ignorabilité ou l'interchangeabilité conditionnelle est une hypothèse formulée par Rubin [**22**] dans le contexte des données manquantes. Elle stipule que:

$$Y^a \perp A|\boldsymbol{W},$$

c'est-à-dire que les contrefactuelles sont indépendantes du traitement reçu conditionnellement à $\boldsymbol{W}$. Ici, $\boldsymbol{W}$ est dénomé facteur de confusion et est l'ensemble des co-variables observées avant le traitement pour lesquelles une fois conditionné, rendent le traitement A et l'issue potentielle Y^a indépendants. Cette hypothèse signifie que dans un même groupe $\boldsymbol{W} = \boldsymbol{w}$, les individus exposés et non-exposés ont la même distribution de leurs issues potentielles. L'hypothèse d'ignorabilité est souvent dénommée l'hypothèse d'absence de facteurs de confusion non mesurés. Dans les études randomisées, cette hypothèse est respectée naturellement vu que le traitement est assigné de manière aléatoire. En effet, l'ignorabilité nous permet d'écrire $P(A = a|Y^a = y,\boldsymbol{W} = \boldsymbol{w}) = P(A = a|\boldsymbol{W} = \boldsymbol{w})$. Ainsi, la probabilité de recevoir l'un ou l'autre des traitements ne dépend pas des Y^a, mais uniquement de $\boldsymbol{W}$. Dans le cas d'une étude dans laquelle la randomisation est faite de manière aléatoire, le traitement auquel un patient est assigné est indépendant des Y^a, i.e $Y^a \perp A$ et c'est la raison pour laquelle une simple différence des moyennes entre les deux groupes suffit pour estimer l'effet moyen du traitement. L'ignorabilité est donc une conséquence directe. Dans le cas des études observationnelles, puisque les groupes n'ont pas été randomisés, cette hypothèse n'est pas généralement respectée, mais peut être contrôlée *a posteriori* lors de l'estimation, si tous les facteurs de confusion sont mesurés.

Sous ces conditions, il est donc possible de définir un paramètre causal avec des données observationnelles. Ainsi, l'issue moyenne observée si tous les individus de la population d'intérêt avaient reçu le traitement $A = a$ peut être identifiée comme suit:

$$\begin{aligned} E(Y^a) &= E_{\boldsymbol{W},0}\{E_0(Y^a|\boldsymbol{W})\} \\ &= E_{\boldsymbol{W},0}\{E_0(Y^a|\boldsymbol{W}, A = a)\} \\ &= E_{\boldsymbol{W},0}\{E_0(Y|\boldsymbol{W}, A = a)\} \end{aligned}$$

avec E_0 une espérance par rapport à l'issue et $E_{\boldsymbol{W},0}$ une espérance par rapport aux co-variables $\boldsymbol{W}$ dans la population selon la présupposition 3).

L'inférence causale peut être donc scindée en deux parties: l'identification et l'estimation. L'identification nous permet de définir ce que nous devons estimer avec des données observationnelles, mais pas comment l'estimer. La deuxième partie n'est qu'un problème d'estimation de fonction.

Pour une issue Y et un traitement binaire A, un autre paramètre d'intérêt qui sera discuté dans cette thèse est l'effet moyen du traitement conditionnel (*conditional average treatment effect*, CATE) définit comme suit

$$CATE(\boldsymbol{w}) = E(Y^1 - Y^0 | \boldsymbol{W} = \boldsymbol{w}).$$

En effet, le CATE mesure l'effet moyen du traitement dans un sous groupe de la population définit par $\boldsymbol{W} = \boldsymbol{w}$. Ce paramètre est intéressant en général, car l'effet individuel pourrait être différent de l'effet causal moyen. Un exemple qu'on voit souvent en pratique est que le traitement n'ait aucun effet en moyenne dans la population, mais par contre que cet effet soit bénéfique dans un sous groupe de cette dernière. D'autres paramètres d'intérêts existent également pour une issue binaire [**69**].

2.2.2. Modèles structurels marginaux

Comme discuté plus haut, les issues potentielles décrivent l'issue dans le cadre d'un intervention hypothétique. En pratique, une seule valeur (Y^1 ou Y^0) est observée pour chaque individu. Par conséquent, poser un modèle pour $E(Y^a)$ n'est pas triviale. Les modèles structurels marginaux (MSM) [**52, 50**] constituent une classe de modèles ou de paramètres causaux permettant de modéliser l'issue contrefactuelle Y^a. Les MSM sont dits marginaux, car ils permettent de modéliser la distribution marginale des issues potentielles, $E(Y^a)$, en absence de confusion, comparés à une régression linéaire standard de l'issue sur le traitement et les co-variables. Par exemple, un MSM linéaire sur le traitement aurait la forme suivante:

$$E(Y^a) = \beta_0 + \beta_1 a,$$

où Y^a est l'issue contrefactuelle qui n'est pas observée. Si a est binaire, β_1 représente l'effet causal moyen. Ce modèle est dit saturé, car il contient deux paramètres inconnus β_0 et β_1 et $E(Y^a)$ prend seulement deux valeurs (pour $a = 0$ et $a = 1$, respectivement). Les MSM se sont montrés efficaces, en particulier lorsqu'il y a médiation entre le traitement et l'issue, dans le cadre de données longitudinales pour la réduction du biais de confusion [**70**].

2.2.3. Modificateur d'effet

Dans plusieurs problèmes de causalité, on s'intéresse à l'effet causal moyen d'un traitement A sur une issue Y. Dans certaines situations, il est possible qu'en moyenne, le traitement A n'ait aucun effet sur Y dans la population, mais que cet effet existe bel et bien dans certains sous-groupes de la population. Par conséquent, il est important de connaître l'effet dans ces sous-groupes et comprendre comment il varie selon les caractéristiques des individus de la population. Ce sont des questions que l'on se pose en médecine personnalisée où l'on désire faire des interventions spécifiques dans certains sous-groupes de la population. Les variables associées à ces sous-groupes dont les effets diffèrent sont appelées des modificateurs d'effet. L'identification d'un modificateur d'effet permet aussi de comprendre les facteurs biologiques, sociaux qui affectent une issue donnée. Plus spécifiquement, on parle de modification d'effet [**111**], quand l'effet d'un traitement A sur une issue Y dépend d'un autre facteur V observé avant le traitement. Tout d'abord, il peut être détecté sous une forme additive ou multiplicative. Supposons que V est binaire. On parle de modification de l'effet additif quand

$$E(Y^1 - Y^0|V = 1) \neq E(Y^1 - Y^0|V = 0)$$

et multiplicatif quand

$$\frac{E(Y^1|V = 1)}{E(Y^0|V = 1)} \neq \frac{E(Y^1|V = 0)}{E(Y^0|V = 0)}.$$

Dans le cadre de ce mémoire, bien qu'il existe d'autre forme sous laquelle il est possible d'identifier un modifcateur d'effet, nous nous concentrons sur la détection d'un modificateur d'effet sous une forme additive. La forme additive est considérée comme la plus facile à

interpréter [**111**] et très utile en santé publique pour évaluer les sous groupes de la population nécessitant une intervention.

2.2.3.1. Détection d'un modificateur d'effet et estimation du CATE

Plusieurs méthodes existent pour détecter un modificateur d'effet. Une étude par stratification est la façon la plus simple d'identifier un modificateur d'effet. Cependant, cette méthode devient irréalisable en présence de plusieurs facteurs V ou quand V est multidimensionnel. Une analyse de régression avec un terme d'interaction dans le modèle ($E(Y|A,V) = \alpha + \beta_A A + \beta_V V + \beta_{AV} A * V$) est aussi une méthode facile pour détecter un modificateur d'effet. Toutefois, cette méthode ne permet pas de modéliser l'issue contrefactuelle Y^a et ne cible pas les paramètres d'un MSM à moins que le vrai modèle de l'issue soit connu et posé, ce qui en pratique est improbable. Par conséquent, la modélisation de la variable Y avec une régression simple ne réprésente pas la modification de l'effet causale. Les modèles structurels marginaux [**49**] sont des méthodes efficaces pour modéliser la modification d'effet. Les MSM peuvent servir à détecter un modificateur de l'effet causal qui n'est pas un confondant. Pour ce faire, il suffit d'inclure le facteur V dans un MSM comme suit:

$$E(Y^a|V) = \beta_0 + \beta_1 a + \beta_2 V + \beta_3 V a.$$

Il est important de noter que ce modèle est conditionnel, car on modélise Y^a conditionnellement à V. Ce modèle nous permet de comprendre comment la moyenne de l'issue contrefactuelle varie en fonction de V. Basé sur ce modèle, V est donc un modificateur d'effet si, $\beta_3 \neq 0$. Bien que Y^a ne soit pas observée et sous les présuppositions causales, il est possible d'estimer correctement les paramètres de ce modèle par des méthodes de modélisation.

D'autres méthodes furent développées autres que celles utilisant les MSM paramétriques ces dernières années pour détecter un modificateur d'effet et estimer un effet conditionnel. Green et Kern [**26**] ont utilisé un arbre de régression [**36**] basé sur le modèle bayésien pour modéliser le CATE. Imai et Ratkovic [**58**] ont adapté les machines à vecteurs de support

[**108**] pour la sélection d'un modificateur d'effet. Nie et Wager [**119**] ont développé un algorithme à deux étapes pour l'estimation de l'effet conditionnel. Luo et al., [**116**] ont utilisé une technique de réduction de dimension pour comprendre l'hétérogénéité. Wager et Athey [**99**] ont proposé une approche non-paramétrique pour l'estimation du CATE en utilisant les forêts aléatoires [**62**]. Powers et al., [**97**] ont adapté les régressions sur les splines [**39**] pour les mêmes fins. Zhao et al., [**84**] ont utilisé la transformation de Robinson [**82**] et la procédure *LASSO (Least Absolute Shrinkage and Selection Operator)* pour la détection d'un modificateur d'effet et l'estimation de l'effet conditionnel. Parmi ces méthodes, nous avons aussi des estimateurs doublement robustes. Par exemple, Lee et al., [**96**] ont développé une procédure permettant d'estimer une bande de confiance uniforme sur le CATE. De plus, Rosenblum et van der Laan [**76**] ont développé une approche pour estimer le CATE basée sur la méthode de maximum de vraisemblance. Plus récemment, Kennedy [**30**] a dérivé des bornes sur l'erreur d'estimation de l'effet conditionnel. Aucune des méthodes présentées ci-dessus n'est doublement robuste pour la sélection de modificateurs d'effet. Dans le cadre de cette thèse au manuscrit 2, nous proposerons une méthode doublement robuste de sélection de modificateurs d'effet et d'estimation du CATE utilisant une extension du LASSO. Notre proposition vient s'ajouter à la petite liste [**33, 114, 84**] de méthodes intégrant la régularisation pour l'estimation de l'effet conditionnel.

2.3. Méthode d'inférence

Dans la section précédente, nous avons donné quelques exemples de paramètres causaux et décrit les présuppositions causales nécessaires pour leurs identifications. La présente section contient quelques méthodes d'estimation des paramètres causaux.

2.3.1. Méthode de pondération inverse

La méthode de pondération inverse (Inverse Probability of Weighting, IPW) est l'une des méthodes populaires pour estimer des paramètres causaux, de par sa simplicité. Elle est populaire également, car elle peut être utilisée avec la majorité des logiciels standards. À

l'origine, la méthode de pondération inverse a été proposée par Horvitz et Thompson (1952, [**23**]) en échantillonnage. Pour estimer le total de la population, Horvitz et Thompson ont proposé de pondérer la valeur de chaque unité de l'échantillon par l'inverse de sa probabilité d'inclusion dans l'échantillon. Cette même technique est utilisée en échantillonnage pour ajuster une issue manquante en multipliant le poids de chaque répondant par l'inverse de sa probabilité de répondre. La même idée est utilisée en inférence causale si l'on suppose que le contrefactuel manquant peut être vu comme une donnée manquante.

En gros, l'IPTW permet de créer une pseudo population dans laquelle la distribution des covariables est similaire dans les deux groupes de traitement et représentative de la population cible. Ainsi, le biais de confusion serait éliminé dans cette population fictive. L'IPTW a été proposé en inférence causale par Robins [**52**] pour des données longitudinales. Dans le contexte d'un traitement A ponctuel et binaire, soit $g_0(1|\boldsymbol{W}) = P(A = 1|\boldsymbol{W})$ le score de propension [**83**]. Dans le but d'estimer l'issue contrefactuelle moyenne $E(Y^a)$, chaque observation est pondérée par $\omega_i = 1/g_n(a|\boldsymbol{W}_i)$, l'inverse de la probabilité de traitement reçu $A = a$ étant donné les facteurs de confusion, avec $g_n(a|\boldsymbol{W}_i) = \widehat{P}(A = a|\boldsymbol{W}_i)$ l'estimation de $g_0(a|\boldsymbol{W})$. Ainsi, l'estimateur de $E(Y^a)$ par la méthode IPTW est donné par:

$$1/n \sum_{i=1}^{n} \omega_i I(A_i = a) Y_i.$$

Par exemple, pour l'effet moyen du traitement, les observations avec $A_i = 1$ (respectivement $A_i = 0$) sont pondérées par $\omega_i = 1/g_n(1|\boldsymbol{W}_i)$ (respectivement $\omega_i = 1/g_n(0|\boldsymbol{W}_i)$). L'estimateur de l'ATE est donc:

$$1/n \sum_{i=1}^{n} (2A_i - 1)\omega_i Y_i.$$

L'estimateur par la méthode de pondération inverse dépend du score de propension et une bonne estimation de cette dernière est nécessaire pour bien estimer $E(Y^a)$. Les méthodes d'apprentissage automatique sont conseillées en pratique pour l'estimation du score de propension [**76**].

2.3.2. Méthode d'augmentation de l'inverse de la probabilité de traitement

2.3.2.1. Estimation des effects causaux par la régression

La méthode d'estimation basée sur la méthode de régression est encore dénommée la méthode de substitution (plug-in estimator) ou le calcul g [**48**]. C'est une méthode traditionnellement utilisée en sciences sociales. Elle est basée sur une formulation de la régression de la variable réponse Y sur le traitement A et la co-variable W. Plus précisément, nous avons démontré que $E(Y^a) = E\{E(Y|\boldsymbol{W}, A = a)\}$. Ainsi, l'effet moyen du traitement peut donc être défini par:

$$\psi_0 = E_{\boldsymbol{W},0}\{\underbrace{E_0(Y|A=1,\boldsymbol{W})}_{\bar{Q}_0(1,\boldsymbol{W})} - \underbrace{E_0(Y|A=0,\boldsymbol{W})}_{\bar{Q}_0(0,\boldsymbol{W})}\}$$

avec E_0 une espérance par rapport à l'issue et $E_{\boldsymbol{W},0}$ une espérance par rapport aux co-variables $\boldsymbol{W}$. Nous notons par $\bar{Q}_0(a,\boldsymbol{W}) = E_0(Y|A = a,\boldsymbol{W})$. L'estimation de l'ATE par le calcul g est donné par:

$$\psi_{g,n} = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n(1,\boldsymbol{w}_i) - \bar{Q}_n(0,\boldsymbol{w}_i)\}$$

où $\bar{Q}_n(1,\boldsymbol{w}_i)$ (respectivement $\bar{Q}_n(0,\boldsymbol{w}_i)$) est estimée par une régression de Y sur $\boldsymbol{W}$ parmi les patients traités (respectivement la régression de Y sur $\boldsymbol{W}$ parmi les patients non traités). $\bar{Q}_n(a,\boldsymbol{w}_i)$ peut être vu comme l'estimé de l'espérance conditionnelle de l'issue contrefactuelle qui est dépourvue de confusion. Ainsi, l'estimation de l'effet moyen du traitement est obtenu par la différence de l'espérance conditionnelle de l'issue contrefactuelle moyenne dans les deux groupes. Le calcul g sert également pour l'estimation des MSM [**53**]. Le calcul g donne une estimation dépourvue de biais tant et si longtemps que le modèle de l'issue est bien spécifié. Bien que la théorie nécessite l'utilisation des méthodes ayant une vitesse de convergence d'au moins $o(n^{-1/2})$ pour la modélisation de l'issue, les méthodes flexibles sont conseillées [**54**]. La procédure de rééchantillonnage est utilisée pour l'estimation de la variance.

2.3.2.2. Méthode d'augmentation de l'inverse de la probabilité de traitement

En pratique, il est souvent difficile de bien estimer la fonction de régression $E(Y|\boldsymbol{W}, A = a)$. Le calcul g produit des estimations biaisées quand cette dernière n'est pas correctement estimée. Dans le but de remédier à cela, des estimateurs semi-paramétriques ont été proposés. La méthode AIPTW [**51**] fut la première méthode de ce genre à être développée. Par ailleurs, nous pouvons citer la méthode d'estimation par maximum de vraisemblance ciblée (Targeted Maximum Likelihood Estimation; TMLE [**73**]). En effet, ces méthodes permettent de réduire le biais introduit dans le modèle de l'issue (suite à une mauvaise modélisation) et assurent la normalité asymptotique sous certaines conditions. L'estimateur AIPTW de l'effet moyen du traitement est le suivant:

$$\psi_{AIPTW,n} = \frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\left[\frac{A_iY_i}{g_n(1|\boldsymbol{w}_i)} - \frac{(1-A_i)Y_i}{1-g_n(1|\boldsymbol{w}_i)}\right]}_{(1)}\right.$$
$$\left.\underbrace{- \frac{(A_i - g_n(1|\boldsymbol{w}_i))}{g_n(1|\boldsymbol{w}_i)(1-g_n(1|\boldsymbol{w}_i))} \times [(1-g_n(1|\boldsymbol{w}_i))\bar{Q}_n(1,\boldsymbol{w}_i) + g_n(1|\boldsymbol{w}_i)\bar{Q}_n(0,\boldsymbol{w}_i)]}_{(2)}\right\}$$

où la première partie de l'équation correspond à l'estimateur IPTW de l'effet moyen du traitement. La seconde partie correspond à l'ajustement ou l'augmentation faite par la somme pondérée des deux fonctions de régression estimées. Il est important de noter que le vecteur de variables $\boldsymbol{W}$ à ajuster dans le score de propension et la fonction de régression ne doit pas nécessairement être similaire. L'estimateur donne la possibilité au chercheur d'ajuster les variables nécessaires dans les deux fonctions pour une estimation efficiente [**68, 104**] et consistante. L'estimateur AIPTW possède beaucoup de propriétés dont la double robustesse et la linéarité asymptotique. Un estimateur est dit doublement robuste s'il possède deux chances d'être consistant. Dans ce cas précis, AIPTW est consistant pour l'effet moyen du traitement si le score de propension ou le modèle de l'issue est correctement estimé. Une preuve de cette double robustesse est bien développée dans Tsiatis (2006, [**5**]).

L'estimateur AIPTW est également asymptotiquement linéaire. En effet, on peut écrire

$$\sqrt{n}(\psi_{AIPTW,n} - \psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IC(O_i) + o_p(1),$$

avec $o_p(1)$, un terme qui converge vers zéro en probabilité, $IC(O_i)$ qui est dénommé la fonction d'influence et est donnée par:

$$IC(O_i) \; = \; \big(\frac{I(A_i = 1)}{g_0(1|\boldsymbol{W}_i)} - \frac{I(A_i = 0)}{g_0(0|\boldsymbol{W}_i)}\big)(Y - \bar{Q}_0(A_i,\boldsymbol{W}_i)) + \bar{Q}_0(1,\boldsymbol{W}_i) - \bar{Q}_0(0,\boldsymbol{W}_i) - \psi_0$$

Ainsi, pour une taille d'échantillon n assez large, on peut éerire:

$$Var(\sqrt{n}(\psi_{AIPTW,n} - \psi_0)) = n \times Var(\psi_{AIPTW,n}) \approx \frac{1}{n} Var\Big(\sum_{i=1}^{n} IC(O_i)\Big).$$

La variable IC est une variable aléatoire dont la moyenne est nulle. Par conséquent, pour n grand, la variance dans l'échantillon de l'estimateur AIPTW est égale à $1/n^2 \sum_{i=1}^{n} IC(O_i)^2$. Si le score de propension et le modèle de l'issue sont correctement modélisés, l'estimateur AIPTW atteint la borne efficiente semi-paramétrique [**5**].

2.3.2.3. Note sur la linéarité asymptotique de l'AIPTW

L'estimateur AIPTW du paramètre ψ_0 est dit asymptotiquement linéaire, car il peut s'écrire ([**14**]) sous la forme:

$$\psi_{AIPTW,n} - \psi_0 = \frac{1}{n} \sum_{i=1}^{n} IC(O_i) + R_n + o_p(n^{-1/2})$$

avec

$$R_n = \int \big[\frac{g_n(1|\boldsymbol{W}) - g_0(1|\boldsymbol{W})}{g_n(1|\boldsymbol{W})}\big][\bar{Q}_n(a,\boldsymbol{W}) - \bar{Q}_0(a,\boldsymbol{W})] dQ_0(\boldsymbol{W})$$

et $dQ_0(\boldsymbol{W})$, la distribution marginale de $\boldsymbol{W}$. La clé pour montrer que l'AIPTW est asymptotiquement linéaire réside dans la convergence vers zéro du second terme $R_n = o_p(n^{-1/2})$. Ce terme atteint cette convergence si nous avons au moins:

$$||\bar{Q}_n(a,\boldsymbol{W}) - \bar{Q}_0(a,\boldsymbol{W})|| = o_p(n^{-1/4})$$

et

$$||g_n(1|\boldsymbol{W}) - g_0(1|\boldsymbol{W})|| = o_p(n^{-1/4})$$

selon la norme $L^2(P)$ définie par $||f||^2_{2,P_0} = \int f(z)^2 dP_0(z)$. Ainsi, nous pouvons écrire $\psi_{AIPTW,n} - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_p(n^{-1/2})$. Par conséquent, en utilisant le théorème limite central, on a:

$$\sqrt{n}(\psi_{AIPTW,n} - \psi_0) \sim N(0,\sigma^2),$$

avec $\sigma^2 = var_{P_0}(IC(O))$.

Il est important de noter que nous obtenons cette convergence dans la situation où les deux modèles (issue et traitement) sont correctement spécifiés et estimés ou si on utilise des modèles flexibles qui convergent assez rapidement afin d'avoir le produit des deux termes à $o_p(n^{-1/2})$. Cependant, si l'un des deux modèles n'est pas correctement estimé, l'un des deux termes de R_n ne converge pas vers zéro. Par conséquent, R_n tend vers zéro lentement et ceci affecte la linéarité asymptotique de l'estimateur [**14**]. Les modèles complètement non paramétriques peuvent être cités en exemple à cause de leur vitesse de convergence qui est plus lente que $n^{-1/4}$. Certains modèles non paramétriques font exceptions à la règle et sont discutés dans Kennedy [**33**]. Les méthodes flexibles comme le *Super Learner* [**28**] permettent aussi de satisfaire cette condition si au moins un des candidats dans la libraire le satisfait, c-à-d que ce candidat est correctement spécifié et tend vers zéro à une vitesse d'au moins $o(n^{-1/4})$.

2.3.3. Méthode d'estimation par maximum de vraisemblance ciblée

La méthode d'estimation par maximum de vraisemblance ciblée [**73**] est une procédure semi-paramétrique servant à construire des estimateurs de substitution (plug-in) efficaces. Contrairement à l'AIPTW qui est une méthode basée sur les équations d'estimations (*estimating equation method*), le TMLE est un estimateur par maximum de vraisemblance. Le TMLE possède les mêmes propriétés asymptotiques que l'estimateur AIPTW décrit dans la section 2.3.2.3. Plus précissement, le TMLE est doublement robuste en ce sens qu'il est

consistant pour l'ATE si le score de propension ou le modèle de l'issue est correctement estimé. Lorsque les deux modèles sont correctement estimés, le TMLE sera efficace, en ce sens que l'estimateur atteint la plus faible variance asymptotique parmi la classe des estimateurs réguliers asymptotiquement linéaires.

Supposons que P_0 qui appartient à une classe de modèle $\mathcal{M}$. Soit $\Psi : \mathcal{M} \to \mathbb{R}$ une fonction gâteaux-différentiable. Le paramètre d'intérêt (par exemple l'ATE) peut être représenté par $\psi_0 = \Psi(P_0)$, comme fonction de P_0. Soit $L(P)$, une fonction de perte qui pourrait être le log de la vraisemblance. L'idée du TMLE est de considérer un sous-modèle paramétrique $P_0(\epsilon)$ de P_0, tel que $P_0(\epsilon = 0) = P_0$ et que le score $\frac{d}{d\epsilon} L(P_0(\epsilon))|_{\epsilon=0}$ soit proportionnel à la fonction d'influence efficace $IC(P)$ (fonction d'influence ayant la variance minimale). En effet, le sous-modèle paramétrique est choisi de telle sorte que le score soit égal à la fonction d'influence efficace. Ainsi, pour un estimateur initial P_n^0 de P_0, on peut estimer $\epsilon_n^0 = \arg\min_\epsilon \sum_i L(P_n^0(\epsilon))(O_i)$ et ceci est équivalent à estimer ϵ_n^0 par la méthode du maximum de vraisemblance. Définisons $P_n^1 = P_n^0(\epsilon_n^0)$ la première étape du TMLE. On peut répéter ce processus jusqu'à obtenir une stabilisation, c'est-à-dire $\epsilon_n^k = \arg\min_\epsilon \sum_i L(P_n^k(\epsilon))(O_i)$ et $P_n^{k+1} = P_n^k(\epsilon_n^k)$ jusqu'à ce que $\epsilon_n^k = 0$. La dernière mise à jour P_n^k est le TMLE de P_0 et sera noté P_n^*. Par conséquent, le TMLE de ψ_0 sera donné par $\psi_{TMLE,n} = \Psi(P_n^*)$. Dans cette procédure d'itération, le TMLE est également solution de la fonction d'influence efficace égale à zéro $\sum_i IC(P_n^*)(O_i) = 0$, ce qui entraine le fait que le TMLE est efficace et asymptotiquement linéaire [**73**].

2.3.3.1. Cas de l'effet moyen du traitement

Dans le cas de l'ATE, le paramètre d'intérêt ψ_0 est fonction de $\bar{Q}_0(P_0)$. Nous pouvons écrire $\psi_0 = \Psi(\bar{Q}_0)$. Comme mentionné dans la section plus haut, un estimateur par substitution (calcul g) utilisant la régression est:

$$\psi_{g,n} = \frac{1}{n} \sum_{i=1}^{n} \{\bar{Q}_n(1,\boldsymbol{w}_i) - \bar{Q}_n(0,\boldsymbol{w}_i)\}.$$

Cet estimateur est considéré comme la première étape de l'estimateur TMLE $\bar{Q}_n^0(a,\boldsymbol{W}) = E_n(Y|A = a,\boldsymbol{W})$. Ainsi, le TMLE se propose de corriger l'erreur faite dans le premier

estimateur en choisissant un modèle paramétrique de fluctuation qui servira à mettre à jour cet estimateur initial. Un exemple de modèle de fluctuation est:

$$\text{logit}\{\bar{Q}_n^1(A,\boldsymbol{W})\} = \text{logit}\{\bar{Q}_n^0(A,\boldsymbol{W})\} + \epsilon_n^0 H_n(A,\boldsymbol{W})$$

avec $H_n(A,\boldsymbol{W}) = \frac{I(A=1)}{g_n(1|\boldsymbol{W})} - \frac{I(A=0)}{g_n(0|\boldsymbol{W})}$. H_n est dénommée la co-variable intelligente (*clever covariate*). Notons que ce modèle paramétrique, servant à mettre à jour l'estimateur initial $\bar{Q}_n^0(a,\boldsymbol{W})$, incorpore l'information sur le traitement, soit le score de propension. Cette mis-à-jour nous permet d'obtenir $\bar{Q}_n^1(a,\boldsymbol{W})$. En pratique, cette mise à jour peut être effectuée en ajustant un modèle de régression logistique de la variable réponse Y sur la co-variable intelligente H_n et le $\text{logit}\{\bar{Q}_n^0(A,\boldsymbol{W})\}$ en intercepte comme une constante dans le modèle. Nous obtenons ainsi ϵ_n^0 qui est un estimateur par maximum de vraisemblance. Ensuite, nous pouvons prédire $\bar{Q}_n^1(a,\boldsymbol{W})$ pour chaque groupe de traitement que nous notons $\bar{Q}_n^*(a,\boldsymbol{W})$. On peut répeter ce processus de mis à jour en ajustant une régression de Y sur H_n avec $\text{logit}\{\bar{Q}_n^{(1)}(A,\boldsymbol{W})\}$ en intercepte. Toutefois, l'estimateur $\epsilon_n^{(1)}$ sera égale à zéro. Ainsi, la convergence du TMLE se fait en une seule étape. Finalement, l'estimateur TMLE de l'effet moyen du traitement est obtenu en substituant la mise à jour $\bar{Q}_n^1(a,\boldsymbol{W})$ à la place de l'estimateur initial et on obtient:

$$\psi_{TMLE,n} = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n^*(1,\boldsymbol{w}_i) - \bar{Q}_n^*(0,\boldsymbol{w}_i)\}$$

Par construction, la fonction d'influence du TMLE décrit plus haut est définit par:

$$IC_n(O_i) = \Big(\frac{I(A=1)}{g_n(1|\boldsymbol{W}_i)} - \frac{I(A=0)}{g_n(0|\boldsymbol{W}_i)}\Big)(Y - \bar{Q}_n^{(1)}(A_i,\boldsymbol{W}_i)) + \bar{Q}_n^{(1)}(1,\boldsymbol{W}_i) - \bar{Q}_n^{(1)}(0,\boldsymbol{W}_i) - \psi_{TMLE}$$

Notons que la fonction d'influence est évaluée pour chacune des n observations O_i. En se basant sur la propriété linéaire asymptotique du TMLE, nous pouvons estimer la variance de l'estimateur par la variance de sa fonction d'influence divisé par la taille de l'échantillon. Par conséquent, nous pouvons construire un intervalle de confiance de type wald sur l'ATE par $\psi_{TMLE} \pm 1.96\sigma_n$, avec $\sigma_n^2 = 1/n^2 \sum_{i=1}^{n} IC_n(O_i)^2$ et faire des tests d'hypothèses.

2.3.4. Inférence en présence de positivité pratique

Les estimateurs de l'effet moyen du traitement développés plus haut reposent sur plusieurs hypothèses. En particulier, la positivité, qui stipule que la probabilité qu'un patient reçoive l'un ou l'autre des traitements conditionnellement aux co-variables doit être plus grande que zéro. Même si cette hypothèse peut être vraie en théorie, elle est souvent violée en pratique. Cette violation de la positivité pratique arrive souvent lorsqu'on est en présence de taille faible d'unités dans certains sous-groupes de la population. L'utilisation des méthodes flexibles pour estimer le score de propension [**71**] peut exacerber le problème, car il entraine l'ajustement des variables fortement associées au traitement, en particulier les variables instrumentales qui sont des purs facteurs de risque du traitement. Une des conséquences de la violation de la positivité pratique est l'obtention de valeurs extrêmes du score de propension, c'est-à-dire valeurs proches de 0 ou 1 et par conséquent des poids $1/g_n(a|W_i)$. Dans ce scénario, l'estimateur peut être instable et la variance de l'ATE peut être grande. Plusieurs approches ont été proposées pour l'estimation d'un paramètre causal lors d'une violation de la positivité en pratique. Parmi ces méthodes, nous pouvons citer la troncature des poids [**107**]. La troncature permet de réduire la variance due au poids extrême et donc limite l'impact des patients associés à ces poids sur le paramètre estimé. Il n'est toutefois pas facile de choisir le seuil de troncature. Dans ce contexte, Bembom et al., [**80**] ont proposé une procédure adaptive de sélection du seuil de troncature qui permet de réduire l'erreur quadratique moyenne. Une autre approche utilisée en présence de violation de la positivité est la rognure [**10**] qui a pour but de supprimer les patients associés aux poids extrêmes de l'analyse. Une telle action changerait inévitablement la population à l'étude. Cette méthode peut aider à réduire la variance mais pourrait augmenter le biais. Les poids stabilisés [**19**] peuvent servir également dans ce contexte. Dans le contexte d'estimateur doublement robuste et pour éviter l'impact de la non-positivité pratique sur les estimations, des procédures collaboratives pour l'estimatiom du score de propension furent developpées et seront présentées dans la section suivante.

2.3.4.1. Estimateur TMLE collaboratif

En pratique, il est souvent difficile pour un analyste de détecter la présence ou non de positivité. Dans le but de combattre l'instabilité induite par la positivités, van der Laan et Gruber, [**75**] ont proposé une version collaborative du TMLE, dénommée C-TMLE (Collaborative Targeted Maximum Likelihood Estimation). Dans l'approche C-TMLE, le score de propension et le modèle de l'issue sont estimés de manière collaborative. En effet, une procédure de sélection de variables pas à pas est utilisée lors de l'estimation du score de propension. Soit $g_{0,n}$ le score de propension obtenu en ajustant une régression du traitement A sur un intercepte. Ce score de propension nous permet de mettre à jour l'estimateur initial $\bar{Q}_n^0$ comme défini dans le TMLE ci-dessus pour obtenir $\bar{Q}_n^1$. Dénotons par K le nombre de co-variables et par $g_{k,n}; k = 1,....K$ le score de propension qui est fonction d'une co-variable $W^{(k)}$. Ces scores de propension sont utilisés pour mettre à jour l'estimateur initial $\bar{Q}_n^0$ pour obtenir K mis à jour $\bar{Q}_{1,n}^2,...,\bar{Q}_{K,n}^2$. Ensuite, on sélectionne $\bar{Q}_{j,n}^2$ (notons $\bar{Q}_n^2$) qui minimise la fonction de perte comparé à celle obtenue la première mise à jour $\bar{Q}_n^1$. Ainsi nous avons $\bar{Q}_n^1$ et $\bar{Q}_n^2$ associés aux scores de propension $g_{0,n}$ et $g_{j,n}$. La procédure continue en intégrant une autre variable parmi les $K-1$ variables restantes jusqu'à ce que toutes les variables soient introduites dans le score de propension. On obtient finalement $\bar{Q}_n^1, \bar{Q}_n^2,...,\bar{Q}_n^K$. Finalement, l'estimateur C-TMLE $\bar{Q}_n^*$ est celle qui minimise la fonction de perte définie dans [**76**] en utilisant une validation croisée. L'algorithme développé par van der Laan et Gruber [**75**] est présenté dans l"article 1. Toutefois, cette méthode présente des inconvénients. Nous pouvons citer par exemple, le temps énorme nécessaire pour l'estimation dû à l'estimation répétée du score de propension et la complexité de faire des inférences [**76**].

Très récemment, Benkeser et al., [**16**] ont proposé une autre version collaborative du TMLE plus efficace que la première et qui est une solution aux inconvénients énumérés ci-dessus. En effet, la méthode proposée par Benkeser et ces collègues assume que le modèle de l'issue est bien estimé avec une vitesse de convergence plus grande que $n^{-1/4}$ selon la norme $L^2(P)$. En se basant sur cette hypothèse, ils proposent un score de propension alternatif qui s'adapte au modèle de l'issue directement. La première version du C-TMLE construit un modèle du

score de propension de manière séquentielle en minimisant l'erreur dans l'estimateur initial. Cette nouvelle version n'est pas séquentielle, mais elle cible un score alternatif et non le vrai score de propension. Ce score alternatif est la probabilité de recevoir le traitement conditionnellement au modèle de l'issue estimée $P(A=a|\bar{Q}_n^1(a,\boldsymbol{W}))$. Ce score entraine une robustesse de la méthode face au problème de positivité pratique. En raison de la propriété de double robustesse de la fonction d'influence, le C-TMLE proposé est consistant pour l'ATE. Un point important qui pourrait causer problème dans cette méthode est le fait que le score alternatif ne sera jamais consistant pour le score de propension vu que la cible est différente. Par conséquent, la linéarité asymptotique de cette méthode en prendra le coup, car le second terme R_n ne sera pas asymptotiquement négligeable. Benkeser et al., [**16**] ont démontré qu'en ciblant la probabilité de recevoir le traitement conditionnellement au modèle de l'issue estimée $P(A=a|\bar{Q}_n^1(a,\boldsymbol{W}))$, R_n sera asymptotiquement négligeable sous certaines conditions raisonnables. La procédure C-TMLE (Benkeser et al., [**16**]) pour $E_0(Y^a)$ est la suivante:

Algorithm 1 Collaborative-TMLE pour $E_0(Y^a)$

1: Construire un estimateur initial "courrant" du modèle l'issue $\bar{Q}_n^1(a,\boldsymbol{W}) = E_n(Y|A=a,\boldsymbol{W})$ et prédire l'issue $\bar{Q}_n^1(a,\boldsymbol{W}_i)$ pour chaque unité $i=1,...,n$
2: Estimer le score alternatif $g_n(a|\bar{Q}_n^1(a,\boldsymbol{W})) = \widehat{P}(A=a|\bar{Q}_n^1(a,\boldsymbol{W}))$ en ajustant une régression de A sur la prédiction $\bar{Q}_n^1(a,\boldsymbol{W}_i)$ et prédire le score alternatif $g_n(a|\bar{Q}_n^1(a,\boldsymbol{W}_i))$ pour chaque unité $i=1,...,n$
3: Modèle de mise à jour: Ajuster un modèle de régression logistique de l'issue Y sur la covariable intelligente $H_n(a,\boldsymbol{W}) = I(A=a)/g_n(a|\bar{Q}_n^1(a,\boldsymbol{W}))$ avec offset $\text{logit}(\bar{Q}_n^1(a,\boldsymbol{W}))$. Notons ϵ_n le coefficient estimé du modèle.
4: Calculer $\bar{Q}_n^*(a,\boldsymbol{W}_i) = \text{expit}\{\text{logit}(\bar{Q}_n^1(a,\boldsymbol{W}_i)) + \epsilon_n H_n(a,\boldsymbol{W}_i)\}$
5: L'estimé final est $\frac{1}{n}\sum_{i=1}^n \bar{Q}_n^*(a,\boldsymbol{W}_i)$.

Toutes les méthodes présentées ci-dessus ont pour objectif de minimiser l'impact de la violation de la positivité en pratique sur les estimations. Toutefois, ces méthodes ne garantissent pas une estimation sans biais. Par conséquent, il est important de quantifier cet impact. Dans la littérature, très peu d'outils existent pour quantifier l'impact de la violation de la positivité en pratique sur l'estimation d'un paramètre causal. Dans la section suivante, nous discuterons de la procédure proposée par Petersen et al., [**78**].

2.3.4.2. Outils de diagnostic paramétrique

Petersen et al., [**78**] ont proposé un algorithme basé sur le rééchantillonnage [**9**] dans l'objectif de mesurer l'impact de la non-positivité en pratique sur l'estimation. La procédure proposée est la suivante:

- Construire un modèle sur l'issue $E(Y|A,\boldsymbol{W})$ noté Q_n d'une part et sur le traitement $E(A|\boldsymbol{W})$ noté g_n d'autre part. Estimer la valeur du paramètre d'intérêt en utilisant une méthode d'estimation comme par exemple le TMLE. Dénotons $\psi(\widehat{P}_0)$ cette estimation.
- Rééchantillonner B fois n unités aléatoire avec remise et supprimer le traitement et l'issue. Ensuite générer Y et A en utilisant (Q_n,g_n).
- Estimer le paramètre causal en utilisant les données simulées et noté $(\psi(\widehat{P}_n^1),...,\psi(\widehat{P}_n^B))$.
- Estimer l'impact de la non-positivité en prenant la différence entre la moyenne des estimations obtenues avec les B données simulées et la valeur $\psi(\widehat{P}_0)$.

En effet, le rééchantillonnage et la simulation effectués au deuxième point ont pour but de générer plusieurs données ayant la même structure que l'originale. En présence de violation de la positivité, on s'attendrait à une grande différence entre la moyenne des estimés bootstrap et la valeur initiale du paramètre estimé. L'approche proposée par Petersen et al., [**78**] est bonne, mais elle ne permet de préserver les associations entre les co-variables et le score de propension. Dans l'article 1 de cette thèse, nous proposons une adaptation de cette procédure qui préserve à la fois, les relations entre les co-variables et les associations entre les co-variables et le score de propension.

2.3.5. Inférence en présence de bias de sélection

Le biais de sélection est un défi important à relever en présence de données non probabilistes. Meng (2018, [**120**]) a développé une formule explicite sans aucune hypothèse sous-jacente pour expliquer l'impact du biais de sélection sur l'erreur d'estimation. Ces résultats démontrent à suffisance qu'un échantillon non probabiliste de grande taille ne permet

pas forcément de réduire l'erreur d'estimation. Un exemple palpable de ceci est la prédiction incorrecte de Alf Landon comme vainqueur de la présidentielle Américaine de 1936 contre Franklin Roosevelt faite par la revue Literary Digest avec un échantillon non probabiliste de 2,3 millions d'Américains [**64**]. La présence d'un biais de sélection vient fondamentalement du fait que les individus se trouvant dans le fichier administratif n'ont pas été sélectionnés aléatoirement. Ce biais peut être également la conséquence de la non-réponse lors d'une enquête échantillon ou de la perte au suivi. Dans la littérature, plusieus approches existent pour faire des inférences dans ce contexte. Soit U la population d'intérêt et $\mathcal{B}$ l'échantillon non-probabiliste ou le fichier administratif qui contient nos variables d'intérêts ($\boldsymbol{W}$,A,Y). En général, la taille de $\mathcal{B}$ est plus petite que celle de U. Soit $\sum_{i\in U} Y_i$ le paramètre d'intérêt correspondant au total de la variable Y dans la population. Une première approche serait de prendre un échantillon probabiliste $\mathcal{A}$ dans la partie non-couverte par les données administratives $U \setminus \mathcal{B}$ et recueillir les valeurs de la variable réponse. Ainsi, le paramètre d'intérêt pourrait être estimé avec une somme pondérée par l'inverse de la probabilité d'inclusion dans l'échantillon combiné $\mathcal{A} \cup \mathcal{B}$. Une seconde approche consiste à modéliser la relation entre la variable d'intérêt Y et les variables auxiliaires W et ensuite à prédire la variable d'intérêt que nous notons Y^* pour chacune des unités hors de l'échantillon non probabiliste $U \setminus \mathcal{B}$. Le total peut être obtenu par $\sum_{i\in\mathcal{B}} Y_i + \sum_{i\in U\setminus\mathcal{B}} Y_i^*$. Dans ce contexte, on suppose que la variable Y et l'indicateur d'inclusion à l'échantillon probabiliste sont indépendants conditionnellement à W. Elle est similaire à l'hypothèse d'ignorabilité en inférence causale. L'appariement statistique pourrait être utilisé dans ce contexte. Supposons qu'on est en mesure de tirer un échantillon probabiliste $\mathcal{A}$ de U. L'appariement statistique consiste à modéliser la relation entre Y et les co-variables $\boldsymbol{W}$, communes aux deux sources, en utilisant les données de l'échantillon non probabiliste. On suppose également l'hypothèse d'ignorabilité. On prédit la variable Y dans l'échantillon probabiliste $\mathcal{A}$ et le total est obtenu par une somme pondérée dans cette dernière. Finalement, Chen, Li & Wu [**121**] ont proposé une approche qui consiste à estimer la probabilité d'inclusion dans l'échantillon probabiliste en utilisant les 2 sources $\mathcal{A}$ et $\mathcal{B}$ ainsi que leurs co-variables communes. Une fois estimé, le

total est obtenu par une somme pondérée dans la source $\mathcal{B}$. L'approche proposé par Chen, Li & Wu [**121**] suppose une forme logistique pour estimer la probabilité d'inclusion dans $\mathcal{B}$ et assume une connaissance des variables à inclure dans ce modèle de régression. Toutefois, en pratique, il est impossible de les connaître et en présence de bases de données de grande dimensions, il est essentiel d'en faire une sélection. Dans ce contexte, Yang, Kim & Song [**101**] ont proposé un algorithme de sélection basé sur le *smoothly clipped absolute deviation* (SCAD; Fan & Li, [**38**]). Dans l'article 3 de cette thèse, nous généralisons la proposition de Chen, Li & Wu [**121**] en introduisant un algorithme de sélection de variables basé sur une extension du LASSO.

2.4. Méthodes d'apprentissage automatique

Les méthodes d'estimation d'un paramètre causal discutées plus haut dépendent de l'estimation des paramètres de nuisance, soit le score de propension et/ou l'espérance conditionnelle de l'issue. Une bonne spécification des ces dernières est primordiale pour une estimation convergente du paramètre causal. Pour atteindre cet objectif, les méthodes d'apprentissage automatique sont conseillées [**76**]. En effet, elles permettent d'éviter les hypothèses sous-jacentes des modèles paramétriques et donc une mauvaise spécification du modèle. Toutefois, elles doivent être utilisées à bon escient car une estimation optimale des paramètres de nuisance n'entraine pas forcément une estimation optimale d'un paramètre causal. En effet, ces méthodes minimisent l'erreur de prédiction et ainsi permettent de prédire au mieux le score de propension et l'espérance conditionnelle de l'issue. Par contre, en inférence causale, l'objectif est de servir des deux paramètre de nuisance pour ajuster les variables nécessaires dans le but d'une estimation convergente et efficiente. Considérons par exemple la méthode du calcul-g qui est basée sur la modélisation de l'espérence de l'issue $\bar{Q}_0(a,W) = E_0(Y|A=a,\boldsymbol{W})$. Une méthode d'apprentissage automatique pourrait par exemple penser que le traitement A n'est pas important pour prédire Y. Par conséquent, on conclurait à tort que A n'a aucun effet sur l'issue Y. D'un autre côté, l'utilisation des méthodes flexibles pour l'esimation du score de propension peut entrainer une sélection de variables instrumentales et par conséquent

diminuer la précision de notre estimateur [**71**]. Il s'ensuit donc que l'objectif de l'inférence causale est différente de celle des méthodes flexibles. Dans les sections suivantes, nous discuterons d'une procédure régularisation LASSO introduite par Tibshirani (1996, [**88**]) et une méthode d'ensemble Super Learner (SL) [**74**].

2.4.1. Super Learner

Les estimateurs doublement robustes nécessitent l'estimation des paramètres de nuisance à savoir le score de propension et le modèle de l'issue. Une régression paramétrique simple comme la régression logistique ou linéaire pourrait être utilisée dans ce contexte. Toutefois, dans le but d'augmenter les chances d'obtenir un modèle correctement ajusté, il est fortement conseillé d'utiliser les méthodes d'apprentissage automatique [**76, 92**]. Dans l'article 1 de cette thèse, nous travaillons avec le Super Learner (SL) [**74**]. Soit Y la variable réponse et $\boldsymbol{W}$ un vecteur de co-variables. Soit $\eta_0(\boldsymbol{W})$ le paramètre d"intérêt obtenu en minimisant une fonction de perte $L(O;\eta)$. on peut écrire $\eta_0 = arg\min_{\eta \in H} E_0 L(O;\eta)$, avec E_0 l"espérance par rapport à Y. Nous pouvons citer par exemple la régression linéaire avec $\eta_0(\boldsymbol{W}) = E_0(Y|\boldsymbol{W})$ et $L(O;\eta) = (Y - \eta(\boldsymbol{W}))^2$. Le SL est une méthode d'ensemble (*ensemble learner*) qui combine des prédictions de divers modèles prédéfinis par l'utilisateur. Ces candidats peuvent inclure des méthodes paramétriques, semi-paramétriques et d'apprentissage automatique. Soit ($\widehat{Y}^{(1)}$, $\widehat{Y}^{(2)}$,..., $\widehat{Y}^{(k)}$) un vecteur de prédictions obtenu par k méthodes en utilisant la validation croisée. En effet, chaque algorithme k est ajusté sur 9/10 de l'ensemble des données. Le modèle obtenu permet de prédire la variable réponse pour les 1/10 restantes. Ainsi, à la fin, chaque algorithme aura produit un vecteur au complet de prédictions via une validation croisée. L'algorithme choisit la meilleure pondération pour combiner ces différentes prédictions. On définit:

$$\widehat{E}(Y|\widehat{Y}^{(r)}) = \alpha_1 \widehat{Y}^{(1)} + \alpha_2 \widehat{Y}^{(2)} + ... + \alpha_k \widehat{Y}^{(k)}.$$

En effet, en utilisant la prédiction obtenue par chaque méthode, $\alpha_k \geq 0$ est déterminée en minimisant l'erreur de validation croisée. Plus spécifiquement, α_k est l'estimation du

coefficient de régression entre Y et le vecteur ($\widehat{Y}^{(1)}$, $\widehat{Y}^{(2)}$,..., $\widehat{Y}^{(k)}$) avec une contrainte $\alpha_k \geq 0$. Asymptotiquement et à erreur de contrôle près, il est démontré que le Super Learner conduit à des résultats au moins aussi bien en terme de prédiction que la meilleure méthode dans la liste prédéfinie par l'utilisateur [**74**]. Ainsi, dans le but de rendre les estimateurs doublement robustes asymptotiquement linéaire, il suffit d'insérer des prédicteurs qui ont au moins une vitesse de convergce de $n^{-1/4}$ dans la liste de modèles prédéfinis.

2.4.2. LASSO

2.4.2.1. Définitions et notations

Le LASSO est l'acronyme de "Least Absolute Shrinkage and Selection Operator". En statistique, les méthodes de régularisation permettent d'éviter la surestimation (*over-fitting*) pour des fins de prédiction. Soit $\{(\boldsymbol{W}_i,Y_i), i = 1,...,n\}$ le vecteur de prédicteurs et la variable réponse respectivement avec $\boldsymbol{W} = (1, W^{(1)},...,W^{(p)})$ le vecteur de co-variables et une constante. La régression ridge [**108**] est la première méthode de régularisation. Cette méthode pose une contrainte sur les coefficients estimés de telle sorte que la norme L_2 des coefficients ne dépasse pas une constante choisie. Ainsi, les coefficients estimés sont rétrécis vers zéro mais ne donne pas des valeurs égales à zéro. Contrairement au ridge, le LASSO utilise une autre forme de contrainte (la norme L_1) sur les coefficients estimés, ce qui entraine que certains coefficients estimés soient nuls. Soit β_j le coefficient associé à la variable $W^{(j)}$. Plus formellement, le LASSO est une régression linéaire dont les coefficients $\boldsymbol{\beta} = (\beta_1,...,\beta_p)$ sont sujets à la contrainte:

$$\sum_{j=1}^{p} |\beta_j| \leq s, avec\ s \geq 0.$$

Le coefficient associé à l'intercepte β_0 n'est pas inclus dans la contrainte. Mathématiquement, la régression LASSO se propose de minimiser la somme des carrés des résidus sous la contrainte ci-dessus. On peut écrire:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \underbrace{||Y - \boldsymbol{W}\beta||^2}_{(1)} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{(2)}.$$

Le terme (1) est la fonction de perte (la somme des carrés des résidus) et le terme (2) est la pénalité qui est la norme L_1. Pour $\lambda = 0$, le LASSO devient exactement la régression linéaire standard. La valeur optimale de λ est en pratique choisie par validation croisée. Pour une valeur grande de λ, certains coefficients β_j, $j = 1,...,p$ sont égaux à zéro. C'est la raison pour laquelle, le LASSO est considéré comme une méthode de sélection de variables. Dans le cas d'une régression logistique LASSO, on cherche $\boldsymbol{\beta}$ qui minimise:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{-Y_i \boldsymbol{W}_i' \boldsymbol{\beta} + \log(1 + exp(-\boldsymbol{W}_i' \boldsymbol{\beta}))\} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

2.4.2.2. Propriété oracle et le LASSO adaptatif

Supposons que les paramètres d'un modèle de régression contiennent à la fois des coefficients nuls et non nuls. Soit $\boldsymbol{\beta}^{(1)}$ le vecteur de coefficients non nuls de dimension p_1 et $\boldsymbol{\beta}^{(2)}$ le vecteur de coefficients nuls de dimension p_2. On peut écrire $\boldsymbol{\beta}$ comme suit:

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(1)}_{(p_1 \times 1)} \\ \boldsymbol{\beta}^{(2)}_{(p_2 \times 1)} \end{pmatrix}.$$

Un modèle de sélection de variables à une propriété oracle si les deux conditions suivantes sont respectées:

- La probabilité d'estimer une valeur nulle pour les coefficients nuls tend vers 1: $Pr(\widehat{\boldsymbol{\beta}}^{(2)} = 0) \to 1$
- Les estimations sont convergentes avec une vitesse de convergence de $\sqrt{n}$ et sont asymptotiquement normales: $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}) \to N(0,C)$

avec C la matrice de covariance de $\boldsymbol{\beta}^{(1)}$. Par conséquent, un modèle ayant cette propriété est capable de sélectionner le vrai sous-ensemble de prédicteurs et de produire de bonnes estimations pour ces derniers pour n suffisamment large. La procédure LASSO ne possède pas la propriété oracle [**35, 81**] car elle n'est pas consistante en termes de sélection de variables. Zhao et Yu [**81**] ont montré qu'il existe une condition à respecter pour le LASSO afin d'avoir la propriété oracle: la condition d'irreprésentativité (*the irrepresentable condition*). La procédure de LASSO est consistante en termes de sélection de variables si les prédicteurs

non pertinents ne sont pas corrélés aux variables contenues dans le vrai modèle.
Plusieurs extensions du LASSO furent développées à travers le temps. En particulier, le LASSO adaptatif [**35**] qui satisfait la propriété oracle quand la fonction de lien est correctement spécifiée ou pas [**115**]. La méthode du LASSO adaptatif est une extension du LASSO qui propose de résoudre le problème de minimisation suivant:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} ||Y - \boldsymbol{W}\beta||^2 + \lambda \sum_{j=1}^{p} \widehat{\omega}_j |\beta_j|,$$

avec $\widehat{\omega}_j = \frac{1}{|\tilde{\beta}_j|^\gamma}$, $\gamma > 0$ et un estimateur $\tilde{\beta}_j$ $\sqrt{n}$-consistent de β_j. Par exemple, on pourrait ajuster un modèle linéaire de Y sur $\boldsymbol{W}$ pour estimer $\tilde{\beta}_j$. Quand n est grand, les poids correspondants aux variables insignifiantes tendent vers l'infini tandis que les poids correspondants aux variables significatives tendent vers une valeur finie. Il s'en suit donc que les coefficients associés aux variables insignifiantes sont rétrécis à zéro et donc ces variables ne sont pas sélectionnées à la fin du modèle. Il est montré [**35**], que la méthode du LASSO adaptatif possède la propriété oracle.

2.4.2.3. LASSO adaptatif basé sur l'issue

En inférence causale, il est important d'ajuster dans le modèle du score de propension les bonnes variables. Dans le but de produire des estimations sans biais, l'hypothèse d'ignorabilité nous suggère de contrôler au minimum les variables qui causent directement l'issue et/ou qui influencent directement la sélection du traitement (VanderWeele & Shpitser, [**110**]). Toutefois, il est montré empiriquement qu'il y a un bénéfice à inclure les purs facteurs de risque de l'issue (Brookhart et al. [**2**]; Shortreed & Ertefaie, [**104**]) et un risque à inclure les variables instrumentales (Schisterman et al.,[**29**]; Schneeweiss et al., [**98**]; van der Laan & Gruber, [**93**]). Plusieurs méthodes furent proposées pour sélectionner les variables appropriées en inférence causale. En particulier, Shortreed & Ertefaie [**104**] ont proposé en 2017, une procédure automatique (OALASSO: Outcome Adaptive LASSO) de sélection des variables confondantes et des purs facteurs de risque de l'issue. L'approche de Shortreed & Ertefaie utilise le LASSO adaptatif [**35**] pour estimer le score de propension. En effet, Shortreed & Ertefaie introduisent un poids $\widehat{\omega}_j$ équivalent à l'inverse du coefficent estimé de

régression entre l'issue Y et le vecteur de co-variables $\boldsymbol{W}$, conditionnellement à A. Soit α_j les coefficients associés à $\boldsymbol{W}$ dans la régression entre Y et $\boldsymbol{W}$, conditionnellement à A. Plus formellement la procédure de Shortreed & Ertefaie, [**104**] résoud le problème de minimisation suivant:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} ||Y - \boldsymbol{W}\beta||^2 + \lambda \sum_{j=1}^{p} \widehat{\omega}_j |\beta_j|,$$

avec $\widehat{\omega}_j = \frac{1}{|\tilde{\alpha}_j|^\gamma}$, $\gamma > 0$ et $\tilde{\alpha}_j$, un estimateur $\sqrt{n}$-consistent de α_j. Shortreed & Ertefaie, [**104**] ont démontré que leur procédure force les coefficients associés aux variables instrumentales à être zéro. De plus, les estimations des coefficients associés aux variables confondantes et aux purs facteurs de risque de l'issue tendent asymptotiquement vers les valeurs qu'ont aurait obtenues si le vrai modèle du score était connu.

2.4.2.4. LASSO hautement adaptatif (Highly Adaptive LASSO; HAL)

Comme discuté plus haut, pour être linéairement asymptotique, les estimateurs doublement robustes nécessitent un estimateur des paramètres de nuisance qui converge avec une vitesse d'au moins $n^{-1/4}$. Toutefois, les méthodes flexibles (non-paramétriques) existantes ont des vitesses de convergence faible. L'estimateur HAL a été développé dans le but de garantir cette vitesse de convergence sans pour autant supposer une forme paramétrique. Des études récentes ont montré également ces avantages pour l'estimation de l'ATE lorsqu'il est basé sur l'issue pour l'estimation du score de propension [**11**].

Le LASSO (LASSO adaptatif) assume une forme paramétrique (linéaire) de la fonction de régression $E(Y|\boldsymbol{W})$. En général, les estimateurs paramétriques assument une forme pour la fonction. Cette supposition peut entrainer un biais lors de la prédiction si la vraie fonction $E(Y|\boldsymbol{W})$ ne suit pas en réalité la forme supposée. Par contre, les estimateurs non paramétriques comme les méthodes d'apprentissage automatiques ne supposent aucune forme. Ils sont dits flexibles. Le LASSO hautement adaptatif est une extension non paramétrique du LASSO adaptatif et est proposé par Benkeser et van der Laan [**15**]. Soit $E(Y|\boldsymbol{W})$, la fonction de régression, avec Y la variable réponse et $\boldsymbol{W}$ le vecteur de co-variables. Considérons une décomposition de $\boldsymbol{W}$ en un ensemble d'indicateurs binaires. Par exemple, si W est

un scalaire, pour une observation w, on génère le vecteur $\boldsymbol{\phi}^*(w) = (\phi_1^*(w),...,\phi_n^*(w))^T$, avec $\phi_i^*(w) = I(w \geq W_i)$, lors $i = 1,...,n$. Si W est un vecteur $\boldsymbol{W} = (W^{(1)},W^{(2)})^T$, nous devons inclure les termes de second ordre $\boldsymbol{\phi}_i^*(\boldsymbol{w}) = I(w_1 \geq W_i^{(1)},w_2 \geq W_i^{(2)})$, pour $i = 1,...,n$. Le LASSO hautement adaptatif HAL [**15**] est un LASSO adaptatif de la variable réponse Y sur les nouvelles variables binaires obtenues après décomposition. Il a été démontré que l'estimateur HAL de la fonction $E(Y|\boldsymbol{W})$ converge vers la vraie fonction sous la norme L_2 avec une vitesse de convergence pas plus lente que $n^{-1/4}$ peu importe la dimension de $\boldsymbol{W}$, et sous la condition selon laquelle la vraie fonction a une norme de variation finie [**15**]. Dans ce mémoire, nous avons utilisé le package *hal9001* [**56**] disponible sur le logiciel R.

2.4.2.5. Inférence sélective

En présence de données de grande dimensions et dans un scénario où le vrai modèle ne contient pas toutes les co-variables, on fait généralement usage des méthodes de sélection de variables. Toutefois, une fois les variables sélectionnées, faire des inférences en ignorant cette étape de sélection donnerait des intervalles de confiance incorrects. L'inférence sélective est une méthode qui permet de construire un intervalle de confiance sur les coefficients estimés après sélection de variables. Soit $\widehat{\boldsymbol{\beta}}'$ un estimateur obtenu par la procédure LASSO. Supposons que $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$ sont les coefficients non nuls du vecteur $\widehat{\boldsymbol{\beta}}'$ avec $\widehat{M} \subseteq \{1,...,p\}$, leurs positions respectives. Dans cette section, on s'intéresse à faire des inférences pour les coefficients non nuls $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$. En réalité, $\widehat{\boldsymbol{\beta}}'$ dépend du modèle sélectionné $\widehat{M}$. Par conséquent, il serait inccorect d'ignorer cette dépendance. Lee et al., [**42**] ont montré que la distribution conditionnelle de $\widehat{\boldsymbol{\beta}}'_M|\{\widehat{M} = M\}$ suit une normale tronquée. Basé sur une statistique pivotale, il est donc possible de construire un intervalle de confiance en inversant un test d'hypothèse. En effet, conditionnellement au fait que le vrai sous-ensemble de variables soit sélectionné, cet intervalle de confiance rapporte un intervalle dans lequel se trouve avec un niveau de confiance $1-\alpha$, la vrai valeur du paramètre. Dans les situations dans lesquelles on intègre des procédures de sélection de variables, en particulier lors de la sélection de modificateurs d'effet, il est important de faire de l'inférence sélective si l'on désire faire des inférences sur les coefficients estimés vu qu'il permet de conditionner sur le modèle sélectionné $\widehat{M}$.

Dans ce mémoire, nous avons utilisé le package *selectiveInference* [**89**] disponible sur le logiciel R.

Chapitre 3

Objectifs

Durant cette thèse, trois articles ont été rédigés. L'objectif général de ces projets est de développer des méthodes d'estimation doublement robustes pour des paramètres marginaux. En particulier, développer des outils de diagnostic et d'estimation pouvant être facilement utilisées et comprises en pharmacoépidémiologie.

Dans le premier article, nous développons un outil de diagnostic pour mesurer l'impact de l'estimation du score de propension en absence de positivité en pratique sur les estimations. Cette procédure identifie quand un estimateur diverge à cause du score de propension estimée par des méthodes d'apprentissage automatique. Cette procédure est une extension de celle proposée par Peterson et al., [**78**].

Dans le second article, nous proposons une procédure à deux étapes, doublement robuste, permettant la sélection de modificateur d'effet et l'estimation d'un effet conditionnel en utilisant les méthodes de régularisation.

Le troisième article développe une procédure de sélection de variables lorsqu'on combine une base de données administrative (échantillon non probabiliste) sujette à un biais de sélection et une enquête probabiliste contenant des co-variables communes dans le but d'avoir des estimations sans biais. Cette méthode est une extension de celle développée par Chen et al., [**121**].

Chapitre 4

Understanding and Diagnosing the Potential for Bias when using Machine Learning Methods with Doubly Robust Causal Estimators

Cet article a été publié dans le journal Statistical Journal for Medical Research.

Préambule: L'étude décrite dans cet article est inspirée par une divergence obtenue lors de l'estimation de l'effet moyen du traitement en utilisant les méthodes paramétriques et flexibles pour l'estimation du score de propension. Plus spécifiquement, en présence de violation de la positivité en pratique, l'estimation de l'ATE peut être déstabilisée par l'estimation du score de propension avec une méthode d'apprentissage automatique. Ainsi, l'objectif principal est de développer un outil de diagnostic pratique qui permet aux analystes de détecter l'influence de l'estimation du score de propension sur l'estimation d'un paramètre d'intérêt. Une attention particulière est portée à l'estimateur par maximum de vraisemblance ciblée. Toutefois, cet outil peut s'appliquer à d'autres estimateurs doublement robustes.

Understanding and Diagnosing the Potential for Bias when using Machine Learning Methods with Doubly Robust Causal Estimators

Asma Bahamyirou, Lucie Blais, Amélie Forget, Mireille E. Schnitzer.

Faculté de pharmacie, Université de Montréal.

Résumé: Dans le but d'estimer des effets causaux marginaux, les méthodes flexibles aux données ont été proposées pour estimer les paramètres de nuisance. Toutefois, en présence de positivité en pratique, ces méthodes peuvent entrainer un manque de chevauchement entre le groupe exposé et non exposé, en termes de densité du score de propension. Ceci peut résulter à une estimation biaisée de l'effet moyen du traitement. Pour motiver notre problème, nous avons évalué l'estimateur du maximum de vraisemblance ciblé (TMLE) grâce à une étude de simulation, pour estimer l'effet moyen du traitement. Nous avons mis en lumière la divergence entre les estimations lorsqu'on utilise une méthode paramétrique et une méthode flexible pour l'estimation du score de propension. Ensuite, nous avons adapté un outil de diagnostic existant dans la littérature pour montrer que les estimations obtenues avec les méthodes flexibles peuvent être entachées de biais et ont une probabilité de recouvrement faible. La procédure bootstrap proposée est à même de détecter cette instabilité et peut être utilisée comme outils diagnostic.

Keys words: Inférence causale, positivité, doublement robuste, IPTW, TMLE, Super Learner.

Abstract: Data-adaptive methods have been proposed to estimate nuisance parameters when using doubly robust semiparametric methods for estimating marginal causal effects. However, in the presence of near practical positivity violations, these methods can produce a separation of the two exposure groups in terms of propensity score densities which can lead to biased estimates of the treatment effect. To motivate the problem, we evaluated the Targeted Minimum Loss-based Estimation procedure using a simulation scenario to estimate the average treatment effect. We highlight the divergence in estimates obtained when using parametric and data-adaptive methods to estimate the propensity score. We then adapted

an existing diagnostic tool based on a bootstrap resampling of the subjects and simulation of the outcome data in order to show that the estimation using data-adaptive methods for the propensity score in this study may lead to large bias and poor coverage. The adapted bootstrap procedure is able to identify this instability and can be used as a diagnostic tool

Keys words: Causal inference, positivity, doubly robust, IPTW, TMLE, Super Learner.

4.1. Introduction

Positivity, or the experimental treatment assumption, is one of the requirements for causal inference, along with conditional exchangeability (i.e., no unmeasured confounders), no interference, and well-defined interventions [**69**]. Positivity requires that the probability of receiving any level of the treatment conditional on the covariates must be positive for each individual in the population. Near practical positivity violations occur when some patients have an estimated probability of receiving some level of treatment close to zero. This can occur even when the theoretical positivity holds, for instance, due to insufficient observations in some covariate strata.

In order to estimate a treatment effect, propensity score methods [**83**], where the propensity score is defined as the conditional probability of receiving a given treatment, have been increasing in popularity. For example, marginal effects such as the average treatment effect (ATE) can be estimated by weighting outcomes by the inverse of the estimated propensity score (IPTW) [**52**]. For these methods, correct specification of the propensity score model is required for unbiased or consistent estimation.

Doubly robust semiparametric methods such as Targeted Minimum Loss-Based Estimation (TMLE,[**76**]), which is closely related to previously existing methods [**25, 32**] have been proposed to remove the dependence on the propensity score model specification. The term doubly robust comes from the fact that the method requires both the estimation of the propensity score and the outcome expectation conditional on treatment and covariates, while only one of which needs to be correctly specified to have consistent estimation. Therefore, when the outcome model is consistent, a correct specification of the propensity score is

unnecessary and vice versa.

To increase the chance of correct specification, Machine Learning (ML) methods [**74**] are often recommended [**76, 87**]. However, flexible modeling of the propensity score may result in the selection of strong predictors of the treatment which may or may not be true confounders [**71**], giving rise to extreme probabilities. TMLE involves the inverse of the propensity score and any near violations of practical positivity can cause unstable parameter estimates and potential bias due to highly variable weights. This can be aggravated, notably by using ML to predict the probability of receiving treatment level (in IPTW/TMLE) [**71**]. In order to resolve these issues, one may use truncation of the weights to reduce the standard error [**107, 78**], though selection of the level of truncation is usually ad-hoc.

In a large covariate space, Collaborative Targeted Minimum Loss-Based Estimation (C-TMLE) [**93**], an extension of TMLE which incorporates a variable selection strategy in the propensity score model, can further improve the mean squared error particularly in the presence of near positivity violations. However, current implementations of C-TMLE rely on parametric estimation of the propensity score.

The first objective of this paper is to show that under a partially misspecified outcome model, flexible modeling of the propensity score can increase bias when there is potential for practical positivity violations. In the second objective of this paper, we have adapted the parametric bootstrap diagnostic tool, proposed by Peterson et al. [**78**] to inform whether, in a given analysis, a doubly robust estimator was likely destabilized by the estimation of the propensity score. The final objective is to demonstrate the usage of the diagnostic tool in a real data exemple.

In Section 2, we use the potential outcomes framework to define the target causal parameter of interest and review standard implementations of IPTW, TMLE and C-TMLE for the estimation of the parameter of interest. In Section 3, we show in a simulation scenario that finite-sample bias can increase under a partially misspecified outcome model and when ML methods are used. In Section 4, we present the adapted version of the diagnostic tool [**78**] and apply this procedure to our simulated data. We present an analysis of the safety

of asthma medications during pregnancy in Section 5. We discuss the results obtained in Section 6.

4.2. Estimators

In this section, we will briefly present the algorithms of IPTW [**23**], TMLE and C-TMLE [**76, 93**].

4.2.1. Targeted estimation

In order to define the target parameter, we use the counterfactual framework of Rubin [**21**]. The observed data can be represented as $O = (W,A,Y)$, where W is the baseline covariates of a patient, A is the treatment which equals 1 if the patient received treatment and 0 otherwise, and Y is the observed continuous outcome. We use $O_i = (W_i,A_i,Y_i)$ to represent the i-th observation of the data. Let Y^a denote the potential outcome that would have occurred under the treatment value $A = a$. In this paper, we focus on the average treatment effect (ATE) which we denote ψ_0. If we assume that we observe $Y = Y^a$ when $A = a$ (consistency) [**106**], no interference [**69**], and no unmeasured confounders [**69**], the target parameter can be defined nonparametrically as:

$$\begin{aligned} \psi_0 &= E_0(Y^1) - E_0(Y^0) \\ &= E_{W,0}\{\underbrace{E_0(Y|A=1,W)}_{\bar{Q}_0(1,W)} - \underbrace{E_0(Y|A=0,W)}_{\bar{Q}_0(0,W)}\}. \end{aligned} \tag{4.2.1}$$

where E_0 is the expectation with respect to the outcome and $E_{W,0}$ is the expectation with respect to the baseline covariates.

4.2.2. Inverse Probability of Treatment Weighting (IPTW)

Horvitz and Thompson [**23**] proposed the idea of weighting observed values by inverse probabilities of selection in the context of sampling methods. The same idea is used in causal inference to estimate the ATE if we consider the counterfactual outcomes which we don't

observe to be missing. Weighting estimators provide ways to obtain large-sample unbiased estimates of the ATE using the propensity score. We denote $g_0(A|W) = P(A = 1|W)$ as the propensity score. Now, in order to estimate the average causal effect, the treated and untreated subjects are assigned the weights $w_i = 1/g_n(1|W_i)$ and $w_i = 1/(1 - g_n(1|W_i))$ respectively, where $g_n(1|W_i) = P_n(A_i = 1|W_i)$ is the estimated probability of treatment for subject i. By weighting subjects, a pseudo-population is created, where the distribution of covariates is comparable between the two treatment groups as in a randomized experiment [**107**]. The IPTW estimator is given by:

$$\psi_n^{IPTW} = \frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)w_i Y_i.$$

IPTW estimators can be unstable when the weights are large for some subjects due to a very low apparent probability of receiving the treatment received. Several methods exist to address this issue such as weight truncation [**107**], in which weights that exceed a specified threshold are each set to that threshold and trimming [**10**], in which subjects with very large weights are dropped from the analysis. These methods can help to reduce the variance but may increase bias in the estimation of the ATE. Data-adaptive methods have also been proposed to select a beneficial truncation level [**80**].

4.2.3. Targeted Minimum Loss-Based Estimation

Targeted Minimum Loss-based Estimation (TMLE) [**73**], is a general framework to produce semiparametric efficient and doubly robust plug-in estimators. TMLE [**76**] is efficient (i.e. minimal variance in large samples) when all models contain the truth. We denote $\bar{Q}_0(a,W) = E_0(Y|A = a,W)$ and let $\bar{Q}_n$ be an estimate of $\bar{Q}_0$. For the estimation of the ATE, TMLE is a one-step procedure where we first obtain an estimate of the outcome model $\bar{Q}_0$ and then use the treatment model g_0 to update the estimate. A TMLE procedure [**73**] for the estimation of $E_0(Y^a)$, where $a = 0$ or 1, is the following:

Algorithm 2 Targeted Minimum Loss-Based Estimation for $E_0(Y^a)$

1: Construct an initial estimate of the outcome expectation $\bar{Q}_n(a,W) = E_n(Y|A=a,W)$ for each subject.
2: Obtain the estimated propensity score $g_n(a|W) = P_n(A=a|W)$ for each subject.
3: Update the initial outcome estimates using the estimated propensity score to obtain $\bar{Q}_n^*(a,W)$ by following steps (a)-(d).
   (a) Define a covariate as $H(a,W) = I(A=a)/g_n(a|W)$.
   (b) Fit an intercept-free logistic regression of $Y \sim$ Offset $\{Logit(\bar{Q}_n(a,W))\} + H(a,W)$.
   (c) Obtain ϵ_n, the estimated coefficient of $H(a,W)$, which is referred as a fluctuation parameter.
   (d) Set $\bar{Q}_n^*(a,W_i) = expit\{logit(\bar{Q}_n(a,W_i)) + \epsilon_n/g_n(a|W_i)\}$
4: The final estimate is $\frac{1}{n}\sum_{i=1}^{n}\bar{Q}_n^*(a,W_i)$.

For a continuous and bounded outcome $Y \in [a,b]$ with $a < b$, Y must first be transformed into $Y^* \in [0,1]$ by shifting and scaling using constants [**94**]. The doubly robust nature of TMLE means that just one of the regression models (propensity or outcome) must be correctly specified to produce large-sample unbiased estimation. In large samples, the variance of TMLE, which is the variance of its influence function [**76**] divided by the sample size, is less than or equal to the variance of all semiparametric estimators, when the initial outcome and the propensity score models are both correctly specified (local efficiency). With an estimate of the standard error σ_n of TMLE, we can construct a Wald-type 95% confidence interval as $\psi_n \pm 1.96\sigma_n$. R[**95, 40**]. A SAS macro for TMLE for a binary point exposure has also been developed [**66**].

4.2.4. Super Learner

Doubly robust estimators involve the estimation of both the conditional outcome and propensity score models. Logistic regression can be used in both cases if we are in the context of a binary outcome and treatment, but to take advantage of the local efficiency of TMLE, we may prefer nonparametric estimators to increase the chances of correct model specification [**76**]. Ensemble learning methods such as Super Learner (SL) [**74**] are often recommended [**92**]. Super Learner combines predictions from a set of user-specified candidate models that may include parametric regression models, semiparametric regression models, and ML methods. The algorithm chooses the best weighted combination of these estimators

using cross-validation and performs generally at least as well as or better than the best candidate estimator in the library in terms of prediction [**74**]. Specifically, each method produces a cross-validated prediction and the optimal weight is determined by minimizing the cross-validated prediction error which is formulated as a regression of the outcome Y on the cross-validated predictions. R packages for implementing SL are available [**27**].

4.2.5. Collaborative Targeted Minimum Loss-Based Estimation

The double robustness property of TMLE guarantees large-sample unbiased estimation if at least one of the models (outcome or treatment) is estimated correctly. In addition, large-sample unbiased estimation occurs when the propensity score model conditions on a set of covariates that explains the residual bias of $\bar{Q}_n$ with respect to $\bar{Q}_0$ even if neither model is correctly specified [**75**]. When estimating the propensity score with data-adaptive methods, optimizing the treatment model fit would favor covariates that may be unrelated to the outcome and strongly predictive of the treatment [**71**] and updating the outcome regression based on this propensity score estimate can inflate estimation variance (or cause computational instability) and potentially bias the estimation [**41**]. C-TMLE [**75**], as an extension of TMLE, has been proposed to avoid such situations by collaboratively building the propensity score based on the outcome model fit. A forwards stepwise variable selection C-TMLE procedure for $E_0(Y^a)$ is the following. Firstly, one needs to define a loss function to evaluate the error in $\bar{Q}_n$. For example the logistic likelihood loss function $L(\bar{Q}_n) = -\sum Y\{\log(\bar{Q}_n) + (1-Y)\log(1-\bar{Q}_n)\}$ can be used for a binary or bounded continuous outcome.

Algorithm 3 Collaborative-TMLE for $E_0(Y^a)$

1: Construct the initial "current" estimate of the outcome model $\bar{Q}_n^c(a,W) = E_n(Y|A = a,W)$.
2: Use a forward selection algorithm to create a sequence of nested g models improving in fit: $g_{1,n},g_{2,n},...,g_{K,n}$ where K is the number of covariates.
   (a) Variables are added to g_n as long as they improve the value of the error of $\bar{Q}_{k,n}^*(a,W)$ (obtained by updating $\bar{Q}_n^c(a,W)$ w.r.t $g_{k,n}$). The variable that offers the greatest improvement is added at each step.
   (b) If no forward selection step improves the error, update the current $\bar{Q}_n^c$ with the current $g_{k,n}$ to obtain a new current $\bar{Q}_n^c$. Then repeat step (a).
3: This procedure creates estimators $\bar{Q}_{1,n}^*(a,W),...,\bar{Q}_{K,n}^*(a,W)$ that are strictly decreasing in error. Use V-fold cross-validation to select the final number of steps, k, that minimizes the error in $\bar{Q}_n^*$.
4: The final estimate is $\frac{1}{n}\sum_{i=1}^n \bar{Q}_{k,n}^*(a,W_i)$.

Because ψ_n can still be asymptotically unbiased if the propensity score model adjusts for the residual bias of $\bar{Q}_n(a,W)$ [**75**], the C-TMLE procedure attempts to select only the set of covariates needed using a forward selection algorithm to fit the propensity score model. This can greatly reduce the variance of the resulting estimator [**76**]. It should be noted that, in the presence of near positivity violations, C-TMLE will generally avoid full adjustment due to a perceived increase in the cross-validated error. This allows for extrapolation using the outcome model which may mask the true incomparability of the treatment groups. One weakness of the above implementation of C-TMLE is that it does not incorporate machine learning methods. However, one may also include non-linear combinations of covariates as additional candidates to be selected to improve the flexibility of the models.

4.3. Simulation Scenario

In this section, we present a simulation scenario to describe the problem and highlight a specific problematic scenario that may arise in practice.

4.3.1. Simulated data

In this simulation, we generated datasets with two confounders, two instruments (pure predictors of treatment), and one pure risk factor of the outcome. The two confounders

$W = (W_1,W_2)$ were generated as bivariate normal with $\mu = (0.5,1)$ and $\Sigma = \begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix}$ and were subsequently bounded between $[-3,4]$. The instruments were generated with normal distributions: $I_1 \sim N(1,2)$, $I_2 \sim N(1,1.9)$, the pure risk factor for the outcome as normal $P \sim N(1,1.5)$ and all were bounded between $[-3,3]$. The treatment mechanism g_0 was set as a Bernoulli with the probability generated nonlinearly in one confounder variable and one instrument.

$$P_0(A = 1|W) = Expit\{0.2 + W_1 + 0.3I_1 + W_1I_1 - 0.2(W_2 + I_2)^2\}$$

The observed outcome Y was Gaussian with a mean generated nonlinearly on the confounders and one pure risk factor.

$$Y = 1 + A + W_1 + 2W_2 + 0.5(W_1 + P)^2 + N(0,1)$$

The true ATE (ψ_0) equals 1. With this treatment mechanism, the probability of treatment will always fall within $[3 \times 10^{-5},1]$, resulting in near or full practical positivity violations for many generated datasets.

4.3.2. Estimation

In this simulation, let us assume that the analyst is not aware of the true data generating mechanism. Suppose then that the analyst uses an outcome model with only main terms in a GLM: $Y \sim A + W_1 + W_2 + P$. The analyst missed the interaction and the squared terms in the regression, which may often happen in practice. This means that the second step of TMLE will be needed, and therefore ϵ_n will be non-null. Using ϵ_n, the second stage of the algorithm updates the initial outcome estimate using the estimated propensity score. For the estimation of the propensity score, the analyst must decide whether to use a parametric model such as a logistic regression of A on all main terms in a GLM or, as suggested in the literature, a more flexible model such as SL. The SL library in this simulation study includes: "glm" for main terms logistic regression, "glm.interaction" for logistic regression with main terms and all first-order interaction terms, "gam" for generalized additive model

and "glmnet" for Lasso with main terms. We generated 500 datasets and ran TMLE and IPTW implemented with these two different approaches for the estimation of the propensity score. While no true bounds exist for the continuous outcome, we nevertheless scaled Y to (0,1) using the sample maximum and minimum values. We then fit the outcome model using a logistic regression as specified above. C-TMLE was implementated with GLM and all main terms and interactions were included in the set of variables to be used in the sequence of propensity score models, thereby allowing the C-TMLE to possibly select the true model.

We present the median and mean statistics in order to summarize the average performance of the estimators. The coverage probability was obtained as the proportion of estimated confidence intervals throughout the 500 generated datasets that contained the true effect, $\psi_0 = 1$. The results for 500 replications are shown in Table 1. We present box plots of the parameter estimates in Figure 1 and density plots of the log of the true and estimated weights in Figure 2.

The IPTW estimators performed poorly whether we used ML or a parametric regression for the estimation of the propensity score with the exception of GLM with 2.5% truncation. IPTW just relies on the propensity score model which was misspecified here. TMLE with a parametric regression for g_n (GLM with all main terms) performed far better than TMLE with SL for g_n across all measures. TMLE with GLM for g_n produced a slightly biased estimate but, overall, the bias and median squared error decreased when we increased the truncation level. When TMLE was fit with SL for g_n, its performance deteriorated across all measures for both sample sizes. C-TMLE with a stepwise variable selection for g_n remained unbiased and achieved the lowest median squared error overall. However, its coverage was sub-optimal for $n = 5000$. From the boxplots in Figure 1, we see again that C-TMLE and TMLE with a GLM for the propensity score model produced estimates with the lowest bias and variability. The density plots of the log of the estimated weights show that the weights obtained using SL were closer to the true weights (i.e. the weights corresponding with the true propensity score) than those estimated using GLM. In particular, large weights were

more prevalent when using SL.

Table 1. Median and mean bias, median squared error and coverage for different bounds of g_n. Estimates taken over 500 generated datasets for different sample sizes, n.

	Bounds on g_n					
	$n = 1000$			$n = 5000$		
Methods	0%	2.5%	5%	0%	2.5%	5%
TMLE-GLM for g_n						
Mean Bias	0.11	0.06	0.07	0.14	0.08	0.08
Median Bias	0.10	0.06	0.07	0.14	0.08	0.10
Median Sq E	0.10	0.09	0.08	0.03	0.02	0.02
Coverage	0.99	0.98	0.97	1.00	0.99	0.96
TMLE-SL for g_n						
Mean Bias	0.38	0.16	0.07	0.37	0.13	0.06
Median Bias	0.37	0.16	0.07	0.36	0.13	0.06
Median Sq E	0.39	0.12	0.06	0.24	0.04	0.01
Coverage	0.47	0.74	0.87	0.34	0.65	0.83
IPTW-GLM for g_n						
Mean Bias	1.90	0.01	0.77	1.95	0.05	0.75
Median Bias	1.80	0.01	0.77	1.92	0.05	0.76
Median Sq E	3.85	0.10	0.60	3.84	0.02	0.60
Coverage	0.28	0.99	0.67	0.02	0.99	0.06
IPTW-SL for g_n						
Mean Bias	1.04	1.50	1.72	0.80	1.47	1.69
Median Bias	1.07	1.51	1.72	0.90	1.47	1.70
Median Sq E	1.26	2.33	3.02	0.86	2.18	2.92
Coverage	0.37	0.00	0.00	0.14	0.00	0.00
C-TMLE-GLM for g_n						
Mean Bias	0.02	0.03	0.03	0.00	0.00	0.00
Median Bias	0.03	0.03	0.03	0.02	0.03	0.03
Median Sq E	0.05	0.05	0.05	0.04	0.04	0.03
Coverage	0.94	0.94	0.94	0.82	0.82	0.83

The simulation study demonstrated that, in the presence of practical positivity violations, when the outcome model is misspecified, using machine learning to predict the treatment mechanism can hurt performance compared to a simple parametric regression. It is known that large weights (caused by near practical positivity violations) can destabilize estimation. TMLE may incorporate the weights when the outcome model is misspecified. So the question remains, how can an analyst detect such instability? In the next section, we suggest an adapted version of the Peterson et al. [**78**] diagnostic tool in order to identify such problems.

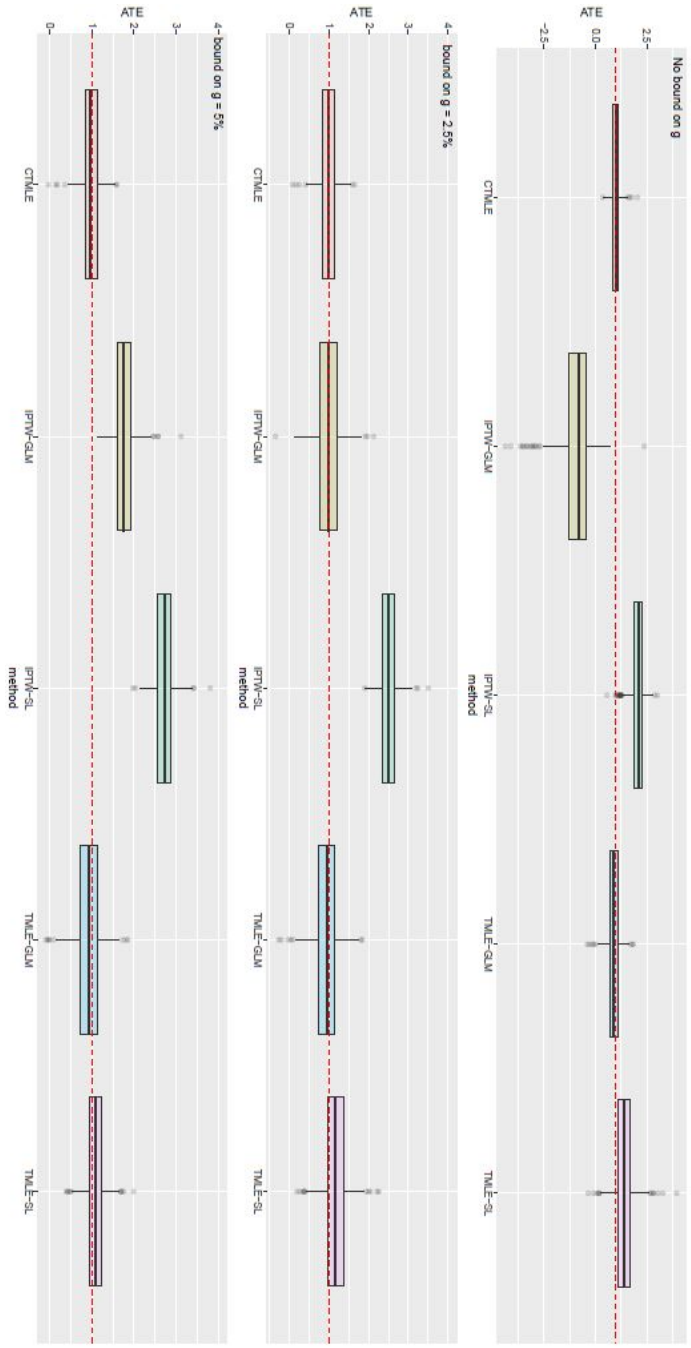

Fig. 1. Boxplots of the ATE with different bounds on g_n for IPTW, TMLE and C-TMLE. $n = 1000$.

4.4. Bootstrap algorithm

In order to introduce the diagnostic tool to inform whether TMLE (or doubly robust estimators) might be destabilized by the use of ML to fit the propensity score, a simple bootstrap simulation of the outcome was employed. Bootstrap resampling [**9**], relies on

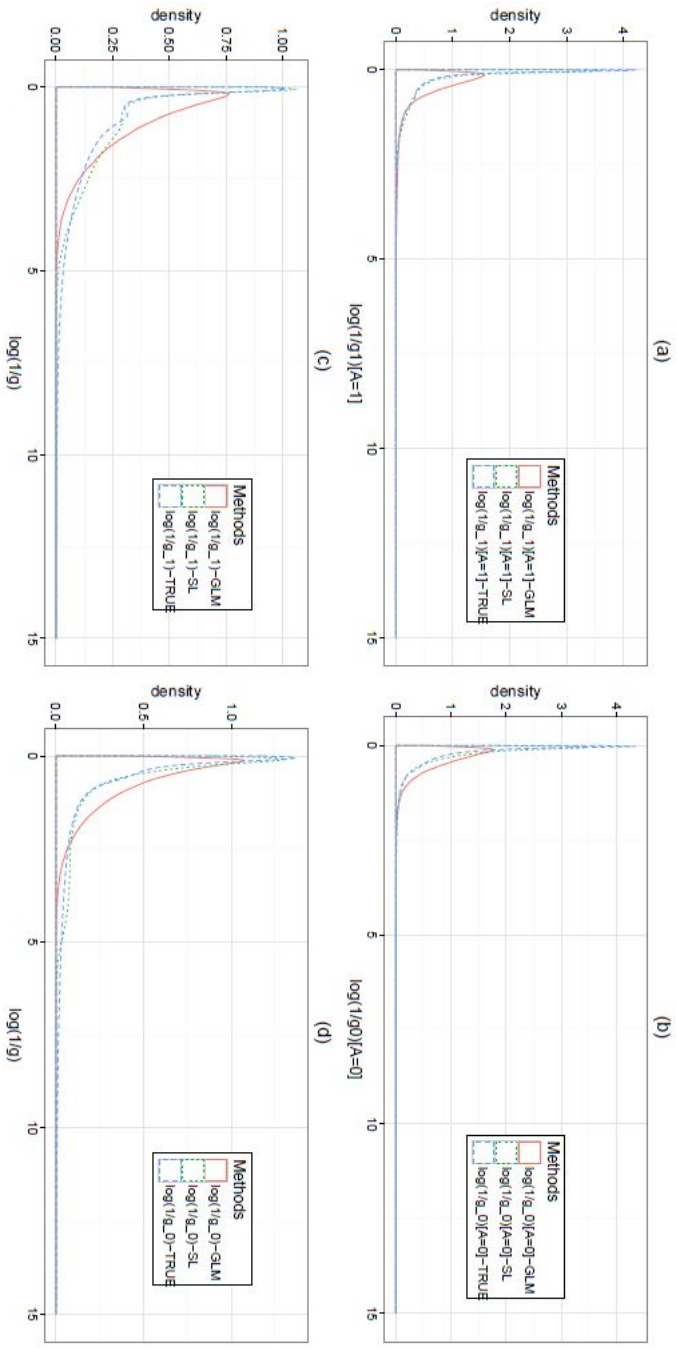

Fig. 2. Density plots of the log of the true and estimated weights for: (a) treatment A=1 in subset of patients with A=1, (b) treatment A=0 in subset of patients with A=0, (c) treatment A=1 for all subjects and (d) treatment A=0 for all subjects.

resampling subjects many times with replacement. The main idea of our simulation follows those of Peterson et al. [**78**], Lendle et al. [**103**] and Franklin et al. [**47**] but instead of simulating both treatment and outcome conditional on the resampled baseline variables, we keep the observed treatment of each resampled subject and only generate the outcome in

order to preserve the associations and structure among covariates and between the covariates and treatment. As the question in our setting is to inform at which point the propensity score estimation can introduce instability, it is important to keep the observed treatment and its natural connection to the observed baseline variables. Since the outcome in our example is continuous, we present this algorithm for use with a continuous Y. However, similar implementations can be easily produced for a binary outcome. Let n denote the sample size of the observed data. The simulation procedure is the following:

Algorithm 4 Adapted Bootstrap Diagnostic Tool (BDT)

1: Consider the two observed subgroups of subjects with $A = 1$ and $A = 0$, respectively. For $a = 1$ and $a = 0$,
   - For subjects with $A = a$, fit a linear regression of Y on W in order to obtain the intercept $\widehat{\beta}_{0a}$, coefficients $\widehat{\beta}_{W_a}$, and $\widehat{\sigma}^2_a$, the estimated conditional variance of Y.
2: Sample n subjects with replacement, and delete the observed outcome values.
3: Using $\widehat{\beta}_{0a}$, $\widehat{\beta}_{W_a}$ and $\widehat{\sigma}^2_a$ obtained in step 1, generate the two potential outcomes from a $\mathcal{N}(\mu_a,\sigma^2_a)$ distribution with $\mu_a = \widehat{\beta}_{0a} + \widehat{\beta}_{W_a}W$, $a \in \{0,1\}$, corresponding to Y^1 and Y^0, for each individual.
4: Taking the resampled data with the simulated outcomes, estimate the parameter of interest with the estimator using a "correct" specification of the outcome model (correct linear regression) and 1) SL and 2) GLM for g_n.
5: Repeat steps 2-4 M times and compute the average bias, variance and Monte-Carlo mean squared error for both approaches.

Since the true data generating distribution is known in the algorithm, the "true" effect in the bootstrap data is known and can be used to assess whether there is a bias increase due to the method used for the estimation of the propensity score. The "true" effect is derived from a contrast of the two potential outcomes, which are computed by simulating exposed and unexposed counterfactual outcomes for all subjects in the population. The average bias is calculated by comparing the mean of the estimator across all bootstrap samples with the true value of the target parameter. The Monte Carlo mean squared error (the squared difference between the true effect and the estimates over all simulations) is used as a measure of estimation variability.

4.4.1. BDT for a single data set

In this section, we apply TMLE, C-TMLE and IPTW on a single dataset obtained using the same data generation and estimation procedure presented in section 3 along with sample size $n = 1000$. We therefore know that the true ATE is $\psi_0 = 1$.

Table 2. Results from one data set (estimates of the average treatment effect and standard error).

Methods	**ATE**	**STD**
TMLE-GLM for g_n	1.01	0.35
TMLE-SL for g_n	1.15	0.17
IPTW-GLM for g_n	1.16	0.51
IPTW-SL for g_n	2.22	0.39
C-TMLE-GLM for g_n	0.98	0.18

TMLE was implemented using both parametric models (GLM) and SL for the estimation of the outcome expectation and propensity score. All of the covariates were included as main terms in the propensity score model as well as in the outcome model. We used the same SL library as in Section 3.2. Table 2 shows the average treatment effect estimates and standard errors based on a single dataset. TMLE and C-TMLE gave an estimate close to the true value when a parametric regression was used for the estimation of the propensity score. However, TMLE with SL for the propensity score exhibited a value 15% larger than the true but reduced the estimated standard error. IPTW produced a larger value than the true and high standard error for both implementations. In our example, we notice that the use of SL increased the point estimate. If we didn't know the true data generating mechanism, we would not know whether the change in estimate produced by using a more flexible method for the propensity score is an improvement in estimation or an instability. We can then use the adapted Bootstrap Diagnostic Tool (BDT) to clarify the change in estimate obtained in Table 2. We also present the results of the bootstrap tool proposed by Peterson et al., where the treatments are simulated using a correctly specified propensity score model (P1) and with an incorrectly specified model that only includes the main covariate terms and no interactions (P2).

Based on $M = 500$ resamples (100 for C-TMLE), the absolute average bias, the mean squared

error (MSE) and the percent coverage (COV) for the estimates of the average treatment effect are tabulated below. Different bounds for the values of g_n were used: 0% (no bounding), 2.5% and 5%. The "true" sample effect obtained by the calculation in the bootstrap data was 0.91. Results are presented in Table 3

When we fit the true outcome expectation (regression of Y on main terms), TMLE remained unbiased overall when we used a parametric regression for the estimation of the propensity score. The average estimated bias was around 0.08 and remained stable when increasing the truncation level. However, using SL for the estimation of the propensity score in the update step of TMLE increased the average bias and decreased the percent coverage. Even though IPTW with GLM for g_n produced a better coverage as compared to TMLE with SL, overall, the TMLE outperformed IPTW. The mean bias and squared error of C-TMLE decreased when increasing the truncation level. The IPTW bias and squared error improved when using SL but the coverage decreased substantially.

Table 3. Results from the BDT and alternative method used on a single simulated data set investigating the absolute average bias, mean squared error and coverage for IPTW and TMLE for different bounds of g_n.

	BDT			P1			P2		
	Bounds on g_n			Bounds on g_n			Bounds on g_n		
Methods	0%	2.5%	5%	0%	2.5%	5%	0%	2.5%	5%
TMLE-GLM for g_n									
Mean Bias	0.08	0.09	0.11	0.12	0.14	0.16	0.12	0.14	0.14
Mean Sq E	0.07	0.07	0.07	0.08	0.08	0.08	0.09	0.09	0.08
Coverage	0.95	0.94	0.92	0.96	0.95	0.92	0.88	0.88	0.88
TMLE-SL for g_n									
Mean Bias	0.42	0.23	0.19	0.38	0.24	0.22	0.13	0.14	0.14
Mean Sq E	0.52	0.21	0.14	0.52	0.22	0.15	0.09	0.09	0.08
Coverage	0.39	0.55	0.62	0.48	0.62	0.69	0.87	0.87	0.87
IPTW-GLM for g_n									
Mean Bias	1.43	0.18	0.84	1.49	0.11	0.76	0.11	0.59	0.92
Mean Sq E	2.39	0.13	0.82	2.56	0.12	0.67	0.19	0.43	0.93
Coverage	0.62	0.99	0.66	0.58	0.99	0.67	0.99	0.84	0.37
IPTW-SL for g_n									
Mean Bias	1.15	1.42	1.64	0.89	1.40	1.60	0.03	0.63	0.95
Mean Sq E	1.43	2.12	2.81	1.04	2.05	2.66	0.17	0.47	0.98
Coverage	0.30	0.03	0.06	0.55	0.01	0.00	0.99	0.80	0.34
C-TMLE-GLM for g_n									
Mean Bias	0.14	0.11	0.09	0.05	0.15	0.14	0.13	0.13	0.13
Mean Sq E	0.12	0.09	0.08	0.08	0.07	0.06	0.07	0.07	0.07
Coverage	0.82	0.85	0.87	0.87	0.86	0.86	0.90	0.91	0.91

Compared with our BDT, the algorithm proposed by Peterson et al. was also able to detect the bias and undercoverage resulting from the usage of SL when the propensity score model used to simulate the treatment was correctly specified. However, when an incorrect propensity score model was used, this method failed to diagnose the same magnitude of bias and low coverage as suggested by the BDT. This is likely due to the generated values of A that do not represent the true relationship between (W,A). In contrast, the BTD does not rely on a specification of the propensity score model and uses the existing values of (W,A) in the procedure, thereby making it a more robust approach.

Table 1 demonstrates how the BDT can be used to investigate the influence of ML for the estimation of the propensity score in the estimation of the treatment effect. Because we used a correct linear regression for Y in the TMLEs, we would expect to have an unbiased estimate of the treatment effect regardless of how the propensity score model was fit. Based of the large amount of bias, MSE and the poor coverage obtained by using ML as compared to the parametric methods, the BDT accurately revealed the fact that the change in the TMLE point estimate obtained in Table 2 was an instability and not an improvement which is due to the estimation of the propensity score.

4.5. Data analysis: Asthma medication during pregnancy

In this section, we use the diagnostic test in an analysis of the safety of asthma medications during pregnancy.

4.5.1. Data description

We used a cohort [**8**] of pregnant women with asthma to study the effect of taking inhaled corticosteroids (ICS) during pregnancy on birth weight. The population of interest is pregnant women with asthma and a singleton delivery in Québec, Canada between 1998-2008, aged ≤ 45 years. This cohort consists of a total of 7,341 pregnancies. Our extraction includes all pregnancies (with at least one diagnosis and prescription of asthma medication in the year

before or during pregnancy) indicated as having mild asthma, as they are clinically eligible to select between taking ICS or not and represent more than 80% of the women in the cohort [**8**]. For simplicity, we considered only the first pregnancy for each woman in this period. Asthma severity was defined according to an index that is based on the Canadian Asthma Consensus Guidelines [**31**]. A total of 4,791 pregnancies in our database fell into this category. All women who filled at least one prescription of ICS during pregnancy were considered exposed, and those who did not were considered unexposed. The outcome of interest is birth weight (continuous in grams). We identified a variety of maternal baseline variables. These potential confounders measured in the year before pregnancy, include demographic characteristics (e.g., provision of income security and place of residence), chronic diseases (e.g., hypertension and diabetes) and variables related to asthma (e.g., at least one hospitalization for asthma, at least one emergency department visit for asthma, and oral corticosteroids). We also included the cumulative daily dose of ICS in the year before pregnancy as a potential confounder. A full list of measured potential confounders can be found in Table 8 in the Appendix [**8, 31**]. The target parameter is the average treatment effect. For our pregnancy cohort, the average treatment effect is the expected difference in the counterfactual birth weight if all women were exposed to ICS during pregnancy versus the counterfactual birth weight if all women were not.

4.5.2. Results of the Analysis

Baseline characteristics of the pregnancy cohort are presented in Table 8. TMLE was implemented using both parametric models (GLM) and SL, for the estimation of the outcome expectation and propensity score. All of the covariates were included as main terms in the propensity score model as well as in the outcome model. The candidate learners in the SL library were: regression (logistic or linear) with main terms, stepwise regression with main terms, and random forests [**62**]. C-TMLE and TMLE were implemented with both a linear regression and SL for the outcome model. Logistic regression was used to estimate

the propensity scores in C-TMLE and in TMLE (with GML for g_n). Results are shown in the Table 4.

Table 4. Estimates of the effect of exposure to ICS on birth weight ($n = 4791$).

Methods	ATE	STD	95% CI	P-value
IPTW–GLM for g_n(trunc 5%)	13.54	86.96	$[-156.90, 183.98]$	0.42
IPTW-SL for g_n(trunc 5%)	18.39	18.18	$[-17.24, 54.03]$	0.27
TMLE-GLM for g_n & $\bar{Q}_n^0$	38.12	30.85	$[-22.35, 98.58]$	0.22
TMLE-GLM for g_n, SL for $\bar{Q}_n^0$	38.19	28.78	$[-17.85, 93.69]$	0.18
TMLE-SL for g_n & $\bar{Q}_n^0$	65.13	11.55	$[38.58, 84.62]$	< 0.01
TMLE-SL for g_n GLM for $\bar{Q}_n^0$	34.58	12.12	$[10.82, 58.34]$	< 0.01
C-TMLE-GLM for $\bar{Q}_n^0$	12.09	17.67	$[-21.83, 44.06]$	0.49
C-TMLE-SL for $\bar{Q}_n^0$	12.75	16.22	$[-19.02, 44.54]$	0.43

IPTW produced a point estimate of 13.54 with a relatively large standard error. However, IPTW with SL for the propensity score produced a point estimate around 18 but improved the estimated standard error. TMLE with a parametric regression for the propensity score produced estimates near 38 with a large reduction in the standard error as compared to IPTW with GLM regardless of how the outcome model was fit. When TMLE was fit with SL for the propensity score and the outcome expectation model, the estimate increased to 65.13, and hypothesis testing concluded that a difference exists. TMLE with SL for g_n and GLM for $\bar{Q}_n^0$ produced a similar significant result with point estimate around 34. The performance of TMLE with SL for g_n produced the smallest estimated standard deviation among all the estimators. C-TMLE limited the variables included in g_n (26 variables selected from the 37). The point estimates of C-TMLE using either a parametric form or SL for the initial outcome expectation model were near 12 with an important improvement in the standard error as compared to TMLE with parametric models. Only the TMLE with SL for g_n concluded that the mean difference in birth weight if all versus no women filled at least one prescription of ICS during pregnancy is different from the null. For an analyst, it is difficult to choose between those models and determine whether the change in estimate produced by the TMLE with SL was due to an instability or due to a true improvement in estimation. We therefore use the BDT algorithm in the next section. While we use this application as a numerical example, we also point out the limitations in a causal interpretation of the results. In

particular, unmeasured confounding may be violated by the absence of a measure of smoking. In addition, we likely have a violation of the well-defined intervention assumption. In our data, exposed women didn't necessarily have the same cumulative dose of ICS, because the outcome may depend on the dose, which likely violates the consistency assumption. Difficulty in assessing exact medication exposure is a common limitation in studies involving electronic health data [**61**].

4.5.3. Bootstrap diagnostic test

Results for the BDT are presented in Table 5. The "true"effect obtained in this bootstrap data is equal to 19.10. Based on $M = 500$ resamples (with a random outcome generation), the absolute average bias, the Monte Carlo standard deviation (STD) and root mean squared error (RMSE) and percent coverage (COV) for the effect estimates are tabulated below. We also ran C-TMLE with a correctly specified outcome model in order to compare its performance.

Table 5. BDT results investigating the absolute average bias, root mean squared error and the percent coverage for TMLE, C-TMLE and IPTW.

Methods	**Bias**	**STD**	**RMSE**	**COV**
TMLE-GLM for g_n & $\bar{Q}_n$	0.05	34.33	34.31	0.88
TMLE-SL for g_n, GLM for $\bar{Q}_n$	13.68	69.23	70.51	0.30
IPTW-GLM for g_n	15.41	17.09	23.01	0.86
IPTW-SL for g_n	13.29	18.01	22.37	0.87
C-TMLE-GLM for $\bar{Q}_n$	5.23	32.14	32.52	0.89

In Table 5 when adjusting for all covariates, TMLE remained unbiased when we fit the true outcome model and used a logistic regression for g_n. The contribution of the propensity score did not impact the bias. However, the bias and the root mean squared error increased when SL was used for g_n. IPTW produced a larger bias and acceptable coverage in the bootstrap simulations. The bias suggests that the parametric model for g_n is misspecified. C-TMLE by its variable selection procedure improved the estimate in root mean squared error by introducing a little bias. TMLE with machine learning for the propensity score produced confidence intervals that failed to cover the "true" effect compared to the TMLE

and C-TMLE that used parametric specifications for g_n. The BDT was able to clarify that SL for the estimation of the propensity score likely did not improve the TMLE point estimate based on the larger bias and poor coverage obtained in the bootstrap data with a correctly specified outcome model. An analyst could then conclude that the adjusted mean difference in birth weight is likely not different from the null.

4.6. Discussion

In this paper, we have exhibited a situation where ML for the treatment mechanism can increase bias of the treatment effect as compared to parametric regression. We then provided an adapted version of the diagnostic tool of Peterson et al. [**78**], to diagnose the instability introduced when machine learning is employed for the estimation of the propensity score. We focused on the application of TMLE to estimate the average treatment effect. We used parametric and data-adaptive (SL) methods for the initial outcome expectation and propensity score models. Through simulation studies and real data analysis, we illustrated that the BDT can help diagnose whether TMLE was likely to be destabilized by the propensity score. The main goal of the BDT is to inform at which point the estimation of the propensity score with ML can hurt performance of the treatment effect. One may also use BDT with IPTW, which is only based on the treatment mechanism, to provide evidence for whether IPTW is producing unbiased estimation.

While the causal interpretation of our example is somewhat limited, the results suggest that the usage of ICS during pregnancy for women with mild asthma does not affect birth weight. The results of the real study are consistent with the results found in another study investigating the safety of ICS during pregnancy [**8**]. However, the blind usage of TMLE with ML would have suggested a reduction in birth weight for the women who didn't receive ICS. The BDT enabled us to conclude that this divergent result was likely due to the instability arising from the weights rather than the improved estimation of the exposure model using machine learning. This paper points to the importance of the new developments [**117, 63, 13**] that produce valid inference and $\sqrt{n}$-convergence speeds even when ML methods are used in

TMLE. In conclusion, the diagnostic tools can provide important insight when using data-adaptive methods to fit the propensity score and all interpretation of the results should be made with caution.

Chapitre 5

Doubly Robust Adaptive LASSO for Effect Modifier Discovery

Cet article est en révision pour une publication dans le journal International Journal of Biostatistics.

Préambule: Les travaux présentés dans l'article 1 peuvent être appliqués sur des problèmes ou l'on s'intéresse à un effet causal moyen d'un traitement A sur une issue Y. Toutefois, il arrive souvent que l'effet causal moyen soit différent de l'effet individuel ou de l'effet dans un groupe. Ainsi, il est important de porter une attention à l'effet dans ces groupes advenant que le traitement puisse leur être bénéfique. L'étude décrite dans cet article se propose de développer une procédure doublement robuste pour sélectionner un modificateur d'effet et estimer un effet conditionnel. Pour ce faire, on utilise une issue modifiée comme celle utilisée dans Kennedy et al., [**30**] qui permet d'intégrer des méthodes d'apprentissage automatique pour l'estimation des paramètres de nuisance. La méthode est évaluée via des études de simulation dans le but d'évaluer sa performance. Finalement, elle est appliquée sur une base de données mesurant l'impact de la prise de corticostéroïdes durant la grossesse sur le poids du nouveau-né.

Doubly Robust Adaptive LASSO for Effect Modifier Discovery

Asma Bahamyirou[1], Mireille E. Schnitzer[1], Lucie Blais[1], Edward H. Kennedy[2], Yi Yang[3].

1 : Faculté de pharmacie, Université de Montréal.

2 : Department of Statistics and Data Science, Canergie Mellon University.

3 : Department of Mathematics and Statistics, McGill University.

Résumé: On parle de modification de l'effet quand l'effet du traitement sur une issue diffère selon les niveaux d'une troisième variable. Cette dernière est dénommée le modificateur d'effet. Une façon naturelle de détecter une modification de l'effet est d'effectuer des analyses de sous-groupes ou d'inclure un terme d'interaction entre le traitement et les co-variables lors d'une analyse de régression. L'analyse de régression toutefois, ne cible pas les paramètres du modèle structurel marginal (MSM) à moins que le vrai modèle de l'issue ne soit spécifié. L'objectif est de développer une méthode flexible pour sélectionner un modificateur d'effet dans un modèle structurel marginal. Une procédure en deux étapes est proposée. En premier lieu, nous estimons les deux paramètres de nuisance (espérance de l'issue conditionnelle et le score de propension) et nous les substituons dans une fonction de perte doublement robuste. En second lieu, nous utilisons le LASSO adaptatif pour sélectionner les modificateurs d'effets et estimer les coefficients du MSM. L'inférence sélective est utilisée pour obtenir les probabilités de recouvrement associées aux variables sélectionnées. Des études de simulations ont été effectuées dans le but d'évaluer la performance de notre estimateur.

Keys words: Estimateur doublement robuste, LASSO adaptatif, modificateur d'effet, Inférence sélective.

Abstract: Effect modification occurs when the effect of the treatment on an outcome differs according to the level of a third variable (the effect modifier, EM). A natural way to assess effect modification is by subgroup analysis or include the interaction terms between

the treatment and the covariates in an outcome regression. The latter, however, does not target a parameter of a marginal structural model (MSM) unless a correctly specified outcome model is specified. Our aim is to develop a data-adaptive method to select effect modifying variables in an MSM with a single time point exposure. A two-stage procedure is proposed. First, we estimate the conditional outcome expectation and propensity score and plug these into a doubly robust loss function. Second, we use the adaptive LASSO to select the EMs and estimate MSM coefficients. Post-selection inference is then used to obtain coverage on the selected EMs. Simulations studies are performed in order to verify the performance of the proposed methods
Keys words: Doubly robust, Adaptive LASSO, Effect modification, Selective inference.

5.1. Introduction

Effect modification occurs when the effect of a treatment on an outcome differs according to the level of some pre-treatment variables (the effect modifier, EM). Detecting variables that are EMs is not a straight-forward task even for a subject matter expert. A natural way to assess effect modification in experimental and observational studies is to perform subgroup analysis, in which observations are stratified based on the potential EMs after which stratum-specific estimates are calculated, though this becomes infeasible with a greater number of potential effect modifiers. One can also include the interaction terms between the treatment and the potential EMs in an outcome regression analysis. With observational data however, this approach does not target a parameter of a marginal structural model (MSM) unless a correct model for the outcome conditional on confounders, treatments, and EMs is specified. In contrast, MSMs can provide a summary of how effect modification occurs in the absence of confounding. Different methods for the estimation of effect modification have been proposed recently. For example, Green and Kern [**26**] used Bayesian Additive Regression Trees [**36**] to model the conditional average treatment effects (CATE). Imai and Ratkovic [**58**] studied EM selection by adapting the support vector machine classifier. Nie and Wager [**119**] developed a two-step algorithm for heterogeneous treatment effect estimation using the marginal

effects and treatment propensities. Luo et al., [**116**] used dimension reduction techniques to learn heterogeneity by estimating a lower dimensional linear combination of the covariates that is sufficient to model the regression causal effects. Wager and Athey [**99**] proposed a nonparametric approach for estimating heterogeneous treatment effects using a random forest algorithm [**62**]. Powers et al.,[**97**] developed an algorithm for heterogeneous treatment effect estimation by adapting the multivariate adaptive regression splines [**39**]. Zhao et al. [**84**] introduced an algorithm based on a semiparametric model that selects the EMs by using Robinson's transformation [**82**] and Least Absolute Shrinkage and Selection Operator (LASSO). Doubly robust semiparametric methods such as Targeted Minimum Loss-Based Estimation (TMLE) [**73, 76**], which is closely related to previously existing methods [**32, 25**] have been proposed. The term doubly robust comes from the fact that the method requires both the estimation of the treatment model and the outcome expectation conditional on treatment and covariates, where only one of which needs to be correctly modeled to allow for consistent estimation of the parameter of interest. However, in a situation where one nuisance parameter is inconsistently estimated, the asymptotic linearity is affected [**14**]. Lee et al. [**96**] developed a doubly robust estimator of the CATE along with a uniform confidence band. Rosenblum and van der Laan [**76**] developed TMLE for MSMs, which can be used to model effect modification, in non-longitudinal settings. Zheng et al. [**118**] developed TMLE for MSMs with counterfactual covariates in longitudinal settings. Most recently, Kennedy [**33**] analyzed a version of the pseudo-outcome regression method for CATE estimation and derives model-free error bounds.

In this paper as in [**84**], we focus on the selection of pre-treatment EMs in a linear MSM for the CATE with a single treatment time-point. Thus, we consider modifiers of the additive effect of a treatment on the mean outcome. We use a component of the efficient influence function of the ATE along with the Adaptive LASSO [**35**] to select EMs. To the best of our knowledge, our paper is one of the first along with [**33, 114**] to investigate and apply a doubly robust two-stage regularization for a CATE model. Our estimation approach can be carried out with standard software implementations, is doubly robust (unlike [**84**]),

can accommodate adaptive methods to estimate the nuisance quantities, and produces estimates of the parameters of an easily interpretable model. A two-stage procedure is thus proposed. First, we estimate two nuisance quantities (the conditional outcome expectation and treatment model) and plug these quantities into a specific function to create a pseudo outcome as developed in [**72, 123, 30**]. Second, we take the pseudo outcome and apply the adaptive LASSO [**35**] to select the EMs and estimate the MSM coefficients. We then apply post-selection inference in order to produce interpretable confidence intervals after the EM selection by adaptive LASSO. We perform simulation studies in order to verify the performance (selection, estimation, double robustness, and post-selection inference) of the proposed method.

The remainder of this article is organized as follows. In Section 2, we use the potential outcomes framework to define the target causal parameter of interest and describe our proposed estimation approach. In Section 3, we conduct a simulation study to verify the performance (selection, MSM coefficient estimation, and double robustness) of the proposed method in both low and high dimensional settings. We present an analysis of the safety of asthma medications during pregnancy in Section 4. A discussion is provided in Section 5.

5.2. Methods

In this section, we present our development of the methodology for the selection of the EMs.

5.2.1. The framework

The observed data, $\{(\boldsymbol{W}_i,A_i,Y_i)\}_{i=1}^n$, are comprised of independent and identically distributed samples of $O = (\boldsymbol{W},A,Y) \sim P_0$, where $\boldsymbol{W}$ is the baseline covariates of a patient, A is the binary treatment which equals 1 if the patient received treatment and 0 otherwise, and Y is the observed outcome (binary or continuous). Let $\boldsymbol{V}$ represent the subset of the variables in $\boldsymbol{W}$ that represents the potential EMs of interest. We use $O_i = (\boldsymbol{W}_i,A_i,Y_i)$ to represent the i-th observation of the data. In order to define the

target parameter, we use the counterfactual framework of Rubin [**21**]. Let Y^a denote the potential (or counterfactual) outcome that would have occurred under the treatment value $A = a$. In this paper, we focus on marginal models for the CATE. If we assume that we observe $Y = Y^a$ when $A = a$ (consistency [**106**], no interference, positivity and no unmeasured confounders [**69**]), the CATE can be defined and identified nonparametrically as:

$$\begin{aligned} \psi_0(\boldsymbol{V}) &= E_0\{Y^1 - Y^0|\boldsymbol{V}\} \\ &= E_{\boldsymbol{W}|\boldsymbol{V}}\{\underbrace{E_0(Y|A=1,\boldsymbol{W})}_{\bar{Q}_0(1,\boldsymbol{W})} - \underbrace{E_0(Y|A=0,\boldsymbol{W})}_{\bar{Q}_0(0,\boldsymbol{W})})|\boldsymbol{V}\} \\ &= E_{W|\boldsymbol{V}}\{\bar{Q}_0(1,\boldsymbol{W}) - \bar{Q}_0(0,\boldsymbol{W})|\boldsymbol{V}\} \end{aligned} \tag{5.2.1}$$

where E_0 is the expectation with respect to the outcome and $E_{W|V}$ is the expectation conditional on the baseline covariates. In this work, we choose to model the CATE using a linear regression model defined as $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \boldsymbol{V}^T\boldsymbol{\beta}_V$ where the relevant subset of $\boldsymbol{V}$ will be selected using adaptive LASSO [**35**]. Our goal here is to identify the true EMs among the set $\boldsymbol{V}$, and estimate their associated coefficients. One could use non-linear models or machine learning methods to estimate $\tilde{\psi}_0(\boldsymbol{V})$, which is important when the goal is prediction [**56**] (e.g. for personalized medicine). However, if interpretation of the coefficient associated with each $V^{(s)}$ is important, it may be beneficial to use a linear model rather than a black box approach.

5.2.2. Adaptive LASSO

The adaptive LASSO [**35**] is an extension of the traditional LASSO of Tibshirani [**88**] that uses coefficient specific weights. Zou [**35**] showed that the adaptive LASSO estimator has the oracle property which roughly means that the algorithm identifies the right subset of variables (consistency of variable selection) and that the coefficient estimators of the selected variables are asymptotically normal. In a prediction (non-causal) setting, let Y be an observed outcome and $\boldsymbol{V}$ a set of covariates. Under the linear model, we can select

predictors of Y by solving the equation below:

$$\arg\min_{\alpha',\boldsymbol{\beta}'} \sum_{i=1}^{n}(Y_i - \alpha' - \boldsymbol{V}_i^T\boldsymbol{\beta}')^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j|\beta'_j| \tag{5.2.2}$$

where $\boldsymbol{\beta}' = (\beta'_1,...,\beta'_p)$, $\widehat{w}_j = 1/|\tilde{\beta}'_j|^\gamma$, for some $\gamma > 0$ and $\tilde{\beta}'_j$ is a $\sqrt{n}$-consistent estimator of β'_j. The selected variables are the positions of the non-zero entries of the solution of Equation (5.2.2) . When the sample size grows, the weights associated with the zero-coefficient predictors tend to infinity, while the weights corresponding to true predictors converge to a constant. Thus, true-zero coefficients are less likely to be selected by the adaptive LASSO than by the standard LASSO, which does not have the oracle property [**35**].

5.2.3. Highly Adaptive LASSO (HAL)

Assume $E(Y|V)$ a regression function where Y is the observed outcome and V is the set of covariates. Consider a map of $\boldsymbol{V}$ onto a set of binary indicator basis functions. For example, if $\boldsymbol{V}$ is scalar, we generate for an observation v, $\boldsymbol{\phi}^*(v) = (\phi_1^*(v),...,\phi_n^*(v))^T$, where $\phi_i^*(v) = I(v \geq V_i)$, for $i = 1,...,n$. With two dimensions, $\boldsymbol{V} = (V^{(1)},V^{(2)})^T$, we need to include the second order basis functions $\boldsymbol{\phi}_i^*(\boldsymbol{v}) = I(v_1 \geq V_i^{(1)},v_2 \geq V_i^{(2)})$, for $i = 1,...,n$. The HAL estimator [**15**] is obtained by fitting a L_1-penalized regression of the outcome Y on these basis functions, with the optimal L_1-norm chosen via cross-validation. The HAL estimator of the regression function $E(Y|\boldsymbol{V})$ converges to the true regression function in L_2-norm no slower than $n^{-1/4}$ regardless of the dimension of $\boldsymbol{V}$, under the assumption that the regression function has bounded variation norm.

5.2.4. Selective inference

Let $\widehat{\boldsymbol{\beta}}'$ be the solution of Equation (5.2.2) and $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$ the non-zero subvector of $\widehat{\boldsymbol{\beta}}'$ where $\widehat{M} \subseteq \{1,...,p\}$ corresponds to the positions of the non-zero entries. Suppose that we are interested in making inference for $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$ in the prediction model of Section 5.2.2. A naive way to obtain inference after selecting the covariates in the model is the standard hypothesis tests for linear regression that treat M, representing the non-zero entries of $\boldsymbol{\beta}'$ and thus

the true model, as known. It is easy to see that $\widehat{\boldsymbol{\beta}}'$ depends on the selected model $\widehat{M}$. Therefore, Lee et al., [**42**] studied the conditional distribution $\widehat{\boldsymbol{\beta}}'_M|\{\widehat{M}=M\}$ and showed that this conditional distribution is a truncated normal Gaussian. They constructed a pivotal statistic for $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$ which can be used for hypothesis testing and therefore by test inversion, to construct a confidence interval. Let $F(y;\mu,\sigma^2,l,u)$ be the CDF of a normal $N(\mu,\sigma^2)$ truncated to the interval $[l,u]$, e_j the unit vector for the j-th coordinate so that $(\widehat{\beta}'_M)_j=\eta_M^T Y$, $\eta_M=[(\boldsymbol{V}_M^T\boldsymbol{V}_M)^{-1}\boldsymbol{V}_M^T]^T e_j$ and $\sigma_*^2=\sigma^2\eta_M^T\eta_M$. In the linear regression setting where $Y\sim N(\mu,\sigma^2 I_n)$, Lee et al., [**42**] showed that $F((\widehat{\beta}'_M)_j;(\beta'_M)_j,\sigma_*^2,\nu^-,\nu^+)|\{\widehat{M}=M\}\sim Unif(0,1)$, where $[\nu^-,\nu^+]$ is defined in [**42**] as a function of Y and the model M. By inverting the hypothesis testing, we can find a $(1-\alpha)$ confidence interval for $(\widehat{\beta}'_M)_j$, conditional on $\widehat{M}=M$, by finding $[L^*,U^*]$ such that

$$F((\widehat{\beta}'_{\widehat{M}})_j;L^*,\widehat{\sigma}_*^2,\nu^-,\nu^+)|\{\widehat{M}=M\}=1-\alpha/2$$

and

$$F((\widehat{\beta}'_{\widehat{M}})_j;U^*,\widehat{\sigma}_*^2,\nu^-,\nu^+)|\{\widehat{M}=M\}=\alpha/2.$$

In this next section, we will explain how this result is applied in our setting.

5.2.5. The model

5.2.5.1. Model definition

Let $\psi_0(\boldsymbol{V})=E_0\{Y^1-Y^0|\boldsymbol{V}\}$ be the CATE. Denote $\bar{Q}_0(a,\boldsymbol{W})=E_0(Y|A=a,\boldsymbol{W})$, the outcome expectation, and $g_0(a|\boldsymbol{W})=P(A=a|\boldsymbol{W})$ as the propensity score. We suggest to use the doubly robust and efficient loss-function proposed by van der Laan [**72**], inspired by Rubin and van der Laan [**17**], $L_{Q_0,g_0}(\psi)(O)=(D(\bar{Q}_0,g_0)(O)-\psi_0(\boldsymbol{V}))^2$ where

$$D(\bar{Q}_0,g_0)(O)=\frac{2A-1}{g_0(A|\boldsymbol{W})}(Y-\bar{Q}_0(A,\boldsymbol{W}))+\bar{Q}_0(1,\boldsymbol{W})-\bar{Q}_0(0,\boldsymbol{W}) \tag{5.2.3}$$

is indexed by the nuisance parameters $(\bar{Q}_0;g_0)$. A similar pseudo-outcome is also used in Zhao et al. [**123**] for estimating optimal individualized treatment rules and Kennedy et al.

[**30**] for the estimation of continuous treatment effects.

The next lemma shows that if one of the two nuisance quantities are consistent, the CATE can be obtained by the conditional expectation of the estimated pseudo-outcome.

Lemma 1. *Let* $||f||^2_{2,P_0} = \int f(z)^2 dP_0(z)$ *denote the* $L^2(P_0)$ *norm. Suppose either* $\bar{Q}_n$ *converges to* $\bar{Q}_0$ *or* g_n *converges to* g_0 *in the sense that* $E||\bar{Q}_n - \bar{Q}_0||^2 = o(1)$ *or* $E||g_n - g_0||^2 = o(1)$ *(not necessarily both). Then* $E(D(\bar{Q}_n,g_n)(O)|\boldsymbol{V}) \to \psi_0(\boldsymbol{V})$ *as* $n \to \infty$.

The preceding lemma shows that the pseudo-outcome we propose for the CATE is doubly-robust in the sense that if at least one nuisance estimator ($\bar{Q}_n$ or g_n) converges to the correct function, but not necessarily both, then a regression of the pseudo-outcome onto the effect modifiers will be consistent for the CATE. Adding and subtracting the true CATE is the key idea to prove Lemma 1. Then, the regression function of the pseudo-outcome on V can be split into two terms: the true CATE and a second term that is a function of both $\bar{Q}_n - \bar{Q}_0$ and $g_n - g_0$. See the Appendix for the proof of Lemma 1.

Suppose that an investigator would like to identify the true EMs amongst multiple suspected effect modifying variables $\boldsymbol{V} = (V^{(1)},...,V^{(p)})$. As described above, to accomplish this we use a linear model for the CATE with corresponding MSM defined as $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \boldsymbol{V}^T\boldsymbol{\beta}_V$ under a least squared error loss function. We then use the adaptive LASSO estimator [**35**] to select amongst the $V^{(j)}$s. More specifically, as suggested by Rubin and van der Laan [**18**], we penalize the aforementioned loss function $L_{\bar{Q}_0,g_0}$ by the adaptive LASSO penalty. Let $D_n = D(\bar{Q}_n,g_n)(O)$ be the estimated pseudo outcome and $D_{i,n}$ the i-th observation of D_n. The parameters of the MSM $\boldsymbol{\beta} = (\beta_0,\beta_1,...,\beta_p)$ are estimated by minimizing the risk function below:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} (D_{i,n} - \tilde{\psi}_0(\boldsymbol{V}_i))^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j| \tag{5.2.4}$$

where $\widehat{\omega}_j = 1/|\tilde{\beta}_j|^\gamma$, for some $\gamma > 0$ and $\tilde{\beta}_j$ is a $\sqrt{n}$-consistent estimator of β_j.

An optimal method would possess the oracle property, able to select the appropriate variables and unbiasedly estimate the selected parameters. Let $\mathbf{A}$ be the set of true variables in the model and $\mathbf{A}^*_n$ be the set selected using adaptive LASSO.

Lemma 2. *Let $D_n = D(\bar{Q}_n, g_n)$ be the estimated pseudo-outcome conditional on the estimated nuisance functions. Assume $E(D_n|\boldsymbol{V}) = \beta_0 + \boldsymbol{V}^T\boldsymbol{\beta}_V$ and $|\mathbf{A}| = p_0 < p$. Suppose that $\lambda/\sqrt{n} \to 0$ and $\lambda n^{(\gamma-1)/2} \to \infty$. Also, assume D_n is obtained by cross-fiting and is consistent in the sense that it belongs to a shrinking neighborhood of D_0 as given in Assumption 3.5 in Semenova and Chernozhukov (2020)* [**114**]*. The proposed estimator $\widehat{\boldsymbol{\beta}}$ inherits the adaptive LASSO oracle properties, i.e.*

- *Consistency in variable selection (i.e. identifies the right subset model):* $\lim_{n\to\infty} P(\mathbf{A}_n^* = \mathbf{A}) = 1$.
- *Asymptotic normality (i.e. has the optimal estimation rate):* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathbf{A}} - \boldsymbol{\beta}_{\mathbf{A}}) \to_d N(0, \Sigma^*)$*, where Σ^* is the covariance matrix knowing the true subset model and $\widehat{\boldsymbol{\beta}}_{\mathbf{A}}$ is the coefficient estimates resulting from the Adaptive LASSO regression of D_n on V.*

As a consequence, our proposed estimator is able to select the correct subset of EMs and produce an unbiased estimate of the MSM coefficients in large samples. See the Appendix for the proof of Lemma 2. This relies on convergence of D_n to D_0 which can result from correct specification of the models for g_n and/or $\bar{Q}_n$ [**114**]. See the Appendix for the proof of Lemma 2.

5.2.5.2. Estimation

In this paragraph, we describe how our proposal can be easily implemented in a two-stage procedure. In the first stage, we construct the pseudo-outcome function by producing estimates $\bar{Q}_n(a, \boldsymbol{W})$ and $g_n(a|\boldsymbol{W})$ of the two nuisance quantities and plugging them into D. Machine Learning (ML) methods are often recommended [**76**] for estimating $\bar{Q}_n$ and g_n. In the second stage, we run the adaptive LASSO regression of the estimated pseudo-outcome $D(\bar{Q}_n, g_n)(O)$ on the set $\boldsymbol{V}$. The selected EMs correspond to the non-zero coefficients of the adaptive LASSO regression.

The proposed algorithm for estimating the parameters in the CATE model with a given value of λ is as follows:

Algorithm 5 Effect modifiers adaptive LASSO algorithm

1: Estimate the outcome expectation $\bar{Q}_n(a,\boldsymbol{W}) = \hat{E}(Y|A=a,\boldsymbol{W})$ for each subject.
2: Obtain the estimated propensity score $g_n(a|\boldsymbol{W}) = \hat{P}(A=a|\boldsymbol{W})$ for each subject.
3: Construct an estimate of the doubly robust function D_n by plugging in the estimated $\bar{Q}_n$ and g_n.
4: Select the effect modifiers by following steps (a)-(d) below:
   (a) Run a linear regression of D_n on $\boldsymbol{V}$ as the set of covariates. Obtain $\tilde{\beta}_j$, the estimated coefficient of $V^{(j)}$, $j=1,...,p$.
   (b) Define the weights $\widehat{\omega}_j = \frac{1}{|\tilde{\beta}_j|^\gamma}$, $j=1,...,p$ for some $\gamma > 0$.
   (c) Run a LASSO regression of D_n on $\boldsymbol{V}$ with $\widehat{\omega}_j$ as the penalty factor associated with $V^{(j)}$ with a given λ.
   (d) The non-zero coefficients of the solution of the adaptive LASSO regression $\{\widehat{\beta}_j\}_{j=1}^p$ are the selected effect modifiers.
5: The final estimate of the CATE is $\psi_n(\boldsymbol{V}) = \widehat{\beta}_0 + \sum_{j=1}^p V^{(j)}\widehat{\beta}_j$.

For the adaptive LASSO tuning parameters, we choose $\gamma = 1$ (Nonnegative Garotte Problem [**67**]) and λ is selected using cross-validation as suggested by Zou [**35**]. The traditional cross-validation minimizes the prediction error knowing the true outcome. In our setting, the Adaptive LASSO is run with the estimated pseudo-outcome as the "true" outcome. We conjecture that if the two nuisance parameters are consistently estimated at fast enough rates, we should be able to use the estimated pseudo-outcome to find an optimal tuning parameter. This conjecture agrees with recent results from Kennedy (2020) [**33**]. Naive inference by ignoring the EM selection would result in incorrect confidence intervals. Zhao et al. [**84**] showed that when the outcome is observed with error, the selective pivotal statistic proposed by Lee et al. [**42**] is still asymptotically valid. Thus we apply their methodology which is expected to produce valid asymptotic results as long as $\bar{Q}_n$ is consistent and both $\bar{Q}_n$ and g_n converge faster than at a $n^{1/4}$ rate in the l_2 norm [**113**]. In order to construct a selective 95%-confidence intervals for the selected submodel, we use the R package **selectiveInference** [**89**] for post-selection inference. The estimated $\widehat{\sigma}^2$ used in the package is the variance of the residual from fitting the full model in $4(a)$.

5.3. Simulation study

5.3.1. Data generation and parameter estimation

To evaluate the performance of the proposed method in finite samples, we conducted a simulation study under four scenarios. We simulated data $O = (\boldsymbol{W}, A, Y)$ representing baseline covariates $\boldsymbol{W}$, a binary exposure A, and a continuous outcome Y. The baseline covariates $\boldsymbol{W}$ include three confounders $(X, V^{(1)}, V^{(2)})$, one instrument Z (pure cause of treatment), and two pure causes of the outcome $(V^{(3)}, V^{(4)})$. All covariates were generated independently with the Bernoulli distribution with success probability p: $X \sim B(p = 0.4)$, $V^{(1)} \sim B(p = 0.5)$, $V^{(2)} \sim B(p = 0.6)$, $V^{(3)} \sim B(p = 0.5)$, $V^{(4)} \sim B(p = 0.7)$ and $Z \sim B(p = 0.45)$.

We varied the strength of the relationship between covariates, outcome and treatment across three low-dimensional scenarios. In the first, we used an outcome model where the covariates were strongly predictive, and a treatment model where the covariates were weakly predictive. The treatment mechanism g_0 was set as a Bernoulli with the probability generated linearly in the three confounder variables and single instrument,

$$P_0(A = 1|X) = \text{expit}\{0.5Z - 0.2X + 0.3V^{(1)}1 + 0.4V^{(2)}\}$$

where $\text{expit}(x) = 1/\{1 + exp(-x)\}$. The observed continuous outcome Y was linearly generated as:

$$Y = 1 + A - 0.5X + 2V^{(1)} + V^{(2)} + V^{(3)} - 0.2V^{(4)} + 4V^{(1)}V^{(2)}V^{(3)} + A(0.5V^{(1)} + V^{(3)}) + N(0,1)$$

The effect modification arises due to interaction between treatment and covariates.

The second scenario has the same data generation except that the coefficient of the interaction term $V^{(1)}V^{(2)}V^{(3)}$ is 0 instead of 4. In the third scenario, we use an outcome model where the covariates are weakly predictive, and a treatment model where the covariates are strongly predictive. We focus here on the first scenario and describe all other simulations settings and results in the Appendix.

We thus have two EMs ($V^{(1)}$,$V^{(3)}$), where the first is a confounder and the second is a pure cause of the outcome. In practice, we are not aware of the true data generating mechanism. So we have a potential set of EMs: $\boldsymbol{V} = (V^{(1)},V^{(2)},V^{(3)},V^{(4)})$. Let $\psi_0(\boldsymbol{V}) = E_{P_0}(Y^1 - Y^0|\boldsymbol{V})$ be the true (nonparametric) CATE, which we model as an MSM: $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \beta_1 V^{(1)} + \beta_2 V^{(2)} + \beta_3 V^{(3)} + \beta_4 V^{(4)}$. Our goal here is to identify among the set $\boldsymbol{V}$, the true EMs and estimate their associated coefficients. Given the data generated, the true values of the coefficients are $\boldsymbol{\beta}_v = (0.5,0,1,0)$. We set $n = 1000$ and then 10000. We also add a smaller sample size $n = 100$ with results in the appendix.

To evaluate the performance of our method in high-dimensional settings, we also extend the first scenario by adding 50 pure binary noise covariates (unrelated to treatment or outcome) to our set of covariates, which are included as potential confounders and EMs. The true values of the coefficients in the MSM are thus $\boldsymbol{\beta}_v = (0.5,0,1,0,...,0)$.

Under each low-dimensional scenario, we tested our proposed method under four different implementations:

(1) Qcgc: Both of the models for $\bar{Q}$ and g are correctly specified using generalized linear models (GLMs).

(2) Qc: Only the GLM for $\bar{Q}$ is correctly specified. g is misspecified using a logistic regression of treatment A on variable X.

(3) gc: Only the GLM for g is correctly specified. $\bar{Q}$ is misspecified using a GLM of treatment Y on variables A and $V^{(3)}$.

(4) HAL: Both $\bar{Q}$ and g are estimated using the Highly Adaptive LASSO (HAL) [32,33] . We use the package default setting.

For comparison, we also tested two implementations of a linear regression model for the outcome to directly assess effect modification:

(5) NLin: Linear regression with main terms (treatment and all covariates) and interactions between treatment and covariates. Only first-order interactions were included.

(6) CLin: Linear regression with a correctly specified outcome model.

Standard confidence intervals are presented for the linear model case and, in our summary, a p-value of less than 0.05 is used as a criterion for a variable to be selected. In the higher dimensional scenario, only HAL was used to estimate $\bar{Q}$ and g.

5.3.2. Simulation results

For each scenario, we produced boxplots of the MSM coefficient estimates. We also present the percent selection, the coverage proportion of the confidence intervals and the false coverage rate in order to summarize the average performance of each estimator and implementation. The percent selection for our LASSO method was obtained as the percentage of estimated coefficients that are non-zero throughout the 1000 generated datasets, and for the linear regression, the percentage of p-values < 0.05. The coverage for each true effect modifier was obtained as the number of times the true model was selected and the corresponding confidence intervals contained the true coefficients, divided by the number of times the true model was selected. For the linear regression, the percent coverage was instead calculated for each coefficient and defined as the proportion of the confidence intervals that contained the true coefficient throughout the 1000 generated datasets. The false coverage rate (FCR) for our LASSO model was obtained as the number of non-covering confidence intervals among the selected coefficients, divided by the number of the selected coefficients throughout the 1000 generated datasets [**42**].

For the first low-dimensional scenario, Figures 1 and 2 contain the boxplots of the MSM coefficient estimates for the true EMs ($V^{(1)}$,$V^{3)}$) and non-EMs ($V^{(2)}$,$V^{(4)}$), respectively. Table 1 (in the Appendix) contains the numerical results. As shown in the first two boxplots in Figures 1 and 2, the implementations (1) Qcgc and (2) Qc performed very well. We obtained unbiased estimates and a coverage of the confidence interval that tended to be around 95% as sample size increased as shown in Figure 5. The FCR was close to the optimal 0.05. In the third boxplot, corresponding to implementation (3) gc, where only the propensity score was correctly specified, the estimator was more biased for both sample sizes but had higher coverage rates and lower FCR. In the fourth

boxplot where the estimator was implemented with HAL, the estimator performed well across all measures. Overall, as shown in Figure 5, the percent coverage of the true EMs ($V^{(1)}$,$V^{3)}$) was around the nominal 95% as sample size increased or when at least the outcome model was correclty specified or machine learning methods were used to estimate both nuisance parameters. In all implementations the true effect-modifiers ($V^{(1)}$,$V^{(3)}$) were selected around 100 percent of the time except when only the propensity score was correctly specified for the smaller sample size (gc). The percent selection of variables that are not effect-modifiers ($V^{(2)}$,$V^{(4)}$) was around 20% for $n = 1000$. In implementations (1), (2), and (4), the percentage was almost halved for $n = 10000$. The FCR was controlled around the nominal 0.05 level in all situations even when only one nuisance model was correctly specified. This supports the double robustness of the proposed estimator and the appropriateness of the post-selection confidence intervals. In implementation (5) NLin, the naive linear model with a misspecified term performed poorly, even when increasing the sample size. On the other hand, when the linear model was correctly specified in implementation (6) CLin, the coefficient estimates were unbiased on average and the coverage was near-optimal. For the two other data generating scenarios described at more length in the Appendix, the results (Tables 2 and 3) look similar to those in the first scenario.

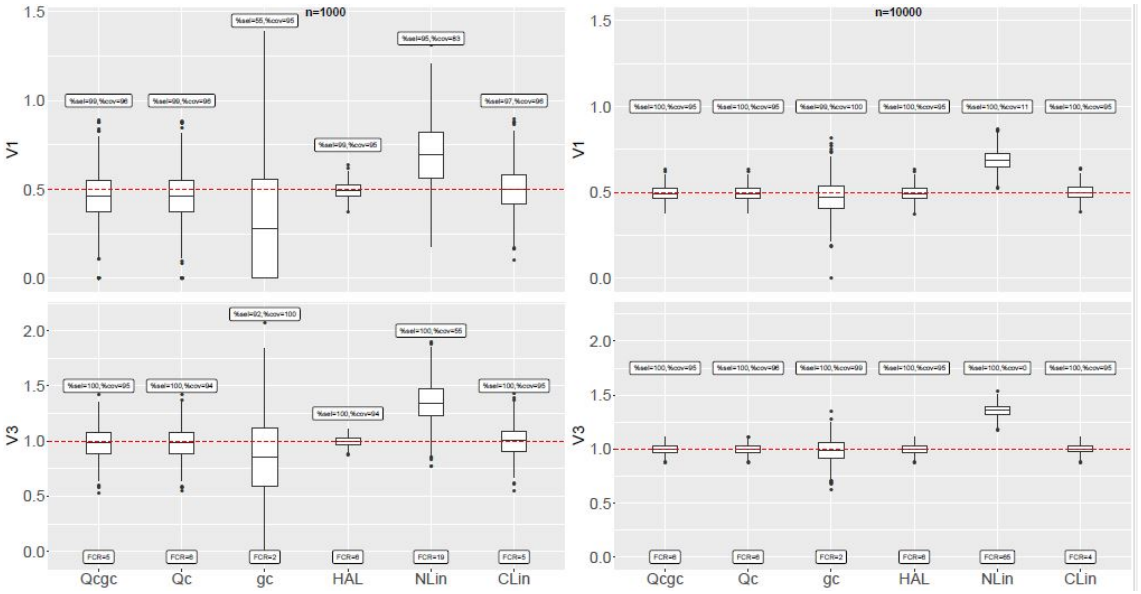

Fig. 1. Simulation results illustrations (Data generating scenario 1). Box plots of 1000 MSM coefficients estimates for the true EMs ($V^{(1)}$,$V^{3)}$). The true values of the coefficients are $(0.5, 1)$. Notation: Qcgc: models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL, NLin: Naive linear model, CLin: Correct linear model. $\%sel$: percent selection of a covariate $\times 100$, $\%cov$: coverage rate of the confidence interval of a coefficient estimate $\times 100$, FCR: False coverage rate of the model $\times 100$.

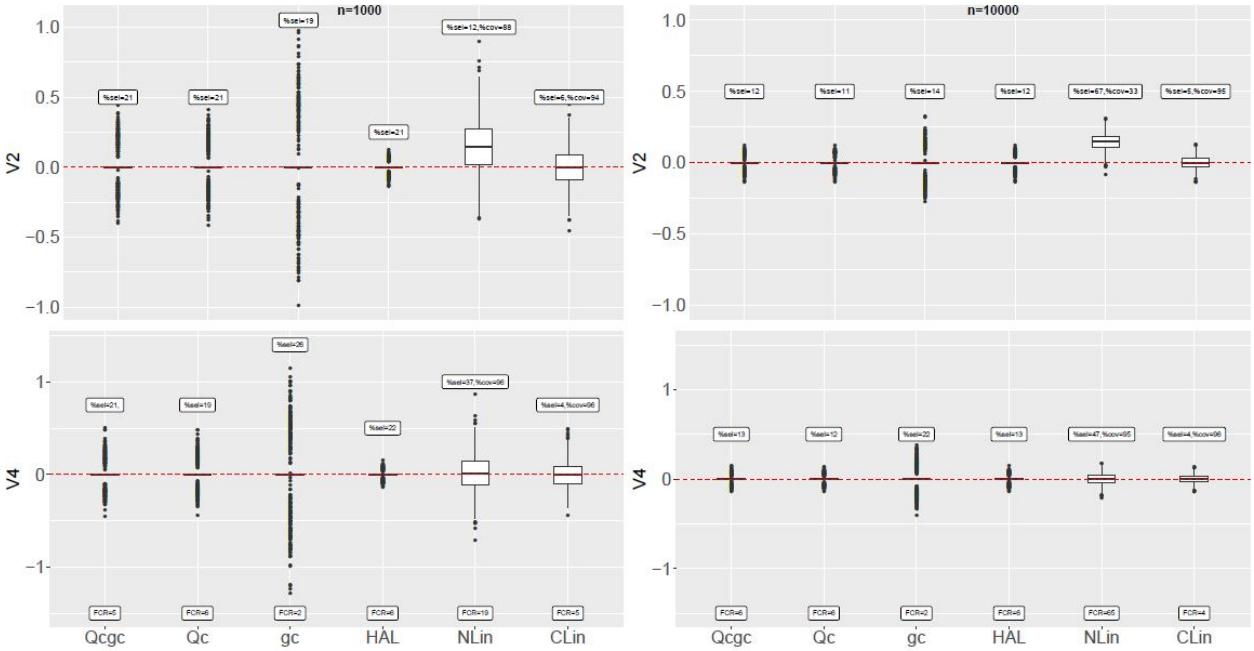

Fig. 2. Simulation results illustrations (Data generating scenario 1). Box plots of 1000 MSM coefficients estimates for the non-EMs ($V^{(2)}$,$V^{4)}$). The true values of the coefficients are $(0,0)$. Notation: Qcgc: models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL, NLin: Naive linear model, CLin: Correct linear model. $\%sel$: percent selection of a covariate $\times 100$, $\%cov$: coverage rate of the confidence interval of a coefficient estimate $\times 100$, FCR: False coverage rate of the model $\times 100$.

Table 4 in the Appendix contains the results with the small sample size $n = 100$. The performance of the proposed methods decreased across all measures except for $V^{(3)}$ where there was a higher coverage rate when $\bar{Q}$ and g were correctly specified or estimated with HAL

The results of the high-dimensional setting are presented in Figures 3 and 4. $\bar{Q}$ and g were estimated with HAL. The estimates were taken over 100 generated datasets and look similar to Figures 1 and 2 for the covariates $V = (V^{(1)},V^{(2)},V^{(3)},V^{(4)})$ in common. For the noise covariate coefficients, the estimates, given in the density plot of Figure 4, were unbiased for 0. The noise covariates had a low percent selection (see Table 5). Using median statistics, the noise covariates were selected around 14% of the time and that proportion decreased to 13% as we increased the sample size. The FCR exceeded the nominal 5% level and was around 15%.

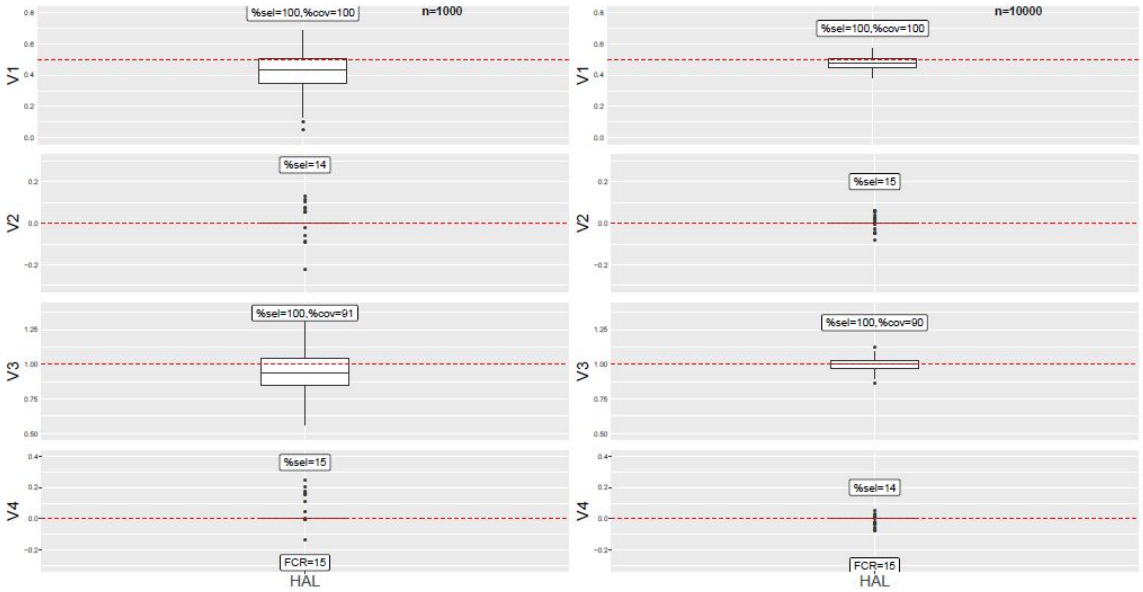

Fig. 3. Simulation results for high-dimensional setting (Data generating scenario 1). Box plots of MSM coefficients estimates over 100 simulations for the potentials EMs $\boldsymbol{V} = (V^{(1)},V^{(2)},V^{(3)},V^{(4)})$. The true values of the coefficients are $(0.5,0,1,0)$. Notations: HAL: $\bar{Q}$ and g are estimated with HAL, $\%sel$: percent selection $\times 100$, $\%cov$: coverage rate $\times 100$, FCR: False coverage rate $\times 100$.

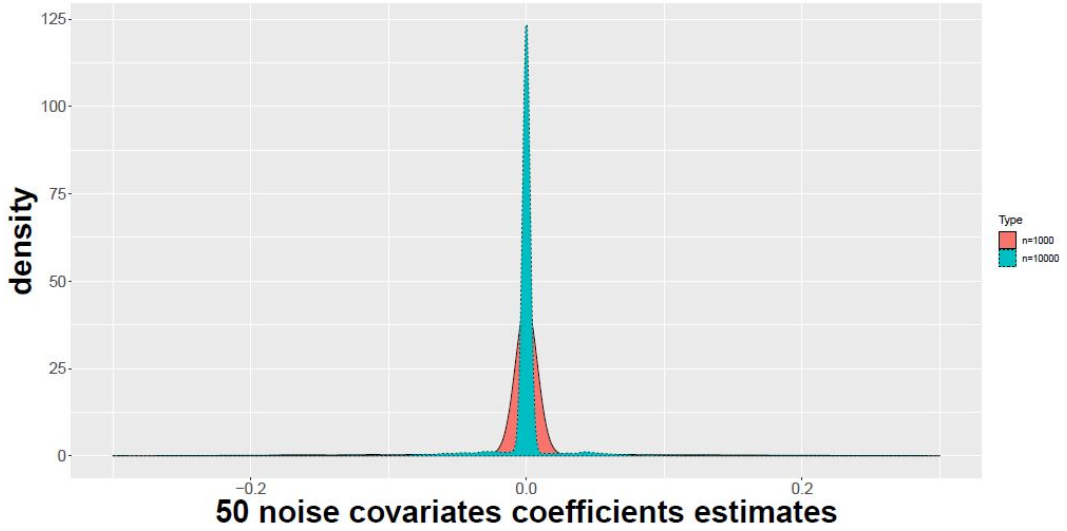

Fig. 4. Illustrations for high-dimensional setting. Box plots of the MSM coefficients estimates over 100 simulations for the 50 noise covariates (for both $n = 1000$ and $n = 10000$). The true values of the coefficients are (0,...,0).

In summary, Table 1 demonstrates that in low-dimensional settings, the proposed algorithm is able to produce unbiased estimates and control the FCR around the nominal level. In contrast, Table 5 demonstrates that in the context of high-dimensional covariates with

many candidate EMs, the FCR is generally much larger than the nominal level. Similar results were obtained by Zhao et al. ([**84**], Figure 2). In addition, at least some non-EMs were always selected by the algorithm at the sample sizes investigated.

5.4. Data analysis: Asthma medication during pregnancy

5.4.1. Data

Our data were obtained from a cohort (Firoozi et al. [**31**]) of deliveries of pregnant women with asthma in order to study the effect of using inhaled corticosteroids (ICS) during pregnancy on birth weight. The population of interest is pregnant women with mild asthma and a singleton delivery in Québec, Canada between 1998-2008, aged ≤ 45 years. For simplicity, We considered only the first delivery for each woman in this period. Asthma severity was defined according to an index that is based on the Canadian Asthma Consensus Guidelines (Cossette et al. [**8**]). A total of 4,707 pregnancies in our database fell into this category. ICS exposure was classified in two categories: "use"(a woman who filled at least one prescription of ICS during pregnancy) and "no use"(a woman who did not fill any prescription of ICS during pregnancy). The outcome of interest is birth weight (continuous in kilograms). We identified a variety of maternal baseline variables. These potential confounders measured in the year before pregnancy include demographic characteristics (e.g. income security provider and place of residence), chronic diseases (e.g. hypertension and diabetes) and variables related to asthma (e.g. at least one hospitalization for asthma, at least one emergency department visit for asthma, and oral corticosteroids). We also included the cumulative daily dose of ICS in the year before pregnancy and sex of the newborn as potential confounders. A full list of measured potential confounders can be found in Table 9 in the Appendix. As we do not know which variables are effect modifiers, we included a wide range of variables in the set $\boldsymbol{V}$, 22 variables in all. Specifically, these variables were: In the year before pregnancy: at least one dose of inhaled short-acting β_2-agonists (SABA)

taken per week, medication for epilepsy, use of warfarin, use of beta blockers, asthma exacerbation, oral SABA use, oral corticosteroids, leukoteriene-receptor antagonists, intranasal corticosteroids, at least one hospitalization for asthma, at least one emergency department visit for asthma, and welfare recipient; At the start of the pregnancy: chronic obstructive disease, cyanotic heart disease, obesity, uterine disorder, antiphospholipid syndrome, sex of the newborn, rural/non-rural residence indicator, hypertension, diabetes, and chromosomal anomalies.

For our pregnancy cohort, the average treatment effect is the expected difference in the mean counterfactual birth weight if all women were exposed to ICS during pregnancy versus the counterfactual birth weight if all women were not [**1**]. The target parameters are the coefficients β_j, $j = 1,...,22$ of the MSM defined as: $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \sum_{j=1}^{22} V^{(j)}\beta_j$, with $\boldsymbol{V} = (V^{(1)},..,V^{(22)})$ the set of potential EMs. Taking the sex of the newborn as an EM for example ($V^{(j)} = sex$), β_j is the difference in the CATE for women having male vs female children.

5.4.2. Results

Baseline characteristics of the pregnancy cohort are presented in Table 9. We first implemented a standard linear regression with main terms for all potential confounders and interaction terms between the treatment and the set $\boldsymbol{V}$. The estimates of the coefficients of the interaction terms are given in Table 7. A variable was considered to be selected as an EM in the standard linear regression if the coefficient of the interaction term between that variable and the treatment had a p-value < 0.05. This model concluded that leukoteriene-receptor antagonists and chromosomal anomalies are EMs. In addition, we implemented our LASSO methods using HAL for the estimation of the outcome expectation and propensity score. All of the covariates were included in the propensity score model as well as in the outcome model. Due to larger weights, a 5% truncation for the values of g_n was used. The selected coefficients of the MSM and their estimated values are presented in Table 8. Three covariates (leukoteriene-receptor antagonists, warfarin one year before pregnancy, and

chromosomal anomalies) were selected using the adaptive LASSO and two of them were significant (leukoteriene-receptor antagonists and chromosomal anomalies) using post-selection inference. Leukoteriene-receptor antagonists and chromosomal anomalies were thus selected as EMs in the association of taking ICS during pregnancy on birth weight. Although the naive linear model and our algorithm generate very similar sets of EMs, the coefficients of the selected EMs are different (compare Table 7 with Table 8). For example, the estimated coefficient of leukoteriene-receptor antagonist is around -0.17 in the adaptive LASSO while it is -0.365 using the linear model.

5.5. Discussion

In this paper, we proposed a doubly robust estimator for selecting effect modifiers (EMs) in an MSM for the CATE. We used the post selection inference method of Lee et al. [**42**] to produce post-selection confidence intervals.

Through simulation studies, we studied the performance of the proposed estimator. As well, we showed that our proposed estimator is doubly robust and performs well in a high dimensional setting but had a higher FCR along with an over-selection of non-EMs. We observed a slower convergence of our estimator when the outcome expectation model was misspecified. Work by Ju et al. [**11**] suggests that better performance might be obtained by incorporating outcome-inverse weighting in the penalty term when using HAL to estimate the propensity score. We also illustrated that the post-selection confidence interval produces good coverage proportions for the selected EMs. In a high dimensional case, we confirmed the observation of Zhao et al. [**84**] concerning the FCR which exceeded the nominal level in the presence of many noise covariates. Debiased Lasso [**3**] could be considered here in a high dimensional case as proposed in Zhao et al (2017). In general, the overall performance of our estimator improved with the sample size. However, the blind usage of traditional methods like a regression with main terms and interactions between treatment and potential effect modifiers may produce biased results.

We also show theoretically that our estimator is doubly robust and also inherits the oracle properties of the adaptive LASSO. Linearity and sparsity are assumptions of lemma 2 and they may be restrictive. However, by modeling the conditional average treatment effect on the linear scale, we are investigating effect modification on the absolute scale (difference between means) which is recommended [**112**]. If the linear model is too restrictive for some applications, we could increase the model capacity by adding higher order terms and interaction terms. Another option could be to use non-linear models or machine learning methods to estimate the pseudo-outcome $\tilde{\psi}_0(V)$. Because of the difficulty for stakeholders to interpret a black-box marginal model [**85**], this approach may not be desireable when the goal is to discover effect modifiers and fit an interpretable model. Machine learning approaches may be more appropriate when the goal is identifying optimal treatment rules.

In our application, the results suggest that leukoteriene-receptor antagonists and chromosomal anomalies may modify the effect of ICS during pregnancy on birth weight for women with mild asthma. The estimated CATE is 0.18 lower for women taking leukoteriene-receptor antagonists. As leukoteriene-receptor antagonists are an addition to ICS, we can suppose that it is a marker for more severe asthma. In the presence of a chromosomal anomaly, the effect of ICS was estimated to be 0.78 lower. The linear regression with standard significance testing suggested the same but with different coefficient estimates. Such discrepancy may be due to the fact that the naive model doesn't target MSM parameters and thus may not be able to model effect modification in the absence of confounding. In this finite sample setting, the regularization may possibly have shrunk the coefficient values relative to the truth. Our results point to the importance of using robust methodologies for selecting effect modifiers in well-defined causal models for estimating the conditional treatment effect.

5.6. Appendix

In the Appendix, we give the numerical results of the simulation study, the baseline characteristics of our pregnancy data, the results of our application and the proof of the two lemmas.

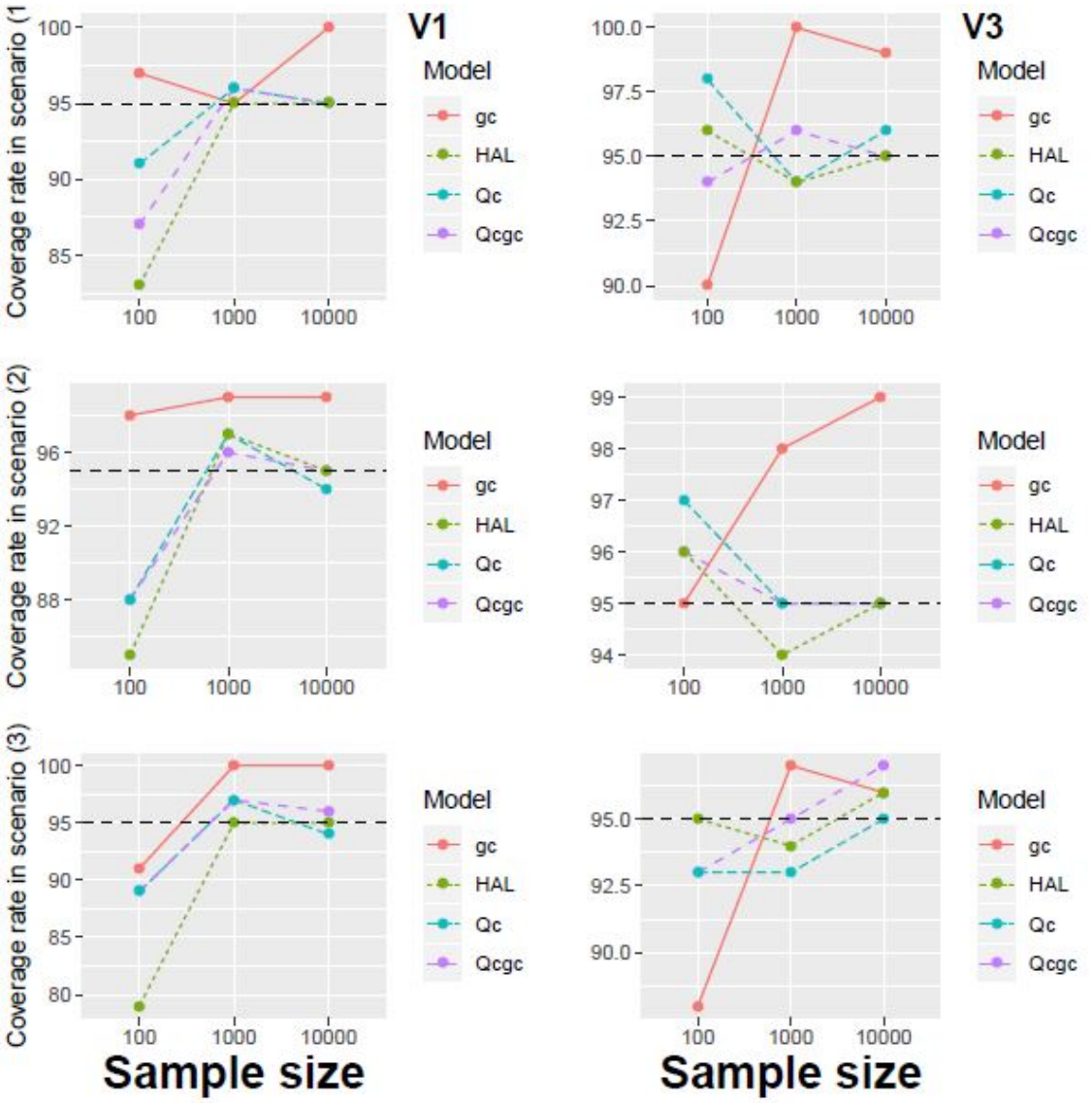

Fig. 5. Percent coverage of the selective confidence interval associated to V_1 and V_3 for different sample size. Notation: Qcgc: models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL..

Table 1. Simulation results (Data generating scenario 1). Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: average estimated value of the coefficients of the MSM, $\%Cov$: percent coverage of the selective confidence interval $\times 100$ (Standard CI for the linear model case), $\%sel$: percent selection of variables $\times 100$, FCR: False coverage rate $\times 100$, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$

		n=1000				n=10000			
Coef	EM	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR
					(1) $\bar{Q}$ & g model are correctly specified				
$V^{(1)}$	T	0.46	98	96		0.49	100	95	
$V^{(2)}$	F	0.00	21		5	0.00	12		6
$V^{(3)}$	T	0.98	100	95		0.99	100	95	
$V^{(4)}$	F	0.00	21			0.00	13		
					(2) $\bar{Q}$ model is correctly specified				
$V^{(1)}$	T	0.46	99	96		0.49	100	95	
$V^{(2)}$	F	0.00	21		6	0.00	11		6
$V^{(3)}$	T	0.98	100	94		0.99	100	96	
$V^{(4)}$	F	0.00	19			0.00	12		
					(3) g model is correctly specified				
$V^{(1)}$	T	0.31	55	95		0.47	99	100	
$V^{(2)}$	F	0.01	19		2	0.00	14		2
$V^{(3)}$	T	0.83	92	100		0.99	100	99	
$V^{(4)}$	F	0.00	26			0.00	22		
					(4) $\bar{Q}$ & g model are estimated using HAL				
$V^{(1)}$	T	0.46	99	95		0.49	100	95	
$V^{(2)}$	F	0.00	21		6	0.00	12		6
$V^{(3)}$	T	0.98	100	94		1.00	100	95	
$V^{(4)}$	F	0.00	22			0.00	13		
					(5) Naive Linear model				
$V^{(1)}$	T	0.69	95	83		0.69	100	11	
$V^{(2)}$	F	0.15	12	88	19	0.15	67	33	65
$V^{(3)}$	T	1.35	100	56		1.36	100	0	
$V^{(4)}$	F	0.01	37	96		0.00	47	95	
					(6) Linear model correctly specified				
$V^{(1)}$	T	0.50	97	96		0.50	100	95	
$V^{(2)}$	F	0.00	6	94	5	0.00	5	95	4
$V^{(3)}$	T	1.00	100	95		1.00	100	95	
$V^{(4)}$	F	0.00	4	96		0.00	4	96	

Table 2. Simulation results (Data generating scenario 2). Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval $\times 100$, %sel: percent selection of variables $\times 100$, FCR: False coverage rate $\times 100$, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$

		n=1000				n=10000			
Coef	EM	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR
					(1) Q & g model are correctly specified				
$V^{(1)}$	T	0.47	99	96		0.49	100	95	
$V^{(2)}$	F	0.00	20		5	0.00	13		5
$V^{(3)}$	T	0.98	100	95		1.00	100	95	
$V^{(4)}$	F	0.00	23			0.00	12		
					(2) Q model is correctly specified				
$V^{(1)}$	T	0.47	99	97		0.49	100	94	
$V^{(2)}$	F	0.00	20		5	0.00	11		6
$V^{(3)}$	T	0.99	100	95		1.00	100	95	
$V^{(4)}$	F	0.00	21.			0.00	11		
					(3) g model is correctly specified				
$V^{(1)}$	T	0.32	55	99		0.47	99	99	
$V^{(2)}$	F	0.01	19		2	0.00	14		2
$V^{(3)}$	T	0.85	94	98		0.99	100	99	
$V^{(4)}$	F	-0.01	24			0.00	21		
					(4) Q & g model are estimated using HAL				
$V^{(1)}$	T	0.47	98	97		0.49	100	95	
$V^{(2)}$	F	0.00	22		5	0.00	12		7
$V^{(3)}$	T	0.98	100	94		1.00	100	95	
$V^{(4)}$	F	0.00	22			0.00	12		
					(6) Linear model correctly specified				
$V^{(1)}$	T	0.50	89	96		0.50	100	95	
$V^{(2)}$	F	0.00	6	94	5	0.00	6	94	5
$V^{(3)}$	T	1.00	100	94		1.00	100	95	
$V^{(4)}$	F	0.00	4	97		0.00	4	96	

Table 3. Simulation results (Data generating scenario 3). Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval $\times 100$, %sel: percent selection of variables $\times 100$, FCR: False coverage rate $\times 100$, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$

		n=1000				n=10000			
Coef	EM	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR
					(1) Q & g **model are correctly specified**				
$V^{(1)}$	T	0.44	94	97		0.49	100	96	
$V^{(2)}$	F	0.00	23		5	0.00	16		5
$V^{(3)}$	T	0.97	100	95		1.00	100	97	
$V^{(4)}$	F	0.00	23			0.00	17		
					(2) Q **model is correctly specified**				
$V^{(1)}$	T	0.45	96	97		0.50	100	94	
$V^{(2)}$	F	0.00	20		6	0.00	13		7
$V^{(3)}$	T	0.98	100	93		1.00	100	95	
$V^{(4)}$	F	0.00	22			0.00	12		
					(3) g **model is correctly specified**				
$V^{(1)}$	T	0.34	74	100		0.49	100	100	
$V^{(2)}$	F	0.01	23		3	0.00	18		4
$V^{(3)}$	T	0.91	99	97		0.99	100	96	
$V^{(4)}$	F	0.00	25			0.00	24		
					(4) Q & g **model are estimated using HAL**				
$V^{(1)}$	T	0.45	95	95		0.49	100	95	
$V^{(2)}$	F	0.00	24		6	0.00	16		5
$V^{(3)}$	T	0.98	100	94		1.00	100	96	
$V^{(4)}$	F	0.00	23			0.00	16		
					(5) **Naive Linear model**				
$V^{(1)}$	T	0.60	89	93		0.59	100	63	
$V^{(2)}$	F	0.10	76	92	10	0.10	35	65	43
$V^{(3)}$	T	1.21	100	81		1.21	100	58	
$V^{(4)}$	F	0.01	38	96		-0.00	44	96	
					(6) **Linear model correctly specified**				
$V^{(1)}$	T	0.50	98	96		0.50	100	95	
$V^{(2)}$	F	0.00	4	96	5	0.00	5	95	5
$V^{(3)}$	T	1.00	100	95		1.00	100	95	
$V^{(4)}$	F	0.00	5	95		0.00	5	95	

Table 4. Simulation results for smaller sample size ($n = 100$). Estimates taken over 500 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval $\times 100$, %sel: percent selection of variables $\times 100$, FCR: False coverage rate $\times 100$, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$

		scenario 1				scenario 2				scenario 3			
Coef	EM	$\widehat{\beta}_V$	$\%sel$	Cov	FCR	$\widehat{\beta}_V$	$\%sel$	Cov	FCR	$\widehat{\beta}_V$	$\%sel$	Cov	FCR
						(1) Q & g model are correctly specified							
$V^{(1)}$	T	0.39	52	87		0.34	49	88		0.30	41	89	
$V^{(2)}$	F	-0.01	22		8	-0.01	25		9	0.02	24		10
$V^{(3)}$	T	0.85	86	94		0.78	80	96		0.78	71	93	
$V^{(4)}$	F	0.01	28.			0.00	25			0.00	24		
						(2) Q model is correctly specified							
$V^{(1)}$	T	0.38	53	91		0.36	50	88		0.29	41	89	
$V^{(2)}$	F	-0.03	27		7	0.00	21		8	0.01	20		10
$V^{(3)}$	T	0.83	85	98		0.79	8	97		0.76	72	93	
$V^{(4)}$	F	-0.02	25			0.00	27			0.00	21		
						(3) g model is correctly specified							
$V^{(1)}$	T	0.24	20	97		0.24	25	98		0.26	25	91	
$V^{(2)}$	F	0.04	16		9	0.04	1		6	0.04	26		9
$V^{(3)}$	T	0.51	29	90		0.59	45	95		0.68	47	88	
$V^{(4)}$	F	0.01	21			0.02	23			0.00	25		
						(4) Q & g model are estimated using HAL							
$V^{(1)}$	T	0.39	54	83		0.36	51	85		0.32	45	79	
$V^{(2)}$	F	0.00	30		10	0.01	27		9	0.00	27		11
$V^{(3)}$	T	0.84	87	96		0.79	81	96		0.80	82	95	
$V^{(4)}$	F	0.00	27			0.01	27			-0.02	24		

Table 5. Simulation results (Data generating scenario 1 with 50 noise covariates). Estimates taken over 100 generated datasets. $\widehat{\boldsymbol{\beta}}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval, %sel: percent selection of variables, FCR: False coverage rate, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0, ..., 0)$

		n=1000				n=10000			
Coef	EM	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR	$\widehat{\beta}_V$	$\%sel$	$\%Cov$	FCR
(1) **Estimates related to the potential EM that are not noise covariates.**									
$V^{(1)}$	T	0.43	100	100		0.48	100	100	
$V^{(2)}$	F	0.00	14		15	0.00	15		15
$V^{(3)}$	T	0.95	100	91		0.99	100	90	
$V^{(4)}$	F	0.01	15			0.00	14		
(2) **Summary of the 50 potential EM that are noise covariates.**									
min		-0.01	7.0			0.00	5		
Q_1		0.00	12			0.00	11		
median		0.00	14			0.00	13		
Q_3		0.00	16			0.00	15		
max		0.01	23			0.00	22		

Table 6. Baseline Characteristics of mothers in the cohort extraction ($N = 4{,}707$) .

Characteristics	**No ICS** **N (%)**	**ICS** **N (%)**
Cohort size	2272 (100)	2435 (100)
Age		
< 18	45 (1.9)	60 (2.4)
18-34	1958 (86.1)	2041 (83.8)
> 34	269 (11.8)	334(13.7)
Sex of the newborn	1149 (51.0)	1271 (52.0)
Welfare recipient	1126 (50.0)	1429 (59.0)
Urban residence	476 (18.0)	407 (20.0)
Hypertension	61 (3.0)	83 (3.0)
Diabetes	73 (3.0)	81 (3.0)
COPD	28 (1.0)	56 (2.0)
Cyanotic heart disease	7 (0.0)	8 (0.0)
Antiphospholipid syndrome	12 (1.0)	13 (1.0)
Uterine disorder	264 (12.0)	331 (14.0)
Epilepsy	18 (1.0)	23 (1.0)
Obesity	87 (4.0)	127 (5.0)
Lupus	1 (0.0)	2 (0.0)
Collagenous vascular disease	6 (0.0)	6 (0.0)
Cushing's syndrome	4 (0.0)	4 (0.0)
Oral corticosteroids one year before pregnancy	234 (10.0)	281(12.0)
Oral SABA use one year before pregnancy	16 (1.0)	8 (0.0)
At least one dose of inhaled SABA taken per week	1523 (67.0)	1332 (55.0)
HIV	3 (0.0)	1 (0.0)
Cytomegalovirus infection	3 (0.0)	12 (0.0)
Leukoteriene-receptor antagonists	33 (1.0)	30 (1.0)
Theophylline use one year before pregnancy	0 (0.0)	0 (0.0)
Intranasal corticosteroids	243 (11.0)	318 (13.0)
Folic acid one year before pregnancy	18 (1.0)	43 (2.0)
Teratogenes taken one year before	0 (0.0)	0 (0.0)
Medication for epilepsy one year before pregnancy	29 (1.0)	48 (2.0)
Warfarin one year before pregnancy	7(0.0)	10 (0.0)
Use of beta-bloqueur one year before pregnancy	19 (1.0)	26 (1.0)
Asthma exacerbation one year before pregnancy	377 (17.0)	411 (17.0)
hospitalization for asthma	1079 (47.0)	809 (33.0)
Chromosomal anomalies	6 (0.0)	4 (0.0)
Cumulative dose of ICS in days (mean (SD))	51.6 (72.8)	54.0 (85.8)
One year cumulative dose of ICS before pregnancy (mean (SD))	151 (32.0)	101.5 (126.3)
At least one emergency department visit for asthma	260 (7.0)	265 (19.0)
At least one hospitalization for asthma	5 (0.0)	8 (1.0)

Table 7. Estimates of the coefficients associated with interaction terms using naive linear model ($n = 4707$).

Variables	Estimate ($\widehat{\beta}_j$)	STD	P-value
Intercept	3.153		
CS:At least one dose of inhaled SABA taken per week	-0.002	0.039	0.940
CS:Leukoteriene-receptor antagonists	-0.365	0.142	0.010*
CS:Intranasal corticosteroids	0.063	0.051	0.214
CS:Folic acid one year before pregnancy	-0.129	0.159	0.415
CS:Medication for epilepsie	-0.136	0.135	0.313
CS:Warfarin	-0.386	0.277	0.164
CS:Beta-blockers	-0.287	0.173	0.097
CS:Asthma exacerbation	0.062	0.069	0.368
CS:At least one hospitalization for asthma	0.017	0,036	0.624
CS:At least one emergency department visit for asthma	0.067	0.055	0.223
CS:COPD	0.141	0.130	0.280
CS:Cyanotic heart disease	-0.345	0.292	0.237
CS:Oral corticosteroids one year before	-0.081	0.081	0.319
CS:Obesity	0.053	0.080	0.508
CS:Uterine disorder	-0.036	0.050	0.460
CS:Oral SABA use one year before	-0.025	0.244	0.918
CS:Antiphospholipid syndrome	0.394	0.227	0.083
CS:Sex of new born	-0.031	0.032	0.335
CS:Welfare recipient	-0.043	0.033	0.1871
CS:Rural/non-rural residence indicator	0.021	0.042	0.602
CS:Hypertension	0.028	0.098	0.774
CS:Diabetes	-0.105	0.092	0.255
CS:Chromosomal anomalies	-1.230	0.361	0.0006*
CS:Cytomegalovirus infection	0.146	0.360	0.683

Table 8. Estimates of the selected MSM coefficients using adaptive lasso ($n = 4707$) with 95% Post selection interval for the selected variables. *: means significant variables

Variables	Estimate ($\widehat{\beta}_j$)	CI Low	CI up
High adaptive LASSO for Q & g			
Intercept	0.018		
Leukoteriene-receptor antagonists*	-0.177	-0.502	-0.031
Warfarin	-0.146	-0.745	0.311
Chromosomal anomalies*	-0.777	-1.420	-0.285

Proof of Lemma 1: Denote $\bar{Q}_n$ (respectively g_n) an estimator of $\bar{Q}$ (respectively g). We have:

$$E_{P_0}(D(\bar{Q}_n,g_n)|\boldsymbol{V}) = E_{P_0}\left\{\frac{2A-1}{g_n(A|\boldsymbol{W})}(Y-\bar{Q}_n(A,\boldsymbol{W}))+\bar{Q}_n(1,\boldsymbol{W})-\bar{Q}_n(0,\boldsymbol{W})|\boldsymbol{V}\right\}$$

$$= E_{P_0}\left\{\frac{2A-1}{g_n(A|\boldsymbol{W})}(Y-\bar{Q}_n(A,\boldsymbol{W})|\boldsymbol{V}\right\}+E_{P_0}(\bar{Q}_n(1,\boldsymbol{W})-\bar{Q}_n(0,\boldsymbol{W})|\boldsymbol{V})+\psi_0(\boldsymbol{V})-\psi_0(\boldsymbol{V})$$

$$= \psi_0(\boldsymbol{V})+E_{P_0}\left\{[\bar{Q}_n(1,\boldsymbol{W})-\bar{Q}_n(0,\boldsymbol{W})]-[\bar{Q}_0(1,\boldsymbol{W})-\bar{Q}_0(0,\boldsymbol{W})]|\boldsymbol{V}\right\}+$$

$$E_{P_0}\left\{\frac{2A-1}{g_n(A|\boldsymbol{W})}(Y-\bar{Q}_n(A,\boldsymbol{W}))|\boldsymbol{V}\right\}$$

$$= \psi_0(\boldsymbol{V})+\int_{\boldsymbol{W}}([\bar{Q}_n(1,\boldsymbol{W})-\bar{Q}_n(0,\boldsymbol{W})]-[\bar{Q}_0(1,\boldsymbol{W})-\bar{Q}_0(0,\boldsymbol{W})]+$$

$$\frac{P_0(1|\boldsymbol{W})}{g_n(1|\boldsymbol{W})}\left\{\bar{Q}_0(1,\boldsymbol{W})-\bar{Q}_n(1,\boldsymbol{W})\right\}-\frac{P_0(0|\boldsymbol{W})}{g_n(0|\boldsymbol{W})}\left\{\bar{Q}_0(0,\boldsymbol{W})-\bar{Q}_n(0,\boldsymbol{W})\right\})dP_0(\boldsymbol{W}|\boldsymbol{V})$$

$$= \psi_0(\boldsymbol{V})+\int_{\boldsymbol{W}}[\frac{P_0(1|\boldsymbol{W})}{g_n(1|\boldsymbol{W})}-1)(\bar{Q}_0(1,\boldsymbol{W})-\bar{Q}_n(1,\boldsymbol{W}))+$$

$$(\frac{P_0(0|\boldsymbol{W})}{g_n(0|\boldsymbol{W})}-1)(\bar{Q}_0(0,\boldsymbol{W})-\bar{Q}_n(0,\boldsymbol{W}))]dP_0(\boldsymbol{W}|\boldsymbol{V})$$

Then $E_{P_0}(D(\bar{Q}_n,g_n)|\boldsymbol{V})\rightarrow\psi_0(\boldsymbol{V})$ if $g_n(A|\boldsymbol{W})$ or $\bar{Q}_n(A,\boldsymbol{W})$ is consistently estimated.

Proof of Lemma 2:

Let $D_n = D(\bar{Q}_n,g_n)$ (respectively $D_0 = D(\bar{Q}_0,g_0)$) represent the estimated pseudo function (respectively the true pseudo-outcome). Our method minimizes the expected risk function below with respect to β:

$$\{(D_n-\sum_j V^{(j)}\beta_j)^2+\lambda\sum_{j=1}^{p}\widehat{w}_j|\beta_j|\}$$

where $\widehat{\omega}_j = 1/|\tilde{\beta}_j|^\gamma$, $j=1,...,p$, for some $\gamma>0$.

Let $\epsilon_n = D_n-\sum_j V^{(j)}\beta_j$ be the residual of the penalized linear regression of D_n on $\boldsymbol{V}$. The proof follows essentially the one of Zou ([**35**]). We have to show that $\epsilon_n^T V/\sqrt{n}$ follows a normal distribution with mean zero and a finite variance.

Indeed, one can write

$$\epsilon_n = (D_n - D_0) + (D_0 - \sum_j V^{(j)}\beta_j).$$

$$\epsilon_n^T V/\sqrt{n} = \underbrace{\sqrt{n}\mathbb{P}_n(D_n - D_0)^T V}_{R_1} + \underbrace{\sqrt{n}\mathbb{P}_n(D_0 - \sum_j V^{(j)}\beta_j)^T V}_{R_2}.$$

with $\mathbb{P}_n$ denotes the empirical measure. $\epsilon_0 = D_0 - \sum_j V^{(j)}\beta_j$ is the residual of the penalized linear regression of the oracle pseudo function D_0 on $\boldsymbol{V}$. Therefore, if we assume $\frac{1}{n}\boldsymbol{V}^T\boldsymbol{V} \to C$ with C a positive definite matrix, we have $R_2 \xrightarrow{d} N(0,\sigma^2 C)$.

One can write

$$\sqrt{n}\mathbb{P}_n(D_n - D_0)^T V \leq \sqrt{n}||\mathbb{P}_n(D_n - D_0)^T V||$$

Semenova and Chernozhukov ([**114**]) showed in Lemma A.3, given their Assumption 3.5, is that

$$\sqrt{n}||\mathbb{P}_n(D_n - D_0)^T V|| = o(1)$$

Therefore, $\sqrt{n}\mathbb{P}_n(D_n - D_0)^T V = o(1)$ which yields the result.

5.7. Addenda

Dans cette section, nous présentons une étude supplémentaire qui ne figure pas dans l'article soumis au journal. Dans cette étude, nous considérons un scénario dans lequel le MSM n'est pas une fonction linéaire des modificateurs d'effet. Les données sont générées de la manière suivante: $X \sim B(p = 0.4)$, $V^{(1)} \sim Unif(0.5,2)$, $V^{(2)} \sim B(p = 0.6)$, $V^{(3)} \sim Unif(-2,2)$, $V^{(4)} \sim B(p = 0.7)$ and $Z \sim B(p = 0.45)$.

Le traitement est généré avec le modèle suivant:

$$P_0(A = 1|X) = \text{expit}\{0.5Z - 0.2X + 0.3V^{(1)} + 0.4V^{(2)}\}$$

avec expit$(x) = 1/\{1+exp(-x)\}$. Soit $Z^{(1)} = exp(V^{(1)}/2)$ et $Z^{(3)} = exp(V^{(3)})$. La variable réponse est générée via le modèle suivant:

$$Y = 1 + A - 0.5X + 2Z^{(1)} + V^{(2)} + Z^{(3)} - 0.2V^{(4)} + A(0.5Z^{(1)} + Z^{(3)}) + N(0,1)$$

Bien que le vrai MSM est fonction des variables $Z^{(1)}$ et $Z^{(3)}$, l'analyste observe les variables suivantes (Y,A,X,Z,$V^{(1)}$,$V^{(2)}$,$V^{(3)}$,$V^{(4)}$). Les résultats de notre procédure sont présentés dans le tableau suivant. Nous considérons uniquement la capacité de notre procédure à sélectionner les bons modificateurs d'effet. En effet, il en résulte qu'asymptotiquement, notre algorithme sélectionne l'ensemble de variables ($V^{(1)}$,$V^{(3)}$) dans la majorité du temps quand l'un ou l'autre des deux paramètres de nuisance sont correctement estimés.

Table 9. Résultat avec un MSM non linéaire. GAM (Generalized Additive Model/Modèle additif généralisé)

	$n = 1000$	$n = 10000$
Méthodes	$\%sel$	$\%sel$
	Q et g sont correctement spécifés	
$V^{(1)}$	88.7	100
$V^{(2)}$	23.0	3.10
$V^{(3)}$	100	100
$V^{(4)}$	20.9	3.20
	GAM pour Q et g	
$V^{(1)}$	86.6	100
$V^{(2)}$	23.3	3.10
$V^{(3)}$	100	100
$V^{(4)}$	23.8	3.90
	GAM pour Q; g mal spécifé	
$V^{(1)}$	87.5	100
$V^{(2)}$	22.2	3.2
$V^{(3)}$	100	100
$V^{(4)}$	22.3	3.40
	GAM pour g; Q mal spécifé	
$V^{(1)}$	42.4	94.6
$V^{(2)}$	20.8	19.1
$V^{(3)}$	100	100
$V^{(4)}$	22.1	21.7

Chapitre 6

Data Integration through outcome adaptive LASSO and a collaborative propensity score approach

Cet article est en révision pour une soumission dans le Journal of Survey Statistics and Methodology (JSSM).

Préambule: Depuis quelque temps, les données administratives sont de plus en plus présentes en pratique pour estimer des paramètres d'intérêt au lieu de procéder à des enquêtes traditionnelles. Toutefois, ces données sont souvent sujettes à un biais de sélection. L'étude décrite dans cet article est une extension de la proposition développée par Chen, Li & Wu [**121**] pour estimer un paramètre d'intérêt marginal avec des données administratives sujettes à un biais de sélection. En effet, le développement de Chen, Li & Wu [**121**] assume une connaissance des co-variables à ajuster dans le modèle du score de propension. Ainsi, l'objectif principal de cet article est de développer une procédure de sélection de variables dans ce contexte pour le score de propension. Pour ce faire, on intègre la procédure du LASSO adaptatif basé sur l'issue de Shortreed & Ertefaie (OALASSO; [**104**]). La méthode est évaluée via des études de simulation dans le but d'évaluer sa performance. Finalement, elle est appliquée sur une base de données mesurant l'impact de la COVID-19 sur les

Canadiens.

Data Integration through outcome adaptive LASSO and a collaborative propensity score approach

Asma Bahamyirou, Mireille E. Schnitzer.

Faculté de pharmacie, Université de Montréal.

Résumé: Les sources de données administratives ou les échantillons non probabilistes sont de plus en plus considérés en pratique pour obtenir des statistiques officielles, vu le gain qu'on en tire (coût moindre, grande taille d'échantillon, etc.) et le déclin des taux de réponse. Toutefois, il est difficile d'obtenir des estimations sans biais provenant de ces bases de données à cause du poids d'échantillonnage manquant. Des méthodes d'estimations ont été proposées récemment qui utilisent l'information auxiliaire provenant d'un échantillon probabiliste représentatif de la population cible. En présence de bases de données de grande dimension, il est difficile d'identifier les variables auxiliaires qui sont associées au mécanisme de sélection. Dans ce travail, nous développons une procédure de sélection de variables en utilisant le LASSO adaptatif et un score de propension collaboratif. Des études de simulations ont été effectuées dans le but de comparer les différentes approches de sélection de variables. Pour terminer, nous avons présenté une application sur l'impact de la COVID-19 sur les Canadiens.

Keys words: Échantillon non-probababiliste, Échantillon probabiliste, LASSO adaptatif, Estimateurs par l'inverse du score de propensionr.

Abstract: Administrative data, or non-probability sample data, are increasingly being used to obtain official statistics due to their many benefits over survey methods. In particular, they are less costly, provide a larger sample size, and are not reliant on the response rate. However, it is difficult to obtain an unbiased estimate of the population mean from such data due to the absence of design weights. Several estimation approaches have been proposed recently using an auxiliary probability sample which provides representative covariate information of the target population. However, when this covariate information is high-dimensional, variable selection is not a straight-forward task even for a subject matter

expert. In the context of efficient and doubly robust estimation approaches for estimating a population mean, we develop two data adaptive methods for variable selection using the outcome adaptive LASSO and a collaborative propensity score, respectively. Simulation studies are performed in order to verify the performance of the proposed methods versus competing methods. Finally, we presented an analysis of the impact of COVID-19 on Canadians.
Keys words: Nonprobability sample, Probability sample, Outcome adaptive LASSO, Inverse weighted estimators.

6.1. Introduction

Administrative data, or non-probability sample data, are being increasingly used in practice to obtain official statistics due to their many benefits over survey methods (lower cost, larger sample size, not reliant on response rate). However, it is difficult to obtain unbiased estimates of population parameters from such data due to the absence of design weights. For example, the sample mean of an outcome in a non-probability sample would not necessarily represent the population mean of the outcome. Several approaches have been proposed recently using an auxiliary probability sample which provides representative covariate information of the target population. For example, one can estimate the mean outcome in the probability sample by using an outcome regression based approach. Unfortunately, this approach relies on the correct specification of a parametric outcome model. Valliant & Dever [**90**] used inverse probability weighting to adjust a volunteer web survey to make it representative of a larger population. Elliott & Valliant [**79**] proposed an approach to model the indicator representing inclusion in the non-probability sample by adapting Bayes' rule. Rafei et al. [**4**] extended the Bayes' rule approach using Bayesian Additive Regression Trees (BART). Chen [**60**] proposed to calibrate non-probability samples using probability samples with the LASSO. Beaumont & Chu [**45**] proposed a tree-based approach for estimating the propensity score (probability that a unit belongs to the non-probability sample) in the same context. Wisniowski et al. [**6**] developed a Bayesian approach for integrating probability and nonprobability samples for the same goal.

Doubly robust semiparametric methods such as the augmented inverse propensity weighted (AIPW) estimator ([**51**]) and targeted minimum loss-based estimation (TMLE; [**73, 76**]) have been proposed to reduce the bias in the regression based approach. The term doubly robust comes from the fact that these methods require both the estimation of the propensity score model and the outcome expectation conditional on covariates, where only one of which needs to be correctly modeled to allow for consistent estimation of the parameter of interest. Chen, Li & Wu [**121**] developed doubly robust inference with non-probability surveys samples by adapting the Newton-Raphson procedure in this setting. Reviews and discussions of related approaches can be found in Beaumont [**44**] and Rao [**55**]. Chen, Li & Wu [**121**] considered the situation where the auxiliary variables are given, i.e. where the set of variables to include in the propensity score model is known. However, in practice or in high-dimensional data, variable selection for the propensity score may be required and it is not a straight-forward task even for a subject matter expert. In order to have unbiased estimation of the population mean, controlling for the variables that influence the selection into the non-probability sample and are also causes of the outcome is important (VanderWeele & Shpitser, [**110**]). Studies have shown that including instrumental variables – those that affect the selection into the non-probability sample but not the outcome – in the propensity score model leads to inflation of the variance of the estimator relative to estimators that exclude such variables (Schisterman et al.,[**29**]; Schneeweiss et al., [**98**]; van der Laan & Gruber, [**93**]). However, including variables that are only related to the outcome in the propensity score model will increase the precision of the estimators without affecting bias (Brookhart et al. [**2**]; Shortreed & Ertefaie, [**104**]). Using the Chen, Li & Wu [**121**] estimator for doubly robust inference with a non-probability sample, Yang, Kim & Song [**101**] proposed a two step approach for variable selection for the propensity score using the smoothly clipped absolute deviation (SCAD; Fan & Li, [**38**]). Briefly, they used SCAD to select variables for the outcome model and the propensity score model separately. Then, the union of the two sets is taken to obtain the final set of the selected variables. To the best of our knowledge, their paper is the first to investigate a variable selection method

in this context.

In causal inference, multiple variable selection methods have been proposed for the propensity score model. We consider two in particular. Shortreed & Ertefaie (OALASSO; [**104**]) developed the outcome adaptive LASSO. This approach uses the adaptive LASSO (Zou; [**35**]) but with weights in the penalty term that are the inverse of the estimated covariate coefficients from a regression of the outcome on the treatment and the covariates. Benkeser, Cai & van der Laan [**16**] proposed a collaborative-TMLE (CTMLE) that is robust to extreme values of propensity scores in causal inference. Rather than estimating the true propensity score, this method instead fits a model for the probability of receiving the treatment (or being in the non-probability sample in our context) conditional on the estimated conditional mean outcome. Because the treatment model is conditional on a single-dimensional covariate, this approach avoids the challenges related to variable and model selection in the propensity score model. In addition, it relies on only sub-parametric rates of convergence of the outcome model predictions.

In this paper, we firstly propose a variable selection approach in high dimensional covariate settings by extending the outcome adaptive LASSO (Shortreed & Ertefaie, [**104**]). The gain in the present proposal relative to the existing SCAD estimator (Yang, Kim & Song [**101**]) is that the OALASSO can accommodate both the outcome and the selection mechanism in a one-step procedure. Secondly, we adapt the Benkeser, Cai & van der Laan [**16**] collaborative propensity score in our setting. Finally, we perform simulation studies in order to verify the performance of our two proposed estimators and compare them with the existing SCAD estimator for the estimation of the population mean.

The remainder of the article is organized as follows. In Section 2, we define our setting and describe our proposed estimators. In Section 3, we present the results of the simulation study. We present an analysis of the impact of COVID-19 on Canadians in Section 4. A discussion is provided in Section 5.

6.2. Methods

In this section, we present the two proposed estimators in our setting: 1) an extension of the OALASSO for the propensity score [**35, 104**] and 2) the application of Benkeser, Cai & van der Laan's [**16**] alternative propensity score.

6.2.1. The framework

Let $U = \{1,2,...,N\}$ be indices representing members of the target population. Define $\{\boldsymbol{X},Y\}$ as the auxiliary and response variables, respectively where $\boldsymbol{X} = (1, X^{(1)},X^{(2)},...,X^{(p)})$ is a vector of covariates (plus an intercept term) for an arbitrary individual. The finite target population data consists of $\{(\boldsymbol{X}_i,Y_i), i \in U\}$. Let the parameter of interest be the finite population mean $\mu = 1/N \sum_{i \in U} Y_i$. Let $\mathcal{A}$ be indices for the non-probability sample and let $\mathcal{B}$ be those of the probability sample. As illustrated in Figure 1, $\mathcal{A}$ and $\mathcal{B}$ are possibly overlapping subsets of U. Let $d_i = 1/\pi_i$ be the design weight for unit i with $\pi_i = P(i \in \mathcal{B})$ known. The data corresponding to $\mathcal{B}$ consist of observations $\{(\boldsymbol{X}_i,d_i) : i \in \mathcal{B}\}$ with sample size $n_{\mathcal{B}}$. The data corresponding to the non-probability sample $\mathcal{A}$ consist of observations $\{(\boldsymbol{X}_i,Y_i) : i \in \mathcal{A}\}$ with sample size $n_{\mathcal{A}}$.

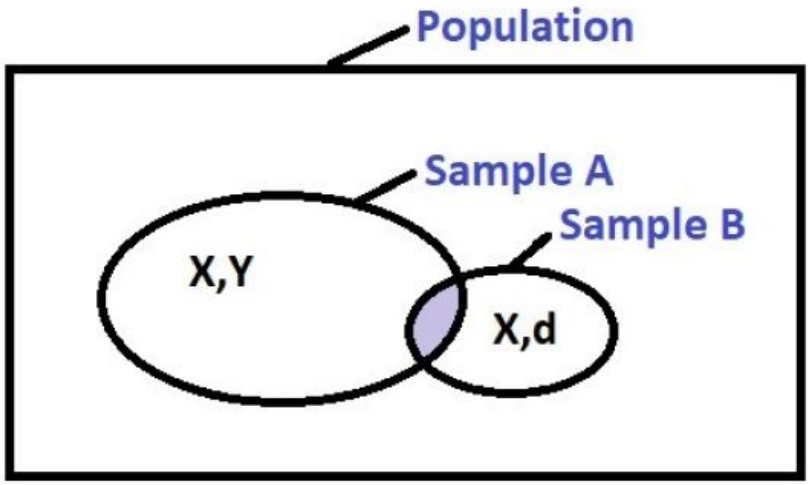

Fig. 1. Population and observed samples.

Sample	X_1	...	X_p	Y	Δ	d
$\mathcal{A}$	.	...	.	.	1	
	.	...	.	.	1	
$\mathcal{B}$	.	...	.		0	.
	.	...	.		0	.
$\mathcal{U} \setminus (\mathcal{A} \cup \mathcal{B})$	.	...	.		0	
	.	...	.		0	

Table 1. Observed data structure

The observed data (Table 1) can be represented as $O = \{\boldsymbol{X}, \Delta, I_{\mathcal{B}}, \Delta Y, I_{\mathcal{B}} d\}$, where Δ is the indicator which equals 1 if the unit belongs to the non-probability sample $\mathcal{A}$ and 0 otherwise, $I_{\mathcal{B}}$ is the indicator which equals 1 if the unit belongs to the probability sample $\mathcal{B}$ and 0 otherwise, and d is the design weight. We use $O_i = \{\boldsymbol{X}_i, \Delta_i, I_{\mathcal{B},i}, \Delta_i Y_i, I_{\mathcal{B},i} d_i\}$ to represent the i-th subject's data realization. Let $p_i = P(\Delta_i = 1 | \boldsymbol{X}_i)$ be the propensity score (the probability of the unit belonging to $\mathcal{A}$). In order to identify the target parameter, we assume these conditions in the finite population: (1) Ignorability, such that the selection indicator Δ and the response variable Y are independent given the set of covariates $\boldsymbol{X}$ (i.e. $\Delta \perp Y | \boldsymbol{X}$) and (2) positivity such that $p_i > \epsilon > 0$ for all i. Note that assumption (1) implies that $E(Y|\boldsymbol{X}) = E(Y|\boldsymbol{X}, \Delta = 1)$, which means that the conditional expectation of the outcome can be estimated using only the non-probability sample $\mathcal{A}$. Assumption (2) guarantees that all units have a non-zero probability of belonging in the non-probability sample.

6.2.2. Estimation of the propensity score

Let's assume for now that the propensity score follows a logistic regression model with $p_i = p(\boldsymbol{X}_i, \boldsymbol{\beta}_0) = exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_0) / \{1 + exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_0)\}$. The true parameter value $\boldsymbol{\beta}_0$ can be estimated as the argument of the maximum (arg max) of the log-likelihood function $\sum_{i=1}^{N} [\Delta_i \log\{p(\boldsymbol{X}_i, \boldsymbol{\beta})\} + (1 - \Delta_i) \log\{1 - p(\boldsymbol{X}_i, \boldsymbol{\beta})\}]$, with summation taken over the target population. One can rewrite this based on our observed data as

$$\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}} \sum_{\mathcal{A}} \boldsymbol{X}_i^T \boldsymbol{\beta} - \sum_{i=1}^{N} \log(1 + e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}). \tag{6.2.1}$$

Equation (6.2.1) cannot be solved directly since $\boldsymbol{X}$ has not been observed for all units in the finite population. However, using the design weight of the probability sample $\mathcal{B}$, $\boldsymbol{\beta}_0$ can be estimated by minimising the pseudo risk function as

$$\arg\min_{\boldsymbol{\beta}} \sum_{\mathcal{A}} \boldsymbol{X}_i^T \boldsymbol{\beta} - \sum_{\mathcal{B}} d_i \log(1 + e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}). \tag{6.2.2}$$

Let $\boldsymbol{X}_{\mathcal{B}}$ be the matrix of auxiliary information (i.e. the design matrix) of the sample $\mathcal{B}$ and $\mathcal{L}(\boldsymbol{\beta})$ the pseudo risk function defined above. Define $U(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta} = \sum_{\mathcal{A}} \boldsymbol{X}_i - \sum_{\mathcal{B}} p_i d_i \boldsymbol{X}_i$, the gradient of the pseudo risk function. Also define $H(\boldsymbol{\beta}) = -\sum_{\mathcal{B}} d_i p_i (1 - p_i) \boldsymbol{X}_i \boldsymbol{X}_i^T = -\boldsymbol{X}_{\mathcal{B}}^T \boldsymbol{S}_{\mathcal{B}} \boldsymbol{X}_{\mathcal{B}}$, the Hessian of the pseudo risk function, where $S_i = d_i p_i (1 - p_i)$ and vector $\boldsymbol{S}_{\mathcal{B}} = (S_i; i \in \mathcal{B})$. The parameter $\boldsymbol{\beta}$ in equation (6.2.2) can be obtained by solving the Newton-Raphson iterative procedure as proposed in Chen, Li & Wu [**121**] by setting $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - H\{\boldsymbol{\beta}^{(t)}\}^{-1} U\{\boldsymbol{\beta}^{(t)}\}$.

6.2.2.1. Variable selection for propensity score

In a high dimensional setting, suppose that an investigator would like to choose relevant auxiliary variables for the propensity score that could help to reduce the selection bias and standard error when estimating the finite population mean. In the causal inference context of estimating the average treatment effect, Shortreed & Ertefaie [**104**] proposed the OALASSO to select amongst the $X^{(j)}$s in the propensity score model. They penalized the aforementioned risk function by the adaptive LASSO penalty (Zou, [**35**]) where the coefficient-specific weights are the inverse of an estimated outcome regression coefficient representing an association between the outcome, Y, and the related covariate.

In our setting, let the true coefficient values of a regression of Y on $\boldsymbol{X}$ be denoted α_j. The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$, corresponding to the covariate coefficients in the propensity score, can be estimated by minimizing the pseudo risk function in (6.2.2) penalized by the adaptive LASSO penalty:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\mathcal{A}} \boldsymbol{X}_i^T \boldsymbol{\beta} - \sum_{\mathcal{B}} d_i \log(1 + e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}) + \lambda \sum_{j=1}^{p} \breve{\omega}_j |\beta_j|. \tag{6.2.3}$$

where $\breve{\omega}_j = 1/|\breve{\alpha}_j|^\gamma$ for some $\gamma > 0$ and $\breve{\alpha}_j$ is a $\sqrt{n}$-consistent estimator of α_j.

Consider a situation where variable selection is not needed ($\lambda = 0$). Chen, Li & Wu [**121**] proposed to estimate $\boldsymbol{\beta}$ by solving the Newton-Raphson iterative procedure. One can rewrite the gradient as $U(\boldsymbol{\beta}) = \sum_{\mathcal{B}}[\Delta_i d_i - p_i d_i]\boldsymbol{X}_i = \boldsymbol{X}_{\mathcal{B}}^T(\boldsymbol{\Sigma} - \boldsymbol{Z}_\beta)$ with vectors $\boldsymbol{Z}_\beta = (p_i d_i; i \in \mathcal{B})$ is a function β and $\boldsymbol{\Sigma} = (\Delta_i d_i; i \in \mathcal{B})$.

The Newton-Raphson update step can be written as:

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + (\boldsymbol{X}_{\mathcal{B}}^T \boldsymbol{S}_\beta \boldsymbol{X}_{\mathcal{B}})^{-1} \boldsymbol{X}_{\mathcal{B}}^T (\boldsymbol{\Sigma} - \boldsymbol{Z}_\beta) \\ &= (\boldsymbol{X}_{\mathcal{B}}^T \boldsymbol{S}_\beta \boldsymbol{X}_{\mathcal{B}})^{-1} \boldsymbol{X}_{\mathcal{B}}^T \boldsymbol{S}_\beta Y^* \end{aligned} \tag{6.2.4}$$

where $Y^* = \boldsymbol{X}_{\mathcal{B}}\boldsymbol{\beta}^{(t)} + \boldsymbol{S}_\beta^{-1}(\boldsymbol{\Sigma} - \boldsymbol{Z}_\beta)$, $X_{\mathcal{B}}^T$ is the matrix of auxiliary variables associated to unit in sample $\mathcal{B}$ and $S_\beta = (S_i; i \in \mathcal{B})$ the vector of weight. Equation (6.2.4) is equivalent to the estimator of the weighted least squares problem with Y^* as the new working response and $S_i = d_i p_i (1 - p_i)$ as the weight associated with unit i. Thus, in our context as well, we can select the important variables in the propensity score by solving a weighted least squares problem penalized with an adaptive LASSO penalty.

6.2.2.2. Implementation of OALASSO

Now we describe how our proposal can be easily implemented in a two-stage procedure. In the first stage, we construct the pseudo-outcome by using the Newton-Raphson estimate of $\boldsymbol{\beta}$ defined in equation (6.2.2) and both samples $\mathcal{A}$ & $\mathcal{B}$. In the second stage, using sample $\mathcal{B}$, we solve a weighted penalized least squares problem with the pseudo-outcome as response variable. The selected variables correspond to the non-zero coefficients of the adaptive LASSO regression. The proposed algorithm for estimating the parameters $\boldsymbol{\beta}$ in equation (6.2.3) with a given value of λ is as follows:

Algorithm 6 OALASSO for propensity score estimation

1: Use the Newton-Raphson algorithm for the unpenalized logistic regression in Chen, Li & Wu [**121**] to estimate $\tilde{\boldsymbol{\beta}}$ in (6.2.2).
2: Obtain the estimated propensity score $\tilde{p}_i = p(\boldsymbol{X}_i, \tilde{\boldsymbol{\beta}})$ for each unit.
3: Construct an estimate of the new working response Y^* by plugging in the estimated $\tilde{\boldsymbol{\beta}}$.
4: Select the useful variables by following steps (a)-(d) below:
   (a) Define $S_i = d_i \tilde{p}_i(1-\tilde{p}_i)$ for each unit in $\mathcal{B}$.
   (b) Run a parametric regression of Y on $\boldsymbol{X}$ using sample $\mathcal{A}$. Obtain $\breve{\alpha}_j$, the estimated coefficient of $X^{(j)}$, $j = 1,...,p$.
   (c) Define the adaptive LASSO weights $\breve{\omega}_j = 1/|\breve{\alpha}_j|^{\gamma}$, $j = 1,...,p$ for $\gamma > 0$.
   (d) Using sample $\mathcal{B}$, run a LASSO regression of Y^* on $\boldsymbol{X}$ with $\breve{\omega}_j$ as the penalty factor associated with $X^{(j)}$ with the given λ.

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\mathcal{B}} S_i (Y_i^* - \boldsymbol{\beta}^T \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} \breve{\omega}_j |\beta_j|$$

   (e) The non-zero coefficient estimate from (d) are the selected variables.
5: The final estimate of the propensity score is $\widehat{p}_i = p(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}) = \frac{exp(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})}{1+exp(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})}$

For the adaptive LASSO tuning parameters, we choose $\gamma = 1$ (Nonnegative Garotte Problem; Yuan & Lin, [**67**]) and λ is selected using V-fold cross-validation in the sample $\mathcal{B}$. The sampling design needs to be taken into account when creating the V-fold in the same way that we form random groups for variance estimation (Wolter, [**59**]). For cluster or stratified sampling for example, all elements in the cluster or stratum should be placed in the same fold.

6.2.2.3. SCAD variable selection for propensity score

Yang, Kim & Song [**101**] proposed a two step approach for variable selection using SCAD. In the first step, they used SCAD to select relevant variables for both the propensity score and the outcome model, respectively. Denote $\mathcal{C}_p$ (respectively $\mathcal{C}_m$) the selected set of relevant variables for the propensity score (respectively the outcome model). The final set of variables used for estimation is $\mathcal{C} = \mathcal{C}_p \cup \mathcal{C}_m$.

6.2.3. Inverse weighted estimators

Horvitz & Thompson [**23**] proposed the idea of weighting observed values by inverse probabilities of selection in the context of sampling methods. The same idea is used to

estimate the population mean in the missing outcome setting. In this approach, inference is conditional on the outcome Y and the covariates X. Recall that based on ignorability, one can write $P(\Delta = 1|Y, \boldsymbol{X}) = P(\Delta = 1|\boldsymbol{X})$ is the propensity score $p(\boldsymbol{X})$. In our context, we take into account (1) a model for the probability for a unit being included in the non-probability sample $\mathcal{A}$ and (2) the selection mechanism for the probabaility sample $\mathcal{B}$. In order to estimate the population mean, the units in the non-probability sample $\mathcal{A}$ are assigned the weights $w_i = 1/\widehat{p}_i$ where $\widehat{p}_i$ is the estimated propensity score obtained using Algorithm 1. Under the joint randomization of the propensity score model and the sampling design for $\mathcal{B}$, the inverse probability weighted (IPW) estimator for the population mean is given by

$$\mu_n^{IPW} = \sum_{i \in \mathcal{A}} w_i Y_i / \widehat{N},$$

where $\widehat{N} = \sum_{i \in \mathcal{A}} w_i$. For the estimation of the variance, we use the proposed variance of Chen, Li & Wu [**121**] which is given by

$$\widehat{V}(\mu_n^{IPW}) = \frac{1}{\widehat{N}_{\mathcal{A}}^2} \sum_{i \in \mathcal{A}} (1 - \widehat{p}_i) \left(\frac{Y_i - \mu_n^{IPW}}{\widehat{p}_i} - \widehat{\boldsymbol{b}}_2^T \boldsymbol{X}_i \right)^2 + \widehat{\boldsymbol{b}}_2^T \widehat{\boldsymbol{D}} \widehat{\boldsymbol{b}}_2$$

with $\widehat{\boldsymbol{D}} = \widehat{N}_{\mathcal{B}}^{-2} \widehat{V}(\sum_{i \in \mathcal{B}} d_i \widehat{p}_i \boldsymbol{X}_i)$, where $\widehat{V}$ is an estimator of $V_p(\mu)$, the design-based variance of the total under the probability sampling design for $\mathcal{B}$ and

$$\widehat{\boldsymbol{b}}_2 = \left\{ \sum_{i \in \mathcal{A}} \left(\frac{1}{\widehat{p}_i} - 1 \right) (Y_i - \mu_n^{IPW}) \boldsymbol{X}_i^T \right\} \left\{ \sum_{i \in \mathcal{B}} d_i \widehat{p}_i (1 - \widehat{p}_i) \boldsymbol{X}_i \boldsymbol{X}_i^T \right\}^{-1}.$$

6.2.4. Augmented Inverse Probability Weighting

Doubly robust semi-parametric methods such as AIPW (Scharfstein, Rotnitzky & Robins, [**25**]) or Targeted Minimum Loss-based Estimation (TMLE, van der Laan & Rubin, [**73**]; van der Laan & Rose, [**76**]) have been proposed to potentially reduce the error resulting from misspecified outcome regressions but also avoid total dependence on the propensity score model specification. In this approach, we take into account (1) a model for the probability for a unit being included in the non-probability sample $\mathcal{A}$, (2) the selection mechanism for the probability sample $\mathcal{B}$ and (3) a model for the outcome. We denote $m(\boldsymbol{X}) = E(Y|\boldsymbol{X})$

the outcome regression model and let $\widehat{m}(\boldsymbol{X})$ be an estimate of $m(\boldsymbol{X})$. Under the joint randomization of the propensity score model, the sampling design for $\mathcal{B}$ and the outcome model, the AIPW estimator proposed in Chen, Li & Wu [**121**] for μ is

$$\mu_n^{AIPW} = \frac{1}{\widehat{N}_{\mathcal{A}}} \sum_{i \in \mathcal{A}} \frac{Y_i - \widehat{m}(\boldsymbol{X}_i)}{p(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})} + \frac{1}{\widehat{N}_{\mathcal{B}}} \sum_{i \in \mathcal{B}} d_i \widehat{m}(\boldsymbol{X}_i)$$

where $\widehat{N}_{\mathcal{A}} = \sum_{i \in \mathcal{A}} 1/p(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})$, $\widehat{N}_{\mathcal{B}} = \sum_{i \in \mathcal{B}} d_i$ and $\widehat{\boldsymbol{\beta}}$ can be estimated using either the Newton-Raphson algorithm in Chen, Li & Wu [**121**] or our proposed OALASSO. One can also use the alternative propensity score proposed by Benkeser, Cai & van der Laan [**16**] and therefore replacing $\widehat{p}_i = p(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})$ by the estimated probability of belonging to the nonprobability sample conditional on the estimated outcome regression $\widehat{P}\{\Delta = 1 | \widehat{m}(\boldsymbol{X}_i)\}$.

For the estimation of the variance, we use the proposed variance of Chen, Li & Wu [**121**] which is given by

$$\widehat{V}(\mu_n^{AIPW}) = \frac{1}{\widehat{N}_{\mathcal{A}}^2} \sum_{i \in \mathcal{A}} (1 - \widehat{p}_i) \left\{ \frac{Y_i - \widehat{m}(\boldsymbol{X}_i) - \widehat{H}_N}{\widehat{p}_i} - \widehat{\boldsymbol{b}}_3^T \boldsymbol{X}_i \right\}^2 + \widehat{W}$$

with $\widehat{H}_N = \widehat{N}_{\mathcal{A}}^{-1} \sum_{i \in \mathcal{A}} \{Y_i - \widehat{m}(\boldsymbol{X}_i)\}/\widehat{p}_i$, $\widehat{t}_i = \widehat{p}_i \boldsymbol{X}_i^T \widehat{\boldsymbol{b}}_3 + \widehat{m}(\boldsymbol{X}_i) - \widehat{N}_{\mathcal{B}}^{-1} \sum_{i \in \mathcal{B}} d_i \widehat{m}(\boldsymbol{X}_i)$, $\widehat{W} = 1/\widehat{N}_{\mathcal{B}}^2 \widehat{V}(\sum_{i \in \mathcal{B}} d_i \widehat{t}_i)$, where $\widehat{V}$ is an estimator of $V_p(\mu)$, the design-based variance of the total under the probability sampling design for $\mathcal{B}$ and

$\widehat{\boldsymbol{b}}_3 = \left[\sum_{i \in \mathcal{A}} \left(\frac{1}{\widehat{p}_i} - 1\right) \{Y_i - \widehat{m}(\boldsymbol{X}_i) - \widehat{H}_N\} \boldsymbol{X}_i^T\right] \left\{\sum_{i \in \mathcal{B}} d_i \widehat{p}_i (1 - \widehat{p}_i) \boldsymbol{X}_i \boldsymbol{X}_i^T\right\}^{-1}$.

6.3. Simulation study

6.3.1. Data generation and parameter estimation

We consider a similar simulation setting as Chen, Li & Wu [**121**]. However, we add 40 pure binary noise covariates (unrelated to the selection mechanism or outcome) to our set of covariates. We generate a finite population $\mathcal{F}_N = \{(\boldsymbol{X}_i, Y_i) : i = 1,...,N\}$ with $N = 10{,}000$, where Y is the outcome variable and $\boldsymbol{X} = \{X^{(1)},...,X^{(p)}\}, p = 44$ represents the auxiliary variables. Define $Z_1 \sim Bernoulli(0.5)$, $Z_2 \sim Uniform(0,2)$, $Z_3 \sim Exponential(1)$ and $Z_4 \sim \chi^2(4)$. The observed outcome Y is a Gaussian with a mean

$\theta = 2 + 0.6X^{(1)} + 0.6X^{(2)} + 0.6X^{(3)} + 0.6X^{(4)}$, where $X^{(1)} = Z^{(1)}$, $X^{(2)} = Z^{(2)} + 0.3X^{(1)}$, $X^{(3)} = Z^{(3)} + 0.2\{X^{(1)} + X^{(2)}\}$, $X^{(4)} = Z^{(4)} + 0.1\{X^{(1)} + X^{(2)} + X^{(3)}\}$, with $X^{(5)}$,...,$X^{(24)} \sim$ *Bernoulli*(0.45) and $X^{(25)}$,...,$X^{(44)} \sim N(0,1)$.

From the finite population, we select a probability sample $\mathcal{B}$ of size $n_{\mathcal{B}} \approx 500$ under a Poisson sampling with probability $\pi \propto \{0.25 + X^{(2)} + 0.03Y\}$. We also consider three scenarios for selecting a non-probability sample $\mathcal{A}$ with the inclusion indicator $\Delta \sim$ *Bernoulli*(p):

- Scenario 1 considers a situation in which the confounders $X^{(1)}$ and $X^{(2)}$ (common causes of inclusion and the outcome) have a weaker relationship with inclusion ($\Delta = 1$) than with the outcome: $P(\Delta = 1|\boldsymbol{X}) = \text{expit}\{-2 + 0.3X^{(1)} + 0.3X^{(2)} - X^{(5)} - X^{(6)}\}$
- Scenario 2 considers a situation in which both confounders $X^{(1)}$ and $X^{(2)}$ have a weaker relationship with the outcome than with inclusion: $P(\Delta = 1|\boldsymbol{X}) = \text{expit}\{-2 + X^{(1)} + X^{(2)} - X^{(5)} - X^{(6)}\}$
- Scenario 3 involves a stronger association between the instrumental variables $X^{(5)}$ and $X^{(6)}$ and inclusion: $P(\Delta = 1|\boldsymbol{X}) = \text{expit}\{-2 + X^{(1)} + X^{(2)} - 1.8X^{(5)} - 1.8X^{(6)}\}$

To evaluate the performance of our method in a nonlinear setting (Scenario 4), we simulate a fourth setting following exactly Kang & Schafer [**43**]. In this scenario, we generate independent $Z^{(i)} \sim N(0,1), i = 1,..4$. The observed outcome is generated as $Y = 210 + 27.4Z^{(1)} + 13.7Z^{(2)} + 13.7Z^{(3)} + 13.7Z^{(4)} + \epsilon$, where $\epsilon \sim N(0,1)$ and the true propensity model is $P(\Delta = 1 \mid \boldsymbol{Z}) = \text{expit}\{-Z^{(1)} + 0.5Z^{(2)} - 0.25Z^{(3)} - 0.1Z^{(4)}\}$. However, the analyst observes the variables $X^{(1)} = \exp\{Z^{(1)}/2\}$, $X^{(2)} = Z^{(2)}/[1 + \exp\{Z^{(1)}\}] + 10$, $X^{(3)} = \{Z^{(1)}Z^{(3)}/25 + 0.6\}^3$, and $X^{(4)} = \{Z^{(2)} + Z^{(4)} + 20\}^2$ rather than the $Z^{(j)}$s.

The parameter of interest is the population mean $\mu_0 = N^{-1}\sum_{i=1}^{N} Y_i$. Under each scenario, we use a correctly specified outcome regression model for the estimation of $m(\boldsymbol{X})$. For the estimation of the propensity score, we perform logistic regression with all 44 auxiliary variables as main terms, LASSO, and OALASSO, respectively. For the Benkeser method, we also use logistic regression for the propensity score. We also implemented IPW with the correct propensity score, and a correctly specified propensity score model, respectively. Because the 4th scenario involves model selection but not variable selection, we only

compare logistic regression with the Benkeser method for the propensity score. We fit a misspecified model and the highly adaptive LASSO (Benseker & van der Lann, [**16**]) for the outcome model.

The performance of each estimator is evaluated through the percent bias ($\%B$), the mean squared error (MSE) and the coverage rate (COV), computed as

$$\%B = \frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{\mu}_r - \mu}{\mu} \times 100$$

$$MSE = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\mu}_r - \mu)^2$$

$$COV = \frac{1}{R}\sum_{r=1}^{R} I(\mu \in \widehat{CI}_r)$$

respectively, where $\widehat{\mu}_r$ is the estimator computed from the rth simulated sample, $\widehat{CI}_r = (\widehat{\mu}_r - 1.96\sqrt{v_r}, \widehat{\mu}_r + 1.96\sqrt{v_r})$ is the confidence interval with v_r the estimated variance using the method proposed by Chen, Li & Wu [**121**] for the rth simulation sample, and $R = 1000$ is the total number of simulation runs.

6.3.2. Results

Tables 2, 3 and 4 contain the results for the first three scenarios. In all three, the IPW estimators performed the worst overall in terms of % bias. Similar to Chen, Li & Wu [**121**], the coverage rates of IPW were suboptimal in all scenarios and the standard error was substantially underestimated in the sense that it is less than the Monte Carlo standard error. We observed the same behaviour even when the propensity score was estimated with a correctly specified model (IPW - True model for PS). However, the standard errors were correctly estimated when the true values of the propensity score were used. The AIPW estimator, implemented with logistic regression, LASSO and OALASSO for the propensity score, performed very well in all scenarios with unbiased estimates and coverage rates close to the nominal 95%. In comparison to IPW and AIPW with logistic regression, incorporating the LASSO or the OALASSO did not improve the bias but did lower the variance and

allowed for better standard error estimation. The Benkeser method slightly increased the bias of AIPW and had underestimated standard errors, leading to lower coverage. The method of Yang et al., [**101**] had the highest bias compared to the other implementations of AIPW and greatly overestimated standard error in all three scenarios.

For the first three scenarios, Figure 2 displays the percent selection of each covariate (1,...,44), defined as the percentage of estimated coefficients that are non-zero throughout the 1000 generated datasets. Overall, the LASSO tended to select the true predictors of inclusion: $X^{(1)}$, $X^{(2)}$, $X^{(5)}$ and $X^{(6)}$. For example, in scenario (2), confounders ($X^{(1)}$, $X^{(2)}$) were selected in around 94% of simulations and instruments ($X^{(5)}$, $X^{(6)}$) around 90%. However, the percent selection of pure causes of the outcome ($X^{(3)}$, $X^{(4)}$) was around 23%. On the other hand, when OALASSO was used for the propensity score, the percent selection of confounders ($X^{(1)}$, $X^{(2)}$) was around 98% and instruments ($X^{(5)}$, $X^{(6)}$) was 64%. However, the percent selection of pure causes of the outcome ($X^{(3)}$, $X^{(4)}$) increased to 83%. When using Yang's proposed selection method, $X^{(1)}$, $X^{(2)}$ and $X^{(3)}$ were selected 100 percent of the time.

Table 5 contains the results of the Kang and Shafer [**43**] setting. AIPW with HAL for the outcome model and either the collaborative propensity score (AIPW - Benkeser method) or propensity score with logistic regression with main terms (AIPW - Logistic (2)) achieved lower % bias and MSE compared to IPW. However, when the outcome model was misspecified, AIPW with logistic regression (AIPW - Logistic (1)) performed as IPW. In this scenario, the true outcome expectation and the propensity score functionals were nonlinear, making typical parametric models misspecified. Consistent estimation of the outcome expectation can be obtained by using flexible models. The collaborative propensity score was able the reduce the dimension of the space and collect the necessary information using the estimated conditional mean outcome for unbiased estimation of the population mean with a coverage rate that was close to nominal.

Table 2. Scenario 1: Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; IPW-LASSO: IPW with LASSO regression for propensity score; IPW-OALASSO: IPW with OALASSO regression for propensity score; AIPW-Logistic: AIPW with logistic regression for propensity score; AIPW-LASSO: AIPW with LASSO regression for propensity score; AIPW-OALASSO: AIPW with OALASSO regression for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score; AIPW-Yang: Yang's proposed AIPW.

Estimator	$\%B$	**MSE**	**MC SE**	**SE**	$\%$**COV**
IPW - True PS	-0.02	0.01	0.08	0.08	96
IPW - True model for PS	-0.11	0.01	0.10	0.07	85
IPW - Logistic	-0.51	0.03	0.12	0.07	79
IPW - LASSO	-1.75	0.03	0.12	0.07	49
IPW - OALASSO	-0.94	0.06	0.25	0.07	42
AIPW - Logistic	0.01	0.01	0.10	0.14	99
AIPW - LASSO	-0.03	0.01	0.10	0.09	93
AIPW - OALASSO	-0.01	0.01	0.10	0.11	94
AIPW - Benkeser	-0.00	0.01	0.10	0.08	90
AIPW - Yang	-0.60	0.01	0.10	0.30	100

Table 3. Scenario 2: Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; IPW-LASSO: IPW with LASSO regression for propensity score; IPW-OALASSO: IPW with OALASSO regression for propensity score; AIPW-Logistic: AIPW with logistic regression for propensity score; AIPW-LASSO: AIPW with LASSO regression for propensity score; AIPW-OALASSO: AIPW with OALASSO regression for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score; AIPW-Yang: Yang's proposed AIPW.

Estimator	$\%B$	**MSE**	**MC SE**	**SE**	$\%$**COV**
IPW - True PS	-0.00	0.00	0.05	0.06	97
IPW - True model for PS	-0.20	0.01	0.10	0.05	73
IPW - Logistic	0.74	0.03	0.18	0.09	66
IPW - LASSO	-2.49	0.03	0.11	0.04	25
IPW - OALASSO	-1.32	0.04	0.18	0.05	37
AIPW - Logistic	0.03	0.01	0.10	0.18	100
AIPW - LASSO	-0.08	0.01	0.10	0.09	94
AIPW - OALASSO	-0.04	0.01	0.10	0.10	95
AIPW - Benkeser	-0.13	0.01	0.10	0.06	83
AIPW - Yang	-1.59	0.02	0.09	0.29	100

Table 4. Scenario 3: Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; IPW-LASSO: IPW with LASSO regression for propensity score; IPW-OALASSO: IPW with OALASSO regression for propensity score; AIPW-Logistic: AIPW with logistic regression for propensity score; AIPW-LASSO: AIPW with LASSO regression for propensity score; AIPW-OALASSO: AIPW with OALASSO regression for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score; AIPW-Yang: Yang's proposed AIPW.

Estimator	$\%B$	**MSE**	**MC SE**	**SE**	$\%$**COV**
IPW - True PS	0.05	0.01	0.10	0.11	97
IPW - True model for PS	-0.12	0.01	0.10	0.06	78
IPW - Logistic	1.11	0.13	0.36	0.29	87
IPW - LASSO	-3.01	0.06	0.15	0.09	44
IPW - OALASSO	-1.71	0.07	0.25	0.11	58
AIPW - Logistic	0.03	0.02	0.15	0.32	100
AIPW - LASSO	0.03	0.01	0.10	0.09	93
AIPW - OALASSO	0.03	0.01	0.10	0.10	94
AIPW - Benkeser	0.06	0.01	0.10	0.07	85
AIPW - Yang	-1.59	0.02	0.10	0.30	100

Table 5. Scenario 4 (non-linear model setting): Estimates taken over 1000 generated datasets. %B (percent bias), MSE (mean squared error), MC SE (monte carlo standard error), SE (square root of the mean variance) and COV (percent coverage). IPW-Logistic: IPW with logistic regression for propensity score; AIPW-Logistic (1): AIPW with logistic regression for propensity score and a misspecified model for the outcome; AIPW-Logistic (2): AIPW with logistic regression for propensity score and HAL for the outcome model; AIPW-Benkeser: AIPW with the collaborative propensity score.

Estimator	$\%B$	**MSE**	**MC SE**	**SE**	$\%$**COV**
IPW - Logistic	3.32	66.56	0.40	0.34	0
AIPW - Logistic (1)	3.33	66.56	0.40	1.15	0
AIPW - Logistic (2)	0.12	2.60	1.59	1.00	86
AIPW - Benkeser	0.12	2.61	1.59	1.28	93

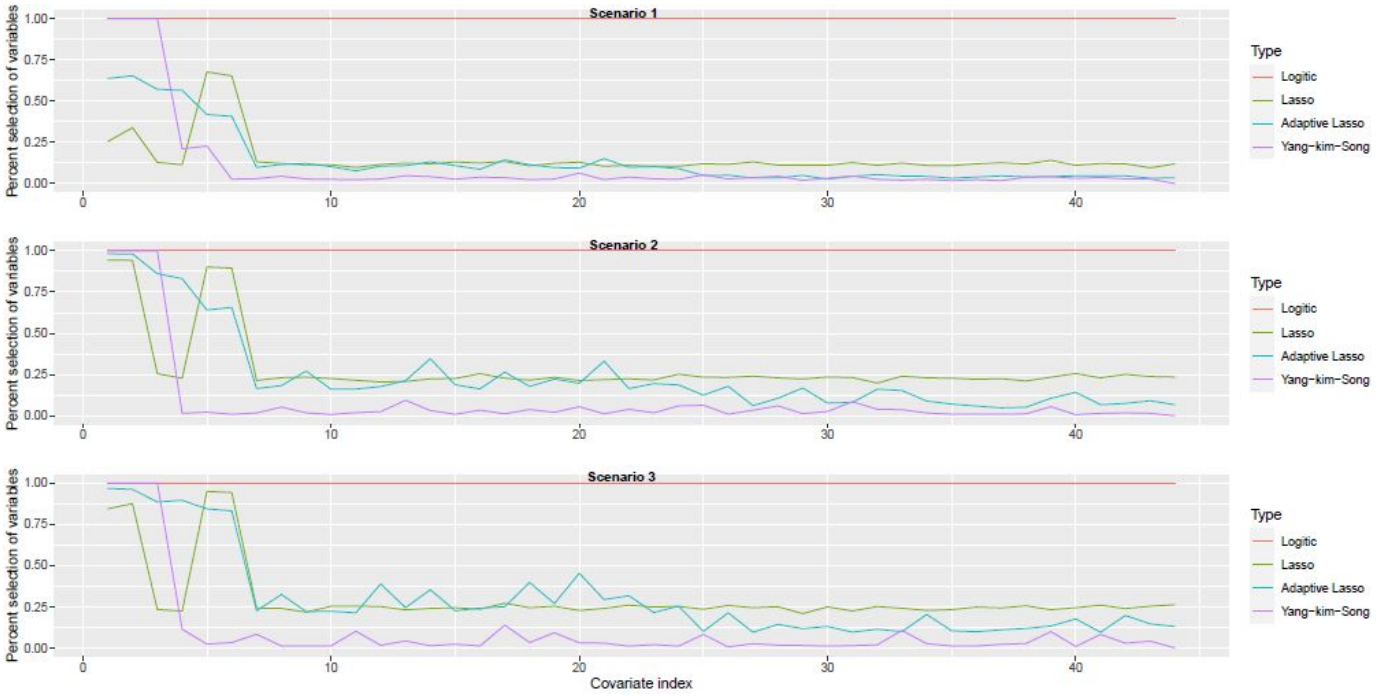

Fig. 2. Percent selection of each variable into the propensity score model over 1000 simulations under scenarios 1-3.

6.4. Data Analysis

In this section, we apply our proposed method to a survey which was conducted by Statistics Canada to measure the impacts of COVID-19 on Canadians. The main topic was to determine the level of trust Canadians have in others (elected officials, health authorities, other people, businesses and organizations) in the context of the COVID-19 pandemic. Data was collected from May 26 to June 8, 2020. The dataset was completely non-probabilistic with a total of 35,916 individuals responding and a wide range of basic demographic information collected from participants along with the main topic variables. The dataset is referred to as Trust in Others (TIO).

We consider Labor Force Survey (LFS) as a reference dataset, which consists of $n_{\mathcal{B}} = 89{,}102$ subjects with survey weights. This dataset does not have measurements of the study outcome variables of interest; however, it contains a rich set of auxiliary information common with the TIO. Summaries (unadjusted sample means for TIO and design-weighted means for LFS) of the common covariates are listed in Table 8 in the appendix. It can be seen that the

distributions of the common covariates between the two samples are different. Therefore, using TIO only to obtain any estimate about the Canadian population may be subject to selection bias.

We apply the proposed methods and the sample mean to estimate the population mean of two response variables. Both of these variables were assessed as ordinal : Y_1, "trust in decisions on reopening, Provincial/territorial government" – 1: cannot be trusted at all, 2, 3: neutral, 4, 5: can be trusted a lot; and Y_2, "when a COVID-19 vaccine becomes available, how likely is it that you will choose to get it?" – 1: very likely, 2: somewhat likely, 3: somewhat unlikely, 4: very unlikely, 7: don't know. Y_1 was converted to a binary outcome which equals 1 for a value less or equal to 3 (neutral) and 0 otherwise. The same type of conversion was applied for Y_2 to be 1 for a value less or equal to 2 (somewhat likely) and 0 otherwise. We used logistic regression, outcome adaptive group LASSO (Wang & Leng, [**34**]; Hastie et al. [**108**]; as we have categorical covariates), and the Benkeser method for the propensity score. We also fit group LASSO for the outcome regression when implementing AIPW. Each categorical covariate in Table 8 were converted to binary dummy variables. Using 5-fold cross-validation, the group LASSO variable selection procedure identified all available covariates in the propensity score model. Table 6 below presents the point estimate, the standard error and the 95% Wald-type confidence intervals. For estimating the standard error, we used the variance estimator for IPW and the asymptotic variance for AIPW proposed in Chen, Li & Wu [**121**]. For both outcomes, we found significant differences in estimates between the naive sample mean and our proposed methods for both AIPW with OA group LASSO and the Benkeser method. For example, the adjusted estimates for Y_1 suggested that, on average, at most 40% (using both outcome adaptive group LASSO or the Benkeser method) of the Canadian population have no trust at all or are neutral in regards to decisions on reopening taken by their provincial/territorial government compared to 43% if we would have used the naive mean. The adjusted estimates for Y_2 suggested that at most 80% using the Benkeser method (or 82% using outcome adaptive group LASSO) of the Canadian population are very or somewhat likely to get the vaccine compared to 83% if we would have used the naive

mean. On the other hand, there was no significant differences between OA group LASSO and group LASSO compared to the naive estimator. The package IntegrativeFPM (Yang, [**100**]) threw errors during application, which is why it is not included.

Table 6. Point estimate, standard error and 95% Wald confidence interval. IPW-Logistic (Grp LASSO/OA Grp LASSO): IPW with logistic regression (Group LASSO/outcome adaptive Group LASSO) for propensity score; AIPW-Logistic (Grp LASSO/OA Grp LASSO): AIPW with logistic regression (Group LASSO/outcome adaptive Group LASSO for propensity score; AIPW-Benkeser: AIPW with the collaborative propensity score.

	Estimator	Mean (SE)	%95 CI
Y_1	Sample mean	0.430 (0.002)	0.424 - 0.435
	IPW - Logistic	0.382 (0.024)	0.330 - 0.430
	IPW - Grp LASSO	0.383 (0.024)	0.335 - 0.431
	IPW - OA Grp LASSO	0.386 (0.024)	0.340 - 0.433
	AIPW - Logistic	0.375 (0.022)	0.328 - 0.415
	AIPW - Grp LASSO	0.372 (0.014)	0.344 - 0.401
	AIPW - OA Grp LASSO	0.373 (0.014)	0.348 - 0.403
	AIPW - Benkeser	0.401 (0.002)	0.396 - 0.406
Y_2	Sample mean	0.830 (0.001)	0.826 - 0.834
	IPW - Logistic	0.820 (0.013)	0.794 - 0.847
	IPW - Grp LASSO	0.810 (0.013)	0.784 - 0.836
	IPW - OA Grp LASSO	0.808 (0.013)	0.784 - 0.833
	AIPW - Logistic	0.810 (0.013)	0.784 - 0.837
	AIPW - Grp LASSO	0.796 (0.012)	0.774 - 0.819
	AIPW - OA Grp LASSO	0.796 (0.011)	0.775 - 0.818
	AIPW - Benkeser	0.788 (0.003)	0.783 - 0.794

6.5. Discussion

In this paper, we proposed an approach to variable selection for propensity score estimation through penalization when combining a non-probability sample with a reference probability sample. We also illustrated the application of the collaborative propensity score method of Benkeser, Cai & van der Laan [**16**] with AIPW in this context. Through the simulations, we studied the performance of the different estimators and compared them with the method proposed by Yang [**101**]. We showed that the LASSO and the OALASSO can reduce the standard error and mean squared error in a high dimensional setting. Also, the IPW plug-in variance estimator proposed by Chen, Li & Wu [**121**] gave a value less than the Monte Carlo standard error overall, leading to confidence interval undercoverage. The collaborative propensity score produced good results but the related confidence intervals were

suboptimal as the true propensity score is not estimated there.

Overall, in our simulations, we have seen that doubly robust estimators generally outperformed the IPW estimators. Doubly robust estimators incorporate the outcome expectation in such a way that can help to reduce the bias when the propensity score model is not correctly specified. Our observations point to the importance of using doubly robust methodologies in this context.

In our application, we found statistically significant differences in the results between our proposed estimator and the corresponding naive estimator for both outcomes. This analysis used the variance proposed by Chen, Li & Wu [**121**] which relies on the true propensity score for IPW estimators. This variance estimator was derived by taking the first-order term in a Taylor series decomposition which assumes that the remainder terms converge rapidly to zero in probability. For future research, it would be quite interesting to develop a variance estimator that is robust to propensity score misspecification/estimation and that can be applied to the Benkeser method. Other possible future directions include post-selection variance estimation in this setting.

6.6. Appendix

Table 7. Distributions of common covariates from the two samples.

Methods	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	$X^{(6)}$	$X^{(7),..,(44)}$
				Scenario 1			
LASSO	25	34	13	11	67	65	24
OALASSO	67	65	57	56	42	40	20
Yang's method	100	100	100	21	23	3	3
				Scenario 2			
LASSO	94	94	25	23	90	90	24
OALASSO	98	98	86	83	64	65	20
Yang's method	100	100	100	1.5	2	1	3
				Scenario 3			
LASSO	84	87	23	23	95	94	24
OALASSO	96	96	88	89	84	83	20
Yang's method	100	100	100	11	2	3	3

Table 8. Distributions of common covariates from the two samples.

Covariates	TIO N (mean)	LFS N (mean)
Sample size	35916	89102
Born in Canada	30867 (85.94%)	72048(71%)
Landed immigrant or permanent resident	149(41%)	15277(26.36%)
Sex (1:Male)	10298(28.67%)	43415(49.35%)
Rural/Urban indicator (1: rural)	4395(12.24%)	14119(8.25%)
Education		
At least High school diploma or equivalency	35588(99.08%)	74288(85.46%)
At least Trade certificate or diploma	32192(89.63%)	51036(59.60%)
At least College - cegep -other non university certificate	30568(85.10%)	41038(50.06%)
At least University certificate or diploma below bachelor	23544(65.55%)	22826(30.50%)
At least Bachelor degree	21299(59.30%)	13865(15.56%)
At least University degree above bachelor	10118(28.17%)	6526(8.97%)
Indigenous identity flag	1047(2.92%)	3689(2.48%)
Province		
Newfoundland and Labrador	328(0.91%)	2965(1.41%)
Prince Edward Island	161(0.45%)	325(0.42%)
Nova Scotia	1762(4.91%)	4695(2.61%)
New Brunswick	794(2.21)	4727(2.04%)
Quebec	5861(16.32%)	16455(22.85%)
Ontario	17177(47.83%)	24978(39.53%)
Manitoba	922(2.57%)	7607(3.33%)
Saskatchewan	890(2.48%)	6104(2.87%)
Alberta	2875(8%)	9265(11.48%)
British Columbia	5146(14.33%)	9981(13.37%)
Age group in increments of 10		
15-24	1113(3.10%)	10902(14.15%)
25-34	6162(17.16%)	12336(16.83%)
35-44	8554(23.82%)	13573(16.16%)
45-54	7309(20.35%)	13912(15.11%)
55-64	7111(19.80%)	6496(16.62%)
65+	5667(15.78%)	1883(21.10%)

Chapitre 7

Conclusion générale et discussion

7.1. Résumé des résultats principaux

Bien que les ECRs constituent le devis par excellence pour tirer des conclusions de cause à effet, il est possible d'en faire autant en utilisant les données administratives sans toutefois oublier les limites de ces dernières, qui sont le biais de confusion et de sélection. Dans cette thèse, nous avons proposé des méthodes d'estimation pour différents paramètres d'intérêts marginaux en inférence causale et pour des données administratives complètes. Plus précisément, nos articles présentés dans cette thèse traitent entre autres de l'instabilité des estimateurs doublement robustes quand il y a violation de la positivité en pratique, la sélection d'un modificateur d'effet, l'estimation de l'effet conditionnel, et finalement, l'ajustement du biais de sélection lors de l'utilisation de données administratives pour l'estimation d'un paramètre d'intérêt. Les méthodes proposées dans ces articles sont reliées dans ce sens qu'elles traitent toutes d'estimateurs doublement robustes pour l'estimation de paramètres marginaux. De plus, l'intégration des méthodes d'apprentissage automatique dans ces estimateurs est possible pour l'estimation des paramètres de nuisance, soit le score de propension et/ou l'espérance conditionnelle de l'issue.

Le premier article démontre comment l'estimateur du maximum de vraisemblance ciblé (TMLE) peut être instable en absence de positivité en pratique et selon la méthode utilisée pour estimer le score de propension. D'autre part, cet article propose un algorithme basé

sur le rééchantillonnage pour identifier une telle instabilité. Un *package R* a été créé par Liu et al., [**122**]. Dans la littérature à ce jour, une seule méthode [**78**] existe pour diagnostiquer l'impact de la positivité sur l'estimation d'un paramètre causal. Toutefois, cette méthode ne considère pas l'association entre le traitement et la co-variable dans son application. Notre algorithme est une avancée dans la littérature, dans le sens où il permet de mesurer l'impact de l'estimation du score de propension en présence de positivité pratique via une procédure bootstrap, tout en gardant l'association entre les co-variables et le traitement. Il permet ainsi d'attribuer la divergence de l'estimateur au score de propension vu que le modèle de l'issue est connu.

Le second article démontre comment un modificateur d'effet peut être identifié par une procédure doublement robuste. Cet article va plus loin en permettant l'estimation de l'effet conditionnel et l'incorporation des méthodes d'apprentissage automatique. Nous avons montré que notre estimateur hérite des propriétés oracle du LASSO adaptatif lorsqu'il est construit avec le vrai modèle de l'issue et du score de propension. Depuis quelques années, comprendre l'hétérogénéité dans les populations et faire de la médecine personnalisée sont de plus en plus courant en pratique. Notre estimateur est une avancée pratique dans la littérature étant donné sa simplicité à comprendre et à implémenter. Par ailleurs, notre procédure est l'une des rares méthodes [**114, 84**] à ce jour à intégrer une méthode de régularisation pour la sélection de modificateurs d'effet.

Le troisième article démontre comment le biais de sélection peut être ajusté en utilisant des données non probabilistes sujettes à la sous-couverture et comment un échantillon probabiliste peut être mis à profit. Cet article propose une procédure d'estimation du score de propension en intégrant une méthode de régularisation pour la sélection de variables. Dans la littérature à ce jour, une seule procédure de sélection de variables [**101**] existe dans ce contexte et a été proposée pour les estimateurs doublement robustes. Notre méthode est une avancée importante dans ce contexte, puisqu'elle est applicable à tout estimateur utilisant le score de propension et non seulement aux estimateurs doublement robustes. De plus, notre

procédure est la deuxième à ce jour permettant la sélection de variables lors de la combinaison entre données administratives et données d'enquêtes.

Les études de simulation effectuées dans les 3 articles permettent d'évaluer la performance de nos algorithmes. Dans le premier article, le scénario étudié présente comment le score de propension, estimé avec des méthodes d'apprentissage automatique, peut augmenter le biais, comparé à un estimateur paramétrique simple comme la régression logistique. De plus, la procédure de diagnostic proposée permet de détecter et d'informer l'utilisateur quand l'estimation obtenue est douteuse. Les études de simulation du second article mettent en évidence la double robustesse de notre algorithme en ce qui concerne la sélection de modificateurs d'effet et comment, l'utilisation des méthodes d'apprentissage automatique pour les paramètres de nuisance peuvent être utiles. En dernier lieu, ces simulations montrent aussi qu'un modèle linéaire, non correctement spécifié pour l'estimation des effets conditionnels, peut donner lieu à des estimations biaisées. Dans le dernier article, les études de simulation révèlent la diminution de la variance et de l'erreur quadratique moyenne, lors de l'estimation de la moyenne de la population, en utilisant les procédures de sélection proposées. De plus, nos simulations démontrent aussi l'importance des estimateurs doublement robustes dans ce contexte. Par ailleurs, les résultats de simulations ont montré que le score collaboratif produit des estimés sans biais. En effet, en conditionnant sur l'espérance conditionnelle de l'issue, le score collaboratif utilise cette dernière comme une sorte de réduction de dimensions. Ainsi, il est une amélioration par rapport au TMLE collaboratif étant donné sa simplicité à implémenter et est robuste par en présence de positivité pratique [**16**].

Le premier et second article de ce mémoire contient des analyses sur les médicaments contre l'asthme durant la grossesse. L'objectif étant d'estimer l'effet de la prise de corticostéroïdes sur le poids du nouveau-né. Dans le premier article, la procédure BDT démontre que l'utilisation du Super Learner pour l'estimation du score de propension fait augmenter le biais associé à l'effet des corticostéroïdes sur le poids du nouveau-né. De plus, cet effet s'avère même significatif, ce qui n'est pas le cas dans la littérature. Le second article utilise ces données dans le but d'estimer l'effet conditionnel. L'antagoniste des récepteurs des leucotriènes

et les anomalies chromosomiques ont été identifiés comme des modificateurs d'effets par notre procédure et également un modèle linéaire avec interaction. Toutefois, les effets conditionnels diffèrent pour la simple raison que le modèle linéaire ne cible pas le même paramètre s'il n'est pas correctement spécifié. Le dernier article présente des analyses sur l'impact de la Covid-19 sur les Canadiens. Cette analyse montre une différence statistiquement significative entre la moyenne naïve et celle obtenue en utilisant nos deux procédures proposées. En effet, notre analyse montre qu'en moyenne, 80% des Canadiens sont très certain ou assez de se faire vacciner contre la Covid-19, mais cette proportion est de 83% en utilisant une simple moyenne. Cette différence peut être expliquée par la présence de biais de sélection dans l'échantillon non-probabiliste. En effet, le Tableau 8 du chapitre 6 nous montre qu'il y a moins de personnes âgées entre 15-24 dans les données administratives comparées à la réalité dans la population. Or, les personnes jeunes sont celles qui se sentent plus intouchables et vont avoir tendance à refuser la vaccination comme on peut observer avec les données de vaccination en ce moment au Canada. Les données administratives montrent également qu'il y a plus de personnes instruites comparées à la réalité dans la population. Or, les personnes instruites sont les plus à même à comprendre et faire confiance à la science et donc accepter la vaccination. Ainsi, en ajustant pour ces différences, la proportion de personnes désirant se faire vacciner devrait diminuer, ce qui explique les résultats obtenus.

7.2. Limites et perspectives

Le premier article démontre à quel point il est important de diagnostiquer et d'utiliser avec précaution les outils et méthodes existants pour l'estimation des effets causaux. Ces résultats prouvent à suffisance la nécessité d'avoir plus d'outils capables d'aider des analystes à vérifier et valider les résultats provenant de différentes méthodes d'estimation. Néanmoins, l'algorithme présenté dans cet article est utile uniquement pour les estimateurs incluant directement l'inverse du score de propension. Il existe d'autres estimateurs doublements robustes qui sont stables face aux problèmes de violation de la positivité pratique. Nous pouvons citer entre autres, l'estimateur de vraisemblance augmentée de Tan [**124**], le

C-TMLE traditionnel [**75**] et d'autres versions du C-TMLE [**14, 24**] pour ne citer qu'eux. Certaines méthodes énumérées ci-dessus diffèrent du TMLE traditionnel dans le sens où elles n'incluent pas directement l'inverse du score de propension [**14, 24**]. Par ailleurs, d'autres l'incluent, mais sont robustes par rapport aux problèmes de violation de la positivité en pratique [**75**]. Ces méthodes peuvent être une solution optimale dans les situations où il y a suspicion de violation de la positivité en pratique et elles permettent aussi d'avoir des gains en termes de variance. En pratique, le TMLE traditionnel demeure une méthode intéressante étant donné sa flexibilité à s'appliquer à différentes structures de données et de paramètres d'intérêts. Toutefois, le C-TMLE est un estimateur irrégulier et l'estimateur de la variance reste un défi de taille. Des travaux futurs sur ce sujet pourraient impliquer l'identification de scénarios dans lesquels les méthodes existantes et fortement utilisées en pratique échouent, dans le but de notifier nos professionnels de recherche sur leur utilisation.

Les résultats présentés dans l'article 2 montrent comment sélectionner un modificateur d'effet et estimer un effet conditionnel de manière doublement robuste. Néanmoins, cette proposition assume une forme linéaire pour arriver à ces fins. Cette hypothèse est toutefois courante en pratique. Dans une situation où l'effet conditionnel serait le paramètre ciblé et non le modificateur d'effet en soi, des méthodes flexibles ou d'apprentissage automatique peuvent être utilisées pour modéliser l'issue modifiée. L'intégration des poids basés sur l'issue dans l'estimateur HAL pourrait également aider à améliorer les performances de notre estimateur comme démontré dans Ju et al., [**11**]. Bien que les résultats présentés dans le second article soient encourageants, d'autres avenues de recherche restent à être explorées. Par exemple, il serait intéressant de comprendre le taux élevé de fausse couverture observé en présence de données de grandes dimensions. L'ajustement croisé (*cross-fitting*) lors de l'estimation de l'issue modifiée pourrait être considéré comme démontré dans Kennedy [**30**]. Par ailleurs, il serait également important de définir un cadre idéal de sélection des paramètres de nuisance dans la validation croisée effectuée dans la procédure de sélection de modificateurs d'effet.

Les travaux présentés dans le troisième article démontrent que la sélection de variables lors

de l'estimation du score de propension quand on est en présence de données de grandes dimensions permet d'avoir des gains en ce qui concerne la variance. Toutefois, dans cet article, nous avons supposé une forme logistique pour le score de propension dans le but d'étendre la théorie du LASSO adaptatif. Il se pourrait qu'en réalité, le vrai score de propension ne suive pas cette forme. Très peu de travaux existent dans ce contexte. La procédure basée sur les arbres de régression de Beaumont & Chu [**45**] pourrait être une solution. Une autre limite observée dans cet article est la sélection des paramètres de nuisance dans la procédure du LASSO adaptatif. En effet, nous avons utilisé la procédure traditionnelle qui consiste à minimiser l'erreur de prédiction. D'autres critères comme la différence absolue de la moyenne pondérée [**104**] pourraient être utilisés. L'article 3 a adopté l'estimateur de variance proposé par Chen, Li & Wu [**121**]. D'autres travaux futurs sur l'impact de la sélection de variables sur l'inférence doivent être menés.

En résumé, les bases de données administratives continueront de jouer un rôle de plus en plus important en pharmacoépidémiologie et pour l'obtention de statistiques officielles. Plus de développements statistiques doivent être mis en place pour mieux guider les utilisateurs dans leur analyse. Également, un travail de vulgarisation et de communication scientifiques doit être fait pour une meilleure compréhension des enjeux. Les estimateurs doublement robustes sont très bénéfiques et sont un atout important étant donné leurs propriétés. Les travaux réalisés dans cette thèse ont montré leur importance et également ont mis en lumière le choix éclairé à faire lors de l'estimation des paramètres de nuisance pour une estimation convergente.

Références bibliographiques

[1] Bahamyirou A, Blais L, Forget A et Schnitzer ME : Understanding and diagnosing the potential for bias when using machine learning methods with doubly robust causal estimators. *Statistical Methods in Medical Research*, 0:1–14, 2019.

[2] Brookhart M A, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J et Sturmer T : Variable selection for propensity score models. *American Journal of Epidemiology*, 163:1149–1156, 2006.

[3] Javanmard A et Montanari A : Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[4] Rafei A, Flannagan AC et Elliott MR : Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8:148–180, 2020.

[5] Tsiatis Anastasios A : *Semiparametric theory and missing data*. Springer, 2006.

[6] Wisniowski A, Sakshaug JW, Ruiz DAP et Blom AG : Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8:120–147, 2020.

[7] Decker AL, Hubbard A, Crespi CM, Seto EYW et Wang MC : Semiparametric estimation of the impacts of longitudinal interventions on adolescent obesity using targeted maximum-likelihood: Accessible estimation with the ltmle package. *Journal of Causal Inference*, 2:95–108, 2014.

[8] Cossette B, Forget A, Beauchesne MF, Rey E, Lemière C, Larivée P, Battista MC et Blais L : Impact of maternal use of asthma-controller therapy on perinatal outcomes. *Thorax*, 68(8):724–730, 2013.

[9] Efron B : Bootstrap methods: Another look at the jackknifes. *The Annals of Statistics*, 7(1):1–26, 1979.

[10] Lee BK, Lessler J et Stuart EA : Weight trimming and propensity score weighting. *PLoS One*, 6(3), 2011.

[11] Ju C, Benkeser D et van der Laan MJ : Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometric Methodology*, 76:109–118, 2020.

[12] Hazelbag CM, Zaal IJ, Devlin IW and; Gatto NM, Hoes AW, Slooter AJC et Groenwold RHH : An application of inverse probability weighting estimation of marginal structural models of a continuous exposure benzodiazepines and delirium. *Epidemiology*, 26(5):52–53, 2015.

[13] Benkeser D, Carone M, van der Laan MJ et Gilbert P : Doubly robust nonparametric inference on the average treatment effect estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 356, 2016.

[14] Benkeser D, Carone M, van der Laan MJ et Gilbert P : Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104:863–880, 2017.

[15] Benkeser D et van der Laan MJ : The highly adaptive lasso estimator. *In 2016 IEEE International Conference on Data Science and Advanced Analytics*, pages 689–696, 2016.

[16] Benkeser D, Cai W et van der Laan MJ : A nonparametric super-efficient estimator of the average treatment effect. *Statistical Sciences*, 35(3):484–495, 2020.

[17] Rubin D et van der Laan MJ : A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3.

[18] Rubin D et van der Laan MJ : Extending marginal structural models through local, penalized, and additive learning. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

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

[20] Rubin DB : Randomization analysis of experimental data : The fisher randomization test comment. *ournal of the American Statistical Association*, 75(371):591–593, 1804.

[21] Rubin DB : Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[22] Rubin DB : Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

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

[24] Ivan Diaz : Doubly robust estimators for the average treatment effect under positivity violations: introducing the e-score. *https://arxiv.org/abs/1807.09148*, 2018.

[25] Scharfstein DO, Rotnitzky A et Robins JM : Adjusting for nonignorable dropout using semiparametric nonresponse models, (with discussion and rejoinder). *Journal of the American Statistical Association*, pages 1121–1146, 1999.

[26] Green DP et Kern HL : Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression tree. *The Public Opinion Quarterly*, 76:496–511, 2012.

[27] Polley EC, LeDell E et van der Laan MJ : Super learner. *https://github.com/ecpolley/SuperLearner*, 2016.

[28] Polley EC et van der Laan MJ : Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 266, 2010.

[29] Schisterman EF, Cole SR et Platt RW : Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(448), 2009.

[30] Kennedy EH : Optimal doubly robust estimation of heterogeneous causal effects. *arXiv:2004.14497v1*, 2020.

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

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

[33] Kennedy Edward H : Semiparametric theory and empirical processes in causal inference. *arXiv:1510.04740*, 2016.

[34] Wang H. et Leng C : A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52:5277–5286, 2008.

[35] Zou H : The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 2006.

[36] Chipman HA, George EI et McCulloch RE : Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[37] Abdollahpour I, Nedjat S, Almasi-Hashiani A, Nazemipour M, Mansournia MA et Luque-Fernandez MA : Estimating the marginal causal effect and potential impact of waterpipe smoking on multiple sclerosis using targeted maximum likelihood estimation method: a large population-based incident case-control study. *American Journal of Epidemiology*, 2021.

[38] Fan J et Li R : Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[39] Friedman J : Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[40] Schwab J, Lendle SD, Petersen ML et van der Laan MJ : Ltmle: Longitudinal targeted maximum likelihood estimation. *Cran*, 2014.

[41] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss C, Rothman KJ, Joffe MM et Glynn RJ : Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11), 2011.

[42] Lee JD, Sun DL, Sun Y et Taylor JE : Exact post-selection inference with application to the lasso. *The Annals of Statistics*, 44:907–927, 2016.

[43] Kang JDY et Shafer JL : Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.

[44] Beaumont JF : Les enquêtes probabilites sont-elles vouées à disparaître pour la production de statistiques officielles? *Survey Methodology*, 46(1):1–30, 2020.

[45] Beaumont JF et Chu K : Statistical data integration through classification trees. *Report paper for ACSM*, 2020.

[46] Edwards JK, Cole SR, Lesko CR Mathews WC, Moore RD, Mugavero MJ et Westreich D : An illustration of inverse probability weighting to estimate policy-relevant causal effects. *American Journal of Epidemiology*, 183(4):336–344, 2015.

[47] Franklin JM, Schneeweiss S, Polinski JM et Rassen JA : Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databasess. *Computational Statistics and Data Analysis*, 72:219–226, 2041.

[48] Robins JM : A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(260):1393–1512, 1986.

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

[50] Robins JM : *Marginal structural models versus structural nested models as tools for causal inference*. Springer, 2000.

[51] Robins JM, Rotnitzky A et Zhao LP : Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(7):846–866, 1994.

[52] Robins JM, Hernan MA et Brumback B : Marginal structural models and causal inference in epidemiology. *Epidemiology*, 115(5):550–560, 2000.

[53] Snowden JM, Rose S et Mortimer KM : Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738, 2011.

[54] Snowden JM, Rose S et Mortimer KM : Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 3(1):33–72, 2014.

[55] Rao JNK : On making valid inferences by integrating data from surveys and other sources. *The Indian Journal of Statistics*, 2020.

[56] Coyle JR, Hejazi NS et van der Laan MJ : hal9001: The scalable highly adaptive lasso. *https://github.com/tlverse/hal9001*, 2020.

[57] Bykov K, Yoshida K, Weisskopf MG et Gagne JJ : Confounding of the association between statins and parkinson disease: systematic review and meta-analysis. *Pharmacoepidemiology and Drug Safety*, 26:294–300, 2017.

[58] Imai K et Ratkovic M : Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7:44–470, 2013.

[59] Wolter KM : *Introduction to variance estimation*. Springer series in Statistics, 2007.

[60] Chen KT : *Using LASSO to Calibrate Non-probability Samples using Probability Samples*. Dissertations and Theses (Ph.D. and Master's), 2016.

[61] Blais L, Beauchesne MF, Rey E, Malo JL et Forget A : Use of inhaled corticosteroids duing the first trimester of pregnancy and the risk of congenital malformations among women with asthmal. *Thorax*, 62:320–328, 2007.

[62] Breiman L : Random forests. *Thorax*, 45(1):5–32, 2007.

[63] Carone M, Diaz I et van der Laan MJ : Higher-order targeted minimum loss-based estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 331, 2014.

[64] Elliot M et Valliant R : Inference for non-probability samples. *Statistical Science*, 32:249–264, 2017.

[65] Pang M, Schuster T, Filion KB, Eberg M et Platt RW : Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*, 27:570–577, 2016.

[66] Pang M, Schuster T, Filion KB, Eberg M et Platt RW : Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*, 27(4):570–577, 2016.

[67] Yuan M et Lin Y : On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B*, 69:143–161, 2007.

[68] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J et Sturmer T : Variable selection for propensity score models. *American Journal of Epidemiology*, 163:1149–1156, 2006.

[69] Hernan MA et Robins JM : *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

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

[71] Schnitzer ME, Lok JJ et Gruber S : Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *International Journal of Biostatistics*, 12(1):97–115, 2016.

[72] van der Laan MJ : Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2013.

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

[74] van der Laan MJ, Polley EC et Hubbard AE : Super learner. *UC Berkeley Division of Biostatistics Working Paper Series*, 22, 2007.

[75] van der Laan M.J et Gruber S : Collaborative double robust targeted maximum likelihood estimation. *International Journal of Biostatistics*, 6(1), 2010.

[76] van der Laan M.J et Rose S : *Targeted learning: causal inference for observational and experimental data.* Springer Series in Statistics, Springer, 2011.

[77] COUPER MK : A review of issues and approaches. *Public Opinion Quarterly*, 64:464–494, 2000.

[78] Petersen ML, Porter KE, Gruber S, Wang Y et van der Laan MJ : Diagnosing and responding to violations in the positivity assumption. *Statistical Methods for Medical Research*, 21(1):31–54, 2012.

[79] Elliott MR et Valliant R : Inference for nonprobability samples. *Statistical Science*, 32(2):249–264, 2017.

[80] Bembom O et van der Laan MJ : Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 230, 2008.

[81] Zhao P et Yu B : On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[82] Robinson PM : Root-n-consistent semiparametric regression. *Econometrica*, 56:931–954, 1999.

[83] Rosenbaum PR et Rubin DB : The central role of the propensity score in observational studies for causal effects. *The Public Opinion Quarterly*, 70(1):41–55, 1983.

[84] Zhao Q, Small DS et Ertefaie A : Selective inference for effect modification via the lasso. *arXiv:1705.08020*, 2018.

[85] Zhao Q et Hastie T : Causal interpretations of black-box models. *Journal of Business and Economic Statistics*, 2019.

[86] Herrera R, Berger U, von Ehrenstein OS, Díaz I, Huber S, Muñoz DM et Radon K : Estimating the causal impact of proximity to gold and copper mines on respiratory diseases in chilean children: An application of targeted maximum likelihood estimation. *International Journal of Environmental Research and Public Health*, 2017.

[87] Pirracchio R, Petersen ML et van der Laan MJ : Improving propensity score estimators robustness to model misspecification using super learner. *American Journal of Epidemiology*, 18(2):108–119, 2015.

[88] Tibshirani R : Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288.

[89] Tibshirani R, Taylor J, Loftus J et Reid S : selectiveinference: Tools for post-selection inference. *https://CRAN.R-project.org/package=selectiveInference*, 2019.

[90] Valliant R et Dever JA : Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40:105–137, 2014.

[91] Farmer RE, Ford D, Mathur R, Chaturvedi N, Kaplan R, Smeeth L et Bhaskaran K : Metformin use and risk of cancer in patients with type 2 diabetes: a cohort study of primary care records using inverse probability weighting of marginal structural models. *American Journal of Epidemiology*, 48(2):527–537, 2019.

[92] Gruber S et van der Laan MJ : Targeted maximum likelihood estimation: A gentle introduction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 252, 2009.

[93] Gruber S et van der Laan MJ : An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *International Journal of Biostatistics*, 6(1)(18), 2010.

[94] Gruber S et van der Laan MJ : A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, 6(26), 2010.

[95] Gruber S et van der Laan MJ : tmle: An r package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(13), 2012.

[96] Lee S, Okui R et Zhang YJ : Doubly robust uniform confidence band for the conditionalaverage treatment effect function. *Journal of Applied Econometrics*, 32:1207–1225, 2017.

[97] Power S, Qian J, Jung K, Schuler A, Shah N, Hastie T et Tibshirani R : Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 20(37(11)):1767–1787, 2018.

[98] Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H et Brookhart MA : High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(512), 2009.

[99] Wager S et Athey S : Estimation and inference of heterogeneous treatment effects using random forests. *The Annals of Applied Statistics*, 112:1228–1242, 2018.

[100] Yang S : Integrativefpm r package. *https://github.com/shuyang1987 IntegrativeFPM/*, 2019.

[101] Yang S, Kim JK et Song R : Doubly robust inference when combining probability and non-probability samples with high-dimensional data. *Journal of the Royal Statistical Society: Series B*, 82(2):445–465, 2020.

[102] Lendle SD, Fireman B et van der Laan MJ : Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology*, 66:91–98, 2013.

[103] Lendle SD, Fireman B et van der Laan MJ : Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology*, 6:91–98, 2016.

[104] Shortreed SM et Ertefaie A : Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

[105] Vouri SM, Thai TM et Winterstein AG : An evaluation of co-use of chloroquine or hydroxychloroquine plus azithromycin on cardiac outcomes: A pharmacoepidemiological study to inform use during the covid19 pandemic. *Research in Social and Administrative Pharmacy*, 17(1):2012–2017, 2021.

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

[107] Cole SR et Hernán MA : Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–665, 2008.

[108] Hastie T, Tibshirani R et Friedman J : *The Elements of Statistical Learning*. Springer, 2008.

[109] VanderWeele TJ : Commentary: Resolutions of the birthweight paradox: competing explanations and analytical insights. *Internation Journal of Epidemiology*, 43:1368–1373, 2014.

[110] VanderWeele TJ et Shpitser I : A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413, 2007.

[111] VanderWeele TJ et Knol MJ : A tutorial on interaction. *Epidemiologic Methods*, 3(1):33–72, 2014.

[112] VanderWeele TL et Knol MJ : A tutorial on interaction. *Epidemiology*, 173(7):731–738, 2014.

[113] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W et Robins J : Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1–68, 2018.

[114] Semenova V et Chernozhukov V : Debiased machine learning of conditional average treatment effects and other causal functions. *Manuscript submitted to The Econometrics Journal*, pages 1–49, 2020.

[115] Lu W, Goldberg Y et Fine JP : On the robustness of the adaptive lasso to model misspecification. *Biometrika*, 9(3):771–731, 2012.

[116] Luo W, Wu W et Zhu Y : Learning heterogeneity in causal inference using sufficient dimension reduction. *Journal of Causal Inference*, 2019.

[117] Zheng W et van der Laan MJ : *Cross-Validated Targeted Minimum-Loss-Based Estimation in Targeted Learning: Causal Inference for Observational and Experimental Dataa*. Springer, 2011.

[118] Zheng W, Luo Z et van der Laan MJ : Marginal structural models with counterfactual effect modifiers. *International Journal of Biostatistics*, 4, 2014.

[119] Nie X et Wager S : Quasi-oracle estimation of heterogeneous treatment effects. *arXiv:1712.04912*, 2017.

[120] Meng XL : Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, 12:685–726, 2018.

[121] Chen Y, Li P et Wu C : Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 0(1):1–11, 2019.

[122] Liu Y, Schnitzer ME et Bahamyirou A : Yan2020729/bdt: Bootstrap diagnostic tool. *https://github.com/Yan2020729/bdt*, 2021.

[123] Zhao Y, Laber EB, Ning Y, Saha S et Sands B : Efficient augmentation and relaxation learning for individualized treatment rules using observational dat. *arXiv:1901.00663*, 2019.

[124] Tan Z : Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.