UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE MÉTODOS ESTATÍSTICOS

Inference on the Average Treatment Effect on the Treated from a Bayesian Perspective

Estelina Serrano de Marins Capistrano

Rio de Janeiro – RJ
2019

Inference on the Average Treatment Effect on the Treated from a Bayesian Perspective

Estelina Serrano de Marins Capistrano

Doctoral thesis submitted to the Graduate Program in Statistics of Universidade Federal do Rio de Janeiro as part of the requirements for the degree of Doctor in Statistics.

Advisors: Prof. Alexandra M. Schmidt, Ph.D.

Prof. Erica E. M. Moodie, Ph.D.

Rio de Janeiro, August 27th of 2019.

Inference on the Average Treatment Effect on the Treated from a Bayesian Perspective

Estelina Serrano de Marins Capistrano

Doctoral thesis submitted to the Graduate Program in Statistics of Universidade Federal do Rio de Janeiro as part of the requirements for the degree of Doctor in Statistics.

Approved by:

Prof. Alexandra M. Schmidt, Ph.D.
IM - UFRJ (Brazil) / McGill University (Canada) - Advisor

Prof. Erica E. M. Moodie, Ph.D.
McGill University (Canada) - Advisor

Prof. Hedibert F. Lopes, Ph.D.
INSPER - SP (Brazil)

Prof. Helio dos S. Migon, Ph.D.
IM - UFRJ (Brazil)

Prof. Mariane B. Alves, Ph.D.
IM - UFRJ (Brazil)

Rio de Janeiro, August 27th of 2019.

CIP - Catalogação na Publicação

C243i

Capistrano, Estelina Serrano de Marins
Inference on the Average Treatment Effect on the Treated from a Bayesian Perspective / Estelina Serrano de Marins Capistrano. -- Rio de Janeiro, 2019.
106 f.

Orientador: Alexandra Schmidt.
Coorientador: Erica Moodie.
Tese (doutorado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós Graduação em Estatística, 2019.

1. Bayesian Inference. 2. Causal Inference. 3. Inverse Probability Weighting. 4. Propensity Score. I. Schmidt, Alexandra , orient. II. Moodie, Erica, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

"Don't worry about failures, worry about the chance you miss when you don't even try."

Jack Canfield

Acknowledgments

First, I would like to thank my advisor, Dr. Alexandra Schmidt, for all the advice and guidance she has given me throughout the last ten years, which has certainly contributed to my professional development. I hope one day I can be half of the researcher she is.

I would like to provide a special thanks to my advisor, Dr. Erica Moodie, whom I have had the pleasure of working with throughout the last four years, for sharing her huge experience and providing encouragement during this research.

I thank Dr. Marina Klein and the researchers of the Canadian Co-Infection Cohort Study for kindly making the data analyzed in Chapter 3 available. I would like to extend my thanks to Dr. Klein for her collaboration and for sharing her clinical knowledge. Further, I thank Leo Wong for kindly providing assistance on the database and extracting data for this specific study.

Additionally, I would like to thank Dr. Olli Saarela for helpful discussions and suggestions about the implementation of the Bayesian bootstrap. I also thank the associate editor and two reviewers for comments that greatly improved the paper presented in Chapter 2.

I would also like to thank the committee members, Dr. Hedibert Lopes, Dr. Helio Migon, and Dr.Mariane Alves for accepting to participate of my thesis committee. In addition, I thank Dr. Fernando Moura for kindly accepting to be the substitute member.

I thank all the staff and colleague of the *Departamento de Métodos Estatísticos* at *Universidade Federal do Rio de Janeiro* for all assistance and support.

I would like to express sincere thanks to the Department of Epidemiology, Biostatistics and Occupational Health at McGill University for welcoming me during my Sandwich Doctorate period.

I acknowledge the financial support of *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq).

I would also like to thank my family and friends for always giving support and encouragement. In particular, I want to thank my parents and siblings. I cannot express how grateful I am for everything they have done for me. None of this would have been possible without their love, understanding and support. Thanks for bravely dealing with my absence, for enduring all stressful moments, and for encouraging me when I felt frustrated.

Finally, I thank God for never leaving me alone.

Abstract

Causal inference from a Bayesian perspective is a challenging issue. In joint likelihood-based approaches, information from the outcome model is used to estimate the parameters in the treatment assignment model, leading to bias due to this feedback. Although some approaches have been developed to address this problem, there is still space for further research and development of new techniques. In particular, this thesis addresses Bayesian propensity score methods; more precisely, weighting using a propensity score function.

Besides that, it is important to note that the current literature on causal inference generally focuses on estimating the average treatment effect. Although the average treatment effect on the treated is a useful and relevant estimand, research on its estimation is far less common in both the frequentist and Bayesian framework. Therefore, this research represents an important contribution to the Bayesian causal inference literature.

A Bayesian approach for estimating the average treatment effect on the treated is developed using propensity score-based weights to adjust for confounding, while avoiding the problem of feedback by considering a weighted likelihood bootstrap strategy to approximate posterior distributions of interest. Simulation studies investigate the impact of sample size and the strength of confounding on the estimation of causal effects. The proposed approach is then extended to accommodate a longitudinal outcome setting, considering a mixed linear model for the outcomes. Thus, the estimation of a time-varying average treatment effect on the treated becomes possible and the analysis of a relevant real data from the Canadian Co-infection Cohort Study is performed.

Additionally, we investigate the ability of the propensity score to provide covariate balance in a two-step Bayesian approach, considering different prior specifications for the parameters in the treatment models. The simulation study encourages Bayesian inference using non-informative prior distributions for coefficients in the treatment models, especially for small sample sizes.

Keywords: Bayesian inference; Causal inference; Inverse probability weighting; Propensity score.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter aims to introduce some key concepts for causal inference and describe how particular parameters that define average causal effects can be computed in conditionally randomized experiments. Furthermore, some conditions under which observational studies lead to valid causal inferences are described and the propensity score methods used to adjust for confounding are presented. Bayesian approaches using propensity score methods for estimating the average causal effect are reviewed.

1.1 Fundamental concepts

Consider Z a dummy variable representing treatment ($z = 1$: treated, $z = 0$: untreated) and Y a variable representing the observed outcome. Set $Y(z)$ as the potential (or counterfactual) outcome under the treatment value z, for $z = 0, 1$. That is, $Y(z)$ represents the outcome that would be obtained if the received treatment was $Z = z$. Thus, the observed outcome for the i-th unit can be expressed as $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$, for all $i = 1, \cdots, n$. Therefore, if $Z_i = z$, then $Y_i = Y_i(z)$; this correspondence is known as consistency.

Individual causal effects are defined as a contrast between the values of potential outcomes. Figure 1.1 was adapted from a table presented in Hernán and Robins (2019) and illustrates an example of the available information for n units, from which the first k units were not treated. Note that for each unit, only one of the potential outcomes will be known – the one that corresponds to the treatment value that was actually received, and therefore, individual causal effects can not be identified. Hence, causal inference has as its main aim to estimate the average causal effect, which can be computed by the difference (or ratio) between the averages of the potential outcomes between suitably balanced or equivalent samples.

Two average causal effects will be considered in this work. The average treatment effect in the population (ATE) is defined as $\Delta_{ATE} = E[Y(1) - Y(0)]$ and is the most commonly studied population causal effect. However, there may be situations in which the exposure of interest cannot be imposed or applied to all eligible units, for example drug abuse or unemployment (Austin and Stuart,

Unit	Z	Y	$Y(0)$	$Y(1)$
1	0	y_1	y_1	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	0	y_k	y_k	?
$k+1$	1	y_{k+1}	?	y_{k+1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	y_n	?	y_n

Figure 1.1: Example of dataset where the first k units are untreated ($Z = 0$) and the following $n-k$ units are treated ($Z = 1$). Values for the observed potential Y are measured. The potential outcome $Y(0)$ is known only for the first k units and is unknown for the other units. In contrast, the potential outcome $Y(1)$ is unknown for the untreated units and known for the last $n-k$ units.

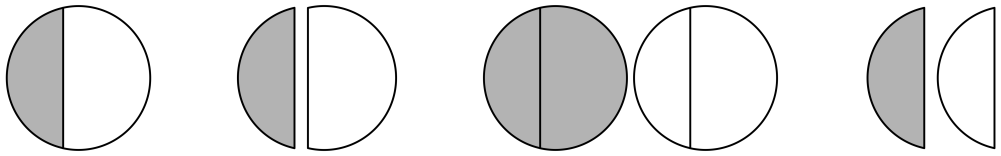
2017). In these contexts, it becomes more interesting to estimate the average treatment effect on the treated (ATT), defined by $\Delta_{ATT} = E[Y(1) - Y(0)|Z = 1]$.

In practice, information is collected only from the outcomes observed through the received treatment. These data are used to compute the expected value of the outcome among those units that received the treatment value z, that is, $E[Y|Z = z]$. Unadjusted averages can be used to obtain the associative effect by estimating $E[Y|Z = 1] - E[Y|Z = 0]$ via simple averages among, respectively, the treated and untreated units in a sample.

Causal effects are often defined in terms of marginal probabilities, where $Y(z)$ would be measured in the entire population. Thus, causal parameters often imply a contrast between the hypothetical situations "if the (entire) population was treated" and "if the (entire) population was not treated", whereas association implies a contrast between the treated and untreated units – which may differ in ways other than simple the treatment. That is, association involves contrasting two disjoint subsets of the population, estimating conditional probabilities for the units that received differing levels of the treatment $Z = z$.

Figure 1.2 was adapted from Hernán and Robins (2019) and depicts differences between causal and associative contrasts, and highlights the conceptual distinction between the ATE and ATT. The population, represented by a circle, is divided into treated (grey area) and untreated (white area). The definition of association implies a contrast between the grey and white areas of the original population. The concept of causality is here represented by the definitions of ATE and ATT. Note that ATE implies a contrast of whole circles, while ATT implies a contrast of the part corresponding to the gray area in the original circle.

An important concept in causal inference is exchangeability. If the treatment groups are exchangeable, then the joint probability distribution of the potential outcomes $Y(z)$ will be the same between the treated and untreated groups. As a consequence, the concept of mean exchangeability becomes



(a) Population (b) Association (c) ATE (d) ATT

Figure 1.2: (a) The population is divided into treated (grey area) and untreated (white area). (b) Association: contrast between treated and untreated units of the original population. (c) ATE: contrast between "if the entire population was treated" and "if the entire population was not treated". (d) ATT: contrast between the observed outcome of treated units and "if the treated units were not treated".

valid: $E[Y(z)|Z=1] = E[Y(z)|Z=0] = E[Y(z)]$, for $z = 0, 1$. That is, the treated and the untreated would have, on average, the same outcome if they had received the same treatment value z. Thus, exchangeability means that the potential outcomes and the received treatment are independent, denoted by $Y(z) \perp Z$, for every treatment value z. The concepts of consistency and mean exchangeability are sufficient to show that $E[Y|Z=z] = E[Y(z)|Z=z] = E[Y(z)]$. In ideal random experiments, randomization makes the potential outcomes jointly independent of the received treatment and, in the presence of exchangeability, the average causal effect can be consistently estimated since it is, mathematically, equal to the associative effect.

Now, suppose that, given a random variable X, we have a random experiment for each value of X. The experiment is said to be conditionally randomized because it uses several probabilities of randomization that depend on the value of X. It can be shown that randomization guarantees that the treatment groups are exchangeable in the subset of units with the same value of X. If this condition holds for all values of X, then conditional exchangeability becomes valid, that is, $Y(z) \perp Z \mid X$, for all z. Therefore, although randomization does not guarantee marginal exchangeability, conditional exchangeability is ensured within the levels of X. Thus, we can compute the average causal effect in each subset of the population defined by the values of X, since the expected value of the potential outcome in each subset equals to the expected value of the observed outcome conditioned to both the received treatment $Z = z$ and the value of X for that subset, that is, $E[Y(z)|X=x] = E[Y|Z=z, X=x]$.

Whether randomization is simple or conditional on covariates, causal inference is assured (assuming, further, no measurement error, simple random sampling from the population, and so on). However, there are many situations where randomized experiments are often unethical, impractical or intrusive (Rosenbaum and Rubin, 1983; Hernán and Robins, 2019), which forces researchers to perform observational studies. The next section will elucidate the assumptions needed for inference to be causal when data do not arise from a randomized trial.

1.2 Causal inference in observational studies

Causal inference from observational data is performed under the assumption that observational studies can be conceptualized as conditionally randomized experiments. For this, the treatment values must correspond to well-defined interventions and the following assumptions are required:

No unmeasured confounding: The conditional probability of receiving any treatment depends only on the measured covariates X. This assumption is equivalent to the definition of conditional exchangeability: $Y(z) \perp Z \mid X$.

Positivity: The conditional probability of receiving any treatment given X is greater than zero, that is, $P(Z = z|X = x) > 0, \forall z$ and $\forall x$ with $P(X = x) \neq 0$.

Together, these two assumptions are known as strong ignorability (Rosenbaum and Rubin, 1983). If either of these conditions do not hold, then one cannot make the analogy between observational studies and conditionally randomized experiments.

In large randomized experiments, all the measured covariates X are equally distributed between the treated and the untreated groups. (Indeed, all unmeasured covariates should also follow the same distribution in the treated and untreated groups.) The distribution of X is said to be balanced between the treatment groups, which implies independence of X and Z. On the other hand, when treatment is not randomly assigned, covariate balance is not guaranteed and the associative effect will have no causal interpretation if the treatment groups differ with respect to the covariates that affect the outcome.

In observational studies, the treatment assignment may be determined by many factors, and therefore, their distribution will vary between the treatment groups. If these factors also affect the outcome, then we say that there is confounding for the effect of treatment on the outcome. Confounding is defined as the bias that arises when the treatment and the outcome share a common cause. Because of lack of exchangeability of the treated and the untreated, the associative effect is not equal the causal effect. Therefore, it is necessary to look for methods to control or adjust for confounding. If adjustment is not made, the estimator of the average causal effect based on simple averages of outcomes in the treated and untreated units will be biased.

Any variable that can be used to help eliminate confounding is referred to as confounder. It is important to note that the observation of confounders X does not exclude the possibility that other unbalanced confounders, unknown to the researchers, remain unmeasured. To improve the justification for the assumption of conditional exchangeability, one must measure many relevant variables, so that the assumption of conditional exchangeability is (approximately) valid within the strata defined by the

combination of all these variables. With this assumption met, one collection of adjustment methods that can be used to remove the effects of confounding and make the treatment groups exchangeable are known as propensity score techniques.

1.3 Propensity score methods

Propensity score methods provide a powerful tool for balancing the distributions of measured covariates in different treatment groups, in order to estimate the causal effects that are not confounded by these covariates. The propensity score is defined as the conditional probability of treatment assignment $Z = 1$ given the measured covariates X (Rosenbaum and Rubin, 1983) and its use is justified by the fact that it has the balancing score property: let $e(X) := P(Z = 1|X)$ be the propensity score, then $Z \perp X \mid e(X)$. This property results in conditional independence between treatment and covariates, given the propensity score. That is, conditioned on the propensity score, the covariates X are balanced between the treated and untreated groups.

It can also be shown that the exchangeability of the treated and the untreated within the levels of the covariates X implies exchangeability within the levels of the propensity score $e(X)$. That is, conditional exchangeability $\{Y(1), Y(0)\} \perp Z \mid X$ implies $\{Y(1), Y(0)\} \perp Z \mid e(X)$. If X is sufficient to adjust for the confounding, then $e(X)$ also will be sufficient. Therefore, under strong ignorability, the propensity score can be used to adjust for confounding and estimate the average causal effects. Note that the use of the propensity score reduces the dimension of the problem of balancing the covariates X between treated and untreated groups to only one dimension.

Four different propensity score methods are commonly used for removing the effects of confounding when estimating the causal effects: propensity score matching, propensity score stratification (or subclassification), regression adjustment using the propensity score as a derived covariate, and weighting – most commonly inverse probability of treatment weighting (IPTW) – using a function of the propensity score. The first three methods were introduced by Rosenbaum and Rubin (1983) and the last one by Rosenbaum (1987). In the following, we briefly describe these propensity score approaches, highlighting the advantages and disadvantages of each. For more details, see Austin (2011a); Hernán and Robins (2019); Deb et al. (2016).

1.3.1 Propensity score matching

In propensity score matching, treated and untreated units with exact or similar propensity scores are matched. The ideal matching aims to create pairs that minimize the total difference in propensity score. Several different criteria can be adopted to define closeness in propensity score matching, such

as nearest neighbor, caliper distance (typically Euclidean distance measured on the propensity score itself); additional choices to be made include whether to match with or without replacement, and whether to perform pair matching or many-to-one or unequal-sized group matching. When matching each treated unit with at least one untreated unit and excluding the unmatched untreated, one is estimating the ATT (Hernán and Robins, 2019). However, if the interest is in estimating the ATE, one also needs to match each untreated unit with at least one treated unit.

In addition to being superior at reducing bias, matching may be more robust to misspecification of the propensity score model (Deb et al., 2016). A further advantage is that the outcomes between paired units are directly compared. On the other hand, matching requires many units. Indeed, when estimating the ATT, many untreated units are required to ensure that all treated units will be paired and thus avoid a significant loss of observations due to unmatched treated units. However, when estimating the ATE, a similar problem can arise if there are few treated units. This disadvantage not only implies information loss but may also introduce selection bias. Nevertheless, propensity score matching is very popular in practice and many applications of this method for different matching procedures have been made. For examples, see Rosenbaum and Rubin (1983, 1985); Imbens (2004); Abadie and Imbens (2008); Austin (2009); Austin and Stuart (2015b, 2017).

1.3.2 Propensity score stratification

Propensity score stratification divides treated and untreated units into mutually exclusive subgroups (strata) according to the propensity scores. The units usually are ranked and stratified into deciles or quintiles based on values of the propensity score, and then estimates of the causal effect are obtained in each stratum. If we are interested in the average causal effect in the entire population, the treatment effects across the subgroups can be computed as a simple or inverse-variance weighted average of the stratum-specific causal effects.

Although the bias reduction is smaller in propensity score stratification as compared to the other approaches discussed in this chapter, the technique may be more robust to slight misspecification of the propensity score model (Deb et al., 2016) and uses all the available units. In the case where one of the treatment groups is small, this technique may result in subgroups with few treated (or control) units and hence produce unstable estimates. Demonstrations of this method are given in Rosenbaum and Rubin (1983, 1984); Imbens (2004); Lunceford and Davidian (2004).

1.3.3 Covariate adjustment using propensity score

Propensity scores can also be used as a covariate in regression models to adjust for confounding (Rosenbaum and Rubin, 1983). The choice of regression model depends on the nature of the outcome: for continuous outcome, a linear model is typically employed; for dichotomous outcome, a logistic regression model may be selected. Although covariate adjustment uses all the available units, this method requires the correct specification of a regression model for the relationship between the outcome and the propensity score. A further challenge of this approach is that it is more difficult to assess the balance between treatment groups than it is in other adjustment methods. The method was employed by, for example, Rosenbaum and Rubin (1983); McCandless et al. (2009, 2010); Austin (2011a); Zigler et al. (2013).

1.3.4 Inverse probability of treatment weighting

Inverse probability of treatment weighting consists of creating a hypothetical population in which the received treatment is independent of the measured covariates; thus, the treatment groups in this pseudo-population are exchangeable. Once exchangeability holds, the causal effect can be found through the associative effect in the pseudo-population. In other words, the associative effect in the pseudo-population equals the causal effect in both the pseudo-population and the original population.

Like matching, inverse probability weighting is superior at reducing bias compared with stratification and covariate adjustment. Further, this method uses all available data, however it may be inefficient as large weights can cause instability in the estimates. A further disadvantage is that misspecification of the propensity score model will may lead to bias. Examples of the use of weighting are given in Hirano and Imbens (2001); Lunceford and Davidian (2004); Kaplan and Chen (2012); Saarela et al. (2015a); Austin (2016); Austin and Stuart (2017).

Essentially, the approach assigns a specific weight to each individual in the population. The weights used for causal adjustment vary according to the causal effect of interest. For instance, the weights when estimating the ATE are defined as the inverse of the conditional probability of receiving the treatment that was actually assigned $Z = z$ given the measured covariates, that is, $\omega^z = 1/P(Z = z|X)$ for $z = 0, 1$; these are known as inverse probability of treatment weights. Thus, for treated units, ω^1 corresponds to the inverse of the propensity score; for untreated units, ω^0 is equivalent to the inverse of 1 minus the propensity score. Note that individuals who are unlikely to be treated (whose propensity score is very close to zero) may receive a large weight, leading to unstable estimates. An improvement is achieved by using stabilized weights.

The stabilized weights to estimate the ATE are $\omega_s^z = \dfrac{P(Z=z)}{P(Z=z|X)}$, while for estimating the ATT, the stabilized weights are $\omega_s^z = \dfrac{P(Z=z)}{P(Z=z|X)} \dfrac{P(Z=1|X)}{P(Z=1)}$ for $z=0,1$. Furthermore, it can be demonstrated that the use of stabilized weights ω_s^z in observational studies will result in a pseudo-population of same size as the original population, while the weights ω^z essentially will double the size of the population (Xu et al., 2010). These weights are summarized in Table 1.1.

Estimands	Weight for treated ($Z=1$)	Weight for untreated ($Z=0$)
ATE	$\omega^1 = \dfrac{1}{P(Z=1\|X)}$	$\omega^0 = \dfrac{1}{P(Z=0\|X)}$
	$\omega_s^1 = \dfrac{P(Z=1)}{P(Z=1\|X)}$	$\omega_s^0 = \dfrac{P(Z=0)}{P(Z=0\|X)}$
ATT	$\omega_s^1 = 1$	$\omega_s^0 = \dfrac{P(Z=0)}{P(Z=0\|X)} \dfrac{P(Z=1\|X)}{P(Z=1)}$

Table 1.1: Propensity-score based weights, by treatment group, according to the causal effect of interest (ATE or ATT).

From a classical point of view, due to the nature of the problem, the estimation of the causal effects using the propensity score methods is often performed in two sequential procedures. First, the propensity score is estimated for each unit from the measured covariates through the treatment model such as $g_Z\{E(Z|X)\} = X\alpha$, where $g_Z(\cdot)$ is a link function (other forms of estimation, e.g. trees, could also be employed). The logistic model is usually used given that the treatment is dichotomous.

Second, the outcomes of treated and untreated units that had similar estimated propensity score values are compared using one of the propensity score methods outlined above. Suppose that an outcome model fitting only the propensity score is given by: $g_Y\{E(Y|Z,X)\} = \beta_0 + \lambda Z + \xi h(\alpha, X)$, where $g_Y(\cdot)$ is a link function and $h(\alpha, X) = X\alpha$, for example. The main objective is to estimate the causal effect of Z on Y, that is, the interest consists on the estimation of the conditional parameter λ. Thus, the treatment effect estimates are computed without considering the uncertainty in the propensity score estimation.

1.4 Bayesian propensity score approaches

Bayesian causal inference has been progressing in recent years. Some Bayesian propensity score methods have been developed, and other alternative approaches using non-parametric models or Bayesian g-formula (e.g. Xu et al. (2018); Roy et al. (2018); Keil et al. (2018)) have also been proposed. However, there remain gaps in the literature, particularly regarding the estimation of ATT. In

the following, the literature on Bayesian propensity score approaches is reviewed. Briefly, we conclude that many of the proposed joint estimation methods are not true propensity score adjustment methods in the sense that they do not retain the balancing property of propensity scores (Saarela et al., 2015a).

McCandless et al. (2009) provided an approach to Bayesian propensity score analysis and studied the impact of uncertainty modeling on propensity scores using stratification. Their approach considers the propensity score as a latent variable, and the authors model the joint likelihood of propensity score and outcome simultaneously via a Markov chain Monte Carlo (MCMC) algorithm. Nevertheless, the proposed approach did not consider an important feature in this context: the feedback. Unlike sequential procedures, which estimate the propensity score using only information on how covariates are associated with treatment, the joint estimation uses information from the outcome model to estimate the propensity score. The feedback from the outcome model arises because the parameter α appears in both likelihood terms, which leads to posterior samples of α involving both the propensity score and outcome model, and which does not reflect the true mechanism of allocation of treatment (data generation), so that the resulting $h(\alpha, X)$ does not provide the desired balancing property. Thus, the Bayesian estimation of the propensity score does not guarantee estimates of λ that reflect the causal effect of treatment. As noted by McCandless et al. (2010), it is important to make it clear that feedback is not inherent to Bayesian modeling and would arise in any analysis that uses the combined likelihood. McCandless et al. (2010) introduced a technique to "cut" the feedback in regression adjustment for the propensity score. They used a sample from the posterior distribution of the propensity score as a covariate in the regression model for the outcome to estimate the treatment effect. The method is called "approximately Bayesian" because it does not involve joint estimation of both models. In this case, the posterior update is not done from the posterior full conditional of α but rather from an approximate conditional distribution that involves only the propensity score model and a prior distribution for α.

Alternatively, Kaplan and Chen (2012) developed a two-step Bayesian approach to propensity score analysis that incorporates information about prior distributions into both the propensity score model and the outcome model. This approach avoids the feedback problem, but it does not provide a mechanism for variance estimation for the point estimator, since the marginal quasi-posterior distribution does not necessarily correspond to a well-defined joint posterior distribution. Therefore, the proposed approach could be not considered as a proper Bayesian method and it does not result in good frequentist properties. Kaplan and Chen (2014) provided a Bayesian model averaging approach via Markov chain Monte Carlo procedure to account for model uncertainty and examined the differences in causal estimates when incorporating informative priors in a model averaging stage.

Covariate balance analysis show that Bayesian model averaging approach provide comparably good covariate balance compared to the two-step Bayesian propensity score approach. Chen and Kaplan (2015) explicitly studied covariate balance in the two-step Bayesian propensity score approach proposed by Kaplan and Chen (2012) using standardized mean differences and variance ratios, and found the propensity score approaches substantially reduce the initial imbalance. They also investigated the use of non-informative priors versus informative priors centered on the true data-generating values, using an approach that samples from the posterior distribution of the propensity score to approximate the posterior distribution of the treatment effect. However, there remains a lack of clarity of the impact of prior specification on the propensity score in Bayesian methods as both Kaplan and Chen (2014) and Chen and Kaplan (2015) considered moderately large samples, and the approach of drawing from the posterior of the propensity scores has since been shown to have poor small sample performance (Saarela et al., 2015a), as covariate balance must be achieved within the observed sample and thus is best achieved by fixing propensity scores to their best estimates (Rosenbaum and Rubin, 1983).

Additionally, Zigler et al. (2013) proposed a strategy based on the increase of the outcome model with the additional adjustment of individual covariates. That is, $g_Y\{E(Y|Z,X)\} = \beta_0 + \lambda Z + \beta X^- + \xi h(\alpha, X)$, with $\beta \neq 0$, where X^- can not included all covariates in X to avoid multicollinearity. Although this modeling also suffers from feedback, the authors have shown that the estimation of α respects the treatment assignment mechanism, maintaining the property of the balancing score and accurately estimating the treatment effect. The authors recommend that the outcome model with propensity score adjustment should be augmented with adjustment for every covariate that appears in the propensity score model. However, this strategy may not be attractive because it requires further parametric specifications and can be inefficient in high dimensions, no longer having the propensity score's advantage of reducing a large covariate vector to a single dimension. Zigler and Dominici (2014) proposed methods for Bayesian model averaging to address uncertainty in the propensity score model specification, and also evaluated the covariate balance in this context.

In summary, such proposed approaches either use a parameterization were the treatment assignment and outcome models become dependent, losing the balancing property of the propensity score, or cut the feedback making inference procedures no longer Bayesian. Another approach to compute the average treatment effect involves marginal structural models, which are models for potential outcomes. The parameter of interest in a marginal structural model corresponds to the average causal effect of a longitudinal sequence of treatments. Specifically, inverse probability of treatment weighting can be used to construct a balanced pseudo-population so that a simple, unadjusted model can be fit. Saarela et al. (2015a) developed a methodology for Bayesian estimation of marginal structural models, which justifies the use of inverse probability of treatment through an importance sampling

argument. The approach suggests that weights should be derived from the posterior predictive probability of treatment assignment. Here, the posterior marginal density of the parameters of the treatment model does not depend on the parameters of the outcome model, and therefore, there is no feedback.

Following this approach, causal inference is seen as a problem of posterior prediction where the estimated inverse probability of treatment weights work as importance sampling weights in predicting the outcome in a pseudo-population. Taking the weighted likelihood bootstrap strategy (Newton and Raftery, 1994) and considering the log-likelihood function of the outcome model as the utility function, we obtain the weighted maximum likelihood estimator. Therefore, the algorithm includes two sequential steps, carried out within a Bayesian bootstrap paradigm: (a) fitting the treatment assignment model and obtaining the propensity score-based weights; (b) obtaining a sample from the posterior distribution of the causal parameters by estimating the parameters of the outcome model using a weighted likelihood.

Compared with ATE estimation, few studies involving ATT estimation are found in the current literature of propensity scores methods. To the best of our knowledge, there is not a Bayesian propensity score approach to estimate the average treatment effect on the treated. Because of that, in this thesis, we focus on making Bayesian inference for the ATT, although the methodology developed can be extended to other causal effects. Results from ATE estimation are presented in the appendices for comparison.

1.5 An overview of the thesis

The thesis is organized as follows: Chapter 2 develops a Bayesian approach for estimating ATT using importance sampling weights to adjust for confounding and the weighted likelihood bootstrap to approximate posterior distributions of interest. Simulation studies evaluate the impact of sample sizes and the strength of confounding on causal effect estimation. The proposed approach is illustrated through applications in two real datasets. This chapter is published under the title 'Bayesian estimation of the average treatment effect on the treated using inverse weighting' by the authors Estelina S. M. Capistrano, Erica E. M. Moodie and Alexandra M. Schmidt in the journal *Statistics in Medicine*, 2019, volume 38, pages 2447-2466.

Chapter 3 analyzes data from the Canadian Co-infection Cohort Study in order to evaluate the effect of successful hepatitis C treatment on lipid profiles. For this purpose, we propose a longitudinal Bayesian approach to estimation of ATT using mixed linear models. This paper is the first to propose a time-dependent ATT. This work, titled 'Successful hepatitis C treatment leads to improved lipid profiles: A longitudinal Bayesian analysis of the average treatment effect on the treated', by authors

Estelina S. M. Capistrano, Erica E. M. Moodie, Alexandra M. Schmidt and Marina B. Klein will soon be submitted for publication.

In Chapter 4, we investigate the ability of the propensity score to provide covariate balance in a two-step Bayesian approach via inverse probability of treatment weighting, considering the impact of sample sizes and different sets of prior distributions. Unlike Chen and Kaplan (2015), weights are obtained from the posterior mean of the parameters of propensity score models. Simulations show that informative prior distributions should be specified carefully to ensure covariate balance, especially for small sample sizes. The manuscript titled 'Can a Bayesian be subjective about propensity score balancing?' by authors Estelina S. M. Capistrano, Alexandra M. Schmidt and Erica E. M. Moodie will soon be submitted for publication.

Finally, Chapter 5 concludes the thesis with a brief discussion and presents our main ideas for future works.

Chapter 2

Bayesian estimation of the average treatment effect on the treated using inverse weighting

Abstract

We develop a Bayesian approach to estimate the average treatment effect on the treated in the presence of confounding. The approach builds on developments proposed by Saarela et al. in the context of marginal structural models, using importance sampling weights to adjust for confounding and estimate a causal effect. The Bayesian bootstrap is adopted to approximate posterior distributions of interest and avoid the issue of feedback that arises in Bayesian causal estimation relying on a joint likelihood. We present results from simulation studies to estimate the average treatment effect on the treated, evaluating the impact of sample size and the strength of confounding on estimation. We illustrate our approach using the classic Right Heart Catheterization dataset, and find a negative causal effect of the exposure on 30-days survival, in accordance with previous analyses of these data. We also apply our approach to the U.S. National Center for Health Statistics Birth dataset, and obtain a negative effect of maternal smoking during pregnancy on birth weight.

Keywords: Bayesian inference; causal inference; inverse probability weighting; observational study; propensity score.

2.1 Introduction

Observational data present challenging questions, especially when causal inferences are the goal of the analysis. Confounding arises when the distribution of covariates that precede and affect both the exposure and the outcome are imbalanced between the two exposure groups. Without adequate adjustment for confounding, bias in the estimation of a treatment effect can arise. Several methods of adjustment can be used; propensity score techniques (Rosenbaum and Rubin, 1983) are widely employed in the observational data context because they decrease the task of covariate balancing to

only one dimension, and offer simplicity of use. In what follows, the terms exposure and treatment shall be used interchangeably.

Propensity score methods are now familiar in statistics and epidemiology, but they have seen relatively little uptake in the Bayesian literature. In the frequentist literature, propensity score analyses are typically performed as two-step procedures, first estimating the parameters of the propensity score then using the estimated exposure probabilities to weight, match, stratify, or as a covariate. In contrast, Bayesian procedures typically perform estimation and inference by considering a joint likelihood function for all parameters of interest (in this case, the exposure and outcome), so that the uncertainty in modeling of the propensity score is propagated into the estimation of the treatment effect. While there has been relatively little literature on Bayesian propensity score approaches to causal inference, that literature has called into question the traditional joint modelling approach; see Rubin (1978) and Zigler (2016) for a review.

McCandless et al. (2009) proposed a Bayesian propensity score analysis using stratification on the propensity score in which the joint likelihood of the propensity score and the outcome were simultaneously modeled via a Markov chain Monte Carlo algorithm. However, the proposed approach did not consider an important feature in the combined likelihood context: the feedback problem, where the joint estimation of the propensity score uses information from the outcome model to estimate the propensity score and does not reflect the true mechanism of allocation of treatment. This feedback can distort the propensity score and reduce its ability to adjust for confounding. To cut feedback, McCandless et al. (2010) later used a sample from the posterior distribution of the propensity score as a covariate in the outcome regression model and updated the parameters of treatment model from an approximate conditional distribution that ignores the likelihood contribution from the outcome.

Kaplan and Chen (2012) developed a two-step Bayesian approach to propensity score analysis that incorporates information about prior distributions into both the propensity score model and the outcome model. Although this approach does not create feedback, it was later shown to be problematic. First, it does not provide a mechanism for variance estimation for the point estimator. The authors suggested a variance estimation approach that assumes the joint quasi-posterior distribution is a well-defined posterior distribution. However, the product of the posterior distributions does not necessarily correspond to a well-defined joint posterior, except in the case when the outcome model is correctly specified, in which case it does not depend on the propensity score. Thus, this particular two-step approach is neither proper Bayesian nor does it result in good frequentist properties. Second, it was shown that the point estimator does not possess good small sample properties (Saarela et al., 2016). Kaplan and Chen (2014) also examined the differences in the causal estimate when incorporating non-informative versus informative priors in a model averaging stage. Zigler et al. (2013) developed

a Bayesian strategy which augments the outcome model with additional individual covariates. While this modeling does allow feedback from the outcome to the exposure model, the authors have shown that the approach maintains the desired balancing property. However, the augmented outcome model with adjustment for every covariate that appears in the propensity score model may not be attractive because this requires further parametric specifications (no longer offering the nice, one-dimensional advantage of propensity score modelling) and can be inefficient in high dimensions.

Recently, Saarela et al. (2015a) proposed a framework that addresses the problem of feedback while providing a fully Bayesian justification for estimation of the average treatment effect (ATE) in a longitudinal setting. Specifically, they developed a methodology for Bayesian estimation of marginal structure models, in which they justified the use of inverse probability of treatment weighting through an importance sampling argument. The inverse probability of treatment weighting approach was used to construct a balanced pseudo-population in which there is no confounding by the covariates, in an approach similar to frequentist inverse probability weighted estimation but which – in contrast to the other approaches noted above – was a proper Bayesian procedure. In this approach, the posterior marginal density of the parameters of the treatment model does not depend on the parameters of the outcome model, and therefore, there is no feedback between the two components of the model. In the rejoinder to a discussion of their proposal, Saarela et al. (2015b) suggested re-estimating weights in predicted resamples via a Bayesian bootstrap approach. In this way, the estimation procedure accounts for the variance reduction due to sample balance obtained through estimation of the treatment weights, similar to a frequentist non-parametric bootstrap approach. Following this approach, Bayesian causal inference can be seen as a problem of posterior prediction where the estimated inverse probability of treatment weights work as importance sampling weights in predicting the outcome in a pseudo-population in which treatment is randomized.

Thus, while progress on Bayesian causal inference has been made in the last few years (including alternatives to propensity score approaches, which we touch upon in section 2.6), there remain lacunae in methodology. In particular, the average treatment effect on the treated (ATT) is a parameter that has seen almost no mention in the Bayesian analyses of medical and epidemiological data; within the frequentist epidemiological literature, estimation has been predominantly via propensity score matching (Austin, 2008, 2009, 2011b; Austin and Stuart, 2015b) with only rare exceptions (e.g. Austin (2016)). However, matching is a highly non-smooth approach, which can be viewed as weighting all individuals with either a 0 or 1. The non-smoothness can lead to poor asymptotic behavior. For instance, standard bootstrap procedures cannot be used to appropriately account for the variability of the resulting estimator (Abadie and Imbens, 2008), although there is an extensive list of papers using matching with bootstrapped standard errors (see Abadie and Imbens (2008)). Thus,

weighting is a very attractive alternative for ATT estimation. Note that, unlike in epidemiology and medicine, estimation of the ATT via various (frequentist) weighting procedures is not so uncommon in econometrics (Hahn, 1998; Rothe and Firpo, 2013; Shinozaki and Matsuyama, 2015).

In this paper, we develop a Bayesian approach to estimate the average treatment effect on the treated that circumvents the disadvantages of matching and retains the attractive properties of Bayesian estimators such as the interpretability of credible intervals and flexibility of posterior probability calculations. The approach relies on importance sampling so as to estimate the parameters of the marginal structural models using inverse probability weighting, extending the ideas of Saarela et al. (2015a). The paper is organized as follows: In Section 2.2, we briefly review propensity scores and associated concepts. In Section 2.3, we introduce the proposed methodology and describe how to estimate the causal effect based on a weighted likelihood bootstrap approach. In Section 2.4, we perform two simulation studies to compare the performance of the Bayesian ATT estimator with a frequentist approach, and in Section 2.5, we apply the Bayesian ATT estimator to the Right Heart Catheterization and U.S. National Center for Health Statistics Birth datasets. Finally, Section 2.6 concludes with a brief discussion.

2.2 Background: notation and concepts

Consider Z an indicator for treatment ($z = 1$: treated, $z = 0$: untreated) and Y a variable representing the (observable) outcome. Let $Y(z)$ denote the potential (or counterfactual) outcome under the treatment value z, for $z = 0, 1$. That is, $Y(z)$ represents the outcome that would be obtained if the received treatment was $Z = z$. Thus, the observable outcome for the i-th subject can be expressed as $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$, for all $i = 1, \cdots, n$.

In observational studies, the treatment assignment may be determined by many factors, and therefore, the distribution of those factors will vary between the treatment groups. If these factors also affect the outcome, then we say that there is confounding of the effect of treatment on the outcome. Confounding is defined as the bias that arises when the treatment and the outcome share a common cause, which results in the associational effect not equalling the causal effect. Therefore, it is necessary to look for methods to control or adjust for confounding.

Causal inference from observational data is performed under the assumption that observational studies can be conceptualized as conditionally randomized experiments, that is, under the following assumptions:

(i) the treatment values correspond to well-defined interventions;

(ii) the conditional probability of receiving any treatment depends only on the measured covariates

X;

(iii) the conditional probability of receiving any treatment is greater than zero.

The second condition is often referred to as no unmeasured confounding and is equivalent to the definition of conditional independence: $Y(z) \perp Z|X$. The third condition is called positivity: $P(Z = z|X = x) > 0, \forall z$ and $\forall x$ with $P(X = x) \neq 0$. Together, they are known as strong ignorability (Rosenbaum and Rubin, 1983; Rubin, 2008). The first condition is required for the other two conditions to be defined.

Individual causal effects are defined as a contrast between individual potential outcome, however these are not, in general, identifiable. Hence, causal inference focuses more often on the *average causal effect*, which can be computed as the difference (or ratio) between the averages of the potential outcomes. Two average causal effects are often considered. The average treatment effect in the population is defined as $\Delta_{ATE} = E[Y(1) - Y(0)]$, whereas the average treatment effect on the treated is given by $\Delta_{ATT} = E[Y(1) - Y(0)|Z = 1]$. This latter quantity may be of interest when studying, for example, the effect of a particular pharmaceutical treatment such as an antidepressant, where there are many indications for use including not only depression but also as a sleep aid, or to treat anxiety, bulimia, and a number of other conditions. Another context in which the ATT may be more relevant than the ATE is when the exposure of interest is harmful and not all of the population would be exposed to it (e.g. drug abuse or unemployment) (Austin and Stuart, 2017).

The propensity score provides a powerful tool for balancing the distributions of measured covariates in different treatment groups, in order to eliminate confounding and to estimate causal effects. The propensity score is defined by Rosenbaum and Rubin (1983) as the conditional probability of treatment assignment $Z = 1$ given the p-dimensional vector of measured covariates X. It has been shown that independence of the treated and the untreated within the levels of the covariates X implies independence within the levels of the propensity score $e(X)$ (Rosenbaum and Rubin, 1983), thereby reducing a p-dimensional balancing problem to a single dimension. Let us denote the propensity score as $e_i = e(X_i) = P(Z_i = 1|X_i)$, for $i = 1, \cdots, n$. Under the assumptions (i)-(iii), the propensity score can be used to adjust for confounding and estimate the average causal effects. Deb et al. (2016) summarize the main propensity score methods, highlighting the advantages and disadvantages of each. See also Hernán and Robins (2019).

In propensity score matching, treated and untreated units with similar propensity scores are matched. Several different criteria can be adopted to define closeness in propensity score matching, such as nearest neighbor or caliper distance, and matching can be performed with or without replacement, one-to-one or one-to-many. It has been suggested that matching may be more robust to misspecification of the propensity score model than other propensity score methods, however it may require many untreated individuals, may exclude many unmatched treated individuals, and – as noted above – the

estimator's asymptotic properties are non-standard (Austin, 2008, 2009, 2011b; Austin and Stuart, 2017). Propensity scores can also be used for stratification (Deb et al., 2016; Rosenbaum and Rubin, 1984) (a course form of group matching) and regression by treatment the propensity score (or some flexible function of the score) as a covariate (McCandless et al., 2009, 2010).

Inverse probability of treatment weighting, the primary estimator for marginal structural models for the ATE, creates a pseudo-population in which the distribution of measured covariates is independent of treatment assignment. The method uses the propensity score to compute weights for each individual. Like matching, it is effective at reducing bias, though unlike matching, inverse weighting uses all available data. This approach has been applied in a Bayesian context for estimating the ATE (Kaplan and Chen, 2012; Saarela et al., 2015a,b).

2.3 A Bayesian weighting estimator of the ATT

2.3.1 A model for the ATT

Suppose as above we have a dichotomous treatment, and let us assume that, conditioned on measured covariates X_i, the distribution of Z_i conditional on X_i follows a Bernoulli distribution with mean e_i, that is, $Z_i \mid X_i \sim Bernoulli(e_i)$. Therefore, we assume the propensity score follows a logistic model

$$\text{logit}(e_i) = \log\left(\frac{e_i}{1-e_i}\right) = \alpha' X_i, \tag{2.3.1}$$

where $X_i = (X_{i1}, \cdots, X_{ip})'$ is a p-dimensional vector of covariates associated with the i-th unit, and the coefficients $\alpha = (\alpha_1, \cdots, \alpha_p)'$ are assumed to be unknown.

If the outcome is continuous, we may assume a conditional model for the potential outcome such that

$$Y_i(z) = \beta' X_i + \lambda' V_i z + \varepsilon_i, \tag{2.3.2}$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ are independent, and identically distributed for $i = 1, 2, \cdots, n$, and V_i is a q-dimensional vector containing a subset of the components of X_i, with both X_i and V_i containing a leading column of ones. Therefore, the difference in potential outcomes under this model can be expressed as: $E[Y_i(1) - Y_i(0) | X_i, \beta, \lambda, \sigma_\varepsilon^2] = \lambda' V_i$, where the coefficients $\lambda = (\lambda_1, \cdots, \lambda_q)'$ are unknown. It follows from the properties of conditional expectation that the average treatment effect on the treated is:

$$\Delta_{ATT} = E_{X|Z=1}\left\{E_{Y|X,Z=1}[Y(1) - Y(0) | X, Z=1, \beta, \lambda, \sigma_\varepsilon^2] | Z=1\right\} = E_{X|Z=1}[\lambda' V | Z=1]. \tag{2.3.3}$$

That is, the ATT is simply the average of the product of the "treatment effect" (parameters λ times their associated covariates), where that average is over the covariate distribution among the *treated* members of the population. Note that the assumption that the outcome is linear in the covariates X is not necessary to the exposition that follows; rather, we assume model 2.3.2 for convenience.

To derive a Bayesian estimator of Δ_{ATT}, we propose an approach motivated by Saarela et al. (2015b) who derived a procedure based on observational and experimental measures for exchangeable observable sequences via a predictive modelling view based on de Finetti's representation (De Finetti, 1974; Saarela et al., 2015a); see Appendix 2.7.1 for details. Saarela et al. borrow ideas from Walker (2010), considering the inference procedure through the lens of a decision analysis framework. In this view, the inferential procedure is based on the maximization of a posterior predictive expected utility function (Saarela et al., 2015b), operationalized through the weighted likelihood bootstrap proposed by Newton and Raftery (1994). Note that the formulation of causal inference as a posterior predictive problem closely resembles the original Bayesian approach by Rubin (1978). We shall, therefore, briefly review the (Bayesian) weighted likelihood bootstrap before describing our proposed approach to obtain samples from the posterior distribution of Δ_{ATT}.

2.3.2 A brief review of the weighted likelihood bootstrap

The bootstrap was proposed by Efron (1979) as a means of generating a sampling distribution of statistics, which can be used, for instance, to estimate the variability of these statistics. When a set of observations $u = (u_1, \cdots, u_n)$ is assumed to be from an independent and identically distributed population, the bootstrap can be constructed by sampling with replacement from the observed dataset. Thus, one bootstrap replication is a simple random sample of size n taken with replacement from the original dataset u. The bootstrap distribution is generated by taking randomly repeated replications from u.

The Bayesian bootstrap was introduced by Rubin (1981) as a natural Bayesian analogue of Efron's (frequentist) bootstrap. Denote the probability of sampling u_i from u by p_i for $i = 1, \cdots, n$. Let f_i be the proportion of times that u_i is drawn in a bootstrap replication, that is, $\hat{p}_i = f_i$ can assume values in the discrete set $\{0, 1/n, \cdots, n/n\}$. Instead of drawing weights from the discrete set, the Bayesian bootstrap approach treats the vector of weights $p = (p_1, \cdots, p_n)$ as unknown parameters and derives a posterior distribution for p.

Let $b = (b_1, \cdots, b_n)$ denote a particular realization of a bootstrap replication; i.e. b is a sample of size n drawn from the original data u with replacement. Note that b can be viewed as a random sample of a categorical variable B that assumes value u with associated probability vector p, that is,

$P(B = u_i|p) = p_i$, for $i = 1, \cdots, n$. Denote by m_i the number of times that the observation u_i appears in a bootstrap replication b such that $\sum_{i=1}^{n} m_i = n$. Rubin (1981) used an improper prior distribution $\pi(p_i) \propto p_i^{-1}$ for each weight p_i that, when combined with the likelihood function $l(p;b) = \prod_{i=1}^{n} p_i^{m_i}$, leads to a Dirichlet distribution with parameters $(m_1, \cdots, m_n)$ as the posterior distribution of p.

If $m_1 = \cdots = m_n = 1$, then the points are uniformly distributed and the uniform Dirichlet weights are obtained from $Dirichlet(1, \cdots, 1)$. Another Bayesian alternative is to use the non-informative prior distribution $\pi(p) \propto 1$ for the vector of weights p. This implies that $p = (p_1, \cdots, p_n)$ is obtained by drawing counts $\xi = (\xi_1, \cdots, \xi_n)$ from a Multinomial distribution with n trials, and probabilities $(n^{-1}, \cdots, n^{-1})$. Then they are normalized to obtain the weights ξ_i/n. (For readers more familiar with the frequentist non-parametric bootstrap, note that the resampling undertaken in the frequentist bootstrap is closely related to Multinomial sampling from the original data.)

However, the interpretation of the resulting distribution is different between the frequentist and Bayesian bootstraps. The Bayesian bootstrap simulates draws from the posterior distribution of the parameter (i.e. the parameter is a random quantity), whereas the frequentist bootstrap simulates an estimated sampling distribution of the statistic of interest (the estimator is the random quantity, with variability due to the sampling procedure). Although there are differences in interpretation, operationally, the frequentist and Bayesian bootstraps differ primarily in how the weights p are obtained; both methods result in a very similar distribution of the weights when n is sufficiently large.

Newton and Raftery (1994) extended the Bayesian bootstrap of Rubin from nonparametric to parametric models and named it the weighted likelihood bootstrap. The idea is as follows: assume that the observed dataset u is generated from a distribution indexed by an unknown parameter ψ. Then, once the data are observed, the likelihood function $l(\psi;u) = \prod_{i=1}^{n} f(u_i|\psi)$ and the maximum likelihood estimate $\hat{\psi}$ are fixed. They proposed to obtain a sample from the posterior distribution of ψ by maximizing a weighted likelihood function $\tilde{l}(\psi;u) = \prod_{i=1}^{n} f(u_i|\psi)^{p_i}$ wherein the weights $p = (p_1, \cdots, p_n)$ have some probability distribution. The authors claimed that the weighted likelihood $\tilde{l}(\psi;u)$ has randomness induced by the distribution of the weights, which is derived through a Bayesian bootstrap procedure.

2.3.3 Proposed Bayesian estimation of the ATT

Following Saarela et al. (2015a,b) and drawing on the de Finetti representation (De Finetti, 1974) of the problem (Appendix 2.7.1), we view estimation of ATT as a decision problem which requires maximizing an expected utility function. Let θ be the parameter vector associated with the outcome model, (y, x, z) be the observed data, and $d_k^* = (y_k^*, x_k^*, z_k^*)$ be hypothetical predictions obtained under

a conceptual (completely) randomized version of the treatment assignment, denoted by $\mathscr{E}$, where the treatment assignment does not depend on the observed covariates. Also, let $P_{\mathscr{E}}(d_k^*|x,y,z)$ and $P_{\mathscr{O}}(d_k^*|y,x,z)$ be the posterior predictive distributions under the experimental setting $\mathscr{E}$, and under the observed (or observational) data generating mechanism $\mathscr{O}$, respectively. Note that the only difference between these settings is our assumption regarding the treatment: under the experimental setting, the distribution of Z is independent of X, whereas under the observational setting, the distribution of Z is conditional on X. However, the marginal distribution of X and the conditional distribution of $Y|X,Z$ are identical under $\mathscr{E}$ and $\mathscr{O}$. Note that the parameters that govern the propensity score model are assumed to be independent of θ, *a priori*. Following the decision analysis theory, and letting $U(\theta;d_k^*)$ be a utility function, our goal is to obtain θ that maximizes the following conditional expectation:

$$E_{\mathscr{E}}[U(\theta;d_k^*)|y,x,z] = \int_{d_k^*} U(\theta;d_k^*)\, P_{\mathscr{E}}(d_k^*|y,x,z)\, \partial d_k^* = \int_{d_k^*} U(\theta;d_k^*)\, \omega_k^*\, P_{\mathscr{O}}(d_k^*|y,x,z)\, \partial d_k^*, \quad (2.3.4)$$

where the weights are defined to be $\omega_k^* = P_{\mathscr{E}}(d_k^*|y,x,z)/P_{\mathscr{O}}(d_k^*|y,x,z)$. The second equality above follows from the connection between the predictive distributions under the experimental and observational representations. This connection arises through the importance sampling identity (Røysland, 2011; Saarela et al., 2016).

The Bayesian estimator for θ is obtained by considering the log-likelihood function $L(\theta;d_k^*) = \log P(y_k^*|z_k^*,x_k^*,\theta)$ as the utility function in equation (2.3.4). Following Walker (2010) and Saarela et al. (2015a), we adopt a weighted likelihood bootstrap strategy such that $P_{\mathscr{O}}(d_k^*|y,x,z)$ is represented as a non-parametric posterior predictive density. That is, $P_n(d_k^*) = \sum_{i=1}^n p_i\, \delta_{d_i}(d_k^*)$, where $p = (p_1,\cdots,p_n)$ are sampling weights following a uniform Dirichlet distribution, i.e. $p \sim Dirichlet(1,\cdots,1)$, and $\delta_{d_i}(d_k^*)$ represents a density function with point mass $1/n$ at each of the observed data points $d_i = (y_i,x_i,z_i)$, for $i = 1,\cdots,n$. Then, the expectation in equation (2.3.4) can be approximated by

$$E[L(\theta;d_k^*)|y,x,z] \approx \int_{d_k^*} \log P(y_k^*|z_k^*,x_k^*,\theta)\, \omega_k^* \sum_{i=1}^n p_i\, \delta_{d_i}(d_k^*)\, \partial d_k^* = \sum_{i=1}^n \omega_i\, p_i\, \log P(y_i|z_i,x_i,\theta), \quad (2.3.5)$$

where $\omega_i = P_{\mathscr{E}}(y_i,x_i,z_i)/P_{\mathscr{O}}(y_i,x_i,z_i)$. As noted above, we assume that $P_{\mathscr{E}}(y|z,x) = P_{\mathscr{O}}(y|z,x)$ and $P_{\mathscr{E}}(x) = P_{\mathscr{O}}(x)$. However, as the treatment assignment does not depend on covariates under $\mathscr{E}$, then $P_{\mathscr{E}}(z|x) = P_{\mathscr{E}}(z) \neq P_{\mathscr{O}}(z|x)$. Therefore, ω_i simplifies to a ratio between marginal and conditional treatment models: $\omega_i = P_{\mathscr{E}}(z_i)/P_{\mathscr{O}}(z_i|x_i)$.

In practice, ω_i in equation (2.3.5) is substituted by a ratio of modelled probabilities $P(z_i|\gamma)/P(z_i|x_i;\alpha)$, where γ and α are the parameters associated with the marginal and conditional treatment models, re-

spectively. We then further substitute estimates of the parameters in the marginal and conditional treatment models, say $\widehat{\omega}_i = P(z_i|\widehat{\gamma})/P(z_i|x_i,\widehat{\alpha})$. The parameter γ of the marginal treatment model can be estimated using the weighted likelihood bootstrap, that is,

$$\widehat{\gamma} = \arg\max_{\gamma} E[L(\gamma; d_k^*)|y,x,z] \approx \arg\max_{\gamma} \left[\sum_{i=1}^{n} p_i \log P(z_i|\gamma)\right]. \tag{2.3.6}$$

Similarly, the parameter vector α of the conditional treatment model is estimated through

$$\widehat{\alpha} = \arg\max_{\alpha} E[L(\alpha; d_k^*)|y,x,z] \approx \arg\max_{\alpha} \left[\sum_{i=1}^{n} p_i \log P(z_i|x_i,\alpha)\right]. \tag{2.3.7}$$

Having obtained weights $\widehat{\omega}_i$, we may now compute the average treatment effect on the population. As shown in Appendix 2.7.2, the average treatment effect on the treated should be computed using stabilized weights $\widehat{\omega}_i = \frac{P(Z=z_i|\widehat{\gamma})}{P(Z=z_i|x_i,\widehat{\alpha})}\frac{P(Z=1|x_i,\widehat{\alpha})}{P(Z=1|\widehat{\gamma})}$. That is, all the treated individuals ($z_i = 1$) will receive a weight of $\widehat{\omega}_i^1 = 1$ and all the unexposed individuals ($z_i = 0$) will receive a weight of $\widehat{\omega}_i^0 = \frac{P(Z=0|\widehat{\gamma})}{P(Z=0|x_i,\widehat{\alpha})}\frac{P(Z=1|x_i,\widehat{\alpha})}{P(Z=1|\widehat{\gamma})}$. Thus, using the weighted likelihood estimates of $\widehat{\omega}_i$, the importance sampling weighted maximum likelihood estimator of θ is obtained as

$$\widehat{\theta} = \arg\max_{\theta} E[L(\theta; d_k^*)|y,x,z] \approx \arg\max_{\theta} \left[\sum_{i=1}^{n} \widehat{\omega}_i\, p_i\, \log P(y_i|z_i,x_i,\theta)\right]. \tag{2.3.8}$$

As Δ_{ATT} is a function of θ, an estimate of Δ_{ATT} may then be obtained for using samples from the posterior distribution of θ through the Bayesian bootstrap. That is, at iteration l of the bootstrap $\Delta_{ATT}^{(l)} = \mathbb{P}[\lambda'^{(l)}V|Z=1]$, where $\mathbb{P}$ denotes the empirical average, taken in this case over those individuals in the sample for whom $Z = 1$ and V is a subset of the covariates in X. A 'point estimate' $\widehat{\Delta}_{ATT}$ is then obtained by taking the mean of its posterior distribution, as approximated by the Bayesian bootstrap. In summary, in the l-th ($l = 1,2,\cdots,L$) replication of the weighted likelihood bootstrap, a sample from the posterior distribution of Δ_{ATT} is obtained as follows:

(a) Sample weights $p^{(l)} = (p_1^{(l)}, p_2^{(l)}, \cdots, p_n^{(l)})$ from the $Dirichlet(1,1,\cdots,1)$ distribution;

(b) Following equations (2.3.6) and (2.3.7), the weighted maximum likelihood estimates of the coefficients from the marginal and conditional treatment models are obtained, respectively, as

$$\widehat{\gamma}^{(l)} = \arg\max_{\gamma} \left[\sum_{i=1}^{n} p_i^{(l)} \left[z_i \log(e^*) + (1-z_i)\log(1-e^*)\right]\right], \text{ where } e^* = \frac{1}{1+\exp(-\gamma)}$$

and

$$\widehat{\alpha}^{(l)} = \arg\max_{\alpha} \left[\sum_{i=1}^{n} p_i^{(l)} \left[z_i \log(e_i) + (1 - z_i) \log(1 - e_i) \right] \right], \text{ where } e_i = \frac{1}{1 + \exp(-\alpha' x_i)};$$

(c) Compute the importance sampling weights, $\widehat{\omega}_i^{(l)} = \dfrac{P(Z = z_i | \widehat{\gamma}^{(l)})}{P(Z = z_i | x_i, \widehat{\alpha}^{(l)})} \dfrac{P(Z = 1 | x_i, \widehat{\alpha}^{(l)})}{P(Z = 1 | \widehat{\gamma}^{(l)})}$;

(d) Following equation (2.3.8), the weighted maximum likelihood estimate of the parameter vector $\theta = (\beta', \lambda', \sigma_\varepsilon^2)'$ in the outcome model is obtained as

$$\widehat{\theta}^{(l)} = \arg\max_{\theta} \left[\sum_{i=1}^{n} \widehat{\omega}_i^{(l)} \; p_i^{(l)} \; \left[-\frac{1}{2\sigma_\varepsilon^2} \left(y_i - \lambda' v_i \right)^2 \right] \right];$$

(e) Compute $\widehat{\Delta}_{ATT}^{(l)}$ as

$$\widehat{\Delta}_{ATT}^{(l)} = \frac{\sum_{i=1}^{n} p_i^{(l)} \widehat{\lambda}^{(l)'} v_i z_i}{\sum_{i=1}^{n} p_i^{(l)} z_i}.$$

Therefore, a Bayesian estimate for the posterior distribution of the average treatment effect on the treated is given by the collection of weighted likelihood estimates $\widehat{\Delta}_{ATT}^{(l)}$. A point estimate can be obtained, for example, by considering the posterior mean of the distribution of Δ_{ATT}, that is, $\widehat{\Delta}_{ATT} = \frac{1}{L} \sum_{l=1}^{L} \widehat{\Delta}_{ATT}^{(l)}$.

The proposed approach is easily adapted to accommodate alternative distributions for the outcome model. For example, if the outcome is dichotomous, then $Y_i \mid X_i \sim Bernoulli(a_i)$, and $\text{logit}(a_i) = \beta' X_i + \lambda' Z_i V_i$ and, as before, V_i is a q-dimensional subset of X_i. Then the average treatment effect on the treated is now computed as

$$\begin{aligned} \Delta_{ATT} &= E_{X|Z}\{E_{Y|X,Z}[Y(1)|X, Z = 1, \beta, \lambda] - E_{Y|X,Z}[Y(0)|X, Z = 1, \beta, \lambda] | Z = 1\} \\ &= E_{X|Z}\{E_{Y|X,Z}[Y|X, Z = 1, \beta, \lambda] - E_{Y|X,Z}[Y|X, Z = 0, \beta, \lambda] | Z = 1\} \\ &= E_{X|Z}[\mu_1 - \mu_0 | Z = 1], \end{aligned} \tag{2.3.9}$$

where $\mu_1 = 1/[1 + \exp\{-(\beta' X + \lambda' V)\}]$ and $\mu_0 = 1/[1 + \exp\{-\beta' X\}]$. Considering equation (2.3.9), because of the nonlinearity of the mean function, the average treatment effect on the treated can be estimated as $\widehat{\Delta}_{ATT} = \frac{1}{L} \sum_{l=1}^{L} \widehat{\Delta}_{ATT}^{(l)}$, where $\widehat{\Delta}_{ATT}^{(l)} = \dfrac{\sum_{i=1}^{n} p_i^{(l)} \left(\widehat{\mu}_{1i}^{(l)} - \widehat{\mu}_{0i}^{(l)} \right) z_i}{\sum_{i=1}^{n} p_i^{(l)} z_i}$, $\widehat{\mu}_{1i}^{(l)} = 1/[1 + \exp\{-(\widehat{\beta}^{(l)'} x_i + \widehat{\lambda}^{(l)'} v_i)\}]$ and $\widehat{\mu}_{0i}^{(l)} = 1/[1 + \exp\{-\widehat{\beta}^{(l)'} x_i\}]$.

The computational algorithm for our approach is described in Figure 2.1. Note that when the outcome is not continuous, a generalized linear model may be used in place of the linear model in estimating $\widehat{\theta}$ in the algorithm and the ATT is computed as a difference in the predicted outcomes under treated and untreated, averaged over the empirical distribution of covariates in the treated. Further, modelling of the outcome y can be allowed to depend on Z, V, and their interactions, and may (but need not) additionally include all of X as shown in equation (2.3.2).

For l in $(1,\cdots,L)\{$

- $p^{(l)} = (p_1^{(l)}, p_2^{(l)}, \cdots, p_n^{(l)}) \leftarrow Dirichlet(1,1,\cdots,1)$;
- $\widehat{\gamma}^{(l)} \leftarrow glm(z \sim 1, weights = p^{(l)}, family = binomial)$
- $\widehat{\alpha}^{(l)} \leftarrow glm(z \sim x, weights = p^{(l)}, family = binomial)$
- $\widehat{\omega}_i^{(l)} \leftarrow \dfrac{P(Z = z_i|\widehat{\gamma}^{(l)})}{P(Z = z_i|x_i, \widehat{\alpha}^{(l)})} \dfrac{P(Z = 1|x_i, \widehat{\alpha}^{(l)})}{P(Z = 1|\widehat{\gamma}^{(l)})}$;
- $\widehat{\theta}^{(l)} \leftarrow lm(y \sim z * v, weights = \widehat{\omega}^{(l)} p^{(l)})$
- $\widehat{\Delta}_{ATT}^{(l)} \leftarrow \dfrac{\sum_{i=1}^n p_i^{(l)} \widehat{\lambda}^{(l)'} v_i z_i}{\sum_{i=1}^n p_i^{(l)} z_i}$.

$\}$

Figure 2.1: Computational algorithm to perform the weighted likelihood bootstrap when estimating the average treatment effect on the treated (ATT) in a Bayesian paradigm via importance sampling.

As described in Subsection 2.3.2, an alternative version of the Bayesian bootstrap algorithm is obtained by replacing p_i by ξ_i/n, where the vector $\xi = (\xi_1, \cdots, \xi_n)$ is generated from a Multinomial distribution, that is, $\xi = (\xi_1, \cdots, \xi_n) \sim Multinomial(n; n^{-1}, \cdots, n^{-1})$. In this case, in the first step of the algorithm shown in Figure 2.1, for each iteration l, we obtain $p^{(l)} = \xi^{(l)}/n$, where $\xi^{(l)} = (\xi_1^{(l)}, \cdots, \xi_n^{(l)})$ is generated from $Multinomial(n; n^{-1}, \cdots, n^{-1})$.

2.4 Simulation studies

2.4.1 A simple artificial dataset

The artificial datasets were generated from the following conditional treatment model: $Z_i \sim Bern(e_i)$, where $logit(e_i) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i}$, with $X_1 \sim N(1,1)$ and $X_2 \sim N(0,1)$, where $N(a,b)$

denotes a normal distribution with mean a and variance b. Then we assume that the outcome Y_i follows a normal distribution with mean μ_i, and variance equal to 0.4, where $\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \lambda_0 Z_i + \lambda_1 X_{1i} Z_i$. The analyses that follow are based on 1000 artificial datasets, each of size n.

In order to obtain the true average treatment effect on the treated, $\Delta_{ATT} = \lambda_0 + \lambda_1 E[X_1|Z=1]$, we need to compute $E[X_1|Z=1]$. Although $P(Z=1|X)$ cannot be computed analytically, the true ATT can be estimated through a very large sample of covariates $X = (X_1, X_2)$ (e.g. $N = 10,000,000$) which can be used to estimate the probability $P(Z=1|X)$ with very high precision.

In the following subsections we perform two sets of simulations studies. The aim is to explore the impact of the strength of confounding and the sample size on the estimation of ATT, and compare results between the Bayesian and frequentist estimators. In order to estimate Δ_{ATT}, we consider three different inferential methods: the Bayesian bootstrap under the Dirichlet and Multinomial sampling distributions and the frequentist bootstrap. In all three approaches, the propensity score is modelled as a function of both X_1 and X_2, while the outcome is modelled as a function of X_1, Z, and their interaction. In the frequentist bootstrap, we sampled individuals randomly with replacement from the original dataset and performed the bootstrap only to obtain an estimate of the variability of the estimator. We report the point estimates $\widehat{\Delta}_{ATT}$, the bias relative (RB) to the true value of Δ_{ATT}, the bias of the standard error estimates relative (SE RB) to the Monte Carlo standard deviation, and the 95% credible or confidence interval coverage probabilities (95% CP) as appropriate.

Study 1.A: The impact of the strength of confounding on ATT estimators

In this study, we fix the sample size at $n = 1000$ and explore different values of the coefficients of the confounders in both treatment and outcome models. To generate the data, we assume $\alpha = (0.5, 0.8, \alpha_2)$, $\beta = (2.0, 0.4, \beta_2)$, $\lambda = (2.0, 1.5)$. We consider $\alpha_2 = -1.0$ or -0.4 (strong or moderate confounding, respectively) and $\beta_2 = -0.6, -0.2$, or 0.0 (strong, moderate, or no confounding, respectively). All the other parameters were fixed at the same values. When $\alpha_2 = -1.0$, the true value of the ATT is $\Delta_{ATT} = 3.753$. When $\alpha_2 = -0.4$, the true value of the ATT is $\Delta_{ATT} = 3.759$. We rely on $L = 1000$ iterations of the bootstrap for all methods.

Figure 2.2 shows the relative bias of the point estimates and standard errors, and 95% coverage probabilities (columns) under different values of α_2 (rows) and different β_2 (x-axis in each panel). The relative biases are similar across the three approaches, regardless of the value of α_2 and β_2. All approaches result in similar absolute values (magnitudes) of the relative bias when $\alpha_2 = -1.0$ or $\alpha_2 = -0.4$. When $\alpha_2 = -0.4$, the Bayesian bootstrap with Multinomial sampling and frequentist approaches produce similar results for the SE relative bias, however the Bayesian bootstrap with Multinomial sampling has greater relative bias of standard errors when $\alpha_2 = -1$. All approaches

provide similar coverage probabilities of credible/confidence intervals, regardless of the values of α_2 and β_2. Coverage is close to the nominal level for all settings.

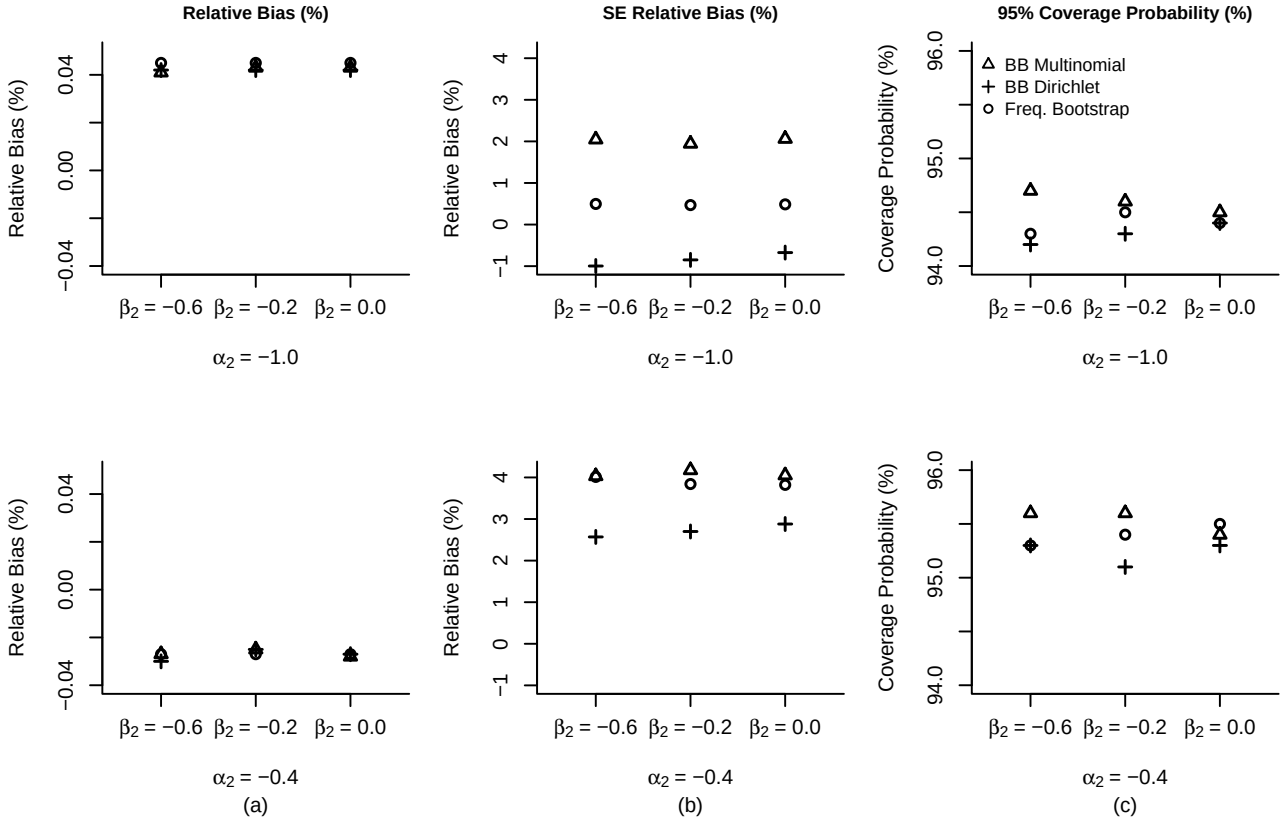


Figure 2.2: ATT estimation for a simple data generation: (a) bias of point estimates relative to the true value of Δ_{ATT}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for different strengths of confounding. Panels in the top row consider stronger confounding, those in the bottom row consider moderate confounding.

Study 1.B: The impact of the sample size on ATT estimators

Here we consider datasets with different sample sizes, n=100, 250, 500, 750, 1000, 2500, 5000, 7500 and 10,000. To generate the data, we assume $\alpha = (0.5, 0.8, -1.0)$, $\beta = (2.0, 0.4, -0.6)$, $\lambda = (2.0, 1.5)$, corresponding to the most extreme scenario (in terms of confounding) presented in Study 1.A. Under this specification, the true value of the ATT is $\Delta_{ATT} = 3.753$. We consider $L = 1000$ iterations of the bootstrap for all three methods.

The results are shown in Figure 2.3. Panel (a) reports the relative bias (in %) of the point estimate, which is small from $n = 750$ onwards for all methods. Panel (b) shows the bias of the standard error estimates relative to the standard deviation (in %), the magnitudes decrease as the sample size increases. Finally, panel (c) shows the 95% credible/confidence interval coverage probabilities. As

expected, the coverage probabilities increases with the sample size. The different methods show very similar coverage probabilities, although for most cases, the Bayesian bootstrap using Multinomial sampling produces slightly better results, specially for smaller sample sizes.

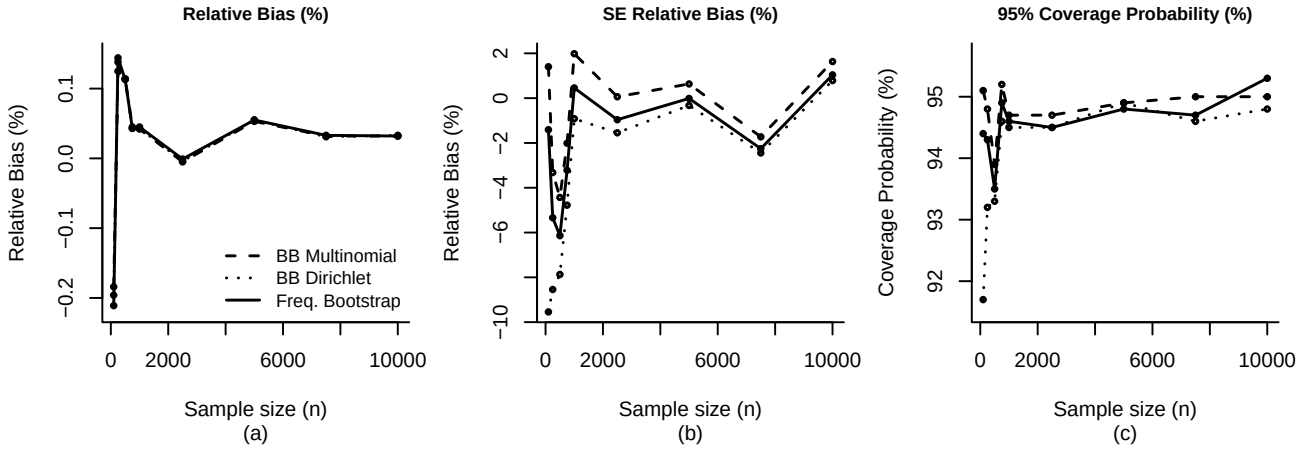


Figure 2.3: ATT estimation for a simple data generation: (a) bias of point estimates relative to the true value of Δ_{ATT}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for varying sample sizes. Results are obtained under the frequentist estimation (solid line), the Bayesian bootstrap (BB) with Multinomial (dashed line), and Dirichlet (dotted line) distributions.

Numerical results related to both studies are shown in Tables 2.2 and 2.3 in Appendix 2.7.3. Section 2.8.1 of the Supplementary Materials shows the results of a similar study performed to estimate the average treatment effect on the population (Δ_{ATE}).

2.4.2 A more complex simulated dataset

Following Kaplan and Chen (2014) and Zigler (2016), we now consider a more complex structure for the treatment model. More specifically, we assume $Z_i \sim Bern(e_i)$, where $logit(e_i) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 |X_{4i}| + \alpha_5 \exp(X_{5i}) + \alpha_6 X_{6i} + \alpha_7 X_{6i}^2$, where $X_1 \sim N(1,1)$, $X_2 \sim Poisson(2)$, $X_3 \sim Bernoulli(0.5)$, $X_4 \sim N(0,1)$, $X_5 \sim N(1,1)$, $X_6 \sim N(0,1)$, and $|X_4|$ represents the absolute value of X_4. We assume that the outcome Y_i follows a normal distribution with mean μ_i, and variance equal to 0.4, where $\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \lambda_0 Z_i + \lambda_1 X_{1i} Z_i + \lambda_2 X_{2i} Z_i$.

Study 2.A: The impact of the strength of confounding on ATT estimators

In this study, we consider 1000 artificial datasets each with sample size $n = 1000$ and explore different values of the coefficients of the confounders in both treatment and outcome models. To generate the data, we assume that $\alpha = (1.0, -1.0, 0.5, \alpha_3, 0.3, -0.1, 1.0, -0.3)$, $\beta = (2.0, -1.0, 0.5, \beta_3)$, $\lambda = (1.6, -0.8, 0.6)$. We consider $\alpha_3 = -1.0$ or -0.5 (strong or moderate confounding, respectively)

and $\beta_3 = -0.8, -0.4$, or 0.0 (strong, moderate, or no confounding, respectively). All the other parameters were fixed at the same values. When $\alpha_3 = -1.0$, the true value of the ATT is $\Delta_{ATT} = 2.626$. When $\alpha_3 = -0.5$, the true value of the ATT is $\Delta_{ATT} = 2.5$. Again, we rely on $L = 1000$ iterations of the bootstrap for all three methods.

Figure 2.4 shows the relative bias of the point estimate and standard error, as well as coverage probabilities of 95% credible or confidence intervals (columns) under each value of α_3 (rows), and different values of β_3 (x-axis in each panel). The values of the relative bias are similar across the three methods, regardless of the values of α_2 and β_2. The relative bias and the SE relative bias (in absolute value) are slightly greater when the effect of the covariate X_3 in the treatment model is moderate ($\alpha_3 = -0.5$). In this setting, the Bayesian bootstrap with Multinomial sampling and frequentist bootstrap provide good estimates for the standard error. The best performances for the coverage probabilities were observed in the Bayesian bootstrap with Multinomial sampling and frequentist approaches.

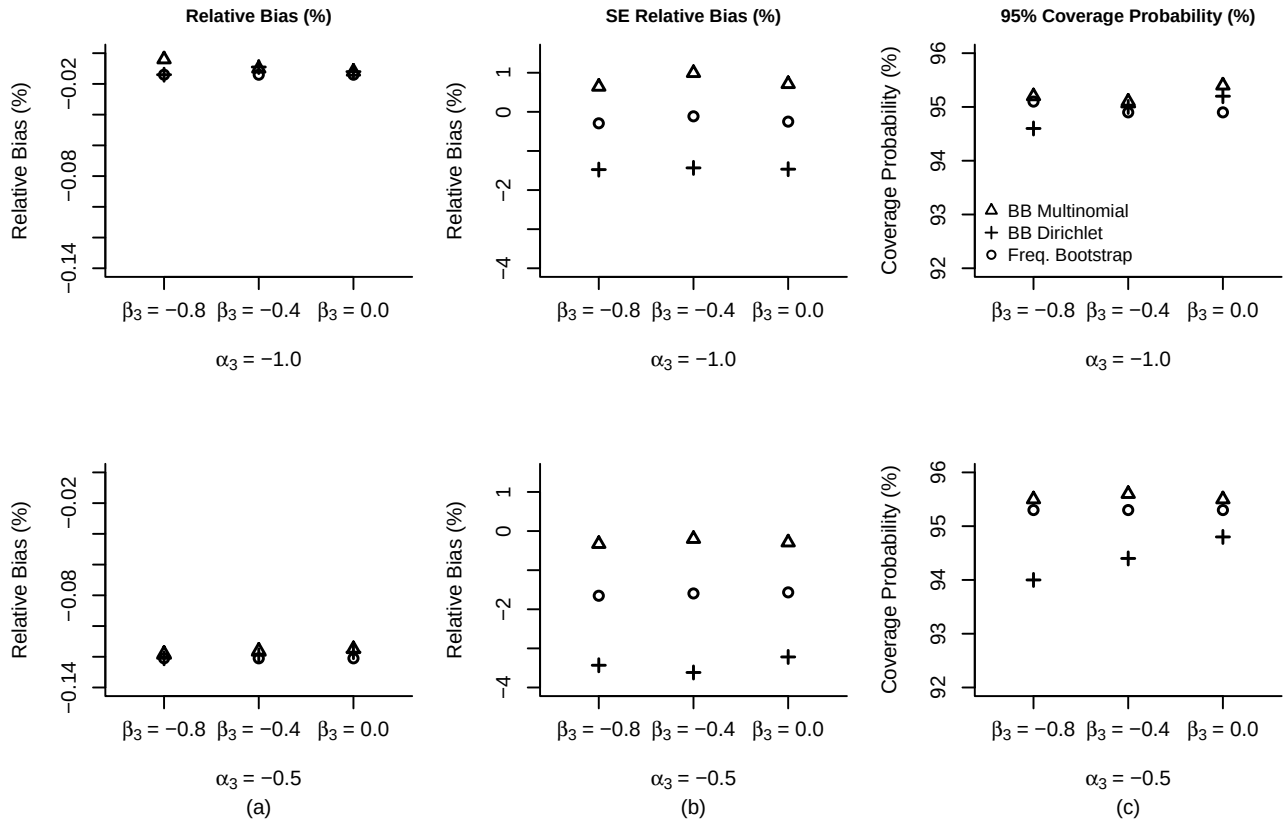


Figure 2.4: ATT estimation for a complex data generation: (a) bias of point estimates relative to the true value of Δ_{ATT}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for different strengths of confounding. Panels in the top row consider stronger confounding, those in the bottom row consider moderate confounding.

Study 2.B: The impact of the sample size on ATT estimators

Now we generate 1000 artificial datasets, considering different sample sizes, n=100, 250, 500, 750, 1000, 2500, 5000, 7500 and 10,000. We assume $\alpha = (1.0, -1.0, 0.5, -0.5, 0.3, -0.1, 1.0, -0.3)$, $\beta = (2.0, -1.0, 0.5, -0.4)$, $\lambda = (1.6, -0.8, 0.6)$. Under this specification, the true average treatment effect on the treated, $\Delta_{ATT} = \lambda_0 + \lambda_1 E[X_1|Z=1] + \lambda_2 E[X_2|Z=1] = 2.5$, was estimated using a very large sample of covariates ($N = 10,000,000$). The study aims to investigate the impact of sample size on the estimation of ATT in the presence of a more complex mean structure in the treatment model. Again, we consider $L = 1000$ iterations of the bootstrap for all three methods.

The results are shown Figure 2.5. Panel (a) reports the relative bias of the point estimates (in %). The relative bias quickly becomes negligible as n increases. Panel (b) shows the relative bias of the standard error estimate (in %), which exhibits the smallest magnitude with the Bayesian bootstrap using Multinomial sampling. Finally, panel (c) shows the coverage probabilities under each sample size n, where there are no notable differences between the three approaches as n increases.

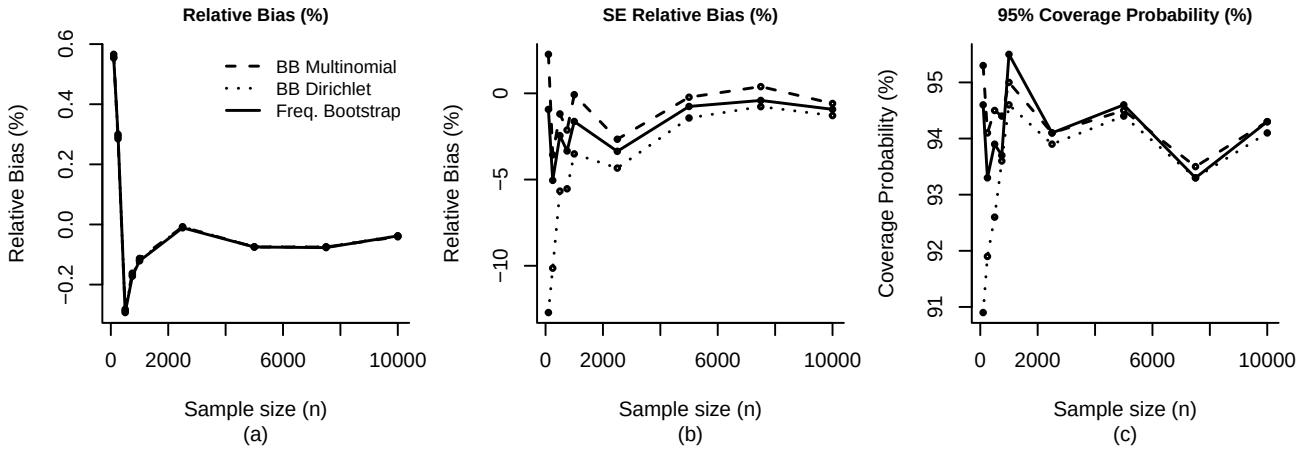


Figure 2.5: ATT estimation for a complex data generation: (a) bias of point estimates relative to the true value of Δ_{ATT}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for varying sample sizes. Results are obtained under the frequentist estimation (solid line), the Bayesian bootstrap (BB) with Multinomial (dashed line), and Dirichlet (dotted line) distributions.

Additional numerical results are provided in Tables 2.4 and 2.5 in Appendix 2.7.3. Section 2.8.1 of the Supplementary Materials gives the results of similar studies of the average treatment effect on the population, Δ_{ATE}.

2.5 Real data analysis

2.5.1 Analysis of the Right Heart Catheterization dataset

We illustrate our proposed approach using data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). SUPPORT collected data on hospitalized adult patients at five medical centers in the U.S., including information about many variables relating to the decision to perform the Right Heart Catheterization (RHC). RHC is a diagnostic procedure used for critically ill patients; its effectiveness in an observational setting was studied by Connors et al. (1996) and since it was first re-analyzed by Hirano and Imbens (2001), it has become a benchmark dataset for causal analyses.

The dataset contains information about 5735 individuals, where 2184 are treated and 3551 are controls. The treatment status is defined as 1 if RHC was performed within 24 hours of admission, and 0 otherwise. The outcome is an indicator of 30 day survival. The dataset also contains information on 72 covariates; we follow De Luna et al. (2011) and consider only 19 of these covariates both in the treatment and outcome models. The variables are listed in Table 2.10 in Section 2.8.2 of the Supplementary Materials; code is provided in Section 2.8.3. The outcome model also included interactions terms between the treatment and each of the following covariates: age, estimate of probability of surviving 2 months, the Glasgow coma score, and transfer more than 24 hours from another hospital, as well as the main effects of these covariates. We obtain estimates of Δ_{ATT} under the Bayesian bootstrap, using both the Dirichlet and the Multinomial sampling distributions. For the sake of comparison, we also obtain estimates of Δ_{ATT} under a frequentist approach and under matching. For each of the weighting approaches, the bootstrap relied on $L = 1000$ samples. The matching was performed with replacement, using 1-1 matching with a caliper distance of 0.1 in the propensity score, and bias adjustment using the same 19 covariates used in the other 3 approaches. Variability of the matching estimator relied on the Abadie-Imbens estimator provided in the `Matching` package in `R` (Sekhon, 2011).

Figure 2.6 shows the summary of the estimates for Δ_{ATT} for the RHC dataset under each of the approaches considered. Solid circles represent the mean and vertical lines correspond to the limits of the 95% posterior credible intervals (under the Bayesian bootstrap) or the limits of the 95% confidence interval (under the frequentist bootstrap and matching). Each of the different methods points to a negative causal effect, that is RHC is reducing the survival at 30 days among those who were treated. This finding agrees with the results described in Connors et al. (1996). The Bayesian bootstraps and frequentist approaches provide similar point estimates and limits of the posterior credible and

confidence intervals are also similar.

We also considered a more parsimonious outcome model that included as covariates only those four variables for which an interaction with the exposure was fit, that is, we consider the following variables in the outcome model: age, estimate of probability of surviving 2 months, the Glasgow coma score, and transfer more than 24 hours from another hospital. Figure 2.6 shows that the results differ only slightly.

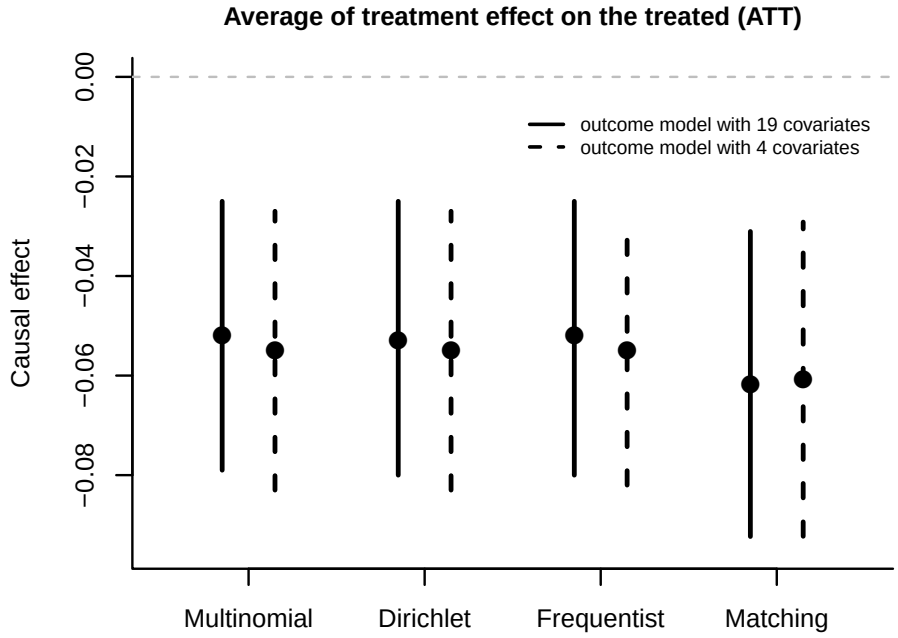


Figure 2.6: ATT estimates and associated measures of variability for the RHC dataset. Solid circles represent the point estimates and vertical lines represent the 95% credible or confidence interval.

Weighted standardized mean differences (SMD) (Austin and Stuart, 2015a) were computed to check the balance in the covariate means between treated and untreated individuals in the sample obtained after weighting. The SMD is calculated as the difference in means of a covariate across the treatment groups, divided by the standard deviation in the treatment groups. The values are below 0.1, often considered a marker of good balance. See Table 2.11 in Section 2.8.2 of the Supplementary Materials for details.

One of the advantages of the Bayesian approach is that we can easily obtain posterior summaries of other quantities of interest. For example, assume one is interested in estimating the posterior probability of RHC reducing 30 day survival by a risk difference of at least 0.05, or equivalently, $P(\Delta_{ATT} < -0.05)$. This is easily achieved by computing the proportion of times that $\Delta_{ATT} < -0.05$ in the sample, that is, $P(\Delta_{ATT} < -0.05) = L^{-1} \sum_{l=1}^{L} I_{\{\Delta_{ATT}^{(l)} < -0.05\}}$, where $I_{\{\Delta_{ATT}^{(l)} < -0.05\}}$ is an indicator function equals 1 if $\{\Delta_{ATT}^{(l)} < -0.05\}$, and 0 otherwise. Here, this probability is estimated at 0.562

under the Multinomial Bayesian bootstrap, and at 0.559 under the Dirichlet Bayesian bootstrap (the probabilities are similar when only four covariates in addition to the exposure are included in the outcome model). Similarly, if interest lies in the posterior probability of RHC reducing 30 day survival by a risk difference of at least 0.07, the estimated values are 0.112 (Multinomial Bayesian bootstrap) and 0.088 (Dirichlet Bayesian bootstrap). The calculation of these probabilities under the frequentist framework are not possible, as the parameter Δ_{ATT} is viewed as a fixed quantity not having a distribution under the frequentist paradigm.

2.5.2 Analysis of the National Center for Health Statistics Birth dataset

We also applied our approach to the Birth dataset from the National Center for Health Statistics (NCHS). The aim is to study the effect of maternal smoking during pregnancy on birth weight. The Birth dataset contains 3,956,112 births that took place in the U.S.A. in 2016. The data are available from https://www.cdc.gov/nchs/index.htm. Available covariates include age, education and health indicators for the mother and father.

In order to have a relatively homogeneous sample, we only consider non-hispanic mothers who were born in the U.S.A. and with both parents having at least 8 years of schooling (Rothe and Firpo, 2013). As we are using this data for illustration purposes only, we randomly selected n=9,980 observations (that corresponds 0.5%) from the original dataset and kept a few covariates in order to decrease the complexity of the problem. The exposure is a dummy variable that equals one if the mother smoked during pregnancy and zero otherwise, whereas the outcome is birth weight measured in grams. In our sample, 801 mothers (8% of total sample) smoked during their pregnancy. The covariates used in this analysis are shown in Table 2.1; summary statistics are given in Table 2.12 of Section 2.8.4 of the Supplementary Materials.

We propose the following conditional exposure model: $Z_i \sim Bern(e_i)$, where $logit(e_i) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 \sqrt{X_{2i}} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \alpha_5 X_{5i} + \alpha_6 X_{6i}$. We assume that the outcome Y_i follows a normal distribution with mean μ_i, and variance of σ^2, where $\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \sqrt{X_{2i}} + \lambda_0 Z_i + \lambda_1 Z_i X_{1i} + \lambda_2 Z_i \sqrt{X_{2i}}$. As in the RHC analysis, four estimators of the ATT are compared: the Bayesian bootstrap under Dirichlet and Multinomial sampling distributions, as well as the weighted frequentist estimator and a matching estimator using the same tuning parameters as above (1-1 matching with replacement, and so on).

Balance is reported in the form of SMDs in Table 2.1, and appears to be adequate with all values below 0.1. Figure 2.7 shows the estimates of Δ_{ATT} and associated credible/confidence intervals under the different approaches. These results show a negative effect of maternal smoking during pregnancy

on birth weight: smoking appears to reduce birth weight by around 150 grams among smokers.

Table 2.1: SMDs computed for the NHCS Birth dataset before and after weighting.

Covariate	Original data	BB/Multimonial	BB/Dirichlet	Frequentist	Matching
X_1: mother's age	0.426	0.024	0.023	0.020	0.084
X_2: weight gain during pregnancy	0.159	0.009	0.009	0.008	0.074
X_3: being married	0.769	0.007	0.007	0.006	0.001
X_4: doing the prenatal care	0.092	<0.001	0.001	0.004	0.019
X_5: pre-pregnancy diabetes	0.031	0.003	<0.001	0.003	0.059
X_6: pre-pregnancy hypertension	0.068	<0.001	<0.001	0.004	0.035

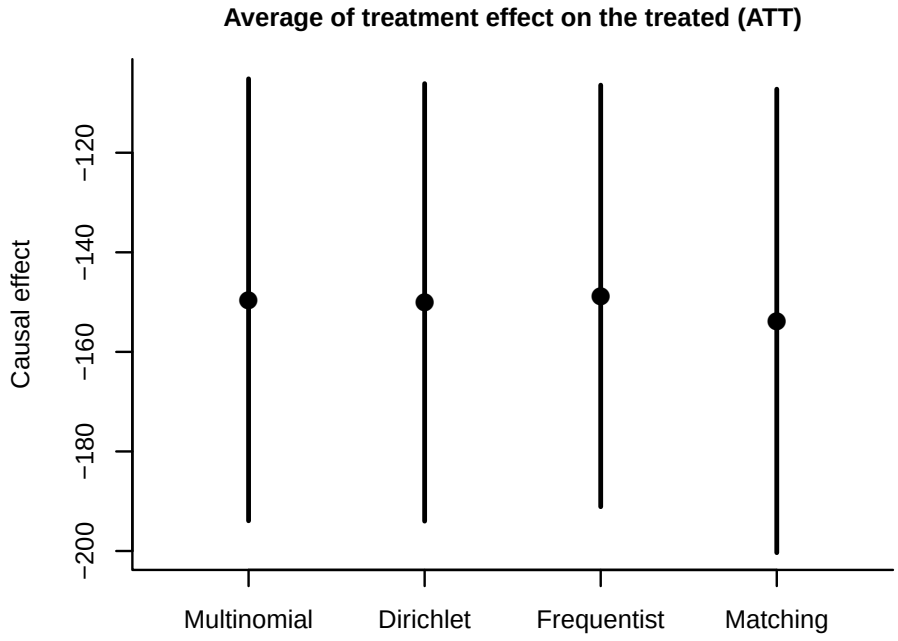


Figure 2.7: ATT estimates and associated measures of variability for the NCHS Birth dataset. Points indicate the point estimates and bars represent the 95% credible or confidence interval.

From the Bayesian approach, we can obtain posterior summaries of parameters, or functions of the parameters. Suppose we are interested in estimating the posterior probability that maternal smoking during pregnancy reduces the birth weight of children born to smokers by at least 150g, or equivalently, $P(\Delta_{ATT} < -150)$. This quantity of interest is easily obtained by computing the proportion of times that $\Delta_{ATT} < -150$ in the posterior sample. In our analyzed dataset, we find that this probability is estimated at 0.484 under the Multinomial Bayesian bootstrap, and at 0.506 under the Dirichlet Bayesian bootstrap. Similarly, if interest lies in the posterior probability of maternal smoking during pregnancy reducing birth weight by at least 180g in children born to smokers, the estimated values are 0.090 and 0.093 for the Bayesian bootstrap with Multinomial and Dirichlet sampling, respectively. Of course, it would not be desirable to consider interventions that make lead to infants being born at the very top end of the birth weight distribution, however this is not a concern in the NCHS Birth data,

where very few children of smokers are heavy, and many are born below the 2,500g threshold for being considered low birth weight, a condition which increases the risk of numerous adverse health conditions.

2.6 Discussion

In addition to the propensity score methods described in the Introduction, alternative Bayesian methods for causal inference have been proposed, and yet the average treatment effect on the treated is a useful estimand that has been largely overlooked in the Bayesian literature. For example, Xu et al. (2018) used Bayesian nonparametric generative models to induce the conditional distribution of the outcome given covariates for causal inference using propensity score. They used a sequential BART approach to impute the missing covariates (Xu et al., 2016) and estimate the quantile causal effects. Gutman and Rubin (2013) described a new procedure called multiple imputation with two subclassification splines (MITSS) that views causal effect estimation as a missing data problem, but only considered binary outcomes and a scalar covariate. Gutman and Rubin (2015) extended this method to situations with continuous outcomes and multiple covariates and used this approach to estimate the ATE, comparing it with other commonly used methods. Roy et al. (2017) parameterized causal effects from a marginal structural model and developed a Bayesian nonparametric approach for continuous and survival outcomes, focus on ATE estimation. Roy et al. (2018) proposed a general Bayesian nonparametric approach to causal inference in the point treatment setting, where the joint distribution of the observed data is modeled using a Dirichlet process and causal effects are obtained using standardization (g-formula). Keil et al. (2018) extended the parametric computation algorithm formula (g-formula) to a Bayesian approach, and showed that the frequentist properties of the Bayesian g-formula seems to improve the accuracy of estimates of causal effects in small samples. Hill (2011) proposed using Bayesian additive regression trees (BART) to flexibly model the response surface. Notice that these studies have focused on the ATE. A notable exception is by Hill and Su (2013) who proposed a method for addressing common support problems and estimating ATT.

In this paper, we have extended the ideas in Saarela et al. (2015a,b) to show how ATT estimation may be accomplished within a Bayesian framework, thus bridging two important statistical domains of research. Using an importance sampling approach to inverse weighting, we have demonstrated how to estimate the ATT within a Bayesian framework while avoiding the problem of feedback, which can corrupt the balance that a propensity score aims to provide. In the simulation study, we found that the Multinomial weighting within the Bayesian bootstrap provided marginally better results than the Dirichlet weighting. In the real data analysis, Bayesian bootstrap with Multinomial and Dirichlet

sampling, and Frequentist approach present very similar results.

A Bayesian approach offers several advantages. The first is the interpretability afforded by posterior probability. Another key advantage of the Bayesian approach is that we can easily obtain posterior summaries of any function of the parameters involved in the model, as we demonstrated in both real-data examples. Although the Bayesian Bootstrap approach implicitly uses only a non-informative prior, it has been suggested (Saarela et al., 2015a) that the inclusion of more informative prior information can be accomplished by combining the Bayesian bootstrap with a sampling-importance resampling approach (Newton and Raftery, 1994). Investigations about the incorporation of prior information into Bayesian causal adjustments provide an avenue of future research.

2.7 Appendices

2.7.1 The De Finetti representation

This appendix contains the de Finetti representation (De Finetti, 1974) of the joint distribution of a random sample of size n from a super-population, under experimental and observational settings. The notation $\mathscr{E}$ indexes the treatment assignment mechanism under experimental setting. The notation $\mathscr{O}$ indexes the data generating mechanism under the observational setting, where the treatment assignment Z can depend on the covariates X. Let (y,z,x) be the observed data and $\pi(\cdot)$ a prior distribution. We assume prior independence among the parameters that govern the treatment and outcome models. The joint distribution under the experimental and observational settings are given, respectively, by

$$\begin{aligned} P_{\mathscr{E}}(y,x,z) &= \int_{\gamma,\theta} P_{\mathscr{E}}(y,z,x|\gamma,\theta)\pi(\gamma)\pi(\theta)\partial\gamma\partial\theta \\ &= \int_{\gamma,\theta} P(y|z,x,\theta)P_{\mathscr{E}}(z|\gamma)P(x)\pi(\gamma)\pi(\theta)\partial\gamma\partial\theta \\ &= \int_{\gamma,\theta} \prod_{i=1}^{n}\Big[P(y_i|z_i,x_i,\theta)P_{\mathscr{E}}(z_i|\gamma)P(x_i)\Big]\pi(\gamma)\pi(\theta)\partial\gamma\partial\theta, \end{aligned} \tag{2.7.10}$$

and

$$\begin{aligned} P_{\mathscr{O}}(y,x,z) &= \int_{\alpha,\theta} P_{\mathscr{O}}(y,z,x|\alpha,\theta)\pi(\alpha)\pi(\theta)\partial\alpha\partial\theta \\ &= \int_{\alpha,\theta} P(y|z,x,\theta)P_{\mathscr{O}}(z|x,\alpha)P(x)\pi(\alpha)\pi(\theta)\partial\alpha\partial\theta \\ &= \int_{\alpha,\theta} \prod_{i=1}^{n}\Big[P(y_i|z_i,x_i,\theta)P_{\mathscr{O}}(z_i|x_i,\alpha)P(x_i)\Big]\pi(\alpha)\pi(\theta)\partial\alpha\partial\theta. \end{aligned} \tag{2.7.11}$$

Considering equations (2.7.10) and (2.7.11), the posterior distribution of parameters of marginal (γ) and conditional (α) treatment models is, respectively,

$$\begin{aligned}
\pi_{\mathscr{E}}(\gamma|y,x,z) &= \int_{\theta} \pi_{\mathscr{E}}(\gamma,\theta|y,x,z)\partial\theta \\
&\propto \int_{\theta} P_{\mathscr{E}}(y,z,x|\gamma,\theta)\pi(\gamma)\pi(\theta)\partial\theta \\
&\propto P_{\mathscr{E}}(z|\gamma)\pi(\gamma) \\
&\propto \pi_{\mathscr{E}}(\gamma|z), \qquad (2.7.12)
\end{aligned}$$

and

$$\begin{aligned}
\pi_{\mathscr{O}}(\alpha|y,x,z) &= \int_{\theta} \pi_{\mathscr{O}}(\alpha,\theta|y,x,z)\partial\theta \\
&\propto \int_{\theta} P_{\mathscr{O}}(y,z,x|\alpha,\theta)\pi(\alpha)\pi(\theta)\partial\theta \\
&\propto P_{\mathscr{O}}(z|x,\alpha)\pi(\alpha) \\
&\propto \pi_{\mathscr{O}}(\alpha|z,x). \qquad (2.7.13)
\end{aligned}$$

Note that the marginal posterior density for γ in equation (2.7.12), and the marginal posterior density for α in equation (2.7.13) do not depend on the outcome y, thus avoiding any problem of feedback.

2.7.2 Computation of ATT weights ω_i

Note that we can write the expected value of the potential outcome on the treated as

$$\begin{aligned}
E[Y(z)|Z=1] &= E[E[Y(z)|X,Z=1]|Z=1] \\
&= \int E[Y(z)|X,Z=1]P_{X|Z}(x|Z=1)d_x \\
&= \int E[Y(z)|X]P_{X|Z}(x|Z=1)d_x \\
&= \int yP_{Y(z)|X}(y|x)P_{X|Z}(x|Z=1)d_xd_y \\
&= \int yI_{\{Z=z\}}P_{Y|X,Z}(y|x,z)P_{X|Z}(x|Z=1)d_xd_y \\
&= z\int yI_{\{Z=z\}}P_{Y|X,Z}(y|x,z)P_{X|Z}(x|Z=1)d_xd_y \\
&\quad +(1-z)\int yI_{\{Z=z\}}P_{Y|X,Z}(y|x,z)P_{X|Z}(x|Z=1)d_xd_y.
\end{aligned}$$

As all treated units have $z = 1$, then

$$\begin{aligned} E[Y(1)|Z=1] &= \int yI_{\{Z=1\}}P_{Y|X,Z}(y|x,1)P_{X|Z}(x|Z=1)d_xd_y \\ &= \int yI_{\{Z=1\}}P_{Y,X|Z}(y,x|Z=1)d_xd_y \\ &= \int yI_{\{Z=1\}}P(d_y,d_x|Z=1) \\ &= E_{Y,X|Z=1}[yI_{\{Z=1\}}]. \end{aligned}$$

Thus, the expected potential outcome among the treated when receiving treatment $z = 1$ is estimated as

$$\widehat{E}[Y(1)|Z=1] = \frac{\sum_{i=1}^{n} y_i I_{\{Z_i=1\}}}{\sum_{i=1}^{n} I_{\{Z_i=1\}}}.$$

If we could observe the potential outcome under no treatment ($z = 0$) within the treated population, the expectation of this distribution would be given by

$$\begin{aligned} E[Y(0)|Z=1] &= \int yI_{\{Z=0\}}P_{Y|X,Z}(y|x,Z=0)P_{X|Z}(x|Z=1)d_xd_y \\ &= \int yI_{\{Z=0\}}\frac{P_{Y|X,Z}(y|x,Z=0)}{P_{Y|X,Z}(y|x,Z=1)}P_{Y|X,Z}(y|x,Z=1)P_{X|Z}(x|Z=1)d_xd_y \\ &= \int yI_{\{Z=0\}}\frac{P_{Y|X,Z}(y|x,Z=0)}{P_{Y|X,Z}(y|x,Z=1)}P_{Y,X|Z}(y,x|Z=1)d_xd_y \\ &= \int yI_{\{Z=0\}}\frac{P_{Y|X,Z}(y|x,Z=0)}{P_{Y|X,Z}(y|x,Z=1)}\frac{P_{Y,X|Z}(y,x|Z=1)}{P_{Y,X|Z}(y,x|Z=0)}P_{Y,X|Z}(y,x|Z=0)d_xd_y \\ &= \int yI_{\{Z=0\}}\frac{P_{X|Z}(x|Z=1)}{P_{X|Z}(x|Z=0)}P_{Y,X|Z}(y,x|Z=0)d_xd_y \\ &= \int yI_{\{Z=0\}}\omega^0 P(d_y,d_x|Z=0) \\ &= E_{Y,X|Z=0}[yI_{\{Z=0\}}\omega^0], \end{aligned}$$

where the weight ω^0 is given by

$$\begin{aligned}
\omega^0 &= \frac{P_{X|Z}(x|Z=1)}{P_{X|Z}(x|Z=0)} \\
&= \frac{P(Z=1|x)P_X(x)/P(Z=1)}{P(Z=0|x)P_X(x)/P(Z=0)} \\
&= \frac{P(Z=1|x)}{P(Z=0|x)}\frac{P(Z=0)}{P(Z=1)} \\
&= \frac{P(Z=1|x)}{P(Z=1)}\frac{P(Z=0)}{P(Z=0|x)}.
\end{aligned}$$

Thus the expected potential outcome if all treated individuals did not in fact receive treatment (i.e. under $z=0$) is given by

$$\widehat{E}[Y(0)|Z=1] = \frac{\sum_{i=1}^{n} y_i I_{\{Z_i=0\}} \omega_i^0}{\sum_{i=1}^{n} I_{\{Z_i=0\}} \omega_i^0}.$$

Note that the stabilized weight for the untreated units used to estimate the Δ_{ATT} is a ratio of the stabilized inverse probability weights for the untreated and treated units used to estimate Δ_{ATE}.

2.7.3 Additional results of the simulation studies

A simple artificial dataset

Now we provide additional results for the estimation of the average treatment effect in the simulated studies of Subsection 2.4.1. Table 2.2 presents the results for Study 1.A, which aimed to measure the impact of the strength of confounding on ATT estimators. The relative biases are greater when the effect of the covariate X_2 in the treatment model is strong ($\alpha_2 = -1.0$). The Bayesian bootstrap with Multinomial sampling yields the greatest relative bias of standard error, regardless of value of α_2. The coverage probabilities are close to the nominal level.

Table 2.3 shows the results for Study 1.B, which aimed to investigate the impact of sample size on ATT estimators. As n increases, the relative bias decreases, the SE relative bias is negligible and the coverage probabilities improve for both Bayesian and frequentist approaches.

Table 2.2: Further results of Study 1.A. ATT estimation for a simple data generation: bias of point estimates relative to the true value of Δ_{ATT} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different strengths of confounding in the treatment and outcome models.

α_2	β_2	RB(%)			SE RB(%)			95% CP		
		BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
	-0.6	0.041	0.042	0.045	2.047	-0.997	0.496	94.7	94.2	94.3
-1.0	-0.2	0.043	0.042	0.045	1.945	-0.849	0.469	94.6	94.3	94.5
	0.0	0.043	0.042	0.045	2.062	-0.674	0.486	94.5	94.4	94.4
	-0.6	-0.027	-0.030	-0.027	4.031	2.569	4.013	95.6	95.3	95.3
-0.4	-0.2	-0.025	-0.025	-0.027	4.175	2.700	3.842	95.6	95.1	95.4
	0.0	-0.028	-0.027	-0.027	4.051	2.880	3.822	95.4	95.3	95.5

Table 2.3: Further results of Study 1.B. ATT estimation for a simple data generation: bias of point estimates relative to the true value of Δ_{ATT} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different sample sizes n.

n	RB(%)			SE RB(%)			95% CP		
	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
100	-0.184	-0.196	-0.211	1.401	-9.543	-1.408	95.1	91.7	94.4
250	0.125	0.138	0.144	-3.320	-8.542	-5.340	94.8	93.2	94.3
500	0.114	0.114	0.113	-4.432	-7.871	-6.138	93.9	93.3	93.5
750	0.045	0.043	0.043	-2.013	-4.776	-3.219	95.2	94.6	94.9
1000	0.042	0.043	0.045	1.985	-0.919	0.455	94.7	94.5	94.6
2500	-0.005	-0.004	-0.001	0.062	-1.547	-0.962	94.7	94.5	94.5
5000	0.053	0.055	0.055	0.634	-0.325	-0.012	94.9	94.9	94.8
7500	0.032	0.031	0.033	-1.727	-2.445	-2.255	95.0	94.6	94.7
10000	0.032	0.033	0.032	1.635	0.785	1.038	95.0	94.8	95.3

A more complex simulated dataset

Here we provide further results for the estimation of the average treatment effect in the simulated studies of Subsection 2.4.2 that consider a more complex structure mean for the treatment model. Table 2.4 presents the results for Study 2.A, which aimed to measure the impact of the strength of confounding on the ATT estimation. Overall, the Bayesian bootstrap with Multinomial sampling performs best, though all three methods perform well.

Table 2.5 shows the results of Study 2.B, to study the impact of sample size n on estimation of the ATT estimation when there is a more complex mean structure in the treatment model. As n increases, the relative bias decreases. The Bayesian bootstrap with Multinomial sampling appears to provide the smallest standard error bias. The 95% interval coverage probabilities are similar for both Bayesian and frequentist approaches.

Table 2.4: Further results of Study 2.A. ATT estimation for a complex data generation: bias of point estimates relative to the true value of Δ_{ATT} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different strengths of confounding in the treatment and outcome models.

α_3	β_3	RB(%)			SE RB(%)			95% CP		
		BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
	-0.8	-0.004	-0.014	-0.014	0.644	-1.477	-0.293	95.2	94.6	95.1
-1.0	-0.4	-0.010	-0.009	-0.014	0.994	-1.433	-0.114	95.1	95.0	94.9
	0.0	-0.012	-0.012	-0.014	0.712	-1.466	-0.250	95.4	95.2	94.9
	-0.8	-0.118	-0.121	-0.121	-0.328	-3.431	-1.654	95.5	94.0	95.3
-0.5	-0.4	-0.116	-0.119	-0.121	-0.200	-3.616	-1.595	95.6	94.4	95.3
	0.0	-0.115	-0.117	-0.121	-0.289	-3.219	-1.567	95.5	94.8	95.3

Table 2.5: Further results of Study 2.B. ATT estimation for a complex data generation: bias of point estimates relative to the true value of Δ_{ATT} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different sample sizes n.

n	RB(%)			SE RB(%)			95% CP		
	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
100	0.566	0.557	0.554	2.263	-12.714	-0.931	95.3	90.9	94.6
250	0.285	0.293	0.300	-3.565	-10.139	-5.047	94.1	91.9	93.3
500	-0.282	-0.288	-0.293	-1.192	-5.670	-2.450	94.5	92.6	93.9
750	-0.162	-0.167	-0.172	-2.128	-5.534	-3.341	94.4	93.6	93.7
1000	-0.113	-0.119	-0.121	-0.078	-3.511	-1.631	95.0	94.6	95.5
2500	-0.008	-0.008	-0.011	-2.662	-4.334	-3.363	94.1	93.9	94.1
5000	-0.075	-0.074	-0.075	-0.225	-1.430	-0.760	94.5	94.4	94.6
7500	-0.075	-0.074	-0.077	0.391	-0.772	-0.412	93.5	93.3	93.3
10000	-0.041	-0.038	-0.038	-0.588	-1.289	-0.927	94.3	94.1	94.3

2.8 Supplementary Materials

2.8.1 Estimation of Δ_{ATE}

In this section, we provide a simulation study of ATE estimation, using the same data generating scenarios described in Section 4 of the main paper. To obtain the true ATE value, note that we may calculate it analytically as $\Delta_{ATE} = \lambda_0 + \lambda_1 E[X_1]$. Note that our simulations are equivalent to a one-interval marginal structural model, and hence these simulations – as expected – are similar to those in Saarela et al. (2016, 2015b).

Study 1.A: The impact of the strength of confounding on ATE estimators for a simple data generation

The results of ATE estimation present a very similar pattern to the results on estimation of the ATT. Figure 2.8 and Table 2.6 show that the magnitude of relative bias is larger when the degree of confounding is greater ($\alpha_2 = -1.0$). However, the SE relative bias is bigger when $\alpha_2 = -0.4$. Coverage probabilities are very close to the nominal level for all methods.

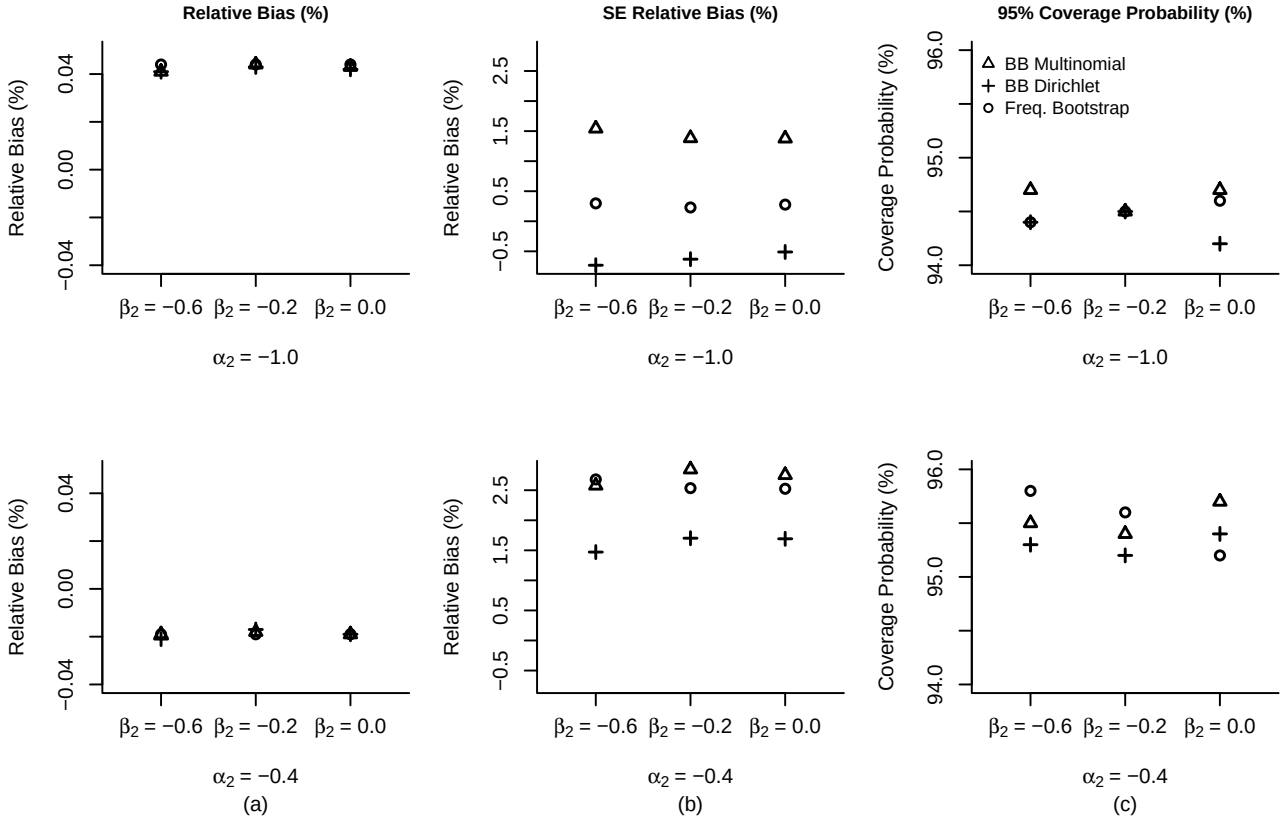


Figure 2.8: ATE estimation for a simple data generation: (a) bias of point estimates relative to the true value of Δ_{ATE}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for different strengths of confounding. Panels in the top row consider stronger confounding, those in the bottom row consider moderate confounding.

Study 1.B: The impact of the sample sizes on ATE estimators for a simple data generation

Again, we observe a very similar pattern between the results obtained for the ATE and ATT estimation. Figure 2.9 and Table 2.7 show that, as the sample size becomes large, the relative bias decreases and 95% coverage probabilities become closer to the nominal level. For n sufficiently large, the frequentist and Bayesian bootstrap tend to provide similar results with respect to the relative bias and SE relative bias.

Table 2.6: Further results of Study 1.A. ATE estimation for a simple data generation: bias of point estimates relative to the true value of Δ_{ATE} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different strengths of confounding in the treatment and outcome models.

α_2	β_2	RB(%)			SE RB(%)			95% CP		
		BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
	-0.6	0.041	0.041	0.044	1.543	-0.732	0.298	94.7	94.4	94.4
-1.0	-0.2	0.044	0.043	0.044	1.383	-0.631	0.230	94.5	94.5	94.5
	0.0	0.043	0.042	0.044	1.378	-0.511	0.277	94.7	94.2	94.6
	-0.6	-0.019	-0.021	-0.019	2.578	1.471	2.680	95.5	95.3	95.8
-0.4	-0.2	-0.018	-0.017	-0.019	2.847	1.702	2.535	95.4	95.2	95.6
	0.0	-0.019	-0.019	-0.019	2.753	1.692	2.525	95.7	95.4	95.2

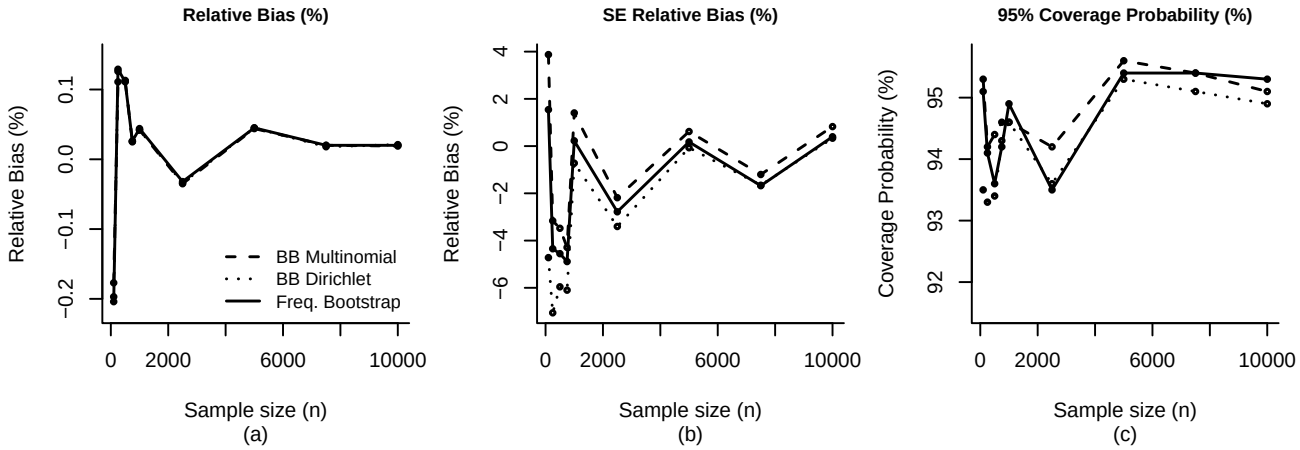


Figure 2.9: ATE estimation for a simple data generation: (a) bias of point estimates relative to the true value of Δ_{ATE}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for varying sample sizes. Results are obtained under the frequentist estimation (solid line), the Bayesian bootstrap (BB) with Multinomial (dashed line), and Dirichlet (dotted line) distributions.

Study 2.A: The impact of the strength of confounding on ATE estimators for a complex data generation

Here we present the results of ATE estimation considering a more complex structure mean for the treatment model. Figure 2.10 and Table 2.8 show that the magnitude (absolute value) of relative bias is smaller when the effect of coefficient of X_3 in the treatment model is moderate ($\alpha_3 = -0.5$). As noted in ATT estimation, the Bayesian bootstrap with Multinomial sampling provides the best estimator of the standard error. We obtain good coverage when the effect of coefficient of X_3 in the treatment model is strong ($\alpha_3 = -1.0$), particularly for the Bayesian bootstrap with Multinomial sampling.

Table 2.7: Further results of Study 1.B. ATE estimation for a simple data generation: bias of point estimates relative to the true value of Δ_{ATE} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different sample sizes n.

n	RB(%)			SE RB(%)			95% CP		
	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
100	-0.177	-0.197	-0.204	3.874	-4.721	1.542	95.3	93.5	95.1
250	0.111	0.126	0.129	-3.163	-7.059	-4.346	94.2	93.3	94.1
500	0.112	0.113	0.111	-3.479	-5.957	-4.554	94.4	93.4	93.6
750	0.026	0.025	0.026	-4.289	-6.099	-4.888	94.6	94.3	94.2
1000	0.042	0.042	0.044	1.404	-0.722	0.224	94.6	94.6	94.9
2500	-0.035	-0.034	-0.032	-2.184	-3.404	-2.781	94.2	93.6	93.5
5000	0.044	0.045	0.045	0.617	-0.069	0.176	95.6	95.3	95.4
7500	0.020	0.018	0.020	-1.202	-1.658	-1.676	95.4	95.1	95.4
10000	0.019	0.021	0.020	0.826	0.335	0.395	95.1	94.9	95.3

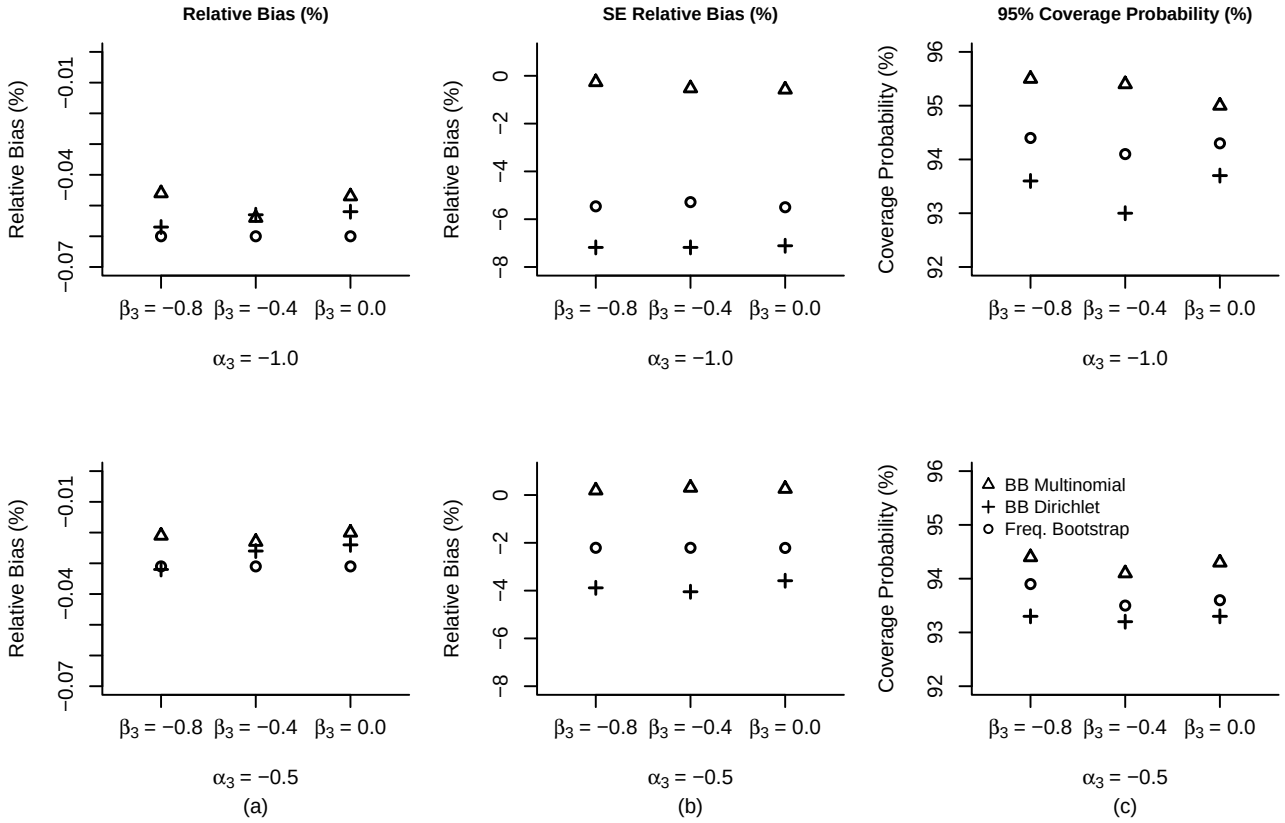


Figure 2.10: ATE estimation for a complex data generation: (a) bias of point estimates relative to the true value of Δ_{ATE}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for different strengths of confounding. Panels in the top row consider stronger confounding, those in the bottom row consider moderate confounding.

Table 2.8: Further results of Study 2.A. ATE estimation for a complex data generation: bias of point estimates relative to the true value of Δ_{ATE} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different strengths of confounding in the treatment and outcome models.

α_2	β_2	RB(%)			SE RB(%)			95% CP		
		BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
	-0.8	-0.046	-0.057	-0.060	-0.264	-7.181	-5.461	95.5	93.6	94.4
-1.0	-0.4	-0.054	-0.053	-0.060	-0.522	-7.180	-5.285	95.4	93.0	94.1
	0.0	-0.047	-0.052	-0.060	-0.574	-7.112	-5.499	95.0	93.7	94.3
	-0.8	-0.021	-0.032	-0.031	0.192	-3.884	-2.207	94.4	93.3	93.9
-0.5	-0.4	-0.023	-0.026	-0.031	0.306	-4.047	-2.207	94.1	93.2	93.5
	0.0	-0.020	-0.024	-0.031	0.260	-3.586	-2.213	94.3	93.3	93.6

Study 2.B: The impact of the sample sizes on ATE estimators for a complex data generation

Figure 2.11 and Table 2.9 present the results of study 2.B, examining the impact of sample size n on the ATE estimator when there is a more complex relationship in the treatment model. As n increases, the relative bias decreases. Although we see that the relative bias of standard error increases (in absolute value) for sample size $n = 10,000$, this is a result of the denominator being very small, and thus exaggerating small differences between the estimated standard error and the standard deviation of the point estimates. The 95% interval coverage probabilities are at about the nominal level, particularly for the Bayesian bootstrap with Multinomial sampling.

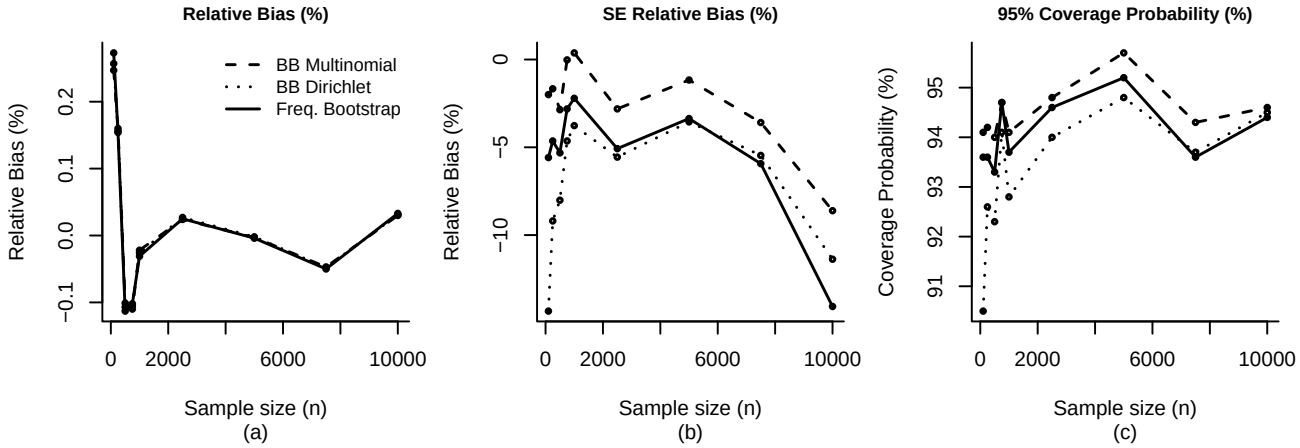


Figure 2.11: ATE estimation for a complex data generation: (a) bias of point estimates relative to the true value of Δ_{ATE}, (b) bias of the standard error estimates relative to Monte Carlo standard deviation, and (c) 95% credible/confidence interval coverage probability for varying sample sizes. Results are obtained under the frequentist estimation (solid line), the Bayesian bootstrap (BB) with Multinomial (dashed line), and Dirichlet (dotted line) distributions.

Table 2.9: Further results of Study 2.B. ATE estimation for a complex data generation: bias of point estimates relative to the true value of Δ_{ATE} (RB), bias of the standard error estimates relative to the Monte Carlo standard deviation (SE RB), and 95% credible/confidence interval coverage probability (95% CP) for different sample sizes n.

n	RB(%)			SE RB(%)			95% CP		
	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.	BB/Mult	BB/Dir	Freq.
100	0.247	0.257	0.273	-1.999	-14.331	-5.585	94.1	90.5	93.6
250	0.160	0.156	0.154	-1.665	-9.195	-4.639	94.2	92.6	93.6
500	-0.101	-0.107	-0.113	-2.858	-8.014	-5.316	94.0	92.3	93.3
750	-0.110	-0.105	-0.102	-0.021	-4.640	-2.815	94.7	94.1	94.7
1000	-0.022	-0.026	-0.031	0.379	-3.765	-2.209	94.1	92.8	93.7
2500	0.024	0.027	0.025	-2.812	-5.559	-5.076	94.8	94.0	94.6
5000	-0.003	-0.002	-0.004	-1.167	-3.552	-3.363	95.7	94.8	95.2
7500	-0.047	-0.048	-0.050	-3.588	-5.462	-5.932	94.3	93.7	93.6
10000	0.030	0.031	0.033	-8.610	-11.376	-14.067	94.6	94.5	94.4

2.8.2 Further information about the RHC dataset

The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) collected data on hospitalized adult patients at five medical centers in the U.S., including information about a big set of variables relating to the decision to perform the Right Heart Catheterization (RHC). More information about the study can be found in Connors et al. (1996). Table 2.10 lists the 19 covariates considered in our analysis.

Table 2.11 shows the standardized mean difference computed for each covariate on the original data before weighting and the weighted SMD considering the ATT weights to check the balancing obtained after weighting. The values are below the cut-offs of 0.1 and indicate the reduction of unbalancing in all covariates.

We also provide ATE results using the RHC dataset from SUPPORT. Figure 2.12 shows the estimates for Δ_{ATE}, where solid circles indicate the point estimates, solid and dashed lines represent the 95% credible/confidence intervals of these estimates. We note that, as in the estimation of ATT, the frequentist estimate is very similar to the Bayesian estimates, regardless of whether multinomial or Dirichlet distributions are used for sampling weights. These results show that RHC is reducing 30-day survival or, equivalently, increasing 30-day mortality among all the patients.

Table 2.10: RHC covariates considered in our analysis. The first column contains the name assigned to covariates in the SUPPORT dataset, which is also how they appear in the R code provided in Section S3 of the Online Supplementary Materials. The last three columns present summaries of covariates in the dataset and in two treatment groups.

Name	Description	Total $n = 5735$	Control $n_0 = 3551$	Treated $n_1 = 2184$
age	Age (years)	61.38 (16.7)	61.76 (17.3)	60.75 (15.6)
income	Under \$11k	56.2%	58.6%	52.4%
	Income \$11 - \$25k	20.3%	20.1%	20.7%
	Income \$25 - \$50k	15.6%	14.1%	18.0%
	Income >\$50k	7.9%	7.2%	8.9%
ins_no	No insurance indicator	5.6%	5.2%	6.2%
cat1	ARF	43.4%	44.5%	41.6%
	CHF	7.9%	7.0%	9.6%
	Cirrhosis	3.9%	4.9%	2.2%
	Colon Cancer	0.1%	0.2%	<0.1%
	Coma	7.6%	9.6%	4.3%
	COPD	8.0%	11.2%	2.7%
	Lung Cancer	0.7%	1.0%	0.2%
	MOSF w=Malignancy	7.0%	6.8%	7.2%
	MOSF w=Sepsis	21.4%	14.8%	32.1%
cat2	Cirrhosis	0.7%	0.8%	0.5%
	Colon Cancer	<0.1%	<0.1%	<0.1%
	Coma	1.6%	2.0%	0.9%
	Lung Cancer	0.2%	0.4%	0.1%
	MOSF w=Malignancy	4.0%	4.8%	2.7%
	MOSF w=Sepsis	14.4%	11.4%	19.2%
	without information	79.1%	80.6%	76.6%
resp	Respiratory diagnosis	36.8%	41.7%	28.9%
neuro	Neurological diagnosis	12.1%	16.2%	5.4%
hema	Hematological diagnosis	6.2%	6.7%	5.2%
dnr1	Do Not Resuscitate status on day 1	11.4%	14.1%	7.1%
surv2md1	Estimate of prob. of surviving 2 months	0.59 (0.2)	0.61 (0.2)	0.57 (0.2)
aps1	APACHE score	54.67 (20.0)	50.93 (18.8)	60.74 (20.3)
scoma1	Glasgow coma score	21.00 (30.3)	22.25 (31.4)	18.97 (28.3)
wtkilo1	Weight	67.83 (29.1)	65.04 (29.5)	72.36 (27.7)
hrt1	Heart Rate	115.18 (41.2)	112.87 (40.9)	118.93 (41.5)
bili1	Bilirubin	2.27 (4.8)	2.00 (4.4)	2.71 (5.3)
psychhx	Psychiatric history, active psychosis or severe depression	6.7%	8.1%	4.6%
malighx	Solid tumor, metastatic disease, chronic leukemia=myeloma, acute leukemia, lymphoma	22.9%	24.6%	20.3%
transhx	transfer (>24 hours) from another hospital	11.5 %	9.4%	15.0%
wt0	weight = 0 indicator	9.0%	10.3%	6.9%

Table 2.11: SMDs computed for the RHC dataset before and after weighting.

Covariate	Original data	BB/Multimonial	BB/Dirichlet	Frequentist	Matching
age	0.061	0.021	0.022	0.021	0.005
income	0.092	0.014	0.015	0.014	0.012
ins_no	0.043	0.004	0.003	0.004	0.015
cat1	0.205	0.026	0.026	0.026	0.009
cat2	0.037	0.002	0.003	0.003	0.026
resp	0.270	0.007	0.007	0.007	0.013
neuro	0.353	0.008	0.007	0.005	0.010
hema	0.062	$<$0.001	0.004	0.004	0.027
dnr1	0.228	0.009	0.008	0.007	0.025
surv2md1	0.198	0.028	0.028	0.028	0.008
aps1	0.501	0.019	0.019	0.017	0.005
scoma1	0.110	0.016	0.018	0.016	0.020
wtkilo1	0.256	0.022	0.019	0.019	0.015
hrt1	0.147	0.012	0.011	0.012	0.017
bili1	0.145	0.009	0.009	0.009	0.017
psychhx	0.143	0.019	0.020	0.019	0.009
malighx	0.101	0.017	0.018	0.019	0.016
transhx	0.170	0.015	0.014	0.016	0.009
wt0	0.119	0.034	0.033	0.033	0.006

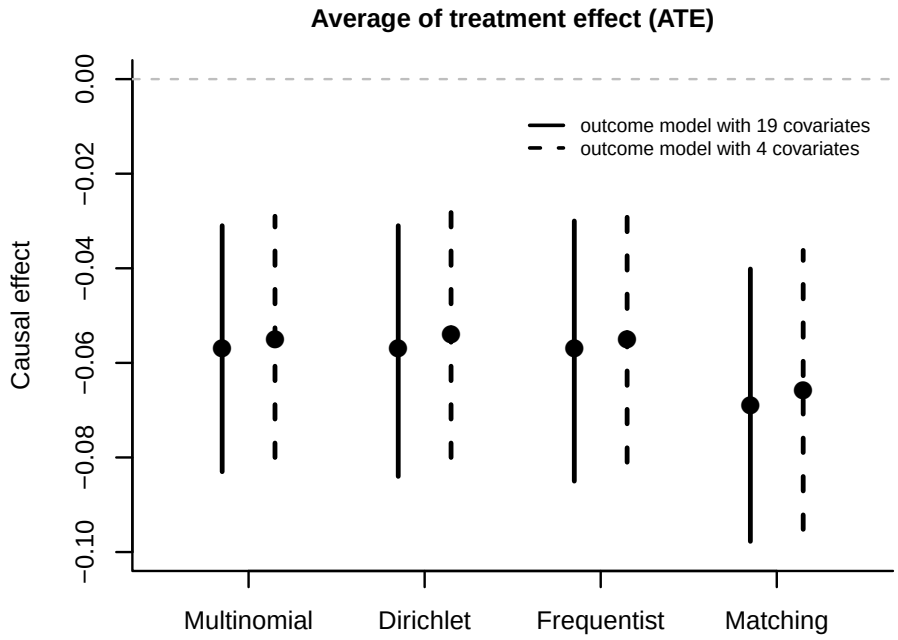


Figure 2.12: ATE estimates and associated measures of variability for the RHC dataset. Solid circles represent the point estimates and vertical lines represent the 95% credible or confidence interval.

2.8.3 R code to estimate ATT

The Supplementary Materials also include R code to estimate the ATT from the RHC dataset.

```
#### R code to accompany "Bayesian estimation for the average treatment effect
#### on the treated using inverse weighting" by Capistrano, Moodie, & Schmidt

##############################################################################
##    Function for estimate ATT using Frequentist and Bayesian Approaches   ##
##############################################################################

My.Optim <- function(dt,var1,var2,var3,weig){

# Conditional model
matXX = as.matrix(cbind(1,dt[,var1]))
p = ncol(matXX)
cpar = coef(glm(z~0+matXX,weight=weig,family="binomial"))
cfitt = 1/(1+exp(-matXX %*% cpar))
cprob = dt$z*cfitt + (1-dt$z)*(1-cfitt)

# Marginal model
mpar = coef(glm(z~1,weight=weig,family="binomial"))
mfitt = 1/(1+exp(-mpar))
mprob = dt$z*mfitt + (1-dt$z)*(1-mfitt)

# Auxiliar matrix
matZX = as.matrix(cbind(1,dt[,var2]))
k = ncol(matZX)
mat.aux = as.matrix(cbind(1,dt[,var3]))
l = k-ncol(mat.aux)

# ATT
ws.z0 = (mprob/cprob)*((1-cprob)/(1-mprob))
ATT.iptw = dt$z*1.0 +(1-dt$z)*ws.z0
```

```
ATT.par = coef(glm(y~0+matZX,weight=ATT.iptw*weig,family="binomial"))
ATT.aux.y0 = matXX[dt$z==1,]%*%ATT.par[1:l]
ATT.aux.y1 = ATT.aux.y0 + mat.aux[dt$z==1,]%*%ATT.par[(l+1):k]
ATT.aux.dif = 1/(1+exp(-ATT.aux.y1)) - 1/(1+exp(-ATT.aux.y0))
ATT.est = sum(weig[dt$z==1]*ATT.aux.dif)/sum(weig[dt$z==1])

return(ATT.est)
}

#############################################################################
##                            RHC dataset study                            ##
#############################################################################

# Propensity score
z = as.numeric(rhc$swang1=="RHC")

# Outcome
y = as.numeric(rhc$dth30=="No")

# Defining covariates
covs = cbind(transhx,age,surv2md1,scoma1,hrt1,bili1,wtkilo1,wt0,dnr1,
resp,neuro,hema,ins_no,income,cat1,cat2NA,psychhx,malighx,aps1)
ncov = ncol(covs)
vars = c("transhx","age","surv2md1","scoma1","hrt1","bili1","wtkilo1","wt0","dnr1",
"resp","neuro","hema","ins_no","income","cat1","cat2NA","psychhx","malighx","aps1")

xz = covs*z
colnames(xz) = paste(vars,".z",sep="")
data = data.frame(z,y,covs,xz)
n = nrow(data)

## ----------------------  Estimating parameters  ------------------------ ##

# Covariates in propensity score and outcome models
```

```
ps.cov = vars
out.cov = c(vars,"z",colnames(xz)[1:4])
eff.cov = vars[1:4]

# Saving the results
Delta.ATT = NULL

# Boostrap
for(l in 1:1000){

# Frequentist
new.data = data[sample(1:n,n,replace=TRUE),]
freq.results = My.Optim(dt = new.data, ps.cov, out.cov, eff.cov, weig = rep(1,n))

# Bayes/Multinomial
mult.weig = as.numeric(rmultinom(1, n, rep(1.0/n,n)))
mult.results = My.Optim(dt = data, ps.cov, out.cov, eff.cov, weig = mult.weig)

# Bayes/Dirichlet
dir.weig = as.numeric(rdirichlet(1, rep(1.0, n))*n)
dir.results = My.Optim(dt = data, ps.cov, out.cov, eff.cov, weig = dir.weig)

# ATT estimates
Delta.ATT = rbind(Delta.ATT,c(freq.results,mult.results,dir.results))
}

# Saving the estimates
func.result = function(x){c(mean(x),var(x))}
ATT.output = as.vector(apply(Delta.ATT,2,func.result))

# Correcting the frequentist mean
freq.mean = My.Optim(dt = data, ps.cov, out.cov, eff.cov, weig = rep(1,n))
ATT.output[1]= freq.mean
```

2.8.4 Further information about the NCHS Birth dataset

The Birth dataset is a very rich database of 3,956,112 births that took place in U.S. in 2016. This dataset was obtained from National Center for Health Statistics (NCHS) and can be public accessed by https://www.cdc.gov/nchs/index.htm. The dataset include variables as age, education and health indicators for the mother and father, among others. Table 2.12 lists the covariates considered in our analysis and their distributions considering the whole sample, and the two groups: non-smoking and smoking mothers.

Table 2.12: Covariates from the NCHS Birth Dataset considered in our analysis. The last three columns present summaries of covariates in the dataset and in two treatment groups.

Covariate	Total $n = 9980$	Non-smoker $n_0 = 9179$	Smoker $n_1 = 801$
Mother's age	29.05 (5.5)	29.23 (5.5)	26.92 (5.4)
Weight gain during pregnancy (in pounds)	5.47 (1.6)	5.39 (1.5)	5.12 (1.8)
Indicator of being married	69.94%	72.81%	37.08%
Indicator of doing the prenatal care	99.19%	99.27%	98.25 %
Indicator of pre-pregnancy diabetes	0.84%	0.82%	1.13%
Indicator of pre-pregnancy hypertension	1.82%	1.74%	2.74%

We also provide ATE results using the NCHS Birth dataset. Figure 2.13 shows the estimates for Δ_{ATE}, where solid circles indicate the point estimates, and solid lines represent the 95% credible/confidence intervals of these estimates. We note that, as in the estimation of ATT, the frequentist estimate is very similar to the Bayesian estimates, regardless of whether multinomial or Dirichlet distributions are used for sampling weights. These results show that the effect of maternal smoking during pregnancy reduces the birth weight by approximately 200 grams per child, on average, among all women in the study population.

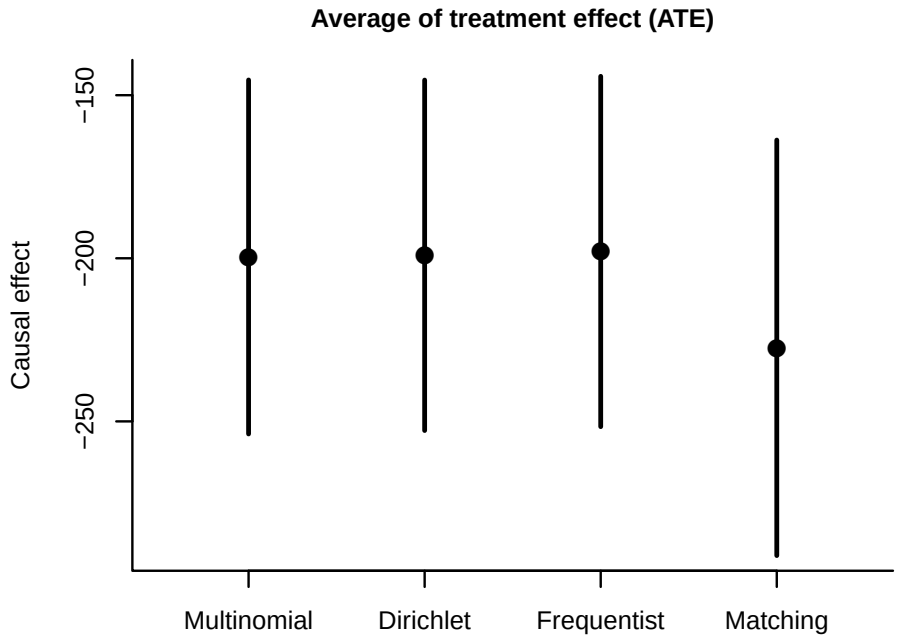


Figure 2.13: ATE estimates and associated measures of variability for the NCHS Birth dataset. Points indicate the point estimates and bars represent the 95% credible or confidence interval.

Chapter 3

Successful hepatitis C treatment leads to improved lipid profiles: A longitudinal Bayesian analysis of the average treatment effect on the treated

Abstract

Co-infection of hepatitis C virus (HCV) in people living with HIV is common, and may lead not only to severe liver diseases but also to negative impacts on cardiovascular health. Efforts have been made to understand the effects of hepatitis C treatment on health beyond its impact on the liver; some studies suggest cardiovascular risk may improve after the HCV cure. We evaluate the effect of sustained virological response (SVR) after hepatitis C virus therapy on serum lipids levels using data from the Canadian Co-infection Cohort Study between 2003 and 2019. Participants who received any hepatitis C virus treatment were followed, with serum lipids measurements collected post-SVR assessment time. We use a novel Bayesian longitudinal approach to estimate a *time-varying* average treatment effect on the treated, using a weighting approach to adjust for confounding factors. In total, 1,149 serum lipid measurements on 272 participants were analyzed. The results show that there is little effect on lipids immediately after achieving SVR, but improvements are achieved over time, suggesting potential benefits of the HCV cure for cardiovascular health.

Keywords: Bayesian inference; causal inference; co-infection; hepatitis C virus; sustained virological response; propensity score.

3.1 Introduction

Hepatitis C virus (HCV) infection affects up to 25% of HIV-infected patients, due largely to the shared routes of infection of injection drug use and men having sex with men (Spradling et al., 2010; Page et al., 2013; Hagan et al., 2015). Chronic HCV infection results in increased rates of cirrho-

sis, liver failure, transplant needs and liver-related death, as well as increased annual health care costs (Centers for Disease Control and Prevention, 1998; Wong et al., 2000; El Saadany et al., 2005). Successful hepatitis C treatment (or HCV cure) is determined by the achievement of a sustained virological response (SVR) defined as hepatitis C viral load being undetectable at 12 weeks after the end of therapy. SVR reduces liver-related complications and mortality in both chronic hepatitis C mono-infected (Coverdale et al., 2004; Veldt et al., 2007) and HIV/HCV co-infected patients (Berenguer et al., 2009). However, less is known about other clinical consequences of the HCV cure, especially in co-infected patients.

Chronic HCV infection is often associated with disorders in lipid metabolism (Kuo et al., 2011; Meissner et al., 2015; Endo et al., 2017; Batsaikhan et al., 2018). Previous studies evaluated the impact of antiviral therapy on serum lipid levels in HCV mono-infected patients. However, there is no consensus in the literature, as different studies came to different conclusions about lipid changes during and after HCV therapy. Findings show that the total cholesterol levels may decrease (Jang et al., 2011), do not change (Kuo et al., 2011; Meissner et al., 2015) or increase during therapy (Endo et al., 2017). In contrast, total cholesterol levels in patients who achieved SVR were reported to increase after treatment (Jang et al., 2011; Kuo et al., 2011; Endo et al., 2017; Batsaikhan et al., 2018). Different results were also found for high-density lipoprotein cholesterol (HDL), finding no change (Meissner et al., 2015) or increase during therapy (Endo et al., 2017). Still other studies indicated that low-density lipoprotein cholesterol (LDL) concentration increased early in HCV therapy, and a further increase was observed after the end of therapy (Meissner et al., 2015; Endo et al., 2017; Batsaikhan et al., 2018). Changes in serum triglycerides also remains an issue of debate: while Endo et al. (2017) show it does not change, Kuo et al. (2011) show that there is an increase during therapy and a decrease in post-treatment period among those who attain SVR, and Meissner et al. (2015) on the other hand, concluded that triglycerides levels decrease during therapy and increase after treatment among those who achieve SVR. However, these studies suffered from some limitations, as they included only one or two measurements of lipids after the end of treatment, and did not considered the more vulnerable subpopulation of HIV/HCV co-infected individuals. The latter point is important to consider, as HIV is itself associated with high total cholesterol, in part due to antiretroviral therapy.

Exceptionally, Townsend et al. (2016) and Mauss et al. (2017) analyzed the effect of HCV therapy on lipid levels in HIV/HCV co-infected patients, and found an increase in total cholesterol and serum LDL levels early in interferon-free therapy, which was maintained during the treatment and at 12 weeks after the end of therapy, while HDL and triglycerides levels were unaffected. In contrast, Mauss et al. (2017) showed that interferon-based therapy decreased the total cholesterol and LDL levels during treatment and there was an increase after the end of therapy. HDL levels decreased

during and after therapy, while triglycerides increased during the treatment. Both studies included a considerable number of measurements of serum lipids during therapy, but had only one measurement after SVR assessment.

We evaluate the effect of the HCV cure on serum lipids levels (total cholesterol, LDL, HDL and triglycerides) among those who achieve SVR during an extended post-SVR assessment period using data from the Canadian Co-infection Cohort Study (Klein et al., 2009). Here, our interest lies on the effects of a point exposure (SVR status) on a longitudinal outcome (each of the four serum lipid levels), and we evaluate the average treatment effect on the treated (ATT) of achieving SVR as a function of time. For this purpose, we provide a Bayesian longitudinal estimator for the ATT to assess the impact of SVR on serum lipids profile over time. To the best of our knowledge, this is the first study to estimate the ATT as a function of time in a longitudinal outcome setting, in either a frequentist or Bayesian paradigm.

3.2 Methods

3.2.1 Study population

We use data from the Canadian Co-infection Cohort, an ongoing prospective study that currently includes more than 1,900 HIV/HCV co-infected participants recruited from HIV clinic populations at 19 centers in Canada since 2003, with follow-up visits scheduled approximately every 6 months (Klein et al., 2009). Socio-demographic, behavioral and quality of life information were self reported at baseline. Information on diagnoses, medical treatments, clinical and laboratory data were updated at each follow-up visit. All participants provided written informed consent. The study was approved by all the research ethics boards of the participating institutions and the community advisory committee of the Canadian HIV Trials Network. Details about the Canadian Co-infection Cohort Study can be found in Klein et al. (2009).

To be eligible for inclusion in the analysis, participants must have received HCV treatment at some point during the follow-up period, and have at least one pre-treatment and at least one post-SVR visit with an available serum lipid measurement. Baseline serum lipids were defined as the last serum lipid level measurement taken prior to the initial date of HCV treatment. If a participant received more than one treatment due to unsuccessful response or reinfection, only the first treatment was considered in the analysis, and only serum lipid measurements taken before the initial date of the second treatment were included. Participants with no SVR status were excluded from the analytic data set.

Exposure

The exposure variable is SVR status. Participants whose HCV treatment was successful, defined as a negative HCV viral load at least 12 weeks after the end of the therapy, are included in the SVR group. Those who failed to achieve an undetecatble HCV viral load by week 12 after treatment (e.g. non-response or relapse), partial or breakthrough, are included in the non-SVR group.

Outcomes

The outcomes of interest are total cholesterol, LDL, HDL and triglycerides levels at each study visit. Participants were followed post-SVR with serum lipids measurements until their last study visit or initial date of the second treatment, if applicable.

Covariates

The baseline variables *a priori* considered as possible confounders of the SVR-lipid relationship are age, gender, ethnicity, current alcohol use, smoking status (ever vs. never), history of injection drug use (ever vs. never), hypertension, HCV viral load, gamma-glutamyl transferase (GGT), platelet count, CD4 cell count, and HIV duration in years. Baseline serum lipid profiles (total cholesterol, LDL, HDL and triglycerides) are also considered as potential confounders.

3.2.2 Statistical analysis

Statistical analyses were performed using software R, version 3.5.0 (R Core Team, Vienna, Austria). We first provide a descriptive analysis, comparing baseline characteristics between the exposure groups. As covariate balance is essential to ensure unbiased estimation of causal effects, standardized mean differences (SMD) (Cohen, 1988) were computed as the difference in mean of a covariate across the exposure groups, divided by the standard deviation in the groups (variations exist for binary or categorical data). As a guideline, 0.1 and 0.25 represent reasonable cut-offs for little and acceptable differences between groups, respectively; larger SMDs indicate lack of balance (Cohen, 1988).

Let Z_i be an indicator for the SVR status, so that $Z_i = 1$ if the i-th participant achieved SVR and $Z_i = 0$ otherwise, for $i = 1, \ldots, m$ participants. Let X_i denote the baseline confounding variables, and assume that all have been accurately recorded.

The effect caused by a HCV cure is a contrast of the outcome that would have been observed if the participant achieved SVR, $Y(1)$, and the outcome that would have been observed if the participants did not achieve SVR, $Y(0)$. The average treatment effect on the treated is defined as $\Delta_{ATT} = E[Y(1) - Y(0)|Z = 1]$, where the average is taken over only who had achieved SVR. We, however, are interested

in a longitudinal setting. Thus, let Y_{ij} be the j-th post-SVR serum lipid level, for $j = 1, \ldots, n_i$, where n_i is the total number of post-SVR serum lipids levels of the i-th participant. Let $Y_{ij}(z)$ denote the potential outcome under the exposure value z, representing the j-th serum lipid that would be obtained by the i-th participant if the exposure was $Z = z$, for $z = 0, 1$. Let T_{ij} be the post-SVR assessment time of the j-th lipid measurement of the i-th participant, where $T_{ij} = 0$ represents the SVR assessment time. We denote by V_i a subset of the components in X_i containing those variables that modify the effect of SVR on lipid levels, and assume a conditional linear mixed model for the potential outcome $Y_{ij}(z)$ such that

$$\begin{aligned} Y_{ij}(z) &= \mu_{ij} + \nu_i + \varepsilon_{ij}, \\ \mu_{ij} &= \beta' X_i + \lambda' V_i z + \delta_0 T_{ij} + \delta_1 T_{ij} z, \end{aligned}$$

with $\nu_i \sim N(0, \sigma_\nu^2)$ an individual-level random effect (random intercept), $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ a random error term, where ν_i and ε_{ij} are mutually independent. In the above equation, we allow both X_i and V_i to contain a leading column of ones to ensure, respectively, that the model contains an intercept and a main effect of exposure. The coefficient vectors β, λ, $\delta = (\delta_0, \delta_1)'$ are unknown and need to be estimated.

Define $Y_i(z,t)$ as the time-indexed potential outcome for the i-th patient under the exposure value z at some post-SVR assessment time $T = t$. Note that the mean difference in potential outcomes can be expressed as a function of time $T = t$, that is, $E[Y_i(1,t) - Y_i(0,t) \mid X_i] = \lambda' V_i + \delta_1 t$. Thus, following the model specification, the effect of HCV cure on serum lipids levels among those who achieved SVR, at time $T = t$, is evaluated as

$$\begin{aligned} \Delta_{ATT}(t) &= E[Y(1,t) - Y(0,t) \mid Z = 1] \\ &= E_{X|Z=1}\left[E\left\{Y(1,t) - Y(0,t) \mid X\right\} \mid Z = 1\right] \\ &= E_{X|Z=1}[\lambda' V \mid Z = 1] + \delta_1 t. \end{aligned}$$

We extend the Bayesian inference approach proposed by Capistrano et al. (2019), following ideas from Saarela et al. (2015a), to a longitudinal outcome setting, allowing estimation of a time-dependent ATT. Following this approach, propensity score-based weighting (Rosenbaum and Rubin, 1983) for ATTs was used to adjust for confounding, where the weights ω_i are equal to 1 for who achieved SVR ($Z_i = 1$) and are obtained by a ratio between marginal and conditional exposure probabilities for those in non-SVR group ($Z_i = 0$). See Appendix 3.5 for details. A Bayesian bootstrap (Newton and Raftery, 1994) was adopted to approximate the posterior distributions of the ATT parameters

λ and δ_1, with sampling weights p_i drawn from a Multinomial distribution, where $p_i = \xi_i/m$ and $\xi = (\xi_1, \ldots, \xi_m) \sim Multinomial(m; m^{-1}, \ldots, m^{-1})$. Thus, exposure probabilities were estimated using baseline covariates via a weighted logistic regression and standardized weights were computed to minimize the baseline covariates differences between the exposure groups. To assess balance in the weighted sample, the sample means and variances in SMD were replaced by their weighted counterparts (Austin and Stuart, 2015a).

The posterior distribution of λ and δ_1, approximated by the Bayesian bootstrap, was used to compute the posterior distribution of $\Delta_{ATT}(t)$. At each iteration l of the bootstrap, we calculated

$$\widehat{\Delta}^{(l)}_{ATT}(t) \quad = \quad \frac{\sum_{i=1}^{m} p_i^{(l)} \widehat{\lambda}'^{(l)} v_i z_i}{\sum_{i=1}^{m} p_i^{(l)} z_i} + \widehat{\delta}_1^{(l)} t.$$

A time-varying 'point estimate' $\widehat{\Delta}_{ATT}(t)$ was then obtained by taking the mean of its posterior distribution at each time t.

3.3 Results

3.3.1 Descriptive analysis

In total, 272 co-infected participants treated for chronic hepatitis C were eligible for participation in our analysis (see Figure 3.1 for details). Together, a total of 1,149 post-SVR study visits with available serum lipids measurements were analyzed. The post-SVR assessment time was measured in months, with a median of 19.9, a mean of 27.8 and a maximum follow-up time of 146 months. Due to the small number of participants in each category of ethnicity, this variable was re-classified as white (1) or other (0). Baseline demographics and metabolic parameters of participants are summarized in Table 3.1. Continuous variables are expressed by their means and standard deviations; binary variables are expressed by their absolute and relative frequencies.

Sustained virological response was achieved by 220 (80.9%) participants. At baseline, the mean age of participants was 49 years. Participants were mostly male, white, current alcohol users, and had a history of smoking and injection drug use. A minority had been diagnosed with hypertension. For comparison purposes, summaries of all characteristics are shown by exposure groups. Overall, the baseline covariates were not similar between the SVR and non-SVR groups. Standardized mean differences for each baseline covariate are also reported in Table 3.1. Distributions of gender, current alcohol use, HIV duration and triglycerides were strongly unbalanced between the exposure groups. Ethnicity, injecting drug use, total cholesterol and LDL exhibited lower SMDs, indicating that their

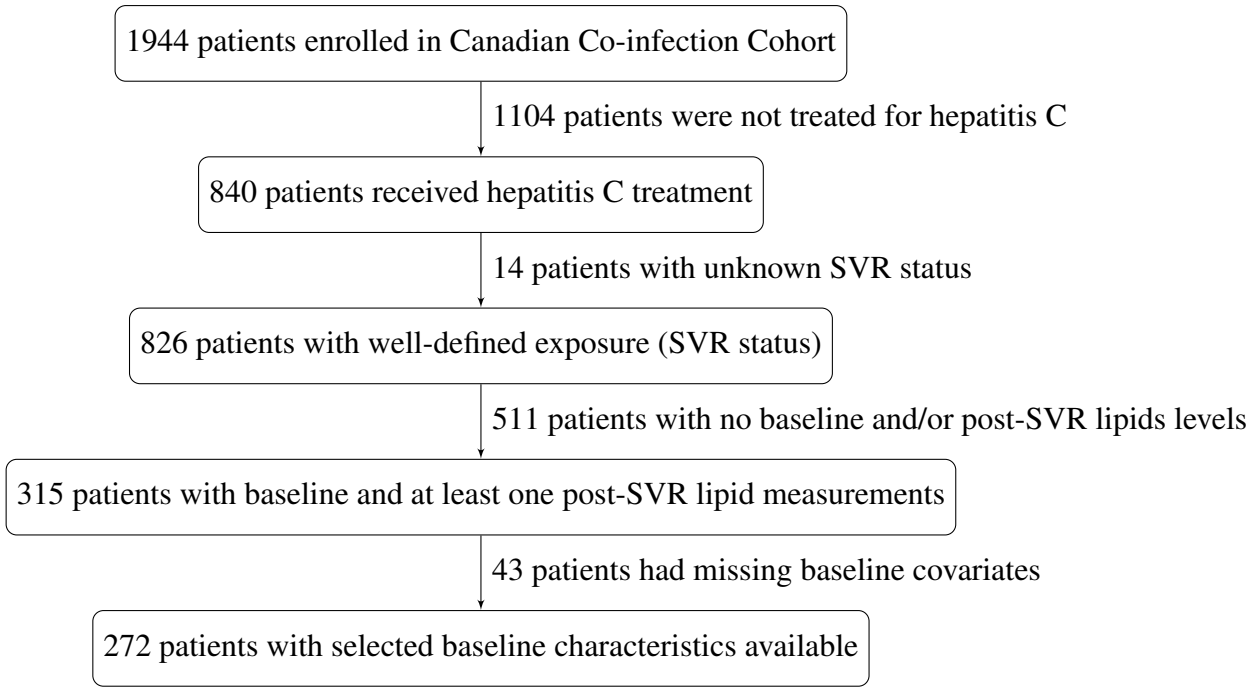


Figure 3.1: Patient flow diagram. As of March 2019, 840 co-infected patients had initiated hepatitis C virus treatment, of whom 826 had well-defined exposure. Of these, 315 patients had at least 1 post-SVR lipid measurement available, of whom only 272 had the selected baseline characteristics available and were included in the analysis. Abbreviation: SVR, sustained virological response.

distributions were acceptably balanced.

3.3.2 Estimation of time-varying causal effects

Due to lower value of SMD, ethnicity, history of injecting drug use, baseline serum total cholesterol and LDL levels were deemed not to be strong confounders and thus we did not included these in the exposure model (although they were included in the outcome model); all other baseline covariates were included in the weighted logistic regression that used sampling weights p_i to estimate the propensity scores and thus compute the standardized weights ω_i. All baseline characteristics in Table 3.1 were included as covariates X_i in the outcome model except LDL due of multicollinearity of the four serum lipids profiles. Interactions between Z and each of age, gender, GGT, and CD4 cell count were also considered; that is, V_i was taken to include these four variables.

After running 1,000 replications of the Bayesian bootstrap, an approximated sample from the posterior distributions of coefficients of the exposure and outcome models was obtained. Weighted SMDs, computed using the posterior median of the coefficients of the exposure models, indicate that differences in baseline characteristics between non-SVR and SVR groups were substantially reduced

Table 3.1: Baseline Demographics and Metabolic Parameters of Eligible Participants From the Canadian Co-infection Cohort Study and Standardized Mean Differences, Canada, 2003-2019.

Characteristics	Total $m = 272$	Non-SVR $m_0 = 52$	SVR $m_1 = 220$	SMD	SMD_w
Age, mean (SD)	49.2 (8.2)	47.8 (8.7)	49.5 (8.0)	0.202	0.067
Gender, no. male (%)	225 (82.7)	48 (92.3)	177 (80.5)	0.351	0.128
Ethnicity, no. white (%)*	222 (81.6)	41 (78.8)	181 (82.3)	0.087	0.040
Current alcohol use, no. (%)	154 (56.6)	23 (44.2)	131 (59.5)	0.310	0.121
Smoking status, no. (%)	244 (89.7)	44 (84.6)	200 (90.9)	0.193	0.243
Injection drug use, no. (%)*	216 (79.4)	43 (82.7)	173 (78.6)	0.103	0.119
Hypertension, no. (%)	40 (14.7)	10 (19.2)	30 (13.6)	0.150	0.042
HCV viral load, mean (SD)	1.23 (0.4)	1.15 (0.4)	1.25 (0.4)	0.224	0.161
GGT, mean (SD)	117.0 (135.2)	146.1 (165.6)	110.2 (126.4)	0.244	0.123
Platelet count, mean (SD)	181.6 (75.7)	167.8 (66.5)	184.8 (77.5)	0.235	0.019
CD4 cell count, mean (SD)	514.4 (271.5)	463.8 (242.4)	526.3 (277.1)	0.240	0.019
HIV duration, mean (SD)	16.1 (8.2)	14.1 (6.7)	16.6 (8.5)	0.336	0.365
Total cholesterol, mean (SD)*	4.02 (1.0)	4.09 (0.9)	4.01 (1.0)	0.086	0.222
LDL, mean (SD)*	2.23 (0.9)	2.21 (0.8)	2.24 (0.9)	0.039	0.131
HDL, mean (SD)	1.14 (0.4)	1.09 (0.4)	1.16 (0.4)	0.167	0.129
Triglycerides, mean (SD)	1.46 (0.7)	1.70 (0.9)	1.40 (0.6)	0.378	0.007

Abbreviations: GGT, gamma-glutamyl transferase; HCV, hepatitis C virus; HDL, high-density lipoprotein cholesterol; LDL, low-density lipoprotein cholesterol; SD, standard deviation; SMD, standardized mean difference; SMD_w, standardized mean difference after weighting; SVR, sustained virological response.
*Variable was not included as covariate in the exposure model.

for the majority of covariates included in the exposure model, as shown in the last column of Table 3.1. While balance was not achieved for all variables, their inclusion in X_i in the outcome model should account for remaining imbalances.

For each serum lipid, an approximate sample from the posterior distribution of coefficients λ and δ was obtained, and posterior mean and 95% credible intervals for ATT were computed for each $t = 0, \ldots, max(T_{ij})$. Figure 3.2 shows the results obtained when each serum lipid of interest (total cholesterol, LDL, HDL and triglycerides, respectively) is considered as the outcome. There was no substantial effect of achieving SVR on serum lipids levels immediately after the SVR assessment time, except for total cholesterol. However, with extended follow-up since HCV cure, a positive effect of SVR on HDL is observed, and there is some suggestion that triglycerides decrease over time. However, there remains little apparent effect of SVR on total cholesterol or on LDL in participants who achieved SVR. On balance, the results suggest that HCV cure has an overall positive impact on the cardiovascular health of participants who achieved SVR in the long term, as measured by serum lipid levels.

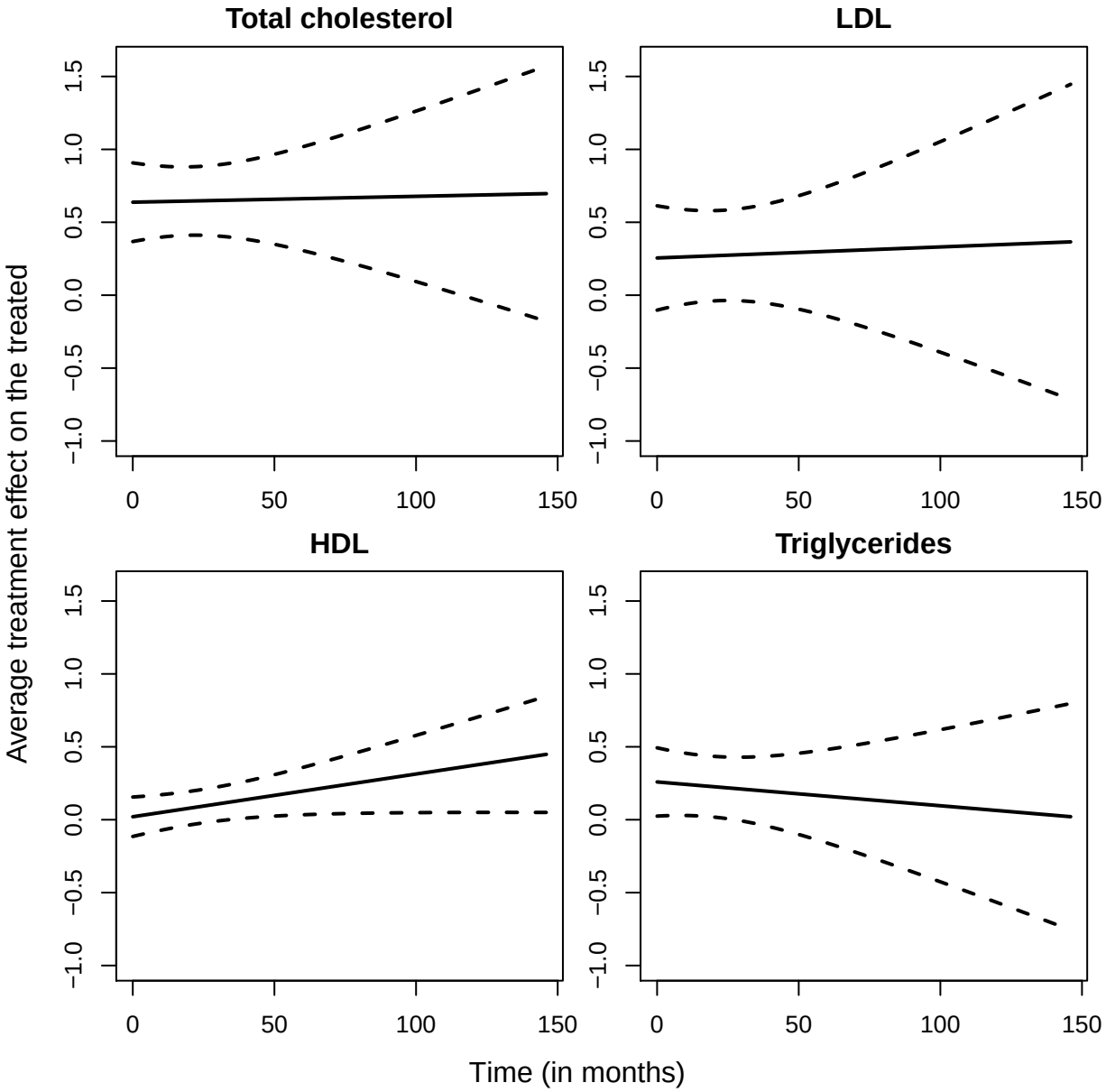


Figure 3.2: Longitudinal ATT (in months) estimated via a Bayesian boostrap approach using multinomial sampling for the effect of SVR on each of four lipids: total cholesterol, LDL, HDL and triglycerides. The solid and dashed lines represent the posterior means and the 95% credible intervals, respectively.

3.4 Discussion

We evaluated the effect of sustained virological response therapy on serum lipids levels among those participants who did, indeed, achieve SVR, using data from the Canadian Co-infection Cohort Study from 2003 to 2019. To do so, we developed a Bayesian estimates of the average treatment effect on the treated in a longitudinal data context wherein the outcomes are time-varying as a function of

post-SVR assessment time. We propose a linear mixed model for each outcome (total cholesterol, LDL, HDL and triglycerides), taking into account the longitudinal nature of the outcomes for each participant who achieved SVR.

Our results identified little effect on lipids immediately after achieving SVR, but suggest that HCV cure has a positive effect on participants who achieved SVR over time, increasing the serum HDL levels and decreasing the triglycerides levels. These findings suggest that achieving SVR improves not only the liver health of treated participants, but also their cardiovascular health.

Although the Canadian Co-infection Cohort data set has a significant number of participants with considerably longer follow-up than any previous study on this topic, our analysis nevertheless suffers from some limitations. First, serum lipid profiles were not available for each follow-up visit, greatly limiting the size of our analytic sample. This more limited sample size, combined with missing information prevented us from considering some additional potential confounders such as HCV genotype, diagnoses of liver diseases including fibrosis and cirrhosis, and use of hypocholesterolaemic drugs.

Among participants who achieved SVR, the sustained virological response may have long-lasting health benefits, not only in the more obvious form of arresting continued liver damage, but also in cardiovascular health through improvements in lipid profiles, by increasing HDL and decreasing triglyceride levels.

3.5 Appendices

3.5.1 Time-varying ATT

Let Z_i be an indicator for the SVR status, so that $Z_i = 1$ if the i-th patient achieved SVR and $Z_i = 0$ otherwise, for $i = 1, \ldots, m$ participants. Let X_i be a p-dimensional vector denoting the baseline confounding variables associated with the i-th patient, and assume that all have been accurately recorded.

Denote by Y_{ij} the j-th post-SVR serum lipid measurement, for $j = 1, \ldots, n_i$, where n_i is the total number of post-SVR serum lipids scores of the i-th patient. Let $Y_{ij}(z)$ denote the potential outcome under the exposure value z, representing the j-th serum lipid that would be obtained by the i-th patient if the exposure was $Z = z$, for $z = 0, 1$. Let T_{ij} be the post-SVR assessment time of the j-th lipid measurement of the i-th patient, where $T_{ij} = 0$ represents the SVR assessment time. We assume a

conditional linear mixed model for the potential outcome $Y_{ij}(z)$ such that

$$\begin{aligned} Y_{ij}(z) &= \mu_{ij} + \nu_i + \varepsilon_{ij}, \text{ with } \nu_i \sim N(0, \sigma_\nu^2) \text{ and } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \\ \mu_{ij} &= \beta' X_i + \lambda' V_i z + \delta_0 T_{ij} + \delta_1 T_{ij} z, \end{aligned}$$

where ν_i is a random intercept, ε_{ij} is a random error term, ν_i and ε_{ij} are mutually independent, and V_i is a q-dimensional subset of X_i containing those variables that modify the effect of SVR on lipid levels, with both X_i and V_i containing a leading column of ones to ensure, respectively, that the model contains an intercept and a main effect of exposure. The coefficient vectors $\beta = (\beta_1, \ldots, \beta_p)'$, $\lambda = (\lambda_1, \ldots, \lambda_q)'$, $\delta = (\delta_0, \delta_1)'$ are also assumed to be unknown and need to be estimated.

Define $Y_i(z,t)$ as the time-indexed potential outcome for the i-th patient under the exposure value z at some post-SVR assessment time $T = t$. Note that the mean difference in potential outcomes can be expressed as a function of time $T = t$, that is, $E[Y_i(1,t) - Y_i(0,t) \mid X_i, \beta, \lambda, \delta, \sigma_\nu^2, \sigma_\varepsilon^2] = \beta' X_i + \lambda' V_i + \delta_0 t + \delta_1 t - (\beta' X_i + \delta_0 t) = \lambda' V_i + \delta_1 t$. Thus, following the model specification, the effect of HCV cure on serum lipids levels among those who achieved SVR, that is, the average treatment effect on the treated, at time $T = t$, is evaluated as

$$\begin{aligned} \Delta_{ATT}(t) &= E_{Y|Z=1}[Y(1,t) - Y(0,t) \mid Z = 1] \\ &= E_{X|Z=1}\left[E_{Y|X,Z=1}\left\{Y(1,t) - Y(0,t) \mid X, Z = 1\right\} \mid Z = 1\right] \\ &= E_{X|Z=1}\left[E_{Y|X,Z=1}\left\{Y(1,t) - Y(0,t) \mid X, \beta, \lambda, \delta, \sigma_\nu^2, \sigma_\varepsilon^2\right\} \mid Z = 1\right] \\ &= E_{X|Z=1}[\lambda' V \mid Z = 1] + \delta_1 t. \end{aligned}$$

3.5.2 Inference procedure

The estimator for a time-dependent ATT builds on the ideas of Capistrano et al. (2019), who derived a Bayesian estimator of Δ_{ATT} motivated by a methodology initially proposed by Saarela et al. (2015a). Saarela et al. (2015a) derived an approach based on observational and experimental measures for exchangeable observable sequences via a predictive modeling view based on de Finetti's representation De Finetti (1974). Considering the estimation of ATT as a decision problem (Walker, 2010), the inference procedure is based on the maximization of a posterior predictive expected utility function (Saarela et al., 2015b). The weighted likelihood bootstrap, proposed by Newton and Raftery (1994), was adopted to approximate posterior distribution of interest and avoid the issue of feedback that arises in Bayesian causal estimation relying on a joint likelihood. Thus, we obtain a non-parametric posterior predictive density using sampling weights $p = (p_1, \cdots, p_m)$ defined by

$p = \xi/m$, where $\xi = (\xi_1, \ldots, \xi_m) \sim Multinomial(m; m^{-1}, \ldots, m^{-1})$. A brief review of the weighted likelihood bootstrap can be found in Capistrano et al. (2019).

Following this approach, propensity-score based weighting (Rosenbaum and Rubin, 1983) was used to adjust for confounding. The weights ω_i are obtained from the marginal and conditional treatment models, $P_{\mathscr{E}}(Z = z_i)$ and $P_{\mathscr{O}}(Z = z_i|X_i)$, respectively. Note that under the experimental setting $\mathscr{E}$, the distribution of Z is independent of X, whereas under the observational setting $\mathscr{O}$, the distribution of Z is conditional on X. In practice, the weights for ATTs are ratios of probabilities

$$\omega_i = \frac{P(Z = z_i|\gamma)}{P(Z = z_i|X_i, \alpha)} \frac{P(Z = 1|X_i, \alpha)}{P(Z = 1|\gamma)},$$

where γ and α are the parameters associated with the marginal and conditional treatment models, respectively. That is, all patients who achieved SVR ($Z_i = 1$) will receive a weight of $\omega_i^1 = 1$ and all patients who do not achieved SVR ($Z_i = 0$) will receive a weight of $\omega_i^0 = \frac{P(Z=0|\gamma)}{P(Z=0|X_i,\alpha)} \frac{P(Z=1|X_i,\alpha)}{P(Z=1|\gamma)}$.

Assume that the marginal distribution of Z_i follows a Bernoulli distribution with mean e^* such that $\text{logit}(e^*) = \log\{e^*/(1-e^*)\} = \gamma$, where the intercept γ is unknown. Suppose that, conditional on X_i, the exposure Z_i follows a Bernoulli distribution with mean $e_i = e(X_i)$, that is, $Z_i \mid X_i \sim Bernoulli(e_i)$. Also, suppose that the propensity score e_i follows a logistic model $\text{logit}(e_i) = \alpha' X_i$, where the coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)'$ are also unknown. The weighted maximum likelihood estimates of the coefficients γ and α are obtained, respectively, as

$$\widehat{\gamma} = \arg\max_\gamma \left[\sum_{i=1}^{m} p_i \log P(z_i|\gamma) \right] = \arg\max_\gamma \left[\sum_{i=1}^{m} p_i \left[z_i \log(e^*) + (1 - z_i) \log(1 - e^*) \right] \right]$$

and

$$\widehat{\alpha} = \arg\max_\alpha \left[\sum_{i=1}^{m} p_i \log P(z_i|x_i, \alpha) \right] = \arg\max_\alpha \left[\sum_{i=1}^{m} p_i \left[z_i \log(e_i) + (1 - z_i) \log(1 - e_i) \right] \right].$$

Then, estimates for importance sampling weights ω_i are given by $\widehat{\omega}_i = \frac{P(Z = z_i|\widehat{\gamma})}{P(Z = z_i|x_i, \widehat{\alpha})} \frac{P(Z = 1|x_i, \widehat{\alpha})}{P(Z = 1|\widehat{\gamma})}$. Therefore, the weighted maximum likelihood estimate of the parameter vector $\theta = (\beta', \lambda', \delta', \sigma_v^2, \sigma_\varepsilon^2)'$ in the outcome model is obtained as

$$\widehat{\theta} \approx \arg\max_\theta \left[\sum_{i=1}^{m} \widehat{\omega}_i \, p_i \sum_{j=1}^{n_i} \log P(y_{ij}|x_i, z_i, t_{ij}, \theta) \right].$$

As $\Delta_{ATT}(t)$ is a function of the parameter vector θ, Bayesian estimates for $\Delta_{ATT}(t)$ are then obtained by samples from the posterior distribution of θ, at each time t. Thus, at each replication l

($l = 1, 2, \cdots, L$) of the weighted likelihood bootstrap, we computed

$$\widehat{\Delta}_{ATT}^{(l)}(t) = \frac{\sum_{i=1}^{n} p_i^{(l)} \widehat{\lambda}^{(l)'} v_i z_i}{\sum_{i=1}^{n} p_i^{(l)} z_i} + \widehat{\delta}_1^{(l)} t.$$

A time-varying 'point estimate' is then obtained, for example, by considering the posterior mean of the distribution of $\Delta_{ATT}(t)$ at each time t, that is, $\widehat{\Delta}_{ATT}(t) = \frac{1}{L} \sum_{l=1}^{L} \widehat{\Delta}_{ATT}^{(l)}(t)$.

Chapter 4

Can a Bayesian be subjective about propensity score balancing?

Abstract

Propensity score methods are a powerful tool for balancing the distributions of measured covariates in different treatment groups, in order to estimate causal effects that are not confounded by measured confounders. Bayesian propensity score approaches have recently been gaining ground as an alternative approach to frequentist estimation. However, few studies have investigated the balancing of covariates in a Bayesian context. We investigate, via simulation, the ability of a propensity score to provide balance by computing and comparing standardized mean differences of confounding variables using different prior distributions and sample sizes. Simulation studies indicate that both frequentist and Bayesian methods improve balance over no adjustment, however, informative priors can lead to poor balance, especially for small sample sizes. A demonstration of the impact of different priors in real data is also provided: considering small random samples of the U.S. National Ambulatory Medical Care Survey reinforces that prior specification should be performed carefully – or avoided in favour of non-informative priors – to ensure covariate balance.

Keywords: Causal inference; imbalance; inverse probability weighting; prior information; propensity score.

4.1 Introduction

4.1.1 Bayesian causal inference

Causal inference from observational data can be drawn under a number of assumptions that permit analysts to view observational studies as conditionally randomized experiments. In randomized experiments, all measured covariates are equally distributed between the treated and the control groups.

When treatment is not randomly assigned, this balancing is not guaranteed and the estimated effect have no causal interpretation if the treatment groups differ with respect to the characteristics that affect the outcome. Without adequate adjustment for confounding, bias in the estimation of a treatment effect can arise. Propensity score techniques (Rosenbaum and Rubin, 1983) are widely employed in the observational data context for balancing the distributions of measured covariates in different treatment groups, in order to estimate causal effects that are not biased by measured confounders.

Although many Bayesian propensity score approaches have been proposed (McCandless et al., 2009, 2010; An, 2010; Hill, 2011; Kaplan and Chen, 2012; Zigler et al., 2013; Saarela et al., 2015a; Roy et al., 2018; Keil et al., 2018; Xu et al., 2018), few have investigated the balancing properties in the context of a Bayesian propensity score adjustment. For instance, Zigler et al. (2013) demonstrated the potential for feedback (McCandless et al., 2009) to distort the nature of the propensity score in a joint Bayesian analysis. They assessed the covariate balance through an average absolute within-stratum difference between the treatment groups. Zigler and Dominici (2014) proposed methods for Bayesian model averaging to address uncertainty in the propensity score model specification. Covariate balance was assessed by comparing covariate prevalence between treatment groups within propensity score subclass. Chen and Kaplan (2015) explicitly studied covariate balance in the two-step Bayesian propensity score approach proposed by Kaplan and Chen (2012) using standardized mean differences and variance ratios; they found, as expected, both Bayesian and frequentist propensity score approaches substantially reduce the initial imbalance. The authors investigated the use of non-informative priors versus informative priors centered on the true data-generating values, using an approach that samples from the posterior distribution of the propensity score to approximate the posterior distribution of the treatment effect. Kaplan and Chen (2014) provided a Bayesian model averaging approach via Markov chain Monte Carlo procedure to account for model uncertainty and examined the differences in causal estimates when incorporating informative priors in the model averaging step. Covariate balance analysis show that Bayesian model averaging approach provide comparably good covariate balance compared to the two-step Bayesian propensity score approach. There remains a lack of clarity of the impact of prior specification on the propensity score in Bayesian methods as: (i) small samples are more sensitive to prior information, and both Chen and Kaplan (2015) and Kaplan and Chen (2014) considered moderately large samples, and (ii) the approach of drawing from the posterior of the propensity scores has been shown to have poor small sample performance (Saarela et al., 2015a), since covariate balance must be achieved within a specific sample and thus is best achieved by fixing propensity scores to their best estimates (Rosenbaum and Rubin, 1983).

In what follows, we provide a more comprehensive investigation of the impact of including prior information on the propensity score into a Bayesian causal inference procedure in small samples.

We employ a two-step procedure to avoid feedback while fixing the propensity score at the posterior mean thus ensuring balance is calibrated to the dataset. Our aim is thus to assess the impact of being subjective when selecting a prior distribution for the parameters of the treatment model.

4.1.2 Background: notation and concepts

Consider Z an indicator for treatment ($z = 1$: treated, $z = 0$: untreated) and Y a variable representing the observed outcome. Let $Y(z)$ denote the potential outcome under the treatment value z, representing the outcome that would be obtained if the received treatment was $Z = z$, for $z = 0, 1$. Thus, the observed outcome for the i-th subject can be expressed as $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$, for all $i = 1, \ldots, n$.

In observational studies, the treatment assignment may be determined by many factors. If these factors also affect the outcome, then there is confounding of the effect of treatment on the outcome. Due to lack of independence between the treatment groups and these confounding factors, the association effect may not be equal a causal effect without further assumptions and adjustments.

Causal inference typically focuses on average causal effects, computed as a contrast between the averages of the potential outcomes. Two effects frequently considered are the average treatment effect (ATE) in the population of interest as a whole, defined by $\Delta_{ATE} = E[Y(1) - Y(0)]$, and the average treatment effect on the treated (ATT), defined by $\Delta_{ATT} = E[Y(1) - Y(0) \mid Z = 1]$.

The propensity score is a powerful tool for balancing the distributions of measured covariates in different treatment groups, in order to estimate causal effects that are not confounded by such covariates. It was defined by Rosenbaum and Rubin (1983) as the conditional probability of treatment assignment $Z = 1$ given the p-dimensional vector of measured covariates X, that is $e(X) = P(Z = 1 \mid X)$. As independence of the treatment groups within the levels of the propensity score $e(X)$ implies independence within the levels of the covariates X, a p-dimensional balancing problem can be reduced to a single dimension. The propensity score can thus be used to adjust for confounding via regression, stratification, matching, or weighting so as to estimate average causal effects. See Deb et al. (2016) for advantages and drawbacks of these approaches.

Inverse probability of treatment weighting uses the propensity score to compute weights for each individual, thus creating a pseudo-population in which the distribution of measured covariates is independent of treatment assignment. The method, effective at reducing bias, uses all available data and has been applied in a Bayesian context for estimating both Δ_{ATE} and Δ_{ATT} (Kaplan and Chen, 2012; Saarela et al., 2015a; Capistrano et al., 2019).

4.2 Estimation

4.2.1 Estimands

Suppose that the distribution of the dichotomous treatment Z_i conditioned on measured covariates X_i follows a Bernoulli distribution with mean $e_i = e(X_i)$, that is, $Z_i \mid X_i \sim Bernoulli(e_i)$. Suppose that the propensity score follows a logistic model $\text{logit}(e_i) = \log\{e_i/(1-e_i)\} = \alpha' X_i$, where X_i is a p-dimensional vector of covariates associated with the i-th unit, and the coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)'$ are assumed to be unknown. Assuming that the marginal distribution of Z_i follows a Bernoulli distribution with mean e^*, we assign a logistic model to e^* such that $\text{logit}(e^*) = \gamma$, where the intercept γ is unknown.

If the outcome is continuous, we may assume a conditional model for the potential outcome such that $Y_i(z) = f(\beta, X_i) + zg(\lambda, V_i) + \varepsilon_i$, for some functions f and g, where $\varepsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed for $i = 1, \ldots, n$, and V_i is a q-dimensional vector containing a subset of components of X_i, with both X_i and V_i containing a column of ones. Under this model, the difference in potential outcomes can be expressed as $E\{Y_i(1) - Y_i(0) \mid X_i, \beta, \lambda, \sigma^2\} = g(\lambda, V_i)$, where λ is assumed unknown. Then $\Delta_{ATE} = E_X\left[E_{Y|X}\left\{Y(1) - Y(0) \mid X, \beta, \lambda, \sigma^2\right\}\right] = E_X\{g(\lambda, V)\}$ and $\Delta_{ATT} = E_{X|Z=1}\left[E_{Y|X,Z=1}\left\{Y(1) - Y(0) \mid X, Z = 1, \beta, \lambda, \sigma^2\right\} \mid Z = 1\right] = E_{X|Z=1}\{g(\lambda, V) \mid Z = 1\}$. For instance, in a linear model, $Y_i(z) = \beta' X_i + \lambda' z V_i + \varepsilon_i$, we have $g(\lambda, V_i) = \lambda' V_i$ and $\Delta_{ATE} = E_X(\lambda' V)$ and $\Delta_{ATT} = E_{X|Z=1}(\lambda' V | Z = 1)$.

4.2.2 A two-step approach to estimate Bayesian causal effects

Estimation of and inference on the causal effects is performed in a two-step inverse probability of weighting procedure. First, the propensity score is estimated for each unit from the covariates measured through the treatment models and weights are computed. Second, the parameters of the outcome model are estimated in the weighted sample and the quantities of interest are computed. Assume that $\pi(\alpha)$ and $\pi(\gamma)$ denote a prior distribution for the parameters of conditional and marginal treatment models, respectively. The posterior distribution for α, $\pi(\alpha \mid x, z)$, is proportional to $\sum_{i=1}^{n} (e_i)^{z_i}(1-e_i)^{(1-z_i)} \times \pi(\alpha)$ with $e_i = \{1 + \exp(-\alpha' x_i)\}^{-1}$. The posterior distribution for γ, $\pi(\gamma \mid z)$, is proportional to $\sum_{i=1}^{n} (e^*)^{z_i}(1-e^*)^{(1-z_i)} \times \pi(\gamma)$ with $e^* = \{1 + \exp(-\gamma)\}^{-1}$. The Bayesian inverse weighting approach is achieved via the following steps:

1. Estimates for α and γ are obtained from the posterior mean of the respective posterior distribution. That is, $\widehat{\alpha} = E(\alpha \mid x, z)$ and $\widehat{\gamma} = E(\gamma \mid z)$.

2. Inverse probability of treatment weights are computed according to the estimand of interest. For Δ_{ATE}, the weights are $\widehat{\omega}_i^{ATE} = P(Z = z_i \mid \widehat{\gamma})/P(Z = z_i \mid x_i, \widehat{\alpha})$. For Δ_{ATT}, the weights are $\widehat{\omega}_i^{ATT} = \{P(Z = z_i \mid \widehat{\gamma})/P(Z = z_i \mid x_i, \widehat{\alpha})\} \times \{P(Z = 1 \mid x_i, \widehat{\alpha})/P(Z = 1 \mid \widehat{\gamma})\}$ (Austin and Stuart, 2015a; Capistrano et al., 2019).

3. The posterior distribution for λ, $\pi(\lambda \mid y, x, z)$, is approximated by a Bayesian procedure on the weighted likelihood function (Capistrano et al., 2019) using a prior distribution $\pi(\lambda)$; that is, $\prod_{i=1}^{n} \left\{P(Y_i \mid x_i, z_i, \beta, \lambda, \sigma^2)\right\}^{\widehat{\omega}_i} \times \pi(\lambda)$. The estimate $\widehat{\lambda}$ is obtained from the posterior mean.

4. The estimand of interest is computed using estimates for λ; that is, $\widehat{\Delta}_{ATE} = E_X\{g(\widehat{\lambda}, V)\}$ and $\widehat{\Delta}_{ATT} = E_{X|Z=1}\{g(\widehat{\lambda}, V) \mid Z = 1\}$. In a linear model, for instance, $\widehat{\Delta}_{ATE} = n^{-1} \sum_{i=1}^{n} \widehat{\lambda}' v_i$ and $\widehat{\Delta}_{ATT} = \sum_{i=1}^{n} \widehat{\lambda}' v_i z_i / \sum_{i=1}^{n} z_i$.

4.2.3 Assessing balance using the standardized mean difference

In causal inference, covariate balance is essential to ensuring unbiased estimation of causal effects. Side-by-side boxplots and empirical cumulative distribution functions can be used to visually compare the distributions of continuous covariates between treatment groups. While a quantification of this difference can be given by a Kolmogorov-Smirnov test statistic defined as the maximum vertical distance between the two empirical cumulative distribution functions of a covariate in the treatment groups, the most common statistic for assessing balance is the standardized mean difference, being calculated as the difference in means of a covariate across the treatment groups, divided by the standard deviation in the treatment groups (variations exist for binary or categorical data). To assess balance in an inverse probability of treatment weighted sample, the sample means and variances are replaced by their weighted counterparts (Austin and Stuart, 2015a). That is, for each covariate X, the standardized mean differences are:

$$smd(X) = \frac{\mid \bar{x}(1) - \bar{x}(0) \mid}{\sqrt{[\{s^2(0) + s^2(1)\}/2]}} \quad \text{and} \quad smd_{\omega}(X) = \frac{\mid \bar{x}_{\omega}(1) - \bar{x}_{\omega}(0) \mid}{\sqrt{[\{s^2_{\omega}(0) + s^2_{\omega}(1)\}/2]}},$$

where, for each treatment group with n_z units that received treatment $Z = z$, $\bar{x}(z)$ and $s^2(z)$ denote the sample mean and variance, $\bar{x}_{\omega}(z) = (\sum_{i=1}^{n_z} \omega_i)^{-1} \sum_{i=1}^{n_z} \omega_i x_i$ is the weighted mean, $s^2_w(z) = \{(\sum_{i=1}^{n_z} \omega_i)^2 - \sum_{i=1}^{n_z} \omega_i^2\}^{-1} \sum_{i=1}^{n_z} \omega_i \times \sum_{i=1}^{n_z} \omega_i (x_i - \bar{x}_{\omega})^2$ is the weighted variance, and ω_i is the inverse probability of treatment weight assigned to the i-th unit.

As a guideline, 0.1 and 0.25 represent reasonable cut-offs for little and acceptable differences between groups, respectively (Cohen, 1988); larger standardized mean differences indicate lack of balance that could suggest confounding is too substantial for unbiased estimation.

4.3 Simulation study

We evaluate the impact of incorporating various priors into a Bayesian propensity score procedure. For simplicity, we considered only two covariates: $X_1 \sim N(1,1)$ and $X_2 \sim N(0,1)$. Treatments are generated as $Z_i \mid X_{1i}, X_{2i} \sim Bernoulli(e_i)$, where $logit(e_i) = 0.5 + 0.8X_{1i} - X_{2i}$, for $i = 1,\ldots,n$. The observed outcome Y_i is generated from a normal distribution with mean $\mu_i = 2.0 + 0.4X_{1i} - 0.6X_{2i} + 2.0Z_i + 1.5Z_iX_{1i}$ and standard deviation 0.4. Therefore, the true values of the causal effects are $\Delta_{ATE} = 3.5$ and $\Delta_{ATT} \approx 3.753$. We generate 100 sets of simulated data, and from each of these, we draw a sub-sample of size $n = 30, 50, 70, 100, 120$ units.

If we wish to use an informative prior for the coefficient vector α of a logistic model, we may specify independent normal prior distributions for each component of α whose parameters are obtained via specification of two quantiles with associated probabilities (Wakefield, 2013), which may be computed on the odds ratio scale for more interpretability. For all analyses, we assign a normal non-informative prior distribution $N(0,100)$ for the intercepts of conditional and marginal treatment models (α_0 and γ, respectively). For the other parameters in the conditional treatment model, we explore the prior distributions on Table 4.1.

Table 4.1: Prior specification for the coefficients of the treatment model for the simulated study

	Prior for α_1	Odd ratio of α_1 (95% credible interval)	Prior for α_2	Odd ratio of α_2 (95% credible interval)
P0	$N(0.0, 100)$	1.00 (0.0;3.2e8)	$N(0.0, 100)$	1.00 (0.0;3.2e8)
P1	$N(1.0, 0.25)$	2.72 (1.02;7.24)	$N(-1.2, 0.5)$	0.30 (0.08;1.20)
P2	$N(0.5, 0.25)$	1.65 (0.62;4.39)	$N(-0.5, 0.5)$	0.61 (0.15;2.43)
P3	$N(0.0, 0.25)$	1.00 (0.38;2.66)	$N(-0.1, 0.5)$	0.90 (0.23;3.62)
P4	$N(-0.3, 0.25)$	0.67 (0.25;1.79)	$N(0.3, 0.5)$	1.35 (0.34;5.40)

For the coefficients in the outcome model we assign independent, non-informative normal prior distributions $N(0,100)$. For the standard deviation of the measurement error we assign an uniform distribution $U(0,10)$ (Gelman, 2006). Samples from the posterior distribution of the parameters were obtained using Markov chain Monte Carlo methods through JAGS (Plummer, 2017). For each dataset we ran 15,000 iterations, used the first 5,000 as burnin and stored every tenth iteration resulting in a sample of size 1,000. Trace plots of the chains suggested that convergence was reached. For the sake of comparison, we also obtain estimates using inverse probability of treatment weighting under a frequentist approach and compute unweighted (naive) estimates from a model without any adjustment for confounding.

The results below focus on the balance of the two covariates across the treatment groups. Figure 4.1 shows the mean and 95% posterior credible intervals (and, in the case of the frequentist estimator, a 95% confidence interval) of the standardized mean differences for the covariates X_1 and X_2. These differences were computed using $smd(X)$ for the original samples and $smd_{\omega}(X)$ after computing the inverse probability of treatment weighting using $\widehat{\omega}^{ATT}$ as weights. On average, and as expected, both frequentist and Bayesian methods substantially reduced the initial imbalance. All

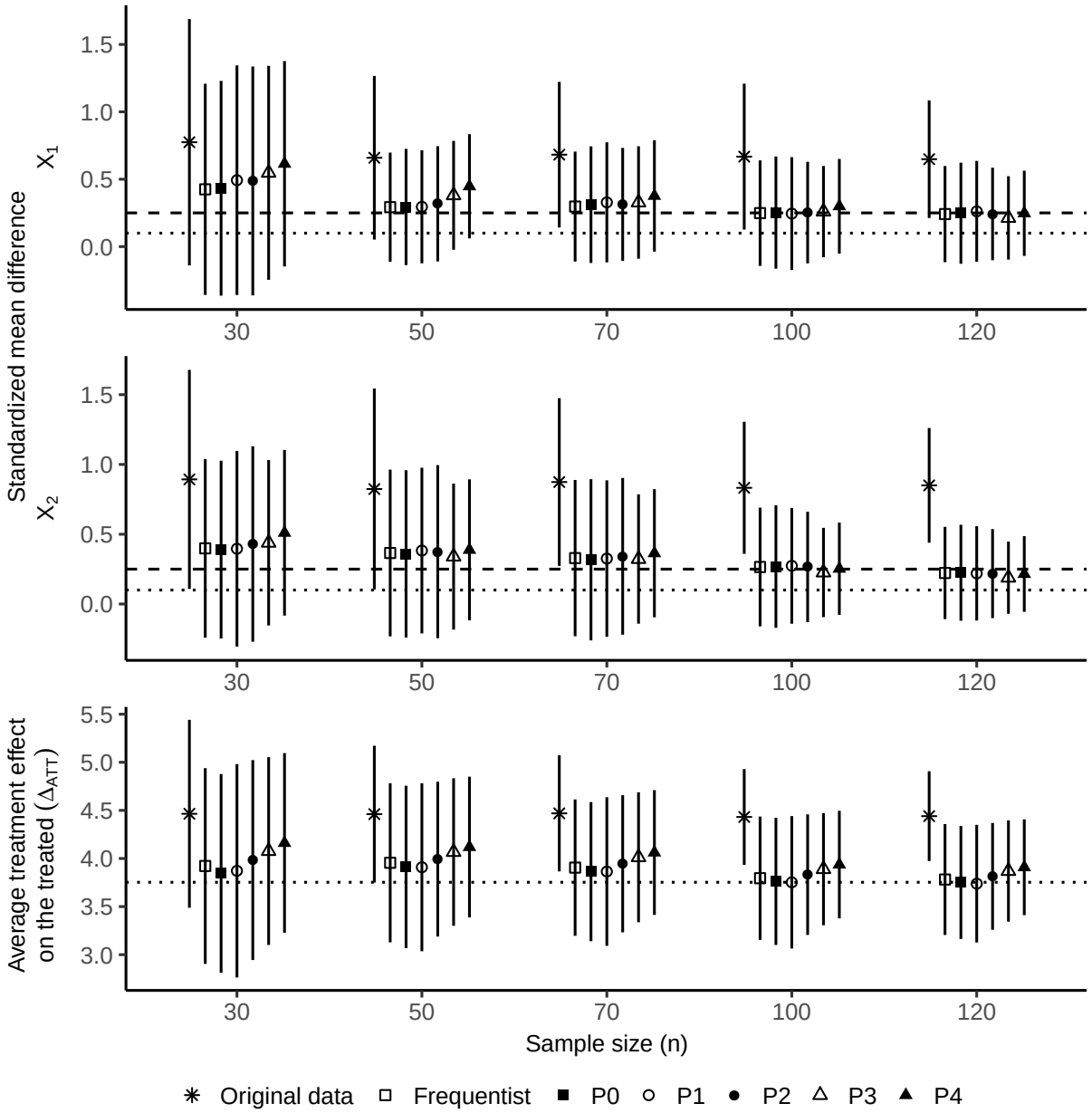


Figure 4.1: Standardized mean differences for X_1 and X_2, and estimates for average treatment effect in simulated data, based on the original sample and following the use of $\widehat{\omega}^{ATT}$ as inverse probability of treatment weights. The vertical bars represent the 95% posterior credible (or frequentist confidence) intervals. The dotted and dashed lines in first panels represent the cut-offs points of 0.1 and 0.25, respectively. In the last panels, the dotted line represents the true value of Δ_{ATT}.

weighted models, regardless of the prior, provide improve balance over the original, unadjusted sample. Using non-informative priors (and frequentist inverse weighting) provide slightly better covariate balance, particularly in the smallest samples. As the sample size increases, the effect of the prior is diminished and the balance improves across all priors. Figure 4.1 also shows the mean and 95% posterior credible intervals for Δ_{ATT} under different prior specifications for the parameters of treatment models (or 95% confidence intervals under frequentist estimator). The use of non-informative priors yield less biased estimators, and even in the larger samples – where balance appeared comparable across the different priors – bias is slightly bigger for informative priors. Using the posterior mode of the propensity score rather than the posterior mean did not significantly alter the results. Summaries of the posterior distribution of the coefficients of the treatment models, and findings for the average treatment effect (Δ_{ATE}) are given in the appendix 4.6.1.

4.4 U.S. National Ambulatory Medical Care Survey data analysis

To examine the potential for balance – or lack thereof – in a more realistic setting, we consider data from the U.S. National Ambulatory Medical Care Survey, which comprises of records of visits to office-based physicians and community health centers conducted by the National Center for Health Statistics. Assume one is interested to investigate the effect of depression on some outcome of interest. Therefore, assume that depression is the exposure of interest, that is $Z = 1$ if the patient was diagnosed with depression, and $Z = 0$ otherwise. Further, we suppose age, gender and smoking are confounders $X = (X_1, X_2, X_3)$.

The original dataset from 2016 has information on 5846 individuals (excluding the missing cases), where 619 (10.58%) have a diagnosis of depression. A random sample of $n = 30, 50, 70, 100, 120$ patients was selected, keeping those from the smallest sample size to the largest one. Due to the small number of treated in the dataset, we opted to keep the proportion of exposed individuals at around 50% to avoid the absence of exposure individuals in any of the selected samples.

We assign a non-informative normal prior distribution $N(0, 100)$ for the intercept in both the conditional and marginal treatment models. For the other coefficients in the conditional treatment model, we explore the prior distributions described on Table 4.2. Samples from the posterior distribution were obtained in the same fashion as for the previous simulation study.

Panels in Figure 4.2 show the standardized mean differences of the three potential confounders in the original sample, and in the weighted sample using $\widehat{\omega}^{ATT}$ as weights. For samples of size equal to, or greater than $n = 70$, the differences computed for age are bellow the cut-off line of 0.25 regardless of the prior specification. However, the results for gender and smoking suggest the weights computed

Table 4.2: Prior specification for the coefficients of the treatment model for the U.S. National Ambulatory Medical Care Survey data

	Prior for α_1	Odd ratio of α_1 (95% credible interval)	Prior for α_2	Odd ratio of α_2 (95% credible interval)	Prior for α_3	Odd ratio of α_3 (95% credible interval)
P0	$N(0.0, 100)$	1.00 (0.0;3.2e8)	$N(0.0, 100)$	1.00 (0.0;3.2e8)	$N(0.0, 100)$	1.00 (0.0;3.2e8)
P1	$N(-0.1, 0.05)$	0.90 (0.58;1.40)	$N(1.0, 0.1)$	2.72 (1.46;5.05)	$N(1.0, 0.1)$	2.72 (1.46;5.05)
P2	$N(0.1, 0.05)$	1.11 (0.71;1.71)	$N(0.7, 0.1)$	2.01 (1.08;3.74)	$N(0.8, 0.1)$	2.23 (1.20;4.14)
P3	$N(-0.4, 0.05)$	0.67 (0.43;1.04)	$N(0.3, 0.1)$	1.35 (0.73;2.51)	$N(0.4, 0.1)$	1.49 (0.80;2.77)
P4	$N(0.4, 0.05)$	1.49 (0.96;2.31)	$N(-0.1, 0.1)$	0.90 (0.49;1.68)	$N(-0.1, 0.1)$	0.90 (0.49;1.68)

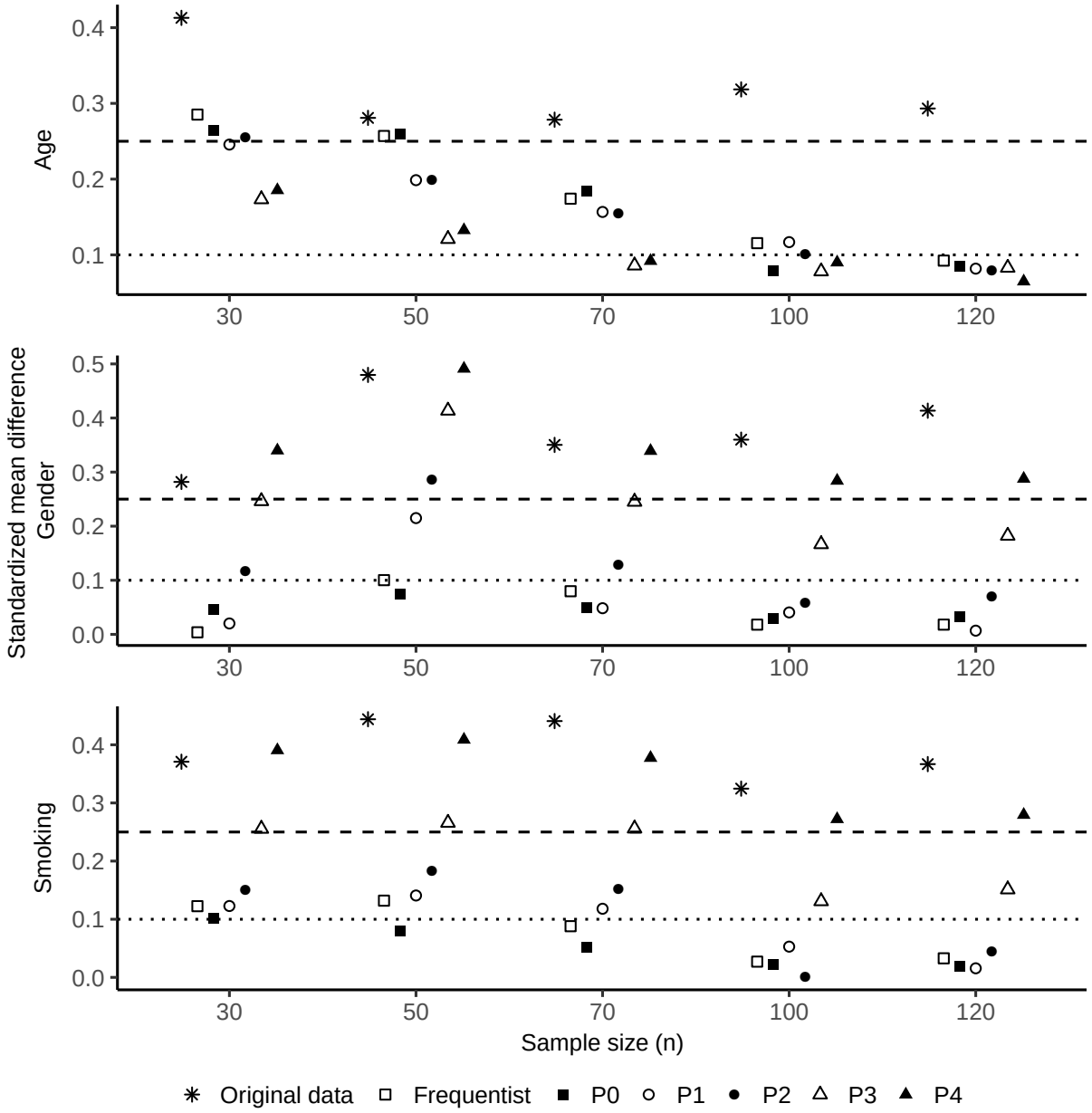


Figure 4.2: Standardized mean differences for the covariates of the U.S. National Ambulatory Medical Care Survey data, based on the original sample and following the use of $\widehat{\omega}^{ATT}$ as inverse probability weights. The dotted and dashed lines represent the cut-offs points of 0.1 and 0.25, respectively.

from informative prior P4 do not provide balance, while balancing from prior P3 for both covariates was achieved for sample sizes greater than 70. In general, standardized mean differences from non-informative prior (and the frequentist approach), as well as informative priors P1 and P2 provide good balance. Appendix 4.6.2 shows the summaries of the posterior distributions of the coefficients in the treatment model and the standardized mean differences computed using $\widehat{\omega}^{ATE}$ as weights.

4.5 Discussion

Prior elicitation plays an important role in Bayesian inference. In causal inference, however, one needs to be more cautious in specifying prior distributions of the parameters in the treatment model. An important step in causal inference is to ensure balance of the distributions of measured covariates in different treatment groups, so that the estimate of causal effects are unbiased. We used the standardized mean difference to investigate how different prior specifications of the coefficients in the treatment model might affect balance. Non-informative priors provided improved balance in smaller samples, while in larger samples there was little difference across the choice of prior. This suggests the routine choice of non-informative priors for propensity score parameters.

4.6 Appendices

4.6.1 Additional results for the simulation study

This appendix includes additional results from the simulation study. Figure 4.3 shows the mean estimates for parameters of treatment models under the frequentist and Bayesian approaches, and their confidence/credibility intervals, for datasets with sample sizes $n = 30, 50, 70, 100, 120$. For the Bayesian approach, it represents the distribution of the posterior mean of the coefficients of treatment models, considering each one of the different prior distributions listed on Table 1 of the paper. As expected, frequentist and non-informative Bayesian estimation are similar and, as n increases, the estimates tend to become more concentrated around the true value of the parameters. It indicates that the prior specification must be made cautiously, especially for data sets with relatively small sample sizes.

Figure 4.4 shows the mean and 95% confidence/posterior credible intervals of the standardized mean differences for the covariates X_1 and X_2, based on the original distribution of the covariates and their respective distributions after balancing using $\widehat{\omega}^{ATE}$ as weights. On average, both frequentist and Bayesian methods substantially reduced the initial imbalance of the original, unadjusted sample.

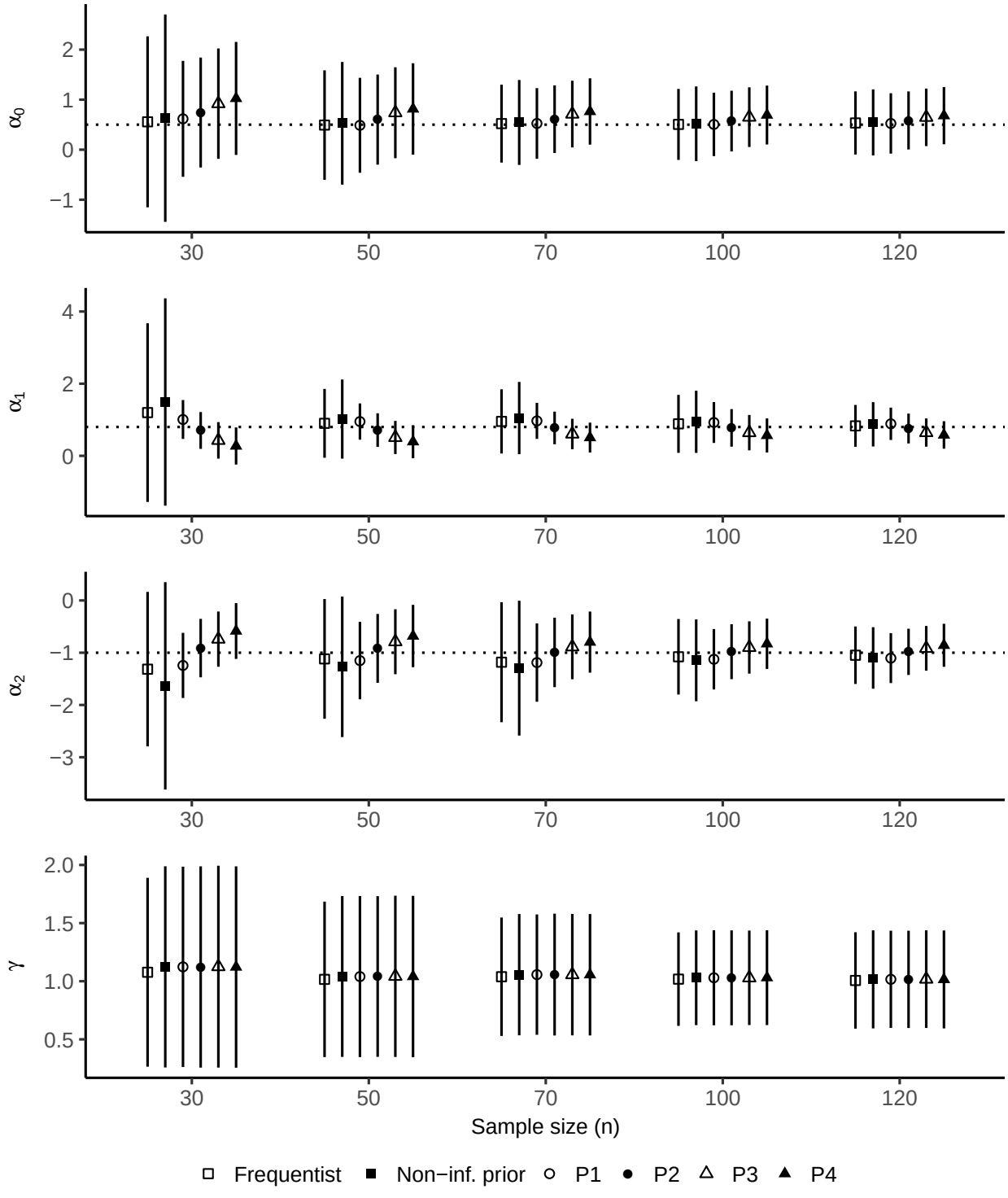


Figure 4.3: Posterior mean of the parameters in the treatment models for the simulated data, under the frequentist and Bayesian approaches, considering different prior distributions and sample sizes n. The vertical bars represent the 95% posterior credible (or frequentist confidence) intervals.

Using non-informative priors (and frequentist inverse weighting) provide slightly better covariate balance, particularly in the smallest samples. As the sample size increases, the effect of the prior is diminished and the balance improves across all priors. Figure 4.4 also shows the mean and 95% confidence/posterior credible intervals for Δ_{ATE} under the frequentist and Bayesian approaches, considering different prior specifications for the parameters in the treatment model and sample sizes $n = 30, 50, 70, 100, 120$. The use of non-informative priors yield less biased estimators, and even in the larger samples – where balance appeared comparable across the different priors – bias is slightly bigger for informative priors.

4.6.2 Additional results for U.S. National Ambulatory Medical Care Survey data analysis

This appendix includes additional results from the analysis of the U.S. National Ambulatory Medical Care Survey data. Figure 4.5 shows the estimates for the parameters in the treatment model under the frequentist and Bayesian approaches, considering samples of sizes $n = 30, 50, 70, 100$ and 120. The Bayesian estimates are based on the posterior mean of the coefficients in the treatment model. As expected, estimates based on the frequentist and Bayesian approaches with non-informative priors tend to be very similar. As n increases, they are also near to those based on informative priors P1 and P2. However, the estimates for α_2 and α_3 based on priors P3 and P4 are still far from the other estimates, even as n increases.

Panels in Figure 4.6 report the standardized mean difference for each of the covariates based on the original sample, and in the weighted sample after balancing using $\widehat{\omega}_{ATE}$ as weights. The differences computed for age exhibit balance (below 0.25), regardless of prior and sample size. However, the results for gender and smoking suggest the weights computed from informative prior P4 do not provide covariate balance. In general, standardized mean differences based on Bayesian approach with non-informative prior and informative priors P1 and P2 are bellow 0.25, providing good balance, as well as the frequentist approach.

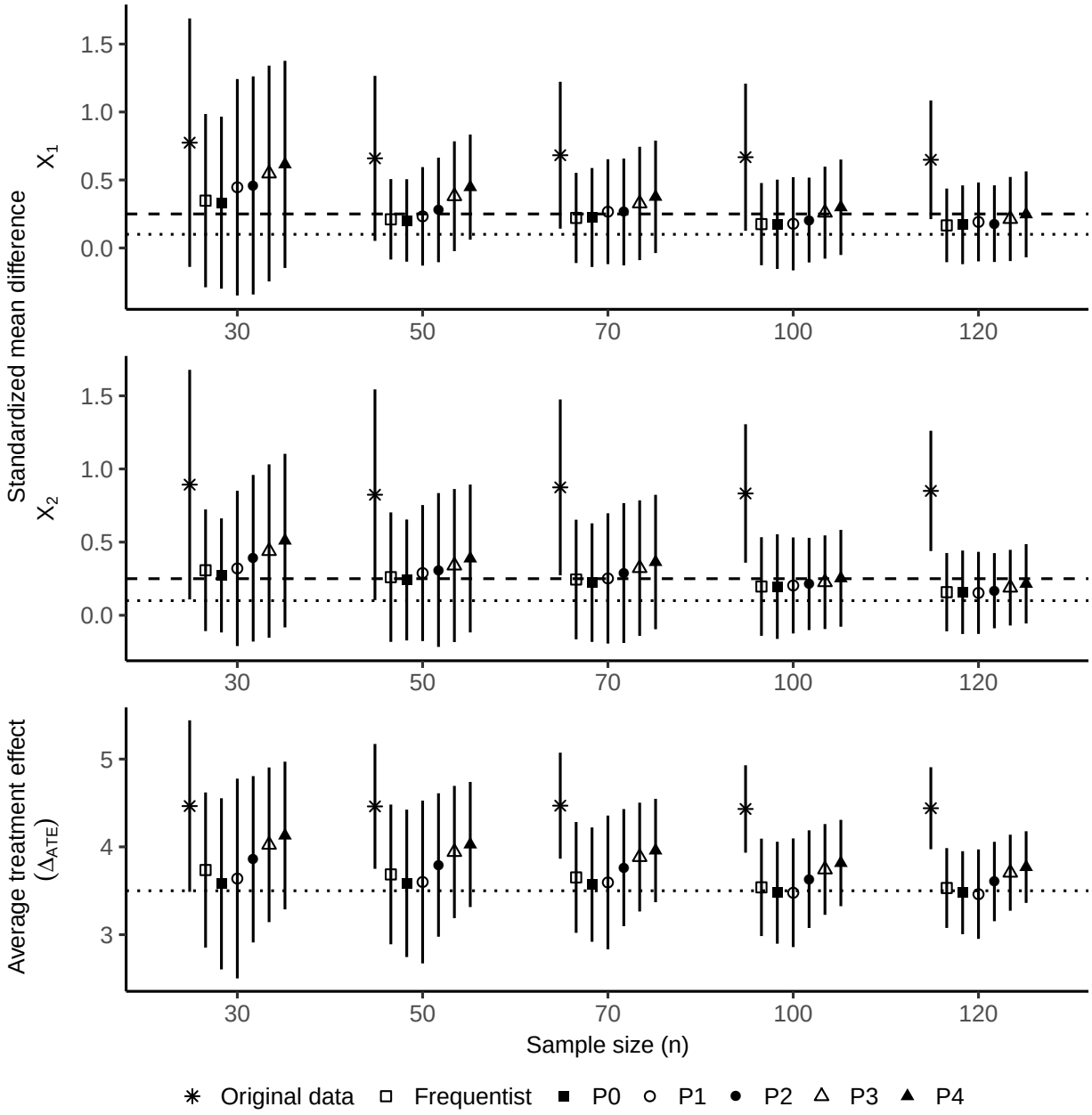


Figure 4.4: Standardized mean differences for X_1 and X_2, and estimates for average treatment effect in simulated data, based on the original sample and following the use of $\widehat{\omega}^{ATE}$ as inverse probability of treatment weights. The vertical bars represent the 95% posterior credible (or frequentist confidence) intervals. The dotted and dashed lines in first panels represent the cut-offs points of 0.1 and 0.25, respectively. In the last panels, the dotted line represents the true value of Δ_{ATE}.

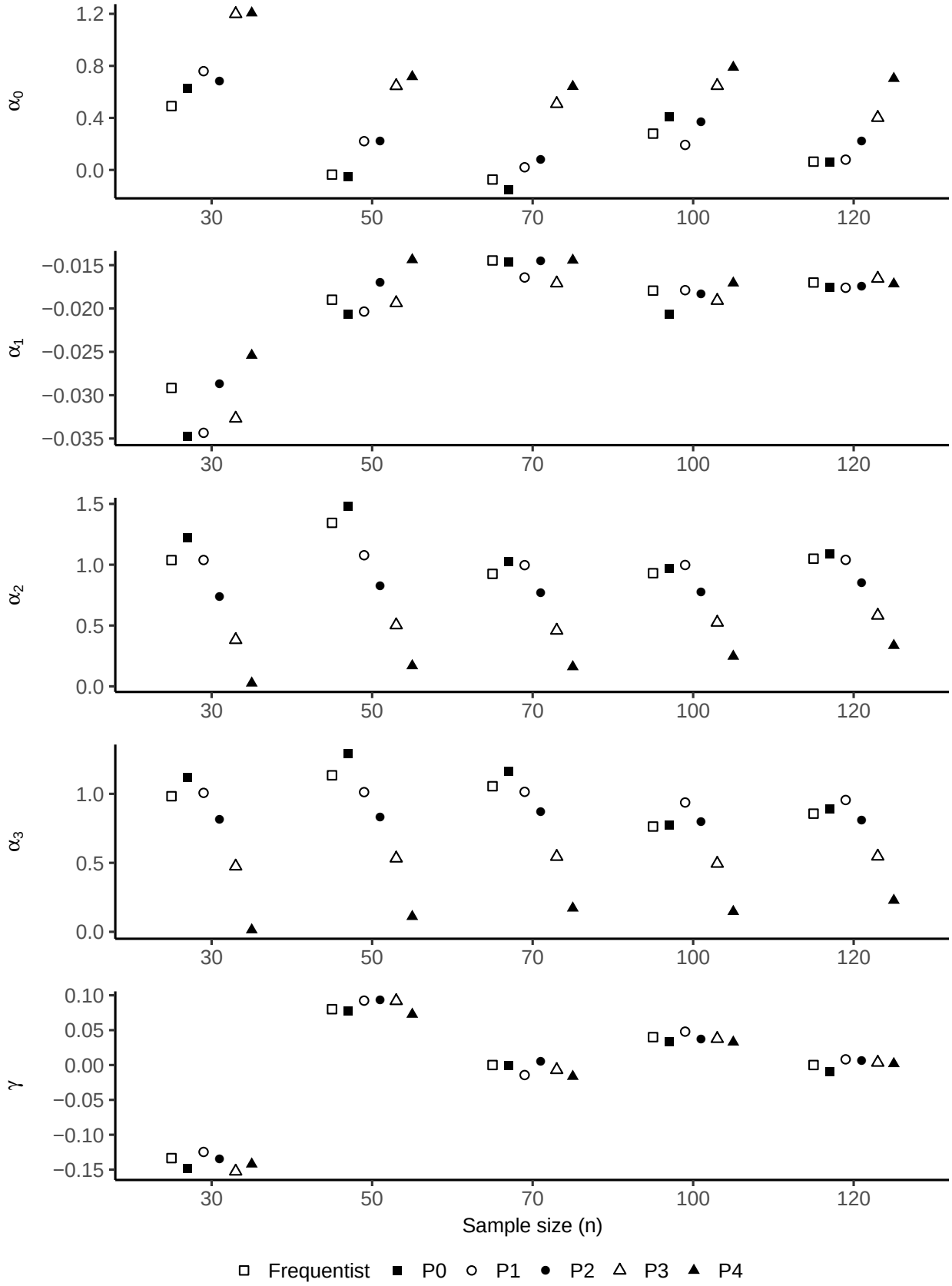


Figure 4.5: Posterior mean of the parameters in the treatment model in the U.S. National Ambulatory Medical Care Survey data, under the frequentist and Bayesian approaches, considering different prior distributions and sample sizes n.

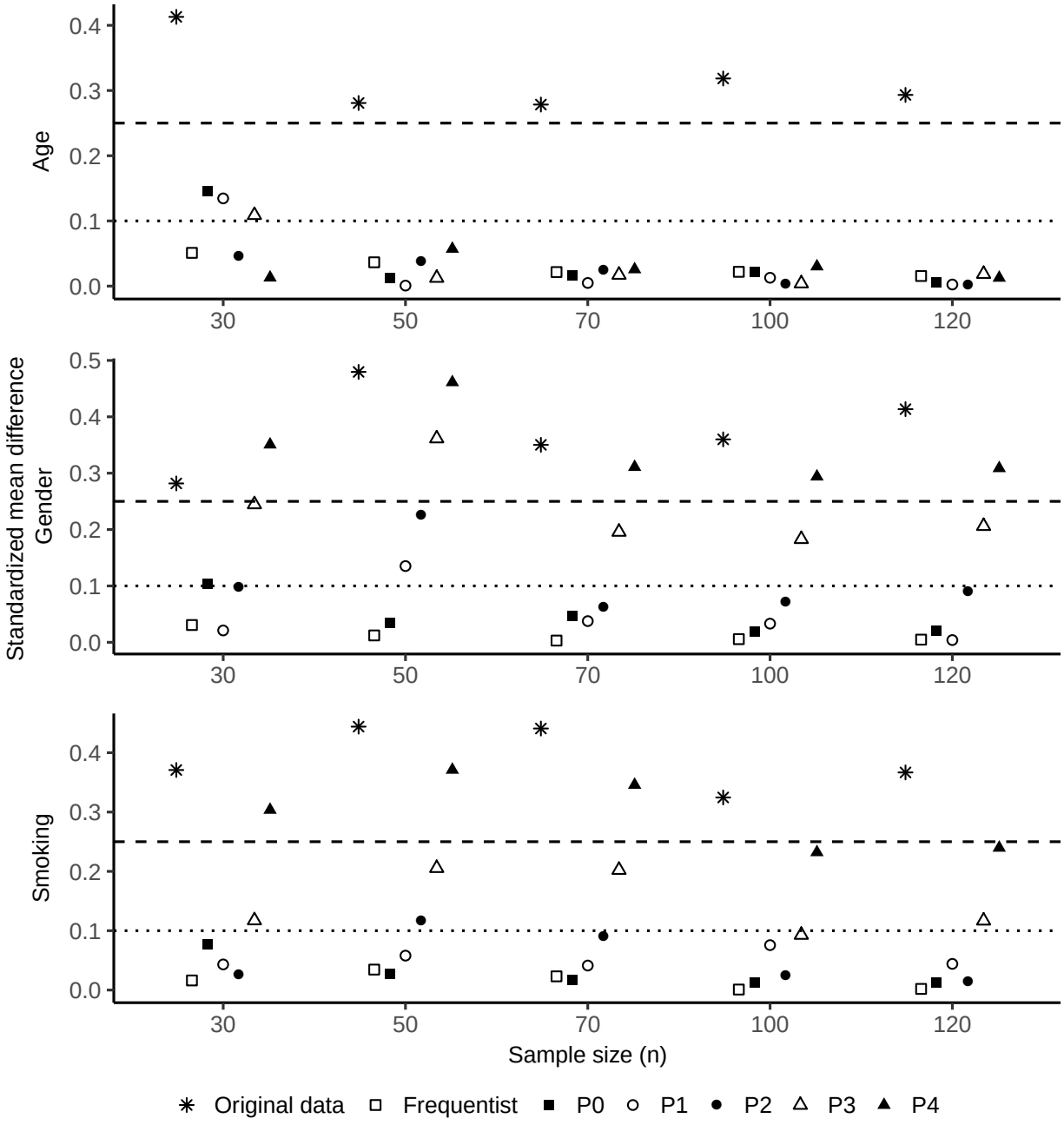


Figure 4.6: Standardized mean differences for the covariates of the U.S. National Ambulatory Medical Care Survey data, based on the original sample and following the use of $\widehat{\omega}^{ATE}$ as inverse probability weights. The dotted and dashed lines represent the cut-offs points of 0.1 and 0.25, respectively.

Chapter 5

Conclusion

This final chapter provides a brief discussion of the research, giving an overall perspective of the three projects developed in this thesis including the main findings, limitations of the proposed approaches, and suggestions for future research.

5.1 Discussion

This work addressed an area of causal inference in which there is still much space for methodological development aimed at solving open problems in analysis of observational data. Causal inference from a Bayesian perspective is a challenging issue. The usual Bayesian approach that combines information from data through a joint likelihood function of the parameters of interest with prior distributions leads to bias due to feedback since, in the causal context, joint estimation uses information from the outcome model to estimate the parameters in the propensity score. Some approaches have been developed to solve this problem, including Bayesian propensity scores methods, Bayesian g-formula and other alternatives approaches using non-parametric models. This thesis focused on the first of these – specifically, weighting using a function of the propensity score.

The current literature on Bayesian causal inference is generally focused on estimating the average treatment effect (ATE). Even within a frequentist framework, research into estimation of the average treatment effect on the treated (ATT) is far less common than those focused on estimating the ATE. Within the Bayesian framework, there is almost no such research on estimating the ATT. Therefore, the present work represents an important contribution to the literature on Bayesian causal inference.

Following ideas from Saarela et al. (2015a,b), a Bayesian approach for estimating ATT was developed in Chapter 2 using propensity score-based weights to adjust for confounding, while avoiding the problem of feedback by considering a weighted likelihood bootstrap strategy to approximate posterior distributions of interest. Simulation studies evaluated the impact of sample size and the strength of confounding on causal effects estimation, and demonstrated that considering sampling weights drawn from a Multinomial distribution within Bayesian bootstrap provided slightly better results than adopting weights from a Dirichlet distribution. In the next chapter, the proposed approach was extended to

accommodate a longitudinal outcome setting, by considering mixed linear models for outcomes and thus allowing the analysis of a relevant real data from the Canadian Co-infection Cohort Study. To the best of our knowledge, this paper is the first to propose a time-varying ATT.

Unfortunately, our Bayesian approach has some limitations. The use of a Bayesian bootstrap implicitly induces non-informative prior distributions, and the proposed approach does not allow direct specification of an informative prior for parameters of the treatment and outcome models. Besides that, the parameters that govern the propensity score model are assumed to be *a priori* independent of the parameters of the outcome model; otherwise, feedback will arise. In addition, the treatment assignment model is assumed to be correctly specified; however, as cited on Chapter 1, the weighting approach can be sensitive to misspecification of the treatment assignment model.

As has been seen, covariate balance is essential to obtain unbiased estimates of causal effects. Chapter 4 investigated the ability of the propensity score to provide covariate balance in a two-step Bayesian approach that allows informative prior distributions for the coefficients of the treatment models. The aim of this work was to investigate how different prior specifications of the parameters in the treatment models might affect covariate balance. The simulation study suggested that Bayesian inference should be performed carefully when assigning informative prior distributions to coefficients in the treatment models, especially for small sample sizes. Non-informative prior distributions provided improved covariate balance in smaller samples, while in larger samples there was little difference across the choice of prior. This encourages Bayesian inference using non-informative prior distributions for propensity score parameters. Note that this two-step approach has one important drawback: one is unable to assess the causal effects on a real data set, since one cannot integrate out the covariates when computing the causal effects.

The next section describes future works that can build on the developments in this thesis, pointing to further gaps in the literature of Bayesian causal inference.

5.2 Future Work

In Chapter 3, we analyzed data from Canadian Co-infection Cohort Study and found that successful hepatitis C treatment can improve serum lipid profiles. Our analysis considered an outcome model separately for each of four lipid measures profile – total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol and triglycerides. However, there is a correlation between these serum lipids and indeed one is a function of the other three. Therefore, it would be interesting to consider a *multivariate* outcome model that allows a correlation structure between the serum lipids.

Motivated by this application, as future work, we will consider a multivariate analysis for a joint

outcome model taking into account a correlation structure that may exist between the different outcomes, while allowing the average treatment effects to be estimated over time in a longitudinal outcome setting.

The methodology developed in this thesis could be also extended to a multilevel models, where units in the same cluster might influence the treatment assignment and/or the outcome of each other. Such interference may imply breaches of strong ignorability assumption. Ignoring the multilevel structure in both the propensity score and outcome models will induce bias in estimating the causal effects. Arpino and Mealli (2011) and Li et al. (2013) show that considering clustering in the propensity score analysis can account for violations that occur at the cluster level. To the best of our knowledge, there is not a Bayesian approach for the average causal effect in the context of multilevel data.

Future research should also investigate how incorporating prior information into the Bayesian bootstrap. Although Saarela et al. (2015a) has stated that informative prior could be incorporated into Bayesian bootstrap using the sampling-importance resampling (Rubin, 1988), this has yet to be implemented. This argument was also used by Newton and Raftery (1994) as a means of aligning weighted likelihood bootstrap samples to a Bayesian posterior.

On the other hand, we considered a two-step approach to assess covariate balance and concluded one should use non-informative prior distributions for the coefficients in the treatment models in order to ensure better covariate balance. However, further research into how to incorporate prior information for the parameters of the outcome model are still lacking.

Furthermore, our work motivates the development of open source software (such as R packages) to facilitate the use of these approaches by researchers. Some improvements may be made in our code to become the estimation process more computationally efficient.

The average treatment effect on treated is an important causal effect to be evaluated in situations where the exposure of interest cannot be imposed or applied to the entire population. However, studies involving the estimation of the ATT seem to be somewhat neglected in the literature, especially from a Bayesian perspective. Despite the difficulties involved in estimating a causal effect, clearly the Bayesian approach offers advantages over the frequentist one. Bayesian methods are useful because allow us to compute quantities that would not be possible in the frequentist framework. We can easily obtain posterior summaries of any function of the parameters involved in the model, for example. Thus, we believe the development of this thesis should motivate further research on Bayesian methods for estimating the ATT.

Bibliography

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

An, W. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40(1):151–189.

Arpino, B. and Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, 55(4):1770–1780.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27(12):2037–2049.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161.

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30):5642–5655.

Austin, P. C. and Stuart, E. A. (2015a). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679.

Austin, P. C. and Stuart, E. A. (2015b). Optimal full matching for survival outcomes: a method that merits more widespread use. *Statistics in Medicine*, 34(30):3949–3967.

Austin, P. C. and Stuart, E. A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when esti-

mating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research*, 26(4):1654–1670.

Batsaikhan, B., Huang, C.-I., Yeh, M.-L., Huang, C.-F., Hou, N.-J., Lin, Z.-Y., Chen, S.-C., Huang, J.-F., Yu, M.-L., Chuang, W.-L., et al. (2018). The effect of antiviral therapy on serum lipid profiles in chronic hepatitis C. *Oncotarget*, 9(30):21313–21321.

Berenguer, J., Álvarez-Pellicer, J., Martín, P. M., López-Aldeguer, J., Von-Wichmann, M. A., Quereda, C., Mallolas, J., Sanz, J., Tural, C., Bellón, J. M., et al. (2009). Sustained virological response to interferon plus ribavirin reduces liver-related complications and mortality in patients coinfected with human immunodeficiency virus and hepatitis C virus. *Hepatology*, 50(2):407–413.

Capistrano, E. S. M., Moodie, E. E. M., and Schmidt, A. M. (2019). Bayesian estimation of the average treatment effect on the treated using inverse weighting. *Statistics in Medicine*, 38(13):2447–2466.

Centers for Disease Control and Prevention (1998). Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. 47(19):1–40.

Chen, J. and Kaplan, D. (2015). Covariate balance in Bayesian propensity score approaches for observational studies. *Journal of Research on Educational Effectiveness*, 8(2):280–302.

Cohen, J. (1988). *Statistical Power Analysis*. Hillsdale, NJ: Erlbaum.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897.

Coverdale, S. A., Khan, M. H., Byth, K., Lin, R., Weltman, M., George, J., Samarasinghe, D., Liddle, C., Kench, J. G., Crewe, E., et al. (2004). Effects of interferon treatment response on liver complications of chronic hepatitis C: 9-year follow-up study. *The American Journal of Gastroenterology*, 99(4):636–644.

De Finetti, B. (1974). *Theory of probability: a critical introductory treatment. Transl. by Antonio Machi and Adrian Smith*. J. Wiley.

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Deb, S., Austin, P. C., Tu, J. V., Ko, D. T., Mazer, C. D., Kiss, A., and Fremes, S. E. (2016). A review of propensity-score methods and their use in cardiovascular research. *Canadian Journal of Cardiology*, 32(2):259–265.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM Review*, 21(4):460–480.

El Saadany, S., Coyle, D., Giulivi, A., and Afzal, M. (2005). Economic burden of hepatitis C in Canada and the potential impact of prevention. *The European Journal of Health Economics*, 6(2):159–165.

Endo, D., Satoh, K., Shimada, N., Hokari, A., and Aizawa, Y. (2017). Impact of interferon-free antivirus therapy on lipid profiles in patients with chronic hepatitis C genotype 1b. *World Journal of Gastroenterology*, 23(13):2355–2364.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534.

Gutman, R. and Rubin, D. B. (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Statistics in Medicine*, 32(11):1795–1814.

Gutman, R. and Rubin, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in Medicine*, 34(26):3381–3398.

Hagan, H., Jordan, A. E., Neurer, J., and Cleland, C. M. (2015). Incidence of sexually-transmitted hepatitis c virus infection in hiv-positive men who have sex with men: A systematic review and meta-analysis. *AIDS (London, England)*, 29(17):2335–2345.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Hernán, M. A. and Robins, J. M. (2019). *Causal inference*. Boca Raton: Chapman & Hall/CRC. (forthcoming).

Hill, J. and Su, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, 7(3):1386–1420.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3):259–278.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

Jang, E. S., Won, J.-E., Jung, J. I., Lee, S.-H., Kim, J. W., and Jeong, S.-H. (2011). The effect of antiviral therapy on serum cholesterol levels in chronic hepatitis C. *Gut and Liver*, 5(3):356–362.

Kaplan, D. and Chen, J. (2012). A two-step Bayesian approach for propensity score analysis: simulations and case study. *Psychometrika*, 77(3):581–609.

Kaplan, D. and Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, 49(6):505–517.

Keil, A. P., Daza, E. J., Engel, S. M., Buckley, J. P., and Edwards, J. K. (2018). A Bayesian approach to the g-formula. *Statistical Methods in Medical Research*, 27(10):3183–3204.

Klein, M. B., Saeed, S., Yang, H., Cohen, J., Conway, B., Cooper, C., Côté, P., Cox, J., Gill, J., Haase, D., et al. (2009). Cohort profile: the Canadian HIV-hepatitis C co-infection cohort study. *International Journal of Epidemiology*, 39(5):1162–1169.

Kuo, Y.-H., Chuang, T.-W., Hung, C.-H., Chen, C.-H., Wang, J.-H., Hu, T.-H., Lu, S.-N., and Lee, C.-M. (2011). Reversal of hypolipidemia in chronic hepatitis C patients after successful antiviral therapy. *Journal of the Formosan Medical Association*, 110(6):363–371.

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19):3373–3387.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Mauss, S., Berger, F., Wehmeyer, M. H., Ingiliz, P., Hueppe, D., Lutz, T., Simon, K. G., Schewe, K., Rockstroh, J. K., Baumgarten, A., et al. (2017). Effect of antiviral therapy for HCV on lipid levels. *Antiviral Therapy*, 21(1):81–88.

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics*, 6(2):1–24.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28(1):94–112.

Meissner, E. G., Lee, Y.-J., Osinusi, A., Sims, Z., Qin, J., Sturdevant, D., McHutchison, J., Subramanian, M., Sampson, M., Naggie, S., et al. (2015). Effect of sofosbuvir and ribavirin treatment on peripheral and hepatic lipid metabolism in chronic hepatitis C virus, genotype 1–infected patients. *Hepatology*, 61(3):790–801.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48.

Page, K., Morris, M. D., Hahn, J. A., Maher, L., and Prins, M. (2013). Injection drug use and hepatitis c virus infection in young adult injectors: using evidence to inform comprehensive prevention. *Clinical Infectious Diseases*, 57(Suppl. 2):S32–S38.

Plummer, M. (2017). Jags version 4.3. 0 user manual [computer software manual]. *See https://sourceforge. net/projects/mcmc-jags/files/Manuals/4. x*.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rothe, C. and Firpo, S. (2013). Semiparametric estimation and inference using doubly robust moment conditions. Technical report, IZA Discussion Paper No. 7564.

Roy, J., Lum, K. J., and Daniels, M. J. (2017). A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics*, 18(1):32–47.

Roy, J., Lum, K. J., Zeldow, B., Dworkin, J. D., Re III, V. L., and Daniels, M. J. (2018). Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics*, 74(4):1193–1202.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics*, 3:395–402.

Rubin, D. B. (2008). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In *Epidemiology and Medical Statistics*, volume 27 of *Handbook of Statistics*, pages 28–63. Elsevier. Edited by C. R. Rao, J .P. Miller and D. C. Rao.

Saarela, O., Belzile, L. R., and Stephens, D. A. (2016). A Bayesian view of doubly robust causal inference. *Biometrika*, 103(3):667–681.

Saarela, O., Stephens, D. A., Moodie, E. E., and Klein, M. B. (2015a). On Bayesian estimation of marginal structural models (with discussion). *Biometrics*, 71(2):279–288.

Saarela, O., Stephens, D. A., Moodie, E. E., and Klein, M. B. (2015b). Rejoinder: On Bayesian estimation of marginal structural models. *Biometrics*, 71(2):299–301.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the Matching package for R. *Journal of Statistical Software*, 42(7):1–52.

Shinozaki, T. and Matsuyama, Y. (2015). Brief report: Doubly robust estimation of standardized risk difference and ratio in the exposed population. *Epidemiology*, 26(6):873–877.

Spradling, P., Richardson, J., Buchacz, K., Moorman, A., Brooks, J., and Investigators, H. O. S. H. (2010). Prevalence of chronic hepatitis B virus infection among patients in the hiv outpatient study, 1996–2007. *Journal of Viral Hepatitis*, 17(12):879–886.

Townsend, K., Meissner, E. G., Sidharthan, S., Sampson, M., Remaley, A. T., Tang, L., Kohli, A., Osinusi, A., Masur, H., and Kottilil, S. (2016). Interferon-free treatment of hepatitis C virus in HIV/hepatitis C virus-coinfected subjects results in increased serum low-density lipoprotein concentration. *AIDS Research and Human Retroviruses*, 32(5):456–462.

Veldt, B. J., Heathcote, E. J., Wedemeyer, H., Reichen, J., Hofmann, W. P., Zeuzem, S., Manns, M. P., Hansen, B. E., Schalm, S. W., and Janssen, H. L. (2007). Sustained virologic response and clinical

outcomes in patients with chronic hepatitis C and advanced fibrosis. *Annals of Internal Medicine*, 147(10):677–684.

Wakefield, J. (2013). *Bayesian and frequentist regression methods*. Springer Science & Business Media.

Walker, S. G. (2010). Bayesian nonparametric methods: motivation and ideas. In *Bayesian nonparametrics*, volume 28, pages 22–34. Cambridge University Press. Edited by N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker.

Wong, J. B., McQuillan, G. M., McHutchison, J. G., and Poynard, T. (2000). Estimating future hepatitis C morbidity, mortality, and costs in the United States. *American Journal of Public Health*, 90(10):1562–1569.

Xu, D., Daniels, M. J., and Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, 17(3):589–602.

Xu, D., Daniels, M. J., and Winterstein, A. G. (2018). A Bayesian nonparametric approach to causal inference on quantiles. *Biometrics*, 74(3):986–996.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., and Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, 13(2):273–277.

Zigler, C. M. (2016). The central role of Bayes' theorem for joint estimation of causal effects and propensity scores. *The American Statistician*, 70(1):47–54.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics*, 69(1):263–273.