Two-phase Targeted Maximum Likelihood Estimation for Mixed Data Meta-Analysis

TWO-PHASE TARGETED MAXIMUM LIKELIHOOD ESTIMATION FOR MIXED DATA META-ANALYSIS

BY

ARMAN ALAM SIDDIQUE, B.Sc.

A THESIS
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Master of Science (2018)
McMaster University

(Mathematics & Statistics)
Hamilton, Ontario, Canada

TITLE: Two-phase Targeted Maximum Likelihood Estimation for Mixed Data Meta-Analysis

AUTHOR: Arman Alam Siddique
B.Sc., (Mathematics and Scientific Computing)
Indian Institute of Technology Kanpur, Uttar Pradesh, India

SUPERVISORS: Dr. Narayanaswamy Balakrishnan
Dr. Mireille Schnitzer

NUMBER OF PAGES: ix, 82

To my parents, Mohammad Khalil and Nasima Khatoon.

Abstract

This thesis focuses on extension of Inverse Probability of Censoring Weighted Targeted Maximum Likelihood Estimation (IPCW-TMLE) which was initially proposed for two-stage sampled data. We adapt this framework to the setting of mixed Aggregate Data (AD) and Individual Patient Data (IPD) meta-analysis. Our methods are motivated by a systematic review investigating treatment effectiveness for Multi-Drug Resistant Tuberculosis (MDR-TB) where studies consist of mixed IPD and AD, and where treatments are not necessarily observed across all studies. We focus on the estimation of the expected potential outcome under a given treatment and then compare the results using different methods in two simulation studies. We also discuss the challenges and demonstrate our estimation approach when there exist studies that do not have access to the treatment of interest, using the concept of transportability. We use the jackknife estimator to estimate the variance and evaluate the coverage probability of different methods in the simulation study. The results showed near unbiased results along with a close to nomial coverage probability when using IPCW-TMLE method.

Acknowledgements

I would like to take this opportunity to thank both my amazing supervisors Dr. Narayanaswamy Balakrishnan and Dr. Mireille Schnitzer, for providing me with both financial and emotional support during my Master's program at McMaster University. Dr. Narayanaswamy Balakrishnan was always available whenever I needed help either with my thesis or coursework.

I would like to show my sincere gratitude to Dr. Mireille Schnitzer for her valuable time and knowledge. Her constant encouragement to attend and present talks in various conferences has really helped me broaden my knowledge in the field of causal inference.

I would also like to thank Dr. Lehana Thabane for accepting to be a part of my thesis defense committee.

I would like to thank my family for their endless encouraging support throughout my thesis and for inspiring me to pursue higher studies. I would also like to thank my friends Anupreet Porwal, Forrest Paton, Michael Thabane and Feras Samain for their support during this period.

Finally, I would like to show my gratitude to the donors of Dr. Sri Gopal Mohanty Graduate Scholarship for their financial support.

Contents

List of Figures

Chapter 1

Introduction

Meta-Analysis is a formal way of quantitatively combining available information from studies to assess treatment effects on a more general scale. This available information maybe present in the form of Aggregate Data (AD), which refers to the summary data reported in the studies, or Individual Patient Data (IPD), which refers to all subject-specific data in each study. A meta-analysis performed on AD-only studies may result in ecological bias, and one limited to only the studies where IPD are available may lead to selection bias. Hence, a desirable approach to meta-analysis uses the information available from both AD and IPD studies, which we refer to here as a mixed data meta-analysis.

Our investigation is motivated by an investigation of drug effectiveness in patients with multi-drug resistant tuberculosis (MDR-TB) [Ahuja et al., 2012]. MDR-TB is a disease caused by the Mycobacterium tuberculosis that is resistant to at least isoniazid and ripamfin, two of the most common drugs prescribed to patients with tuberculosis. MDR-TB patients are treated using multiple alternative microbial agents with the current guideline recommending the intake of five or more microbial agents concurrently [World Health Organization, 2016]. A recent systematic review of the literature performed in 2012 identified articles which summarized the findings obtained in the analysis of 67 unique observational studies. The IPD were obtained after contacting the authors of the identified studies. Not all authors responded, however, and so the IPD from multiple studies are not available for the meta-analysis. The available studies consist of 32 IPD and 35 AD-only studies where the IPD studies were performed in 23 World Health Organization health regions [Ahuja et al., 2012]. Across all studies, patients were observed to be taking any of 15 medications, and with patients taking multiple medications concurrently, the analysis of this kind of data poses a challenge in estimating the causal effects of the observed treatments.

We tackle this problem of mixed AD and IPD two-stage data structure here which has hardly been discussed in the literature [Idris and Misran, 2015]. We justify our procedure using a counterfactual approach to causal inference, which is commonly employed to estimate such causal quantities as the average treatment effect. Available research has not demonstrated the identifiability of these causal parameters in mixed AD and IPD meta-analysis. The treatment effects of interest are defined using counterfactuals, and to obtain consistent estimation, we formulate a novel application of the inverse probability of censoring weighted targeted maximum likelihood estimator (IPCW- TMLE). IPCW-TMLE also incorporates a partial double robustness property which we demonstrate analytically as well as through Monte Carlo simulation.

One specific complication that arises for IPD data analysis of multiple treatments deals with cases when treatments are unevenly distributed across studies. This complication has already been addressed [Wang, 2018], and the concept of transportability of causal effects was developed for these situations under stringent assumptions. We incorporate this concept in our IPCW-TMLE method.

In this thesis, we focus on the use of IPCW-TMLE for estimating treatment effects for mixed data meta-analysis. In Chapter 2, we review the basics required for performing causal inference in observational studies and provide an overview of TMLE and the previous methods used for estimating treatment effects in meta-analysis. We formulate the methodology for the use of IPCW-TMLE for our thesis in Chapter 3, by first explicitly describing the data structure and the identification of the target parameter and then providing the estimation algorithm and proof of double robustness. In Chapter 4 we perform a couple of simulation studies, motivated by the MDR-TB data, to evaluate the performance of the IPCW-TMLE against some simpler approaches. We conclude the thesis in Chapter 5 and suggest some further research that can be carried out in this area.

Chapter 2

Literature Review

2.1 Causal Inference

This section lays down the necessary foundation for performing causal inference on a dataset. Under specific assumptions, one can demonstrate the identifiability of a causal parameter of interest using observed data. We first elaborate two types of studies used in the clinical literature followed by defining the standard causal assumptions required for simple observational studies. We then define the propensity score and describe its use in medical studies followed by Targeted Maximum Likelihood Estimator (TMLE) and the Inverse Probability of Censored Weighting-Targeted Maximum Likelihood Estimator (IPCW-TMLE) for estimating causal quantities with different data structures.

2.1.1 Observational Studies and Randomized Controlled Trials

Observational studies and randomized controlled trials (RCTs) are two different types of study designs used in clinical research. RCTs are regarded as "gold standard" for estimating treatment effects in clinical studies [Byar et al., 1976]. RCTs are studies performed on a strictly defined population under ideal treatment conditions which include randomization of treatments given to patients in the study. For instance, suppose a doctor wishes to test the effectiveness of a new drug A to an existing drug B for the same illness. The doctor then randomly assigns each patient to one of the drugs irrespective of the characteristics of that patient. This is an example of a RCT.

An observational study is a type of study design where inference is based on the data in which the researcher has no control over the treatments taken by individuals in the study. For instance, suppose a farmer wants to buy fertilizer to improve his crop yield. To choose the best fertilizer for the crop, he compares the performance of different fertilizers with the crop based on the data from his neighbors who used different fertilizers over the past years.

Due to the absence of randomization of treatments in observational studies, the covariates of individuals taking each treatment may largely differ and possibly also affect the observed outcome. These covariates, referred to as confounders, may result in confounding bias in observational studies in contrast to studies conducted in a RCT that are not subject to such a bias [Hannan, 2008].

However, RCTs are time-consuming and also require additional costs compared to observational studies. Another major drawback of RCTs is that one needs to take into account the ethical issues involved in carrying out the study. For instance, the individuals taking part in the study must have the right to retract themselves at any phase in the study, individuals must be aware of the associated risks of taking part in the study, the possible risk that might occur during the study should be balanced by the benefits of the study, *etc.* The institutional review board should approve all of the ethical issues arising in RCTs before the study is performed [Sarker, 2014]. Due to these drawbacks, one might consider the use of an observational study for assessing treatment effects.

2.1.2 Counterfactual Model

Suppose Mark just had an interview with Google for a data analyst position but was unprepared for it. As a consequence, he was not offered a job after the interview. He might think, "If I had prepared for the interview, I probably would have stood a chance for the job position." This is a speculative thought, and there is no way to observe what exactly would have happened had he prepared well for the interview. These kinds of statements are called *counterfactual statements* and are formally defined as statements with a false premise followed by the outcome that would have occurred, had the premise been true [Brady, 2002].

Consider the data of the form $O = \{X, A, Y\}$, where X represents a vector of covariates, A represents a binary categorical treatment wherein $A = 1$ denotes that patient i was exposed to the treatment and $A = 0$ denotes that the patient was not exposed to the treatment. As stated by Greenland and Brumback [2002], the counterfactual model for the above data assumes that:

- Individuals in the study could hypothetically have been exposed to any treatment levels ($A = 0$ or 1) at the beginning of the study;

- Outcomes for every individual exist under both treatment levels and are denoted by $Y(A = 1)$ and $Y(A = 0)$ which represent the observed outcome if the individual received the treatment $A = 1$ and $A = 0$, respectively.

At the end of the study, the outcome is only observed for a single treatment value corresponding to the treatment received; in otherwords, for an individual who was exposed to treatment, we observe the outcome $Y(A = 1)$, corresponding to the treatment exposure $A = 1$, whereas $Y(A = 0)$ is unobserved for that individual. The unobserved or hypothetical outcome is called the *counterfactual* or *potential* outcome.

In general, the choice of treatment does not affect an individual if the counterfactual outcomes of every available treatment in the study are the same as the observed outcome. Practical use of the counterfactual model in causal inference is to assess the effectiveness of a treatment by estimating the difference of the average of the potential outcomes across the study population based on the observed covariates and the outcomes of individuals in the study [Greenland and Brumback, 2002].

Counterfactuals are also used to estimate the average treatment effect (ATE), which is defined as the difference of the expected mean when one intervenes on the study by exposing everyone to the treatment against no one receiving the treatment. For instance, for a binary treatment A, the ATE is defined as:

$$ATE = \mathbb{E}(Y(A = 1)) - \mathbb{E}(Y(A = 0)).$$

To estimate the ATE, we state the causal assumptions required in this simple observational study structure followed by defining the propensity score model and the TMLE algorithm for the estimation of causal quantities.

2.1.3 Assumptions for Observational Studies for Inference

In order to demonstrate identifiability (defined in Section 2.1.4) for the parameter of interest in observational studies, we require assumptions which relates counterfactual outcomes to the observed data. The assumptions are described as follows:

Stable Unit Treatment Value Assumption: Stable Unit Treatment Value Assumption (SUTVA) was introduced by Rubin [1980], but was also previously discussed by Cox [1958] informally. Within the field of epidemiology, SUTVA is referred to as the *treatment variation irrelevance* and the *consistency* assumption [Schwartz et al., 2012].

The *treatment variation irrelevance* assumption states that the response of any individual is only dependent on the treatment provided to the individual and is independent of any other individual's treatment.

The *consistency* assumption states that the true observed outcome of an individual is same as the counterfactual outcome under actual treatment (A_i) provided to the individual. Consider any patient i, who is exposed to treatment A_i after which we observe the outcome Y_i. Hence, the consistency statement can be mathematically written in the following form:

$$Y_i = Y_i(A = A_i), \tag{A1}$$

where $Y_i(A = A_i)$ denotes the potential outcome under treatment A_i.

Positivity: This assumption states that for all individuals in the study, the probability of receiving any treatment based on the covariates should be positive, *i.e.*, for the data structure presented in Section 2.1.2, $Pr(A = 0|X) > 0$ and

$Pr(A = 1|X) > 0$. The assumption of positivity is violated when it is theoretically impossible to assign the treatment of interest to some patients due to their set of pre-treatment covariates, for instance, some individuals in a study may be contraindicated for one of the treatments. Practical positivity violations occur when some patients in the study have an arbitrarily small estimated probability of receiving the treatment of interest. This commonly occurs in studies with small sample sizes. The presence of these cases in a model leads to an increase in bias and variance in the estimation of the causal parameter of interest [Petersen et al., 2012].

Conditional Exchangeability: Conditional Exchangeability can be mathematically represented as

$$Y(A = a) \perp\!\!\!\perp A|X. \tag{A3}$$

This assumption states that conditional on the pre-treatment covariates, the counterfactual outcome of an individual is independent of the treatment received, *i.e.*, $\mathbb{E}(Y(A = a)|A = 1, X) = \mathbb{E}(Y(A = a)|A = 0, X) = \mathbb{E}(Y(A = a)|X)$. In epidemiology, we say that X contains all of the confounders in the model affecting A and Y and there are no unmeasured confounders. However, conditional exchangeability cannot be empirically tested in an observational study [Hernán, 2012].

2.1.4 Identifiability of $\mathbb{E}(Y(A = a))$

A parameter is said to be identifiable if it can be determined using infinite samples of the available data. For the data structure O defined in Section 2.1.2, the proof of identifiability of the target parameter $\mathbb{E}[Y(A=a)]$, using the assumptions mentioned in Section 2.1.3, is as follows:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|A=a,X]] &= \mathbb{E}[\mathbb{E}[Y(A=a)|A=a,X]] && \text{using A1,}\\
&= \mathbb{E}[\mathbb{E}[Y(A=a)|W]] && \text{using A3,}\\
&= \mathbb{E}[Y(A=a)] && \text{using law of iterated expectation (LIE).}
\end{aligned}$$

Therefore, $\mathbb{E}[Y(A=a)]$ can be estimated using the available data. The following sections describe the estimation of the target parameter using a semi-parametric estimation structure.

2.1.5 Propensity Score

The propensity score is defined as an individual's probability of receiving a treatment based on their vector of covariates [Rosenbaum and Rubin, 1983]. When considering a binary treatment A, the propensity score can be mathematically defined as

$$g(A=a) = Pr(A=a|X), \tag{2.1}$$

where $a \in \{0,1\}$. In RCTs, this probability is fixed and known. For instance, treatments being assigned to patients by flipping a coin, where the occurrence of heads and tails indicates that the patient is exposed to the treatment or not, respectively, corresponds to $Pr(A=a|X) = 0.5$. On the other hand, in observational studies, this probability is unknown and likely to depend on the pre-treatment covariates of the individual. Due to this, the treatment arms in observational studies are improperly balanced regarding the patient characteristics in each arm [Braitman and Rosenbaum, 2002], where treatment arms refer to the subset of patients exposed to a particular treatment.

The propensity score may, for instance, be estimated using logistic regression, classification, and regression trees analysis, or random forests [Lee et al., 2010]. The propensity score has been used to estimate treatment effects using stratification [Rosenbaum and Rubin, 1984], matching [Rosenbaum and Rubin, 1985] or regression adjustment [d'Agostino, 1998].

2.1.6 Efficient Influence Function

Semi-parametric models are defined as models that contain a parametric part with a finite dimensional space and a non-parametric part with a finite or infinite dimensional space, which in some cases is referred to as the nuisance parameter or the model noise [Powell, 1994]. The influence function is an essential characteristic of semi-parametric estimators, which are used for analysis of a dataset.

Consider a statistical model with O_1, O_2,..., O_n, independent and identical random vectors. Let ψ denote the parameter of interest. Suppose there exists a consistent estimator $\hat{\psi}$ of ψ and a random vector $\phi(O)$, such that $\mathbb{E}(\phi(O)) = 0$ and,

$$\sqrt{n}(\hat{\psi}_n - \psi_0) = \frac{1}{\sqrt{n}} \sum_i \phi(O_i) + o_p(1), \tag{2.2}$$

where $o_p(1)$ denotes a random variable that converges in probability to zero as the sample size converges to infinity, and ψ_0 denotes the true value of the parameter.

For the above random vector $\phi(O)$, provided $\mathbb{E}(\phi(O)\phi(O)^T)$ is finite and singular, ψ is described as an asymptotically linear estimator and $\phi(O)$ is called the influence function of $\hat{\psi}$ [Tsiatis, 2007]. Further, using Central Limit Theorem, we have that

$$\frac{1}{\sqrt{n}}\sum_i \phi(O_i) \xrightarrow{D} \mathbb{N}(0, \mathbb{E}(\phi(O)\phi(O)^T)),$$

which, from the use of Slutsky's Theorem, yields

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{D} \mathbb{N}(0, \mathbb{E}(\phi(O)\phi(O)^T)),$$

since $\frac{1}{\sqrt{n}}\sum_i \phi(O_i) \xrightarrow{P} \sqrt{n}(\hat{\psi}_n - \psi_0)$ using (2.2). Every asymptotically linear estimator is characterized by a unique influence function though any number of estimators may share the same influence function. Further, $\hat{\psi}$ is said to be asymptotically robust if $(\phi(O_i) - \phi(o))/n$ is bounded, where o represents any value of O and $\phi(O)$ is an influence function of $\hat{\psi}$ [Van der Vaart, 2014].

For every regular and asymptotically linear estimator $\hat{\psi}$ of ψ, there exists an influence function $\phi_1(O)$ such that $\mathbb{E}(\phi(O)\phi(O)^T)$ - $\mathbb{E}(\phi_1(O)\phi_1(O)^T)$ is non-negative definite for every other influence function $\phi(O)$ of $\hat{\psi}$. This $\phi_1(O)$ is called the efficient influence function of $\hat{\psi}$ [Tsiatis, 2007].

Efficient influence functions also provide a tool for variance estimation, and a typical variance estimate of the parameter ψ is given by $\frac{1}{n}\sum_i(\phi_1(O_i))^2$ [Rocke and Downs, 1981].

2.1.7 Targeted Maximum Likelihood Estimation

Targeted Maximum Likelihood Estimation (TMLE) is a semi-parametric estimation framework, proposed by Van der Laan and Rubin [2006], for the estimation of causal quantities.

We have the data denoted by $O = \{X, A, Y\}$, where X, A and Y represent covariates, treatment and outcome respectively, with a true probability distribution $\mathbb{P}_0$, a statistical model space M which consists of all possible distributions for $\mathbb{P}_0$, and the parameter of interest denoted by $\psi = \Psi(\mathbb{P})$, where Ψ is a mapping defined from the probability distribution $\mathbb{P}$ to the target parameter. $\psi_0 = \Psi(P_0)$ represents the true value of the parameter of interest. Below a typical algorithm is provided for the estimation of a causal parameter using TMLE.

Algorithm

Suppose the parameter of interest is $\psi = \mathbb{E}[\mathbb{E}[Y_i|A_i = a, X_i]]$, which also represents the causal effect under treatment $A = a$ denoted by $\mathbb{E}[Y(A_i = a)]$ as proved in Section 2.1.3. We also define $Q_0 = \{\bar{Q}_0, Q_{X,0}\} \in \mathbb{P}_0$, where $Q_{X,0}$ denotes the full distribution of X and $\bar{Q}_0$ denotes the expectation of the outcome conditional on the covariates. The efficient influence curve $\phi_1(O_i)$ at $\mathbb{P}_0$ is given by

$$\phi_1(O_i) = \frac{I(A_i = a)}{g_0(X_i)}\{Y_i - \bar{Q}_0(X_i)\} + \bar{Q}_0(X_i) - \Psi(\mathbb{P}_0),$$

where $g_0(X_i)$ is the true propensity score [Rose and Van der Laan, 2011].

TMLE in this setting consists of two steps, essentially referred to as the initial step and the update step. The initial step consists of initializing estimates of $\bar{Q}_0(X_i)$

and $g_0(X_i)$ which are defined as follows:

$$\bar{Q}_0(X_i) = \mathbb{E}[Y_i|A_i = a, X_i],$$
$$g_0(X_i) = Pr(A_i = a|X_i),$$

over $\mathbb{P}_0$. For obtaining an initial estimate $\bar{Q}_n^0(X_i)$ of $\bar{Q}_0(X_i)$, we can, for example, fit a logistic regression model for Y against X for all individuals with treatment $A = a$. This model is then used to predict $\mathbb{E}(Y_i|A_i = a, X_i)$ for all individuals in the dataset using the pre-treatment covariates of the individuals. The estimate of the propensity score is denoted by $g_n(X_i)$ which can be obtained similarily by fitting a logistic regression model for A against X and using this model to predict the probability of assignment of the treatment $A = a$ based on the pre-treatment covariates of an individual.

We define a new variable H_n^*, which is called the clever covariate [Van der Laan and Rubin, 2006], as

$$H_n^* = H(X_i, a) = \frac{I(A_i = a)}{g_n(X_i)}.$$

We then model Y against the clever covariate H_n^* using a logistic regression with an offset of $\text{logit}(\bar{Q}_n^0(X_i))$ and with no intercept term in the resulting model. The resulting coefficient for H_n^* in the obtained model is denoted by ϵ_n which represents the maximum likelihood estimate for the logistic regression model. The next step involves updating $\bar{Q}_n^0(X_i)$ to a new estimate $\bar{Q}_n^1(X_i)$, which is given by

$$\text{logit}(\bar{Q}_n^1(X_i)) = \text{logit}(\bar{Q}_n^0(X_i)) + \frac{\epsilon_n}{g_n(X_i)}. \tag{2.3}$$

As mentioned earlier, $Q_{X,0}$ is the probability distribution of X_i, and its estimate is obtained using an empirical distribution. The obtained estimate is independent of the fluctuation parameter ϵ_n and therefore is not updated [Rose and Van der Laan, 2011]. The TMLE estimate for Q_0 is then given by $Q_n^* = \{\bar{Q}_n^1, Q_{X,0}\}$. The target parameter ψ depends on $\mathbb{P}_0$ through $\bar{Q}_0$ and its estimate ψ_n^{TMLE} is given by

$$\psi_n^{TMLE} = \frac{1}{n}\sum_i \bar{Q}_n^1(X_i), \tag{2.4}$$

where n denotes the sample size. As mentioned in Van der Laan and Rose [2011], one can use the efficient influence curve for the estimation of the variance of the parameter of interest, provided that one of $\bar{Q}_n^0(X_i)$ or $g_n(X_i)$ is consistently estimated. The estimate $\hat{\phi}_1(O_i)$ of the efficient influence function $\phi_1(O_i)$ is given by

$$\hat{\phi}_1(O_i) = \frac{I(A_i = a)}{g_n(X_i)}\{Y_i - \bar{Q}_n^1(X_i)\} + \bar{Q}_n^1(X_i) - \psi_n^{TMLE}, \tag{2.5}$$

and the variance estimate for ψ_n^{TMLE} using the efficient influence curve is given by

$$Var\{\psi_n^{TMLE}\} = \frac{1}{n}\sum_i (\hat{\phi}_1(O_i))^2.$$

Apart from using the efficient influence curve, one can also use bootstrp method to estimate the variance of the parameter of the parameter of interest [Gruber and van der Laan, 2013, Schnitzer et al., 2015].

Properties

TMLE targets the maximum likelihood estimate of the target parameter by performing an additional bias step by updating $\bar{Q}_n^0(X_i)$. It is a well-defined estimator, *i.e.*, there exists a unique solution for this algorithm.

TMLE is asymptotically unbiased, *i.e.*, the obtained estimate converges to the true value as the sample size increases to infinity, provided one of $Pr(A_i = a|X_i)$ or $\mathbb{E}(Y_i|A_i = a, X_i)$ is consistently estimated. This is also referred to as the *double-robustness* property. When both models for $Pr(A_i = a|X_i)$ and $\mathbb{E}(Y_i|A_i = a, X_i)$ are correctly specified, the resulting algorithm is asymptotically efficient [Van der Laan and Gruber, 2016]. The proof of the double-robustness property is given in Section 2.1.7.

TMLE is a substitution or plug-in estimator, *i.e.*, it is obtained by plugging the estimate $\bar{Q}_n^1$ of the data generating function $\bar{Q}_0$ into the mapping function Ψ:

$$\psi_n^{TMLE} = \Psi(\bar{Q}_n^1) = \mathbb{E}(\bar{Q}_n^1(A_i = 1, X_i)) = \frac{1}{n}\sum_i \bar{Q}_n^1(A_i = 1, X_i).$$

Due to the above reason, TMLE is more robust to outliers and sparsity [Van der Laan and Rose, 2011].

Double Robustness of TMLE

TMLE solves the efficient influence function equation given by

$$\sum_i \frac{\mathbb{I}(A_i = a)}{g_n(X_i)}\{Y_i - \bar{Q}_n^1(X_i)\} + \bar{Q}_n^1(X_i) - \psi_n^{TMLE} = 0. \tag{2.6}$$

Proof : As shown is Section 2.1.7, the update step of the TMLE corresponds to the logistic regression of Y against the clever covariate H_n^* using an offset of $\text{logit}(\bar{Q}_n^0(X_i))$ and with no intercept term. The score function of the logistic regression model can thus be written as

$$\begin{aligned}
&\sum_i \frac{\mathbb{I}(A_i = a)}{g_n(X_i)}(Y_i - \text{expit}(\text{logit}(\bar{Q}_n^0(X_i)) + \frac{\epsilon_n}{g_n(X_i)})) = 0 \\
\Rightarrow &\sum_i \frac{\mathbb{I}(A_i = a)}{g_n(X_i)}(Y_i - \bar{Q}_n^1(X_i)) = 0 && \text{by Equation (2.3)} \\
\Rightarrow &\sum_i \frac{\mathbb{I}(A_i = a)}{g_n(X_i)}\{Y_i - \bar{Q}_n^1(X_i)\} + \bar{Q}_n^1(X_i) - \psi_n^{TMLE} \quad = 0 && \text{by Equation (2.4).}
\end{aligned}$$

Hence, Equation (2.6) holds. Further, provided one of $Q_n^1(X_i) \to \mathbb{E}(Y_i|A_i = 1, X_i)$ or $g_n(X_i) \to Pr(A_i = 1|X_i)$, we can claim that

$$\sum_i \frac{\mathbb{I}(A_i = a)}{g_n(X_i)}\{Y_i - \bar{Q}_n^1(X_i)\} + \bar{Q}_n^1(X_i) - \psi_0 \xrightarrow{P} 0, \text{ as } n \to \infty. \tag{2.7}$$

To prove the above claim, we consider two cases with an assumption that one of $Q_n^1(X_i)$ or $g_n(X_i)$ is correctly specified and show that Equation (2.7) is satisfied.

i) Let's assume $Q_n^1(X_i) \to \mathbb{E}(Y_i|A_i = 1, X_i)$ and $g_n(X_i) \to \widetilde{g}(X_i)$, where $\widetilde{g}(X_i)$ is not necessarily equal to $Pr(A_i = a|X_i)$. Therefore, by the Weak Law of Large Numbers

(WLLN), the left hand side of Equation (2.7) can be written as

$$
\begin{aligned}
&\sum_i \frac{\mathbb{I}(A_i = a)}{g_n(X_i)}\{Y_i - \bar{Q}_n^1(X_i)\} + \bar{Q}_n^1(X_i) - \psi_0 \\
&\xrightarrow{P} \mathbb{E}\left(\frac{\mathbb{I}(A_i = a)}{\widetilde{g}(X_i)}\{Y_i - \mathbb{E}(Y_i|A_i = a, X_i)\} + \mathbb{E}(Y_i|A_i = a, X_i) - \psi_0\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\left\{\frac{\mathbb{I}(A_i = a)}{\widetilde{g}(X_i)}\{Y_i - \mathbb{E}(Y_i|A_i = a, X_i)\} + \mathbb{E}(Y_i|A_i = a, X_i) - \psi_0)\right\}|A_i = a, X_i\right)\right) \\
&\hspace{25em}\text{by LIE,} \\
&= \mathbb{E}\left(\frac{\mathbb{I}(A_i = a)}{\widetilde{g}(X_i)}\{\mathbb{E}(\{Y_i - \mathbb{E}(Y_i|A_i = a, X_i)\}|A_i = a, X_i)\} + \mathbb{E}(Y_i|A_i = a, X_i) - \psi_0\right) \\
&= \mathbb{E}\left(\frac{\mathbb{I}(A_i = a)}{\widetilde{g}(X_i)}\{\mathbb{E}(Y_i|A_i = a, X_i) - \mathbb{E}(Y_i|A_i = a, X_i)\} + \mathbb{E}(Y_i|A_i = a, X_i) - \psi_0\right) \\
&= \mathbb{E}\left(\frac{\mathbb{I}(A_i = a)}{\widetilde{g}(X_i)}\{0\} + \mathbb{E}(Y_i|A_i = a, X_i) - \psi_0\right) \\
&= \mathbb{E}(\mathbb{E}(Y_i|A_i = a, X_i)) - \psi_0 \\
&= 0.
\end{aligned}
\tag{2.8}
$$

ii) Now, let's assume that $Q_n^1(X_i) \to \widetilde{Q}_n(X_i)$ and $g_n(X_i) \to Pr(A_i = 1|X_i)$, where $\widetilde{Q}_n(X_i)$ is not necessarily equal to $\mathbb{E}(Y_i|A_i = 1, X_i)$. Therefore, by the WLLN, the left hand side of Equation (2.7) can be written as

$$
\begin{aligned}
&\sum_i \frac{\mathbb{I}(A_i = a)}{g_n(X_i)}\{Y_i - \bar{Q}_n^1(X_i)\} + \bar{Q}_n^1(X_i) - \psi_0 \\
&\xrightarrow{P} \mathbb{E}\left(\frac{\mathbb{I}(A_i = a)}{Pr(A_i = 1|X_i)}\{Y_i - \widetilde{Q}_n(X_i)\} + \widetilde{Q}_n(X_i) - \psi_0\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\left\{\frac{\mathbb{I}(A_i = a)}{Pr(A_i = 1|X)}\{Y_i - \widetilde{Q}_n(X_i)\} + \widetilde{Q}_n(X_i) - \psi_0\right\}|Y(A_i = a), X_i\right)\right) \\
&\hspace{20em} \text{by LIE,} \\
&= \mathbb{E}\left(\mathbb{E}\left(\left\{\frac{\mathbb{I}(A_i = a)}{Pr(A_i = 1|X_i)}\{Y(A_i = a) - \widetilde{Q}_n(X_i)\} + \widetilde{Q}_n(X_i) - \psi_0\right\}|Y(A_i = a), X_i\right)\right) \\
&\hspace{20em} \text{by (A1),} \\
&= \mathbb{E}\left(\frac{Pr(A_i = 1|X_i)}{Pr(A_i = 1|X_i, Y(A_i = a))}\{Y(A_i = a) - \widetilde{Q}_n(X_i)\} + \widetilde{Q}_n(X_i) - \psi_0)\right) \\
&= \mathbb{E}\left(\frac{Pr(A_i = 1|X_i)}{Pr(A_i = 1|X_i)}\{Y(A_i = a) - \widetilde{Q}_n(X_i)\} + \widetilde{Q}_n(X_i) - \psi_0)\right) \\
&= \mathbb{E}(Y(A_i = a) - \widetilde{Q}_n(X_i) + \widetilde{Q}_n(X_i) - \psi_0))) \\
&= \mathbb{E}(Y(A_i = a)) - \psi_0 \\
&= 0
\end{aligned}
$$

Therefore, we conclude that TMLE is doubly-robust. Double-robustness is a desirable property for an estimator because of its consistent estimation of the target quantity when any one of the two components is correctly specified. The double-robustness property also makes TMLE stable [Kang et al., 2007, Porter et al., 2011].

2.1.8 Transportability

The generalization of a conclusion obtained from a particular study and using those conclusions to make inference on a more general population is known as *external validity*. One particular form of generalization involving the transfer of causal effects from an investigational study to a different target population is called transportability [Pearl and Bareinboim, 2014]. For example, consider a small group where patients are treated with a particular treatment for a disease. Even though the treatment might be effective in this study group, the same can't be said for a different target population. This generalization requires similarities between the experimental group and the target population.

Pearl and Bareinboim [2011] laid the ground rules and circumstances under which the transportability of these causal associations is possible with the application of *selection diagrams* which are used to compare the similarities and dissimilarities between the source study and the target study. Further, Bareinboim and Pearl [698-704, 2012] gave a formal graphical algorithm to decide whether the similarities between the study and the target population allow for the transfer of causal effects between studies.

Hernán and VanderWeele [2011] justified transportability of the causal effect estimates while dealing with different versions of the same category of treatment in studies when the corresponding studies have similar characteristics regarding effect modification, interference in the studies and versions of the compound treatments.

Data fusion, formally defined as the combining of experimental results from multiple studies, led to a generalization of the transportability theory to multiple heterogeneous studies, is one of the essential concepts used in the thesis. Data fusion synthesizes the results obtained from multiple studies to infer the causal effects on the target study [Bareinboim and Pearl, 2016]. Previous works have used transportability to estimate treatment importance in MDR-TB using targeted learning [Wang, 2018].

In the next section, we provide an overview of meta-analysis and the properties of different types of data on which meta-analysis is performed.

2.2 Systematic Review and Meta-Analysis

Systematic Review is the review of information based on devising a detailed selection strategy plan for the exploration of relevant articles on a particular topic of interest. It aims to perform an unbiased search for all the available articles in the literature. Three critical standards, which are looked upon while performing a systematic review are framing a review question, identification of the relevant articles and inclusion criteria for selection of the obtained articles [Khan et al., 2003].

Meta-Analysis, a term coined by Eugene Glass in 1976 [DerSimonian and Laird, 2015], is defined as a formal way of quantitatively combining data from previous studies and providing a general estimate of a quantity [Haidich, 2010]. Meta-Analysis may or may not be based on the data collection using a systematic review. However, results obtained by Meta-Analysis on articles acquired using systematic reviews are often considered to be more reliable [Ryś et al., 2009].

Meta-Analysis strengthens the evidence of the treatment efficiency by examining multiple articles with the same objective [DerSimonian and Laird, 1986]. Meta-Analysis provides inference about the effectiveness of a treatment on a more general level as opposed to individual studies [Mansfield et al., 2016].

However, meta-analysis has some drawbacks. Studies containing results with more statistical importance or positive results are often more likely to be published than studies containing less significant treatment effects. Performing a meta-analysis which excludes such studies often leads to misguided interpretations about the evidence of the treatment effect, and this is referred to as *publication bias* [Begg and Berlin, 1989]. Previous literature suggests the use of graphical models [Duval and Tweedie, 2000] or modeling methods [Hedges, 1992] to assess publication bias.

Another drawback of meta-analysis is the presence of *heterogenity* in the collected data information. Heterogeneity occurs due to two sources in the collected studies, known as *within-study variability* and *between-study variability. Within-study variability* accounts for the sampling variability, differences in patient characteristics within the study, different version of treatment, *etc. Between-study variability* (true heterogeneity) occurs in meta-analysis due to factors like the population differences between the studies, region-specific treatment variations, methodological study quality, *etc.*, which accounts for the difference in treatment effects between the studies [Huedo-Medina et al., 2006, Montori et al., 2003]. Meta-Analysis of studies should account for the availability for both sources of heterogeneity and more details on modeling methods that incorporate this information are provided in Section 2.2.1.

Meta-Analysis can be based on two types of data, called *aggregate data* or *individual patient data*, which are briefly explained and compared in the following sections.

2.2.1 Aggregate Data Meta-Analysis

Aggregate Data Meta-Analysis (AD-MA) is a type of Meta-Analysis performed on articles containing summary data information of the study, such as sex ratio, percentage of patients cured by the treatment and mean age. These articles also contain the summary results for the study, such as the average treatment effect or odds ratio. One does not need the consent of the publisher to perform Meta-Analysis on the articles, which is the main reason behind the extensive usage of AD-MA.

One of the most popular methods used for Meta-Analysis in clinical trials, proposed by DerSimonian and Laird [1986], summarizes the available evidence of the effectiveness of a treatment using a random effects model with normally distributed study specific intercepts. This random effects model allows for heterogeneity of the treatment effect across various studies and also overcomes the difficulties arising from the usage of different weights for each study in a fixed effects model.

However, AD-MA has many limitations. Firstly, AD-MA is susceptible to publication bias. Secondly, due to the differences in the characteristics of the individuals in each study and the usage of different designed statistical models for each study, AD-MA can lead to discrepancies in the resulting estimate of the treatment effect. Specifically, in observational studies, the presence of different confounding factors across studies leads to an additional bias [Blettner et al., 1999]. In cases where studies use different measurement scales to represent the outcome, one needs to normalize these outcomes to perform AD-MA [Stewart and Tierney, 2002].

Due to the above drawbacks of AD-MA, one cannot always rely on the statistical conclusions, and hence it is suggested to obtain individual patient data for studies whenever feasible, since it may reduce the bias and also provide reliable and consistent estimates.

2.2.2 Individual Patient Data Meta-Analysis

Individual Patient Data Meta-Analysis (IPD-MA) is considered to be one of the best approaches to analyze systematic reviews and is often regarded as the "gold standard" to perform meta-analysis [Stewart and Tierney, 2002]. Individual patient data (IPD) contains information about all of the individuals participating in any study. Usually, IPD-MA is performed using one of the two different approaches: One-Stage Approach and Two-Stage Approach [Burke et al., 2017].

In an One-Stage Approach IPD-MA, IPDs from all the individual studies are pooled together, and the overall effect for the pooled data is estimated. In a Two-Stage Approach IPD-MA, each study is analyzed separately to obtain the study-specific estimates which are then combined using an appropriate AD-MA model to obtain the pooled estimate [Burke et al., 2017].

IPD-MA, however, has many advantages over AD-MA. The collection of individual patient data includes direct contact with the authors of the article who are experts in the research topic and can also provide information about unpublished work in the research area. IPD-MA considers information about the patient-level factors as well as the study-level factors, possibly decreasing the extent of heterogeneity in the meta-analysis [Broeze et al., 2010]. Also, IPD-MA allows one to estimate subgroup effects by performing the analysis on those individuals who satisfy the criterion of interest, for instance, estimating the treatment effect for the subset of patients with age > 30 [Stewart and Tierney, 2002]. IPD can also be used in some cases even-though, the article using the IPD does not contain the relevant association to the Meta-Analysis. For instance, suppose we have an IPD article which contains two treatments, A_1 and A_2, and summarizes the treatment effect of A_2 for the study.

Suppose we are interested in performing Meta-Analysis on articles which provide association of the treatment on the study. One can use the IPD to first estimate the treatment effectiveness of A_1 for the study and then perform a second-stage AD-MA. We can also pool together all the IPDs and perform an IPD-MA.

However, IPD-MA also has some disadvantages. It requires large amounts of time and resources to obtain the data from the authors and then cleaning the data according to the requirement of the analysis. Sometimes, data gets lost or destroyed, or it also might be possible that the author may not be willing to share the data due to confidentiality reasons. Also, the quality of the individual patient data cannot be increased, and the studies might consist of poorly-designed trials [Riley et al., 2010].

IPD-MA is also susceptible to publication bias; however, this bias is exacerbated in particular due to selection bias when we do not have access to the IPD studies for some available AD studies.

2.2.3 Partial Individual Patient Data Meta Analysis

Because of the disadvantages of meta-analysis using IPD or AD alone, a reasonable way to perform meta-analysis is to base it on a dataset consisting of IPD and AD, the latter corresponding to datasets in which the IPD cannot be retrieved. We refer to these kinds of datasets as Partial IPD. A systematic review performed by Riley et al. [2007] identified 33 applied articles and 8 methodological articles which combined IPD and AD data to perform a meta-analysis. The analysis in the obtained articles was performed using a two-stage method, partial reconstruction of IPD, multilevel modeling, and Bayesian Hierarchical Regression.

In the two-stage method, all the IPD information is converted to AD, and the final analysis is performed on the AD for every study [Tudur et al., 2001, Simmonds, 2005, Idris and Misran, 2015]. This allows the analyst to perform the same version of analysis on each dataset. The multilevel model assumes that every IPD study can be viewed as a multilevel model where the highest level of the model represents the study and the lowest level of the model represents observations of individuals in the study. This modeling framework can be extended to include AD studies which are considered to contribute only to the highest level of the model [Goldstein et al., 2000]. Bayesian Hierarchical Regression uses a Markov Chain Monte Carlo method to estimate common regression parameters of the AD and the IPD models. The obtained regression parameters are used in performing meta-analysis [Jackson et al., 2008, Sutton et al., 2008].

In this thesis, we interpret partial IPD as a two-stage sampling structure which is defined briefly in Section 2.3. In the next section, we describe various methods that are used for estimating causal parameters in these kinds of datasets and focus mainly on a TMLE approach.

2.3 Two-Stage Data

Consider the problem wherein one aims to estimate average patient outcomes in a population consisting of cancer patients who visited a particular hospital. To estimate the average outcomes under treatment we consider interviewing the patients who were exposed to the treatments. Since bringing in patients and conducting surveys costs money and time, the interviews are only conducted on a random subset of the population and clinical data information are obtained for this subset. The data extracted from this kind of study is called two-stage data [Neyman, 1938].

Previous works have proposed various ways to perform analysis on data sampled with two-stage designs. White [1982] used weighted least-square techniques to estimate the log-odds ratio for a two-stage design data. Breslow and Cain [1988] proposed a modified logistic regression to provide inference for two-stage data by taking into account the data available in the first stage. Flanders and Greenland [1991] proposed a generalization of the weighted-likelihood estimator for the estimation of absolute rates and risk in two-stage designs. Robins et al. [1994] devised semi-parametric estimators using inverse probability weighting by considering two-stage designs as a data structure with observations missing at random. Zhao and Lipsitz [1992] provided a general overview of the estimators developed by Breslow and Cain [1988] and Flanders and Greenland [1991] for the analysis of two-stage data using some simulation studies and provided recommendations for the usage of these methods in different situations. Wang et al. [2009] proposed an enriched doubly robust estimator for the estimation of the treatment effectiveness for a binary treatment in the two-stage design of observational studies. Rose and Van der Laan [2011] developed the concept of combining TMLE with the inverse probability of censoring weighting to adjust for the missingness of the units which were not sampled at the second stage. We discuss this last estimator in detail in the next subsection.

2.3.1 Inverse Probability of Censoring Weighted Targeted Maximum Likelihood Estimator

Inverse Probability of Censoring Weighted Targeted Maximum Likelihood Estimator (IPCW-TMLE) is a methodology proposed by Rose and Van der Laan [2011] for estimating the causal target parameter in two-stage sample designs. This estimator uses TMLE to obtain the estimates for the observations collected in the second stage sampling followed by an IPCW estimate to adjust for the observations which were not selected from the first stage to the second stage.

Consider the data $O = \{Y, \Delta, \Delta W\}$, where Y denotes the outcome of the observations sampled in the first stage, Δ denotes the inclusion of the observations from the first stage to the second stage, *i.e.*, $\Delta_i = 1$ if the i^{th} observation from the first stage is resampled during the second stage sampling, else $\Delta_i = 0$, and $W = \{X, A, Y\}$ denotes the additional details for the observations sampled during the second stage. Let the probability distribution of the above model be given by $\mathbb{P}_0$ and our parameter of interest denoted by $\psi_0 = \Psi(\mathbb{P}_0) = \mathbb{E}[\mathbb{E}[Y_i|A_i = a, X_i]]$, where Ψ is a mapping defined from the probability distribution $\mathbb{P}_0$ to the target parameter. The full data efficient influence function curve for the above model is given by

$$D(Q_0, g_0) = \Big(\frac{I(A = a)}{g_0(X_i)}\Big)(Y - \bar{Q}_0(X_i))) + \bar{Q}_0(X_i) - \Psi(Q_0),$$

where $Q_0 = \{\bar{Q}_0, Q_{X,0}\} \in \mathbb{P}_0$, where $Q_{X,0}$ denotes the full distribution of X and $\bar{Q}_0(X_i) = \mathbb{E}(Y_i|X_i, A_i)$ and $g_0(X_i) = Pr(A_i = a|X_i)$ is our propensity score as defined in Section 2.1.7 for $a = 1$. We also define $\pi_n(Y_i)$, the probability of inclusion of the studies from the first stage to the second stage, as

$$\pi_n(Y_i) = Pr(\Delta_i = 1|Y_i).$$

For the analysis of this data, we first fit a TMLE for the observations with $\Delta = 1$. To do so, we first obtain an initial estimate $\bar{Q}_n^0(X_i)$ of $\bar{Q}_0(X_i)$ using a regression model for Y based on the covariates in the subset of observations with $A = a$, with weights as $1/\pi_n(Y_i)$. We also denote $I(A_i = a)/g_n(X_i)$ as the clever covariate given by $H_n(X_i, a)$, where $g_n(X_i)$ is an estimate of $g_0(X_i)$. The next step is to update $\bar{Q}_n^0(X_i)$ in order to obtain the new estimate $\bar{Q}_n^1(X_i)$ using the following regression equation:

$$\text{logit}(\bar{Q}_n^1(X_i)) = \text{logit}(\bar{Q}_n^0(X_i)) + \epsilon H_0(X_i, a),$$

where ϵ is obtained using the weighted maximum likelihood estimation as shown in Section 2.1.7, with weights of $1/\pi_n(Y_i)$. After obtaining $\bar{Q}_n^1(X_i)$, we obtain our estimate for ψ_0, $\psi_n^{IPCW-TMLE} = \Psi(Q_n)$, where $Q_n = \{\bar{Q}_n^1, Q_{X,0}\}$ is our IPCW-TMLE estimate, which is given by

$$\psi_n^{IPCW-TMLE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\Delta_i}{\pi_n(Y_i)}(\bar{Q}_n^1(X_i))\right\}.$$

As mentioned by Rose and Van der Laan [2011], IPCW-TMLE is a locally efficient way of dealing with two-stage data as it takes into account the data collected in both the first as well as the second stage sampling. IPCW-TMLE inherits the double robustness property of the complete data TMLE by solving the weighted complete-data efficient influence function provided that we consistently estimate $\pi_n(Y_i)$.

2.3.2 Variance Estimation in two-stage sampling

Sitter [1997] proposed jackknife and bootstrap estimators to estimate the variance and standard errors for the parameter of interest in two-stage sampling. Based on the available data, the jackknife estimator is shown to perform well because it is asymptotically equivalent to a *robust variance estimator* [Ziegler, 1997]. Here, we make use of the jackknife estimator which is described below.

Jackknife Estimator for Two-Stage Sampling

Consider a sample of N observations. Suppose the first stage sample s_1' consists of n_1 observations which are sampled from the initial N observations without replacement. The second stage sample s_2' consists of n_2 observations sampled without replacement from the n_1 observations sampled in the first stage. As defined earlier in Section 2.1.7, our target parameter ψ can be written as

$$\psi = \Psi(P)(O)$$

Using a slightly different notation, the above expression is written as

$$\psi = \bar{\Psi}(O) = \bar{\Psi}(O_1, O_2, ..., O_{n_1}),$$

where O_i denotes the i^{th} observation in the first stage sampling. Further, we denote ψ_{-j} as

$$\psi_{-j} = \hat{\Psi}(O_1, ..., O_{j-1}, O_{j+1}, ...O_{n_1}).$$

Then, the jackknife variance estimate for ψ is given by

$$Var(\psi) = \frac{n_1 - 1}{n_1} \sum_{j \in s_1'} \{\psi_{-j} - \sum_{k \in s_1'} \psi_{-k}\}^2.$$

We extend the concept of IPCW-TMLE which was originally used by Rose and Van der Laan [2011] for observation to a context in mixed data of AD and IPD in Chapter 2. To show the effectiveness of this approach, we use some simulation studies in Chapter 3 and compare this method to other simpler methods. We also use the jackknife estimator as an alternative to estimate the standard errors and compare them with Monte Carlo standard errors.

Chapter 3

Methods

We define the parameter of interest as the expected potential outcome, fixing a given treatment, without intervention on other available treatments. Our data are defined as comprising both Aggregate Data (AD) and Individual Patient Data (IPD). As mentioned in Chapter 2 and demonstrated by simulation studies in Chapter 4, estimation of the target parameter using only the IPD studies may lead to selection bias incurred due to the exclusion of the relevant APD studies. In order to deal with this selection bias, we use the IPCW-TMLE algorithm described in Chapter 2.

This chapter lays down the identifiability of the parameter of interest and its estimation using the IPCW-TMLE algorithm. We introduce the multi-drug resistant tuberculosis example in Section 3.1, which serves as a motivating example for performing the simulation studies in this thesis. Section 3.2 provides a general context and interpretation of the partial IPD as a two-stage design structure. Section 3.3 provides an overview of the data structure along with the definition of the parameter of interest. Section 3.4 lists the assumptions required for the identifiability of the parameter of interest given the data structure and the proof of identifiability of the parameter of interest is demonstrated in Section 3.5. Section 3.6 describes each component of the IPCW-TMLE algorithm and gives the proof of consistency of the IPCW-TMLE algorithm for the defined data structure under the stated assumptions.

3.1 Multi-Drug Resistant Tuberculosis

Multi-Drug Resistant Tuberculosis (MDR-TB) is a type of a tuberculosis (TB) which is caused by the *Mycobacterium Tuberculosis* that is resistant to isoniazid and ripafmin, the most commonly prescribed drugs for TB. The Collaborative Group for Meta-Analysis of Individual Patient Data in Multidrug-Resistant Tuberculosis (IPD-MDRTB) collected IPD from 31 observational studies comprising 9,290 individual MDR-TB patients [Ahuja et al., 2012]. Further, findings of 36 studies were also reported, though the investigators were not able to secure the IPD for these studies. In the available IPD, patients with MDR-TB were observed to be treated with combinations of 15 different antimicrobial agents.

The collected IPD contains study level information on the year when the study was conducted and the country where it was conducted and patient-specific information such as the age of the individual, sex, HIV status, binary indicators for the 15 different antimicrobial agents and a binary outcome for each individual. The binary outcome was defined as either treatment success (the treatment was completed and cured the infection) or failure (patient still tested culture positive for MDR-TB, died, or defaulted on treatment/were lost to follow-up).

3.2 Two-Stage Sampling

Meta-Analysis requires investigators to perform a systematic review and select relevant articles regarding the topic of interest. The investigator proceeds by first obtaining the published study data (AD) for the selected articles. Next, one obtains the IPD using different sources, but sometimes the IPD is not available for all of the studies. This leaves us with IPD and AD for some studies and AD only for other studies. We refer to this kind of data as mixed AD and IPD.

Our goal is to make consistent inference in a meta-analysis with mixed AD and IPD. We think of this data structure as a two-stage sampling problem. To demonstrate the conceptual sampling strategy followed, a flowchart is given below followed by the description of its usage in the thesis:

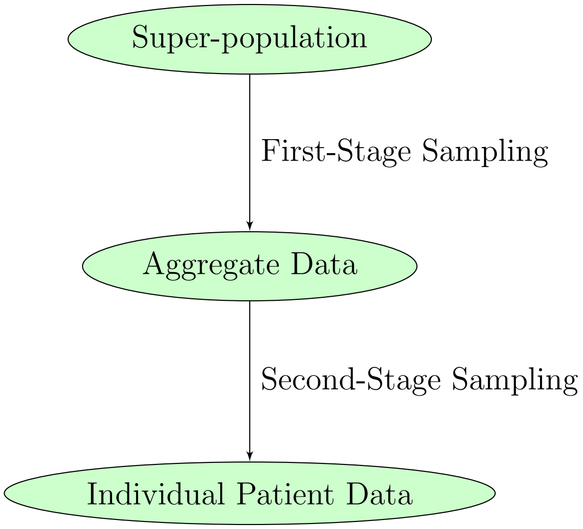


Figure 3.1: Flowchart for two-stage sampling structure.

Each study is assumed to be sampled from an infinite population of potential studies referred to as the super-population. This is the first-stage sample. The data available from this sampling contains AD information about the sampled studies in the first stage. Next, the second stage-sampling involves obtaining the IPD for the studies obtained in the first stage. The full data structure obtained using the two-stage sampling is formulated in Section 3.3. The final analysis is then performed on the obtained AD and IPD using methods described in Section 3.6.

3.3 Data Structure and Parameter of Interest

We consider a data structure similar to that of the MDR-TB data defined in Chapter 1. The data consist of 67 studies where the AD are always available and IPD are sometimes available. Each study i has study level variables jointly denoted by V_i, which represents the population information of the study data such as data such as the year of the study, information about the country in which the study took place, *etc.* The data comprise of K different binary treatments which are denoted by $A_{ij}^{(k)}$, where k = 1,2,..,K. The participants in a given study may or may not have had availability to any individual treatment. We use binary indicators , $D_i^{(k)}$ for k = 1,2,,..,K, to represent this availability. For instance, $D_i^{(k)} = 1$ indicates that individuals in the i^{th} study had access to the k^{th} treatment and $D_i^{(k)} = 0$ indicates that individuals in the i^{th} study didn't have access to the k^{th} treatment. We assume that $D_i^{(k)}$ is observable such that if treatment k had been available in study i, at least one patient would have been exposed to it. The measured characteristics of any individual j in study i are denoted by W_{ij} and their outcome is indicated by a binary variable Y_{ij}. We denote Δ_i for the binary selection indicator of the i^{th} study in the second stage sampling. We set $\Delta_i = 1$ if we have access to the IPD for the i^{th} study. The sample size for study i is denoted by n_i. Then the full observed data structure

O_i is given by

$$O_i = (D_i^{(1)}, \Delta_i, X_{ij}, \Delta\{A_{ij}^{(1)}, Y_{ij}, \quad j \in S_i\}), \text{ where}$$
$$X_{ij} = (V_i, \{D_i^{(k)}; k = 2, .., K\}, \Delta(W_{ij}, \{A_{ij}^{(k)}; k = 2, .., K\}; j \in S_i)), \quad i = 1, ..., 67,$$

where S_i denotes index set of the patients in the i^{th} study. We further denote $\bar{O}_i$ for the data corresponding to the low dimensional measure of each study or the AD availability in every study. $\bar{O}_i$ contains information measured over the course of the study. Here, we represent $\bar{O}_i$ by

$$\bar{O}_i = \{V_i, \{D_i^{(k)}, \bar{A}_i^{(k)}; k = 1, 2, .., K, j \in S_i\}, \bar{W}_i, \bar{Y}_i\}, \quad i = 1, ..., 67,$$

where $\bar{W}_i$, $\bar{A}_i^{(k)}$, and $\bar{Y}_i$ represent the average summary for W_{ij}, $A_{ij}^{(k)}$, and Y_{ij}, respectively, over all the individuals in the i^{th} study.

The aim of this thesis is to estimate the expected potential outcome, setting $A_{ij}^{(1)}$ $= 1$, with no interventions on $A_{ij}^{(k)}$ for $k \neq 1$. The parameter of interest is defined by $\psi_0 = \mathbb{E}(Y\{A_{ij}^{(1)} = 1, A_{ij}^{(2)}, A_{ij}^{(3)}\})$. One can use this to estimate the add-on effect of treatment $A^{(1)}$ for the study population by obtaining various measures such as $\mathbb{E}(Y\{A_{ij}^{(1)} = 1, A_{ij}^{(2)}, A_{ij}^{(3)}\})$ - $\mathbb{E}(Y\{A_{ij}^{(1)} = 0, A_{ij}^{(2)}, A_{ij}^{(3)}\})$ or $\mathbb{E}(Y\{A_{ij}^{(1)} = 1, A_{ij}^{(2)}, A_{ij}^{(3)}\})$ - $\mathbb{E}(Y\{A_{ij}^{(1)}, A_{ij}^{(2)}, A_{ij}^{(3)}\})$. This generalization can be made to all the K treatments in the study. However, in this thesis, we only estimate ψ_0. The proof of identifiability for ψ_0 is shown in Section 3.5 under the model assumptions stated below in Section 3.4.

3.4 Causal Assumptions

In order to identify and estimate the parameter of interest based on the available data, we make some assumptions similar to those mentioned in Chapter 1 for the ATE. The causal assumptions are as follows:

- **SUTVA:** The outcome of an individual in a study is independent of the treatment exposure of other individuals in the study. Mathematically,

$$Y_{ij}(A_{ij}^{(1)}, A_{ij}^{(2)}, A_{ij}^{(3)}) \perp\!\!\!\perp A_{ij'}^{(1)}, A_{ij'}^{(2)}, A_{ij'}^{(3)} \quad, \forall j' \in \{S_i - j\}. \tag{B1}$$

  Further, the consistency statement in SUTVA states that the potential outcome of an individual under treatment exposure $A_{ij}^{(1)} = 1$ is the same as the true outcome of the individual provided that the individual was exposed to the treatment in the study. Mathematically,

$$Y_{ij}(A_{ij}^{(1)} = 1, A_{ij}^{(2)}, A_{ij}^{(3)}) = Y_{ij}, \quad \{j \in S_i \mid A_{ij}^{1} = 1\}; \tag{B2}$$

- **Positivity:** For the parameter of interest defined here, we need to have two positivity conditions to hold, which are as follows:
  i) Every study in the model has a positive probability of having access to all the treatments in the model conditional on the study level variables V_i within studies where the IPD are available, *i.e.*,

$$Pr(D_i^{(1)} = 1 | V_i, \Delta_i = 1) > 0, \quad i = 1, ..., 67, \tag{B3}$$

  This assumption is important in order to allow for the transportability of the parameter of interest.

ii) Provided that a study i has availability of the treatment $A^{(1)}$, every patient in that study has a positive probability of being exposed to $A^{(1)}$ conditional on the pre-treatment covariates X_{ij} within studies where the IPD are available, *i.e.*

$$Pr(A_{ij}^{(1)} = 1 | D_i^{(1)} = 1, X_{ij}, \Delta_i = 1) > 0, \quad \forall j \in S_i; \tag{B4}$$

- **Conditional Exchangeability:** The potential outcome of an individual in study i is independent of the exposure to treatment $A^{(1)}$ conditional on the pre-treatment covariates of the individual within studies where the IPD are available, *i.e.*,

  $$Y_{ij}(A_{ij}^{(1)} = 1, A_{ij}^{(2)}, A_{ij}^{(3)}) \perp\!\!\!\perp A_{ij}^{(1)} | X_{ij}, \Delta_i = 1. \tag{B5}$$

  This assumption eliminates confounding bias by adjusting for all the covariates which act as a confounder. Equivalently, X_{ij} is sufficient to estimate the potential outcomes for studies where the IPD is available;

- **Transportability:** The potential outcome of an individual is independent of the treatment indicator of selection for $A^{(1)}$ conditional on the covariates and treatment exposure to the individuals within studies where the IPD are available, *i.e.*,

  $$Y_{ij}(A_{ij}^{(1)} = 1, A_{ij}^{(2)}, A_{ij}^{(3)}) \perp\!\!\!\perp D_i^{(1)} | A_{ij}^{(1)} = 1, X_{ij}, \Delta_i = 1. \tag{B6}$$

  This assumption states that the potential outcomes of patients in the IPD can be estimated using X_{ij} irrespective of the fact that the patients have availability to the medication $A^{(1)}$ or not;

- **Selection Assumption for Δ (Missing at Random):** Conditional on the low dimensional measure of every study ($\bar{O}_i$), the binary selection indicator Δ_i is independent of the counterfactual outcomes of an individual. Mathematically,

$$\Delta_i \perp\!\!\!\perp \bar{Y}_i(A^1_{ij} = 1, A^2_{ij}, A^3_{ij})|\bar{O}_i. \tag{B7}$$

  This assumption states that the low dimensional measure of the studies $\bar{O}_i$ is sufficient enough to estimate the average potential outcome in study i, which is represented by $\bar{Y}_{ij}(A^1_{ij} = 1, A^2_{ij}, A^3_{ij})$;

- The treatment availability for $A^{(1)}$ is independent of the individual level confounders $\{X_{ij} \backslash V_i\}$ conditional on the study level variables V_i. Mathematically,

$$D^{(1)}_i \perp\!\!\!\perp \{X_{ij} \backslash V_i\}|V_i. \tag{B8}$$

  This assumption states that study-level variables V_i are sufficient enough to predict the treatment availability $D^{(1)}_i$.

The above causal assumptions are incorporated to achieve the identifiability of our target parameter and to show the consistency of the IPCW-TMLE algorithm adapted here.

3.5 Identifiability of ψ_0

For simplicity reasons, we shall now denote the parameter of interest $\psi_0 = \mathbb{E}(Y(A_{ij}^{(1)} = 1, A_{ij}^{(2)}, A_{ij}^{(3)}))$ as $\psi_0 = \mathbb{E}(Y_{ij}(A = 1))$. In order to prove identifiability of ψ_0 using the observed dataset O, we decompose ψ_0 as follows:

$$\psi_0 = \mathbb{E}(\bar{Q}_{Y,\Delta}), \tag{3.1}$$

where $\bar{Q}_{Y,\Delta} = \mathbb{E}(Y_{ij}(A = 1)|X_{ij}, \Delta_i = 1)$. The proof for Equation (3.1) is as follows:

$$\begin{aligned}
\psi_0 &= \mathbb{E}(\bar{Q}_{Y,\Delta}) \\
&= \mathbb{E}(\mathbb{E}(\bar{Q}_{Y,\Delta}|\bar{O}_i, \Delta_i = 1)) && \text{using LIE} \\
&= \mathbb{E}(\mathbb{E}(\mathbb{E}(Y_{ij}(A = 1)|X_{ij}, \Delta_i = 1)|\bar{O}_i, \Delta_i = 1)) \\
&= \mathbb{E}(\mathbb{E}(Y_{ij}(A = 1)|\bar{O}_i, \Delta_i = 1)) && \text{using LIE since } \bar{O}_i \in X_{ij}, \\
&= \mathbb{E}(\mathbb{E}(\bar{Y}_i(A = 1)|\bar{O}_i, \Delta_i = 1)) \\
&= \mathbb{E}(\mathbb{E}(\bar{Y}_i(A = 1)|\bar{O}_i)) && \text{using (B7)}, \\
&= \mathbb{E}(\bar{Y}_i(A = 1)) && \text{using LIE} \\
&= \mathbb{E}(\bar{Y}_{ij}(A = 1)).
\end{aligned} \tag{3.2}$$

Following Equations (3.1) and (3.2), it is sufficient to show that ψ_0 is identifiable if $\bar{Q}_{Y,\Delta}$ is identifiable, where the proof for latter is as follows:

$$\begin{aligned}
\bar{Q}_{Y,\Delta} &= \mathbb{E}(Y_{ij}(A=1)|X_{ij}, \Delta_i = 1) \\
&= \mathbb{E}(Y_{ij}(A=1)|\Delta_i = 1, X_{ij}, A^{(1)}_{ij} = 1) && \text{using (B5)}, \\
&= \mathbb{E}(Y_{ij}(A=1)|\Delta_i = 1, X_{ij}, A^{(1)}_{ij} = 1, D^{(1)}_{ij} = 1) && \text{using (B6)}, \\
&= \mathbb{E}(Y_{ij}|\Delta_i = 1, X_{ij}, A^{(1)}_{ij} = 1, D^{(1)}_{ij} = 1) && \text{using (B2)}.
\end{aligned} \tag{3.3}$$

We further denote ψ_Δ as

$$\psi_\Delta = \mathbb{E}(\bar{Q}_{Y,\Delta}|\Delta_i = 1) = \mathbb{E}(\mathbb{E}(Y_{ij}(A=1)|X_{ij}, \Delta_i = 1)|\Delta_i = 1), \tag{3.4}$$

where ψ_Δ is interpreted as the expected outcome under setting $A^{(1)} = 1$ within the population represented by the IPD and is also identifiable as it is an estimable function of $\bar{Q}_{Y,\Delta}$.

3.6 IPCW-TMLE Algorithm for Two Stage Design

In order to estimate the target parameter ψ_0, we extend the IPCW-TMLE algorithm mentioned in Chapter 1 to our context of mixed AD and IPD meta-analysis. The original formulation was developed for a two-stage sample of individuals rather than study data. The algorithm for two-stage data designs involves two phases, where the first phase of the algorithm uses TMLE for the IPD studies and the second phase uses IPCW to account for the studies which were not selected in the second stage.

In this section, we first define the Q and g components used in the IPCW-TMLE algorithm and then provide the general framework of the IPCW-TMLE algorithm for mixed data meta-analysis.

3.6.1 Q Component

The Q component is defined as the expectation of the counterfactual outcome of an individual conditional on the covariates and exposure to the counterfactual treatment $A^{(1)}$ within studies which were selected in the second stage, *i.e.*, studies with $\Delta_i = 1$. We refer to the Q-component as $\bar{Q}(X_{ij})$ which is defined as follows:

$$\bar{Q}(X_{ij}) = \mathbb{E}(Y_{ij}(A = 1)|A_{ij}^{(1)} = 1, X_{ij}, \Delta_i = 1).$$

In order to estimate $\bar{Q}(X_{ij})$, we take the subset of the IPD studies where the treatment $A_{ij}^{(1)}$ was available, *i.e.*, studies with $D_i^{(1)} = 1$. Using this subset, we fit a logistic regression model for Y_{ij} on X_{ij}, for all those individuals who had $A_{ij}^{(1)} = 1$ with weights of $1/\pi_n(\bar{O}_i)$, where $\pi_n(\bar{O}_i)$ denotes the probability of selection in the second stage conditional on $\bar{O}$, *i.e.*, the conditional probability of obtaining the IPD in the second stage. $\pi_n(\bar{O}_i)$ is an estimate of $\pi(\bar{O}_i)$, mathematically represented as

$$\pi(\bar{O}_i) = Pr(\Delta_i = 1|\bar{O}_i),$$

where we recall that $\bar{O}_i$ contains the study specific AD. The obtained regression model is then used to predict Y_{ij} based on the covariates of an individual for all the studies in the data with $\Delta_i = 1$. We refer to this prediction as $\bar{Q}_n(X_{ij})$, the estimate of $\bar{Q}(X_{ij})$. The prediction step for the Q-component uses transportability since our aim is to predict the counterfactual outcomes for the treatment exposure $A^{(1)}$ for all the individuals in studies with $\Delta_i = 1$ irrespective of the reality that the studies had access to the treatment $A^{(1)}$ or not. The proof of identifiability for $\bar{Q}(X_{ij})$ has already been shown earlier in Equation (3.3) which allows $\bar{Q}(X_{ij})$ to be estimated based on the studies with $\Delta_i^{(1)} = 1$.

3.6.2 g-Component

The g-component is usually referred to as the propensity score and is herein defined as the probability of an individual being exposed to treatment $A^{(1)}$ conditional on the covariates of the individual and the binary selection indicator Δ_i =1. We refer the g-component as $g(X_{ij})$ and is defined as

$$g(X_{ij}) = Pr(A_{ij}^{(1)} = 1 | X_{ij}, \Delta_i = 1).$$

In order to obtain $g(X_{ij})$ from the available data, we decompose the above expression into a simpler form as follows:

$$\begin{aligned}
g(X_{ij}) &= Pr(A_{ij}^{(1)} = 1 | X_{ij}, \Delta_i = 1) \\
&= Pr(A_{ij}^{(1)} = 1, D_i^{(1)} = 1 | X_{ij}, \Delta_i = 1) \\
&= Pr(A_{ij}^{(1)} = 1 | \Delta_i = 1, D_i^{(1)} = 1, X_{ij}) \cdot Pr(D_i^{(1)} = 1 | \Delta_i = 1, X_{ij}) \\
&= Pr(A_{ij}^{(1)} = 1 | \Delta_i = 1, D_i^{(1)} = 1, X_{ij}) \cdot Pr(D_i^{(1)} = 1 | \Delta_i = 1, V_i) \quad \text{using (B8)} \\
&= g^{(1)}(X_{ij}) \cdot g^{(2)}(X_{ij}). \qquad (3.5)
\end{aligned}$$

Each of the above decomposed quantities can be easily estimated from the available data. The first component in Equation (3.5), $g^{(1)}(X_{ij})$, is estimated using a logistic regression on the IPD, while the second component $g^{(2)}(X_{ij})$ is estimated using a logistic regression on the AD selected in the second stage.

3.6.3 IPCW-TMLE for mixed AD and IPD

We formulate the IPCW-TMLE algorithm using the above defined Q and g-components to obtain an estimate of our target parameter defined in Section 3.3. We denote the initial estimate of the Q-component and g-component by $\bar{Q}^0_n(X_{ij})$ and $g_n(X_{ij})$, respectively.

After obtaining $\bar{Q}^0_n(X_{ij})$ and $g_n(X_{ij})$, we fit a logistic regression for Y on $1/g_n(X_{ij})$ with an offset of $\text{logit}(\bar{Q}^0_n(X_{ij}))$ and with no intercept term with weights of $1/\pi_n(\bar{O}_i)$. The above model is only fitted for individuals with $A^{(1)}_{ij} = 1$ in our IPD studies. The coefficient for $1/g_n(X_{ij})$ in the above model is then termed as our fluctuation parameter and is represented by ϵ. We then use its estimate, ϵ_n, to update $\bar{Q}^0_n(X_{ij})$ as follows:

$$\bar{Q}^1_n(X_{ij}) = \text{expit}\Big(\text{logit}(\bar{Q}^0_n(X_{ij})) + \frac{\epsilon_n}{g_n(X_{ij})}\Big).$$

The above update is performed for all the individuals in our IPD studies. Had our data only consisted of IPD studies, one would obtain the plug-in estimate $\psi_{TMLE,n}$ of the target parameter ψ_0 as

$$\psi_{TMLE,n} = \frac{1}{n} \sum_{\substack{i=1 \\ \Delta_i=1}}^{67} \sum_{j=1}^{n_i} \bar{Q}^1_n(X_{ij}),$$

where $n = \sum_{i=1}^{67} n_i$. We refer to the above estimation step of $\bar{Q}^1_n(X_{ij})$ as the first phase estimate and this procedure is similar to the TMLE described in Chapter 1. $\psi_{TMLE,n}$ can be also written as

$$\psi_{TMLE,n} = \mathbb{E}(\bar{Q}_n^1(X_{ij})|\Delta_i = 1).$$

$\psi_{TMLE,n}$ is a biased estimate of ψ. In order to adjust for this bias, we use IPCW for $\bar{Q}_n^1(X_{ij})$. We utilize $\pi_n(\bar{O}_i)$ to obtain the estimate of ψ_0 using the IPCW step in the second phase of our algorithm. Now, we introduce a new variable $\bar{Q}_{i,n}^*$ which is defined as follows:

$$\bar{Q}_{i,n}^* = \frac{1}{n_i}\sum_{j=1}^{n_i} \bar{Q}_n^1(X_{ij}).$$

$\bar{Q}_{i,n}^*$ represents the estimate for the parameter of interest over the IPD study i. Finally, one obtains the plug-in estimate $\psi_{IPCW-TMLE,n}$ of the target parameter ψ_0 as

$$\psi_{IPCW-TMLE,n} = \sum_{i=1}^{n} \frac{\Delta_i}{\pi(\bar{O}_i)}\bar{Q}_{i,n}^*.$$

$\psi_{IPCW-TMLE,n}$ is obtained by taking a weighted average of $\bar{Q}_{i,n}^*$ using the second-stage sampling probability. This weighting strategy reduces the selection bias obtained by only using the IPD information for estimating ψ_0 in the first stage.

3.6.4 Consistency of IPCW-TMLE

Provided that we have IPD for all studies such that $\pi(\bar{O}_i) = 1$, IPCW-TMLE solves the following equations:

$$\frac{1}{N}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\mathbb{I}(A_{ij}=1)}{g_n(X_{ij})}(Y_{ij}-\bar{Q}_n^1(X_{ij}))=0, \tag{3.6}$$

$$\frac{1}{N}\sum_{i=1}^{n}\Big(\Big(\frac{1}{n_i}\sum_{j=1}^{n_i}\bar{Q}_n^1(X_{ij})\Big)-\psi_0\Big)=\frac{1}{N}\sum_{i=1}^{n}\Big(\bar{Q}_{i,n}^{*}-\psi_0\Big)=0. \tag{3.7}$$

Equation (3.6) refers to the weighted score equation for the first phase estimation and Equation (3.7) refers to the weighted mean equation for the second phase estimation. The standard full data EIF is given by

$$\phi_1(O)=\frac{1}{N}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\Big(\frac{\mathbb{I}(A_{ij}=1)}{g_n(X_{ij})}(Y_{ij}-\bar{Q}_n^1(X_{ij}))+\bar{Q}_n^1(X_{ij})-\psi_0)\Big), \tag{3.8}$$

which is also equivalent to

$$\phi_1(O)=\frac{1}{N}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\Big(\frac{\mathbb{I}(A_{ij}=1)}{g_n(X_{ij})}(Y_{ij}-\bar{Q}_n^1(X_{ij}))+\frac{1}{n_i}\sum_{j=1}^{n_i}\bar{Q}_n^1(X_{ij})-\psi_0)\Big),$$

since $\frac{1}{n_i}\sum_{j=1}^{n_i}\bar{Q}_n^1(X_{ij})$ is invariant with respect to j. As demonstrated in Chapter 2, this EIF is doubly robust, *i.e.*, we obtain consistent estimates for ψ_0 provided one of $\bar{Q}_n^1(X_{ij})$ or $g_n(X_{ij})$ is consistently estimated.

Following the article by Rose and Van der Laan [2011], IPCW-TMLE solves the weighted influence function which in this case is given by

$$\frac{1}{N}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_n(\bar{O}_i)}\frac{1}{n_i}\sum_{j=1}^{n_i}\Big(\frac{\mathbb{I}(A_{ij}=1)}{g_n(X_{ij})}(Y_{ij}-\bar{Q}_n^1(X_{ij}))+\frac{1}{n_i}\sum_{j=1}^{n_i}\bar{Q}_n^1(X_{ij})-\psi_0)\Big)=0.$$

IPCW-TMLE inherits the double robustness property of the full data EIF in Equation (3.8) provided that $\pi_n(\bar{O}_i)$ is consistently estimated, *i.e.*, $\pi_n(\bar{O}_i) \rightarrow Pr(\Delta_i = 1|\bar{O}_i)$.

3.6.5 Variance Estimation

For the estimation of the variance of the parameter of interest obtained using IPCW-TMLE algorithm we employ the jackknife estimator previously described in Section 2.3.2. Let us assume that our parameter of interest ψ is a function of all the studies irrespective of whether they fall into the category of IPD or AD. We denote ψ_n, the estimate of ψ, by

$$\psi_n = \bar{\Psi}(O_1, O_2, ..., O_{67}).$$

We define a new variable $\psi_{-i,n}$ which denotes the estimate of ψ obtained using the full data excluding the i^{th} study. We denote $\psi_{-i,n}$ by

$$\psi_{-i,n} = \bar{\Psi}(O_1, ..., O_{i-1}, O_{i+1}, ..., O_{67}).$$

Then, the jackknife variance estimate of ψ_n is given by

$$Var(\psi_n) = \frac{66}{67} \sum_{i=1}^{67} \Big(\psi_{-i,n} - \sum_{k=1}^{67} \psi_{-k,n}\Big)^2.$$

We evaluate the performance of the IPCW-TMLE and the jackknife estimator using some simulation studies in the next chapter.

Chapter 4

Simulation Study

In this chapter, we perform some simulation studies, motivated by the MDR-TB example, and demonstrate the properties of IPCW-TMLE for mixed IPD and AD scenarios used in the simulation studies. Our goals for performing the simulation studies are to demonstrate the double robustness property of IPCW-TMLE, the importance of the second phase update step which eliminates selection bias due to neglecting the AD-only studies in the first phase, and to show the validity of the variance estimation. Section 4.1 provides the entire data generation structure of the performed simulation studies. Section 4.2 lists various estimation methods used in the analysis of the simulated data. We present the results of the analysis in Section 4.3.

4.1 Data Generation

For the data generation of the simulation, we generate 67 studies with n_i denoting the sample size of each study. We consider two scenarios for our data generation which are

- Fixed and identical sample size $n_i = 200$, for every study;
- Sample size for each study is drawn from a uniform distribution.

4.1.1 Data Generation Scenario I

We generate a clustered dataset of 67 studies with a total of 13400 patient observations. Every study, denoted by i includes 200 patient observations, denoted by j. We generate two study level variables, $V_i^{(1)}$ and $V_i^{(2)}$, one unmeasured individual level variable, $U_{ij}^{(1)}$, and two individual level covariates, $W_{ij}^{(1)}$ and $W_{ij}^{(2)}$. We generate three study level binary treatment availability indicators, $D_i^{(1)}$, $D_i^{(2)}$ and $D_i^{(3)}$. We generate three individual level binary indicators for exposure to treatments, $A_{ij}^{(1)}$, $A_{ij}^{(2)}$ and $A_{ij}^{(3)}$, which can only be non-zero if the corresponding availability indicator is 1. We generate a binary outcome variable, Y_{ij}, conditional on both patient and study-specific covariates, including the unobserved U_{ij}. Finally, the indicator of selection for study i in the second stage is given by Δ_i, where $\Delta_i = 1$ indicates that we have access to the IPD for study i. We denote the full data structure by O_i, which is given by

$$\begin{aligned} O_i &= (D_i^{(1)}, \Delta_i, X_{ij}, \Delta\{A_{ij}^{(1)}, Y_{ij}, \quad j \in S_i\}), \text{ where} \\ X_{ij} &= (V_i, \{D_i^{(k)}; k = 2, 3\}, \Delta(W_{ij}^{(1)}, W_{ij}^{(2)}, \{A_{ij}^{(k)}; k = 2, 3\}; j \in S_i)), \\ V_i &= \{V_i^{(1)}, V_i^{(2)}\} \quad i = 1, ..., 67, \end{aligned}$$

where S_i denotes the index set of patients in the i^{th} study. The above data structure is similar to that of the MDR-TB data mentioned in Chapter 3, though the number of treatments in particular is greatly reduced for simplicity. Table 4.1 represents the entire data generating function for these variables, which is also described below.

We first generate V_i and U_i for each study using a normal distribution with fixed mean and standard deviation. Every individual in study i is then allocated V_{ij} and U_{ij} as the same generated value of V_i and U_i, respectively, for each study. $U_{ij}^{(1)}$ is then generated using a normal distribution with mean as a function of U_{ij} and a fixed standard deviation. The treatment availability indicators $D_i^{(k)}$, k = 1,2,3, are then generated using three binomial distributions with probability set to be a function of V_{ij} along with the additional constraint that for every study i, $D_i^{(1)} + D_i^{(2)} + D_i^{(3)} > 0$, *i.e.*, every study has access to at least one of the 3 available treatments. Based on the above study level variables, we then generate individual level variables $W_{ij}^{(1)}$, $W_{ij}^{(2)}$, $A_{ij}^{(k)}$, k = 1,2,3 and Y_{ij}. Specifically, $W_{ij}^{(1)}$ and $W_{ij}^{(2)}$ are generated using a normal distribution with mean as a function of the variables V_{ij} and a fixed standard deviation. For every study i with $D_i^{(k)} = 1$, $A_{ij}^{(k)}$ is generated using a Bernoulli distribution with mean as a function of V_{ij}, $W_{ij}^{(1)}$, and $W_{ij}^{(1)}$. For every study i with $D_i^{(k)} = 0$, $A_{ij}^{(k)}$ was set to be 0 for each $j \in S_i$. The outcome variable Y_{ij} is then generated using a Bernoulli distribution with mean as a function of U_{ij}, $W'_{ij}s$ and $A'_{ij}s$. As mentioned previously, $U_{ij}^{(1)}$ is an unmeasured individual level variable, which is used to create random effect heterogeneity between studies. We further discuss the implications of $U_{ij}^{(1)}$ in Section 4.3.

For the purpose of simulation study, we take $\bar{O}_i$ to be

$$\bar{O}_i = (V_i, D_i^{(1)}, D_i^{(2)}, D_i^{(3)}, \bar{W}_i^{(1)}, \bar{W}_i^{(2)}, \bar{A}_i^{(1)}, \bar{A}_i^{(2)}, \bar{A}_i^{(3)}, \bar{Y}_i),\ i = 1, ..., 67,$$

where $\bar{W}_i^{(1)}$, $\bar{W}_i^{(2)}$, $\bar{A}_i^{(1)}$, $\bar{A}_i^{(2)}$, $\bar{A}_i^{(3)}$, $\bar{Y}_i$ denote the average values of the variables $W_{ij}^{(1)}$, $W_{ij}^{(2)}$, $A_{ij}^{(1)}$, $A_{ij}^{(2)}$, $A_{ij}^{(3)}$ and Y_{ij}, respectively, over study i. Δ_i, for every study i, is generated using a binomial distribution with a mean dependent on the components of $\bar{O}_i$. $\Delta_i = 1$ denotes that study i was selected in the second stage, allowing access to the individual patient data for that study.

4.1.2 Data Generation Scenario II

The variable generation for Scenario II was similar to the variable generation for Scenario I, except for the fact that the sample size of each study was obtained using an uniform distribution. In other words, for every study i, in Scenario I, the value of n_i was fixed to be 200 whereas, in Scenario II, we generate n_i using a discrete uniform distribution with a range from 50 to 1000. This simulation scenario was created to take into account the random nature of the sample size in any given study.

4.2 Analysis

After generating data as demonstrated in Section 4.1, we carry out an analysis using the methods described in Chapter 3. This section describes the aims of the simulation study, followed by description of the different estimation methods used.

4.2.1 Aims of the simulation study

The simulation studies aim to

- Estimate the target parameter ψ, defined as follows:

$$\psi = \mathbb{E}(Y(A_{ij}^{(1)}) = 1, A_{ij}^{(2)}, A_{ij}^{(3)}) = \mathbb{E}(Y(A_{ij} = 1));$$

- Verify the double-robustness property for the first phase estimation step for

IPCW-TMLE;

- Compare the results obtained by using IPCW-TMLE on the entire dataset against using TMLE only on the available IPD studies, which demonstrates the importance of the second phase;

- Check the validity of the Jackknife method for the variance estimation.

4.2.2 True Value of the Target Parameter

To obtain the true value of ψ, we increase the number of studies from 67 to 10^4 and fix the sample size of each study at 10^4. The data is then generated similarly as described in Section 4.1 except everyone in the studies to be deterministically exposed to $A^{(1)}$, and $\Delta_i = 1$ is set for all studies. The true value ψ_0 of ψ is approximated by

$$\psi_0 \approx \frac{1}{10^8} \sum_{i=1}^{10^4} \sum_{j=1}^{10^4} Y_{ij}.$$

4.2.3 Methods used for estimating ψ

For the estimation of ψ, we first obtain estimates for the Q and g components as discussed in Chapter 3. We describe the various model specification for the components which are either correctly specified or incorrectly specified.

- **Estimate for the correctly specified Q-component**: To obtain $\bar{Q}^0_{ij,n}$, we fit a logistic regression model for Y against X_{ij} on the subset of individuals in the IPD studies with $A^{(1)}_{ij} = 1$ with weights of $1/\pi_n(\bar{O}_i)$, where $\pi_n(\bar{O}_i)$ is an estimate of $\pi(\bar{O}_i)$, given by:

$$\pi(\bar{O}_i) = Pr(\Delta_i = 1|\bar{O}_i).$$

The above obtained model is then used to predict $\bar{Q}^0_{ij,n}$ for all individuals in the IPD studies using the observed pre-treatment covariates of the individuals.

- **Estimate for the incorrectly specified Q-component**: We obtain the average of Y over all the patient observations in the IPD with $A^{(1)}_{ij} = 1$. We then set $\bar{Q}^0_{ij,n}$ to be this average for all individuals in the IPD studies.

- **Estimate for correctly specified g-component**: To obtain $g_n(X_{ij})$, we first independently obtain estimates for $g^{(1)}(X_{ij})$ and $g^{(2)}(X_{ij})$ as defined in Chapter 3. For obtaining the estimate of $g^{(1)}(X_{ij})$, we fit a logistic regression model for $A^{(1)}_{ij}$ against X_{ij}, for all individuals in the IPD studies who had availability to $A^{(1)}$ with weights of $1/\pi_n(\bar{O}_i)$. This model is then used to predict values of $g^{(1)}_n(X_{ij})$ using the pre-treatment covariates of individuals in IPD study.

  For obtaining the estimate of $g^{(2)}(X_{ij})$, we fit a logistic regression model for $D^{(1)}_i$ against the study level variables V_i for all IPD studies with weights of $1/\pi_n(\bar{O}_i)$ and use this model to obtain the probability of treatment availability of $A^{(1)}$ based on the study level variables V_i for each IPD study. $g^{(2)}_n(X_{ij})$ is then set to be the above estimated probability.

  The estimate $g_n(X_{ij})$ is then taken to be the product of $g^{(1)}_n(X_{ij})$ and $g^{(2)}_n(X_{ij})$.

- **Estimate for incorrectly specified g-component**: To obtain $g_n(X_{ij})$, we again independently obtain estimates for $g^{(1)}(X_{ij})$ and $g^{(2)}(X_{ij})$, where the latter is estimated using the same procedure as for the correctly specified g-component. For obtaining the incorrectly specified estimate of $g^{(1)}(X_{ij})$, we obtain the proportion of observations with $A_{ij}^{(1)} = 1$ within studies with $D_i^{(1)} = 1$ and $\Delta_i = 1$. We then set $g_n^{(1)}$ to be this proportion for all individuals in studies with $\Delta_i = 1$. The final estimate $g_n(X_{ij})$ is then given by

$$g_n(X_{ij}) = g_n^{(1)}(X_{ij}) \cdot g_n^{(2)}(X_{ij}).$$

The above components are used to estimate ψ using TMLE or IPCW-TMLE algorithm. The complete list of estimation methods which are used for the analysis of the simulation studies are as follows:

- **IPCW-TMLE using correct Q and g components (IPCW-TMLE_Qcgc)**: This estimate is obtained using correct Q and g components in IPCW-TMLE;
- **IPCW-TMLE using correct Q component and incorrect g component (IPCW-TMLE_Qcgi)**: This estimate is obtained using correct Q component and incorrect g component in IPCW-TMLE;
- **IPCW-TMLE using incorrect Q component and correct g component (IPCW-TMLE_Qigc)**: This estimate is obtained using the incorrect Q component and correct g component in IPCW-TMLE;
- **IPCW-TMLE using incorrect Q and g components (IPCW-TMLE_Qigi)**: This estimate is obtained using the incorrect Q and g components in IPCW-TMLE;
- **Phase I TMLE using correct Q and g components (TMLE_Phase_I)**:

This estimate is obtained using the correct Q and g components in IPCW-TMLE, but omitting the IPCW update step or the second phase estimation of IPCW-TMLE. This estimation method models the transportability of the causal effects within the IPD, but does not adjust for selection bias of the second stage subsample of IPD;

- **TMLE using correct Q and g component on studies with $\boldsymbol{\Delta = 1}$ and had availability to treatment $\boldsymbol{A^{(1)}}$ (TMLE_Available)**: The estimate is obtained using TMLE on the IPD studies which had availability to treatment $A^{(1)}$. This method does not involve updating the estimate using IPCW;
- **Mean of outcomes for patients exposed to** $A^{(1)}$ **(Mean)**: This estimate is obtained by taking the sample mean of individuals exposed to the treatment $A^{(1)}$ in the IPD studies.

We use 1000 different randomly generated seed values, thereby obtaining 1000 different datasets. Our mean estimate for ψ using each method described above is obtained by taking the mean over the corresponding obtained estimate for each dataset.

We expect to obtain consistent estimates for the estimation methods using IPCW-TMLE when at least one of the Q or g components is correctly specified, due to the double-robustness property of IPCW-TMLE for the first phase, as demonstrated in Chapter 3. We expect the other estimation methods to produce biased estimates since they either omit the second phase estimation, ignore transportability, or obtain the sample mean of the population.

The results obtained using different estimation methods are compared using box plots of the Monte Carlo estimates in Section 4.3. The true value ψ_0 is represented by the green line, with the top and the bottom part of each boxes representing the 25^{th} and 75^{th} percentiles, respectively, of the estimates. The median of the estimates is portrayed by the dark black line within the box. The outliers in the box-plots are represented by the symbol $\circ$.

The standard error of the estimate in each simulated dataset is obtained using the jackknife estimator. The median jackknife standard error estimate is compared to the Monte Carlo standard error. We also construct the jackknife confidence intervals using

$$CILB = \psi_{IPCW-TMLE,n} - 1.96 * Var(\psi_{IPCW-TMLE,n}),$$
$$CIUB = \psi_{IPCW-TMLE,n} + 1.96 * Var(\psi_{IPCW-TMLE,n}),$$

where $CILB$ and $CIUB$ denote the confidence interval lower and upper bounds, respectively, and $Var(\psi_{IPCW-TMLE,n})$ denotes the obtained estimate of the jackknife variance for $\psi_{IPCW-TMLE,n}$. A 95% coverage probability is obtained for all jackknife confidence intervals and presented in the next section.

4.3 Results

The true value of ψ was estimated to be 0.52 for both the scenarios. Figure 4.1 shows the box plots for the obtained estimates under different estimation methods for Scenario I. Under the correct specification of either one of the Q or g components, the estimate obtained using the IPCW-TMLE algorithm is almost unbiased. A significant degree of bias is observed in the estimate obtained using IPCW-TMLE with an incorrect specification of both components thereby verifying the double-robustness property of the first phase of IPCW-TMLE.

The estimate obtained using only TMLE while incorporating transportability, **TMLE_Qcgc**, is seen to exhibit a low degree of bias whereas a significant amount of bias is observed for **TMLE_Available** since it excluded the IPD studies which didn't have availability of $A^{(1)}$. The estimate obtained by taking the mean outcomes of individuals exposed to the treatment $A^{(1)}$ is even more biased. We observe more outliers with IPCW-TMLE as compared to other estimation methods.

The estimates of ψ obtained using these estimation methods rarely produced estimates greater than 1 which is caused due to the higher sampling weights used in the second phase estimation. This problem occurred due to a low number of studies and we expect this problem to vanish when there are a large number of studies.

Figure 4.2 shows the box plots for the obtained estimates using different estimation methods for Scenario II with similar trends as in Scenario I.

Table 4.2 displays the results for the mean estimates, Monte Carlo standard errors, median jackknife standard errors and the coverage probabilities of the confidence intervals using the estimation methods described in Section 4.2.3 for the simulation data in Scenario I. As observed in the box plot for Scenario I, IPCW-TMLE estimation methods with atleast one correctly specified Q or g component produces consistent estimates, whereas the rest of the methods has a significant amount of bias in their estimates.

For correctly specified Q and g components, IPCW-TMLE produces a coverage probability of 0.95, whereas when one of the Q or g components is correctly specified, IPCW-TMLE produces a coverage proportion slightly higher than 0.95, whereas the rest of the estimation methods exhibit poor coverage due to the bias in the estimates.

Similar Monte Carlo and median jackknife standard errors were obtained for all the methods. We omitted the mean jackknife standard errors since the obtained jackknife standard error contains outliers which overestimates the mean for the jackknife variance.

Table 4.3 displays the results for the analysis carried out on the dataset generated by Scenario II. The estimation methods produced similar mean estimates as seen in Scenario I. This scenario also produces optimal coverage for the estimation methods using IPCW-TMLE with atleast one correct Q or g component.

We constructed this simulation study to demonstrate the advantages of the IPCW-TMLE algorithm for mixed IPD and AD. We misspecified or eliminated components in the model to demonstrate the potential importance different modeling components has. We specifically generated a random effects model by using the unobserved study level confounder $U_{ij}^{(1)}$, for the outcome generation in the simulation study. In the absence of this random effects model, instead of using the IPCW-TMLE algorithm, ψ can be estimated using simple TMLE for studies with availability to treatment $A^{(1)}$. Treatment effect heterogeneity is incorporated into the data generating mechanism by making sure that each study in the dataset had a different expected outcome for individuals exposed to $A^{(1)}$ given the measured covariates.

We also demonstrated the importance of modeling transportability for estimating the treatment effect for partial IPD. As shown in the simulation studies, ignoring the differential treatment availability leads to additional bias in the estimation of the causal parameter ψ. The transportability assumption holds in the generated data because all study level confounders involved in simulating the treatment availability were considered to be observed. Generating unobserved study or individual level confounders would lead to a violation of the transportability and the conditional exchangeability assumptions, resulting in an inestimable causal effect. Further, the inclusion of unobserved study level variables in the generation of the binary indicator of selection Δ would lead to a violation of the Missing at Random (MAR) assumption making it impossible to identify the target parameter ψ from the observed data.

Table 4.1: Complete Data Generating Mechanism for Scenario 1.

	Generation Mechanism
Study Level Variables	
V_{ij}	$V_i^{(1)} \sim \mathcal{N}(mean = 0.45, s.d. = 1)$
	$V_i^{(2)} \sim \mathcal{N}(mean = 0.5, s.d. = 1), \quad i =$ 1,...,67
	Now we set $V_{ij}^{(1)} = V_i^{(1)}$ & $V_{ij}^{(2)} = V_i^{(2)}$ for each j in study i
D_{ij}	$D_i^{(1)} \sim Bin(expit(0.33 + 1.1V_i^{(1)}))$
	$D_i^{(2)} \sim Bin(expit(0.76 + 0.5V_i^{(2)}))$
	$D_i^{(3)} \sim Bin(expit(0.55 + 0.6V_i^{(1)})), \quad i =$ 1,...,67
	an additional constraint that for each i, $D_i^{(1)} + D_i^{(2)} + D_i^{(3)} >$ 0
	Now we set $D_{ij}^{(1)} = D_i^{(1)}$, $D_{ij}^{(2)} = D_i^{(2)}$ & $D_{ij}^{(3)} = D_i^{(3)}$ for each j in study i
Individual Level Variables	for $i =$ 1,...,67 and $j =$ 1,...,200
$U_{ij}^{(1)}$	$U_i \sim \mathcal{N}(mean = 0.55, s.d. = 0.7), \quad i =$ 1,...,67
	Now set $U_{ij} = U_i$ for each j in study i
	Now $U_{ij}^{(1)} = \mathcal{N}(mean = 0.4U_{ij}, s.d. = 0.8)$
W_{ij}	$W_{ij}^{(1)} \sim \mathcal{N}(mean = 0.1 + 0.35V_{ij}^{(1)}, s.d. = 0.5)$
	$W_{ij}^{(2)} \sim \mathcal{N}(mean = 0.15V_{ij}^{(2)}, s.d. = 0.6)$
A_{ij}	$A_{ij}^{(1)} \sim Bin(D_{ij}^{(1)}(expit(-0.45 + V_{ij}^{(1)} + 0.4W_{ij}^{(1)} + 1.8W_{ij}^{(2)})))$
	$A_{ij}^{(2)} \sim Bin(D_{ij}^{(2)}(expit(-0.55 + 2V_{ij}^{(1)} + W_{ij}^{(1)} + W_{ij}^{(2)})))$
	$A_{ij}^{(3)} \sim Bin(D_{ij}^{(3)}(expit(-0.1 + 1.4V_{ij}^{(1)} + 0.35W_{ij}^{(1)} + 1.6W_{ij}^{(2)})))$
Y_{ij}	$Y_{ij} \sim Bin(expit(0.7 - 3.2W_{ij}^{(1)} - 1.1U_{ij}^{(1)}A_{ij}^{(1)} + 0.15A_{ij}^{(2)} - 0.45A_{ij}^{(3)}))$
Binary Selection Indicator	
Δ_i	$\Delta_i \sim Bin(expit(0.32 + 1.3\bar{W}_i^{(1)} * D_i^{(1)} - 1.2\bar{W}_i^{(2)} - 0.5\bar{A}_i^{(1)} + 0.52\bar{A}_i^{(2)} - 0.7\bar{Y}_i)),$
	$i =$ 1,...,67

Table 4.2: Results for Scenario I.

	$\mathbb{E}(\hat{Y}(A^1 = 1))$	Monte Carlo	Median Jackknife	Coverage
		standard error	standard error	Probability
Estimating Methods				
IPCW-TMLE_Qcgc	0.51	0.07	0.07	0.95
IPCW-TMLE_Qcgi	0.52	0.07	0.06	0.98
IPCW-TMLE_Qigc	0.52	0.08	0.07	0.96
IPCW-TMLE_Qigi	0.45	0.07	0.06	0.73
TMLE_Phase_I	0.48	0.03	0.03	0.81
TMLE_Available	0.44	0.03	0.03	0.34
Mean	0.39	0.03	0.03	0.01

Table 4.3: Results for Scenario II.

	$\mathbb{E}(\hat{Y}(A^1 = 1))$	Monte Carlo	Median Jackknife	Coverage
		standard error	standard error	Probability
Estimating Methods				
IPCW-TMLE_Qcgc	0.51	0.08	0.08	0.94
IPCW-TMLE_Qcgi	0.52	0.08	0.08	0.94
IPCW-TMLE_Qigc	0.52	0.08	0.08	0.94
IPCW-TMLE_Qigi	0.46	0.08	0.07	0.80
TMLE_Phase_I	0.48	0.03	0.03	0.77
TMLE_Available	0.44	0.03	0.03	0.40
Mean	0.39	0.03	0.03	0.03

Figure 4.1: Box plots for the estimate ψ, observed using different estimation methods for Scenario I.

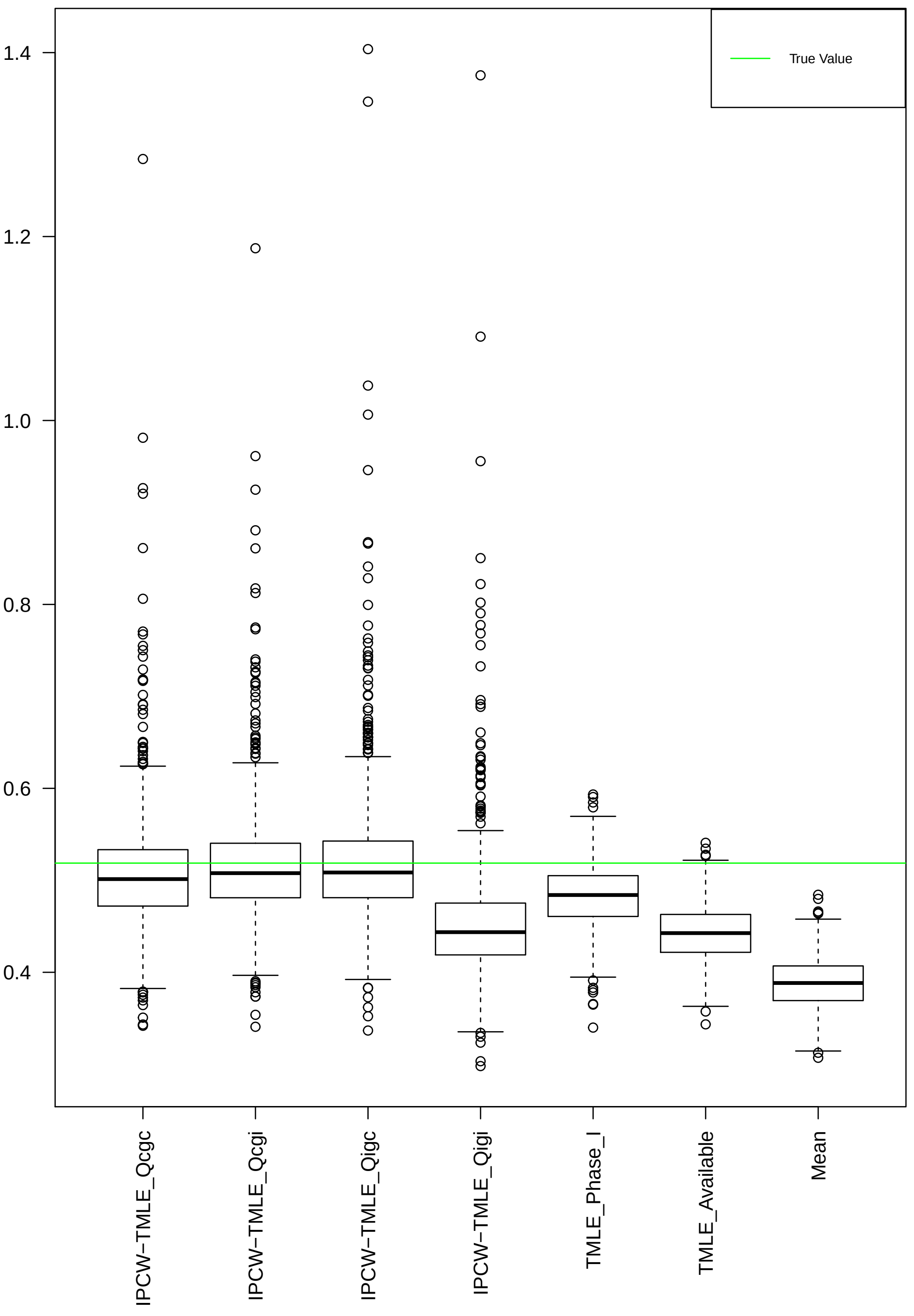
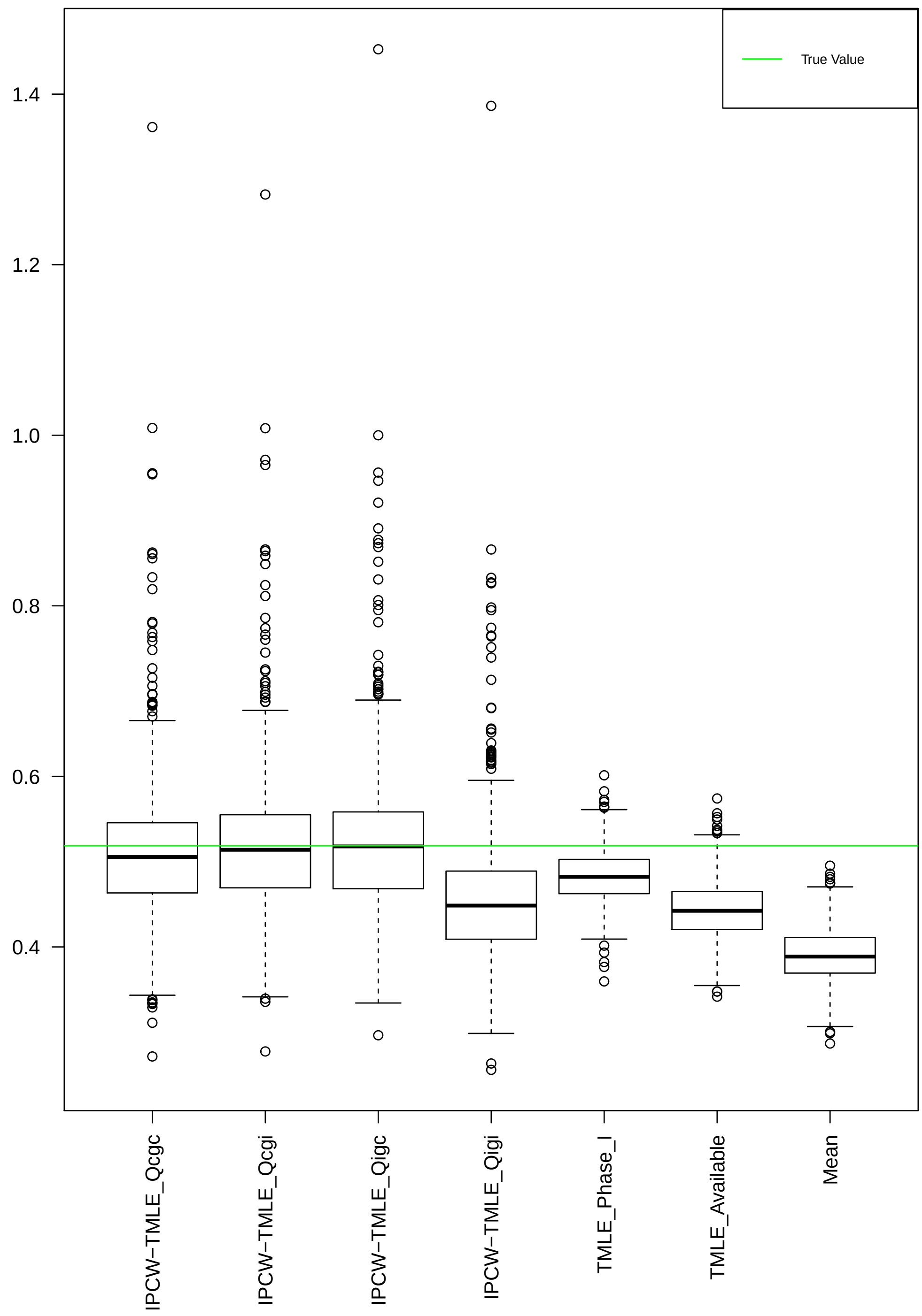


Figure 4.2: Box plots for the estimate ψ, observed using different estimation methods for Scenario II.



Chapter 5

Discussion

In this thesis, we have proposed a novel application of IPCW-TMLE for mixed AD and IPD meta-analysis. Previous work on mixed AD and IPD did not demonstrate nonparametric identifiability of the causal parameter of interest. We have established the nonparametric identifiability of the causal parameter and have shown that we can obtain consistent estimates using causal estimation methods. In particular, the related two-stage method [Tudur et al., 2001] did not adjust for the study-level covariates. In addition, they relied on parametric modeling whereas TMLE in principle can also be performed under non-parametric setting. However, for simplicity, we have used parametric working models here when fitting the TMLE.

We demonstrated the finite-sample performance of our method using simulation studies. We verified the double robustness property of IPCW-TMLE for mixed AD and IPD. In the presence of a random selection of the IPD studies in the second stage one does not need to perform the IPCW step, since the single phase TMLE would be consistent. Further, provided that specific studies do not have access to the treatment of interest, following the transportability procedure reduces the bias in the estimation of the causal parameter in the more general population.

In our algorithm, the first phase uses TMLE, which is a bounded estimator. Without the second phase, estimation would therefore be bounded. The weighting step in the second phase concludes with a weighted sum of the estimates of the IPD studies. This results in an unbounded final estimate of ψ_0. A possible solution to reduce the variability of the estimate may use data-adaptive truncation of the weights, which has previously been proposed for standard propensity scores [Ju et al.]. In practice, one may set the estimates for $\psi_{IPCW-TMLE,n}$ to be 1, provided that the value exceeds 1, for the estimate as well as the jackknife procedure. Note that the upper confidence bound may also exceed 1.

Future work includes implementation of this method to estimate the expected potential outcome on real-life data such as the MDR-TB data. This data includes additional complexity such as a higher number of available treatments, patients that are resistant to some treatments, *etc.* Furthermore, this proposed method can also be used to estimate the effects by regimen, rather than by single treatment, where regimen is defined to be some specific combination of the treatments [Siddique et al., 2018].

Appendix A

R-codes

A.1 Data Generation Code for Scenario-I

```
set.seed(2014)
  N_Sampled <- 67
  Sample_Size <- rep(200,N_Sampled)
  V1 <- rnorm(N_Sampled,0.45,1)
  V2 <- rnorm(N_Sampled,0.5,1)
  U1 <- rnorm(N_Sampled,0.55,0.7)
  #U2 <- rnorm(N_Sampled,0.6,0.9)
  D1 <- NA
  D2 <- NA
  D3 <- NA
  i <- 1
  while(i <= N_Sampled){
    D1[i] <- rbinom(1,1,plogis(0.33+1.1*V1[i]))
    D2[i] <- rbinom(1,1,plogis(0.76+0.5*V2[i]))
    D3[i] <- rbinom(1,1,plogis(0.55+0.6*V1[i]))
    if((D1[i]+D2[i]+D3[i])>0)
      i <- i+1
  }
  V1_ij <- c()
  V2_ij <- c()
  U_ij <- c()
  D1_ij <- c()
  D2_ij <- c()
  D3_ij <- c()
  Study_ID <- c()
  for(i in 1:N_Sampled){
    V1_ij <- append(V1_ij,rep(V1[i],Sample_Size[i]))
    V2_ij <- append(V2_ij,rep(V2[i],Sample_Size[i]))
    U_ij <- append(U_ij,rep(U1[i],Sample_Size[i]))
    #U2_ij <- append(U2_ij,rep(U2[i],Sample_Size[i]))
    D1_ij <- append(D1_ij,rep(D1[i],Sample_Size[i]))
    D2_ij <- append(D2_ij,rep(D2[i],Sample_Size[i]))
    D3_ij <- append(D3_ij,rep(D3[i],Sample_Size[i]))
    Study_ID <- append(Study_ID,rep(i,Sample_Size[i]))
  }
  N <- sum(Sample_Size)
  U1_ij <- rnorm(N,mean=0.4*U_ij,sd = 0.8)
  W1_ij <- rnorm(N,mean=0.1+0.35*V1_ij,sd=0.5)
  W2_ij <- rnorm(N,mean=0.25*V2_ij,sd=0.6)
```

```
A1_ij <- rbinom(N,1,D1_ij*plogis(-0.45+V1_ij+0.4*W1_ij+
                1.8*W2_ij))
A2_ij <- rbinom(N,1,D2_ij*plogis(-0.55+2*V1_ij+W1_ij+W2_ij))
A3_ij <- rbinom(N,1,D3_ij*plogis(-0.1+1.4*V1_ij+0.35*W1_ij+
                1.6*W2_ij))
Y_ij <- rbinom(N,1,plogis(0.7-3.2*W1_ij+1.1*U1_ij*A1_ij+
              0.15*A2_ij-0.45*A3_ij*W2_ij))


Full_IPD <- data.frame(Study_ID,W1_ij,W2_ij,A1_ij,A2_ij,A3_ij,
                      Y_ij,D1_ij,D2_ij,D3_ij,V1_ij,V2_ij)
Full_AD <- c()
for(i in 1:N_Sampled){
  IPD <- Full_IPD[which(Full_IPD$Study_ID==i),]
  Full_AD <- rbind(Full_AD,colMeans(IPD))
}
colnames(Full_AD) <- c("Study_ID","W1_bar","W2_bar","A1_bar",
                      "A2_bar","A3_bar","Y_bar","D1_bar",
                      "D2_bar","D3_bar","V1","V2")
Full_AD <- data.frame(Full_AD)
pInd <- plogis(0.25+1.3*Full_AD$W1_bar*Full_AD$D1 -
              1.2*Full_AD$W2_bar - 0.5*Full_AD$A1_bar +
              0.52*Full_AD$A2_bar - 0.7*Full_AD$Y_bar)
Indicator_bar <- rbinom(N_Sampled,1,pInd)
```

A.2 Data Generation Code for Scenario-II

```
set.seed(683447215)
  N_Sampled <- 67
  Sample_Size <- floor(runif(67,50,1000))
  V1 <- rnorm(N_Sampled,0.45,1)
  V2 <- rnorm(N_Sampled,0.5,1)
  U1 <- rnorm(N_Sampled,0.55,0.7)
  D1 <- NA
  D2 <- NA
  D3 <- NA
  i <- 1
  while(i <= N_Sampled){
    D1[i] <- rbinom(1,1,plogis(0.33+1.1*V1[i]))
    D2[i] <- rbinom(1,1,plogis(0.76+0.5*V2[i]))
    D3[i] <- rbinom(1,1,plogis(0.55+0.6*V1[i]))
    if((D1[i]+D2[i]+D3[i])>0)
      i <- i+1
  }
  V1_ij <- c()
  V2_ij <- c()
  U_ij <- c()
  D1_ij <- c()
  D2_ij <- c()
  D3_ij <- c()
  Study_ID <- c()
```

```
for(i in 1:N_Sampled){
  V1_ij <- append(V1_ij,rep(V1[i],Sample_Size[i]))
  V2_ij <- append(V2_ij,rep(V2[i],Sample_Size[i]))
  U_ij <- append(U_ij,rep(U1[i],Sample_Size[i]))
  D1_ij <- append(D1_ij,rep(D1[i],Sample_Size[i]))
  D2_ij <- append(D2_ij,rep(D2[i],Sample_Size[i]))
  D3_ij <- append(D3_ij,rep(D3[i],Sample_Size[i]))
  Study_ID <- append(Study_ID,rep(i,Sample_Size[i]))
}
N <- sum(Sample_Size)
U1_ij <- rnorm(N,mean=0.4*U_ij,sd = 0.8)
W1_ij <- rnorm(N,mean=0.1+0.35*V1_ij,sd=0.5)
W2_ij <- rnorm(N,mean=0.25*V2_ij,sd=0.6)
A1_ij <- rbinom(N,1,D1_ij*plogis(-0.45+V1_ij+0.4*W1_ij+
                1.8*W2_ij))
A2_ij <- rbinom(N,1,D2_ij*plogis(-0.55+2*V1_ij+W1_ij+W2_ij))
A3_ij <- rbinom(N,1,D3_ij*plogis(-0.1+1.4*V1_ij+0.35*W1_ij+
                1.6*W2_ij))
Y_ij <- rbinom(N,1,plogis(0.7-3.2*W1_ij+1.1*U1_ij*A1_ij+
              0.15*A2_ij-0.45*A3_ij*W2_ij))

Full_IPD <- data.frame(Study_ID,W1_ij,W2_ij,A1_ij,A2_ij,A3_ij,
                      Y_ij,D1_ij,D2_ij,D3_ij,V1_ij,V2_ij)
Full_AD <- c()
for(i in 1:N_Sampled){
  IPD <- Full_IPD[which(Full_IPD$Study_ID==i),]
  Full_AD <- rbind(Full_AD,colMeans(IPD))
}
colnames(Full_AD) <- c("Study_ID","W1_bar","W2_bar","A1_bar",
                      "A2_bar","A3_bar","Y_bar","D1_bar",
                      "D2_bar","D3_bar","V1","V2")
Full_AD <- data.frame(Full_AD)
pInd <- plogis(0.25+1.3*Full_AD$W1_bar*Full_AD$D1 -
              1.2*Full_AD$W2_bar - 0.5*Full_AD$A1_bar +
              0.52*Full_AD$A2_bar - 0.7*Full_AD$Y_bar)
Indicator_bar <- rbinom(N_Sampled,1,pInd)
```

A.3 Code for obtaining the true value of the causal quantity for the simulation data

```
set.seed(2001)
Sample_size <- 10000
V_1 <- rnorm(10000,0.45,1)
V_2 <- rnorm(10000,0.5,1)
U_1 <- rnorm(10000,0.55,0.7)
M1 <- 0.1+0.35*V_1
M2 <- 0.25*V_2
M3 <- 0.4*U_1
pInd2 <- plogis(0.76+0.5*V_2)
pInd3 <- plogis(0.55+0.6*V_1)
In2 <- rbinom(10000,1,prob = pInd2)
In3 <- rbinom(10000,1,prob = pInd3)

Estimate_Population <- function(Sample_size,Mean1,Mean2,Mean3,
                                Ind2,Ind3,V1,V2){
  W1 <- rnorm(Sample_size,Mean1,0.5)
  W2 <- rnorm(Sample_size,Mean2,0.6)
  U1 <- rnorm(Sample_size,Mean3,0.8)

  A1 <- rep(1,Sample_size)

  pA2 <- plogis(-0.55+2*V1+W1+W2)
  pA3 <- plogis(-0.1+1.4*V1+0.35*W1+1.6*W2)
  A2 <- rbinom(Sample_size,1,prob = pA2*Ind2)
  A3 <- rbinom(Sample_size,1,prob = pA3*Ind3)

  pY <- plogis(0.7-3.2*W1+1.1*U1*A1+0.15*A2-0.45*A3*W2)
  Y <- rbinom(Sample_size,1,prob = pY)

  return(mean(Y))
}

Causal_Estimate <- c()
for(i in 1:10000){
  Causal_Estimate[i] <- Estimate_Population(Sample_size,M1[i],
                             M2[i],M3[i],In2[i],In3[i],
                             V_1[i],V_2[i])
  print(i)
}
True_Value_Study <- mean(Causal_Estimate)
```

Bibliography

S. D. Ahuja, D. Ashkin, M. Avendano, et al. Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: an individual patient data meta-analysis of 9,153 patients. *PLoS Medicine*, 9(8):e1001300, 2012.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

E. Bareinboim and J. Pearl. Transportability of causal effects: Completeness results. *Proceedings of The Twenty-Sixth Conference on Artificial Intelligence (AAAI 2012)*, 698-704, 2012.

C. B. Begg and J. A. Berlin. Publication bias and dissemination of clinical research. *JNCI: Journal of the National Cancer Institute*, 81(2):107–115, 1989.

M. Blettner, W. Sauerbrei, B. Schlehofer, T. Scheuchenpflug, and C. Friedenreich. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology*, 28(1):1–9, 1999.

H. E. Brady. Models of causal inference: Going beyond the Neyman-Rubin-Holland theory. In *Annual Meetings of the Political Methodology Group*, 2002.

L. E. Braitman and P. R. Rosenbaum. Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of Internal Medicine*, 137(8):693–695, 2002.

N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20, 1988.

K. A. Broeze, B. C. Opmeer, F. Van der Veen, P. M. Bossuyt, S. Bhattacharya, and B. W. Mol. Individual patient data meta-analysis: a promising approach for evidence synthesis in reproductive medicine. *Human Reproduction Update*, 16(6): 561–567, 2010.

D. L. Burke, J. Ensor, and R. D. Riley. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in Medicine*, 36(5):855–875, 2017.

D. P. Byar, R. M. Simon, W. T. Friedewald, J. J. Schlesselman, D. L. DeMets, J. H. Ellenberg, M. H. Gail, and J. H. Ware. Randomized clinical trials: perspectives on some recent ideas. *New England Journal of Medicine*, 295(2):74–80, 1976.

D. R. Cox. *Planning of experiments.* Oxford, England: Wiley, 1958.

R. B. d'Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19): 2265–2281, 1998.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, 45:139–145, 2015.

S. Duval and R. Tweedie. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463, 2000.

W. D. Flanders and S. Greenland. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5):739–747, 1991.

H. Goldstein, M. Yang, R. Omar, R. Turner, and S. Thompson. Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):399–412, 2000.

S. Greenland and B. Brumback. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31(5):1030–1037, 2002.

S. Gruber and M. J. van der Laan. An application of targeted maximum likelihood estimation to the meta-analysis of safety data. *Biometrics*, 69(1):254–262, 2013.

A. B. Haidich. Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29, 2010.

E. L. Hannan. Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations. *JACC: Cardiovascular Interventions*, 1(3):211–217, 2008.

L. V. Hedges. Modeling publication selection effects in meta-analysis. *Statistical Science*, 7(2):246–255, 1992.

M. A. Hernán. Beyond exchangeability: the other conditions for causal inference in medical research. *Statistical Methods in Medical Research*, 21(1):3–5, 2012.

M. A. Hernán and T. J. VanderWeele. Compound treatments and transportability of causal inference. *Epidemiology*, 22(3):368, 2011.

T. B. Huedo-Medina, J. Sánchez-Meca, F. Marín-Martínez, and J. Botella. Assessing heterogeneity in meta-analysis: Q statistic or i^2 index? *Psychological Methods*, 11 (2):193, 2006.

N. R. N. Idris and N. A. Misran. A modified two-stage method for combining the aggregate-data and individual-patient-data in meta-analysis. *Journal of Applied Sciences*, 15(10):1231, 2015.

C. Jackson, A. N. Best, and S. Richardson. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1): 159–178, 2008.

C. Ju, J. Schwab, and M. J. Van der Laan. On adaptive propensity score truncation in causal inference. *Technical Report*. University of California, Berkeley, 2017.

J. D. Y. Kang, J. L. Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.

K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes. Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, 96(3):118–121, 2003.

B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, 2010.

K. E. Mansfield, J. Sim, J. L. Jordan, and K. P. Jordan. A systematic review and meta-analysis of the prevalence of chronic widespread pain in the general population. *Pain*, 157(1):55, 2016.

V. M. Montori, M. F. Swiontkowski, and D. J. Cook. Methodologic issues in systematic reviews and meta-analyses. *Clinical Orthopaedics and Related Research*, 413: 43–54, 2003.

J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938.

J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 540–547. IEEE, 2011.

J. Pearl and E. Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. Van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.

K. E. Porter, S. Gruber, M. J. Van der Laan, and J. S. Sekhon. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics*, 7(1):1–34, 2011.

J. L. Powell. Estimation of semiparametric models. In *Handbook of Econometrics*. 4:2443-2521. Elsevier, 1994.

R. D. Riley, M. C. Simmonds, and M. P. Look. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology*, 60(5):431–e1, 2007.

R. D. Riley, P. C. Lambert, and G. Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal*, 340:c221, 2010.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

D. M. Rocke and G. W. Downs. Estimating the variances of robust estimators of location: influence curve, jackknife and bootstrap: Robust estimators of location. *Communications in Statistics-Simulation and Computation*, 10(3):221–248, 1981.

S. Rose and M. J. Van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*, 7(1):1–21, 2011.

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

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

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

D. B. Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Przemysław Ryś, Magdalena Władysiuk, Iwona Skrzekowska-Baran, and Maciej T Małecki. Review articles, systematic reviews and meta-analyses: which can be trusted? *Polskie Archiwum Medycyny Wewnetrznej*, 119(3):148–156, 2009.

J. Sarker. Ethical issues of randomized controlled trials. *Bangladesh Journal of Bioethics*, 2014.

M. E. Schnitzer, J. J. Lok, and R. J. Bosch. Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring. *Biostatistics*, 17(1):165–177, 2015.

S. Schwartz, N. M. Gatto, and U. B. Campbell. Extending the sufficient component cause model to describe the Stable Unit Treatment Value Assumption (SUTVA). *Epidemiologic Perspectives & Innovations*, 9(1):3, 2012.

A. A. Siddique, M. E. Schnitzer, A. Benedetti, et al. Causal inference with multiple concurrent medications: a comparison of methods and an application in multidrug-resistant tuberculosis. *Statistical Methods in Medical Research*, 2018. To be resubmitted.

M. C. Simmonds. *Statistical Methodology of Individual Patient Data*. PhD thesis, University of Cambridge, 2005.

R. R. Sitter. Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92(438):780–787, 1997.

L. A. Stewart and J. F. Tierney. To ipd or not to ipd? advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions*, 25(1):76–97, 2002.

A. J. Sutton, D. Kendrick, and C. A. C. Coupland. Meta-analysis of individual-and aggregate-level data. *Statistics in Medicine*, 27(5):651–669, 2008.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2007.

C. Tudur, P. R. Williamson, S. Khan, and L. Y. Best. The value of the aggregate data approach in meta-analysis with time-to-event outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(2):357–370, 2001.

M. J. Van der Laan and S. Gruber. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *The International Journal of Biostatistics*, 12(1):351–378, 2016.

M. J. Van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.

M. J. Van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

A. Van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 29(4):679–686, 2014.

G. Wang. *Estimating Treatment Importance in Multidrug-resistant Tuberculosis Using Targeted Learning: An Observational Individual Patient Data Network Meta-analysis*. PhD thesis, McGill University, 2018.

W. Wang, D. Scharfstein, Z. Tan, and E. J. MacKenzie. Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):947–969, 2009.

J. E. White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128, 1982.

World Health Organization. *WHO Treatment Guidelines for Drug-Resistant Tuberculosis 2016 Update*. World Health Organization, 2016.

L. P. Zhao and S. Lipsitz. Designs and analysis of two-stage studies. *Statistics in Medicine*, 11(6):769–782, 1992.

A. Ziegler. Practical considerations of the jackknife estimator of variance for generalized estimating equations. *Statistical Papers*, 38(3):363–369, 1997.