Estimating treatment importance in multidrug-resistant tuberculosis using Targeted Learning: an observational individual patient data network meta-analysis

Guanbo Wang

Master of Science

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal,Quebec

2017-08-11

Report of thesis carried out as a requirement of the M.Sc. Thesis Program in Biostatistics, Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

DEDICATION

This thesis is dedicated to my parents, Hongguang Jia and Zhiyou Wang.

ACKNOWLEDGEMENTS

I am very pleased to be able to thank the many people that have assisted me in producing this thesis.

I would sincerely thank my supervisors Dr. Andrea Benedetti and Dr. Mireille Schnitzer for their guidance, encouragement, dedication, understanding, tolerance and financial support during my graduate study at McGill University. I would also thank my friends Dr. Yi Yang, Menglan Pang, Shomoita Alam, Alexander Levis and Bo Chen who supported me a lot in this long period. I deeply value their help. Finally, my dissertation work would not go smoothly without the support from my family. Particularly, I would thank my parents.

I would also like to express sincere thanks to the faculties Drs. Robert Platt, James Hanley, Erica Moodie and Tibor Schuster and fellow students in the Department of Epidemiology, Biostatistics and Occupational Health at McGill University for fostering a welcoming and engaging atmosphere for study. I would also like to thank our wonderful department administrators who help create a warm and friendly environment in our workplace.

ABSTRACT

Multi-drug-resistant tuberculosis (MDR-TB) is defined as strains of tuberculosis (TB) that do not respond to at least the two most powerful anti-TB drugs. Nowadays, MDR-TB continues to emerge and thus accurate assessment of the importance of treatments for MDR-TB is a critical issue. MDR-TB is often treated with multiple first and second line antibiotics.

Our data consists of individual patient data from 31 international observational studies which measured patient demographic information, medical history, medications used and therapeutic outcomes. In this study, we defined an adjusted variable importance metric to compare the apparent contribution of each medication to the overall recovery rate among patients who are not known to be resistant to the given antibiotic.

While individual studies are able to evaluate subpopulation-specific associations between treatment and outcome, our individual patient data network meta-analysis (IPD-NMA) allows for a global perspective on average medication importance in the treatment of MDR-TB.

To these ends, we develop identifiability criteria and apply targeted maximum likelihood estimation (TMLE) to estimate the adjusted recovery rate means for each treatment amongst patients who were not known to be resistant to the treatment. TMLE is a semi-parametric and double robust method. Throughout the analysis, transportability is utilized to translate the estimation from studies where a given antibiotic was used to studies where it wasn't. Finally, we adopted a clustered sandwich

estimator derived from the efficient influence function to compute variance estimates. Simulation studies were conducted to assess the performance of our estimator and verify the theoretical double robustness property. These simulations were also used to evaluate the validity of the sandwich estimator for variance estimation and the coverage rate of the derived Wald-type confidence intervals.

The results show that Ciprofloxacin has the greatest treatment importance, followed by Amikacin and High-generation Quinolones. They also show that Para-aminosalicylic acid, Pyrazinamide, and Group 5 level drugs are the least important.

ABRÉGÉ

La tuberculose multirésistante est une souche de la tuberculose qui ne répond pas au moins aux deux médicaments les plus puissants contre cette dernière. La tuberculose multirésistante continue de se propager de nos jours, et par conséquent, il est important de savoir évaluer avec précision l'importance des traitements pour cette souche. La tuberculose multirésistante est souvent traitée par de multiples antibiotiques de première et de seconde ligne.

Notre base de données se compose des données individuelles des patients provenant de 31 études observationnelles et contient des caractéristiques démographiques, l'historique médical, les médicaments utilisés et les issues thérapeutiques. Dans cette étude, nous définissons une métrique mesurant l'importance d'une variable afin de comparer la contribution apparente de chaque médicament au taux de récupération global parmi les patients qui ne sont pas connus pour être résistant à l'antibiotique donné.

Bien que les études individuelles puissent évaluer les associations spécifiques entre le traitement et les résultats dans la sous-population, notre méta-analyse sur les données individuelles des patients permet une perspective globale de l'importance moyenne des médicaments dans le traitement de la tuberculose multirésistante.

Pour ce faire, nous avons développé des critères d'identification et appliqué l'estimation par maximum de vraisemblance ciblée (TMLE) pour estimer le taux moyen de récupération ajusté pour chaque traitement chez les patients qui ne sont pas connus pour être résistants au traitement. TMLE est une méthode semi-paramétrique et doublement robuste. Au cours de l'analyse, la transportabilité est utilisée pour

transférer l'estimation des études où un traitement a été observé aux études où ce traitement n'a pas été observé. Enfin, nous avons adopté un estimateur sandwich dérivé de la fonction d'influence efficiente pour estimer leur variance associée.

Des études de simulation sont menées pour prouver la validité de notre estimateur et vérifier la double robustesse de notre estimateur. En outre, ils montrent également que notre méthode d'estimation de la variance est appropriée avec un taux de couverture adéquat.

Les résultats montrent que la ciprofloxacine a la plus grande importance au traitement, suivie de l'amikacine et des quinolones à haute génération. Ils montrent également que les médicaments à base d'acide para-aminosalicylique, de pyrazinamide et de groupe 5 sont les moins importants.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

Multidrug-resistant Tuberculosis (MDR-TB) is a type of Tuberculosis (TB) that is resistant to at least Isoniazid and Rifampin, the two most commonly prescribed TB drugs [56]. MDR-TB is more challenging to treat than common TB, because MDR-TB requires a combination of antibiotics, most of which cause serious side effects [3]. Therefore, identification of the most effective antibiotic regimens for MDR-TB is urgently needed.

In this setting, statistical analysis is complicated because patients with MDR-TB usually take multiple treatments and each subject's MDR-TB may be resistant to various treatments. Nevertheless, many analyses have investigated treatment effectiveness in MDR-TB [27, 46, 76]. In particular, meta-analysis is often employed to obtain a global conclusion from numerous studies. In this analysis, we use individual patient data (IPD) from multiple studies consisting of observations of over 9000 subjects. Methodology for the estimation of treatment importance in MDR-TB using IPD meta-analysis has not yet been well developed, and is complicated by the fact that not all treatments were available in every study.

Randomized Clinical Trials (RCTs) are considered to be the "gold standard" in the estimation of causal effects. In ideal circumstances, the difference in outcomes between the different treatment groups represents the relative effects of the treatment [72]. However, RCTs are not always feasible due to their high cost, limited study

participation, treatment contamination, and potential ethical issues. In terms of MDR-TB, for example, forcing some patients to have certain medications to satisfy the RCT implementation is not ethical, and it is very expensive to implement a RCT with large samples due to the high cost of MDR-TB medications. On the other hand, observational studies are easier to implement and may lead to valid conclusions on a more diverse population. This thesis uses IPD from multiple observational studies in order to investigate the treatment importance of multiple MDR-TB treatments.

The data source of meta-analysis can be either aggregated data (AD) or IPD. Because multiple antibiotics are available, network meta-analysis (NMA), which can compare multiple treatment effects through both direct and indirect evidence, may be an appropriate approach to analyze this type of data. Some researchers have used AD to perform NMA [39, 45]. However, this kind of analysis has several disadvantages. First, AD is dependent on the quality of reporting in the individual studies and they do not allow for the investigation of subgroup effects if the included studies did not report subgroup analyses. A standard random effects network meta-analysis using IPD was performed by Ahuja et al. in 2012 to investigate relative treatment effectiveness in MDR-TB [11]. The study suggested that the use of later generation quinolones, ofloxacin, and ethionamide/prothionamide as part of multi-drug regimens were associated with recovery from MDR-TB [11]. They used a mixed effects model to compare the odds ratios of various treatments. However, the parametric models used required assumptions such as the independence and normal distribution of the within-group errors and they did not take into consideration the selection bias arising from the usage of only a subset of treatments in each study. Under violation

of the parametric modeling assumptions and if the usage of treatments across studies was non-random, the results from such a study may be biased.

Targeted Maximum Likelihood Estimation (TMLE) was proposed by Van der Laan and Rubin [83]. TMLE is based on semiparametric theory, which relaxes the strong assumptions made by parametric models. It only requires the specification of certain components of the likelihood [86]. By adopting TMLE, one can obtain a regular, asymptotically linear estimator. Under correct specification of the certain components, the estimator has the lowest asymptotic variance in its class of semiparametric estimators [86]. In addition, TMLE is doubly robust, meaning that if either of the two model components is correctly specified, the estimator is consistent [86]. Furthermore, TMLE can incorporate flexible prediction methods in its estimation. SuperLearner, an ensemble learning method, is often recommended to improve the modeling in TMLE [86].

This thesis investigates MDR-TB treatment importance by conducting an IPD-NMA using TMLE. Treatment importance here is defined as the difference between the average recovery rate of the patients who used the treatment and the average recovery rate of all the patients while controlling for confounding and other treatments. As such it is the nonparametric analogue of the coefficient of each treatment of interest in an adjusted linear regression. Since different bacterial strains are resistant to different antibiotics, we only used the data from the patients who were not or not known to be resistant. Because many treatments were not available to the patients in every study, we develop transport models to extrapolate the information from

available studies to the unavailable studies. Transportability, whereby we can generalize statistical conclusions from one population to another, is a concept formalized by Pearl and Bareinboim [61].

While we can evaluate the treatment effects of MDR-TB in a causal way under some assumptions, we believe that some of these assumptions do not hold in this setting. For example, the time-ordering assumption states that the covariates precede the treatment to be investigated, which precede the outcome. Given that the use of other treatments can affect both the use of the treatment to be investigated and the outcome, we regard other treatments as confounders. However, we cannot assume a patient has had the treatment that we investigate ahead of all other treatments, hence we define our parameter of interest as the treatment importance rather than as a causal effect.

In Chapter 2, a comprehensive literature review is conducted. After introducing the methodology in Chapter 3, Chapter 4 illustrates how to perform an IPD-NMA with TMLE to investigate the treatment importance of MDR-TB. A simulation study to evaluate the proposed estimator is conducted and described in Chapter 5. Two methods are employed to calculate the variance of the estimators: a bootstrap that accounts for the clustered nature of the data and the sandwich estimator approach. We then discuss the limitations of this study and suggest some improvements to our analysis in Chapter 6.

CHAPTER 2
Literature Review

2.1 Multiple Drug Resistant Tuberculosis (MDR-TB)

Tuberculosis (TB) is an infectious disease caused by *mycobacterium tuberculosis bacterium (M.tb.)*. TB usually affects the lungs and is easily transmitted via the airborne route [1]. Not everyone infected develops symptoms, but TB can be fatal if treated improperly [1]. The World Health Organization (WHO) reported that as of October 2016, TB is one of the top 10 causes of death worldwide [6]. In 2015, 10.4 million people were infected with TB and 1.8 million died from the disease [6]. In most cases, patients with active TB are treated by a six-month drug regimen under the strict supervision of doctors [4]. Drug resistance emerges when anti-TB medicines are used inappropriately, through incorrect prescription by health care providers, poor quality drugs, and patients stopping treatment prematurely [6, 2]. Multidrug-resistant TB (MDR-TB) is a type of TB that is resistant to at least Isoniazid and Rifampin, the two most commonly prescribed TB drugs [56]. Individuals may develop MDR-TB when resistant strains of TB are directly transmitted from an infected person to an uninfected person. This mode of infection accounts for most of the MDR-TB cases [56]. Globally in 2014, there were an estimated 3.3% of new cases (480,000 people) and 20% of previously treated cases that developed MDR-TB, with 190,000 deaths from the illness [5].

The treatment of MDR-TB may include the use of second line reserve drugs which

are much more expensive, less effective and have more severe side effects than first line drugs [19]. Using them also requires a longer treatment duration than using first-line drugs [58]. MDR-TB is difficult to treat in some settings because medicine is limited and expensive, appropriate drugs are not always accessible, and patients experience many adverse effects from these treatments [4].

The relative effectiveness of different antibiotics for the treatment of MDR-TB has so far exclusively been evaluated through observational studies usually with small sample sizes [11]. However, a single study is usually not considered adequate to establish the effectiveness of the treatments. Combining multiple studies together can give researchers more accurate and robust results, and allows for better investigation of subgroup effects. The investigation of comparative drug effectiveness in MDR-TB has been performed through the use of both aggregated data and individual patient data sources from multiple studies [26, 76, 90]. An analysis of 9,153 MDR-TB patients in 32 populations was performed by Ahuja et al. in 2012. The study showed that later generation quinolones, ofloxacin, and ethionamide/prothionamide as part of multi-drug regimens were all associated with treatment success (with an alternative of treatment failure, disease relapse, or death) as were the combined use of four or more likely effective drugs during the initial intensive treatment phase and the use of three or more likely effective drugs during the continuation phase [11]. However, they did not adjust for the associations amongst the multiple treatments. In addition, not all antibiotics were available in every study; ignoring this could lead to potential between-study confounding of treatments (or selection bias). Furthermore, their analysis used only parametric models that have strong assumptions which may

not be true. We discuss various statistical approaches to the analysis of pooled data in the next section.

2.2 Analysis of Pooled Data

2.2.1 Meta-Analysis (MA)

Meta-analysis (MA) is a formal quantitative method to systematically assess existing studies in order to derive global conclusions about the research questions of interest [10]. Results from a MA are generally more precise than those from an individual study [10]. The main benefit of MA is to enlarge sample size, combine results from similar studies, improve statistical power, and thus obtain a consistent estimate [25].

However, MA also has its drawbacks. Statistically significant and positive results are more likely to be published than their counterparts; this is called publication bias [60, 22], and can result in biased results in the MA. In addition, the studies included in the MA may have been conducted in different time periods and have different exclusion criteria. The aggregation of such data could also produce misleading results [77]. The differences among studies can be due to multiple reasons, such as the investigation of different populations, outcome definitions, intervention definitions, study designs, and methods [66]. These differences often contribute to heterogeneity in the effect estimation [37]. The challenge is then to decide how to incorporate this heterogeneity in an appropriate way, which may depend on the aims of the synthesis and the observed directions and magnitudes of the effects of the individual studies

[40].

2.2.2 Aggregate Data Meta-analysis (AD-MA)

Traditional methods for meta-analysis use aggregate study-level data collected from study publications or the authors. Such aggregate data often consists of treatment effect estimates (odds ratios, risk differences, risk ratios, proportion and mean differences, etc.) and their associated uncertainty (standard errors or confidence intervals) [67]. The analysis of aggregate data is called Aggregate Data Meta-analysis (AD-MA). In this kind of meta-analysis, the aggregate data are synthesized using statistical methods to estimate some global effect. Generally, pooling the aggregate data together is considered as leading to more robust statistical estimation than the measures from any individual study, especially when some of the studies are too small to draw valid conclusions [51].

Nonetheless, AD-MA may lack robustness for several reasons. AD-MA depends on the quality of reporting of the original studies. Since researchers of AD-MA have no access to the individual patient data, potential for bias in the individual studies may exist. Furthermore, AD-MA has difficulty in dealing with the situation where different studies used different measurement scales or classifications.

It is important to note that the treatment effect may vary across patient subgroups. For example, a treatment may be more effective in younger patients and less so in the elderly or vice versa. Investigating subgroup effects in AD-MA is challenging as it requires that each study investigated the same subgroup [78].

Due to the above limitations and drawbacks of AD-MA, one cannot always easily

interpret the results obtained from AD-MA, especially when the study designs are not identical among the studies. Thus, using the individual patient data to perform a meta-analysis is preferable if they are accessible.

2.2.3 Individual Patient Data Meta-analysis (IPD-MA)

An alternative and increasingly popular approach is individual patient data meta-analysis (IPD-MA), whereby the raw individual-level data for each study, instead of the published estimates of the effects, are used to conduct the analysis [77, 67]. In the process of conducting IPD-MA, one must take the correlation within studies into consideration since there may be observed or unobserved similarities within studies. Therefore, studies may be considered as clusters in the analysis [67], and must be accounted for no matter the approach used.

In the two-stage approach, each study is analyzed separately, and as similarly as possible. Next, the effect estimates are analyzed via an appropriate AD-MA to produce a pooled result. In the one-stage approach, the IPD from all studies are analyzed simultaneously to obtain the pooled result, during which the study clusters are taken into account (e.g. via a mixed-effect regression model.) [67].

The two-stage approach is the most popular one because it is straightforward and easy to implement [81, 68]. The one-stage approach may be a more accurate approach because it takes both within- and between-study heterogeneity into consideration simultaneously [15]. However, it may increase the complexity of the statistical analysis and modeling [67].

Both one-stage and two-stage IPD-MA can counterbalance some of the drawbacks

of AD-MA as discussed in Section 2.2.2. The usage of IPD ensures that the data to be used in estimation include only the relevant observations that are consistent with the aim of study. In particular, they allow for the investigation of subgroup-specific treatment effects. For example, suppose one wants to compare the relative effectiveness of treatments a and b in male patients. AD-MA cannot achieve this goal if the AD failed to report the treatment effectiveness in men. Compared to AD-MA, IPD analysis allows for more flexibility in translating between the different measurement or classification scales used in different studies [78]. Moreover, analysts can enforce common models, inclusion and exclusion criteria, baseline factor adjustments, and statistical methods in the pooled or second-stage analysis which may reduce heterogeneity. Also, missing and censored data can be addressed in a uniform way by IPD-MA.

2.2.4 Individual Patient Data Network Meta-Analysis (IPD-NMA)

When more than two treatments have been used to treat the same disease and were evaluated in different studies, one can use network meta-analysis (NMA) to conduct a comprehensive analysis in order to compare the treatments [50]. One approach, called pairwise meta-analysis focuses on comparing two treatments (e.g. treatment a and b) at a time when more than two treatments are available. This contrast is then made for each pair of available treatments [89]. Alternative approaches that compare three or more treatments have recently been developed [57]. One straight-forward method is indirect comparison of pairwise meta-analysis. For example, consider a study that compares the drug effectiveness of treatments a and

b, and another study that compares the effectiveness of treatments b and c, then one can use indirect comparison of pairwise meta-analysis to analyze the comparative effectiveness of treatments a and c. A more complex extension is multiple comparison modeling (often referred to simply as NMA) which allows for the simultaneous analysis of multiple studies involving different treatments [16, 50]. Multiple comparison modeling is generally considered to be superior to both pairwise meta-analysis and indirect comparison meta-analysis.

Mixed-effect models are often employed in analyzing IPD, given that they take within- and between-study heterogeneity into consideration. NMA performed with mixed-effect models can improve the accuracy of point estimates, infers the relationships between interventions, and ranks the interventions according to their effectiveness [16, 50, 38].

As with usual MA, NMA can be performed using both AD or IPD. NMA based solely on AD is called aggregated data network meta-analysis (AD-NMA). The relative disadvantages of AD-NMA include those disadvantages associated with AD-MA (see Section 2.2.2), but also include highly contested modeling strategies [34, 24]. However, performing Individual Patient Data Network Meta-Analysis (IPD-NMA), which uses IPD to conduct NMA, may avoid these biases. Moreover, IPD-NMA can account for between-study heterogeneity and incorporate multivariate models in order to obtain more accurate results [42]. In addition, it can improve the precision of the estimation over an aggregate approach [23].

Most IPD-NMA use standard statistical approaches and do not make direct reference to causal inference. As we are interested in integrating causal inference into an

IPD-NMA as an extension of our methodological approach, we introduce fundamental causal inference concepts in the next section.

2.3 Causal Inference

This section contains a discussion of some fundamental concepts of causal inference as relevant to this thesis. We begin with the counterfactual model, then introduce the assumptions of causal inference, and explain propensity score and Inverse Probability of Treatment Weighting (IPTW).

2.3.1 Counterfactual Model

The Rubin causal model (RCM) [72] is a framework to define causal effects through the use of counterfactual (or potential) outcomes. The simplest data scenario is defined as follows: observed data $O = (X, A, Y)$, where X are covariates, A is some binary treatment and Y is the outcome of interest. In the RCM, an individual i is considered to have had the potential to have been exposed to two different interventions. Let $A_i = 0$ indicate that individual i received one of the interventions and $A_i = 1$ indicate having received the other. Then the two possible outcomes of individual i under exposure of or not are denoted as $Y_i(A_i = 1)$ and $Y_i(A_i = 1)$ respectively. These are referred to as counterfactual outcomes. We generally observe only one of these two counterfactual outcomes for a single individual. Notwithstanding, in the RCM, the causal effect for individual i is defined as the contrast between $Y_i(A_i = 1)$ and $Y_i(A_i = 0)$. The mean causal effect for the population is defined as $E(Y_i(A_i = 1) - Y_i(A_i = 0))$. If

one's interest is the causal effect in a certain group B, under some classical assumptions, which I describe later, the average causal difference in group B is defined as $E(Y_i(A_i = 1) - Y_i(A_i = 0)|B) = E(Y_i(A_i = 1)|B) - E(Y_i(A_i = 0)|B)$ [52].

Randomized Clinical Trials (RCTs) are the "gold standard" to estimate average causal effects. In an RCT, treatments are randomly allocated rather than being selected by the study participants [8]. The purpose of randomization is to ensure that the characteristics of the participants in different intervention groups are similar enough at the start of the comparison [79]. Under randomization, all known and unknown factors in different intervention groups are the same in expectation. No other study design allows researchers to balance these factors. As a result, the average causal effect is estimated as the mean difference between the outcomes of populations that received different treatments [79].

However, RCTs have some limitations. Because of their high cost compared to observational studies as well as the strict inclusion and exclusion criteria, the sample sizes of most of RCTs are relatively small. As a consequence, randomization may not perfectly balance all factors in different groups and the selected participants may not represent the whole population of interest [33]. Moreover, some ethical issues may exist in the planning of RCTs. For example, forcing some MDR-TB patients to have certain antibiotics to satisfy the RCT implementation is not ethical if theses antibiotics would not have been recommended by their clinicians given the particularities of their infection. Therefore, due to the high cost, small sample size and ethical issues, RCTs have not been used in MDR-TB. Thus, observational studies are currently considered to be the best source of information for evaluating treatment

effectiveness for MDR-TB.

2.3.2 Causal Assumptions for Observational Studies

Observational data are generally easier to implement and less expensive to collect compared to RCTs. However, causal effects are identifiable if and only if several assumptions hold in the observational study. For the simple scenario described in section 2.3.1, the following assumptions are required: (i) time-ordering, (ii) non-interference, (iii) positivity, (iv) consistency, and (v) conditional exchangeability. The first three assumptions allow for the definition of the counterfactual model and marginal causal effect, and the last two assumptions allow for identifiability of the causal effect. I discuss each assumption in turn below.

(i) The time-ordering assumption: The covariates precede treatments, which precede outcomes [86]. Logically, it requires that the three components are separated and ordered by time. As an example of where this would not hold, consider two treatments a and b that may be taken concurrently. Because the use of either treatment may not be mutually independent, the treatment a may be regarded as a confounder when we estimate treatment b's causal effect relative to not taking treatment b. It would not be valid to make the assumption that the patients necessarily received treatment a first when no time-ordering exists. If we then wanted to estimate the causal effect of treatment a relative to not taking a, we would be tempted to include b as a confounder, which would essentially require an opposing time-ordering assumption.

(ii) The non-interference assumption states that the potential outcomes of one individual are unaffected by the treatment assignment of other individuals [69, 20].

This assumption, does not mean that all of the individuals are independent, it just requires that the potential outcomes do not depend on the treatment assignment of different units.

(iii) The positivity assumption means that for all individuals, there is a positive probability of taking each treatment given the distribution of confounders, such that $Pr(A = a|X) > 0$. This assumption is not guaranteed to hold in observational studies if some patients never have access to certain treatments. Even with access, some treatments may rarely be taken by certain types of individuals. Due to this sparsity, one might estimate zero or near-zero probabilities of treatment for some individuals. These are referred to as "practical positivity violations" and may cause estimation problems.

(iv) The fourth assumption is consistency [41]. We define individual i's counterfactual outcome $Y_i(A_i)$ under exposure a as the outcome that would have been observed if individual i had received exposure a [18]. The consistency assumption states that the observed outcome for individual i, Y_i is the same as the counterfactual outcome $Y_i(A_i)$, under an intervention, when the intervention is set to the observed exposure. Mathematically, $Y_i = Y_i(A_i)$, if $A_i = a$ [18]. Rubin proposed an additional component to the consistency assumption called "no versions of treatment" [21]. The "versions of treatment" are the means to obtain the exposures. In the context of MDR-TB, let $A = a$ denote that a patient received the treatment a. This statement assumes that all the patients who receive treatment $A = a$ received it in the same way [35]. That is, there should be only one version of treatment a. If there were multiple versions, say different versions, k, of receiving the treatment (e.g. intramuscular

injection, intravenous injection, or oral etc.), then we should revise the definition of individual i's counterfactual outcome under exposure a as $Y_i(A = a, K^a = k)$, which means the counterfactual outcomes are obtained through the k^{th} version of a. Thus the consistency assumption should be $Y_i = Y_i(A = a, K^a = k)$, if $A_i = a$, no matter the version K^a [84].

(v) The last assumption is conditional exchangeability, also called no unmeasured confounding [20]. Mathematically, $Y(A) \perp\!\!\!\perp A|X$, which means that the counterfactual outcomes are independent of treatment received, conditional on the measured covariates (i.e. the treated and untreated individuals are conditionally exchangeable within strata of the measured covariates). Essentially, we must believe that if the patients in the treated group had not received the treatments, they would have had the same outcomes on average as with the patients in the untreated groups who have the same covariate values, and vice versa [35]. This is equivalent to saying that X is sufficient to control for confounding. When treatment is not randomly assigned, the reasons for receiving treatment are likely to be associated with some patient characteristics, and so the distribution of confounders will generally vary between the treated and untreated groups [35]. In making this assumption, the vital question is whether X contains all relevant variables that are distributed unequally between the treated and the untreated group. In reality, it is statistically impossible to ensure there are no unmeasured covariates, but it is required to make causal claims [32].

Based on these assumptions, I discuss some basic concepts and ways to adjust for confounding that are relevant to this thesis.

2.3.3 Propensity Score and Inverse Probability of Treatment Weighting (IPTW)

The propensity score is defined as the probability of receiving treatment a given the covariates X: $Pr(A = a|X)$ [70]. The propensity score can be used for controlling for confounding because treatment assignment is independent of the covariates conditional on the true propensity score [70]. In addition, if the conditional exchangeability assumption holds, the counterfactual outcomes are independent of the treatment assignment given the propensity score [70]. As a result, conditioning on the propensity score is sufficient to control for confounding and to identify the marginal causal contrast $E(Y_i(A_i = 1)) - E(Y_i(A_i = 0))$.

Adjusting for confounding via the propensity score may be accomplished via inverse probability of treatment weighting (IPTW). Under the causal assumptions, the average treatment effect can be written as $E\left(\frac{Y_i \mathbb{1}(A_i=a)}{Pr(A_i=a|X_i)}\right)$ because:

$$
\begin{aligned}
& E\left(\frac{Y_i \mathbb{1}(A_i = a)}{Pr(A_i = a|X_i)}\right) \\
=& E\left(\frac{Y_i(A_i) \mathbb{1}(A_i = a)}{Pr(A_i = a|X_i)}\right) && \text{by consistency} \\
=& E\left\{ E\left(\frac{Y_i(A_i) \mathbb{1}(A_i = a)}{Pr(A_i = a|X_i)}\right) \;\middle|\; X_i \right\} && \text{by the law of iterated expectation} \\
=& E\left\{ E(Y_i(A_i)|X_i) \frac{E\left(\mathbb{1}(A_i = a)|X_i\right)}{Pr(A_i = a|X_i)} \right\} && \text{by conditional exchangeability} \\
=& E\big\{ E\left(Y_i(A_i)|X_i\right) \big\} \\
=& E(Y_i(A_i)),
\end{aligned}
$$

where $\mathbb{1}()$ is the logical variable, if the conditions in the braket is true, the variable is 1, 0 otherwise.

Here, the outcomes are weighted by the estimated propensity score. By the Law of Large Numbers, the estimator $\frac{1}{n}\left(\sum_{i=1}^{n}\frac{Y_i\mathbb{1}(A_i=a)}{Pr(A_i=a|X_i)}\right) \rightarrow E\left(\frac{Y_i\mathbb{1}(A_i=a)}{Pr(A_i=a|X_i)}\right)$ [36], where n is the total sample size. Because this method emulates the data that would have been observed if the confounders had not been used to decide the probability of treatment, we often say that this method removes confounding [35].

Again, practical positivity is essential for modelling the causal effect in this way [36]. If this assumption is violated, then the above estimates may approach infinity because the estimated propensity score is the denominator. For this reason, the estimated propensity score is often truncated, though at the expense of finite sample bias.

2.4 SuperLearner

Prediction is important in statistics. One of an important components of this thesis is to predict the recovery rate of MDR-TB patients. In this section, we describe a recently developed prediction technique that we use to these ends.

SuperLearner is an ensemble learning estimation method [64]. In general, we have little knowledge of our data generating mechanism, so one might use machine learning to adaptively learn a model from the data. Moreover, there may be several possible models suggested by various experts, but none of which may be correct [64].

SuperLearner offers us a method that can combine several algorithms together by weighting each algorithm to produce a "best" prediction [86]. One can include all algorithms that may be useful in prediction in the SuperLearner library. SuperLearner

fits each algorithm and calculates its estimated risk which represents its prediction error (for example, mean squared error) using cross-validation [64]. To decide the values of weights, each algorithm is first used to obtain cross-validated predictions of the outcome. Then, the weights are estimated by running a restricted and intercept-free regression of the true outcome on the cross-validated outcome predictions, which minimizes the cross-validated risk of the ensemble prediction [64]. After obtaining the weights of the algorithms, SuperLearner combines the full-sample predictions using the weights in order to obtain the full-sample ensemble predictions [86].

The weights, which range from zero to one with the sum of the weights equal to one, represent the importance of each algorithm in the final prediction [86]. For example, if the weight for an algorithm is 0.5, then the prediction values of SuperLearner are composed of 0.5 times the prediction values of this algorithm plus the weighted prediction values of other algorithms. Hence, the researcher can be "adventurous" in selecting the algorithms since an algorithm's weight could be close to zero if it is not beneficial for prediction [64]. V-fold cross-validation is used in SuperLearner because it needs less computation than other types of cross-validation (like exhaustive cross-validation), and it preserves the desirable finite sample and asymptotic optimality properties [64].

What is the criterion for the "best" prediction? SuperLearner is a loss-based method, using a loss function to assign a measure of performance to each candidate function [85]. The loss function is selected based on the data type and the analyst's preference. Examples include the L_1 absolute error loss function, the L_2 squared error loss function and the negative log loss function for a binary outcome [85].

Therefore, the SuperLearner algorithm provides an approach for combining several algorithms together into an improved estimator, with the goal of optimizing prediction, defined as a minimization of the cross-validated error.

2.5 Semiparametric Estimation

2.5.1 Semiparametric Models in Observational Studies

In this thesis, the goal is to develop a semiparametric approach for the analysis of pooled observational data (i.e. pooled IPD). Parametric statistical approaches for data analysis involve firstly assuming a parametric model and then estimating the model parameters (using, for instance, maximum likelihood estimation). This assumes that the parametric model chosen is representative of the truth, which is untestable in most cases. If this assumption is violated, then the parameter itself has questionable meaning. In addition, the estimation will be based on an incorrect model, further confounding the results [86]. In addition, in parametric models, the evaluation of the estimation performance is based on the overall fit of the whole probability distribution rather than the probability distribution of the parameter of interest, and thus may perform a suboptimal bias-variance trade-off [86].

Semiparametric models make fewer assumptions about the data-generating mechanism by solely assuming parametric forms for subcomponents of the likelihood. Here, we use the term semiparametric to indicate both nonparametric and semiparametric. Let O represent the observed data that has probability distribution P_0. Specifically (as in our example), the dataset O contains the treatment, outcome and covariates. The true data-generating distribution P_0 is considered to be a component of the

statistical model $\mathcal{M}$, which is a collection of probability distributions. We write $O \sim P_0$, and $P_0 \in \mathcal{M}$. The parameter of interest ψ can be regarded as a function of the probability distribution P_0 so that we can write the true value of parameter of interest as $\psi_0 = \Psi(P_0)$.

In causal inference, one usually wants to know the effect of a treatment. However, in semiparametric statistics,there are parameters of interest that are not causal effects. For example, the importance of various treatments and their ranking, as defined by the residual marginal association between the treatment after adjustment for other covariates may also be of interest. Here, importance can be defined depending on the investigators' interest. In particular, measures of treatment-outcome associations where no causal assumptions are made are often referred to as variable importance measures [86]. In the previous example, the data structure is $O = (X, A, Y)$, and the importance of treatment a can then be defined as $E(E(Y_i|A_i = a, X_i) - E(Y_i|X_i))$. Under the causal inference assumptions, this parameter could be written counterfactually as $E(Y_i(A_i)) - E(Y)$ and be interpreted as the population mean difference in the outcome obtained by treating everyone with treatment $A = a$ versus no intervention. But as discussed in section 2.3.2, the time-ordering causal assumption would be violated if we have multiple treatments. The importance measures can be estimated in semiparametric models and this thesis focuses on them as a parameter of interest, denoted as ψ.

2.5.2 Asymptotic Linearity and Influence Curve

In this section, I present some properties of semiparametric estimators and a useful semiparametric tool to analyze the data.

Let P_0 and P_n denote the true and empirical distribution, respectively, of the observed data $O_1, O_2 ... O_n$, n independent and identically distributed random variables. Let ψ_n be the estimate of the parameter of interest and ψ_0 be the true value of the parameter of interest. Then, a plug-in estimator is defined as $\psi_n = \Psi(P_n)$. We represent the estimator as a function of an empirical process $(P_n f : f \in \mathcal{F})$ for some class $\mathcal{F}$ of functions of O [86]; that is, the estimator maps a "vector" of empirical means onto the estimate.

Estimators that are empirical means of a function of $O_1, O_2 \cdot O_n$ are called linear in P_n, since they are asymptotically consistent and normally distributed due to the central limit theorem. The functional delta method [87, 31] establishes the asymptotic linearity and normality of the estimator.

An estimator ψ_n is considered an asymptotically linear estimator if there exists a function $IC(O_i)$, such that $E\{IC(O_i)\} = 0$, and

$$n^{\frac{1}{2}}(\psi_n - \psi_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} IC(O_i) + o_p(1). \tag{2.1}$$

Here, $o_p(1)$ is a term that converges in probability to zero as n goes to infinity and $E(IC(O_i) \cdot IC(O_i)^T)$ is finite and nonsingular [9]. This equation arises from a first-order Taylor expansion of $\psi_n = \Psi(P_n)$ [9]. Any asymptotically linear estimator has a unique asymptotic influence curve [9].

The function $IC(O_i)$ is referred to as the influence curve, which determines the

asymptotic properties of semiparametric estimators because the asymptotic variance-covariance matrix of the estimator is determined by the influence curve. By the central limit theorem, we have that:

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} IC(O_i) \xrightarrow{\mathrm{D}} N(0, E(IC(O_i) \cdot IC(O_i)^T)),$$

and by Slutsky's theorem,

$$n^{\frac{1}{2}}(\psi_n - \psi_0) \xrightarrow{\mathrm{D}} N(0, E(IC(O_i) \cdot IC(O_i)^T)).$$

Thus, one can estimate the influence curve $IC(O_i)$ to estimate the standard error of ψ_n. In another sense, the influence curve can be regarded as a measure of the influence of each observation on the estimator, and can assist in analyzing the robustness of the estimator [86].

2.5.3 Local Efficiency and Efficient Influence Curve

An estimator ψ_n is considered to be regular if the distribution of $n^{\frac{1}{2}}(\psi_n - \psi_0)$ does not rely on the local data generating process [9]. The variance of any regular, asymptotically linear influence curve has a lower bound and the semiparametric regular, asymptotically linear estimator with variance achieving this bound is said to be locally efficient. However, we should note that this efficiency property holds only when all required model components are correctly specified.

Equation (2.1) shows us that the asymptotic behaviour of the asymptotically linear estimator is determined by its influence curve. Therefore, improvements of the efficiency of the estimator can be obtained by finding estimators associated with an

influence curve that has minimal variance. The efficient influence curve, which we denote $IC^*(O)$, is defined as the unique influence curve of the parameter ψ with minimal variance [9]. The semiparametric variance bound is a generalization of the Cramér-Rao lower bound theorem that applies to parametric models. Thus, by finding the efficient influence curve, one can derive an estimator attaining the asymptotic minimal variance in its class of regular, asymptotically linear semiparametric estimators with the same model space $\mathcal{M}$.

2.5.4 Efficient Influence Curve Estimating Equation

Let $f(O_i, \psi, P_n)$ denote some continuously differentiable function with parameter of interest ψ that satisfies $E(f(O_i, \psi, P_n)) = 0$, where P_n is the estimates of the density of O. Define $h(\psi) = \frac{1}{n}\sum_{i=1}^{n} f(O_i, \psi, P_n)$. Then the equation $h(\psi_n) = 0$ is an estimating equation, and the estimator ψ_n that satisfies this estimating equation is asymptotically linear under appropriate conditions [59, 86].

If we assign function $f(O_i, \psi, P_n)$ to be the efficient influence curve $IC^*(O, \psi, P_n)$, then $h(\psi) = \frac{1}{n}\sum IC^*(O, \psi, P_n)$, and the estimating equation $h(\psi_n) = 0$ turns to be $\frac{1}{n}\sum IC^*(O, \psi, P_n) = 0$, which is referred to as the efficient influence curve estimating equation.

An estimator that solves this estimating equation achieves the minimum variance bound when the necessary components of P_n are correctly modeled (called *local semiparametric efficiency*) [80]. The variance of such a semiparametric estimator can be estimated by $\frac{1}{n}\sum(IC^*(O, \psi, P_n))^2$ if the estimator consistently estimates the

parameter of interest.

2.6 Targeted Maximum Likelihood Estimation (TMLE)

Targeted Maximum Likelihood Estimation (TMLE) was proposed by Van der Laan and Rubin as a general framework for the estimation of statistical quantities [83].

2.6.1 Roadmap of TMLE

To approach statistical estimation according to their proposed guidelines, one may follow a "roadmap". To begin with, one must define the research question, the data structure O, the probability distribution P_0, the model space, and the parameter of interest ψ_0. Let $\mathcal{M}$ denote the statistical model space. $\mathcal{M}$ contains not only the probability distribution P_0, but also some nontestable assumptions placed on this distribution. The parameter of interest $\psi_0 = \Psi(P_0)$ can be regarded as a feature of P_0 , and the function Ψ maps P_0 onto the parameter of interest.

The second step is estimation, which contains two procedures. The first step is an initial estimation of the relevant models using an ensemble machine learning method SuperLearner which we described above. Then, TMLE updates this initial estimation, for the goal of balancing the bias and variance for the estimation of the parameter of interest (described with technical details further on).

The inference step is our final step. One can estimate the variance of the estimator empirically (i.e. using the influence curve) or with the bootstrap method. Then we may interpret the parameter of interest under the nontestable assumptions if we

believe them to hold or as an associational measure otherwise.

2.6.2 Example Algorithm

In order to illustrate the procedure of TMLE, we rephrase our simple example: we have data $O = (X, A, Y)$, where X, A and Y are the covariates, treatment and outcome respectively. We define our parameter of interest as $\psi_0 = \Psi(P_0) = E[E(Y|A = a, X)]$, which simplifies to the average causal effect of A on Y under the causal assumptions. We denote the conditional mean of Y given $A = a$ and X, $E(Y|A = a, X)$ as $\bar{Q}_0$ and the estimate of $\bar{Q}_0$ as $\bar{Q}_n$. In this case, the target parameter $\Psi(P_0)$ only depends on P_0 through a relevant (infinite-dimensional) parameter $Q_0 = Q_0(P_0)$ of P_0, so that we can also write $\psi_0 = \Psi(Q_0)$. In this simple example, the factorization of Q_0 would be $Q_0 = (\bar{Q}_0, Q_{X,0})$, where $Q_{X,0}$ is the true marginal distribution of X.

Two components must be estimated in the initial estimation for this simple example. The first is $\bar{Q}_n$, the initial estimate of $\bar{Q}_0$. The second component that needs to be estimated is the propensity score $Pr(A = a|X)$, which is denoted g_0. Its estimate is denoted g_n. Both $\bar{Q}_0$ and g_0 may be estimated via SuperLearner. The component $Q_{X,0}$ is generally estimated using its empirical distribution.

The next step is to update $\bar{Q}_n$ with g_n, which is called the fluctuation step, we let

$$logit\{\bar{Q}_n^*(\epsilon)\} = logit\{\bar{Q}_n\} + \frac{\epsilon}{g_n},$$

where ϵ is chosen such that the fluctuation minimizes a logistic regression loss function $L(Y, \bar{Q}_n^*(\epsilon))$. Let ϵ_n denote the resulting estimate of ϵ. Specifically, ϵ_n is estimated

by running a logistic regression of Y on $\frac{\mathbb{1}(A=a)}{g_n}$ with offset $logit\{\bar{Q}_n\}$, and without intercept. Then ϵ_n is the estimated coefficient of $\frac{\mathbb{1}(A=a)}{g_n}$.

The TMLE update step automatically solves the efficient influence curve estimating equation. This occurs because the logistic regression update step minimizes a corresponding loss function by solving the logistic regression score equations.

$$\frac{\partial L(Y, \bar{Q}_n^*(\epsilon))}{\partial \epsilon} \Big|_{\epsilon=0} = 0 = \frac{1}{n} \sum IC^*(O). \tag{2.2}$$

Then $\bar{Q}_n^*(\epsilon_n)$ is our updated TMLE estimator:

$$\bar{Q}_n^* = \bar{Q}_n^*(\epsilon_n) = expit\left(logit\{\bar{Q}_n\} + \frac{\epsilon_n}{g_n}\right),$$

Finally, our estimator is obtained by setting $\psi_n = \frac{1}{n}\sum \bar{Q}_n^*$.

To calculate its associated variance as we discussed in section 2.5.4, we may use the efficient influence curve. The efficient influence curve for this simple example is [71]

$$\begin{aligned} IC^*(O) &= \frac{\mathbb{1}(A=a)}{Pr(A=a|X)}\{Y - E(Y|A=a,X)\} + E(Y|A=a,X) - \psi_n \\ &= \frac{\mathbb{1}(A=a)}{g_n}\{Y - \bar{Q}_n^*\} + \bar{Q}_n^* - \psi_n. \end{aligned}$$

Thus, the uncertainty about estimating ϵ is accounted for in the $IC^*(O)$ estimation by including the $\bar{Q}_n^*$ in $IC^*(O)$. Under regularity conditions, the variance is estimated by $\frac{1}{n}\sum(IC^*(O))^2$ if the TMLE consistently estimates ψ_0.

2.6.3 Properties

SuperLeaner is expected to more closely approximate $\bar{Q}_0$ than the parametric models. By combining TMLE updating procedure with nonparametric approximation of $\bar{Q}_0$ and g_0, we aim to get an estimator that closely approximates the parameter of interest.

TMLE is a substitution or "plug-in" estimator, and can therefore constrain the estimator to the target parameter's model space. Substitution estimators are more robust to outliers and sparsity which may cause an estimator to lie outside of the possible range of the parameter of interest [83].

TMLE estimators are well-defined; they are constructed by minimizing the loss over the statistical model, and thus have only one solution in the model space. In addition, the substitution estimator respects the global constraints implied by the statistical model and targeted parameter mapping. Furthermore, by incorporating SuperLearner, TMLE can avoid parametric assumptions.

The TMLE we described is a doubly robust as well as locally efficient estimator. An estimator is doubly robust if when either at least one of the two component models is correctly specified (consistent), the estimator is consistent for ψ_0. In the simple example, if the model for either g_0 or $\bar{Q}_0$ is correctly specified, then the TMLE is consistent. Local efficiency means that the estimator attains the minimal variance bound when both g_0 and $\bar{Q}_0$ are correctly estimated as we discussed in Sections 2.5.3 and 2.5.4. This is a result of the estimator solving the efficient influence curve estimating equation as shown in equation (2.2) in section 2.6.2. When correctly specified, the TMLE estimator is therefore the estimator with minimal variance in

its class of regular, asymptotically linear semiparametric estimators. In particular, IPTW (as defined in section 2.3.3) is less efficient than TMLE as it is in the same class and has a larger asymptotic variance

2.7 Transportability

Generalization in statistics is the way to make statistical conclusions about a much broader population than the sample actually represents [61]. It is not an unfamiliar concept in statistics because the data we obtain is always a sample of the whole population and we desire to know the relationships of interest in a more general population.

Transporting in this context is defined as transferring the statistical associations measured in one population to another population [61]. Generalization is therefore a type of transporting. If we deem that there are some similarities between the two study populations, the one in which the full data can be observed, and the target population where inference is of interest, then transporting may be used [13].

Causal effects and relationships between two variables are examples of statistical relationships that can be transported. If one wants to obtain reliable results in a target population, it is necessary to adjust for covariates during the transporting process when there is heterogeneity across populations. Because these covariates can modify the effect of the treatment and the outcome in both the study population and the target population, they need to be adjusted for if we want to transport the causal effect across the two populations [74].

A formal approach for analyzing transportability did not appear until 2011. Pearl

and Bareinboim [61] established "licensing assumptions", which are formal conditions under which the transport of results across diverse environments is acceptable. Multiple types of datasets can be addressed using transportability, such as data from observational studies, RCTs conducted on a representative sample of the population, nonrepresentative randomized clinical trials (with selection bias) and dissimilar populations [13]. Each of these require different assumptions to implement transport. In this thesis, we focus on the fusion of multiple observational studies.

Pearl and Bareinboim defined observational transportability as so: Given two populations, study population and target population–with their own probability distributions, their causal relationships and a set of assumptions—a statistical relationship is said to be *observationally transportable* from one population to another if the statistical relationship of the target population is identifiable from the above information [61]. How to conduct transportability analysis from multiple populations has been discussed [62, 14, 12].

Rudolph and Van der Laan developed a TMLE for transporting intervention effets from one population to a target population [74]. They considered various definitions of average treatment effects as their parameters of interest. One of the parameters of interest was the $E[\{E(Y|S=1,X,A=1)-E(Y|S=1,X,A=0)\}|S=0]$, where Y, S, X, A are outcome, population indicator (target and study), covariates and treatment respectively. The parameter represents the average treatment effect in the target population ($S=0$). Their simulation study showed that the proposed TMLE estimators have the double robustness property in this setting as well [74]. In addition, the bias of TMLE estimators was often smaller than that of the competing

approaches except when all the models were misspecified [74]. They showed these estimators are applicable to observational studies. This work closed an important gap by demonstrating an estimation approach for the transportability problem proposed by Pearl and Bareinboim. [61, 74].

CHAPTER 3
Objectives

In general, this work aims to provide a doubly robust and locally efficient estimation method for estimating the treatment importance of multiple antibiotics for treating MDR-TB. In particularly, the objectives of this thesis are threefold: theoretical, methodological and application-based.

Objective 1

The theoretical objective is to derive the assumptions necessary for the identifiability of the parameters of interest (treatment importance) in different scenarios: 1) We demonstrate the identifiability of the treatment importance parameter under the transportability framework in the context of the fusion of multiple observational datasets with multiple treatments where any given treatment may not be available in all studies. 2) In addition, we prove that the average treatment effect under the causal framework is identifiable in under more stringent assumptions.

Objective 2

The methodological objective of this thesis is to develop a method for estimating treatment importance using TMLE with transportability models in an individual patient data network meta-analysis context with the existence of multiple treatments and multi-drug resistance. The variance of the estimator is estimated using

two different methods (clustered bootstrap and influence curve sandwich estimation). Simulation studies are conducted to show the validity of the estimator and demonstrate its double robustness statistical properties.

Objective 3

The application in this thesis considers the estimation of treatment importance for 15 antibiotics used to treat MDR-TB. Therefore, the conclusions may provide relevant information for clinicians and guidance for further research on MDR-TB.

CHAPTER 4
Study Summary

The data used in this thesis are MDR-TB IPD, which were collected from world-wide MDR-TB studies after 1970. In this section, we first describe how the data were collected and the relevant work that has been done previously. Then we provide a brief description of the data.

4.1 Data Extraction

A detailed description of how the data were collected is presented in Ahuja et al. (2012)[11]. Studies were identified from three previous systematic reviews [**?**, 58, 44]. We describe the objectives, inclusion/exclusion criteria, methods and results of these systematic reviews and then illustrate the contribution of the work done by Ahuja. et al.

4.1.1 Three Systematic Reviews

Orenstein et al. conducted a systematic review in 2009 [58] to investigate how treatment success was related to treatment regimen, study methodology, and patient population.

They included the studies with: 1) patients with MDR-TB; 2) treatment outcome definitions specified by mycobacterial culture endpoints; 3) clearly defined treatment protocols including second-line drugs; and 4) outcomes reported according to WHO

classifications of success, failure, default, and death. Studies in which all patients had extensively drug-resistant tuberculosis (XDR-TB) were excluded.

The systematic review included 34 clinical reports (with 8500 patients) and the subsequent analysis found that treatment duration longer than 18 months and directly observed therapy were associated with a greater proportion of successfully treated patients.

Another systematic review was performed by Johnston et al. in 2009 [44]. They estimated the pooled treatment outcomes of MDR-TB using 36 studies and identified risk factors associated with poor outcomes in patients with MDR-TB.

The inclusion criteria of this systematic review were: original studies; reported in English; reported treatment outcomes in a population of adults, culture-confirmed MDR-TB patients; and reported outcomes presented in a format allowing for comparison with other studies. The exclusion criteria were: exclusive surgical series; and exclusive use of first-line therapy in the treatment protocol.

The outcomes were defined in the same way following the WHO guidelines [43, 27]. This study included 36 studies encompassing 7575 patients. In addition to presenting the proportion with the defined treatment outcome, they also concluded that factors associated with successful outcome were surgical intervention, no previous treatment, and fluoroquinolone use. Factors associated with a greater likelihood of a failed treatment outcome were: male gender, alcohol abuse, low BMI, smear positivity at diagnosis, fluoroquinolone resistance and the presence of an XDR resistance pattern.

The last systematic review was conducted by Akcakir in 2010 [**?**]. The objectives of

this study were: 1) to use MA to estimate the rates of treatment outcomes in MDR-TB; 2) to explore the correlation between these outcomes and population, disease, treatment characteristics and health setting.

The following criteria were applied to include: 1) studies that reported treatment outcomes (with definition of success, failure, relapse, death or default) for MDR-TB patients; 2) studies that contained at least 25 subjects; 3) studies that were published in peer-reviewed journals in English, French, or Spanish and 4) studies that were published after 1970. The exclusion criteria were: 1) studies reporting exclusively on extensively drug-resistant tuberculosis and 2) studies reporting exclusively on extra-pulmonary MDR-TB.

As a result, 74 articles which contained 64 unique cohorts (with 8046 individuals) were analyzed in this study. The results showed the overall pooled rate of cumulative treatment outcome and how those variables correlated with the treatment outcomes.

4.1.2 Previous Work

The Collaborative Group for Meta-Analysis of Individual Patient Data in MDR-TB, founded by Dr. Dick Menzies, conducted an IPD-MA using the studies identified in the above three systematic reviews in 2012. They aimed to assess how the type, number and duration of treatments were associated with the treatment outcomes.

They set some additional specific criteria for this meta-analysis when the three systematic reviews were screened: the authors of the studies were still reachable; at least 25 patients treated for MDR-TB were included in the cohort; and, at least treatment success (defined by WHO & Laserson Criteria [43, 27]) was reported.

They identified 67 unique cohorts from the three systematic reviews. Authors were contacted to participate in the IPD-MA. Authors who agreed to participate provided both study-level information and patient level information.

Applying the inclusion/exclusion criteria, 2 authors refused to participate the study and 34 studies were excluded for the above exclusion criteria. 32 cohorts were left with 9153 patients. They estimated the odds of treatment success as opposed to one of the three outcomes: 1) treatment failure or relapse; 2) treatment failure, relapse or death; and 3) treatment failure, relapse, death or default. Random effects multi-variable logistic regression via both penalized quasi-likelihood and adaptive quadrature were used to estimate the adjusted odds and 95% CIs of treatment success associated with different treatment covariates. Five covariates were included to adjust the treatment effects: age, gender, HIV co-infection, extent of disease, and past history of TB treatment. Missing values of these covariates were imputed by the means of patients at the same center with non-missing information.

Specifically, they classified amoxicillin-clavulanate, macrolides (azithromycin, roxithromycin, and clarithromycin), clofazimine, thiacetazone, imipenem, linezolid, high dose isoniazid, and thioridazine as group 5 drugs. We have the same definition of group 5 drugs in this thesis.

Results showed that the use of certain drugs was associated with treatment success compared to failure or relapse. Similarly, they also investigated which groups of drugs were associated with treatment success compared to other alternative combination of outcomes. In addition, in the initial intensive phase of treatment, the odds of success were greater with the use of four or more drugs compared to the use of

three or fewer treatments. In the continuation phase, the use of three or more drugs was associated with a higher odds of success compared to use of two or fewer likely effective drugs. Furthermore, longer duration of the initial intensive phase up to a duration of 7.0 to 8.4 months and longer total duration of therapy up to 24.6–27.5 months were both associated with a greater odds of treatment success.

This study is the largest combined analysis of treatment of MDR-TB, and the first IPD-MA of treatment outcomes in drug resistant TB. Using this data set, the Collaborative Group for Meta-Analysis of Individual Patient Data in MDR-TB investigated further, and they also concluded that there is no improvement in treatment success among patients taking Group 5 drugs [28]. Standard multivariable and propensity-score based methods of adjusting for individual patient covariates yielded similar effect estimates [29]. In XDR-TB patients, regimens containing more drugs than those recommended in MDR-TB but given for a similar duration were associated with the highest odds of success [26].

The collaborative group provided estimates of the impact of specific drugs, number of drugs, and duration of treatment on clinical outcomes of patients with pulmonary MDR-TB and assisted in the development of WHO clinical guidelines.

4.2 Data Structure

In our analysis, we use the same data as Ahuja et al. However, one of the 32 studies was not available because the author is not reachable. Thus, we have 31 cohorts left with a total of 9290 patients. As shown in Figure 4-1, with the same

inclusion/exclusion criteria, they identified 93 studies from the three systematic reviews, excluded 26 studies which contained the same or overlapping cohorts and identified 67 unique cohorts. They excluded cohorts from the studies for the following issues: no author response, no longer had access to data; inadequate outcome data; refusals; no response following an initial contact; no data on drug sensitivity testing; agreed to forward data-but data was never sent; cohort with less than 25 patients. In this study, 31 cohorts were included and within these cohorts, patients with XDR-TB, extra-pulmonary TB and the patients without treatment information were excluded. Consequently, 9290 patients were eligible to be included in this study. With the data, we have the following information: study-level information: the publication year of each study, and the income group of the country of each study, individual –level information: patients' age, sex, HIV status, acid fast bacilli (AFB) smear status at start of MDR-TB treatment regimen, past history of TB, cavitation status of a patient, if the patients recover or not after receiving the treatment, 15 treatment usage indicator, and resistance status for each treatment.

Some missing data exist in the individual –level information. We deleted the missingness in age and sex because they only accounted for 0.065% and 0.301% respectively of 9290 subjects.

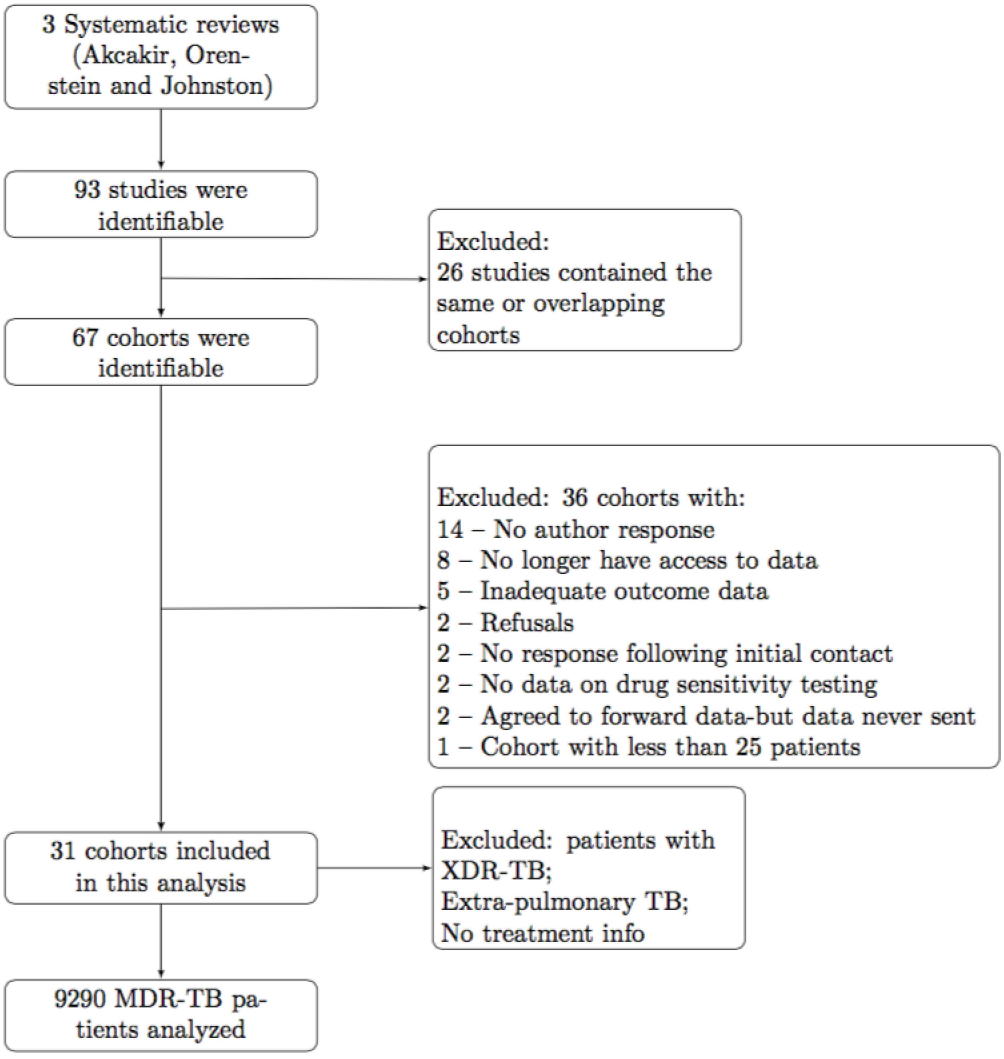

Figure 4–1: Data collection.

CHAPTER 5
Methods

With the data described in Section 4.2, we aimed to rank the 15 treatments by their importance. However, there were some major challenges that complicated the design of the methods. One of them was that we have multiple treatments, so some causal assumptions might be violated if we analyze the data in a traditional causal way. In addition, patients may be resistant to different antibiotics, which made the estimation more difficult because it would not make sense to investigate a treatment using the data of the patients who were resistant to the treatment. Furthermore, we have multiple studies, and not all of the treatments were observed in each study. Moreover, we wanted to obtain a robust and efficient estimator, so the TMLE algorithm was adopted in our analysis.

In this chapter, we first introduce the data structure, then state the assumptions needed for estimation, along with the parameter of interest and its identifiability. Then we also describe a related causal parameter of interest with required identifiability assumptions. Next, we explain the models and algorithms used in our methods, including the methods used to estimate the estimation variance. Finally, a detailed illustration of how to conduct the analysis using the data is given.

5.1 Data Structure

As we described in section 4.2, in our 31 studies, each patient i of study j has a binary outcome Y_{ij} (Y_{ij}=1 if the patient recovered after receiving the treatments and 0 otherwise), indicators for the use of 15 treatments A_{ij}^k (A_{ij}^k=1 if the patient used the treatment k), where k=1...15 is the index of treatment, 2 study-level covariates V_{ij}, 6 individual-level covariates W_{ij}, and resistance status R_{ij}^k (R_{ij}^k=1 if the patient i was resistant to medication k). In addition, we introduced the treatment availability D_{ij}^k. We say that the treatment k is available to individual i in study j if at least one of the patients in study j used the treatment k and then D_{ij}^k=1, otherwise D_{ij}^k=0. In terms of missing data, there are 2.8% missing values in the outcome Y_{ij}. We defined an indicator for missing outcomes C_{ij} such that if the outcome was missing (i.e., Y_{ij} =NA), we assign C_{ij} as 1, otherwise 0.

Thus we define our data structure as

$$O_{ij} = (Y_{ij}, V_{ij}, W_{ij}, , C_{ij}, \{A_{ij}^k, R_{ij}^k, D_{ij}^k\}; k = 1...15); i \in S_j, j = 1...31,$$

where S_j is the set of indices of subjects in study j. We assume the clusters are independent to each other but there may exist some correlation pattern within cluster. Furthermore, we define the covariates needed to be adjusted as confounders when estimating the variable importance of treatment k:

$$X_{ij}^k = (V_{ij}, W_{ij}, A_{ij}^{k^*}; k^* \neq k).$$

We include in this set all study-level as well as individual-level covariates. We also include $\{A_{ij}^{k^*}; k^* \neq k\}$ in this set as we believe that other treatment use is correlated

with the use of the treatment under investigation and also affects the outcome. Note that not all of studies include all the treatments. We therefore have no information about certain treatments' effects in some populations. In order to produce an estimate of a generalized treatment effect that can be applied to a global population, we must assume a type of transportability. In particular, we assume that the outcome models fit with only studies where a treatment was available are applicable to all studies. We say "available study" to refer to studies where treatment k was observed and "unavailable study" to denote otherwise. Thus, we make a transportability assumption to extrapolate the estimate from the available studies to unavailable studies. We define the counterfactual outcome and counterfactual treatment exposure as $Y_{ij}(D_{ij}^k = 1)$ and $A_{ij}^k(D_{ij}^k = 1)$, respectively. These are the outcome and treatment exposure of patient i in population j if treatment k had been available in population j.

5.2 Assumptions

The assumptions below are made for the identifiability and interpretation of our parameter of interest. We say a parameter is identifiable if the parameter can be derived from the observed data via available models [49]. Most of assumptions below are similar to the traditional assumptions but not exactly identical. These assumptions would allow us to interpret our parameter of interest in a non-causal way.

A1. Time ordering: Treatment use and covariates precede outcomes in time. More specifically, the data-generating distribution conforms to a specific non-parametric structural equation model.

A2. Positivity: In this context, we must make three positivity assumptions:

(a) There would be a positive probability of receiving the treatment k, had the treatment k been available to study j, over the distribution of confounders among the patients who are not known to be resistant for the treatment k, i.e., $Pr(A_{ij}^k(D_{ij}^k = 1) = 1|X_{ij}^k, R_{ij}^k = 0) > 0$, $\forall X_{ij}^k$ s.t. $Pr(X_{ij}^k, R_{ij}^k = 0) > 0$.

(b) There is a positive probability for every study (given its characteristics V_{ij}) to have had access to treatment k, i.e., $Pr(D_{ij}^k = 1|V_{ij}) > 0$, $\forall V_{ij}$ s.t. $Pr(V_{ij}) > 0$.

The idea is that this probability will be used to balance out the study-level covariates so that we remove the study-level confounding.

(c) There is a positive probability of observing all outcomes given the confounders among the patients who are not known to be resistant for and use the treatment k, i.e., $Pr(C_{ij} = 0|A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) > 0$, $\forall X_{ij}^k$ s.t. $Pr(A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) > 0$.

A3. Consistency: We estimate treatment availability for the individual based on the study-wide data. We must assume that the treatment and outcome of the individual - had the treatment been available to the individual - are the same when we observe that the treatment was in fact available in their study. In other words:

(a) When treatment k is observed to be available in study j, the counterfactual outcome is exactly the same with what we would have observed had treatment k been available to the individual. i.e., $Y_{ij}(D_{ij}^k = 1) = Y_{ij}$ when $D_{ij}^k = 1$.

(b) When treatment k is observed to be available in study j, the counterfactual treatment received by the individual is exactly the same as what we would have observed if treatment k had been available to the individual. I.e.,$A_{ij}^k(D_{ij}^k = 1) = A_{ij}^k$ when $D_{ij}^k = 1$.

The second consistency assumption may fail when not all patients in the same study had access to the same treatments (due to the studies being conducted across multiple centers or over large periods of time). Then, it is possible that in studies where treatment k was observed, a subject who did not, in fact, received this treatment. These situations would have had they personally had access to the treatment k. This issue may also invalidate A3(a) in a similar way.

A4. Exchangeability: The counterfactual outcomes are independent of treatment availability and censoring of outcome conditional on confounders in the subset of $\{R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1\}$. It is the same to say the observed treatment availability and censoring of outcome provide no information about the counterfactual outcomes conditional on confounders in the subset of $\{R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1\}$. I.e.,

$$Y_{ij}(D_{ij}^k = 1) \perp\!\!\!\perp \{D_{ij}^k, C_{ij}\} | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1,$$

and therefore

$$\begin{aligned} &Pr(Y_{ij}(D_{ij}^k = 1) = 1 | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1) \\ =&Pr(Y_{ij}(D_{ij}^k = 1) = 1 | D_{ij}^k = 1, C_{ij} = 0, X_{ij}^k, R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1). \end{aligned}$$

This assumption may fail if the analyst did not account for a covariate which was correlated with the outcome, and the treatment availability or the censoring of the outcome simultaneously. For example, if the social economics status of the patients was not included in the confounders, then it should be considered as an unmeasured confounder and this assumption violated.

A5. The treatment availability is independent of confounders and resistant status given the study-level covariates.:$D_{ij}^k \perp\!\!\!\perp X_{ij}^k, R_{ij}^k = 0 | V_{ij}$, such that,

$$Pr(D_{ij}^k = 1 | X_{ij}^k, R_{ij}^k = 0) = Pr(D_{ij}^k = 1 | V_{ij}).$$

This assumption says the study-level covariates are sufficient to explain the treatment availability. We thus need to include all study-level covariates which are correlated with the treatment availability.

A6. Independence: Studies are independent between each other and individuals in a given study have an arbitrary dependence pattern.

A7. Missing at random: The missingness of outcomes is independent of the unobserved outcome values, conditional on covariates and treatment use among the patients who were not known to be resistant to the treatment k. I.e, $C_{ij} \perp\!\!\!\perp Y_{ij} | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k$.

5.3 Parameter of Interest

We incorporate a definition of treatment importance as our parameter of interest. The treatment importance of treatment k is the average adjusted difference among the patients who were not known to be resistant to treatment k if all patients used the treatment k and the overall recovery rate.

$$\psi^k = E\left[E\left(Y_{ij}(D_{ij}^k = 1) | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1\right) - E\left(Y_{ij} | X_{ij}^k, A_{ij}^k, R_{ij}^k = 0\right) | R_{ij}^k = 0\right].$$

Under the assumptions above, we rewrite our parameter of interest as:

$$\begin{aligned}
\psi^k =& E\big[E\big(Y_{ij}(D_{ij}^k = 1) | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1\big) \\
& - E\big(Y_{ij} | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k\big) | R_{ij}^k = 0\big] \\
=& E\big[E\big(Y_{ij}(D_{ij}^k = 1) | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k(D_{ij}^k = 1) = 1, D_{ij}^k = 1, C_{ij} = 0\big) \\
& - E\big(Y_{ij} | X_{ij}^k, , R_{ij}^k = 0, A_{ij}^k\big) | R_{ij}^k = 0\big] \qquad \text{by A4}
\end{aligned}$$

$$
\begin{aligned}
=&E\big[E\big(Y_{ij}(D_{ij}^k=1)|X_{ij}^k,R_{ij}^k=0,A_{ij}^k(D_{ij}^k=1)=1,D_{ij}^k=1,C_{ij}=0\big)\\
&-E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k,C_{ij}=0\big)|R_{ij}^k=0\big] && \text{by A7}\\
=&E\big[E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k=1,D_{ij}^k=1,C_{ij}=0\big)\\
&-E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k,C_{ij}=0\big)|R_{ij}^k=0\big] && \text{by A3}\\
=&E\big[E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k=1,C_{ij}=0\big)\\
&-E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k,C_{ij}=0\big)|R_{ij}^k=0\big]\\
&(\text{since } A_{ij}^k=1 \text{ implies } D_{ij}^k=1)\\
=&E\big[E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k=1,C_{ij}=0\big)|R_{ij}^k=0\big]\\
-&E\big[E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k,C_{ij}=0\big)|R_{ij}^k=0\big]\\
=&\tau^k-\mu^k.
\end{aligned}
$$

Because $E\left[E\left(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k=1,C_{ij}=0\right)|R_{ij}^k=0\right]$ and $E\big[E\big(Y_{ij}|X_{ij}^k,R_{ij}^k=0,A_{ij}^k,C_{ij}=0\big)|R_{ij}^k=0\big]$ are both estimable from the observed data, ψ^k is identifiable. This definition of treatment importance is analogous to the coefficient of the treatment of interest in a linear regression that also adjusts for the other treatments and covariates. However, we incorporate censoring and resistance information into the definition and also do not prespecify an estimation method.

This parameter could also roughly be interpreted (if the additional causal assumptions hold, see Section 5.4) as treatment k's contribution to the average recovery rate in the population of individual who are not resistant to the treatment k, had all of these patients been forced to use this treatment on top of all other treatments they were initially given. If a treatment possesses a large treatment importance, it

is likely an effective add-on medication. The treatment therefore should be recommended more frequently than an antibiotic whose treatment importance is less. For instance, assume a clinician has two options: to prescribe antibiotics a and b together or b and c together to a patient, and that the clinician believes that these two combinations would have a similar effect. But if treatment a has a larger treatment importance than treatment c, then the clinician is suggested to prescribe treatments a and b together rather than the combination of b and c. Further, it is possible that a treatment has a negative treatment importance. Suppose a treatment has strong side effects or drug interactions. When all patients take this antibiotic possibly in conjunction with other treatments, the average recovery rate may decrease. Since this treatment may still be useful in particular circumstances, it may be inappropriate to say that this treatment is not important at all; rather, we would say this treatment should be carefully used.

5.4 Causal Assumptions & Interpretation

We can also interpret our parameter of interest as the attributed recovery rate (i.e., add-on effect) in a fully causal framework. However, the assumptions needed are more stringent:

B1. Time ordering: X_{ij}^k precedes A_{ij}^k and A_{ij}^k precedes Y_{ij} in time: In particular, when investigating treatment k, $A_{ij}^{k^*}$ precedes A_{ij}^k, $k \neq k^*$, in time. Because we regarded $A_{ij}^{k^*}$ as confounders which have an effect on A_{ij}^k and Y_{ij} at the same time, they need to be adjusted for in the analysis.

B2. Non-interference: For each individual, the potential outcome depends only upon whether or not he or she received the treatment and is independent of all other individuals' treatments, such that $Y_{ij}(A_{ij}^k = 1) \perp\!\!\!\perp A_{i^*j}^k$; where $i^* \neq i$.

B3. Positivity:

(a) There is a positive probability of receiving each of the treatments over the distribution of X_{ij}^k among the patients who are not known to be resistant to this treatment. i.e.,$Pr(A_{ij}^k = 1|X_{ij}^k, R_{ij}^k = 0, D_{ij}^k = 1) > 0$, $\forall X_{ij}^k$ s.t. $Pr(X_{ij}^k, R_{ij}^k = 0, D_{ij}^k = 1) > 0$.

(b) There is a positive possibility of observing available studies over the distribution of study-level covariates, i.e., $Pr(D_{ij}^k = 1|V_{ij} = 0) > 0$, $\forall V_{ij}$ s.t. $Pr(V_{ij}) > 0$.

(c) There is a positive possibility of observing non-misssing outcomes over the distribution of confounders among the patients who are not known to be resistant for and use the treatment k, i.e., $Pr(C_{ij} = 0|A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) > 0$, $\forall X_{ij}^k$ s.t. $Pr(A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) > 0$.

Note that, because

$$Pr(A_{ij}^k = 1|X_{ij}^k, R_{ij}^k = 0, D_{ij}^k = 1) > 0,$$

then $Pr(A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) = Pr(A_{ij}^k = 1, D_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0)$ is non-zero for $\forall X_{ij}^k$, s.t.

$$Pr(X_{ij}^k, R_{ij}^k = 0, D_{ij}^k = 1) > 0.$$

Also note that

$$\begin{aligned}
&Pr(A_{ij}^k = 1, C_{ij} = 0|X_{ij}^k, R_{ij}^k = 0) \\
=&Pr(A_{ij}^k = 1|D_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) \cdot Pr(D_{ij}^k = 1|V_{ij}) \cdot Pr(C_{ij} = 0|A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0),
\end{aligned}$$

so we assume that all these three probabilities are positive.

B4. Consistency: An individual's potential outcome under the observed exposure is precisely the observed outcome. i.e., $Y_{ij}(A_{ij}^k = 1) = Y_{ij}$ when $A_{ij}^k = 1$.

B5. Exchangeability: Conditional on all measured confounders among the patients who were not known to be resistant to treatment k, the patients who received the treatment k and had missing outcomes have the same distributions of potential outcomes $Y_{ij}(A_{ij}^k = 1)$ with those that did not, which means there is no unmeasured confounding. Equivalently, A_{ij}^k and C_{ij} is conditionally independent of the counterfactual outcome, given X_{ij}^k and $R_{ij}^k = 0$. I.e., $Y_{ij}(A_{ij}^k = 1) \perp\!\!\!\perp \{A_{ij}^k, C_{ij}\} | X_{ij}^k, R_{ij}^k = 0$.

B6. Independence: See Assumption A6.

B7. Missing at random: See Assumption A7.

Note that if a patient uses the treatment k, the treatment is automatically available in the study for the patient by the definition of treatment availability, i.e., $Pr(A_{ij}^k = 1) = Pr(A_{ij}^k = 1, D_{ij}^k = 1)$. Thus, the counterfactual outcome in causal setting $Y_{ij}(A_{ij}^k = 1) = Y_{ij}(A_{ij}^k = 1, D_{ij}^k = 1)$.

Under the above assumptions, the attributed recovery rate ψ_C^k is:

$$
\begin{aligned}
\psi_C^k =& E\left(Y_{ij}(A_{ij}^k = 1) | R_{ij}^k = 0\right) - E\left(Y_{ij} | R_{ij}^k = 0\right) \\
=& E\left[E\left(Y_{ij}(A_{ij}^k = 1) | X_{ij}^k, R_{ij}^k = 0\right) | R_{ij}^k = 0\right] - E\left[E\left(Y_{ij} | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k\right) | R_{ij}^k = 0\right] \\
& \text{(By the law of iterated expectations)} \\
=& E\left[E\left(Y_{ij}(A_{ij}^k = 1) | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k = 1, C_{ij} = 0\right) | R_{ij}^k = 0\right] \\
& - E\left[E\left(Y_{ij} | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k\right) | R_{ij}^k = 0\right] \qquad \text{by B5} \\
=& E\left[E\left(Y_{ij}(A_{ij}^k = 1) | X_{ij}^k, R_{ij}^k = 0, A_{ij}^k = 1, C_{ij} = 0\right) | R_{ij}^k = 0\right]
\end{aligned}
$$

$$
\begin{aligned}
&-E\left[E\left(Y_{ij}|X_{ij}^k, R_{ij}^k = 0, A_{ij}^k, C_{ij} = 0\right)|R_{ij}^k = 0\right] && \text{by B7}\\
&=E\left[E\left(Y_{ij}|X_{ij}^k, R_{ij}^k = 0, A_{ij}^k = 1, C_{ij} = 0\right)|R_{ij}^k = 0\right]\\
&-E\left[E\left(Y_{ij}|X_{ij}^k, R_{ij}^k = 0, A_{ij}^k, C_{ij} = 0\right)|R_{ij}^k = 0\right] && \text{by B4}\\
&=\tau_C^k - \mu_C^k.
\end{aligned}
$$

Therefore, the causal parameter of interest is identifiable under the assumptions B. However, as we discussed in section 2.3.2, in our application, it is not credible that the treatments satisfy the time-ordering assumption. We have multiple treatments that are not mutually independent. One treatment use may be regarded as a confounder when we estimate another's treatment effect. In reality, it would not be valid to assume that the patient received treatments in an ordered way. Therefore, rather than interpreting our point estimates as the attributed recovery rate, we interpreted them as the treatment importance, as described in section 5.3

Both parameter definitions lead to the same estimation procedures but different interpretations.

5.5 Models & Algorithm

In this section, we first focus on estimating τ^k using TMLE. An almost identical procedure was applied to estimate μ^k (see Section 5.7 for details). (τ^k and μ^k are the same as we defined in Section 5.3) As we discussed in section 2.6.2, to proceed with the TMLE algorithm, there are two quantities needed to be modeled: Q (conditional expectation of the outcome) and g (propensity score). We describe how to adapt these quantities in order to estimate τ^k. We also show that either Q or g can

separately be used to estimate τ^k.

5.5.1 Outcome Model (Q Model)

We can use the Q model to construct a consistent estimator for τ^k with the aid of transportability and G-computation [41].

Our Q model for τ^k is defined as:

$$Q_{ij}^{\tau k} = Pr(Y_{ij}(D_{ij}^k = 1) = 1 | A_{ij}^k (D_{ij}^k = 1) = 1, X_{ij}^k, R_{ij}^k = 0),$$

which is the probability of having a successful counterfactual outcome conditional on the counterfactual treatment exposure and covariates among the patients who were not known to be resistant to treatment k when all the studies are available. However, treatment k is not observed in all studies, therefore, we use transportability (see Section 2.7) to estimate this probability.

Under the assumptions made above, we write $Q_{ij}^{\tau k}$ as:

$$\begin{aligned}
Q_{ij}^{\tau k} &= Pr(Y_{ij}(D_{ij}^k = 1) = 1 | A_{ij}^k (D_{ij}^k = 1) = 1, X_{ij}^k, R_{ij}^k = 0) \\
&= Pr(Y_{ij}(D_{ij}^k = 1) = 1 | A_{ij}^k (D_{ij}^k = 1) = 1, X_{ij}^k, R_{ij}^k = 0, D_{ij}^k = 1, C_{ij} = 0) \quad \text{by A4} \\
&= Pr(Y_{ij} = 1 | A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0, D_{ij}^k = 1, C_{ij} = 0) \quad \text{by A3} \\
&= Pr(Y_{ij} = 1 | A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0, C_{ij} = 0) \\
&\text{(since } A_{ij}^k = 1 \text{ implies } D_{ij}^k = 1)
\end{aligned}$$

This last probability is estimable taking only patients who were not known to be resistant to treatment k for $Q_{ij}^{\tau k}$. We can then fit a model for $Q_{ij}^{\tau k}$ to estimate the

quantity:

$$\tau^k = E[E(Y_{ij} = 1|A^k_{ij} = 1, X^k_{ij}, R^k_{ij} = 0, C_{ij} = 0)|R^k_{ij} = 0].$$

We can use the data in the subset of subjects with $\{A^k_{ij} = 1, R^k_{ij} = 0, C_{ij} = 0\}$ to construct the model $Q^{\tau k}_{ij}$ and then predict on the patients in the subset of $\{R^k_{ij} = 0\}$. Transportability is thus used to transport the estimation from available studies to unavailable ones. One of the reasons that we use transportability is that we believe patients in these two subsets share some similarities. Had we not believed this to be the case, we could not attempt to generalize the treatment importance over all study populations.

We use SuperLearner (see Section 2.4) to estimate $Q^{\tau k}_{ij} = Pr(Y_{ij} = 1|A^k_{ij} = 1, X^k_{ij}, R^k_{ij} = 0, C_{ij} = 0)$ using the data in the subset $\{A^k_{ij} = 1, R^k_{ij} = 0, C_{ij} = 0\}$, and then predict on the subset $\{R^k_{ij} = 0\}$. The prediction value is then defined as $Q^{\tau k}_{nij}$, which is the estimation of $Q^{\tau k}_{ij}$. SuperLearner is adopted to better guarantee the consistency of the TMLE (i.e. so that it is more likely that the model for Q is correctly specified.). Finally, G-computation is conducted to estimate the quantity $\tau^k = E[E(Y_{ij} = 1|A^k_{ij} = 1, X^k_{ij}, R^k_{ij} = 0, C_{ij} = 0)|R^k_{ij} = 0]$.

$$\tau^{Gcomp,k}_n = \frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R^k_{ij=0}}} Q^{\tau k}_{nij},$$

where n^k is the number of subjects who were not known to be resistant to treatment k, and $\tau^{Gcomp,k}_n$ is the G-computation estimation of τ^k.

5.5.2 Propensity Score Model (g Model)

The g model, or propensity score model defined below enables us to use IPTW (see Section 2.3.3) to estimate τ^k.

We define the probability as:

$$g_{ij}^{\tau k} = Pr(A_{ij}^k = 1, C_{ij} = 0 | X_{ij}^k, R_{ij}^k = 0).$$

Under the assumptions described, we divide g model for τ^k into three submodels:

$$\begin{aligned}
g_{ij}^{\tau k} =& Pr(A_{ij}^k = 1, C_{ij} = 0 | X_{ij}^k, R_{ij}^k = 0) \\
=& Pr(A_{ij}^k = 1, D_{ij}^k = 1, C_{ij} = 0 | X_{ij}^k, R_{ij}^k = 0) \\
=& Pr(A_{ij}^k = 1 | D_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) \cdot Pr(D_{ij}^k = 1 | X_{ij}^k, R_{ij}^k = 0) \\
& \cdot Pr(C_{ij} = 0 | A_{ij}^k = 1, D_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) \\
=& Pr(A_{ij}^k = 1 | D_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) \cdot Pr(D_{ij}^k = 1 | V_{ij}) \\
& \cdot Pr(C_{ij} = 0 | A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0) \\
=& g_{1ij}^{\tau k} \cdot g_{2ij}^{\tau k} \cdot g_{3ij}^{\tau k}.
\end{aligned}$$

The consistency of IPTW relies on the law of Large Numbers as shown below:

$$\frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} Y_{ij} \frac{\mathbb{1}\{A_{ij}^k = 1, C_{ij} = 0\}}{g_{ij}^{\tau k}} \xrightarrow{n^k \to \infty} E\left(Y_{ij} \frac{\mathbb{1}\{A_{ij}^k = 1, C_{ij} = 0\}}{g_{ij}^{\tau k}} \,\middle|\, R_{ij}^k = 0 \right).$$

Note that $g_{ij}^{\tau k} > 0$ by the positivity Assumption A2.

We then prove that the IPTW estimator is consistent to τ^k as below:

$$
\begin{aligned}
& E\left(Y_{ij}\frac{\mathbb{1}\{A_{ij}^k=1, C_{ij}=0\}}{g_{ij}^{\tau k}} \,\middle|\, R_{ij}^k=0\right) \\
=& E\left[E\left(\frac{Y_{ij}\mathbb{1}\{A_{ij}^k=1, C_{ij}=0\}}{g_{ij}^{\tau k}} \,\middle|\, A_{ij}^k=1, X_{ij}^k, R_{ij}^k=0\right) \,\middle|\, R_{ij}^k=0\right], \\
& \text{(by the law of iterated expectation)}, \\
=& E\left[E\left(\frac{Y_{ij}\mathbb{1}\{A_{ij}^k=1, D_{ij}^k=1, C_{ij}=0\}}{g_{ij}^{\tau k}} \,\middle|\, A_{ij}^k=1, D_{ij}^k=1, X_{ij}^k, R_{ij}^k=0\right) \,\middle|\, R_{ij}^k=0\right], \\
& \text{(since } A_{ij}^k=1 \text{ implies } D_{ij}^k=1\text{)}, \\
=& E\left[E\left(\frac{Y_{ij}(D_{ij}^k=1)\mathbb{1}\{A_{ij}^k(D_{ij}^k=1)=1, D_{ij}^k=1, C_{ij}=0\}}{g_{ij}^{\tau k}}\right.\right. \\
& \left.\left. \,\middle|\, A_{ij}^k(D_{ij}^k=1)=1, D_{ij}^k=1, X_{ij}^k, R_{ij}^k=0\right) \,\middle|\, R_{ij}^k=0\right], && \text{by A3} \\
=& E\left[\frac{E\left(Y_{ij}(D_{ij}^k=1) \,\middle|\, A_{ij}^k(D_{ij}^k=1)=1, D_{ij}^k=1, X_{ij}^k, R_{ij}^k=0\right)}{g_{ij}^{\tau k}}\right. \\
& \left. \cdot E\left(\mathbb{1}\{A_{ij}^k=1(D_{ij}^k=1), D_{ij}^k=1, C_{ij}=0\}\right) \,\middle|\, R_{ij}^k=0\right], && \text{by A4} \\
=& E\left[\frac{E\left(Y_{ij} \,\middle|\, A_{ij}^k=1, D_{ij}^k=1, X_{ij}^k, R_{ij}^k=0, C_{ij}=0\right)}{g_{ij}^{\tau k}}\right. \\
& \left. \cdot E\left(\mathbb{1}\{A_{ij}^k=1, D_{ij}^k=1, C_{ij}=0\}\right) \,\middle|\, R_{ij}^k=0\right], && \text{by A3} \\
=& E\left[\frac{E\left(Y_{ij} \,\middle|\, A_{ij}^k=1, X_{ij}^k, R_{ij}^k=0, C_{ij}=0\right)}{g_{ij}^{\tau k}} \cdot E\left(\mathbb{1}\{A_{ij}^k=1, C_{ij}=0\}\right) \,\middle|\, R_{ij}^k=0\right],
\end{aligned}
$$

(since $A_{ij}^k = 1$ implies $D_{ij}^k = 1$),

$$= E\left[\frac{E\left(Y_{ij} \,\middle|\, A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0, C_{ij} = 0\right)}{g_{ij}^{\tau k}} \cdot E\left[E\left(\mathbb{1}\{A_{ij}^k = 1, C_{ij} = 0\} \,\middle|\, X_{ij}^k, R_{ij}^k = 0\right)\right] \,\middle|\, R_{ij}^k = 0\right],$$

(by the law of iterated expectation),

$$= E\left[E\left(Y_{ij} \,\middle|\, A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0, C_{ij} = 0\right) \cdot \frac{Pr(A_{ij}^k = 1, C_{ij} = 0 | X_{ij}^k, R_{ij}^k = 0)}{Pr(A_{ij}^k = 1, C_{ij} = 0 | X_{ij}^k, R_{ij}^k = 0)} \,\middle|\, R_{ij}^k = 0\right],$$

$$= E\left[E\left(Y_{ij} \,\middle|\, A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0, C_{ij} = 0\right) \,\middle|\, R_{ij}^k = 0\right],$$

$$= \tau^k.$$

In summary, we can use the IPTW estimator

$$\tau_n^{IPTW,k} = \frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} Y_{ij} \frac{\mathbb{1}\{A_{ij}^k = 1, C_{ij} = 0\}}{g_{nij}^{\tau k}}$$

to estimate τ^k, where $g_{nij}^{\tau k}$ is the estimation of $g_{ij}^{\tau k}$.

Note that we can also estimate τ_C^k with the assumptions in Section 5.4 using IPTW.

5.5.3 TMLE Algorithm

Even though either Q or g can be used to estimate τ^k, it is less robust because if Q or g is misspecified, then the estimation is biased. Thus we prefer to use a more robust estimator. TMLE produces doubly robust and locally efficient estimation based on both Q and g (see Section 2.6.3). This section elaborates on the details of

the TMLE algorithm in our case.

After estimating the two models as discussed above, we can update $Q_{nij}^{\tau k}$ with $g_{nij}^{\tau k}$ using a logistic regression as we described in section 2.6.2:

$$logit\{Q_{nij}^{\tau k^*}(\epsilon^\tau)\} = logit\{Q_{nij}^{\tau k}\} + \frac{\epsilon^\tau}{g_{nij}^{\tau k}}, \tag{5.1}$$

where ϵ^τ is chosen such that the above fluctuation minimizes a logistic regression loss function $L(Y, Q_{nij}^{\tau k^*}(\epsilon^\tau)) = \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} -(Y_{ij} log(Q_{nij}^{\tau k^*}(\epsilon^\tau)) + (1 - Y_{ij}) log(1 - Q_{nij}^{\tau k^*}(\epsilon^\tau)))$. We denote ϵ_n^τ as the estimate of ϵ^τ, and ϵ_n^τ is estimated by a logistic regression:

$$glm(Y_{ij} \sim -1 + \frac{A_{ij}^k}{g_{nij}^{\tau k}}); \qquad offset = logit(Q_{nij}^{\tau k}). \tag{5.2}$$

ϵ_n^τ is then the coefficient of the covariate $\frac{A_{ij}^k}{g_{nij}^{\tau k}}$.

Recall that TMLE is a substitution estimator. Thus the estimate of τ^k is

$$\tau_n^{TMLE,k} = \frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} Q_{nij}^{\tau k^*}.$$

For the estimation of μ^k, we define the outcome model as

$$Q_{ij}^{\mu k} = Pr(Y_{ij} = 1 | X_{ij}^k, A_{ij}^k, R_{ij}^k = 0, C_{ij} = 0).$$

Since we only need to correct for censoring, the propensity score model is

$$g_{ij}^{\mu k} = Pr(C_{ij} = 0 | X_{ij}^k, A_{ij}^k, R_{ij}^k = 0).$$

Following similar procedures, first we conduct a logistic regression

$$glm(Y_{ij} \sim -1 + \frac{A_{ij}^k}{g_{nij}^{\mu k}}); \qquad offset = logit(Q_{nij}^{\mu k}), \tag{5.3}$$

in order to estimate ϵ^μ. We define ϵ_n^μ as the estimate of ϵ^μ, which is the coefficient of $\frac{A_{ij}^k}{g_{nij}^{\mu k}}$.

We then update $Q_{nij}^{\mu k}$ using the logistic regression fit

$$logit\{Q_{nij}^{\mu k *}(\epsilon_n^\mu)\} = logit\{Q_{nij}^{\mu k}\} + \frac{\epsilon_n^\mu}{g_{nij}^{\mu k}}, \tag{5.4}$$

Then the estimator of μ^k is

$$\mu_n^{TMLE,k} = \frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} Q_{nij}^{\mu k *}.$$

Finally, we get the estimate of the parameter of interest:

$$\psi_n^{TMLE,k} = \tau_n^{TMLE,k} - \mu_n^{TMLE,k},$$

where $\psi_n^{TMLE,k}$ is the substitution estimate of ψ^k.

5.6 TMLE Asymptotics

TMLE is a semiparametric estimator, as described in Section 2.5. To better understand the algorithm, we first write the estimator in its asymptotically linear form (2.1):

$$(n^k)^{\frac{1}{2}}(\psi_n^{TMLE,k} - \psi^k) = (n^k)^{-\frac{1}{2}} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} IC_{ij}^{\psi k} + o_p(1),$$

where $IC_{ij}^{\psi k}$ is the value of the influence curve of ψ^k for subject i in study j.

By Central Limit Theorem we have that,

$$(n^k)^{\frac{1}{2}}(\psi_n^{TMLE,k} - \psi^k) \xrightarrow{\text{D}} N(0, E(IC_{ij}^{\psi k} \cdot (IC_{ij}^{\psi k})^T), \tag{5.5}$$

So that the large-sample properties of the estimator can be characterized by its influence curve.

To obtain $IC_{ij}^{\psi k}$, we first need to know $IC_{ij}^{\tau k}$ and $IC_{ij}^{\mu k}$, the influence curve of τ^k and μ^k respectively. In this case,

$$IC_{ij}^{\tau k} = \frac{\mathbb{1}(A_{ij}^k = 1, R_{ij}^k = 1, C_{ij} = 0)}{g_{nij}^{\tau k}}(Y_{ij} - Q_{nij}^{\tau k *}) + Q_{nij}^{\tau k *} - \tau^k, \tag{5.6}$$

$$IC_{ij}^{\mu k} = \frac{\mathbb{1}(C_{ij}^k = 1, R_{ij}^k = 0)}{g_{nij}^{\mu k}}(Y_{ij} - Q_{nij}^{\mu k *}) + Q_{nij}^{\mu k *} - \mu^k. \tag{5.7}$$

We use the Delta method to show that $IC_{ij}^{\psi k} = IC_{ij}^{\tau k} - IC_{ij}^{\mu k}$ [86] below.

Let ϕ be a function, which is differentiable at ψ and assume $\sqrt{n}(\phi(\psi_n) - \phi(\psi))$ converges in distribution,

$$\begin{aligned}
\sqrt{n}(\phi(\psi_n) - \phi(\psi)) =& \phi_\psi^{'}(\sqrt{n}(\psi_n - \psi)) + o_p \quad \text{by the Delta method} \\
=& \frac{1}{\sqrt{n}} \phi_\psi^{'} \sum IC(\psi) + o_p \quad \text{by asymptotic linearity} \\
=& \frac{1}{\sqrt{n}} \sum \phi_\psi^{'} IC(\psi) + o_p, \quad \text{by the continuous mapping theorem}
\end{aligned}$$

where $\phi_{\psi}^{'}$ is the derivative of $\phi(\psi)$ evaluated at ψ.

This means the influence curve of the transformed estimator $\phi(\psi_n)$ is $\phi_{\psi}^{'} IC(\psi)$.

In our case, $\psi^k = \phi(\tau^k - \mu^k) = \tau^k - \mu^k$, so $\phi_{\psi}^{'} = (1, -1)$.

Define $IC_{\tau^k - \mu^k, ij}$ as the influence curve of ψ^k. Thus,

$$IC_{ij}^{\psi k} = C_{\tau^k - \mu^k, ij} = \phi_{\psi}^{'} IC(\psi) = (1, -1) IC(\psi) = IC_{ij}^{\tau k} - IC_{ij}^{\mu k}.$$

Referring to section 2.5.4, let $IC_{ij}^{\psi k}$ denote the efficient influence curve of ψ^k, then the efficient influence curve estimating equation in this case is:

$$\frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} IC_{ij}^{\psi k} = 0. \tag{5.8}$$

Recall from Section 2.6.2 that the TMLE update step in (5.1) and (5.2) automatically solves the efficient influence curve estimating equation (5.8). This occurs because the logistic regression update step minimizes a corresponding loss function by solving the logistic regression score equations. This is done separately for the two components

$$\frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} IC_{ij}^{\tau k} = 0,$$

and

$$\frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} IC_{ij}^{\mu k} = 0,$$

resulting in the solution to function (5.8).

5.7 Variance

Clustering is considered in the variance estimation. We assume that studies are independent between each other and that individuals in a given study have an arbitrary dependence pattern (see Section 5.2, assumption A7). Thus, ignoring the clustering would cause bias in the estimation of the variance of the TMLE.

By Central Limit Theorem, with sufficient clusters, we can estimate the variance of $\psi_n^{TMLE,k}$ using the efficient influence curve of ψ^k in our example. Recall from (5.5), the asymptotic variance of ψ^k is [52] (see Section 2.5.2)

$$\begin{aligned}(\sigma_\psi^k)^2 &= \frac{1}{(n^k)^2} E\left(\sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} IC_{ij}^{\psi k}\right)^2 \\ &= \frac{1}{(n^k)^2} \sum_{j=1}^{31} \left[\sum_{\substack{i,m \in S_j \\ s.t. R_{ij=0}^k}} E(IC_{ij}^{\psi k} \cdot IC_{ij}^{\psi k}) \mathbb{1}(i \neq m) + E(IC_{ij}^{\psi k})^2 \mathbb{1}(i = m)\right],\end{aligned} \tag{5.9}$$

where $IC_{ij}^{\psi k}$ is defined in Section 5.6. Following the formula above, we can obtain the variance estimation of $\psi_n^{TMLE,k}$ in Section 5.3.3.

Note that the variance is being computed over clusters and according to Central Limit Theorem, the approximation depends on a large number of clusters. We believe that 31 is large enough as we evaluated 30 clusters in the simulation study and it produced adequate results.

So far, we discussed the method of constructing a TMLE estimator and estimating its associated variance. As mentioned before, the TMLE estimator is doubly robust, which means either the Q model or the g model is correctly specified, the estimation is consistent. More specifically, if either $Q_{ij}^{\tau k}$ or $g_{ij}^{\tau k}$ is correctly specified, the estimation of τ^k is consistent. In addition, if all of the models are correctly specified, the TMLE algorithm produces an estimator that has the minimum variance among the regular semiparametric estimators in its class.

5.8 Real Data Analysis

In this section, we describe the analysis that we performed based on the aforementioned methods. We used RStudio (Version 1.0.44) for the data analysis and utilized the parallel computing facilities from Compute Canada.

5.8.1 Data Preparation

Since we had only a few missing values in the variable sex and age (0.06% and 0.3% respectively), we dropped all subjects without this information. However, missing values accounted for 2.8%–27.1% in other covariates, so we added in a missingness indicator for each of the other covariates[47, 54]. For example, HIV status is a binary covariate (positive or negative). We then introduced 2 dummy variables indicating HIV-positive and HIV-negative status into the model, so that the reference level is a missing HIV status.

5.8.2 Models and Estimation

As data are limited, we use a combination of parametric and nonparametric models to estimate the Q and g probabilities. Table 5-1 gives the models used, the estimation subset and the prediction subset for each of the quantities to be estimated for the estimation of τ^k.

Model for g

$g_{ij}^{\tau k}$ was estimated using either logistic regression or least absolute shrinkage and selection operator (LASSO) for each of the three components. Referring to section 4.5, $g_{1ij}^{\tau k}$ is estimated on the subset of $\{D_{ij}^k = 1, R_{ij}^k = 0\}$ using logistic regression and predicted for all the subjects with $R_{ij}^k = 0$. For the estimation of $g_{2ij}^{\tau k} = Pr(D_{ij}^k = 1|V_{ij})$, there were only 31 data points (31 studies in total) that could be used to fit the model because the patients in the same study shared the same D_{ij} and V_{ij}. We therefore used LASSO in order to be able to include all study-level covariates V_{ij}. However, the treatments Ethambutol and Pyrazinamide were only unavailable in one study, Kanamycin was unavailable in three studies, and Ofloxacin was unavailable in four studies. Because of the difficulty of fitting the corresponding models, we considered $g_{2ij}^{\tau k}; k \in \{$Ethambutol, Pyrazinamide, Kanamycin, Ofloxacin$\}$ as constants, corresponding with 30/31, 30/31, 28/31, and 27/31 respectively.

$g_{3ij}^{\tau k}$ was estimated in the subset of $\{A_{ij}^k = 1, R_{ij}^k = 0\}$ and predicted on $\{R_{ij}^k = 0\}$ using logistic regression. Note that there were insufficient data to fit this model for the treatment Amikacin as there were few censored observations. In this case, we

Table 5–1: Model specification for the estimation of τ^k.

Quantity	**Model**	**Estimation Subset**	**Prediction Subset**
$Q_{ij^{\tau k}} = Pr(Y_{ij} = 1 \mid A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0, C_{ij} = 0)$	SuperLearner[a]	$\{A_{ij}^k = 1, R_{ij}^k = 0, C_{ij} = 0\}$	$\{R_{ij}^k = 0\}$
$g_{1ij}^{\tau k} = Pr(A_{ij}^k = 1 \mid D_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0)$	logistic regression	$\{D_{ij}^k = 1, R_{ij}^k = 0)\}$	$\{R_{ij}^k = 0)\}$
$g_{2ij}^{\tau k} = Pr(D_{ij}^k = 1 \mid V_{ij})$	constant or LASSO[b]	31 data points [c]	$\{R_{ij}^k = 0)\}$
$g_{3ij}^{\tau k} = Pr(C_{ij} = 0 \mid A_{ij}^k = 1, X_{ij}^k, R_{ij}^k = 0)$	logistic regression[d]	$\{A_{ij}^k = 1, R_{ij}^k = 0)\}$	$\{R_{ij}^k = 0)\}$

NOTE: [a]The algorithm library of SuperLearner includes k-nearest neighborhood, random forest, logistic regression, and generalized linear model with penalty maximum likelihood.

[b] For treatment Ethambutol, Pyrazinamide, Kanamycin and Ofloxacin, proportion $\frac{\mathbb{1}(D_{ij}^k=1)}{31}$ was used;
For other treatments, LASSO was used.

[c]Because we included 31 studies in total, for each treatment, the patients in the same study share the same D_{ij}^k.

[d]For treatment Amikacin, LASSO is used because of the sparsity of outcomes.

again used LASSO instead of logistic regression.

Logistic regression for the g model was used to maintain the stability of the TMLE. Because of the sparsity of the outcomes for these models, machine learning methods might produce extreme values of estimation of the g model [75]. We used LASSO as an exception when the logistic regression did not converge.

Multiplying each of the three values together for each subject, we obtained $g^{\tau k}_{nij} = g^{\tau k}_{n1ij} \cdot g^{\tau k}_{n2ij} \cdot g^{\tau k}_{n3ij}$, for the estimation of $g^{\tau k}_{ij}$.

Model for Q

$Q^{\tau k}_{nij}$ was estimated using SuperLearner [63] (see Section 2.4). We used cluster-specific cross-validation in SuperLearner, so that the cross-validation error is computed by resampling the clusters. We estimated $Q^{\tau k}_{nij}$ with $X^k_{ij} = (W_{ij}, V_{ij}, A^{k^*}_{ij}; k^* \neq k)$ as our covariates, in the subset of $\{A^k_{ij} = 1, R^k_{ij} = 0, C_{ij} = 0\}$ and then predicted on the subset of $\{R^k_{ij} = 0, C_{ij} = 0\}$. In SuperLearner library, we included the algorithms k-nearest neighbors [88], Random Forests [48], logistic regression [65] and LASSO [30].

TMLE Algorithm

Following the TMLE procedure, we then fit a logistic regression (5.2) to estimate ϵ^τ_n, and then we updated $Q^{\tau k}_{nij}$ with $g^{\tau k}_{nij}$ using (5.1). Therefore, we got

$$\tau^{TMLE,k}_n = \frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R^k_{ij=0}}} Q^{\tau k^*}_{nij}$$

as the TMLE for τ^k.

Similarly, for the estimation of μ^k, we used logistic regression to obtain both $Q_{nij}^{\mu k}$ and $g_{nij}^{\mu k}$. $Q_{ij}^{\mu k}$ and $g_{ij}^{\mu k}$ were fit on the subset $\{R_{ij}^k = 0, C_{ij} = 0\}$ and $\{R_{ij}^k = 0\}$ respectively, and the predictions were made on the subset $\{R_{ij}^k = 0\}$ for both. Hence, we obtain values of $Q_{nij}^{\mu k}$ and $g_{nij}^{\mu k}$ for all subjects with $R_{ij}^k = 0$.

Then we fit a logistic regression (5.3) to estimate ϵ_n^{μ}, and updated $Q_{nij}^{\mu k}$ with $g_{nij}^{\mu k}$ using (5.4).

Therefore, we get

$$\mu_n^{TMLE,k} = \frac{1}{n^k} \sum_{j=1}^{31} \sum_{\substack{i \in S_j \\ s.t. R_{ij=0}^k}} Q_{nij}^{\mu k *}$$

as the TMLE for τ^k.

Then the estimate for the parameter of interest is:

$$\psi_n^{TMLE,k} = \tau_n^{TMLE,k} - \mu_n^{TMLE,k}.$$

5.8.3 Variance Calculation

Recall (5.6) to (5.9), the variance of $\psi_n^{TMLE,k}$, $(\sigma_\psi^k)^2$ is estimated by

$$\frac{1}{(n^k)^2} \sum_{j=1}^{31} \left[IC^{\psi k}_{R_{ij}^k=0, C_{ij}=0} \cdot (IC^{\psi k}_{R_{ij}^k=0, C_{ij}=0})^T \right).$$

Implementing the same procedures for all of the 15 treatments, we obtain the estimated treatment importance.

CHAPTER 6
Simulation

This simulation study aims to: 1) demonstrate the consistency of our estimator, 2) verify the double robustness of our estimator, 3) show that our method for estimating the variance is appropriate, and 4) demonstrate the coverage of the Wald-type confidence intervals computed with the estimated variance. The methods and results are elaborated below.

6.1 Methods

6.1.1 Data generation

We simulated data with a similar structure to our real-life data, (excluding all missing data and censoring). The sample size of each generated dataset is 9000, comprising 30 study-clusters with 300 individuals in each cluster. We generated a binary outcome of patient i in study j, Y_{ij}, three binary indicators of antibiotic use A_{ij}^k, k =1,2,3, one continuous study-level covariate V_{ij}, one individual level continuous covariate W_{ij}, an indicator of resistance to antibiotic k, R_{ij}^k, and one continuous unmeasured cluster-level variable U_{ij}. As in the real dataset, we consider that not all treatments are available in all studies (or clusters). We denote treatment availability for treatment k as D_{ij}^k, k=1,2,3, and note that patients in the same study share a same D_{ij}^k. Thus, the simulated data structure is

$$O_{ij} = (Y_{ij}, V_{ij}, W_{ij}, U_{ij}, \{A_{ij}^k, R_{ij}^k, D_{ij}^k\}, k = 1...3) \overset{i.i.d}{\sim} P_O; i \in S_j, j = 1...30.$$

Note that U_{ij} is not included in the data observed by the analyst. We denote $D_{ij} = (D_{ij}^1, D_{ij}^2, D_{ij}^3)$ to be the vector indicating treatment availability, and similarly for A_{ij} and R_{ij}.

Two scenarios were investigated. In the first, treatment availability was completely random and in the other one, treatment availability was dependent on the study-level covariate. Table 6-1 shows the data generating mechanisms for these two scenarios. In the first scenario, we first randomly and independently generated the study-level covariate V_{ij} with a Gaussian distribution, and individual resistance status R_{ij}^k with a Bernoulli distribution. Then we generated the individual-level covariate W_{ij} dependent on V_{ij} with Normal distributions, and study-level unmeasured variable U_{ij} (all the subjects in the same study had the same value of U_{ij}, so we define $U_j = U_{ij}$, if $i \in S_j$, the same for V_{ij}) dependent on V_{ij} with Normal distributions. In the process of generating the treatment availability D_{ij}^k, we first draw, for each study, how many drugs are available to this study (with a minimum of one), and then we generated which one(s) is(are) available to this study. For example, we sample a number from 1 to 3, say 2. Therefore, for study j, we have 2 medications that are available to this study. Then, we sample two drugs (say the first and the third) from 3 drugs. If the first and the third medications are available in study j, then vector R_{ij}= (1,0,1). Everyone in the same study shares the same D_{ij}. Next, we generated individual treatments where the probability of receiving $A_{ij}^k = 1$ depends on V_{ij}, W_{ij} and R_{ij}^k. We then forced $A_{ij}^k = 0$ when $D_{ij}^k = 0$. Finally, the binary outcome Y_{ij} was sampled from a Bernoulli distribution with probability dependent on V_{ij}, W_{ij}, U_{ij} and $A_{ij}^k(1 - R_{ij}^k)$, $k = 1, 2, 3$. See Appendix 1 for the data generating code for

scenario 1.

In the second scenario, the probability of $D_{ij}^k = 1$ was dependent on V_{ij}. See Appendix 2 for the data generating code for scenario 2.

6.1.2 Data analysis

Our parameter of interest in the simulation study is:

$$\psi^1 = E[E(Y_{ij}|X_{ij}^1, R_{ij}^1 = 0, A_{ij}^1 = 1)|R_{ij}^1 = 0],$$

where $X_{ij}^1 = (V_{ij}, V_{ij}, A_{ij}^{k^*}), k^* \neq 1$. (This corresponds to the first component of the treatment importance parameter.) In order to find the true value of ψ^1, we generated data as above, but with sample sizes greater than 10^7, and forced all A_{ij}^1 equal to 1. Then we took the average of the generated Y_{ij} within the subset of individuals who were not resistant to medication 1. After enlarging the sample size, the value of $E(Y_{ij}|R_{ij}^1 = 0)$,converged to a fixed value, which is regarded as the true value of our parameter of interest.

We used logistic regressions to fit $g_{1ij}^1 = Pr(A_{ij}^1|D_{ij}^1 = 1, X_{ij}^1, R_{ij}^1 = 0)$ In the first scenario. We took the proportion of $D_{ij}^1 = 1$ to be the estimate of g_{ij}^2, since in this scenario, D_{ij}^1 is independently generated. In the second scenario, we modeled $g_{ij}^2 = Pr(D_{ij}^1|V_{ij})$ using a logistic regression. Then the TMLE algorithm was applied to update the predicted values of Q_{ij}^1. 1000 random seeds were drawn and stored. 1000 datasets were then generated from these seeds and the analyses conducted on the datasets. See Appendix 3 for the modeling code.

Table 6–1: The data generating mechanism of two scenarios.

Variable	**Generating Mechanism**
V_{ij}	$V_{ij} \sim N(mean = 0.3, sd = 0.3, n = 30)$ Set $V_{ij} = V_{ij}$ for all i in S_j
R_{ij}^k	$R_{ij}^1 \sim Ber(p = 0.25, n = 300)$ $R_{ij}^2 \sim Ber(p = 0.30, n = 300)$ $R_{ij}^3 \sim Ber(p = 0.25, n = 300)$
W_{ij}	$W_{ij} \sim N(mean = 0.1V_{ij}, sd = 0.1, n = 300)$
U_{ij}	$U_{ij} \sim N(mean = 0.2V_{ij} + 0.1, sd = 0.5, n = 30)$ Set $U_{ij} = U_{ij}$ for all i in S_j
D_{ij}^k	**Scenario 1** Within the same study, a random number d=1,2, or 3 of available treatments was generated. Randomly select d treatments and set $D_{ij}^k = 1$ for these treatments. Set $D_{ij}^k = D_{ij}^k$ for all i in S_j **Scenario 2** $D_j^1 \sim Bin(logit(p) = 1 + 2V_{ij}, n = 30)$ $D_j^2 \sim Bin(logit(p) = 0.5 + 1.5V_{ij}, n = 30)$ $D_j^3 \sim Bin(logit(p) = 1.5 + 0.3V_{ij}, n = 30)$ Within the same study, $D_j^1 + D_j^2 + D_j^3 > 0$ Set $D_{ij}^k = D_{ij}^k$ for all i in S_j
A_{ij}^k	$A_{ij}^1 \sim Bin(logit(p) = -0.75 + 2.4V_{ij} + 1.8W_{ij} - 0.1R_{ij}^1, n = 9000)$ $A_{ij}^2 \sim Bin(logit(p) = -1 + V_{ij} + 1.7W_{ij} - 0.15R_{ij}^1, n = 9000)$ $A_{ij}^1 \sim Bin(logit(p) = -1.5 + 1.7V_{ij} + W_{ij} - 0.16R_{ij}^1, n = 9000)$
Y_{ij}	$Y_{ij} \sim Bin(logit(p) = 2 + 3.5V_{ij} + 0.3W_{ij} - 0.1U_{ij} + 2.2A_{ij}^1(1 - R_{ij}^1) +$ $0.12A_{ij}^2(1 - R_{ij}^2) + 0.05A_{ij}^3(1 - R_{ij}^3)$

In order to verify the double robustness property of our proposed estimator, we varied the model specifications used to estimate $g^1_{ij} = g^1_{1ij} \cdot g^1_{2ij}$ and Q^1_{ij}. In the first scenario, $Q^1_{ij}$and g^1_{ij} were misspecified as null models (the outcome regressed on the intercept only). In the second scenario, the misspecification of g^1_{ij} implied that g^1_{2ij} was also assigned a null model. In order to save space, we only provide the code when all the models in scenario 2 were correctly specified in Appendix 3 and all models were misspecified in Appendix 4.

The standard error of the TMLE was estimated using the influence curve as discussed in Sections 5.6 and 5.7. We then compared these standard error estimates with the estimates of the standard error computed using a cluster bootstrap [17]. Then, the coverage rates for different sample sizes were also calculated when all models were correctly specified. In addition, to show that we need to account for clustering when we estimate the variance, we also provided the coverage rates with the variance estimated with and without clusters taken into account.

6.2 Results

The true values for both scenarios were found to be 0.74 (up to two decimal places). In the first scenario, we present the mean TMLE estimate with four different model specifications: where both Q^1_{ij} and g^1_{ij} are correctly specified ($Qvgv$); where Q^1_{ij} is correctly specified but g^1_{ij} is misspecified ($Qvgx$); where Q^1_{ij} is misspecified but g^1_{ij} is correctly specified ($Qxgv$); and where both Q^1_{ij} and g^1_{ij} are misspecified ($Qxgx$). For each of these scenarios, we also present the percentage of mean bias, Monte-carlo standard error and sandwich estimator standard error in Table 6-2.

Table 6–2: Simulated estimation with different model specifications and their bias percentage in scenario 2.

Situation	**Model specification**[a]	**Average Estimate**	**Bias (%)**	**Monte-Carlo SE**	**Sandwich estimator SE**
1	$Qvgv$	0.74	0.0	0.035	0.033
2	$Qvgx$	0.74	0.0	0.035	0.033
3	$Qxgv$	0.74	0.0	0.040	0.038
4	$Qxgx$	0.80	8.1	0.033	0.031

NOTE: [a] v indicates the model is correctly specified and x indicates the model is misspecified.

The table shows that when either the Q_{ij}^1 or g_{ij}^1 model was correctly specified, the estimate of the parameter was unbiased at this sample size. The estimate was biased only if both of the models were misspecified. Similarly, Figure 6-1 contains the boxplots of the simulated estimations with different model specifications. The red line is the true value in scenario 1. We can see that when Q_{ij}^1 is misspecified and g_{ij}^1 is correctly specified, the estimate is consistent, although the variance is a little bit larger than the estimates when Q_{ij}^1 is correctly specified. Unsurprisingly, the mean estimate when all models are misspecified diverges from the true value. Also note that the sandwich estimator slightly underestimates the Monte-Carlo standard error.

In addition, we used two estimation methods to calculate the associate standard error when all of the models are correctly specified in scenario 1: constructing a sandwich estimator by using efficient influence curves; and estimating by cluster bootstrap [17]. As shown in Table 6-2, the sandwich estimator standard error was 0.033, while the estimation from cluster bootstrap was 0.032 (a difference of 3%).

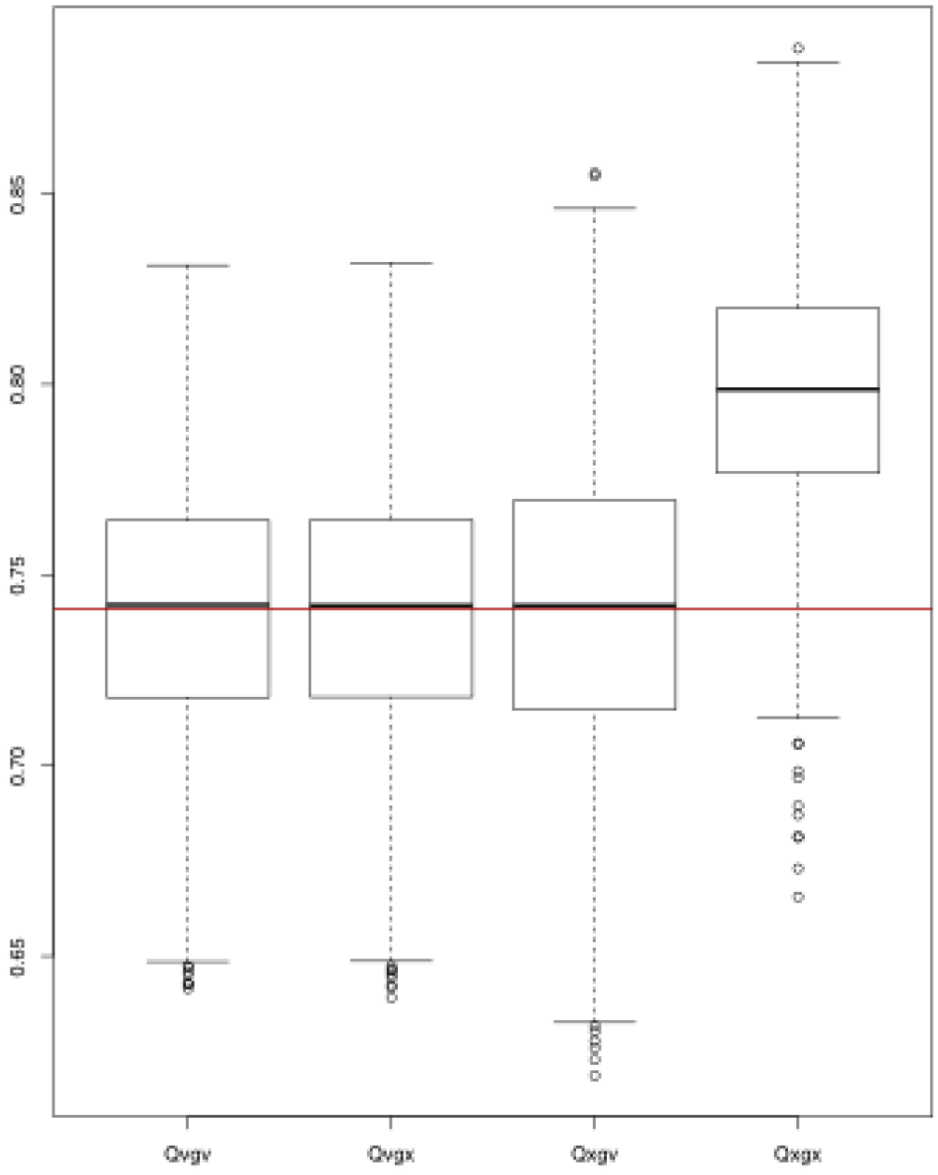

Figure 6–1: Boxplots of the estimated values in the simulation study with different model specifications in scenario 1.

Table 6–3: Simulated estimation with different model specifications and their bias percentage in scenario 2.

Situation	**Model specification**[a]	**Average Estimate**	**Bias (%)**	**Monte-Carlo SE**	**Sandwich estimator SE**
1	$Qvg1vg2v$	0.74	0.0	0.035	0.033
2	$Qvg1xg2v$	0.74	0.0	0.034	0.032
3	$Qvg1vg2x$	0.74	0.0	0.034	0.033
4	$Qvg1xg2x$	0.74	0.0	0.034	0.032
5	$Qxg1vg2v$	0.74	0.0	0.035	0.036
6	$Qxg1vg2x$	0.76	2.7	0.032	0.033
7	$Qxg1xg2v$	0.80	8.1	0.027	0.028
8	$Qxg1xg2x$	0.81	9.5	0.028	0.026

NOTE: [a] v indicates the model is correctly specified and x indicates the model is misspecified.

In the second scenario, where treatment availability D_{ij}^k was dependent on study-level covariate V_{ij}, we must also model g_{2ij}^1. Therefore, we have 8 different model specifications to verify the double robustness property, which should hold only when Q_{ij}^1 is correctly specified or when both g_{1ij}^1 and g_{2ij}^1 are correctly specified. We again used a logistic regression to correctly specify $Q_{ij}^1 = Pr(D_{ij}^1|R_{ij}^1 = 0)$ and a null model to misspecify it. Other procedures were similar to scenario 1. The results are shown in Table 6-3 and Figure 6-2.

The results show that when Q_{ij}^1 was misspecified, misspecifying either g_{1ij}^1 or g_{2ij}^1 resulted in some bias (although in this data generating scenario, the misspecification of g_{2ij}^1 resulted in less bias than g_{1ij}^1). As expected, when Q_{ij}^1 or g_{ij}^1 was correctly specified, the results were unbiased. The bias was largest when all three models were misspecified.

Table 6-4 shows the coverage rate of the Wald-type confidence intervals calculated

Table 6–4: Coverage rate with different sample size.

cluster size	**subjects in each cluster**	**coverage (%)**
30	300	92
60	300	94
30	600	92
60	600	95

using the sandwich estimator of the variance for different sample sizes. We can see that the coverage increases to the optimal rate with the sample size. In particular, we see that the number of clusters drives the convergence, which corresponds with the knowledge that the validity of the first-order estimation in this context depends on the number of clusters (see Section 2.5). Therefore, usage of the efficient influence function for the estimation of the variance is only effective with a sufficient number of clusters.

Furthermore, the coverage rate when the variance is estimated with clusters was 91.3%, while the coverage rate when the variance is estimated without clusters was 42.5%, which showed us that we need to take clustering into consideration when we estimate the variance of the parameter of interest.

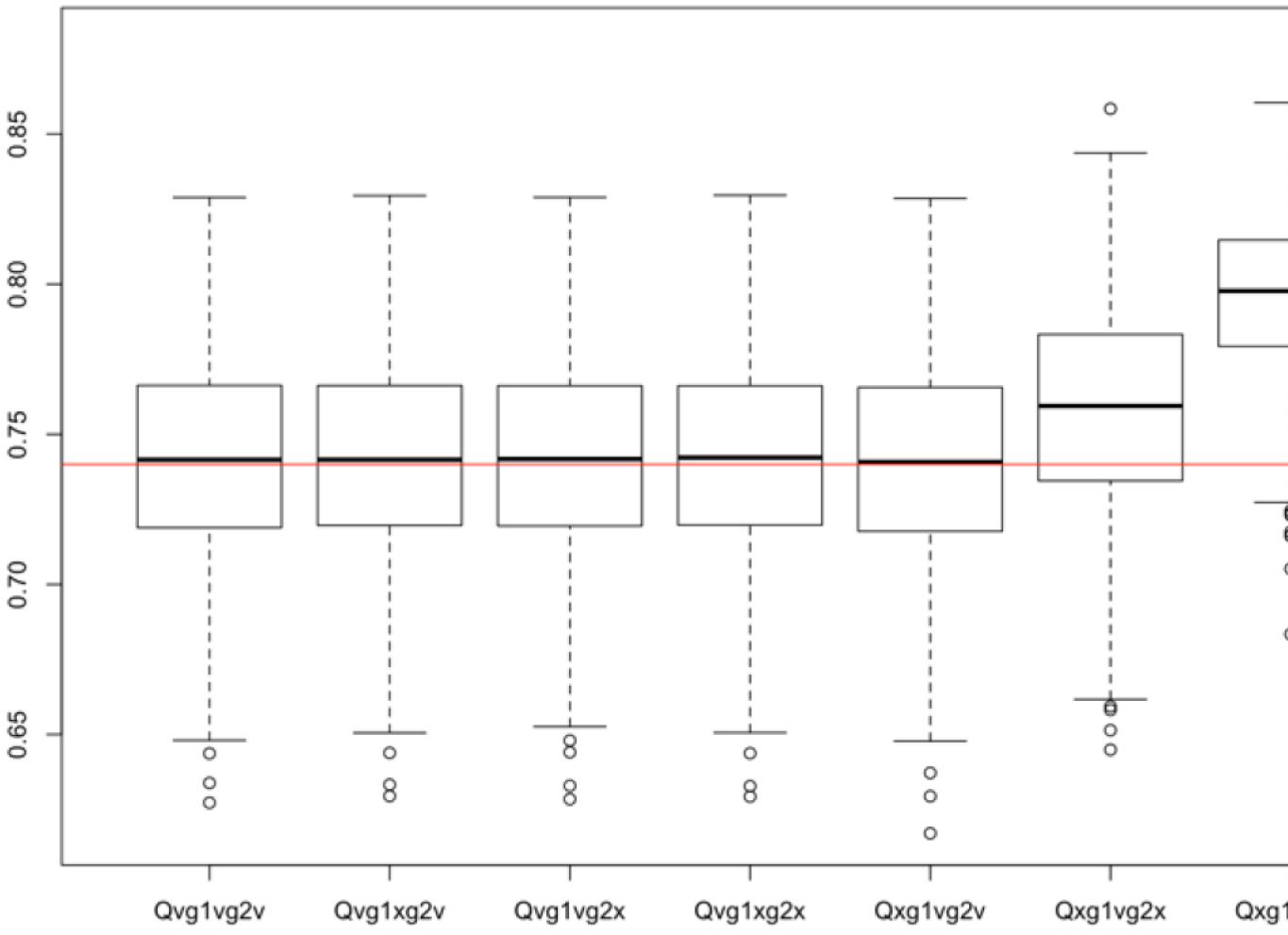

Figure 6–2: Boxplots of simulated estimation with different model specifications in scenario 2.

CHAPTER 7
Results

In this chapter, we present the data summary by descriptive statistics and results of the analysis of the MDR-TB data.

7.1 Data Description

The data set consists of 31 studies with a total of 9290 patients. Very few observations were missing sex and age (0.065% and 0.301% respectively). For the convenience of analysis, we dropped the observations with missing sex and age information. As a result, 31 studies (9258 subjects) were included in this analysis. The data consists of the outcome (defined as treatment success or not); drug treatments used; two study-level covariates: the income group of the country of the study and the year of the study; and 6 individual-level variables: age, sex, HIV status, acid fast bacilli (AFB) smear status at start of MDR-TB treatment regimen, past history of TB, and cavitation status of a patient. In addition, we also have the patients' resistance status. The data contained resistance information for 8 treatments and we defined the resistance status of patients of the other 7 treatments as not available. Study level information including the numbers of patients and treatments used in each cohort, and the mean (with standard error) or the number (with proportion) of individual-level covariates, are presented in Figure 7-1. Some studies included only 3 treatments while others used all available treatments documented in the dataset.

The publication year ranged from 1995 to 2009. The number of patients in different studies varied from 43 to 2211, and the average number of patients in a study was 300. Studies were conducted in countries that belonged to high, upper middle and lower middle income groups. The range of the mean age of the participants in each study varied from 31.1 to 47.6, while one pediatric study's mean age was only 7. 47.4% to 96% patients were male. Most patients were HIV-negative, and just 9 studies contained HIV-positive patients. Furthermore, the proportions of the patients with positive acid fast bacilli smear, past TB, and cavitation on X-ray all ranged from 0 to 100%.

Individual level information for each covariate and outcome are shown in Table 7-1. The year of study ranged from 1995 to 2009, with the median 2004 and inter-quartile range from 2002 to 2004. The median age of patients was 38, with inter-quartile range from 29 to 48. 404 (4.4%) and 3106 (33.4%) of patients came from low middle income group countries and upper middle income group countries respectively as defined by the World Bank [7], and the remaining 5780 (62.2%) patients came from high income group countries. Male patients accounted for 32.1% with 6305 patients. The number of patients with positive HIV, AFB smear, positive past TB, and positive cavitation X-ray were 1193 (12.8%), 5836 (62.8%), 6489 (69.8%) and 4623 (50.0%), respectively. In terms of outcome, 4847 (52.2%) patients had a clinically successful outcome, while 260 (2.8%) patients failed to report their outcome.

Study name	Year	IG[a]	N.[b]	Txts. used[c]	Age mean (SD)	Male N. (%)[d]	HIV N. (%)[e]	smear N. (%)[f]	TB N. (%)[g]	Cav. N. (%)[h]
Ahuja	1997	H	823	a-h,j-o	41.4 (12.0)	561 (68.2)	488 (59.3)	509 (61.8)	0 (0.0)	0 (0.0)
Avenda	2009	H	72	a-h,j-o	36.3 (15.3)	43 (59.7)	1 (1.4)	67 (93.1)	39 (54.2)	0 (0.0)
Burgos	2000	H	48	a-h,j-o	47.2 (14.8)	32 (66.7)	11 (22.9)	36 (75.0)	31 (64.6)	21 (43.8)
Chan	1998	H	203	a,c-h,j-o	42.0 (14.3)	116 (57.1)	0 (0.0)	203 (100.0)	196 (64.4)	120 (59.1)
Chiang	1996	H	125	a,e,g-m,o	46.1 (15.2)	90 (72.0)	0 (0.0)	109 (87.2)	114 (91.2)	0 (0.0)
Cox	2005	LM	77	a,c,e-h,j-m,0	36.9 (11.2)	47 (61.0)	0 (0.0)	76 (98.7)	77 (100.0)	0 (0.0)
Garcia	2009	UM	47	a,j-o	47.6 (16.4)	26 (55.3)	0 (0.0)	42 (89.4)	14 (29.8)	20 (42.6)
Granic	2006	H	104	a-h,j-m	40.3 (19.5)	61 (58.7)	1 (1.0)	75 (72.1)	62 (59.6)	0 (0.0)
Koh	2005	H	155	a,c-e,g-i,k-o	40.9 (14.4)	82 (52.9)	0 (0.0)	131 (84.5)	137 (88.4)	0 (0.0)
Leung	1997	H	99	e,g,l-m	46.1 (16.2)	74 (74.7)	0 (0.0)	78 (78.8)	62 (62.6)	46 (46.5)
Miglio	2004	H	101	a-i,k-o	39.4 (14.7)	61 (60.4)	6 (5.9)	80 (79.2)	58 (57.4)	48 (47.5)
Mitnic	2002	UM	732	a-o	31.1 (12.0)	436 (59.6)	8 (1.1)	508 (69.4)	720 (98.4)	550 (75.1)
Narita	1997	H	81	a-h,j-m,o	40.2 (11.8)	55 (67.9)	41 (50.6)	0 (0.0)	53 (65.4)	33 (40.7)
Palmer	1999	H	114	a,c-i,k-m,o	35.3 (13.3)	54 (47.4)	0 (0.0)	108 (94.7)	76 (66.7)	103 (90.4)
Pasvol	2004	H	45	a,b,d,e,g-o	36.9 (15.9)	21 (51.2)	0 (0.0)	29 (70.7)	24 (58.5)	0 (0.0)
Pena	2000	H	25	a,h,o	41.2 (13.3)	24 (96.0)	0 (0.0)	25 (100.0)	22 (88.0)	16 (64.0)
Perez	1995	UM	34	a,b,d,g,i,k-m,o	42.1 (12.4)	21 (61.8)	0 (0.0)	34 (100.0)	34 (100.0)	22 (64.7)
Quy	2000	LM	157	a,j-l	39.5 (11.4)	121 (77.1)	4 (2.5)	157 (100.0)	102 (65.0)	0 (0.0)
Riekst	2004	H	1027	a-c,e,g-i,k-o	42.3 (12.7)	780 (75.9)	32 (3.1)	269 (26.2)	648 (63.1)	704 (68.5)
Robert	1999	H	45	a,e-h,j-m,o	41.7 (15.6)	24 (53.3)	9 (20.0)	33 (73.3)	24 (53.3)	31 (68.9)
Schaaf	2002	UM	39	a,b,e-g,j-m,o	7.0 (5.4)	20 (51.3)	6 (15.4)	9 (23.1)	12 (30.8)	14 (35.9)
Seung	2002	H	142	a,e,g-i,k-m,	43.9 (15.4)	117 (82.4)	0 (0.0)	142 (100.0)	142 (100.0)	86 (60.6)
Shim	2002	H	1364	a,e,g-o	42.8 (14.9)	1014 (74.3)	1 (0.1)	927 (68.0)	977 (71.6)	569 (41.7)
Shin	2004	H	608	a-c,e-i,l-o	35.8 (11.3)	506 (83.2)	5 (0.8)	497 (81.7)	592 (97.4)	368 (60.5)
Shirai	2007	H	61	a,c-f,h,j-o	46.4 (11.9)	46 (75.4)	0 (0.0)	0 (0.0)	0 (0.0)	61 (100.0)
Tabars	2006	UM	43	a,b,e,g-i,l,m,o	44.4 (19.1)	27 (62.8)	0 (0.0)	42 (97.7)	43 (100.0)	43 (100.0)
Tupasi	2003	LM	170	a-o	39.2 (12.4)	106 (62.3)	0 (0.0)	107 (62.9)	164 (96.5)	138 (181.2)
VanDer	2008	H	43	a,b,d-h,j,l-m,o	32.9 (18.3)	31 (73.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Vander	2004	UM	2211	a,e-g,l,m	36.6 (10.9)	13965 (62.5)	565 (25.9)	1376 (63.0)	1963 (89.9)	1247 (57.1)
Viikle	2002	H	284	a-c,e,g-m,o	43.0 (13.6)	201 (70.8)	9 (3.2)	153 (53.9)	0 (0.0)	206 (72.5)
Yimkim	2007	H	211	a-c,e,g-o	39.3 (15.8)	124 (58.8)	0 (0.0)	0 (0.0)	84 (39.8)	160 (75.8)

NOTE: [a] IG: Income group of the country of each study. LM indicates the country belongs to the lower middle income group, UM indicates the country belongs to the upper middle income group, H indicates the country belongs to the high income group.
[b] N.: The number of patients in each study.
[c] We use letters to present the treatments. a: Ethambutol; b:Amikacin; c:Capreomycin; d:Ciprofloxacin; e:Cycloserine; f:Ethionamide/Prothionamide; g:Ofloxacin; h:Para-Aminosalicylic Acid; i:Protionamide; j:Rifabutin; k:Streptomycin; l:Pyrazinamide; m:Kanamycin n:High- generation quinolones; o:Group 5 level drugs respectively.
[d] Male N. (%): Number and percentage of male in each study.
[e] HIV N. (%): Number and percentage of positive HIV patients in each study.
[f] smaer N. (%): Number and percentage of positive Acid fast bacilli smear patients in each study.
[g] TB N. (%): Number and percentage of patients with past positive TB strains in each study.
[h] Cav. N. (%): Number and percentage of patients with positive cavitation on X Ray in each study.

Figure 7–1: Study Information.

Table 7–1: Descriptive statistics of covariates and outcome.

Covariates	**Summary**		Missing N.(%)
Year of Study	Median	2004	
	IQR[a]	(2002,2004)	
Age	Median	38	28 (0.3)
	IQR	(29,48)	
IG[b] N.(%)	LM	404(4.4)	
	UM	3106 (33.4)	
	H	5780 (62.2)	
Sex N.(%)	Male	2979 (32.1)	6 (0.06)
	Female	6305 (67.9)	
positive HIV N. (%)		1193 (12.8)	1369 (14.7)
positive smear N. (%)		5836 (62.8)	1439 (15.5)
positive past TB N. (%)		6489 (69.8)	524 (5.6)
positive cavitation N. (%)		6489 (69.8)	2521 (27.1)
Success outcome N. (%)		4847 (52.2)	260 (2.8)

NOTE: [a]IQR: inter-quartile range.

[b]IG: Income group of the country of each study. LM indicates the country belongs to the lower middle income group, UM indicates the country belongs to the upper middle income group, H indicates the country belongs to the high income group. rest of countries belongs to the high income group.

For each treatment, the number of patients who used the treatments and were resistant to them are also provided in Table 7-2. The proportion of patients who used a certain treatment ranged from 6.5% (Amikacin) to 70.6% (Ofloxacin). Among the studies that had the resistance information, the number of patients who were resistant to a certain treatment ranged from 480 (5.2%) to 4153 (44.7%).

7.2 Results of Analysis

As discussed in Chapter 4, out of 9290 patients, 9258 subjects in 31 studies were analyzed after excluding observations with missing sex or age data.

Table 7–2: The number and proportion of 9290 patients who used and are resistant to the 15 treatments, respectively.

Treatment	**N. used (%)**	**N. resistant (%)**
High-generation quinolones	930 (10.0)	
Ciprofloxacin	1031 (11.1)	
Protionamide	3341 (36.1)	
Amikacin	605 (6.5)	
Cycloserine	5729 (61.9)	480 (5.2)
Kanamycin	5093 (55.0)	1821 (19.6)
Ofloxacin	6538 (70.6)	
Group 5 level drugs	2205 (23.8)	
Pyrazinamide	6263 (67.6)	2771 (29.8)
Ethambutol	4325 (46.7)	4130 (44.5)
Para-Aminosalicylic Acid	4005 (43.3)	1092 (11.8)
Capreomycin	1956 (21.1)	630 (6.8)
Ethionamide/Prothionamide	4005 (43.3)	1763 (19.0)
Streptomycin	1418 (15.3)	4153 (44.7)
Rifabutin	1371 (14.8)	

Following the procedures described in Chapter 5, the treatment importance of all 15 treatments was estimated with the TMLE algorithm, for which SuperLearner was used to estimate the outcome model, and logistic regression and LASSO were used to fit the treatment model. Table 7-3 shows the estimates, ordered by treatment importance, with their associated influence curve based standard errors and confidence intervals.

As shown in the table, Ciprofloxacin had the greatest estimated treatment importance, with a confidence interval not containing zero. Other treatments with at least moderately large positive treatment importance were Amikacin, later-generation quinolones, Capreomycin, Ethionamide, Streptomycin, and Cycloserine. Among these, hypothesis tests for the treatment importance of Streptomycin and Cycloserine also rejected the null where the null corresponded with (zero importance). Therefore, we may infer that Ciprofloxacin would contributed the most to the average recovery rate for treating MDR-TB in the context of multiple treatment use, and there is evidence that Streptomycin and Cycloserine both have moderately positive treatment importance. In contrast, Para-aminosalicylic acid, Pyrazinamide and Group 5 level drugs (defined in section 4.1.2) had negative importance in our analysis. Note that the treatment importance of Para-aminosalicylic acid and Pyrazinamide were only slightly less than zero, while the estimate for Group 5 level drugs was substantially more negative. The estimated standard errors, based on the efficient influence function, for these treatment importance estimates ranged from 0.017 to 0.102.

Table 7–3: Treatment importance, associated standard error and confidence interval of 15 treatment.

Treatment	**Estimate**	**Standard Error**	**Confidence Interval**
Ciprofloxacin	0.134	0.056	(0.024, 0.243)
Amikacin	0.091	0.056	(-0.018, 0.200)
High-generation quinolones	0.084	0.102	(-0.116, 0.283)
Capreomycin	0.070	0.064	(-0.055, 0.196)
Ethionamide	0.068	0.040	(-0.011, 0.147)
Streptomycin	0.063	0.027	(0.011, 0.116)
Cycloserine	0.054	0.028	(0.000, 0.109)
Prothionamide	0.047	0.154	(-0.255, 0.348)
Ofloxacin	0.023	0.032	(-0.040, 0.085)
Ethambutol	0.020	0.022	(-0.022, 0.063)
Kanamycin	0.020	0.024	(-0.027, 0.067)
Rifabutin	0.014	0.071	(-0.125, 0.153)
Para-aminosalicylic acid	-0.002	0.019	(-0.038, 0.035)
Pyrazinamide	-0.005	0.017	(-0.038, 0.028)
Group 5 level drugs	-0.035	0.037	(-0.108 ,0.038)

CHAPTER 8
Discussion

In this chapter, we aim to summarize the work that has been done in this thesis. In addition, we elaborate on the contribution of this thesis, acknowledge the limitations of our methods, and discuss future directions.

8.1 Summary of the work

The literature review introduced the current global MDR-TB situation. Then, we briefly described meta-analytical methods, from traditional aggregated meta-analysis to individual patient data network meta-analysis. Next, we discussed some basic concepts in causal inference, most notably, propensity scores and Inverse Probability of Treatment Weighting, which are related to the TMLE algorithm. Super-Learner, an ensemble statistical learning prediction method used in TMLE, was introduced before describing the semiparametric theory behind TMLE. Several important definitions were explained, including local efficiency and the efficient influence curve of regular asymptotically linear estimators, which laid the foundation for deriving the TMLE estimator and its asymptotic variance. An example TMLE algorithm was described for the estimation of the average treatment effect with a binary exposure, and its properties were discussed. Finally, a brief overview of transportability was given, which paved the way for developing the methods used in this thesis.

The MDR-TB data were introduced and summarized in Chapter 4. After elaborating on the source of the data, we described the structure of the data (IPD) in terms of the studies, covariates, and treatments involved.

Then, the application of our methods was illustrated for the data set described above. After defining the data structure, we presented the assumptions that are required in order to identify the treatment importance parameter of interest. In addition, we described a closely related causal parameter and its required identifiability assumptions. Furthermore, the TMLE algorithm and variance estimation for the MDR-TB data were described.

A simulation study was conducted to demonstrate the consistency of our estimator and the validity of our variance estimation. We found that our proposed estimator can precisely target the parameter of interest without bias if all relevant models are correctly specified. Influence curve based variance estimates performed similarly to the variance estimated using cluster bootstrap (with a 3% difference) when all models were correctly specified. In addition, the double robustness property of our estimator was verified by modifying the model specifications. Two scenarios were generated: independent and dependent treatment availability. In both scenarios, the estimates were biased only if both the outcome and at least one component of the treatment models were misspecified. Finally, the coverage of the estimator was presented; the coverage rate was 92% when the number of clusters was 30 with 300 subjects in each cluster. When we increased the sample size (60 clusters with 600 subjects in each cluster), the coverage was approximately 95%.

Results from the MDR-TB data were described in chapter 7. In this section, we also

discussed the explanation and interpretation of our proposed parameter of interest – treatment importance. Based on our analysis, Ciprofloxacin is the antibiotic that has the greatest treatment importance in a non-resistant population, which suggests that it is a more effective treatment for MDR-TB. Other antibiotics that were found to have a relatively large treatment importance were Amikacin, later-generation quinolones, Capreomycin, Ethionamide, Streptomycin. On the other hand, based on our analysis, Para-aminosalicylic acid, Pyrazinamide and Group 5 level drugs should not be recommended in lieu of other antibiotics, unless they must be used for certain indications. The ranking of treatment importance is shown in Table 7-3, which may be useful for clinicians and may guide future MDR-TB treatment research.

The results are not entirely in line with the findings of The Collaborative Group for Meta-Analysis of Individual Patient Data in MDR- TB [11], which used a similar data extraction to analyze which antibiotics as part of multi-drug regimens were associated with treatment success compared to failure, relapse or death during the initial intensive treatment phase. They showed the use of later generation Quinolones, Ofloxacin, Ethionamide/Prothionamide and Group 4 level drugs were more likely effective. These antibiotics all had positive treatment importance in our analysis. However, we focused on a different parameter of interest, and we focused on which antibiotics contributed more to the average recovery rate (see Section 5.3), whereas their study was interested in which drugs were associated with the treatment success. We took other treatments' effects and the overall recovery rate into account, whereas

their study only explored the correlation between the individual treatments and outcomes. Finally, we also adjusted for censoring, resistance status, and the selection bias that may arise due to some treatments not being available in some studies.
Another study [28] conducted by this group showed that there was no improvement in treatment success among patients taking Group 5 drugs, while our study showed that Group 5 drugs ranked as the last in treatment importance.

8.2 Contribution & Limitations

Our analysis – in contrast to a traditional meta-analysis that targets more frequently used parameters (e.g., risk differences or odds ratios between treatment pairs) – investigated the treatment importance [53, 55] of the 15 antibiotic types observed across studies of the treatment of MDR-TB.
Instead of analyzing the data study by study and then aggregating the results [10, 67], we transported statistical information from available studies to unavailable studies by incorporating the concepts of transportability and treatment availability (see Section 2.7 and 5.1). We used the information from patients who used the treatment of interest to analyze the relationships between covariates and outcome, and then extrapolate these relationships to patients who did not use the treatment to obtain a global estimate.

8.2.1 Parameter of Interest

Treatment importance, as we have defined it, is an interpretable quantity with a calculation-friendly estimator. As we discussed in Chapter 5, treatment importance

can be simply formulated as $\tau^k - \mu^k$ (where τ^k represents the adjusted recovery rate when all the patients using the treatment k, and μ^k is the overall recovery rate) for treatment k under the set of reasonable assumptions we made. It is roughly analogous to the coefficients of each treatment in a linear regression that adjusts for all treatments and baseline covariates. We chose to estimate this parameter because of the challenges in dealing with multiple treatment regimens in a causal inference setting. As we described, making the causal assumption of variable ordering (i.e. that the treatment follows covariates in time) when individually evaluating multiple treatments lacked of validity.

Our parameter of interest was described in the literature [53, 55], and defined in our context. Rather than the risk difference between two treatments, treatment importance is roughly the difference between the recovery rate if all the patients used the treatment and the overall recovery rate.

Two challenges inherent to our data when we estimate treatment importance were treatment resistance and the simultaneous use of multiple treatments. We analyzed each treatment in the stratum of patients who were not resistant to that treatment and considered all other treatments as confounders.

This parameter can then be interpreted as the contribution of a treatment to the average recovery rate when everyone uses this treatment. The larger the treatment importance, the more effective the drug is as an add-on medication. A medication with a large importance should therefore be more commonly recommended by clinicians.

Additional causal assumptions are needed to interpret this parameter as an add-on

effect. In either case, the parameter can be estimated through the same TMLE procedure. However, we do not believe these additional assumptions to be satisfied in this context. The identifiability assumptions are laid out in section 5.4.
Furthermore, it is true that we could alternatively define treatment importance for treatment k simply as the average recovery rate had all patients taken medication k. However, note that the "baseline" for each antibiotic (the average recovery rate where there is no treatment given) was different since we condition on not being resistant to the treatment, and the collection of patients resistant to each drug is distinct. Therefore, targeting this parameter would not yield comparable estimates, and the resulting treatment importance ranking would be meaningless.

8.2.2 Missing Data

When we analyzed the data, we took a pragmatic approach to dealing with missing covariates. We completely removed observations with missing data in sex and age because the missingness was rare (0.3% and 0.06%, respectively). We accounted for missing data in other covariates (ranging from 5.6% to 27.1% missing), by introducing missing indicators. This approach allowed for unbiased estimation when conditional exchangeability was satisfied conditional on the incomplete covariates in addition to the missingness indicators which is sometimes considered to be implausible [82]. A more robust approach might involve producing multiple imputations prior to performing the TMLE analysis on each completed data set [93, 92]. However, we did adjust for the missing (or censored) outcome data, as the g_3^k model considered the probability of censoring given treatment, confounders and resistance

status. This approach is valid under the outcome missingness at random assumption, as described in the list of assumptions.

8.2.3 Adjustment Variables and Outcomes

We had the country of each study as one of the adjustment variables, which is a study-level variable in our analysis. Including it as a country-specific indicator covariate would have produced a large number of levels if we included them all. Therefore, we categorized countries based on the income group, resulting in only three levels. The other study-level covariate was published year of the studies. There are many possible outcomes of a given treatment assignment, including completing or not completing treatment, cure, death, and relapse. For the convenience of analysis, we defined treatment success as completing treatment and being cured of the disease, and considered all other outcomes as unsuccessful. We could have investigated other outcome definitions as well, but the priority of this thesis was to provide a framework for obtaining treatment importance estimates in this type of data. Nearly identical analyses could be repeated on the differently defined outcomes.

8.2.4 Transportability

Transportability in this situation involved two main elements: 1) the generalizability of the treatment-specific outcome models, which were estimated in studies in which the treatments were observed, and used to predict mean counterfactual outcomes in studies where the treatments were not observed, and 2) the appropriateness of estimating the probability at treatment availability of the study level. Transportability, in our specific context of data fusion, enabled us to obtain a global estimate

with the assumptions described in Section 5.2.

8.2.5 TMLE

To achieve robust estimation of treatment importance parameters, we made use of the TMLE algorithm [53] in tandem with the concept of transportability [73].
One of the advantages of using TMLE is that it is a semiparametric method, which allowed us to avoid some strong assumptions necessary for parametric models [86]. In addition, it has the very favorable double robustness and local efficiency properties. With the aid of SuperLearner, TMLE can provide precise estimates with lower variance [86, 64, 85].
SuperLearner was preferred over the parametric regression approach partially because of its ability to adapt to the small numbers of events that occurred in the analysis of certain treatments.

8.2.6 Models

In the estimation of the treatment and outcome models, we did not use a clustered model such as a mixed model. SuperLearner is flexible in that it may place larger weights on more complex models when the sample size and number of events are greater, and may exclude models that are too complex (and may overfit) in these settings. While the prediction models used within the SuperLearner did not respect clustering, the cluster-specific cross-validation used to compute the model weights did.
In terms of the study-level model $g_2^{\tau k}$ for the probability of treatment availability,

to avoid the problem of rank deficiency, we assumed that given the study-level covariates, individual-level covariates were independent of treatment availability. We believe that this is plausible because treatment availability (in a given region in a certain time period) was also a study-level variable, and therefore each patient in a given study shared the same treatment availability. We used logistic regression for this model in order to avoid overprediction by a more flexible model and maintain the stability of the TMLE. Because of the small sample size for this model, machine learning methods might produce extreme values of estimation of the g model [75]. We used LASSO as an exception when the logistic regression did not converge.

8.2.7 Simulation Studies

The simulation studies we performed demonstrated the double robustness of our estimator as well as the appropriate coverage rate of the sandwich variance estimator based on the efficient influence function. We also showed the consistency of both the sandwich and cluster bootstrap variance estimators. Therefore, we only used the sandwich estimator in our data analysis as it was far less computationally complex than running a bootstrap. Furthermore, the necessity of considering clustering when estimating the variance was also briefly demonstrating in the simulated studies.

8.2.8 Confounding by Indication

We adjusted for many variables that describe the severity of TB (whether the patients had prior TB for example), but confounding by indication may still exist in the models we estimated. This type of confounding arises from the fact that patients

who are given a medication are inherently different from those who are not [91]. This bias can be eliminated by using a RCT, but our data was obtained from an observational study.

8.2.9 Heterogeneity Across Studies

We included 31 studies in the data. However, there might exist heterogeneity across the studies. We transported the estimation from available studies to unavailable studies, assuming the studies were independent of each other. We did not include some algorithms in the library of SuperLearner due to the data constraint and the limitation of SuperLearner. We could develop a wrapper, such like generalized estimating equation or generalized linear mixed model in SuperLearner that can take heterogeneity across studies into account in the future.

8.3 Future work

There are many topics that may be worth exploring following this thesis. This study is a fundamental step in understanding which antibiotics, on average, are important in the treatment of MDR-TB. However, we may also be interested in knowing which antibiotics would be most effective for a specific patient given some measured covariates. We would benefit greatly from understanding which patient characteristics affect treatment success for different combinations of antibiotics, and which treatment strategies would maximize the success of treatment for MDR-TB. To accomplish this goal, we may develop more complicated models, such as a propensity score model for multiple treatment regimes, models that can automatically select

the covariates for different regimes, and ensemble learning methods for obtaining the optimal regime. These promising extensions could advance our knowledge of data fusion and multiple treatment issues. The clinical contribution of our analysis is important as well. It may provide insight for clinicians to tailor the choice of treatments based on patient characteristics. More ambitiously, the identification of optimal treatment regimens may lead to yet more successful patient outcomes. Future work can be done by investigating which regimen is more effective based on various patients' characteristics, then clinicians can prescribe appropriate regimens for different patients. In addition, if we know which characteristics are more important for the recovery of MDR-TB (the treatment duration or the adherence of patients for example), the clinicians can then make efforts to intervene on them.

APPENDIX A
Appendix 1

The code of generating the data in scenario 1

```
# Generate the data--Simulation
# Scenario 1: D is independent
  N=9000;n=300;clusterN=N/n
  study_id<-c(rep(1:clusterN,each=n))
  ind_id<-c(1:N)
  v1_bar<-rnorm(clusterN,mean=0.3,sd=0.3)
  v1<-rep(v1_bar,each=n)
  #r1=1-->resistant to a1
  r1<-rbinom(N,1,0.25)
  r2<-rbinom(N,1,0.30)
  r3<-rbinom(N,1,0.25)
  D <-NULL
  for (i in 1:clusterN)
  {
    #d_index: how many medications are accessible
    d_index <- sample(x=c(1:3),size=1)
    d_zero <- c(0,0,0)
```

```
  #d: fill 1 into c(0,0,0), sample the positions of 1s with the
   number of accessible medications
  d <- replace(d_zero,sample(c(1:3),d_index), 1)
  D <- rbind(D, d)
}
d1_bar<-as.numeric(D[,1])
d1<-rep(d1_bar,each=n)
d2_bar<-as.numeric(D[,2])
d2<-rep(d2_bar,each=n)
d3_bar<-as.numeric(D[,3])
d3<-rep(d3_bar,each=n)
w1<-rnorm(N,mean=0.1*v1,sd=0.1)
#u:study-level radom unmeasured confounding
u_bar<-rnorm(clusterN,mean=0.2,sd=0.6)
u<-rep(u_bar,each=n)

library(faraway)
a1_full<-rbinom(N,1,ilogit(-0.75+2.4*v1+1.8*w1-0.1*r1))
a1<-ifelse(d1==1,a1_full,0)
a2_full<-rbinom(N,1,ilogit(-1+1*v1+1.7*w1-0.15*r2))
a2<-ifelse(d2==1,a2_full,0)
a3_full<-rbinom(N,1,ilogit(-1.5+1.7*v1+1*w1-0.16*r3))
a3<-ifelse(d3==1,a3_full,0)
```

```
y0<-rbinom(N,1,ilogit(-2+3.5*v1+0.3*w1-0.005*u
            +2.2*a1*(1-r1)+0.12*a2*(1-r2)+0.05*a3*(1-r3)))

data_full<-data.frame(ind_id,study_id,v1,r1,r2,r3,d1,d2,d3
                        ,w1,u,a1,a2,a3,y0)
BSD=data_full
```

APPENDIX B
Appendix 2

the code of generating the data in scenario 2

```
# Generate the data--Simulation
# Scenario 2: D is dependent of V
  N=9000;n=300;clusterN=N/n
  study_id<-c(rep(1:clusterN,each=n))
  ind_id<-c(1:N)
  v1_bar<-rnorm(clusterN,mean=0.3,sd=0.3)
  v1<-rep(v1_bar,each=n)
  #r1=1-->resistant to a1
  r1<-rbinom(N,1,0.25)
  r2<-rbinom(N,1,0.30)
  r3<-rbinom(N,1,0.25)

  d1_bar <- rbinom(clusterN,1,ilogit(1+2*v1_bar))
  d2_bar <- rbinom(clusterN,1,ilogit(0.5+1.5*v1_bar))
  d3_bar <- rbinom(clusterN,1,ilogit(1.5+0.3*v1_bar))
  D <- data.frame(d1_bar,d2_bar,d3_bar)
  D <- D [sum(d1_bar+d2_bar+d3_bar)!=0,]
```

```
d1_bar<-as.numeric(D[,1])
d1<-rep(d1_bar,each=n)
d2_bar<-as.numeric(D[,2])
d2<-rep(d2_bar,each=n)
d3_bar<-as.numeric(D[,3])
d3<-rep(d3_bar,each=n)

w1<-rnorm(N,mean=0.1*v1,sd=0.1)
#u:study-level radom unmeasured confounding
u_bar<-rnorm(clusterN,mean=0.2,sd=0.6)
u<-rep(u_bar,each=n)

library(faraway)
a1_full<-rbinom(N,1,ilogit(-0.75+2.4*v1+1.8*w1-0.1*r1))
a1<-ifelse(d1==1,a1_full,0)
a2_full<-rbinom(N,1,ilogit(-1+1*v1+1.7*w1-0.15*r2))
a2<-ifelse(d2==1,a2_full,0)
a3_full<-rbinom(N,1,ilogit(-1.5+1.7*v1+1*w1-0.16*r3))
a3<-ifelse(d3==1,a3_full,0)

y0<-rbinom(N,1,ilogit(-2+3.5*v1+0.3*w1-0.005*u
           +2.2*a1*(1-r1)+0.12*a2*(1-r2)+0.05*a3*(1-r3)))
```

```
data_full<-data.frame(ind_id,study_id,v1,r1,r2,r3,d1,d2,d3
                      ,w1,u,a1,a2,a3,y0)
BSD=data_full
```

APPENDIX C
Appendix 3

The code for model fitting (all models are correctly specified in scenario 2)

```
seeds<-as.vector(read.table("sim_seed.txt"))
results<-NULL
for (i in 1:1000){
  set.seed(seeds[i,])
  source("Generate_S2_30.R") #the code in Appendix 2
#################### g1   ##########
data.g1.a1<-data.frame(BSD$a1,BSD$a2,BSD$a3,BSD$v1,
                       BSD$w1,BSD$r1,BSD$d1)
g1.a1.glm<-glm(BSD.a1~BSD.v1+BSD.w1+BSD.a2+BSD.a3,
                       family="binomial",
                       data=data.g1.a1[which(BSD$d1==1  &
g1.a1.pre<-predict(g1.a1.glm,newdata=data.frame(BSD$v1,BSD$w1,
#################### g2   ##########
g2.a1.glm<-glm(d1_bar~v1_bar,family="binomial")
g2.a1.pre<-predict(g2.a1.glm,type="response")
g2.a1.pre<-rep(g2.a1.pre,each=n)
g.a1.pre_all<-g1.a1.pre*g2.a1.pre
```

```
g.a1.pre<-g.a1.pre_all[which(BSD$r1==0)]
##################### Q ###########
interaction_a2<-BSD$a2*(1-BSD$r2)
interaction_a3<-BSD$a3*(1-BSD$r3)
data.Q1.a1<-data.frame(BSD$y0,BSD$a1,BSD$a2,BSD$a3,BSD$v1,
newdata<-data.frame(BSD$v1,BSD$w1,interaction_a2,interaction_a3)
Q1.a1.glm<-glm(BSD.y0~BSD.v1+BSD.w1+interaction_a2+interaction_a3,
Q1.a1.pre<-predict(Q1.a1.glm,newdata=newdata,type="response")
##################### TMLE #################
#(1)
y.a1 <- BSD$y0[BSD$r1==0]
h.a1<-as.numeric(as.logical(BSD$a1==1)[BSD$r1==0])/g.a1.pre
e.glm.a1<-glm(y.a1~-1+h.a1,offset=logit(Q1.a1.pre),family="binomial")
e.a1<- coef(e.glm.a1)["h.a1"]
#(2)
up.logitQ.a1<-logit(Q1.a1.pre)+(e.a1/g.a1.pre)
up.Q.a1<-ilogit(up.logitQ.a1)

#(3)
tau.Q.a1<-mean(up.Q.a1)

######### SD estimation ##############
diff.yQ.a1<- y.a1-up.Q.a1
```

```
ic_a1<-h.a1*diff.yQ.a1+up.Q.a1-tau.Q.a1
ic_a1.study<-split(ic_a1,study_id[BSD$r1==0])
var.san.a1<-sum(unlist(lapply(ic_a1.study,FUN=function(x)
                (length(BSD$a1[which(BSD$r1==0)])[1])^2
sd.san.a1<-sqrt(var.san.a1)
ci.san.a1<-c(tau.Q.a1-1.96*sqrt(var.san.a1),tau.Q.a1
                +1.96*sqrt(var.san.a1))
#coverage<-as.numeric(as.logical(0.741>ci.san.a1[1])&0.741
                <ci.san.a1[2])
result<-c(tau.Q.a1,sd.san.a1)
results<-rbind(results,result)
}
devout=paste("TMLE_Q_correct_g_correct_30.csv", sep="")
write.csv(results, devout, row.names = F)
#mean(est)
mean(results[,1],na.rm=TRUE)
#mean(sandwish estimator SD)
mean(results[,2],na.rm=TRUE)
#Monte-Carlo SD
sd(results[,1],na.rm=TRUE)
```

APPENDIX D
Appendix 4

Code for model fitting (all the models are misspecified in scenario 2)

```
seeds<-as.vector(read.table("sim_seed.txt"))
results<-NULL
for (i in 1:1000){
  set.seed(seeds[i,])
  source("Generate_S2_30.R") #the code in Appendix 2
#################### g1   ##########
data.g1.a1<-data.frame(BSD$a1,BSD$a2,BSD$a3,BSD$v1,
                       BSD$w1,BSD$r1,BSD$d1)
g1.a1.glm<-glm(BSD.a1~1,family="binomial",
               data=data.g1.a1[which(BSD$d1==1  & BSD$r1==0),])
g1.a1.pre<-predict(g1.a1.glm,newdata=data.frame(BSD$v1,BSD$w1,
               BSD$a2,BSD$a3),type="response")
#################### g2   ###########
g2.a1.glm<-glm(d1_bar~v1_bar,family="binomial")
g2.a1.pre<-predict(g2.a1.glm,type="response")
g2.a1.pre<-rep(g2.a1.pre,each=n)
g.a1.pre_all<-g1.a1.pre*g2.a1.pre
```

```
g.a1.pre<-g.a1.pre_all[which(BSD$r1==0)]
##################### Q ###########
data.Q1.a1<-data.frame(BSD$y0,BSD$a1,BSD$a2,BSD$a3,BSD$v1,
                       BSD$w1,BSD$r1,BSD$r2,BSD$r3,
                      BSD$study_id)[which(BSD$r1==0 & BSD$a1==1),]
newdata<-data.frame(BSD$a2,BSD$a3)[which(BSD$r1==0),]
Q1.a1.glm<-glm(BSD.y0~1,family="binomial",data=data.Q1.a1)
Q1.a1.pre<-predict(Q1.a1.glm,newdata=newdata,type="response")
##################### TMLE #################
#(1)
y.a1 <- BSD$y0[BSD$r1==0]
h.a1<-as.numeric(as.logical(BSD$a1==1)[BSD$r1==0])/g.a1.pre
e.glm.a1<-glm(y.a1~-1+h.a1,offset=logit(Q1.a1.pre),
                family="binomial")
e.a1<- coef(e.glm.a1)["h.a1"]
#(2)
up.logitQ.a1<-logit(Q1.a1.pre)+(e.a1/g.a1.pre)
up.Q.a1<-ilogit(up.logitQ.a1)

#(3)
tau.Q.a1<-mean(up.Q.a1)

######### SD estimation ##############
```

```
diff.yQ.a1<- y.a1-up.Q.a1
ic_a1<-h.a1*diff.yQ.a1+up.Q.a1-tau.Q.a1
ic_a1.study<-split(ic_a1,study_id[BSD$r1==0])
var.san.a1<-sum(unlist(lapply(ic_a1.study,FUN=function(x)
        sum(x %*% t(x)))))/
        (length(BSD$a1[which(BSD$r1==0)])[1])^2
sd.san.a1<-sqrt(var.san.a1)
ci.san.a1<-c(tau.Q.a1-1.96*sqrt(var.san.a1),tau.Q.a1+
                1.96*sqrt(var.san.a1))
#coverage<-as.numeric(as.logical(0.741>
        ci.san.a1[1])&0.741<ci.san.a1[2])
result<-c(tau.Q.a1,sd.san.a1)
results<-rbind(results,result)
}
devout=paste("/TMLE_Q_x_g1xg2v_30.csv", sep="")
write.csv(results, devout, row.names = F)
#mean(est)
mean(results[,1],na.rm=TRUE)
#mean(sandwish estimator SD)
mean(results[,2],na.rm=TRUE)
#Monte-Carlo SD
sd(results[,1],na.rm=TRUE)
```

References

[1] CDC basic tb facts. https://www.cdc.gov/tb/topic/basics/default.htm. 2016 updated.

[2] Mdr-tb fact sheet. http://www.who.int/tb/challenges/mdr/mdr_tb_factsheet.pdf. 2016 updated.

[3] Tbfacts.org–treatment of drug resistant tb. http://www.tbfacts.org/treatment-of-drug-resistant-tb/. Page last updated: November 21.

[4] What is multidrug-resistant tuberculosis (mdr-tb) and how do we control it? http://www.who.int/features/qa/79/en/. 2016 updated.

[5] Who drug-resistant tb: global situation. http://www.who.int/tb/areas-of-work/drug-resistant-tb/global-situation/en/. 2016 updated.

[6] Who media center: Tuberculosis fact sheets. http://www.who.int/mediacentre/factsheets/fs104/en/. 2016 updated.

[7] World bank country and lending groups. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. 2017 updated.

[8] Jadad A. *Randomised controlled trials*. BMJ London, 1998.

[9] Tsiatis A. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[10] Haidich A.B. Meta-analysis in medical research. *Hippokratia*, 14(1):29–37, 2011.

[11] Ashkin D. Avendano-M. Banerjee R. Bauer M. Bayona J.N. Becerra M.C.-Benedetti A. Burgos M. Centis R. Ahuja, S.D. et al. Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: an individual patient data meta-analysis of 9,153 patients. *PLoS med*, 9(8):e1001300, 2012.

[12] Pearl J. Bareinboim, E. Transportability from multiple environments with limited experiments: Completeness results. *Advances in Neural Information Processing Systems*, pages 280–288, 2014.

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

[14] Pearl J. Bareinboim, E. et al. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1(1):107–134, 2013.

[15] Ensor J. Riley-R.D. Burke, D.L. 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.

[16] Ades A.E. Higgins-J. Caldwell, D.M. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ: British Medical Journal*, 331(7521):897, 2005.

[17] Gelbach J.B. Cameron, A.C. and L. Miller, Douglas. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

[18] Frangakis C.E. Cole, S.R. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.

[19] Chaulet P. Maher-D. Grosset J. Harris W. Horne N. Iseman M.-Watt B. Crofton, S.J. et al. Guidelines for the management of drug-resistant tuberculosis. 1997.

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

[21] Rubin D.B. Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[22] Moons K.GM. Abo-Zaid G.M.A. Koffijberg H. Riley R.D. Debray, T.PA. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PloS one*, 8(4):e60650, 2013.

[23] Schuit E. Efthimiou-O. Reitsma J.B. Ioannidis J.PA. Salanti G. Moons K.GM. Debray, T.PA. et al. An overview of methods for network meta-analysis using

individual participant data: when do benefits arise? *Statistical Methods in Medical Research*, page 0962280216660741, 2016.

[24] Ades A.E. Dias, S. Absolute or relative effects? arm-based synthesis of trial data. *Research synthesis methods*, 2015.

[25] Staessen J.A. Thijs-L. Fagard, R.H. Advantages and disadvantages of the meta-analysis approach. *Journal of Hypertension*, 14:S9–S13, 1996.

[26] Gandhi N. Migliori-G.B. Sotgiu G. Cox H. Holtz T.H. Hollm-Delgado-MG. Keshavjee S. DeRiemer K. Centis R. Falzon, D. et al. Resistance to fluoroquinolones and second-line injectable drugs: impact on mdr-tb outcomes. *European Respiratory Journal*, pages erj01347–2012, 2012.

[27] Centers for Disease Control, Prevention (CDC, et al. Emergence of mycobacterium tuberculosis with extensive resistance to second-line drugs–worldwide, 2000-2004. *MMWR. Morbidity and mortality weekly report*, 55(11):301, 2006.

[28] Benedetti A. Cox-H. Koh W.J. Viiklepp P. Ahuja S. Pasvol-G.-Menzies D. Fox, G..J. Group 5 drugs for multidrug-resistant tuberculosis: individual patient data meta-analysis. *European Respiratory Journal*, 49(1):1600993, 2017.

[29] Benedetti A. Mitnick-C.D. Pai M. Menzies D. Collaborative Group for Meta-Analysis of Individual Patient Data in MDR-TB Fox, G.J. et al. Propensity score-based approaches to confounding by indication in individual patient data meta-analysis: non-standardized treatment for multidrug resistant tuberculosis. *PloS one*, 11(3):e0151724, 2016.

[30] Hastie T. Tibshirani-R. Friedman, J. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[31] Wellner J.A. Præstgaard-J. Gill, R.D. Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics*, pages 97–128, 1989.

[32] Robins J.M. Greenland, S. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986.

[33] Mackenzie F.J. Grossman, J. The randomized controlled trial: gold standard, or merely standard? *Perspectives in biology and medicine*, 48(4):516–534, 2005.

[34] Scott D.A. Woods-B. Hawkins, N. Arm-based parameterization for network meta-analysis. *Research synthesis methods*, 2015.

[35] Robins J.M. Hernán, M.A. Causal inference. `https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2016/09/hernanrobins_v1.10.31.pdf`. Page last updated: 2016-09-11.

[36] Robins J.M. Hernán, M.A. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health*, 60(7):578–586, 2006.

[37] Thompson S.G. Higgins, J. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558, 2002.

[38] Whitehead A. Higgins, J. Borrowing strength from external trials in a meta-analysis. *Statistics in medicine*, 15(24):2733–2749, 1996.

[39] Ross R.R. Izzo, R.L. Meta-analysis of rehabilitation programs for juvenile delinquents: A brief report. *Criminal Justice and Behavior*, 17(1):134–142, 1990.

[40] Higgins J. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International journal of epidemiology*, 37(5):1158–1160, 2008.

[41] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period?application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

[42] Riley R. White-I.R. Jackson, D. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

[43] E Jaramillo. *Guidelines for the programmatic management of drug-resistant tuberculosis:[emergency update 2008]*. Number 616-002.5. Stop TB Initiative (OMS) Gran Bretaña. Estados Unidos., 2008.

[44] Shahidi N.C. Mohsen-S. Fitzgerald J.M. Johnston, J.C. Treatment outcomes of multidrug-resistant tuberculosis: a systematic review and meta-analysis. *PloS one*, 4(9):e6914, 2009.

[45] Krahn U. Binder-H. König, J. Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons. *Statistics in medicine*, 32(30):5414–5429, 2013.

[46] Noeske J. Rieder-H.L. Ait-Khaled N. Abena Foe J.L. Trébucq A. Kuaban, C. High effectiveness of a 12-month regimen for mdr-tb patients in cameroon. *The International Journal of Tuberculosis and Lung Disease*, 19(5):517–524, 2015.

[47] Song X. Gray-RH. Li, X. Comparison of the missing-indicator method and conditional logistic regression in 1: m matched case-control studies with missing exposure values. *American journal of epidemiology*, 159(6):603–610, 2004.

[48] Wiener M. Liaw, A. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[49] Heidenreich W.F. Li-G. Little, M.P. Parameter identifiability and redundancy: theoretical considerations. *PloS one*, 5(1):e8915, 2010.

[50] Ades A.E. Lu, G. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in medicine*, 23(20):3105–3124, 2004.

[51] Kuderer N.M. Lyman, G.H. The strengths and limitations of meta-analyses based on aggregate data. *BMC medical research methodology*, 5(1):14, 2005.

[52] Schnitzer M. Targeted maximum likelihood estimation for longitudinal data. `http://digitool.Library.McGill.CA:80/R/-?func=dbin-jump-full&object_id=114242&silo_library=GEN01`, 2013. Ph.D. thesis, McGill University, Canada.

[53] Van der Laan M.J. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

[54] Jones M.P. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230, 1996.

[55] Oswald F.L. Nimon-K. Nathans, L.L. Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, 17(9), 2012.

[56] Nunn P. Uplekar-M. Floyd-K. Jaramillo E. Lönnroth K. Weil D.-Raviglione M. Nathanson, E. Mdr tuberculosis-critical steps for prevention and control. *New England Journal of Medicine*, 363(11):1050–1058, 2010.

[57] Chaimani A. Veroniki-A.A. Vasiliadis-H.S. Schmid C.H. Salanti G. Nikolakopoulou, A. Characteristics of networks of interventions: a description of a database of 186 published networks. *PLoS One*, 9(1):e86754, 2014.

[58] Basu S. Shah-N.S. Andrews-J.R. Friedland G.H. Moll A.P. Gandhi N.R.-Galvani A.P. Orenstein, E.W. Treatment outcomes among patients with multidrug-resistant tuberculosis: systematic review and meta-analysis. *The Lancet infectious diseases*, 9(3):153–161, 2009.

[59] Mukhopadhyay P. *An introduction to estimating functions*. Alpha Science Int'l Ltd., 2004.

[60] McCulloch M. Gorman-J.D. Pai-N. Enanoria W. Kennedy G. Tharyan P.-Colford J John M. Pai, M. Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National medical journal of India*, 17(2):86–95, 2003.

[61] Bareinboim E. Pearl, J. Transportability of causal and statistical relations: A formal approach. pages 540–547, 2011.

[62] Bareinboim El. Pearl, J. et al. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

[63] LeDell E. Kennedy-C. Van der Laan M.J. Polley, E. *SuperLearner: Super Learner Prediction*.

[64] Van der Laan-M.J. Polley, E.C. Super learner in prediction. 2010.

[65] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[66] Glas A.S. Rutjes-A.WS. Scholten R.JPM. Bossuyt P.M. Zwinderman A.H. Reitsma, J.B. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of clinical epidemiology*, 58(10):982–990, 2005.

[67] Lambert P.C. Abo-Zaid G. Riley, R.D. Meta-analysis of individual participant data: rationale, conduct, and reporting. *Bmj*, 340:c221, 2010.

[68] Takwoingi Y. Trikalinos-T. Guha A. Biswas A. Ensor J. Morris R.K.-Deeks J.J. Riley, R.D. Meta-analysis of test accuracy studies with multiple and missing

thresholds: a multivariate-normal model. *Journal of Biometrics & Biostatistics*, 5(3):1, 2014.

[69] Greenland S. Robins, J. The probability of causation under a stochastic model for individual risk. *Biometrics*, pages 1125–1138, 1989.

[70] Rubin D.B. Rosenbaum, P.R. The central role of the propensity score in observational studies for causal effects. *Biometrika*, pages 41–55, 1983.

[71] Robins J.M. Scharfstein-D.O. Rotnitzky, A. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339, 1998.

[72] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[73] Van der Laan-M.J. Rudolph, K.E. Double robust estimation of encouragement-design intervention effects transported across sites. 2015.

[74] Van der Laan-M.J. Rudolph, K.E. Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

[75] Lok J.J. Gruber-S. Schnitzer, M.E. Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *The international journal of biostatistics*, 12(1):97–115, 2016.

[76] Centis R. DÁmbrosio-L. Alffenaar JW.C. Anger H.A. Caminero J.A. Castiglia P.-De Lorenzo S. Ferrara G. Koh WJ. Sotgiu, G. et al. Efficacy, safety and tolerability of linezolid containing regimens in treating mdr-tb and xdr-tb: systematic review and meta-analysis. *European Respiratory Journal*, 40(6):1430–1442, 2012.

[77] Parmar M.KB. Stewart, L.A. Meta-analysis of the literature or of individual patient data: is there a difference? *The Lancet*, 341(8842):418–422, 1993.

[78] Tierney J.F. Stewart, L.A. 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.

[79] Norman G. Trop-I. Stolberg, H.O. Fundamentals of clinical research for radiologists. *AJR*, 183:1539–1544, 2004.

[80] Van der Laan-M.J. Susan, G. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1):1–18, 2010.

[81] Benedetti A. Kloda-L.A. Levis B. Nicolau I. Cuijpers P. Gilbody S.-Ioannidis J.PA. McMillan D. Patten S.B. Thombs, B.D. et al. The diagnostic accuracy of the patient health questionnaire-2 (phq-2), patient health questionnaire-8 (phq-8), and patient health questionnaire-9 (phq-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Systematic reviews*, 3(1):124, 2014.

[82] García R. Luis-A. Hernán M.A. Toh, S. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. *Pharmacoepidemiology and drug safety*, 21(S2):13–20, 2012.

[83] Daniel-R. Van der Laan, M.J. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):1–40, 2006.

[84] Haight-T.J. Tager I.B. Van Der Laan, M.J. Van der laan et al. respond to hypothetical interventions to define causal effects? *American Journal of Epidemiology*, 162(7):621–622, 2005.

[85] Polley-E.C. Hubbard A.E. Van der Laan, M.J. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[86] Rose-S. Van der Laan, M.J. Targeted learning: Causal inference for observational and experimental data. 2011.

[87] Wellner-J.A. Van Der Vaart, A.W. Weak convergence. pages 16–28, 1996.

[88] Ripley B.D. Venables, W.N. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[89] Soobiah C. Tricco-A.C. Elliott M.J. Straus S.E. Veroniki, A.A. Methods and characteristics of published network meta-analyses using individual patient data: protocol for a scoping review. *BMJ open*, 5(4):e007103, 2015.

[90] Sandven-P. Brubakk O. Von der Lippe, B. Efficacy and safety of linezolid in multidrug resistant tuberculosis (mdr-tb)?a report of ten cases. *Journal of Infection*, 52(2):92–96, 2006.

[91] A.M. Walker. Confounding by indication. *Epidemiology*, 7(4):335–336, 1996.

[92] Edwards J.K. Cole-S.R. Platt R.W. Mumford S.L. Schisterman E.F. Westreich, D. Imputation approaches for potential outcomes in causal inference. *International journal of epidemiology*, 44(5):1731–1737, 2015.

[93] Royston P. Wood-A.M. White, I.R. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.