Mireille E Schnitzer[1] / Russell J Steele[2] / Michèle Bally[3] / Ian Shrier[4]

A Causal Inference Approach to Network Meta-Analysis

[1] Université de Montreal, Faculté de pharmacie, Montreal, Quebec, Canada, E-mail: mireille.schnitzer@umontreal.ca
[2] Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada
[3] Department of Pharmacy, Centre de recherche du Centre hospitalier de l'Université de Montréal, Montreal, Quebec, Canada
[4] Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, 3755 Cote Sainte Catherine Road, Montreal, Quebec H3T 1E2, Canada

Abstract: While standard meta-analysis pools the results from randomized trials that compare two treatments, network meta-analysis aggregates the results of randomized trials comparing a wider variety of treatment options. However, it is unclear whether the aggregation of effect estimates across heterogeneous populations will be consistent for a meaningful parameter when not all treatments are evaluated on each population. Drawing from counterfactual theory and the causal inference framework, we define the population of interest in a network meta-analysis and define the target parameter under a series of nonparametric structural assumptions. This allows us to determine the requirements for identifiability of this parameter, enabling a description of the conditions under which network meta-analysis is appropriate and when it might mislead decision making. We then adapt several modeling strategies from the causal inference literature to obtain consistent estimation of the intervention-specific mean outcome and model-independent contrasts between treatments. Finally, we perform a reanalysis of a systematic review to compare the efficacy of antibiotics on suspected or confirmed methicillin-resistant *Staphylococcus aureus* in hospitalized patients.

Keywords: g-formula, identifiability, network meta-analysis, nonparametric structural equation, propensity score, systematic review, TMLE

DOI: 10.1515/jci-2016-0014

1 Introduction

While individual studies are rarely used to inform scientific or medical decision making [1], multiple sources of evidence may be aggregated in order to offer more generalizable and precise comparisons between treatments [2–5]. Meta-analysis, which is the statistical synthesis of multiple study results, is often considered the highest form of quantitative evidence due to its ability to combine all relevant information in the scientific literature. However, because of such issues as effect heterogeneity across study populations and methodology that does not necessarily account for all sources of bias, the status of meta-analysis as the "gold standard" of medical knowledge has been questioned [6].

Standard meta-analysis compares two treatments of interest (or, for instance, an active treatment and placebo). When many treatments for a common condition are tested and made available over time, the medical literature may then contain multiple randomized controlled trials (RCTs) with various treatment comparisons on potentially different populations. Without additional guidance, clinicians and patients are left to informally synthesize information in the available studies in order to determine an optimal treatment decision. A *network meta-analysis* statistically aggregates the results from the relevant RCTs in order to obtain an estimate of the contrast between each pair of treatments. In particular, this type of analysis can produce estimates of contrasts even when no RCT directly compared the two treatments of interest directly.

Each RCT in the network may be performed on populations that differ in terms of their baseline characteristics. These population-specific variables may affect the average response to treatment so that in order to combine inference involving the means, it might be beneficial to control for such variables [7]. Furthermore, it has been noted that if these characteristics not only differentially affect response to treatment, but also the initial study design choice of which treatments to compare, then these variables may confound the overall effect estimate [6, 8]. As an example, Jansen et al. [8],suggest that the baseline severity of patients recruited into a study can be related to the type of treatments investigated in the study and also affect the average outcome at the end of the study. As we demonstrate in this paper, such "study-level confounding" must be adjusted for in order to obtain consistent estimation of average treatment effects.

Mireille E Schnitzer is the corresponding author.

In this paper, we consider the setting where individual patient data are not available so that the observed data is limited to average covariate and outcome values in addition to study-level information (which we refer to as "aggregate" or study-level data). We begin by describing past parametric approaches to network meta-analysis where the parameter of interest is dependent on the model specification and where the absence of effect heterogeneity is often required a priori. Using the counterfactual framework, we propose a novel definition of a marginal and model-independent causal parameter of interest in network meta-analysis and delineate the assumptions required to estimate this parameter in the presence of measured study-level confounders. We are then able to clarify conditions under which a network meta-analysis is appropriate and when it might mislead decision making regardless of estimation method used. We describe several marginal estimation methods adapted from the single study causal inference setting, including a doubly robust and semiparametric locally efficient Targeted Maximum Likelihood Estimator, and then compare these methods in a simulation study. Finally, we perform a reanalysis of the systematic review by Bally et al. [9],to compare the efficacy of antibiotics on suspected or confirmed methicillin-resistant *Staphylococcus aureus* (MRSA) in hospitalized patients.

2 The observed data

Each RCT is assumed to randomly sample subjects from a wider population, called a *superpopulation*. Within the RCT, randomization assigns subjects to two or more groups, each one receiving a treatment. These groups are often referred to as *treatment arms*. Due to randomization and random sampling, each group is a representative sample from the superpopulation. Therefore, each arm can be thought of as a distinct study on the same superpopulation. The superpopulations targeted by the RCTs may differ in terms of their characteristics due to, for example, each trial's physical and temporal location, the individual inclusion and exclusion criteria, and the recruitment sample size targets. Therefore, if effect heterogeneity exists (i. e. if the relative treatment effects at the subject level depend on baseline covariate values), one would not expect the average relative treatment effects to necessarily be equal across superpopulations.

More formally, the superpopulation is the conceptual group of essentially infinite size from which the study sample is selected [10]. A measure of some outcome (Y) is taken on each subject in the RCT arm. In this article, we will generally consider the example where the sample mean and standard deviation of Y are the summary statistics computed in each RCT.

Let A_{ij} be the intervention received by subjects in arm j of a particular RCT indexed by i. For this arm, we observe an estimated mean outcome $\bar{Y}_{ij}$ and standard deviation S_{ij}. Let $O_i = (W_i, n_i, \{N_{ij}, A_{ij}, \bar{Y}_{ij}, S_{ij}\}; j = 1, ..., n_i), i = 1, ..., N$ where W_i is study baseline information and n_i is the number of arms in the study. For the $j-$th arm of RCT i, let N_{ij} be the number of subjects and N be the total number of RCTs in the sample.

Because we are interested in summarizing effects across multiple superpopulations, we are arguably attempting to estimate effects in a *metapopulation* that contains the individual superpopulations from each study. For the purpose of this paper, we define the metapopulation as the union of possible study superpopulations and define our parameters of interest with respect to this metapopulation. In particular, we assume that the individual O_i vectors are independently drawn from the metapopulation and identically distributed.

3 Past approaches to network meta-analysis

Standard approaches in network meta-analysis where only aggregate data are observed place a hierarchical model on either the study-specific contrasts (e. g. the difference in means, $\bar{Y}_{i1} - \bar{Y}_{i2}$) or the arm-specific outcomes ($\bar{Y}_{ij}$) and specify a within-study correlation structure [3, 5, 11, 12]. As the absence of effect heterogeneity is often required, a priori [13] and post-hoc [14] investigation of this assumption is routinely recommended. The reader is referred to published guidance [15, 16] and to an example of how heterogeneity was accounted for in an economic analysis [17]. There has been recent heated debate about the appropriateness of arm-based estimation methods [11, 18].

The effect targeted in a hierarchical model depends on the contrast-type chosen and the parametrization of the model, and may or may not correspond to a marginal effect as we define further on. For binary outcomes, due to the non-collapsibility of the logistic regression model [19] in particular, adjustment for covariates in such a model changes the true value of the "effect" parameter being estimated. This type of modeling strategy may therefore be biased for the estimation of a marginal effect. Even in linear models, the inclusion of treatment interactions with covariates can also bias the value of the coefficient of treatment relative to the marginal effect. Zhang et al. [12],and Zhang et al. [20],take a missing data perspective and model the arm-specific outcomes

using a Bayesian hierarchical model to estimate marginal parameters. While neither approach has yet been extended to incorporate covariates, the former paper assumes that treatments are applied to studies at-random while the latter allows for estimation in a not-at-random context by explicitly specifying the unobservable selection mechanism.

While adjustment for covariates is rare in practice, Jansen et al. [8],introduced the notion of adapting Pearl's causal directed acyclic graphs (DAGs) to this setting [21] in order to assist in covariate selection. As a general rule, Jansen et al. [8],advocate for the adjustment of all modifiers of the relative treatment effects across comparisons. They also discourage adjustment for covariates that are not effect modifiers due to the fact that they may *induce* bias in the meta-analysis.

4 The counterfactual approach

Let Y^a be the potential (or counterfactual) outcome of a random subject drawn from the metapopulation had that subject received treatment $A = a$. In an RCT, each study arm produces an estimate of the superpopulation-specific mean of the outcome Y^a under the treatment assigned. Let the true mean of the potential outcome under treatment a for the superpopulation targeted in study i be denoted $M_i^a := E(Y^a|P_i)$ where P_i represents the superpopulation targeted in study i. Let $\Sigma_i^a := \sqrt{Var(Y^a|P_i)}$ be the standard deviation of the potential outcomes in P_i under treatment a. Now suppose that each superpopulation is independently drawn from a metapopulation, $\mathcal{P} = \bigcup_{i \in_p} P_i$, the union of all possible study superpopulations indexed by the set S_P. A marginal target parameter in a meta-analysis is $M^a := E(Y^a) = E(M_i^a)$, which represents the mean outcome under treatment a on the metapopulation. The standard deviation of the overall outcome distribution is $\Sigma^a := \sqrt{Var(Y^a)} = \sqrt{E\{Var(Y^a|P_i)\} + Var\{E(Y^a|P_i)\}} = \sqrt{E(\Sigma_i^{a2}) + Var(M_i^a)}$, representing the within and between study heterogeneity in the outcome under treatment. Due to treatment arm randomization and random sampling, $\bar{Y}_{ij}$ is an unbiased estimate of $M_i^{A_{ij}}$, the mean potential outcome under the observed treatment, and S_{ij} is a consistent estimate of $\Sigma_i^{A_{ij}}$, the potential standard deviation under the observed treatment.

For two treatments, $A = a$ and b, with corresponding means M^a and M^b, we can define a causal effect as the contrast between the mean outcome when the entire metapopulation is treated according to one treatment versus another. For instance, for binary outcomes we may define the causal risk difference as $M^a - M^b$ and the causal risk ratio as M^a/M^b.

The patient sample in any given study arm may not be representative of the metapopulation, for which the effect of interest is defined. In addition, because treatment was not randomly allocated across different RCTs, the collection of mean outcomes observed under a given treatment a may not be representative of the metapopulation under treatment a. At the design stage, the decision of which treatments to include as arms within an RCT may be influenced by the characteristics of the superpopulation on which the study is taking place. For instance, consider the example of planning a study for a superpopulation with higher disease severity from Jansen et al. [8]. Studies including patients with severe disease are more likely to include an arm with an aggressive treatment. If this occurs, the mean outcome under the aggressive treatment may be different than in a less severe superpopulation. In this situation, we would say that the treatment-mean outcome relationship is confounded at the study level by severity.

4.1 A causal directed acyclic graph (DAG) for network meta-analysis

Similar to Alonso et al. [22], we assume that heterogeneity in the different superpopulations targeted in the individual RCTs implies that each RCT estimates a different causal effect. Like Zhang et al. [12], we take an "arm-based" approach to the problem. Like Jansen et al. [8], we draw a causal DAG in order to conceptualize the relationship between treatment, study results, and population-specific characteristics. We arbitrarily choose to intervene on the arm labeled j in each study. We write $N_i = \{N_{ij}, j = 1, ..., n_i\}$, the vector of sample sizes across arms. We will also define $A_i = \{A_{ij}, j = 1, ..., n_i\}$, the vector of treatment assignments evaluated in study i, and $A_{i\backslash j}$ to mean the treatment vector excluding some arm j.

Many of the assumptions presented in detail in Section 4.3 are drawn explicitly using the study-level DAGs in Figure 1(a). The nodes of the DAG represent variables measured at the level of the RCT and the arrows between them represent the effect of the parent on the child node. For example, the absence of an arrow from $A_{i\backslash j}$ to $\bar{Y}_{ij}$, S_{ij} represents a component of the "no interference" assumption that the treatment in one arm will not affect the outcome in another. The arrow from N_{ij} to $\bar{Y}_{ij}$, S_{ij} is present because the sample size within a study arm will affect the distribution of the outcome summary statistics.

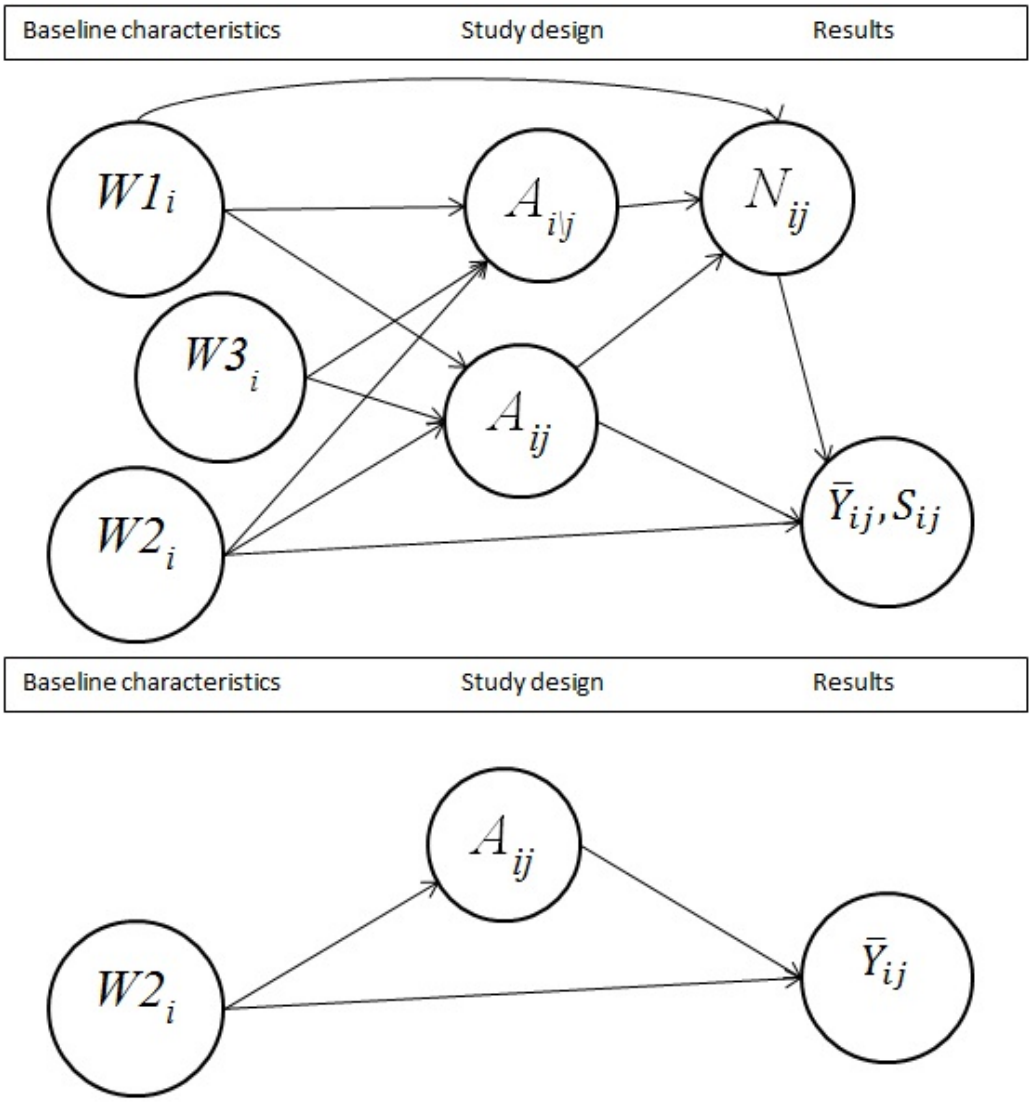

Figure 1: (a) The study-level DAG reflecting the unconfoundedness and time-ordering assumptions made in Section 4.3 and Section 4.2 without assuming independence between the sample mean and standard deviation within a study arm. (b) The simplified DAG that arises from assuming the independence between the sample mean and standard deviation. Here, $W1_i$, $W2_i$ and $W3_i$ are baseline covariates, A_{ij} and $A_{i\setminus j}$ are the treatments assigned to arms j and the non-j arm(s), respectively, N_{ij} is the sample size of arm j, $\bar{Y}_{ij}$ is the mean outcome and S_{ij} is the estimated standard error of the outcome of arm j.

The sample size node N_{ij} is determined by the sample size calculation made in the study design phase and also by the success of recruitment. This calculation is inherently conditional on the superpopulation being evaluated, as superpopulation characteristics are taken into account when hypothesizing an effect size and standard error. This calculation is also conditional on the treatments being compared.

Causal DAGs can be used as a tool to identify which variables must be controlled for in the meta-analysis in order to estimate the treatment-specific metapopulation mean outcome. Depending on some underlying statistical assumptions that we will investigate in detail in the following sections, these DAGs may simplify to Figure 1(b). This happens because we can ignore the mediation path through N_{ij} in order to estimate the total effect of the treatment on outcome. Under these conditions, assuming independence between the variables in W_i, the analysis must adjust for all common causes of treatment selection and study outcome distribution.

Note that the recommendations based on this DAG differ from those of Jansen et al. [8], who say that the analysis must adjust exclusively for effect modifiers. The assumptions that we list in Section 4.3 are explicitly

required in the steps we take in Section 4.2 in order to obtain identifiability of the meta-analysis parameter of interest.

4.2 The G-formula and nonparametric identifiability

Suppose we observe the aggregate data $O_i, i = 1, ..., N$, independently drawn and identically distributed. Using the nonparametric structural equation modeling (NPSEM) of Pearl [21], the metapopulation mean outcome, M^a, can be shown to be identifiable (that is, known with infinite data) under the several conditions outlined and discussed in Section 4.3.

4.2.1 The observed data generation

At the study design stage for RCT i, the superpopulation P_i is randomly drawn from the metapopulation P. The selection of P_i determines the population-level covariates W_i. The number of study arms n_i and the treatments compared in the study, the multivariate $A_i = \{A_{ij}, j = 1, ..., n_i\}$, are drawn conditional on W_i. The sample size calculation is carried out based on the choice of treatment comparison and on the sub-population characteristics (i. e. based on the expected effect and precision in that sub-population). This calculation is approximate and the resulting sample size also depends on the success of recruitment. Therefore, the sample sizes for the treatment arms, N_i, are not deterministic, but are drawn conditional on A_i, n_i and W_i.

The second stage operates at the individual level once subjects are recruited and randomly assigned treatment. Suppose each subject k in arm j of study i has continuous outcome $Y_{ijk}, k = 1, ..., N_{ij}$ (under treatment A_{ij}). Each Y_{ijk} is independently drawn from a distribution with mean $M_i^{A_{ij}}$ and standard deviation $\Sigma_i^{A_{ij}}$. The empirical mean outcome in arm j of study i is therefore $\bar{Y}_{ij} = 1/N_{ij} \sum_k Y_{ijk}$. The standard deviation is estimated as $S_{ij}^2 = 1/(N_{ij} - 1) \sum_k (Y_{ijk} - \bar{Y}_{ij})^2$. In addition, subject recruitment yields summary characteristics of the superpopulation, which we assume to include complete information about the covariates W_i that were known at study conception and contributed to the treatment choice. We assume in the following that we do not observe the subject-level data.

Let ω_{ij} represent the set of estimated summary statistics of the outcome variable from study i arm j. For instance, we might have that $\omega_{ij} = \{\bar{Y}_{ij}, S_{ij}\}$. Correspondingly, let ω_{ij}^a be the set of estimates of the counterfactual summary statistics that would arise had arm j been assigned treatment a.

Assuming no interference between arms and that the distribution of ω_{ij} in one arm of a study is conditionally independent of the outcomes in the others and also independent of the total number of arms, the NPSEM that we assume can then be written as

$$
\begin{aligned}
W_i &= f_W(\varepsilon_W) \\
n_i &= f_n(W_i, \varepsilon_n) \\
A_i &= f_A(n_i, W_i, \varepsilon_A) \\
N_i &= f_N(A_i, n_i, W_i, \varepsilon_N), \text{for} j = 1, ..., n_i \\
\omega_{ij} &= f_\omega(N_{ij}, A_{ij}, W_i, \varepsilon_\omega), \text{for} j = 1, ..., n_i
\end{aligned}
$$

The probability density function $f(O_i)$ arising from the NPSEM without intervention can be decomposed as

$$
\begin{aligned}
f(O_i) = Q_W(W_i) Q_n(n_i|W_i) g_A(A_i|n_i, W_i) \times \\
Q_N(N_i|A_i, n_i, W_i) \prod_{j=1}^{n_i} Q_\omega(\omega_{ij}|N_{ij}, A_{ij}, W_i)
\end{aligned}
$$

where $Q_W(W_i)$ corresponds to the density function for W_i, $Q_n(n_i|W_i)$ corresponds to the density function for n_i conditional on W_i, and $g_A(A_i|n_i, W_i)$ corresponds to the conditional density function for A_i. Within each RCT, $Q_N(N_i|A_{ij}, A_{i\setminus j}, n_i, W_i)$ corresponds to the conditional density function forN_i and $Q_\omega(\omega_{ij}|N_{ij}, A_{ij}, W_i)$ is the conditional (joint) density for the measured summary statistic(s) in arm j.

4.2.2 The counterfactual distribution

Define an intervention as the assignment of treatment strategy a to an arbitrary arm in each study. In other words, for all i we set $A_{ij} = a$ for a single arbitrary arm j. The remaining non-j arms receive potential treatments $A^a_{i\backslash j}$. The joint density for the counterfactual data $O^a_i = (W_i, n_i, A^a_{i\backslash j}, \{\omega^a_{ij^*}, N^a_{ij^*}; j^* = 1, ..., n_i\})$ can be obtained through the G-formula [23]. This joint density function can be written as

$$f(O^a_i) = Q_W(W_i)Q_n(n_i|W_i)g_{A\backslash j}(A^a_{i\backslash j}|n_i, W_i)Q_N(N^a_i|A^a_{i\backslash j}, n_i, W_i)Q_\omega(\omega^a_{ij}|N^a_{ij}, W_i)\times \prod_{j^*\neq j} Q_\omega(\omega^a_{ij^*}|N^a_{ij^*}, A^a_{ij^*}, W_i)Q_N(N^a_{ij^*}|A^a_{i\backslash j}, n_i, W_i)$$

where $g_{A\backslash j}(A^a_{i\backslash j}|n_i, W_i)$ is defined as the conditional (joint) density of the treatments assigned to non-j arms.

4.2.3 Identifiability for conditionally independent $\bar{Y}$ and S

Suppose we have that $\omega_{ij} = \{\bar{Y}_{ij}, S_{ij}\}$, meaning that each study reported the sample means and sample standard deviations of a continuous outcome. We then make the key structural assumption that $\bar{Y}^a_{ij} \perp\!\!\!\perp S^a_{ij}|N^a_{ij}, W_i$ where N^a_{ij} is the counterfactual sample size in study i arm j. Let Y^a_{ijk} represent an individual recruited into study i arm j in the counterfactual scenario. The independence assumption arises naturally from the distributional assumption that $Y^a_{ijk} \sim N(M^a_i, (\Sigma^a_i)^2)$ because $\bar{Y}^a_{ij}$ and S^a_{ij} are the sample mean and standard deviation in superpopulation P_i when a is the treatment assigned. Asymptotically, we have that $\bar{Y}^a_{ij}$ and S^a_{ij} are independent normal variables when the subject-level outcomes are drawn from a distribution with zero skew, such that $E\{(Y^a_{ijk})^3\} = 0$ [24], p. 46]. We show in Appendix A.1 that under this assumption $f(O^a_i)$ can be decomposed in such a way that the mean outcome, $E(\bar{Y}^a_{ij}) = M^a$, can be written independently of the non-j arms, resulting in the simple equality $M^a = \int_W E(\bar{Y}^a_{ij}|N^a_{ij}, W_i)Q_W(W_i)dW$. Under the unconfoundedness assumption $\bar{Y}^a_{ij} \perp\!\!\!\perp A_{ij} = a|N^a_{ij}, W_i$, and under the consistency assumption (see next section) we may write the G-formula [23] $M^a = \int_{W_i} E(\bar{Y}_{ij}|W_i, A_{ij} = a)Q_W(W_i)dW_i$. Therefore, this quantity is identifiable from the data.

Identifiability without assuming this structural independence is possible, and we describe the additional causal assumptions required for this setting in Appendix A.2.

4.2.4 Identifiability for binary outcomes

If the original study outcomes are binary (such that $Y_{ijk} = \{0, 1\}$), the study means $\bar{Y}_{ij}$ are the proportions of subjects with the indicated outcome. Therefore, $N^a_{ij}\bar{Y}^a_{ij}$ has a binomial distribution with true probability of outcome $M^a_i = E(Y^a|P_i)$. Then, $\Sigma^a_i = \sqrt{Var(Y^a|P_i)} = \sqrt{M^a_i(1 - M^a_i)}$. Similarly, the study arm estimate of the standard deviation is $S^a_{ij} = \sqrt{\bar{Y}^a_{ij}(1 - \bar{Y}^a_{ij})}$. In this case, the likelihood will not include a component for S_{ij} so no independence assumption is necessary. The resulting G-formula is still $M^a = \int_{W_i} E(\bar{Y}_{ij}|N_{ij}, W_i, A_{ij} = a)Q_W(W_i)dW_i$ and will rely on the same unconfoundedness assumption that $\bar{Y}^a_{ij} \perp\!\!\!\perp A_{ij} = a|N_{ij}, W_i$.

4.3 Assumptions

For convenience, here we list the assumptions needed for the identification of M^a, corresponding with the NPSEM in Section 4.2.1 and the DAGs in Figure 1. We also comment on the meaning and plausibility of these assumptions in the hypothetical situation where each individual RCT has full compliance. Under full compliance, each RCT arm produces a consistent estimate of the mean outcome in the superpopulation under full adherence to the assigned treatment.

No interference. The use of the above counterfactual notation presupposes that the treatment assigned to one study does not affect the counterfactual outcome of another study [25]. A secondary level of interference within an individual study involves the treatment in one study *arm* affecting the outcomes in another study arm. This means that the estimates $\bar{Y}^a_{ij}$ and S^a_{ij} do not depend on the treatment received by another arm of the same RCT. The assumption of no interference will generally not hold for certain studies of infectious disease. For example,

an effective vaccine in one arm may impact the outcome of an unvaccinated subject in the control arm, because the unvaccinated subject will be less likely to be exposed to the disease through herd immunity.

Unconfoundedness. (Weak) unconfoundedness [26] is required for the identification of M^a. In this context, unconfoundedness is the assumption that the counterfactual sample means under a treatment a are independent of the true treatment received conditional on measured covariates. Specifically, this means that $\bar{Y}^a_{ij} \perp\!\!\!\perp A_{ij} = a | N^a_{ij}, W_i$. In the example DAG of Figure 1(a), this corresponds to measuring all the components of node $W2_i$. The validity of this assumption is entirely dependent on the subject-matter, how RCTs in the field are designed, and on the information reported in the RCTs.

Consistency. The consistency assumption in this context states that the counterfactual mean of a study arm under a given treatment is the same as the observed result. With notation, this is equivalent to stating that $\bar{Y}^a_{ij} = \bar{Y}_{ij}$ when $A_{ij} = a$. Having different definitions of treatment across studies may violate this assumption if all are categorized under the same treatment type and this variation has an impact on the outcome [27]. For example, there may be different drug dosages and lengths of follow-up across studies. Disregarding these differences will violate consistency if the various treatment-types have differential effects on the patient outcomes. With some additional unconfoundedness requirements, one might surmount this obstacle using the approach described in VanderWeele and Hernán [28]. [Note that this definition of consistency corresponds with the causal assumption and is distinct from the network meta-analysis meaning of the term in e. g. [14].]

Positivity. Finally, we need to evaluate both theoretical and practical positivity. Theoretical positivity is the assumption that, *conditional only on variables required for unconfoundedness*, all studies had a positive probability of being assigned each treatment under investigation. Practical positivity is the condition that for every level of the characteristics W_i, there is an *estimated* positive probability of receiving treatment.

It is important to note that treatment comparisons are based on the same P and that the target parameter $M^a = E(M^a_i)$ relies on the definition of this metapopulation. If positivity does not hold on some subpopulations it would be necessary to exclude all studies (and corresponding superpopulations) that contain such subpopulations.

It is furthermore important to note that the positivity assumption is not the same as requiring that all studies could have realistically been assigned each treatment. In particular, certain treatments may not have been available when some older trials were carried out. If year of study is not required to unconfound the analysis, then the *unconditional* probability may still be non-zero.

5 Estimation of the treatment-specific metapopulation mean outcome

5.1 G-Computation

G-Computation procedures based on the G-formula in Section 4.2 can be used to estimate the target parameter. Here we define a simple procedure resulting from the data requirement that the sample mean and standard deviation are independent within a study arm. This procedure allows for simple frequentist estimation of the mean effect of treatment.

This procedure requires estimates for the conditional expectation $E(\bar{Y}_{ij}|W_i, N_{ij}, A_{ij} = a)$ for a given value of treatment. First we must note that while the conditional mean of $\bar{Y}_{ij}$ is independent of N_{ij}, its distribution is not. In particular, we have that

$$Var(\bar{Y}_{ij}|W_i, N_{ij}, A_{ij}) = \frac{1}{N_{ij}} Var(Y_{ijk}|W_i, N_{ij}, A_{ij}) = \frac{1}{N_{ij}} (\Sigma_i^{A_{ij}})^2.$$

Because S^2_{ij} is a consistent estimate of the superpopulation-level variance under treatment A_{ij}, we are able to estimate this variance.

A model for the regression on $\bar{Y}_{ij}$ may be fit by pooling over all arms regardless of treatment assignment. In order to obtain the Best Linear Unbiased Estimator, we can weight by N_{ij}/S^2_{ij}. Using this model fit, we predict $\hat{\bar{Y}}^a_i = \hat{E}(\bar{Y}_{ij}|W_i, A_{ij} = a)$, i. e. the predicted mean under treatment a for each study. The G-Computation estimate is then $\hat{M}^a_{GCOMP} = 1/N \sum_{i=1}^{N} \hat{\bar{Y}}^a_i$.

The standard error for the G-Computation estimate is usually computed through nonparametric bootstrap methods [29]. Bootstrap resampling must be done by resampling studies, rather than arms, similar to what is done in a study with clustering [30].

5.2 Inverse probability of treatment weighting

Likelihood methods, such as G-Computation, require correct parametric specification of the outcome model, which may be difficult to specify. An alternative approach is to utilize propensity score methods, which require the estimation of a model for the treatment received by the arm. For a given treatment type a, let $g_a(W_i)$ be an estimate of the probability $P(a \in A_i|W_i)$, called the generalized propensity score [26].

Despite the small sample size in standard network meta-analysis, one might attempt inverse probability of treatment weighting (IPTW) for the estimation of the marginal parameter. Let $\bar{Y}_i^a$ represent the observed outcome of the arm of study i that received treatment a (or N/A if no arm of study i received treatment a). An IPTW estimator for multiple treatments [26] can be represented as

$$\hat{M}^a_{IPTW} = 1/N \sum_{i=1}^{N} \frac{\mathbb{I}(a \in A_i)\bar{Y}_i^a}{g_a(W_i)}.$$

Intuitively, this estimator takes a mean of $\bar{Y}_{ij}$ with only the arms treated according to $A_{ij} = a$. It then adjusts this estimate to remove the confounding bias caused by the baseline variables.

The consistency of this estimator can be shown as follows.

$$\hat{M}^a_{IPTW} \xrightarrow{P} E\left[\frac{\bar{Y}_i^a \mathbb{I}(a \in A_i)}{P(a \in A_i|W_i)}\right] = E\left\{\bar{Y}_i^a E\left[\frac{\mathbb{I}(a \in A_i)}{P(a \in A_i|W_i)}\middle|\bar{Y}_i^a, W_i\right]\right\} = E\left(\bar{Y}_i^a\right) = M^a.$$

5.3 Targeted minimum loss-based estimation

Targeted Minimum Loss-based Estimation (TMLE) [31, 32] is a framework for the construction of semi-parametric estimators generally applied to the estimation of causal quantities. The TMLE procedure is carried out by first fitting a model for the expected value of the arm-based means, $E(\bar{Y}_{ij}|W_i, A_{ij} = a)$ which, under the causal assumptions, can equivalently be written as the expectation of the potential outcome had the study evaluated treatment a, $E(\bar{Y}_i^a|W_i, a \in A_i)$. As in the G-Computation procedure, this model can be estimated by weighing each observation by N_{ij}/S^2_{ij}. For each arm in the study, we use this model to obtain $\hat{\bar{Y}}_i^a$, predictions of the sample mean of each trial i under treatment a. These predictions are then updated by fitting a no-intercept logistic regression using study arms that evaluated treatment a. This logistic regression is fit with outcome $\bar{Y}_{ij}$, offset $logit(\hat{\bar{Y}}_i^a)$, and single covariate $g_a^{-1}(W_i)$, corresponding with the inverse probability weights. Denote the estimate of the coefficient from this regression as $\hat{\varepsilon}$. The updated predictions are then $logit(\hat{\bar{Y}}_i^{a,*}) = logit(\hat{\bar{Y}}_i^a) + \hat{\varepsilon}/g_a(W_i)$, which is calculated for each study. The final targeted estimate for M^a is $\hat{M}^a_{TMLE} = 1/N \sum_{i=1}^{N} \hat{\bar{Y}}_i^{a,*}$. Note that in order to perform the update step, the means and outcome must be transformed to (0,1) and then subsequently transformed back to the original scale [33]. This can be done using real or empirical bounds.

This TMLE is consistent under correct specification of the propensity score model or the model for the expected value of the mean outcome (the property of *double robustness*). If both of these models are correct, then TMLE is asymptotically efficient in the class of regular, asymptotically linear estimators in the semiparametric model space [32]. More details and a proof of consistency are included in Appendix A.3.

6 Simulation study

In this section we demonstrate that we can obtain consistent estimation of the target parameter $M^a = E(Y^a)$ under the NPSEM using the proposed estimators. We also compare the efficiency of each approach.

While the proposed estimators do not restrict the number of study arms, we fix all simulated studies to have exactly two treatment arms for simplicity. We are interested in estimating the mean outcome of the metapopulation under treatment for each of four treatments of interest. For each study $i = 1, ..., N$, we generate the population average characteristic, W_i from a Poisson distribution with mean 2. The probabilities of receiving a given treatment are calculated conditional on the value of W_i. Two treatment options A_i are then sampled without replacement using the calculated probabilities. Treatments 2 and 4 are generated to be less likely to be chosen with larger W_i. The sample size N_i (which we allowed to be common to both arms in the study) is drawn from a Poisson distribution with mean linear in W_i and A_i. For each subject within each arm, we draw a baseline covariate X_{ijk} from a Gaussian distribution with mean W_i and constant variance. We set $\beta = (0.8, 0.2, 1, -0.05)$ to

be the treatment-specific coefficients. Outcome values Y_{ijk} are drawn from a Gaussian: $Y_{ijk} \sim N(X_{ijk} + \beta[A_{ij}], 1)$. A summary of the data-generation is presented in Table 1.

Table 1: Simulation study: data generation.

Variable	Study design: for each $i = 1, ..., N$
Number of arms	$n_i = 2$
Study-level covariate	$W_i \sim Poisson(\mu = 2)$
	$A_i = (A_{i1}, A_{i2})$ sampled without replacement with probabilities
	$p_1 = logit^{-1}(0.4W_i)$
	Treatments $p_2 = logit^{-1}(-0.4W_i)$
	$p_3 = logit^{-1}(0.8W_i)$
	$p_4 = logit^{-1}(-0.8W_i)$
	Sample size $N_i \sim Poisson(\mu = 5000exp(-0.4W - sum(\gamma[A_{ij}])))$
	(study recruitment) where $\gamma = (-1.5, 1, -1, 1)$
	Within-study: for each $j = 1, 2, k = 1, ..., N_i$
	Subject-level covariate $X_{ijk} \sim N(\mu = W_i, \sigma^2 = 4)$
Subject-level outcome	$Y_{ijk} \sim N(\mu = X_{ijk} + \beta[A_{ij}], \sigma^2 = 1)$
	where $\beta = (0.8, 0.2, 1, -0.5)$
	Observed data: for each $i = 1, ..., N, j = 1, 2$
	W_i, A_i, and $\bar{Y}_{ij}$ where
	Study-level information $\bar{Y}_{ij} = 1/N_i \sum_{k=1}^{N_i} Y_{ijk}$

The sample statistics from each study arm are calculated by taking the mean and standard deviation of Y_{ijk} within each arm. The true treatment-specific superpopulation means are $M^1 = 2.80, M^2 = 2.20, M^3 = 3.00, M^4 = 1.95$. We are interested in estimating a subset of the contrasts between the treatments, specifically marginal mean differences $M^2 - M^1 = -0.60$, $M^3 - M^1 = 0.20$, and $M^4 - M^1 = -0.85$. Note that random effects were not generated in this simple simulation study.

We tested the three methods described in the text (G-Computation, IPTW and TMLE) for $N = 15$ and 50 simulated studies. We used logistic regression models conditional on the covariate for the generalized propensity score for IPTW and TMLE. We ran two scenarios: incorrect and correct outcome model specification. For the correct scenario, linear regression models for the outcome adjusting for treatment type and covariate were used in G-Computation and TMLE. For the incorrect scenario, the outcome was scaled to $(0, 1)$ and logistic regression models were used. We also display results for an unadjusted estimator that merely takes the mean difference in treatment-specific outcomes when available. Variance and confidence intervals were estimated using the nonparametric cluster bootstrap [30] where study is considered the cluster (and arms are the individual observations). In Table 2, we present statistics describing the quality of the estimation of all contrasts with treatment 1. These statistics are the percent finite sample bias ("% Bias"), the standard deviation of the estimates over the simulated data ("SE-MC"), the bootstrap-estimated standard error ("SE-BS"), and the percentage of the 95% confidence intervals that capture the true effect size ("% Cov"). Bootstrap resamples that did not allow for an estimate of the contrast (i. e. if either of the treatments did not appear in the resampled data set) were discarded, potentially biasing this standard error estimate.

The unadjusted estimator was greatly biased for the first and third contrasts, indicating that those two contrasts were highly confounded by the simulated study-level covariate. The correctly specified G-Computation estimator had the lowest bias throughout, the smallest standard errors, and near optimal confidence interval coverage. This is to be expected as G-Computation is a function of maximum likelihood parameter estimates with correct parametric specification of the necessary component of the likelihood (namely, the conditional mean of the outcome). However, with an incorrectly specified outcome model, the estimator was biased which caused the coverage to suffer for the third contrast.

IPTW was the most biased estimator and also had the largest variance. The bias largely dissipated when the sample size was increased to $N = 50$ studies. IPTW had good coverage except for the third treatment contrast where treatment 4 was rare. The slower convergence of IPTW in the contrast involving treatment 4 can be explained by a higher variance of the estimated weights for that treatment compared to the others. The performance of IPTW has previously been seen to suffer when data support for certain exposure levels is sparse (i. e. under near practical positivity violations) [33]. Truncation of the propensity score at 5% and 10% respectively (that is, replacing the bottom p% of the propensity score with the pth percentile) [34] increases the bias for the first and third contrasts while reducing the variance, with no effect on the coverage (results not shown).

TMLE with correct outcome model specification had bias comparable to G-Computation but slightly higher for $N = 15$. For $N = 50$, the standard error of TMLE was comparable to that of G-Computation but for $N = 15$

it was up to 80 times larger. Regardless, correctly specified TMLE had good coverage throughout. Notably, the bootstrap standard error estimates were comparable to the Monte-Carlo standard error for $N = 50$ but diverged for IPTW and TMLE when $N = 15$. Certain implementations of TMLE are more sensitive to near practical positivity violations [33, 35, 36], hence the need for the robust version that involves the logistic regression for the update of the predictions [as described in [33],and for our specific setting in Section 5.3]. When the outcome model was misspecified, TMLE also accrued bias for the first and third contrasts, with magnitude comparable to the misspecified G-Computation. This bias decreased with more studies due to the double robustness of TMLE (making this estimator consistent even when the outcome model is misspecified). Coverage only suffered for the third contrast which was the most biased.

Table 2: Simulation: quality of treatment contrast estimation with a Gaussian outcome (two-arm studies, 1,000 simulated datasets).

		% Bias		SE-MC		SE-BS(% Cov)	
	N	15	50	15	50	15	50
				Correctly specified models			
$M^2 - M^1 = -0.6$							
	G-Comp	0	0	0.04	0.02	0.04(91)	0.02(94)
	IPTW	40	9	0.57	0.46	0.61(89)	0.41(92)
	TMLE	4	0	0.27	0.03	0.44(96)	0.05(92)
$M^3 - M^1 = 0.20$							
	G-Comp	1	0	0.04	0.02	0.04(91)	0.02(95)
	IPTW	–2	0	0.10	0.04	0.25(99)	0.05(97)
	TMLE	–1	–1	0.17	0.04	0.29(96)	0.04(93)
$M^4 - M^1 = -0.85$							
	G-Comp	0	0	0.04	0.02	0.04(89)	0.02(93)
	IPTW	81	3	0.76	0.74	0.62(63)	0.80(68)
	TMLE	–9	0	0.81	0.11	0.74(94)	0.21(95)
				Misspecified outcome model			
$M^2 - M^1 = -0.6$	No adjustment	101	103	0.65	0.35	0.61(75)	0.34(52)
	G-Comp	2	12	0.20	0.13	0.24(98)	0.11(94)
	TMLE	–8	–2	0.33	0.09	0.46(97)	0.11(96)
$M^3 - M^1 = 0.20$	No adjustment	5	–7	0.37	0.20	0.38(92)	0.20(93)
	G-Comp	–1	7	0.20	0.12	0.18(99)	0.11(95)
	TMLE	0	0	0.15	0.05	0.28(99)	0.05(96)
$M^4 - M^1 = -0.85$	No adjustment	126	125	0.69	0.38	0.61(51)	0.36(18)
	G-Comp	36	33	0.53	0.29	0.48(88)	0.24(80)
	TMLE	44	–24	0.86	0.38	0.75(87)	0.36(75)

7 Application: Antibiotic use on methicillin-resistant *Staphylococcus aureus* infection

We illustrate this causal inference approach and the adapted estimation methods in network meta-analysis with an example from infectious disease research. An increase in MRSA has spurred investigation of comparative efficacy of different antibiotic treatment options. While the antibiotic vancomycin has been the standard treatment for decades, treatment failures have been noted in patients with serious infections [37]. Interest therefore lies in whether alternative antibiotics are as effective as the standard. Bally et al. [9] performed a systematic review and Bayesian network meta-analyses of RCTs of parenteral antibiotics used for treating hospitalized adults with complicated skin and soft-tissue infections (cSSTIs) and hospital-acquired or ventilator-associated pneumonia.

We consider the target population of interest to be the population of clinical trial participants with suspected or confirmed MRSA cSSTIs or pneumonia, with corresponding studies published until May 2012. The site of infection and confirmation of MRSA represent important differences in the entrance criteria of the various studies. 24 studies were found. Patients were randomized based on suspicion of MRSA in all but three studies for which the protocol specified confirmation of presence of MRSA at baseline. 14 studies enrolled subjects with cSSTIs, 7 studies enrolled subjects with hospital-acquired or ventilator-associated pneumonia, and 3 studies allowed for either indication. The original network meta-analysis of Bally et al. [9],analyzed each infection site in separate analyses and therefore obtained stratified estimates. Based on the theory we developed, we can account

for the potentially different treatment effects in each subpopulation by controlling for subpopulation type as a covariate in the analysis. By doing so, we ask a higher-level yet still clinically interesting question: "Are the alternative therapies as effective as the standard antibiotic for the treatment of suspected or confirmed MRSA?" Because infection site, MRSA confirmation, and study year can potentially affect the choice of investigated therapies and the outcomes, these three covariates (labeled W_i) should be adjusted for in order to minimize confounding bias.

The outcome of interest is clinical test of cure for all subjects who received at least one dose of treatment (a standard measure in infectious disease research). Four papers evaluated the outcome only on a subset of patients selected post-randomization; as this does not conform to our definition of the RCT-specific parameter of interest, we considered these outcomes missing. For our analysis, we chose to compare vancomycin with the two most prevalent alternatives: telavancin and linezolid. In total, 47 study arms evaluated one of these three treatments and 36 had an observed outcome. Of the remaining treatments, tigecycline, daptomycin, and ceftaroline were each evaluated in three study arms, and a regime of quinupristin/dalfopristin was evaluated in one arm. All of this information is available in the data extraction Table 3.

Table 3: Data extraction table for the network meta-analysis of antibiotic use on methicillin-resistant *Staphylococcus aureus* infection.

Publication	Events	Ni	Ai	StudyID	Year	Infection	Confirmed MRSA at baseline
Katz et al., 2008	42	48	vancomycin	1	2007	cSSTI	0
	36	48	daptomycin	1	2007	cSSTI	0
Arbeit et al., 2004	162	266	vancomycin	2	2001	cSSTI	0
	165	264	daptomycin	2	2001	cSSTI	0
	235	292	vancomycin	3	2000	cSSTI	0
	217	270	daptomycin	3	2000	cSSTI	0
Breedt et al., 2005	216	250	vancomycin	4	2003	cSSTI	0
	212	253	tigecycline	4	2003	cSSTI	0
Sacchi-danand et al., 2005	196	255	vancomycin	5	2003	cSSTI	0
	203	268	tigecycline	5	2003	cSSTI	0
Stryjewski et al., 2008	307	429	vancomycin	6	2006	cSSTI	0
	309	426	telavancin	6	2006	cSSTI	0
	360	489	vancomycin	7	2006	cSSTI	0
	348	472	telavancin	7	2006	cSSTI	0
Stryjewski et al., 2006	81	95	vancomycin	8	2004	cSSTI	0
	82	100	telavancin	8	2004	cSSTI	0
Corey et al., 2010	297	347	vancomycin	9	2007	cSSTI	0
	304	351	ceftaroline	9	2007	cSSTI	0
Wilcox et al., 2010	289	338	vancomycin	10	2007	cSSTI	0
	291	342	ceftaroline	10	2007	cSSTI	0
Talbot et al., 2007	26	32	vancomycin	11	2005	cSSTI	0
	59	67	ceftaroline	11	2005	cSSTI	0
Weigelt et al., 2005	402	573	vancomycin	12	2003	cSSTI	0
	439	583	linezolid	12	2003	cSSTI	0
Stevens et al., 2002	54	87	vancomycin	13	1999	cSSTI	0
	64	99	linezolid	13	1999	cSSTI	0
	16	32	vancomycin	14	1999	pneumonia	0
	20	39	linezolid	14	1999	pneumonia	0
Wunderin-ket al., 2003	128	302	vancomycin	15	2000	pneumonia	0

	135	321	linezolid	15	2000	pneumonia	0
Rubenstein et al., 2001	73	192	vancomycin	16	1999	pneumonia	0
	85	203	linezolid	16	1999	pneumonia	0
Rubenstein et al., 2011	221	374	vancomycin	17	2007	pneumonia	0
	214	372	telavancin	17	2007	pneumonia	0
	228	380	vancomycin	18	2007	pneumonia	0
	227	377	telavancin	18	2007	pneumonia	0
Fagon et al., 2000	67	148	vancomycin	19	1996	pneumonia	0
	65	150	quin-upristin/ dalfopristin	19	1996	Pneumo nia	0
Lin et al., 2008	NA	33	linezolid	20	2005	cSSTI	0
	NA	29	vancomycin	20	2005	cSSTI	0
	NA	38	linezolid	21	2005	pneumonia	0
	NA	40	vancomycin	21	2005	pneumonia	0
Kohno et al., 2007	NA	51	linezolid	22	2004	cSSTI	0
	NA	26	vancomycin	22	2004	cSSTI	0
	NA	31	linezolid	23	2004	pneumonia	0
	NA	17	vancomycin	23	2004	pneumonia	0
Florescu et al., 2008	NA	70	tigecycline	24	2005	cSSTI	0
	NA	23	vancomycin	24	2005	cSSTI	0
Itani et al., 2010	223	276	linezolid	25	2007	cSSTI	1
	196	266	vancomycin	25	2007	cSSTI	1
Wunderink et al., 2008	NA	30	linezolid	26	2005	pneumonia	1
	NA	20	vancomycin	26	2005	pneumonia	1
Wunderink et al., 2012	102	186	linezolid	27	2010	pneumonia	1
	92	205	vancomycin	27	2010	pneumonia	1

We ran four methods to obtain estimates of the counterfactual relative risk of both contrasts with the comparator vancomycin. The methods are 1) a ratio of the unadjusted mean outcomes using all available arms (called "No Adjust"), 2) a random effects regression for the arm-specific study outcomes using a log-link and a study-specific intercept ("RE Arm"), 3) G-Computation where a random effects logistic regression weighted by the inverse standard errors is used to predict the conditional mean outcomes, and 4) TMLE with a weighted logistic random effects model for the outcome and LASSO-penalized logistic regressions (to handle the sparse data) for the propensity score and a missing data model using the R library glmnet Friedman et al. [38] The missing outcomes required that the TMLE algorithm include fitting a model to estimate the probability of a missing outcome in each study; the TMLE update step was therefore modified to use a product of the propensity score and the probability of observing the outcome in place of $g_a(W_i)$. To estimate the standard errors and confidence intervals, the built-in functions in the library lme4 were used for the random effects model, and the clustered nonparametric bootstrap (1,000 times 54 resamples of 27 studies with replacement) was used for the other methods.

The results of the network meta-analysis are presented graphically in Figure 2 (and numerically in the Appendix Figure 4). We also included the results of the studies that contrasted the two treatments directly. For the comparison of telavancin versus vancomycin, all estimators include the null in the confidence interval. The random effect regression and G-Computation produce estimates of the relative risk close to one, indicating near equivalence of treatments while the point estimate of TMLE was further from the null (in the direction of the superiority of vancomycin). Notably, the confidence interval for the TMLE in the first contrast is much wider than the others. The unadjusted method produced a point estimate in the direction of the superiority of telavancin, demonstrating that the correction for study-level confounding impacted the analysis. For the comparison of linezolid versus vancomycin, the random effects regression, G-Computation and TMLE agree on the superiority of linezolid. The original study by Bally et al. [9],also found some suggestion of a superior effect of linezolid compared to vancomycin but for both subpopulations the confidence intervals were large and spanned the null.

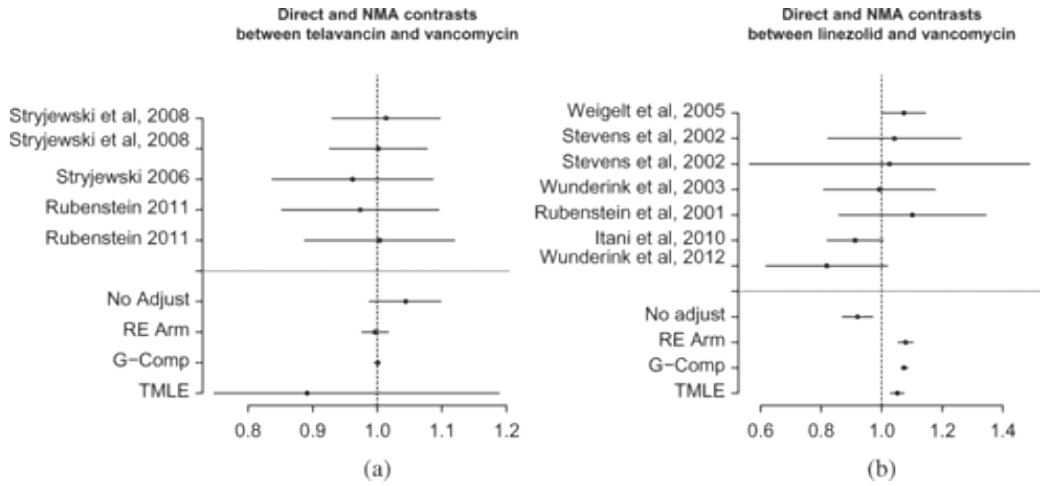

Figure 2: Risk ratio estimates and confidence intervals for clinical success at test of cure for all studies with direct comparisons and all network meta-analysis methods for the contrasts between a) telavancin and vancomycin and b) linezolid and vancomycin. Risk ratio values below one indicate superiority of vancomycin.

We can also easily obtain estimates of the contrast between telavancin and linezolid. The G-Computation and TMLE produce risk ratios for clinical success of 0.94 (95% confidence interval = 0.92,0.94) and 0.85 (0.71,1.12), respectively, with G-Computation concluding the superiority of linezolid. As no RCT directly contrasted these two antibiotics, this demonstrates another general advantage of network meta-analysis, which is the ability to formally compare treatments using only indirect evidence of their relative performance.

If we are to interpret the summary statistics as estimates of the relative causal effects of antibiotic choice on successful treatment, the causal assumptions in Section 4.3 need to be satisfied. Each of the studies evaluated the clinical efficacy of the treatments, which is defined on patients who had received at least one dose of the study drug. Because randomized treatment was first-line therapy (administered intravenously in-hospital) and the success of treatment was determined clinically, each trial estimated the relative effect under full adherence. *No interference*: No interference is credible in this case because all subjects were already suspected or confirmed to have MRSA upon entry to the study. Therefore, the choice of treatment in the other arm wouldn't have an effect on existing infections nor the success of treatment. *Unconfoundedness*: The unconfoundedness assumption relies on whether year, infection type, and whether MRSA was confirmed were sufficient to control for confounding at the study-level. This assumption could be violated if prognostic demographic variables were involved in the study design stage. However, prognostic markers such as diabetes and peripheral vascular disease (for cSSTI) and mechanical ventilation, APACHE II score, clinical markers of severity, and presence of organ dysfunction [for pneumonia) are unlikely to determine the choice of initial therapy [39], Niederman [40]. *Consistency*: The dosage regimens varied somewhat across studies but were all considered to be at therapeutic levels. However, the length of time to the evaluation time point for each treatment type varied within and between studies (e. g. 7–14 days for telavancin versus 12–28 days for linezolid]. If this corresponds to meaningfully different treatment durations (and/or periods of time lapsed before evaluation), this would indicate different definitions of interventions across studies, and thus a violation of the consistency assumption. *Positivity*: All subjects in the study were indicated to receive any of the treatments evaluated.

8 Summary

In this paper, we nonparametrically define the parameter of interest in a network meta-analysis with direct and indirect comparisons using the counterfactual framework often employed in causal inference. This definition of the parameter of interest is model-independent and is interpretable on what we define as a metapopulation, the union of all superpopulations. Such an approach allows for a straight-forward description of what is being estimated, which is accessible even without an understanding of the estimation methods being used. In particular, we can interpret the marginal effects defined in this paper as the relative mean outcome had all subjects in the metapopulation been assigned to each treatment versus another. If a specific population is of interest and not represented by the metapopulation, with some conditions it may also be possible to more generally transport effect estimation, as described by Bareinboim and Pearl [41].

We have presented a set of conditions under which identifiability of the parameter of interest is possible. Identifiability allows for a clear description of when the parameter of interest can and cannot be estimated. For instance, the non-interference requirement casts doubt on the synthesis of studies that allow for treatment

switching, crossover, or group contamination. The assumptions that we made allowed for the simplification of the relevant components of the observed data likelihood so that arm-based inference is possible.

One might alternatively specify the RCT-estimated contrast as the "outcome of interest" (rather than use the arm-specific outcome as we did). However, under this alternative, the propensity score would then be defined as the probability of a trial directly contrasting a given treatment pair. For standard network meta-analysis sample sizes, this would most often produce practical positivity problems, indicating the need for extrapolation using the outcome model (and thereby creating estimators that are very sensitive to model misspecification). In particular, two treatments that had never been directly compared would have no data support in this model.

If all treatments are selected completely at random into studies (or if only two treatments have ever been available to compare) then a standard unadjusted analysis using those arms assigned the desired treatments would be consistent. If we weaken this assumption and replace it with conditional exchangeability, then the estimators introduced in this paper are appropriate in that they allow for the adjustment of study-level covariates.

Our methods also allow for a wider inclusion criteria of studies in a systematic review. It is often the case that systematic reviews will exclude studies because they do not evaluate the exact desired clinical endpoint. Using our proposed methods, we can avoid selection bias due to studies excluded only for this reason. To do so, we would artificially censor the outcomes of studies that do not estimate the desired outcome-type of interest. The censored outcomes of these studies might then be considered "missing at random" conditional on the study baseline information which should still be included in the analysis (both in the propensity score model and the missing data model).

For the analysis of continuous individual-level outcomes, we assumed independence between the sample mean and standard deviation within each study arm. While we chose to present our identifiability argument under this assumption, it is not ultimately necessary. However, it is not straight-forward to propose a valid Monte Carlo or Bayesian estimation approach to the setting with dependent sample means and standard deviations. In some cases, it may be possible to transform the individual-level data to remove the skew, but this relies on access to each study's raw data, in which case an individual patient data analysis would be preferable.

In the simulation study, we show that certain estimators adopted from the causal inference literature can produce valid estimates of effect contrasts under the identifiability conditions described. In particular, G-Computation and TMLE might lend themselves well to network meta-analysis, which is characterized by small sample sizes and low prevalence for certain treatments. IPTW was seen to be sensitive to rare treatment assignment and G-Computation and TMLE were seen to be somewhat sensitive to model misspecification. Some general benefits of using TMLE are that it is double robust and can incorporate nonparametric (or machine learning) estimation of the propensity score and outcome model which can help avoid bias from model misspecification vander Laan and Rose [32]. More methods development and investigations are needed to address extremely rare treatments and how (or whether) TMLE can be adapted to be robust in this setting.

The application we presented compared the results of random effects regression, G-Computation, and TMLE in a network meta-analysis of the relative efficacy of treatment options for MRSA infection. The random effects regression and G-Computation produced small confidence intervals relative to the direct contrasts of the individual RCTs though TMLE only did for one comparison investigated. In contrast to the analysis in the original article that used unadjusted contrast-based hierarchical Bayesian modeling on the separate subpopulations of infection types, our analyses concluded that there is evidence to support the superiority of linezolid over vancomycin. We also noted the poor stability of IPTW in this example and generally do not recommend this estimator when the data support for certain treatment levels is sparse. Finally, using this data example, we demonstrated how the causal assumptions should be listed and critiqued in order to stimulate discussion about the appropriateness of causal interpretations in specific contexts.

The framework we present formally assumes that we are restricting our analyses to studies evaluating a common parameter-type. If there was only partial-adherence in the RCTs, our framework does not allow for the mixing of intent-to-treat parameter estimates with adherence-adjusted parameter estimates. [Estimation of the adherence-adjusted parameters in RCTs is described in [42]. The same restriction applies to the results of observational studies if the parameter type estimated in the observational study is not the same as in the clinical trials. Specifically, treatment adherence and outcome need to be defined identically across studies, and all studies whose endpoints are included must estimate the same mean treatment-specific counterfactual outcome. Although it is common practice to include different parameter types in a meta-analysis, our formalization of the target parameter reveals that a causal interpretation of the resulting effect estimate may be quite challenging.

In addition to the issues we describe, there are many other concerns about aggregating study results in various settings. For instance, one might question the independence between RCTs happening close in time, or the systematic review inclusion criteria. We believe our framework provides additional structure to the ongoing discussion about the validity of network meta-analysis and will help stimulate solutions to the remaining challenges.

References

1. Slavin RE. Best evidence synthesis. An intelligent alternative to meta-analysis. J Clin Epidemiol 1995;48:9–18.
2. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2004;21:2313–2324.
3. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105–3124.
4. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. BMJ 2005;331:897–900.
5. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res 2008;17:279–301.
6. Berlin JA, Golub RM. "Meta-analysis as evidence: Building a better pyramid. J Am Med Assoc 2014;312:603–606.
7. Salanti G, Marinho V, Higgins JPT. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. J Clin Epidemiol 2009;62:857–864.
8. Jansen JP, Schmid CH, Salanti G. Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons. J Clin Epidemiol 2012;65:798–807.
9. Bally M, Dendukuri N, Sinclair A, Ahern SP, Poisson M, Brophy J. A network meta-analysis of antibiotics for treatment of hospitalised patients with suspected or proven meticillin-resistant Staphylococcus aureus infection. Int J Antimicrob Agents 2012;40:479–495.
10. Robins JM. Confidence intervals for causal parameters. Stat Med 1988;7:773–785.
11. Dias S, Sutton AJ, Ades AE, Welton NJ. A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. ed Decis Making 2013a;33:607–617.
12. Zhang J, Carlin BP, Neaton JD, Soon GG, Nie L, Kane R. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. Clin Trials 2014;11:246–262.
13. Cope S, Zhang J, Saletan S, Smiechowski B, Jansen JP, Schmid P. A process for assessing the feasibility of a network meta-analysis: A case study of everolimus in combination with hormonal therapy versus chemotherapy for advanced breast cancer. BMC Med 2014;12(93).
14. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006;101:447–459.
15. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: Heterogeneitysubgroups, meta-regression, bias, and bias-adjustment. Med Decision Making 2013b;33:618–640.
16. Jansen PJ, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: An ispor-amcp-npc good practice task force report. Value in Health 2014;17:157–173.
17. Welton NJ, Soares MO, Palmer S, Ades AE, Harrison D, Shankar-Hari M. Accounting for heterogeneity in relative treatment effects for use in cost-effectiveness models and value-of-information analyses. Med Decision Making 2015;35:608–621.
18. Hong H, Chu H, Zhang J, Carlin BP. Rejoinder to the discussion of "A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons," by S. Dias and A.E. Ades. Res Synth Methods 2016;7:29–33.
19. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 1984;71:431–444.
20. Zhang J, Chu H, Hong H, Virnig BA, Carlin BP. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. Stat Methods Med Res 2015. DOI:10.1177/0962280215596185.
21. Pearl J. Causality: Models, Reasoning, and Inference, 2nd ed. . New York, NY: Cambridge University Press, 2009.
22. Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. Biometrics 2015;71:15–24.
23. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. Math Model 1986;7:1393–512.
24. Ferguson TS. Texts in statistical science. A course in large sample theory. London, UK: Chapman & Hall/CRC, 1996.
25. Rubin DB. Randomization analysis of experimental data: The fisher randomization test comment. J Am Stat Assoc 1980;75:591–593.
26. Imbens GW. The role of the propensity score in estimating dose-response functions. Biometrika 2000;87:706–710.
27. Cole SR, Frangakis CE. The consistency statement in causal inference: A definition or an assumption? Epidemiology 2009;20:3–5.
28. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. Journal of Causal Inference 2013;1:1–20.
29. Snowden JM, Rose S, Mortimer KM. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. Am J Epidemiol 2011;173:731–738.
30. Efron B, Tibshirani RJ. Monographs on Statistics and Applied Probability. An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC, 1994.
31. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat 2006;2. Article 11.
32. van der Laan MJ, Rose S. Springer Series in Statistics. Targeted Learning: Causal Inference for Observational and Experimental Data. New York, NY: Springer, 2011.
33. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Int J Biostat 2010;6. Article 26.
34. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008;168:656–664.
35. Schnitzer ME, Moodie EEM, Platt RW. Targeted maximum likelihood estimation for marginal time-dependent treatment effects under density misspecification. Biostatistics 2013;14:1–14.
36. Porter KE, Gruber S, van der Laan MJ, Sekhon JS. The relative performance of targeted maximum likelihood estimators. Int J Biostat 2011;7:1–34.
37. Liu C, Bayer A, Cosgrove SE, Daum RS, Fridkin SK, Gorwitz RJ. Clinical practice guidelines by the infectious diseases society of america for the treatment of methicillin-resistant staphylococcus aureus infections in adults and children. Clin Infect Dis 2011;52. e18–e55.
38. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22. http://www.jstatsoft.org/v33/i01/.

39. Lipsky BA, Itani KM, Weigelt JA, Joseph W, Paap CM, Reisman A. The role of diabetes mellitus in the treatment of skin and skin structure infections caused by methicillin-resistant staphylococcus aureus: Results from three randomized controlled trials. Int J Infect Dis 2011;15. e140–e146.
40. Niederman MS. Hospital-acquired pneumonia, health care-associated pneumonia, ventilator-associated pneumonia, and ventilator-associated tracheobronchitis: Definitions and challenges in trial design. Clin Infect Dis 2010;51(Suppl 1). S12–S7.
41. Bareinboim E, Pearl J. Meta-transportability of causal effects: A formal approach. Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, 2013.
42. Hernán MA, Hernández-Díaz S. Beyond the intention to treat in comparative effectiveness research. Clin trials 2012;9:48–55.
43. Tsiatis AA. Springer Series in Statistics. Semiparametric Theory and Missing Data. New York, NY: Springer, 2006.
44. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. Springer series in statistics. New York: Springer Verlag, 2003.

A Appendix

A.1 Proof of identifiability under structural independence

The joint counterfactual distribution can be decomposed as $f(O_i^a) = f_1(O_i^a)f_2(O_i^a)$ where $f_1(O_i^a) = Q_W(W_i)Q_{\bar{Y}}(\bar{Y}_{ij}^a|N_{ij}^a, W_i)$ and

$$f_2(O_i^a) = Q_n(n_i|W_i)g_{A\setminus j}(A_{i\setminus j}^a|n_i, W_i)Q_N(N_{ij}^a|A_{i\setminus j}^a, n_i, W_i)Q_S(S_{ij}^a|N_{ij}^a, W_i)\times \prod_{j^*\neq j} Q_{\bar{Y},S}(\bar{Y}_{ij^*}^a, S_{ij^*}^a|N_{ij^*}^a, A_{ij^*}^a, W_i)Q_N(N_{ij^*}^a|A_{i\setminus j}^a, n_i, W_i).$$

Let A be the set of possible treatments. The target of our analysis is the study arm counterfactual outcome under treatment a, or $E(\bar{Y}_{ij}^a) = M^a$. This mean can be written as

$$\begin{aligned}
&\int_{W_i} \sum_{n_i=1}^{\infty} \sum_{A_{i\setminus j}\in\{A\setminus a\}} \sum_{N_{i\setminus j}=1}^{\infty} \sum_{N_{ij}=1}^{\infty} \int_0^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty} \bar{Y}_{ij}^a f(O_i^a) d\bar{Y}_{ij}^a dS_{ij}^a d\bar{Y}_{ij^*}^a dS_{ij^*}^a dW_i \\
&= \int_{W_i} \sum_{n_i=1}^{\infty} \sum_{A_{i\setminus j}\in\{A\setminus a\}} \sum_{N_{i\setminus j}=1}^{\infty} \sum_{N_{ij}=1}^{\infty} \int_0^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty} \bar{Y}_{ij}^a Q_{\bar{Y}}(\bar{Y}_{ij}^a|N_{ij}^a, W_i) d\bar{Y}_{ij}^a \times f_2(O_i^a)Q_W(W_i) dS_{ij}^a d\bar{Y}_{ij^*}^a dS_{ij^*}^a dW_i \\
&= \int_{W_i} \sum_{n_i=1}^{\infty} \sum_{A_{i\setminus j}\in\{A\setminus a\}} \sum_{N_{i\setminus j}=1}^{\infty} \sum_{N_{ij}=1}^{\infty} \int_0^{\infty}\int_{-\infty}^{\infty}\int_0^{\infty} E(\bar{Y}_{ij}^a|N_{ij}^a, W_i) f_2(O_i^a)Q_W(W_i) dS_{ij}^a d\bar{Y}_{ij^*}^a dS_{ij^*}^a dW_i
\end{aligned}$$

where the integral for W_i can be a multiple integral, taken over the domain of potentially multivariate W_i. Now we note that for identically distributed and conditionally independent draws Y_{ijk}

$$E(\bar{Y}_{ij}^a|N_{ij}^a, W_i) = E(\frac{1}{N_{ij}^a}\sum_{k=1}^{N_{ij}^a} Y_{ijk}^a|N_{ij}^a, W_i) = E(Y_{ijk}^a|W_i)$$

because we assume that the study size has no effect on the individual-level outcome. It follows that $E(\bar{Y}_{ij}^a|N_{ij}^a, W_i)$ is conditionally independent of N_{ij}^a. The expression for M^a then simplifies to $\int_{W_i} E(\bar{Y}_{ij}^a|N_{ij}^a, W_i)Q_W(W_i)dW_i$. In order for the conditional expectation to be estimable from the observed data, we require the unconfoundedness assumption $\bar{Y}_{ij}^a \perp\!\!\!\perp A_{ij} = a|N_{ij}^a, W_i$. With respect to the example DAG in Figure 1(a), this corresponds to having measured all components of $W2_i$. If this assumption holds in addition to the consistency of treatment for $\bar{Y}_{ij}$, we may write $M^a = \int_W E(\bar{Y}_{ij}|W_i, A_{ij} = a)Q_W(W_i)dW$ to establish identifiability.

A.2 Identifiability without assuming structural independence

It may not be plausible to assume conditional independence between $\bar{Y}_{ij}^a$ and S_{ij}^a. In this case, the relevant part of the distribution of the observed data counterfactuals is

$$f_3(O_i^a) = Q_W(W_i)Q_n(n_i|W_i)g_{A\setminus j}(A_{i\setminus j}^a|n_i, W_i)Q_N(N_{ij}^a|A_{i\setminus j}^a, n_i, W_i)Q_{\bar{Y},S}(\bar{Y}_{ij}^a, S_{ij}^a|N_{ij}^a, W_i).$$

The target parameter can be estimated as a multiple integral over each $\bar{Y}_{ij}^a$ and each density component in $f_3(O_i^a)$. Identifiability in this case requires a list of unconfoundedness assumptions: $A_{i\setminus j}^a \perp\!\!\!\perp A_{ij} = a|n_i, W_i$, $N_{ij}^a|A_{ij} = a|A_{i\setminus j}^a, n_i, W_i$, and $\bar{Y}_{ij}^a, S_{ij}^a \perp\!\!\!\perp A_{ij} = a|N_{ij}^a, W_i$. Assuming the DAG in Figure 1(a) in the main manuscript, this requires having measured all components of $W1_i$, $W2_i$, and $W3_i$. It also requires the consistency assumption for $A_{i\setminus j}$, N_{ij}, $\bar{Y}_{ij}$ and S_{ij}. Under these assumptions, we can rewrite the relevant density component as

$$f_3(O_i^a) = Q_W(W_i)Q_n(n_i|W_i)g_{A\setminus j}(A_{i\setminus j}|A_{ij} = a, n_i, W_i)Q_N(N_{ij}|A_{ij} = a, A_{i\setminus j}, n_i, W_i)\times$$
$$Q_{\bar{Y},S}(\bar{Y}_{ij}, S_{ij}|A_{ij} = a, N_{ij}, W_i).$$

Since each component of this density is estimable from the data, we have identifiability of the target parameter in this case as well.

A.3 Efficiency and Consistency of TMLE

The local semiparametric efficiency and estimation consistency of the TMLE we describe can be derived very similarly to the standard observational data setting (with a single categorical exposure variable) for the estimation of the average treatment effect [32]. To give more insight into how this extends to the network meta-analysis case, we present some additional details and a proof of double robustness.

The efficient influence function for parameter of interest M^a with only aggregate observed data is

$$D_{ij}^*(O) = \left\{\bar{Y}_{ij} - E(\bar{Y}_i^a|W_i, a \in A_i)\right\}\frac{\mathbb{I}(a \in A_i)}{g_a(W_i)} + E(\bar{Y}_i^a|W_i, a \in A_i) - M^a.$$

Note that the TMLE update step produces values of $\hat{\bar{Y}}_{ij}^{a,*}$ that solve the empirical efficient influence function equation:

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i}(\bar{Y}_{ij} - \hat{\bar{Y}}_{ij}^{a,*})\frac{\mathbb{I}(a \in A_i)}{g_a(W_i)} + (\hat{\bar{Y}}_{ij}^{a,*} - \hat{M}_{TMLE}^a) = 0$$

so that it follows that the TMLE is a locally efficient estimator [43, 44]. Specifically, the logistic regression update step with single covariate $X_i = \mathbb{I}(a \in A_i)/g_a(W_i)$ solves the score equation $\sum_{i=1}^{N}\sum_{j=1}^{n_i} X_i(\bar{Y}_{ij} - \hat{\bar{Y}}_{ij}^{a,*}) = 0$ and in the last TMLE step we set $\hat{M}_{TMLE}^a = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\hat{\bar{Y}}_{ij}^{a,*}$.

First suppose that for increasing values of $\sum_{i=1}^{N}\mathbb{I}(a \in A_i)$, the generalized propensity score $g_a(W_i)$ converges to some $\tilde{g}_a(W_i) \neq P(a \in A_i|W_i)$ but that $\hat{\bar{Y}}_{ij}^{a,*}$ converges to the true values $E(\bar{Y}_{ij}|W_i, a \in A_i)$. We then have that

$$\begin{aligned}
&E\left[\left\{\bar{Y}_{ij} - E(\bar{Y}_i^a|W_i, a \in A_i)\right\}\times\frac{\mathbb{I}(a \in A_i)}{\tilde{g}_a(W_i)} + E(\bar{Y}_i^a|W_i, a \in A_i) - M^a\right]\\
&= E\left[E\left\{\bar{Y}_{ij} - E(\bar{Y}_i^a|W_i, a \in A_i)|W_i, a \in A_i\right\}\times\frac{\mathbb{I}(a \in A_i)}{\tilde{g}_a(W_i)} + E(\bar{Y}_i^a|W_i, a \in A_i) - M^a\right]\\
&= E\left[0\times\frac{\mathbb{I}(a \in A_i)}{\tilde{g}_a(W_i)}\right] + 0 = 0
\end{aligned}$$

Now suppose that $g_a(W_i)$ converges to the true values $P(a \in A_i|W_i)$ but that $\hat{\bar{Y}}_{ij}^{a,*}$ converges to some function $\tilde{Q}_a(W_i) \neq E(|Y_{ij}|W_i, a \in A_i)$. We then have that

$$
\begin{aligned}
&E\left[\left\{\bar{Y}_{ij}-\tilde{Q}_a(W_i)\right\}\times\frac{\mathbb{I}(a\in A_i)}{P(a\in A_i|W_i)}+\tilde{Q}_a(W_i)-M^a\right]\\
&=E\left[\left\{\bar{Y}_{ij}-\tilde{Q}_a(W_i)\right\}\times E\left\{\frac{\mathbb{I}(a\in A_i)}{P(a\in A_i|W_i)}\right\}+\tilde{Q}_a(W_i)-M^a\right]\\
&=E\left[\left\{\bar{Y}_{ij}-\tilde{Q}_a(W_i)\right\}\times 1+\tilde{Q}_a(W_i)-M^a\right]\\
&=E\left[\bar{Y}_{ij}-M^a\right]=0
\end{aligned}
$$

Therefore, if either of the models for $E(\bar{Y}_i^a|W_i, a\in A_i)$ or $P(a\in A_i|W_i)$ are consistent, then the TMLE for M^a is also consistent as the efficient influence function equation is consistent for M^a.

A.4 Data extraction information and numerical results for the example of antibiotic use on methicillin-resistant *Staphylococcus aureus* infection

Table 3 presents the full study list from the systematic review of Bally et al. [9] and the data that we used in the analysis in Section 7. Table 4 presents the numerical results that we obtained from our analyses, corresponding with Figure 2. The full reference list is below.

Table 4: Risk ratio estimates, standard errors and 95% confidence intervals for relative effects of antibiotics telavancin (TEL), linezolid (LIN), and mainstay therapy vancomycin (VAN)

	TEL vs VAN			LIN vs VAN			TEL vs LIN		
Method	Est	SE	95% CI	EST	SE	95% CI			
No Adjust	1.04	0.028	(0.99,1.10)	0.92	0.027	(0.87,0.97)	1.13	0.045	(1.05,1.22)
RE Arm	1.00	0.010	(0.98,1.02)	1.08	0.012	(1.05,1.10)	0.92	0.014	(0.89,0.95)
G-Comp (RE)	1.00	0.003	(1.00,1.00)	1.06	0.006	(1.06,1.09)	0.94	0.005	(0.92,0.94)
TMLE (RE)	0.89	0.106	(0.75,1.19)	1.05	0.012	(1.03,1.07)	0.85	0.102	(0.71,1.12)

References

Arbeit R. D., Maki D., Tally F. P., Campanaro E., Eisenstein B. I. Daptomycin 98-01 and 99-01 Investigators. Clinical Infectious Diseases Vol. 38, 2004:1673–1681. The safety and efficacy of daptomycin for the treatment of complicated skin and skin-structure infections.

Breedt J., Teras J., Gardovskis J., Maritz F. J., Vaasna T., Ross D. P., Gioud-Paquet M., Dartois N., Ellis-Grosse E. J., Loh E. and Tigecycline 305 cSSSI Study Group. Safety and efficacy of tigecycline in treatment of skin and skin structure infections: Results of a double-blind phase 3 comparison study with vancomycin-aztreonam. Antimicrobial Agents and Chemotherapy 2005;49:4658–4666.

Corey G. R., Wilcox M. H., Talbot G. H., Thye D., Friedland D., Baculik T. and CANVAS 1 investigators. Canvas 1: The first phase iii, randomized, double-blind study evaluating ceftaroline fosamil for the treatment of patients with complicated skin and skin structure infections,". Journal of Antimicrobial Chemotherapy 2010;65(Suppl 4):iv41–51.

Fagon J., Patrick H., Haas D. W., Torres A., Gibert C., Cheadle W. G., Falcone R. E., Anholm J. D., Paganin F., Fabian T. C., Lilienthal F. "Treatment of gram-positive nosocomial pneumonia. prospective randomized comparison of quinupristin/dalfopristin versus vancomycin. nosocomial pneumonia group," American Journal of Respiratory and Critical Care Medicine 2000;161:753–762.

Florescu I., Beuran M., Dimov R., Razbadauskas A., Bochan M., Fichev G., Dukart G., Babinchak T., Cooper C. A., Ellis-Grosse E. J., Dartois N., Gandjini H. Efficacy and safety of tigecycline compared with vancomycin or linezolid for treatment of serious infections with methicillin-resistant staphylococcus aureus or vancomycin-resistant enterococci: a phase 3, multicentre, double-blind, randomized study,". Journal of Antimicrobial Chemotherapy 2008;62(1):i17–28. 307 Study GroupSuppl.

Itani K. M., Dryden H. M. S., Bhattacharyya M.J., Kunkel A., Baruch M., Weigelt J. A. Efficacy and safety of linezolid versus vancomycin for the treatment of complicated skin and soft-tissue infections proven to be caused by methicillin-resistant staphylococcus aureus. American Journal of Surgery 2010;199:804–816.

Katz D. E., Lindfield K. C., Steenbergen J. N., Benziger D. P., Blackerby K. J., Knapp A. G., Martone W. J. "A pilot study of high-dose short duration daptomycin for the treatment of patients with complicated skin and skin structure infections caused by gram-positive bacteria," International Journal of Clinical Practice 2008;62:1455–1464.

Rubinstein E., Cammarata S., Oliphant T., Wunderink R. and Linezolid Nosocomial Pneumonia Study Group. Linezolid (pnu-100766) versus vancomycin in the treatment of hospitalized patients with nosocomial pneumonia: A randomized, double-blind, multicenter study. Clinical Infectious Diseases 2001;32:402–412.

Rubinstein E., Lalani T., Corey G. R., Kanafani Z. A., Nannini E. C., Rocha M. G., Rahav G., Niederman M. S., Kollef M. H., Shorr A. F., Lee P. C., Lentnek A. L., Luna C. M., Fagon J. Y., Torres A., Kitt M. M., Genter F. C., Barriere S. L., Friedland H. D., Stryjewski M. E., Study Group ATTAIN. "Telavancin versus vancomycin for hospital-acquired pneumonia due to gram-positive pathogens," Clinical Infectious Diseases 2011;52:31–40.

Sacchidanand S., Penn R. L., Embil J. M., Campos M. E., Curcio D., Ellis-Grosse E., Loh E., Rose G. "Efficacy and safety of tigecycline monotherapy compared with vancomycin plus aztreonam in patients with complicated skin and skin structure infections. Results from a phase 3, randomized, double-blind trial," International Journal of Infectious Diseases 2005;9:251–261.

Stryjewski M. E., Chu V. H., O'Riordan W. D., Warren B.L., Dunbar L.M., Young D.M., Vall´ee M., Fowler V.G. J., Morganroth J., Barriere S., Kitt M. M., Corey G. R. and FAST 2 Investigator Group. Telavancin versus standard therapy for treatment of complicated skin and skin structure infections caused by gram-positive bacteria: Fast 2 study. Antimicrobial Agents and Chemotherapy 2006;50:862–867.

Stryjewski M. E., Graham D. R., Wilson S. E., O'Riordan W., Young D., Lentnek A., Ross D. P., Fowler V.G., Hopkins A., Friedland H. D., Barriere S. L., Kitt M. M., Corey G. R. Assessment of Telavancin in Complicated Skin and Skin-Structure Infections Study (2008): "Telavancin versus vancomycin for the treatment of complicated skin and skin-structure infections caused by gram-positive organisms. Clinical Infectious Diseases 1683–1693;46.

Talbot G. H., Thye D., Das A., Ge Y. Phase 2 study of ceftaroline versus standard therapy in treatment of complicated skin and skin structure infections,". Antimicrobial Agents and Chemotherapy 2007;51:3612–3616.

Wilcox M. H., Corey G. R., Talbot G. H., Thye D., Friedland D., Baculik T. and CANVAS 2 investigators. Canvas 2: The second phase iii, randomized, double-blind study evaluating ceftaroline fosamil for the treatment of patients with complicated skin and skin structure infections,". Journal of Antimicrobial Chemotherapy 2010;65(4):iv53–65. Suppl.

Wunderink R. G., Cammarata S. K., Oliphant T. H., Kollef M. H. and Linezolid Nosocomial Pneumonia Study Group. Continuation of a randomized, double-blind, multicenter study of linezolid versus vancomycin in the treatment of patients with nosocomial pneumonia. Clinical Therapeutics 2003;25:980–992.

Wunderink R. G., Mendelson M. H., Somero M. S., Fabian T. C., May A. K., Bhattacharyya H., Leeper K. V. J., Solomkin J. S. Early microbiological response to linezolid vs vancomycin in ventilator-associated pneumonia due to methicillin-resistant staphylococcus aureus. Chest 2008;134:1200–1207.

Wunderink R. G., Niederman M. S., Kollef M. H., Shorr A. F., Kunkel M. J., Baruch A., McGee W. T., Reisman A., Chastre J. "Linezolid in methicillin-resistant staphylococcus aureus nosocomial pneumonia. A randomized, controlled study," Clinical Infectious Disease 2012;54:621–629.