Population-Adjusted Indirect Treatment Comparisons with Limited Access to Patient-Level Data

Antonio Remiro Azócar

Thesis for Submission at University College London
Research Degree: Statistical Science

DECLARATION

I, Antonio Remiro Azócar, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Antonio Remiro Azócar

Date: May 20, 2021

Para mi madre, María. Estoy orgulloso de ser hijo tuyo.

ACKNOWLEDGEMENTS

This thesis is the result of thousands of infinitesimal steps, which I hope comprise something much greater in aggregate. I have taken these steps thanks to my parents: my mother María and my father Antonio. My mother has been a source of unwavering support. My father has been a constant source of inspiration. They both gave me the intellectual curiosity, education and tools I needed to succeed in the most ambitious project that I have ever set out to achieve.

Probably, one of the best decisions of my life has been selecting Gianluca Baio and Anna Heath as my PhD supervisors. Their attitude has brought deep meaning to my research. I remember having a chat with Gianluca for the first time and realizing that we got along. I really can't thank him enough for his mentorship. There are so many ways he's gone above and beyond what I could have ever reasonably asked for from a supervisor. Anna has also been fantastic. She has always been very constructive in her feedback and has pushed me on my ideas. Despite their kind criticism, the two have always been supportive of what I think, and have made it easier to push through the setbacks.

Many impressive people have contributed to the quality of work in this thesis. Tim Morris provided very helpful comments after evaluating my proposal defense. His comments on G-computation and marginalization helped motivate Chapter 4 of the thesis. I am also very grateful to Anthony Hatswell for discussions during the first year of my PhD that contributed to the direction I ended up taking. I am hugely thankful to the editors and anonymous peer reviewers of the articles that I have submitted for publication. Their comments have been extremely insightful and have considerably improved the underlying motivation of this dissertation. I am grateful to David Phillippo, who has provided very valuable feedback to this work, and has helped in substantially improving my research. Finally, I acknowledge Andrea Gabrio's advice and expertise in missing data, which was very helpful for Chapter 4.

Thank you to the United Kingdom Doctoral Training Centre in Financial Computing and Analytics (Philip Treleaven and Yonita Carter), and to governmental institutions such as the Engineering and Physical Sciences Research Council of the United Kingdom. This research was possible thanks to their funding, in the form of a PhD scholarship. A special thank you to Yonita for her commitment towards making our experience at UCL more enjoyable and for helping me out when dealing with some tribulations at the beginning of the program. Speaking of tribulations, I am particularly thankful to Steven Davies at the University of Bath and Steven Allen at Runnymede for believing in my academic potential in crucial moments of my life and education.

My acknowledgements also go to the organizations, and people in these organizations, that supported me economically, employing me as a statistical consultant during my PhD. My time at IQVIA, with Andreas Karabis, Ines Guerra and Adam Lloyd, was highly enriching because it gave me exposure to the practical problems addressed in this research. So was my time

at ICON plc, with Victoria Paly, and at the Hospital for Sick Children (SickKids), with Petros Pechlivanoglou.

Finally, I cannot imagine doing something like this without amazing people by my side to share the triumphs with. Many friends were not at the UCL Department of Statistical Science with me but made my life a lot more fun during different stages of the PhD. Thank you to Salvador for dragging me out to drinks. Thank you to my cousin Ana for her huge support. In no particular order, many thanks to Seyd, Tim, Marina, Cassandra and Sam.

ABSTRACT

Health technology assessment systems base their decision-making on health-economic evaluations. These require accurate relative treatment effect estimates for specific patient populations. In an ideal scenario, a head-to-head randomized controlled trial, directly comparing the interventions of interest, would be available. Indirect treatment comparisons are necessary to contrast treatments which have not been analyzed in the same trial.

Population-adjusted indirect comparisons estimate treatment effects where there are: no head-to-head trials between the interventions of interest, limited access to patient-level data, and cross-trial differences in effect measure modifiers. Health technology assessment agencies are increasingly accepting evaluations that use these methods across a diverse range of therapeutic areas. Popular approaches include matching-adjusted indirect comparison (MAIC), based on propensity score weighting, and simulated treatment comparison (STC), based on outcome regression. There is limited formal evaluation of these methods and whether they can be used to accurately compare treatments. Thus, I undertake a review and a simulation study that compares the standard unadjusted indirect comparisons, MAIC and STC across 162 scenarios.

This simulation study assumes that the trials are investigating survival outcomes and measure continuous covariates, with the log hazard ratio as the measure of effect — one of the most widely used setups in health technology assessment applications. MAIC yields unbiased treatment effect estimates under no failures of assumptions. The typical usage of STC produces bias because it targets a conditional treatment effect where the target estimand should be a marginal treatment effect. The incompatibility of estimates in the indirect comparison leads to bias as the measure of effect is non-collapsible. When adjusting for covariates, one must integrate or average the conditional model over the population of interest to recover a compatible marginal treatment effect.

I propose a marginalization method based on parametric G-computation that can be easily applied where the outcome regression is a generalized linear model or a Cox model. In addition, I introduce a novel general-purpose method based on the ideas underlying multiple imputation, which is termed multiple imputation marginalization (MIM) and is applicable to a wide range of models, including parametric survival models. The approaches view the covariate adjustment regression as a nuisance model and separate its estimation from the evaluation of the marginal treatment effect of interest. Both methods can accommodate a Bayesian statistical framework, which naturally integrates the analysis into a probabilistic framework, typically required for health technology assessment.

Another simulation study provides proof-of-principle for the methods and benchmarks their performance against MAIC and the conventional STC. The simulations are based on scenarios with binary outcomes and continuous covariates, with the log-odds ratio as the measure of

effect. The marginalized outcome regression approaches achieve more precise and more accurate estimates than MAIC, particularly when covariate overlap is poor, and yield unbiased marginal treatment effect estimates under no failures of assumptions. Furthermore, regression-adjusted estimates of the marginal effect provide greater precision and accuracy than the conditional estimates produced by the conventional STC, which are systematically biased because the log-odds ratio is a non-collapsible measure of effect.

The marginalization methods outlined in this thesis are necessary and important for health technology assessment more generally, because marginal treatment effects should be the preferred inferential target for reimbursement decisions at the population level. Treatment effectiveness inputs in health economic models are often informed by the treatment coefficient of a multivariable regression. An often overlooked issue is that this has a conditional interpretation, and that the coefficients of the regression must be marginalized over the target population of interest to produce a relevant estimate for reimbursement decisions at the population level.

IMPACT STATEMENT

There are two sides to the story told by this thesis. One addresses a substantive problem in health technology assessment, which is the application of population-adjusted indirect comparisons. These are regularly used to adjust for cross-trial differences in covariates where there is limited access to patient-level data. The other side of the story is methodological. It highlights the importance of carefully considering whether a marginal or conditional treatment effect is of interest in health technology assessment.

The initial objective of the thesis was to investigate the former. However, the latter ensued when evaluating the distinct methodologies for population-adjusted indirect comparisons. Different methodologies estimate different measures of effect, yet marginal effects should be the preferred inferential target for decisions at the population level. The typical usage of regression adjustment, in the context of indirect treatment comparisons, targets a conditional treatment effect. I propose marginalization methods that make the effect estimate compatible in indirect treatment comparisons and relevant for population-level decision-making. The methods are not limited to the primary context of this thesis and are applicable, more generally, in health technology assessment.

The findings of this thesis can help stakeholders evaluate the value of new health technologies. Potential stakeholders include a wide range of organizations and professionals within these organizations:

- National and local regulatory bodies, and health technology assessment agencies. The research advances some of the latest concepts and techniques in health technology assessment, and can be used to update reporting standards and identify best practices.
- Small and large pharmaceutical, biotechnology, or medical device companies, with the research describing statistical developments in health technology assessment, and providing recommendations on potential methodological and strategic initiatives.
- Contract research organizations can benefit from this research to help unlock perspectives and generate insights for their clients.

OUTPUTS

Papers

- **Remiro-Azócar, A.**, Heath, A., and Baio, G. "Methods for Population Adjustment with Limited Access to Individual Patient Data: A Review and Simulation Study". *Research Synthesis Methods*, 12(6), 2021. Available at: https://doi.org/10.1002/jrsm.1511
- **Remiro-Azócar, A.**, Heath, A., and Baio, G. "Marginalization of Regression-Adjusted Treatment Effects in Indirect Comparisons with Limited Patient-Level Data". Working Paper. Available at: https://arxiv.org/abs/2008.05951
- **Remiro-Azócar, A.**, Heath, A., and Baio, G. "Parametric G-computation for Compatible Indirect Treatment Comparisons with Limited Individual Patient Data". Submitted to Research Synthesis Methods. Available at: https://arxiv.org/abs/2108.12208
- **Remiro-Azócar, A.**, Heath, A., and Baio, G. "Conflating marginal and conditional treatment effects: Comments on "Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study"". *Statistics in Medicine*, 40(11), 2021. Available at: https://doi.org/10.1002/sim.8857
- **Remiro-Azócar, A.**, Heath, A., and Baio, G. "Effect modification in anchored indirect treatment comparisons: Comments on "Matching-adjusted indirect comparisons: Application to time-to-event data"". *In press, Statistics in Medicine*. Available at: https://arxiv.org/abs/2012.05127
- **Remiro-Azócar, A.** "Target estimands in population-adjusted indirect comparisons". Submitted to Statistics in Medicine. Available at: https://arxiv.org/abs/2112.08023
- **Remiro-Azócar, A.**, Heath, A., and Pechlivanoglou, P. "A catalogue of assumptions and potential sources of bias in matching-adjusted indirect comparisons". Report commissioned by the Canadian Agency for Drugs and Technologies in Health (CADTH).
- van Oostrum, I., Ouwens, M., **Remiro-Azócar, A.**, Baio, G. Postma, M., Buskens, E., and Heeg, B. "Comparison of parametric survival extrapolation approaches incorporating general population mortality for adequate health technology assessment of new oncology drugs". *Value in Health*, 24 (9), 2021. Available at: https://doi.org/10.1016/j.jval.2021.03.008

Poster presentations

- Mohr, P., Larkin, J., Paly, V. F., **Remiro-Azócar, A.**, Baio, G., Kurt, M., Amadi, A., Rizzo, J.I., Johnson, H.M., Moshyk, A., Kotapati, S., and Middleton, M. "Estimating long-term survivorship in patients with advanced melanoma treated with immune-checkpoint inhibitors: Analyses from the phase III CheckMate 067 trial". Presented at the ESMO Virtual Congress 2020.
- Paly, V. F., Mohr, P., Larkin, J., Middleton, M., Youn, J., **Remiro-Azócar, A.**, Baio, G., Moshyk, A., Kotapati, S., Hamilton, M., Kurt, M. "Assessing the impact of modeling non–disease-related mortality on long-term survivorship rates in previously untreated advanced melanoma: a case study from CheckMate 067". Presented at Virtual ISPOR 2021.
- **Remiro-Azócar, A.**, Heath, A., and Baio, G. "Predictive-adjusted indirect comparison (PAIC): A novel method for population-adjusted indirect comparison". Presented at ISPOR Europe 2019 in Copenhagen, Denmark.

Invited presentations

- R for Health Technology Assessment (R-HTA) Workshop — July 2021
- UCL Priment Clinical Trials Unit Statistical Seminar — November 2020
- UCL Statistics for Health Economic Evaluation Seminar — June 2020
- Health Economics Study Group Winter Meeting, Newcastle, UK — January 2020
- Spanish Health Economics Association Conference, Albacete, Spain — June 2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ADEMP Aims, data-generating mechanisms, estimands, methods, and performance measures

ALD Aggregate-level data

ANCOVA Analysis of covariance

DAG Directed acyclic graph

EMA European Medicines Agency

ESE Empirical standard error

ESS Effective sample size

FDA United States Food and Drug Administration

HTA Health technology assessment

ICER Incremental cost-effectiveness ratio

INLA Integrated nested Laplace approximation

IPD Individual patient data

MAE Mean absolute error

MAIC Matching-adjusted indirect comparison

MCMC Markov chain Monte Carlo

MCSE Monte Carlo standard error

MI Multiple imputation

MIM Multiple imputation marginalization

ML Maximum-likelihood

ML-NMR Multilevel network meta-regression

MSE Mean square error

NICE National Institute for Health and Care Excellence

RCT Randomized controlled trial

STC Simulated treatment comparison

SUTVA Stable unit treatment value assumption

TA Technology appraisal

UCL University College London

1

CHAPTER 1: INTRODUCTION

1.1 MOTIVATION FOR THE THESIS

The development of novel pharmaceuticals requires several stages, which include regulatory evaluation and, in several jurisdictions (including the United Kingdom), health technology assessment (HTA) [1]. To obtain regulatory approval at the licensing stage, a new technology must demonstrate efficacy and randomized controlled trials (RCTs) are the most reliable design for this purpose [2], due to their potential in limiting bias in the study sample [3]. Evidence supporting regulatory approval is often provided by a two-arm RCT, typically comparing the new technology to placebo or standard of care, but not necessarily against other active interventions. Then, in certain jurisdictions, HTA addresses whether the health care technology should be publicly funded by the health care system. For HTA, manufacturers must convince payers that their product offers the best "value for money" of all available options in the market. This demands more than a demonstration of efficacy [4] and will often require comparing the effectiveness of treatments that have not been trialed against each other [5].

This evaluation of alternative health care interventions lies at the heart of HTAs, such as those commissioned by the National Institute of Health and Care Excellence (NICE),[1] the body responsible for providing guidance on whether health care technologies should be publicly funded by the National Health Service in England and Wales [5]. Other well-regarded HTA agencies issuing recommendations include the Canadian Agency for Drugs and Technologies in Health, and the Pharmaceutical Benefits Advisory Committee in Australia.

In the absence of data from head-to-head RCTs, indirect treatment comparisons (ITCs) are at the top of the hierarchy of evidence when assessing the relative effects of interventions and can inform treatment and reimbursement decisions, being very prevalent in HTA [6]. Standard ITCs use indirect evidence obtained from RCTs through a common comparator arm [6, 7]. These techniques are compatible with both individual patient data (IPD) and aggregate-level data (ALD), with IPD considered the gold standard [8]. However, they assume that there are no cross-trial differences in the distributions of effect measure modifiers, i.e., that relative treatment effects are constant across study populations. Therefore, they almost always produce biased estimates when these differences exist [9].

The motivation for the thesis is as follows. In many HTA processes, there are: (1) cross-trial imbalances in effect measure modifiers, implying that relative treatment effects are not constant

1 Originally set up as the National Institute for Clinical Excellence, which explains the NICE acronym.

across studies; (2) no head-to-head trials comparing the interventions of interest; and (3) IPD available for at least one intervention (e.g. from the trial of the manufacturer submitting evidence), but only published ALD for the relevant comparator(s). Several methods, labeled *population-adjusted indirect comparisons*, have been introduced to estimate relative treatment effects in this scenario, requiring access to IPD from at least one of the trials. These methods include matching-adjusted indirect comparison (MAIC) [10–12], based on inverse propensity score weighting [13], and simulated treatment comparison (STC) [14], based on outcome regression [15]. There is also a simpler alternative, crude direct post-stratification (also known as non-parametric standardization, subclassification or direct adjustment) [16], but this fails if any of the covariates are continuous or where there are several covariates for which one must account [17], in which case it is also statistically inefficient (i.e., inaccurate).

The NICE Decision Support Unit has published a technical support document with formal submission guidelines for population adjustment with limited access to IPD, which provide recommendations on the use of MAIC and STC in HTA [9, 18]. Various reviews [9, 18–20] define the relevant terminology and assess the theoretical validity of these methodologies but do not express a preference. Questions remain about the correct application of the methods and their validity in HTA [9, 18, 21]. Thus, Phillippo et al. [9] state that current guidance can only be provisional, as more thorough understanding of the properties of population-adjusted indirect comparisons is required.

As remarked by Phillippo et al. [9, 18], further research must: (1) examine these methods through comprehensive simulation studies; and (2) develop novel methods for population adjustment. In addition, recommendations have highlighted the importance of embedding the methods within a Bayesian framework, which allows for the principled propagation of uncertainty to the wider health economic model [22], and is particularly appealing for "probabilistic sensitivity analysis" [23], used to characterize the impact of the uncertainty in the model inputs on the decision-making process. This component is often mandatory in the normative framework of HTA bodies such as NICE [22].

Consequently, several simulation studies have been conducted since the release of the NICE technical support document to assess population-adjusted indirect comparisons [24–29]. These have primarily assessed the performance of MAIC relative to standard ITCs in a limited number of simulation scenarios. In general, the studies set relatively low covariate imbalances and do not vary these, even though MAIC is prone to imprecision when high imbalances lead to poor covariate overlap [30], i.e., where the degree of similarity in the covariate ranges across studies is low. Most importantly, existing simulation studies typically consider binary covariates at non-extreme values, not close to zero or one. In these scenarios, MAIC is likely to perform well as covariate overlap is strong.

Propensity score weighting methods such as MAIC are known to be highly sensitive to scenarios with poor overlap [31–34], in which case they are not statistically precise because of their inability to extrapolate beyond the covariate space observed in the patient-level data. With poor overlap, extreme weights may produce unstable treatment effect estimates with high variance. Hence, it is important to evaluate the performance of MAIC in the face of practical

scenarios with poor overlap between the studies' covariate distributions. A related problem in finite samples is that feasible weighting solutions may not exist [35], e.g. where sample sizes are small and the number of covariates is large [36, 37].

A recent simulation study by Phillippo et al. [34] comprehensively examines MAIC, STC and a novel method by the authors called multilevel network meta-regression [38] in practical scenarios with poor covariate overlap. However, the target estimand of the simulation study is a conditional treatment effect as opposed to a marginal treatment effect, which should be the target for population-level reimbursement decisions in HTA [39].

1.2 AIMS AND STRUCTURE OF THE THESIS

This thesis seeks to address the following research objectives:

- To review methods currently used for population-adjusted indirect comparisons, evaluating and comparing their statistical performance through comprehensive simulation studies;
- To develop novel outcome modeling methodologies that improve the performance of the existing population adjustment methods and can be embedded within a Bayesian framework;
- To influence practice by making recommendations on the way and circumstances in which population-adjusted indirect comparisons should be applied;
- To provide clarifications on what the target of the analysis, i.e. the estimand, should be for population-adjusted indirect comparisons, given that these are used to inform reimbursement decisions at the population level in HTA.

I now describe the structure of the thesis and the research questions tackled by each individual chapter. In Chapter 2, I carry out an up-to-date review of MAIC and STC. I demonstrate that MAIC and the typical usage of STC, as described by HTA guidance and recommendations [18], target different estimands. STC targets a conditional treatment effect as opposed to a marginal estimand, which is the appropriate target for HTA decisions at the population level. In addition, the conditional estimand cannot be combined in any indirect treatment comparison or compared between studies because conditional estimands vary across different covariate adjustment sets. This is a recurring problem in meta-analysis and is particularly troublesome where the measure of effect is non-collapsible [40, 41].

In Chapter 3, I conduct a comprehensive simulation study to benchmark the performance of MAIC and the typical usage of STC against the standard ITC. The simulation study provides proof-of-principle for the methods and is based on scenarios with survival outcomes, continuous covariates and the Cox proportional hazards regression as the outcome model, with the log hazard ratio as the measure of effect. This is one of the most common setups in HTA applications [30]. The methods are evaluated in a wide range of settings; varying the

trial sample size, effect-modifying strength of covariates, prognostic effect of covariates, imbalance/overlap of covariates and the level of correlation in the covariates. 162 simulation scenarios are considered. An objective of the simulation study is to inform the circumstances under which population adjustment should be applied and which specific method is preferable in a given situation.

In this simulation study, MAIC yields treatment effect estimates that are unbiased and relatively accurate, with its potential for bias reduction outweighing the loss of precision when all effect modifiers are accounted for in the adjustment. Robust sandwich standard errors slightly underestimate the variability when effective sample sizes are small. The simulation study demonstrates that the conventional version of STC produces systematically biased estimates with inappropriate coverage rates because it targets the wrong estimand, which is incompatible in the indirect comparison. As the (log) hazard ratio is non-collapsible, marginal and conditional estimands do not coincide. As a result, the typical usage of STC produces bias.

The crucial element that has been missing from the application of STC is the marginalization of treatment effect estimates. When adjusting for covariates, one must integrate or average the conditional estimates over the joint covariate distribution to recover a marginal treatment effect that is compatible in the indirect comparison. In Chapter 4, I develop several methods to accomplish this and present these methods in detail. Firstly, I propose a marginalization method based on parametric G-computation [42, 43] or model-based standardization [44–47], often applied in observational studies in epidemiology and medical research where treatment assignment is non-random. In addition, I introduce a novel general-purpose method based on the ideas underlying multiple imputation [48], which I term *multiple imputation marginalization* (MIM) and is applicable to a wide range of models, including parametric survival models.

Both parametric G-computation and multiple imputation marginalization can be viewed as extensions to the conventional STC, with all methods making use of effectively the same outcome model. The novel methodologies are outcome regression approaches, thereby capable of extrapolation, that target marginal treatment effects. They do so by separating the covariate adjustment regression model from the evaluation of the marginal treatment effect of interest. The conditional parameters of the regression are viewed as nuisance parameters, not directly relevant to the research question. The methods can be implemented in a Bayesian statistical framework, which explicitly accounts for relevant sources of uncertainty, allows for the incorporation of prior evidence (e.g. expert opinion), and naturally integrates the analysis into a probabilistic framework.

The development of outcome regression approaches that target compatible marginal treatment effects is appealing and impactful. These methodologies tend to be more efficient than weighting, providing more stable estimators [49]. MAIC is a weighting method that cannot extrapolate where the overlap between studies is insufficient. Conversely, outcome regression models can extrapolate beyond the covariate space observed in the patient-level data, thereby overcoming some limitations of MAIC. We view extrapolation as an advantage because poor overlap and small sample sizes are pervasive issues in HTA [30]. While extrapolation can also

be viewed as a disadvantage if it is not valid, in our case it expands the range of scenarios in which population adjustment can be used.

In Chapter 5, I carry out a simulation study to benchmark the performance of the novel "marginalized" outcome regression methods against MAIC and the conventional STC. The simulations provide proof-of-principle and investigate scenarios with binary outcomes and continuous covariates, with the log-odds ratio as the measure of effect. The novel approaches achieve greater precision and accuracy than MAIC and are unbiased under no failures of assumptions. Furthermore, the marginalized regression-adjusted estimates provide greater statistical precision that the conditional estimates produced by the conventional version of STC. While this precision comparison is irrelevant, because it is made for estimators of different estimands, it supports previous research on non-collapsible measures of effect [44, 50].

Finally, Chapter 6 contains some concluding remarks. I highlight the importance of carefully considering what the target estimand should be for population-adjusted indirect comparisons. I clarify why marginal treatment effects should be the preferred inferential target. Avenues for future work are also discussed. While this thesis intends to influence applied practice, I note that a real case study demonstrating the application of the methodologies is missing. I provide proof-of-principle through simulation studies and code for simulated examples for a range of population adjustment methods in the supplementary appendices. Nevertheless, the application of the methodologies to real examples is a key priority for future research.

2

CHAPTER 2: METHODS FOR POPULATION ADJUSTMENT WITH LIMITED ACCESS TO INDIVIDUAL PATIENT DATA: A REVIEW

In Section 2.1, I establish the context for population-adjusted indirect comparisons. In Section 2.2, I outline the data structure/requirements for the methods. In Sections 2.3 and Sections 2.4, I present an updated review of MAIC and STC, respectively. In Section 2.5 I clarify that each methodology targets a different estimand, something that is currently overlooked by the literature. Finally, Section 2.6 provides some brief concluding remarks. Part of the content of this chapter is included in the article "Methods for Population Adjustment with Limited Access to Individual Patient Data: A Review and Simulation Study" (Remiro-Azócar et al., 2021).[1]

2.1 CONTEXT

HTA often takes place late in the drug development process, after a new medical technology has obtained regulatory approval, typically based on a two-arm RCT that compares the new intervention to placebo or standard of care before the licensing stage. At the licensing stage, the question of interest is whether or not the drug is effective. In HTA, the relevant policy question is: "given that there are finite resources available to finance health care, which is the best treatment of all available options in the market?". In order to answer this question, one must evaluate the relative effectiveness of interventions that may not have been trialed against each other.

The following scenario is common in the appraisal of new oncology drugs. Consider an active treatment A, which needs to be compared to another active treatment B for the purposes of reimbursement. Treatment A is new and being tested for cost-effectiveness, while treatment B is typically an established intervention, already on the market. Both treatments have been evaluated in a RCT against a common comparator C, e.g. standard of care or placebo, but not against each other. Indirect treatment comparison methods are performed to estimate the relative treatment effect of A vs. B for a specific outcome. The objective is to perform the analysis that would be conducted in a hypothetical head-to-head RCT between A and B, which indirect treatment comparisons seek to emulate.

The RCT is widely considered the gold standard design to evaluate the efficacy of interventions [2] due to its high internal validity [3], i.e., its potential for limiting bias within the study

1 The article has been published in Research Synthesis Methods and is available at: https://doi.org/10.1002/jrsm.1511

sample. Appropriate randomization guarantees covariate balance on expectation, so that the treatment groups are comparable and confounding is limited. Therefore, assuming no structural issues (e.g. no dropout, informative missingness, measurement error, etc.), RCTs allow for unbiased estimation of the relative treatment effect within the study.

RCTs may target marginal or conditional estimands. These are calibrated at different hierarchical levels. The *marginal* or *population-average* effect is calibrated at the population level. It quantifies how mean outcomes change when moving all randomized individuals between two hypothetical worlds: from one where everyone receives treatment B to one where everyone receives treatment A [51–53]. Conversely, *conditional* effects are calibrated at the subgroup or individual level. A conditional effect compares average treatment outcomes when switching the treatment of an individual in the trial from B to A, fully conditioned on the average combination of subject-level covariates, or the average effect across sub-populations of patients who share the same covariate values.

The marginal effect is typically, but not necessarily, estimated by an "unadjusted" analysis. This may be a simple comparison of the expected outcomes for each group or a univariable regression including only the main treatment effect. RCTs typically report unadjusted analyses, which rely on measured and unmeasured covariates being balanced between treatment groups due to randomization.

The conditional treatment effect is often estimated as the treatment coefficient of an "adjusted" analysis, e.g., a multivariable regression of outcome on the main effects of randomized treatment and a set of baseline covariates, such as prior medical history, demographic factors or physiological status. In this analysis, the target estimand is a weighted average of (also conditional) individual-level or subgroup-specific effects. These effects are conditional on the baseline covariates that have also been included in the model. The covariates are pre-specified in the protocol or analysis plan and are likely to be prognostic variables, associated with the clinical outcome of interest.

Note that, as highlighted by Daniel et al. [50], "the words conditional and adjusted (likewise marginal and unadjusted) should not be used interchangeably". A recurring theme throughout this thesis is that marginal need not mean unadjusted because covariate-adjusted analyses may also target marginal estimands [54]. Population-adjusted indirect comparisons are used to inform reimbursement decisions in HTA at the population level, where interest lies in the impact of a health technology on the target population for the decision problem. Therefore, marginal treatment effect estimates are required [54].

The indirect comparison between treatments A and B is typically carried out on the "linear predictor" scale [6, 7]; namely, assuming additive effects for a given linear predictor, e.g. log-odds ratio for binary outcomes or log hazard ratio for survival outcomes. Indirect treatment comparisons can be "anchored" or "unanchored". Anchored comparisons make use of a connected treatment network. In this case, this is available through the common comparator C (Figure 1). Unanchored comparisons use disconnected treatment networks or single-arm trials and require much stronger assumptions than their anchored counterparts [9]. The use of unanchored comparisons where there is connected evidence is discouraged and often labeled

as problematic by HTA agencies [9, 18]. This is because it does not respect within-study randomization and is not protected from imbalances in any covariates that are prognostic of outcome (in essence implying that absolute outcomes can be predicted from the covariates, a heroic assumption). This set of covariates is, almost invariably, a larger set of covariates than the set of effect measure modifiers. Hence, our focus in this thesis is on anchored comparisons.

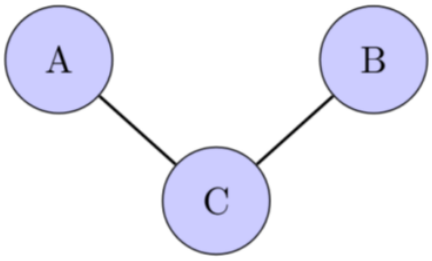


Figure 1: Diagram of the connected network in an anchored indirect comparison.

In the standard anchored scenario, a company submitting evidence for reimbursement to HTA bodies has access to patient-level data from its own trial that compares its product A against standard intervention C. However, as disclosure of proprietary, confidential patient-level data from industry-sponsored clinical trials is rare, IPD on baseline covariates, treatment and outcomes for the competitor's trial, comparing the relative efficacy or effectiveness of treatment B against C, are regularly unavailable. This is the case for both the manufacturer submitting evidence for reimbursement and the national HTA agency evaluating the evidence. For this study, only summary outcome measures and marginal moments of the covariates, e.g. means with standard deviations for continuous variables or proportions for binary and categorical variables, as found in so-called "Table 1" of clinical trial publications, are available. We consider, without loss of generality, that IPD are available for a trial comparing intervention A to intervention C (denoted AC) and published ALD are available for a trial comparing B to C (BC).

We briefly introduce some notation. Let Z denote a treatment indicator. Active treatment A is denoted $Z = 1$, active treatment B is denoted $Z = 2$, and the common comparator C is denoted $Z = 0$. In addition, consider that S denotes a specific study. The AC study, comparing treatments A and C is denoted $S = 1$. The BC study is denoted $S = 2$. The true relative treatment effect between Z and Z' in study population S is indicated by $\Delta^{(S)}_{ZZ'}$ and is estimated by $\hat{\Delta}^{(S)}_{ZZ'}$.

Standard methods for indirect comparisons such as the Bucher method [7], a special case of network meta-analysis, allow for the use of ALD to give a standard ITC. These estimate the marginal A vs. B treatment effect as:

$$\hat{\Delta}_{12} = \hat{\Delta}^{(1)}_{10} - \hat{\Delta}^{(2)}_{20}, \tag{1}$$

where $\hat{\Delta}^{(1)}_{10}$ is the estimated relative treatment effect of A vs. C in the AC population, and $\hat{\Delta}^{(2)}_{20}$ is the estimated relative treatment effect of B vs. C in the BC population. Standard ITC methods do not typically explicitly specify the target population for the A vs. B treatment effect

estimate $\hat{\Delta}_{12}$, hence the lack of superscript in the notation, regardless of whether the analysis is based on ALD or on IPD from each study [55].

The available patient-level data can be used to estimate $\hat{\Delta}_{10}^{(1)}$ and its variance, e.g. by fitting a univariable regression of outcome on treatment. The estimate $\hat{\Delta}_{20}^{(2)}$ and an estimate of its variance may be directly published or derived non-parametrically from aggregate outcomes made available in the literature. The majority of RCT publications will report an estimate $\hat{\Delta}_{20}^{(2)}$ targeting a marginal treatment effect, typically derived from a simple regression of outcome on a single independent variable, treatment assignment. In addition, the estimate $\hat{\Delta}_{12}$ should target a marginal treatment effect for reimbursement decisions at the population level. Therefore, $\hat{\Delta}_{10}^{(1)}$ should target a marginal treatment effect that is compatible with $\hat{\Delta}_{20}^{(2)}$. As the indirect comparison is based on relative effects observed in separate RCTs, the within-trial randomization of the originally assigned patient groups is preserved. The within-trial relative effects are statistically independent of each other; hence, their variances are simply summed to estimate the variance of the marginal A vs. B treatment effect.

One can also take a Bayesian approach to estimating the indirect treatment comparison, e.g. using Markov chain Monte Carlo (MCMC) simulation, in which case variances would be derived empirically from draws of the posterior density. In my opinion, a Bayesian analysis is helpful because simulation from the posterior distribution provides a framework for probabilistic decision-making, directly allowing for both statistical estimation and inference, and for principled uncertainty propagation [6].

Standard ITCs such as the Bucher method [7] assume that there are no differences across trials in the distribution of *effect measure modifiers*. A variable is an effect measure modifier, *effect modifier* for short, if the relative effect of a particular intervention on the outcome, as measured on a specific scale (e.g. the linear predictor), varies at different levels of the variable. For instance, if women react differently to a drug therapy than men on the log-odds ratio scale, then gender modifies the effect of the drug on such scale. Within the biostatistics literature, effect modification is usually referred to as heterogeneity or interaction, because effect modifiers are considered to alter the effect of treatment by interacting with it on a specific scale [56], and are typically detected by examining statistical interactions [57].

In the Bucher method, one assumes that the relative effect of A vs. C in the AC population (denoted $\Delta_{10}^{(1)}$) is equivalent to that which would have occurred in the BC population (indicated as $\Delta_{10}^{(2)}$). Again, the Bucher method and most conventional network meta-analysis methods do not explicitly specify a target population of policy interest (whether this is AC, BC or otherwise) [55]. Hence, they cannot account for differences in covariates across study populations. The Bucher method is only valid when either: (1) the A vs. C treatment effect is homogeneous, such that there is no effect modification; or (2) the distributions of the effect modifiers are the same in both studies.

If the A vs. C treatment effect is heterogeneous and the effect modifiers are not equidistributed across trials, relative treatment effects are no longer constant across the trial populations, except in the pathological case where the bias induced by different effect modifiers is in opposite directions and cancels out. Hence, the assumptions of the Bucher method are broken. In

this scenario, a standard ITC between A and B is liable to produce biased and overprecise estimates of the treatment effect [58]. These features are undesirable, particularly from the economic modeling point of view, as they impact negatively on probabilistic sensitivity analysis.

As a result, population adjustment methodologies such as MAIC and STC have been introduced and have become increasingly popular in HTA. These target the A vs. C treatment effect that would be observed in the BC population, thereby performing an adjusted indirect comparison in such population. MAIC and STC implicitly assume that the target population is the BC population. The population-adjusted A vs. B treatment effect is estimated as:

$$\hat{\Delta}_{12}^{(2)} = \hat{\Delta}_{10}^{(2)} - \hat{\Delta}_{20}^{(2)}, \tag{2}$$

where $\hat{\Delta}_{10}^{(2)}$ is the estimated relative treatment effect of A vs C (mapped to the BC population), and $\hat{\Delta}_{20}^{(2)}$ is the estimated marginal treatment effect of B vs. C (in the BC population). Again, the estimate $\hat{\Delta}_{12}^{(2)}$ should target a marginal treatment effect for reimbursement decisions at the population level. Therefore, $\hat{\Delta}_{10}^{(2)}$ should target a marginal treatment effect that is compatible with $\hat{\Delta}_{20}^{(2)}$ [39].

Variances are combined in the same way as the Bucher method. As the relative effects, $\hat{\Delta}_{10}^{(2)}$ and $\hat{\Delta}_{20}^{(2)}$, are specific to separate studies, the within-trial randomization of the originally assigned patient groups is preserved. Because the estimates are based on different study samples (IPD are unavailable for BC), the within-trial relative effects are assumed statistically independent of each other. Hence, their variances are simply summed to estimate the variance of the A vs. B treatment effect.

A reference intervention is required to define the effect modifiers. In the methods considered in this thesis, we are selecting the effect modifiers influencing treatment A with respect to C (as opposed to the treatment effect modifiers of B vs. C). This is because we have to adjust for these in order to perform the indirect comparison in the BC population, implicitly assumed to be the target population. If one had access to IPD for the BC study and only published ALD for the AC study, one would have to adjust for the factors modifying the effect of treatment B with respect to C, in order to perform the comparison in the AC population.

Those studying the generalizability of treatment effects often make a distinction between sample-average and population-average marginal effects [17, 59–62]. Typically, another implicit assumption made by population-adjusted indirect comparisons is that the marginal treatment effects estimated in the BC sample, as described by its published covariate moments in the case of $\hat{\Delta}_{10}^{(2)}$, coincide with those that would be estimated in the target population of the trial. Namely, either the study sample on which inferences are made is the study target population, or it is a simple random sample (i.e., representative) of such population, ignoring sampling variability in their descriptive characteristics. Throughout the text, when referring to the AC and BC "populations", we are in fact referring to the AC and BC study samples. We do not view these as samples of the trial populations, but as the populations themselves.

In indirect treatment comparisons, a challenge also arising from treatment effect heterogeneity is inconsistency [63]. This concerns the relationship between direct and indirect pairwise treatment comparisons. While inconsistency is also induced by imbalances in effect modifiers

(between the direct and indirect evidence), it is a property specific to loops of evidence. In the network displayed in Figure 1, there are no such loops — direct evidence for a comparison between A and B is unavailable. Therefore, inconsistency is not covered by this thesis.

The use of population adjustment in HTA, both in published literature as well as in submissions for reimbursement, and its acceptability by national HTA bodies, e.g. in England and Wales, Scotland, Canada and Australia [21], is increasing across diverse therapeutic areas [21, 30, 64, 65]. As of January 20, 2022, a search among titles, abstracts and keywords for "matching-adjusted indirect comparison" and "simulated treatment comparison" in Scopus, reveals at least 174 peer-reviewed applications of MAIC and STC and conceptual papers about the methods.

To capture the use of population-adjusted indirect comparisons in submissions for reimbursement, the NICE website[2] was queried for published technology appraisals (TAs) using the terms "matching-adjusted indirect comparison" and "simulated treatment comparison". NICE TAs are recommendations on the clinical and cost-effectiveness of treatments in the National Health Service in England and Wales. Manufacturer submissions, evidence review group reports and NICE committee feedback documents completed between 1 January, 2010 and January 20, 2022 were systematically reviewed.

A total of 55 TAs using MAIC or STC were identified in the search — of these, 48 have been published since 2017. Figure 2 shows the rapid growth of peer-reviewed publications and NICE TAs featuring MAIC or STC since the introduction of these methods in 2010. MAIC and STC are mainly applied in the evaluation of cancer drugs, as 45 of the 55 NICE TAs using population adjustment have been in the oncology area. MAIC is used more predominantly than STC. Of the 55 identified appraisals, 50 employed MAIC and 9 employed STC to support the submission for reimbursement.

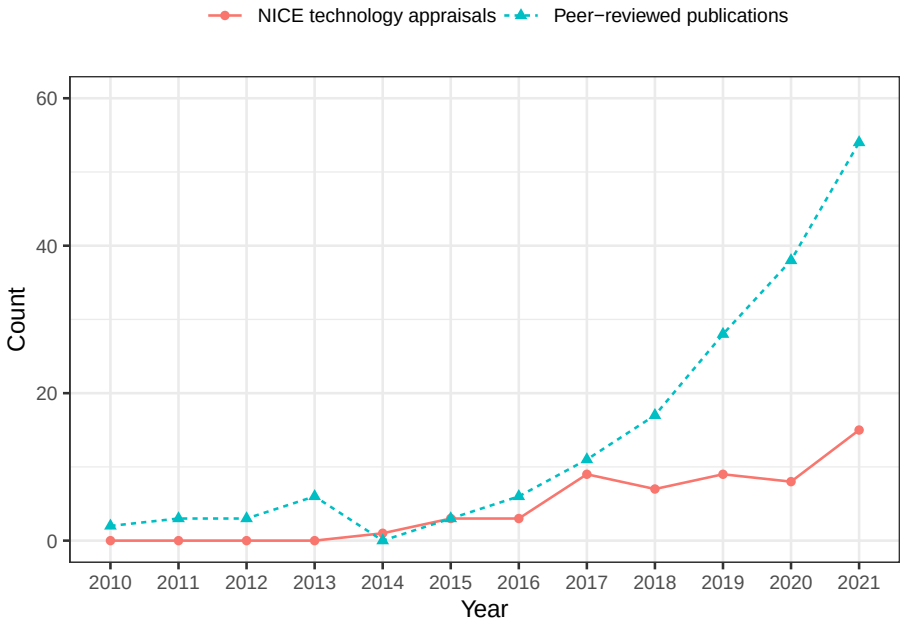


Figure 2: Number of peer-reviewed publications and technology appraisals from the National Institute for Health and Care Excellence (NICE) using population-adjusted indirect comparisons per year.

2 Appraisal consultation documents are publicly available in the website https://www.nice.org.uk/.

2.2 DATA STRUCTURE

We briefly outline the data requirements for the MAIC and STC methodologies. Consider that IPD are available for a randomized trial of size N comparing two interventions: treatment A and treatment C. Treatment A is a novel active intervention being tested for cost-effectiveness, which for reimbursement purposes needs to be compared to other established active treatments that are already on the market. The manufacturer submitting evidence to HTA bodies has access to IPD from its AC trial. We shall assume that the following data are available for the $n-$th subject ($n = 1, \ldots, N$) in the trial:

- A covariate vector of K baseline characteristics $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,K})$, e.g. age, gender, comorbidities;

- A treatment indicator z_n. Without loss of generality, we assume here for simplicity that $z_n \in \{0, 1\}$ for the common comparator and active treatment, respectively;

- An observed outcome y_n, e.g. a time-to-event or binary indicator for some clinical measurement.

Given this information, one can compute an estimate $\hat{\Delta}_{10}^{(1)}$ of the A vs. C treatment effect in the AC population, and an estimate of its variance. In the Bucher method, such estimate would be plugged in to Equation 1. On the other hand, MAIC and STC generate a population-adjusted estimate $\hat{\Delta}_{10}^{(2)}$ of the A vs. C treatment effect in the BC population that would be plugged in to Equation 2.

Treatment A needs to be compared to another active intervention, treatment B, but there are no head-to-head trials. A randomized trial comparing B and a common comparator C has been conducted by a competitor company. IPD are unavailable for this BC trial but published ALD are available. For the BC trial, the available data consist of the following components:

- A vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ of published summary values for the baseline characteristics. For ease of exposition, we shall assume that these are means and are available for all K covariates (alternatively, one would take the intersection of the available covariates).

- An estimate $\hat{\Delta}_{20}^{(2)}$ of the B vs. C treatment effect in the BC population, and an estimate of its variance, either published directly or derived from aggregate outcomes in the literature.

Each baseline characteristic $k = 1, \ldots, K$ can be classed as a prognostic variable (a covariate that affects outcome), an effect modifier, both or none. For simplicity in the notation, it is assumed that all available baseline characteristics are prognostic of the outcome and that a subset of these, $\boldsymbol{x}_n^{(EM)} \subseteq \boldsymbol{x}_n$, are selected as effect modifiers (of treatment A) on the linear predictor scale. Similarly, for the published summary values, $\boldsymbol{\theta}^{(EM)} \subseteq \boldsymbol{\theta}$.

2.3 MATCHING-ADJUSTED INDIRECT COMPARISON

Matching-adjusted indirect comparison (MAIC) is a population adjustment method based on inverse propensity score weighting [13]. IPD from the AC trial are weighted so that the means and, potentially, higher moments of specified covariates match those in the BC trial. The weights for the subjects in the IPD are estimated using a propensity score logistic regression model:

$$\ln(w_n) = \alpha_0 + \boldsymbol{x}_n^{(EM)}\boldsymbol{\alpha_1},$$

where α_0 and $\boldsymbol{\alpha_1}$ are the logistic regression parameters, and the weight w_n assigned to each individual n represents the "trial selection" or "trial assignment" odds, i.e., the odds of being enrolled in the BC trial as opposed to being enrolled in the AC trial. These are defined as a function of the baseline characteristics modifying the effect of treatment A, $\boldsymbol{x}_n^{(EM)}$ for subject n. Note that in standard applications of propensity score weighting, e.g. in observational studies, the propensity score logistic regression is for the *treatment group* assigned to the subject. In MAIC, the objective is to balance covariates across studies so the propensity score model is for the *trial* in which the participant is enrolled.

The logistic regression parameters cannot be derived using conventional methods such as maximum-likelihood estimation because IPD are not available for BC. Signorovitch et al. [10] propose using a method of moments to estimate the model parameters by setting the weights so that the mean effect modifiers are exactly balanced across the two trial populations. After centering the AC effect modifiers on the published BC means, such that $\boldsymbol{\theta}^{(EM)} = \mathbf{0}$, the weights are estimated by minimizing the objective function:

$$Q(\boldsymbol{\alpha_1}) = \sum_{n=1}^{N} \exp\left(\boldsymbol{x}_n^{(EM)}\boldsymbol{\alpha_1}\right),$$

where N represents the number of subjects in the AC trial. $Q(\boldsymbol{\alpha_1})$ is a convex function that can be minimized using standard algorithms, e.g. Broyden–Fletcher–Goldfarb–Shanno [66], to yield a unique finite solution $\hat{\boldsymbol{\alpha}}_{\mathbf{1}} = \text{argmin}[Q(\boldsymbol{\alpha_1})]$. Then, the estimated weight for subject n is:

$$\hat{w}_n = \exp(\boldsymbol{x}_n^{(EM)}\hat{\boldsymbol{\alpha}}_{\mathbf{1}}).$$

Consequently, the mean outcomes under treatment $z \in \{0,1\}$ in the BC population are predicted as the weighted average:

$$\hat{\mu}_z = \frac{\sum_{n=1}^{N_z} y_{n,z}\hat{w}_n}{\sum_{n=1}^{N_z} \hat{w}_{n,z}},$$

where N_z represents the number of subjects in arm z of the AC trial, $y_{n,z}$ denotes the outcome for patient n receiving treatment z in the patient-level data, and $\hat{w}_{n,z}$ is the weight assigned to participant n under treatment z. Note that we have summary data from the BC trial to estimate absolute outcomes under C. However, in this anchored scenario, we do not focus on

the absolute outcomes as the objective is to generate a relative effect for A vs. C in the BC population.

Such relative effect is typically estimated by fitting a weighted model, i.e., a model where the contribution of each subject to the likelihood is weighted. For instance, if the outcome of interest is a time-to-event outcome, an "odds-weighted" Cox model can be fitted by maximizing its weighted partial likelihood [67]. In this case, a subject n from the AC trial, who has experienced an event at time t, contributes the following term to the partial likelihood function:

$$\left(\frac{\exp(\beta_z z_n)}{\sum_{j \in R(t)} \hat{w}_j \exp(\beta_z z_j)} \right)^{\hat{w}_n}, \tag{3}$$

where $R(t)$ is the set of subjects without the event and uncensored prior to t, i.e., the risk set. Here, the fitted coefficient $\hat{\beta}_z$ of the weighted regression (i.e., the value of the parameter maximizing the partial likelihood in Equation 3) is the estimated relative effect for A vs. C, such that $\hat{\Delta}_{10}^{(2)} = \hat{\beta}_z$.

In the original MAIC approach, covariates are balanced for active treatment and control arms combined and standard errors are computed using a robust sandwich estimator, which allows for heteroskedasticity [10, 68]. Typically, implementations of this estimator do not explicitly account for the fitting of the logistic regression model for the weights, assuming these to be fixed.

Terms of higher order than means can also be balanced, e.g. by including squared covariates in the method of moments to balance variances. However, this decreases the degrees of freedom and may increase finite-sample bias [69]. Balancing both means and variances (as opposed to means only) appears to result in more biased and less accurate treatment effect estimates when the variances of covariates differ across trials [25, 27].

A proposed modification to MAIC uses entropy balancing [70] instead of the method of moments to estimate the weights [25, 28]. Entropy balancing has the additional constraint that the weights are as close as possible to unit weights. Potentially, it should penalize extreme weighting schemes and provide greater precision. However, Phillippo et al. recently demonstrated that weight estimation via entropy balancing and the method of moments are mathematically identical [71]. Other proposed modifications to MAIC include balancing the covariates separately for active treatment and common comparator arms [25, 28, 72], and using the bootstrap [73, 74] to compute standard errors [75], which does not rely upon strong assumptions about the estimation of the MAIC weights. Balancing the covariates separately seems to provide greater precision in simulation studies [25]. However, we do not recommend this approach because it may break randomization, distorting the balance between treatment arms A and C on covariates that are not accounted for in the weighting. If these covariates are prognostic of outcome, this would compromise the internal validity of the within-study treatment effect estimate for A vs. C.

As MAIC is a reweighting procedure, it will reduce the effective sample size (ESS) of the AC trial. The approximate ESS of the weighted IPD is estimated as $\left(\sum_n \hat{w}_n\right)^2 / \sum_n \hat{w}_n^2$; the reduction in ESS can be viewed as a rough indicator of the lack of overlap between the AC

and BC covariate distributions. For relative effects to be conditionally constant and eventually produce an unbiased indirect comparison, one needs to include all effect modifiers in the weighting procedure, whether in imbalance or not [18] (see Supplementary Appendix A for a non-technical overview of the full set of assumptions made by MAIC, and more generally, by population-adjusted indirect comparisons). The exclusion of balanced covariates does not ensure their balance after the weighting procedure. Including too many covariates or poor overlap in the covariate distributions can induce extreme weights and large reductions in ESS. This is a pervasive problem in NICE TAs, where most of the reported ESSs are small with a large percentage reduction from the original sample size [30].

Propensity score mechanisms are very sensitive to poor overlap [31–33]. In particular, weighting methods are unable to extrapolate — in the case of MAIC, extrapolation beyond the covariate space observed in the AC IPD is not possible. Almost invariably, the level of overlap between the covariate distributions will decrease as a greater number of covariates are included. Therefore, no purely prognostic variables should be balanced to avoid loss of effective sample size and consequent inflation of the standard error due to over-balancing [9]. Cross-trial imbalances in purely prognostic variables should not produce bias as relative treatment effects are unaffected due to within-trial randomization [18].

2.4 SIMULATED TREATMENT COMPARISON

While MAIC is a reweighting method, simulated treatment comparison (STC) [14] is a population adjustment method based on outcome regression [15]. Outcome regression methods are promising because they may increase precision and statistical power with respect to propensity score-based methodologies [76–78]. Contrary to most propensity score methods, outcome regression mechanisms are able to extrapolate beyond the covariate space where overlap is insufficient, using the linearity assumption or other appropriate assumptions about the input space. However, the validity of the extrapolation depends on the accuracy in capturing the true covariate-outcome relationships.

In STC, IPD from the AC trial are used to fit a regression model describing the observed outcomes in terms of the relevant baseline characteristics and the treatment variable. STC has different formulations [9, 14, 18, 19]. In the conventional version described by the NICE Decision Support Unit Technical Support Document 18 [9, 18], the covariates are centered at the published mean values $\boldsymbol{\theta}$ from the BC population. Under a generalized linear modeling framework, the following linear predictor is fitted to the observed AC IPD:

$$g(\mu_n) = \beta_0 + (\boldsymbol{x}_n - \boldsymbol{\theta})\,\boldsymbol{\beta_1} + \left[\beta_z + \left(\boldsymbol{x}_n^{(EM)} - \boldsymbol{\theta}^{(EM)}\right)\boldsymbol{\beta_2}\right]\mathbb{1}(z_n = 1), \tag{4}$$

where, for a generic subject n, μ_n is the expected outcome on the natural scale, e.g. the probability scale for binary outcomes, $g(\cdot)$ is an invertible canonical link function, β_0 is the intercept, $\boldsymbol{\beta_1}$ is a vector of K regression coefficients for the prognostic variables, $\boldsymbol{\beta_2}$ is a vector of interaction coefficients for the effect modifiers (modifying the effect of treatment A vs. C) and

β_z is the A vs. C treatment coefficient. For binary outcomes in logistic regression, one uses the $\text{logit}(\mu_n) = \ln[\mu_n/(1-\mu_n)]$ link function, but other choices are possible in practice, e.g. the identity link for standard linear regression with continuous-valued outcomes, or the log link for Poisson regression with count outcomes. Covariates are sometimes centered separately for active treatment and common comparator arms. We do not recommend this approach because it may break randomization.

The regression in Equation 4 models the conditional outcome mean given treatment and the centered baseline covariates. Because the IPD covariates are centered at the published mean values from the BC population ($\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(EM)}$, respectively), the estimated $\hat{\beta}_z$ is directly interpreted as the A vs. C treatment effect in the BC population or, more specifically, in a pseudopopulation with the BC covariate means and the AC correlation structure. Typically, analysts set $\hat{\Delta}_{10}^{(2)} = \hat{\beta}_z$ in Equation 2, inputting this coefficient into the health economic decision model [79, 80]. For uncertainty quantification purposes, the variance of said treatment effect is obtained from the standard error estimate of the treatment coefficient in the fitted model [9, 18]. In a Cox proportional hazards regression framework, a log link function could be employed in Equation 4 between the hazard function and the linear predictor component of the model.

For relative effects to be conditionally constant across studies, one needs to include all imbalanced effect modifiers in the model. In addition, the relationship between the effect modifiers and outcome must be correctly specified; in the case of this chapter, the effect modifiers must have an additive interaction with treatment on the linear predictor scale. It is optional to include (and to center) imbalanced variables that are purely prognostic. These will not remove bias further but a strong fit of the outcome model may increase precision. The NICE technical support document [18] suggests adding purely prognostic variables if they increase the precision of the model and account for more of its underlying variance, as reported by model selection criteria (e.g. residual deviance or information criteria). However, such tools should not guide decisions on effect modifier status, which must be defined prior to fitting the outcome model. As effect-modifying covariates are likely to be good predictors of outcome, the inclusion of appropriate effect modifiers should provide an acceptable fit.

Alternative "simulation-based" formulations to STC have been proposed [19, 81]. These are outlined as follows. The joint distribution of BC covariates is approximated under certain parametric assumptions to characterize the BC population, e.g. simulating continuous covariates at the individual level from a multivariate normal with the BC means and the correlation structure observed in the AC IPD. A regression of the outcome on the predictors is fitted to the AC patient-level data (this time, the covariates are not centered at the mean BC values). Then, the coefficients of this regression are applied to the simulated subject profiles and the linear predictions for patients under A and under C in the BC population are averaged out. The treatment effect for A vs. C is given by subtracting the average linear prediction under C from the average linear prediction under A. Neither the original conceptual publications nor the NICE technical support document provide detailed information about variance estimation, which is likely to be complicated and probably requires bootstrapping or similar approaches.

It is worth noting that, in the linear predictor scale, the arithmetic mean of the average linear predictor (the average linear predictor for patients sampled under the centered covariates) and its geometric mean (the linear predictor evaluated at the expectation of the centered covariates) coincide. Therefore, provided that the number of simulated subjects is sufficiently large (i.e., in expectation or ignoring sampling variability), the "covariate simulation" approach generates estimates that are equivalent to those of the "plug-in" methodology.

2.5 CLARIFICATION OF ESTIMANDS

An important issue that has not been discussed in the literature is that MAIC and the typical usage of STC target different types of estimands. In MAIC, as is typically the case for propensity score methods, $\hat{\Delta}_{10}^{(2)}$ targets a *marginal* treatment effect [51, 82, 83]. In biostatistics [84–87] and epidemiology [88–90], this marginal effect is also known as a *population-average* or *population-level* treatment effect, as it measures the average treatment effect for A vs. C at the population level. MAIC targets a marginal treatment effect for A vs C, in the BC population, because the weighted regression is a simple regression of outcome on treatment assignment alone. Therefore, assuming a reasonably large sample size and appropriate randomization in the AC trial, the fitted regression coefficient $\hat{\beta}_z$ in Equation 3 estimates a relative effect between subjects that have the same distribution of baseline characteristics (corresponding to the BC population), assuming that trial AC is reasonably large and has been appropriately randomized [91].

On the other hand, in the version of STC outlined by the NICE Decision Support Unit, $\hat{\Delta}_{10}^{(2)}$ targets a *conditional*, rather than a marginal treatment effect. The conditional treatment effect denotes the average effect at the individual or subgroup level [51, 92]. STC targets a conditional treatment effect because the estimate is the regression coefficient extracted from the fitted multivariable regression in Equation 4, conditional on the baseline covariates included as predictors, that have also been adjusted for. While the treatment coefficient $\hat{\beta}_z$ in STC targets an *average* treatment effect, it does not target a population-level measure, contrary to the marginal effect, which is the effect of moving all trial participants from one treatment to the other. While there is only one marginal effect for a specific population (as described by its covariate distribution), there may be many average conditional effects for a given population, one for every possible combination of covariates and model specification considered for adjustment.

Conditional measures of effect are clinically relevant in medical research, where one desires to apply the results of RCTs to individual patients. If there is treatment effect heterogeneity and this is accounted for by the inclusion of treatment-by-covariate interactions, conditional effect estimates are relevant as patient-centered evidence in a clinician–patient context, e.g. in precision or personalized medicine. Here, decision-making relates to the treatment benefit for an individual subject with specific covariate values. Conditional estimands are typically not of interest when making decisions at the population level in HTA and health policy, as they characterize effects at the unit or subgroup level.

A measure of effect is said to be *collapsible* if marginal and conditional effects coincide in the absence of confounding bias [88, 93]. The property of collapsibility is closely related to that of linearity [94, 95], e.g. mean differences in a linear regression are collapsible [51, 88, 92, 93]. However, most applications of population-adjusted indirect comparisons are in oncology and are typically concerned with time-to-event outcomes, or rate outcomes modeled using logistic regression [30]. These yield non-collapsible measures of treatment effect such as (log) hazard ratios [51, 88, 92, 96] or (log) odds ratios [51, 87, 88, 92, 93, 96, 97].

With non-collapsible effect measures, there may be sizable differences between marginal and conditional estimands for non-null effects do not coincide [94], even with covariate balance and in the absence of confounding [88, 93, 98–100]. For both collapsible and non-collapsible measures of effect, maximum-likelihood estimators targeting distinct estimands will have different standard errors [50]. Therefore, marginal and conditional estimates quantify parameter uncertainty differently, and conflating these will lead to the incorrect propagation of uncertainty to the wider health economic decision model, which will be problematic for probabilistic sensitivity analyses.

Therefore, the relative effect estimate $\hat{\Delta}_{10}^{(2)}$ in STC is unable to target a marginal treatment effect and the comparison of interest, a comparison of compatible marginal effects, cannot be performed. A comparison of conditional effects is not of interest for decisions at the population level, and also, cannot be carried out. A compatible conditional effect for B vs. C is unavailable because its estimation requires fitting the non-centered version of Equation 4, adjusting for the same set of covariates and with the same outcome regression specification, to the BC patient-level data [50]. Such data are unavailable and it is unlikely that the estimated treatment coefficient from this model is available in the clinical trial publication.

Hence, $\hat{\Delta}_{10}^{(2)}$ is incompatible with $\hat{\Delta}_{20}^{(2)}$ in the indirect comparison (Equation 2) for STC, even if all effect modifiers are accounted for and the outcome model is correctly specified. If we intend to target a marginal estimand for the A vs. C treatment effect (in the BC population) and naively assume that STC does so, $\hat{\Delta}_{12}^{(2)}$ may produce a biased estimate of the marginal treatment effect for A vs. B, even if all the assumptions in Supplementary Appendix A are met. None of the reviewed technology appraisals and peer-reviewed publications that use STC discuss the estimand that is targeted or take any steps for "marginalization". Neither do any of the simulation studies that have evaluated the performance of STC in the anchored scenario [24, 26, 34, 101].

On the other hand, $\hat{\Delta}_{10}^{(2)}$ targets a marginal treatment effect in MAIC. There are no compatibility issues in the indirect treatment comparison as $\hat{\Delta}_{10}^{(2)}$ and $\hat{\Delta}_{20}^{(2)}$ target comparable estimands of the same form. In the Bucher method, if the estimate $\hat{\Delta}_{10}^{(1)}$ is derived from a simple comparison of group means or from an univariable regression of outcome on treatment in the AC IPD, this targets a marginal effect and there are no compatibility issues in the indirect treatment comparison either.

2.6 CONCLUDING REMARKS

In this chapter, I have carried out an up-to-date review of MAIC and STC. I have demonstrated that MAIC and the typical usage of STC, as described by HTA guidance and recommendations, target different estimands. STC targets a conditional treatment effect as opposed to a marginal estimand, which is the appropriate target for HTA decisions at the population level. In addition, the conditional estimand cannot be combined in any indirect treatment comparison or compared between studies because conditional estimands vary across different covariate adjustment sets. This is a recurring problem in meta-analysis and is particularly troublesome where the measure of effect is non-collapsible.

3

CHAPTER 3: METHODS FOR POPULATION ADJUSTMENT WITH LIMITED ACCESS TO INDIVIDUAL PATIENT DATA: A SIMULATION STUDY

In this chapter, I conduct a comprehensive simulation study to benchmark the performance of MAIC and the typical usage of STC against the standard indirect treatment comparison. The simulation study provides proof-of-principle for the methods and is based on scenarios with survival outcomes, continuous covariates and the Cox proportional hazards regression as the outcome model, with the log hazard ratio as the measure of effect.

Section 3.1 describes the simulation study, which evaluates the properties of the approaches described in Chapter 2 under a variety of conditions. Section 3.2 presents the results of the simulation study. An extended discussion of my findings and their implications is provided in Section 3.3. Finally, Section 3.4 provides some brief concluding remarks. Part of the work in this chapter is included in the article "Methods for Population Adjustment with Limited Access to Individual Patient Data: A Review and Simulation Study" (Remiro-Azócar et al., 2021).[1]

3.1 SIMULATION STUDY DESIGN

3.1.1 *Aims*

The objectives of the simulation study are to evaluate MAIC, STC and the Bucher method across a wide range of scenarios, thereby benchmarking and comparing the statistical performance of existing methods for unadjusted and population-adjusted indirect comparisons, and providing proof-of-principle for the methodologies. For each estimator, we assess the following properties [102]: (1) unbiasedness; (2) variance unbiasedness; (3) randomization validity;[2] and (4) precision. The selected performance measures evaluate these criteria specifically (see 3.1.5). The simulation study is reported following the ADEMP (Aims, Data-generating mechanisms, Estimands, Methods, Performance measures) structure [102]. All simulations[3] and analyses were performed using R software version 3.6.3 [103]. Supplementary Appendix C lists the

1 The article has been accepted for publication in Research Synthesis Methods and is available at: https://doi.org/10.1002/jrsm.1511

2 In a sufficiently large number of repetitions, $(100 \times (1-\alpha))$% confidence intervals based on normal distributions should contain the true value $(100 \times (1-\alpha))$% of the time, for a nominal significance level α.

3 The files required to run the simulations are available at http://github.com/remiroazocar/population_adjustment_simstudy.

specific settings of each simulation scenario and Supplementary Appendix E presents example R code implementing MAIC, STC and the Bucher method on a simulated example.

3.1.2 *Data-generating mechanisms*

In line with the typical oncology application of MAIC and STC, we consider survival or time-to-event outcomes (e.g. overall or progression-free survival), using the log hazard ratio as the measure of effect.

For trials AC and BC, we follow Bender et al. [104] to simulate Weibull-distributed survival times under a proportional hazards parametrization.[4] Using the notation for the AC trial data, survival time t_n (for subject n) is generated according to the formula:

$$t_n = \left(\frac{-\ln u_n}{\lambda \exp[\boldsymbol{x}_n \boldsymbol{\beta_1} + (\beta_z + \boldsymbol{x}_n^{(EM)} \boldsymbol{\beta_2}) \mathbb{1}(z_n = 1)]} \right)^{1/\nu}, \tag{5}$$

where u_n is a uniformly distributed random variable, $u_n \sim \text{Uniform}(0, 1)$. We set the inverse scale of the Weibull distribution to $\lambda = 8.5$ and the shape to $\nu = 1.3$ as these parameters produce a functional form reflecting frequently observed mortality trends in metastatic cancer patients [27] (as illustrated in Figure 3 and Figure 4, which display the survival curves implied by the parameters). Four correlated or uncorrelated continuous covariates $\boldsymbol{x}_n$ are generated per subject using a multivariate Gaussian copula [106]. Two of these are purely prognostic variables; the other two ($\boldsymbol{x}_n^{(EM)}$) are effect modifiers, modifying the effect of both treatments A and B with respect to C on the log hazard ratio scale, and prognostic variables.

We introduce random right censoring to simulate loss to follow-up within each trial. Censoring times $t_{c,n}$ are generated from the exponential distribution $t_{c,n} \sim \text{Exp}(\lambda_c)$, where the rate parameter $\lambda_c = 0.96$ is selected to achieve a censoring rate of 35% under the active treatment at baseline (with the values of the covariates set to zero), considered moderate censoring [107]. We fix the value of λ_c before generating the datasets, by simulating survival times for 1,000,000 subjects with Equation 5 and using the R function `optim` (Brent's method [108]) to minimize the difference between the observed and targeted censoring proportion.

The number of subjects in the BC trial is 600, under a 1:1 active treatment vs. control allocation ratio. This sample size corresponds to that of a reasonably large Phase III RCT [109]. Different values are not explored as preliminary results showed that these drive performance less than the number of subjects in the AC trial. While the number of subjects in BC contributes to sampling variability, the reweighting or regressions are performed in the AC patient-level data. For the BC trial, the individual-level covariates and outcomes are aggregated to obtain summaries. The continuous covariates are summarized as means — these would typically be available to the analyst in the published study as a table of baseline characteristics. The

4 At baseline, this formulation has a hazard function $h_0(t) = \lambda \nu t^{\nu-1}$, a cumulative hazard function $H_0(t) = \lambda t^{\nu}$, a density function $f_0(t) = \lambda \nu t^{\nu-1} \exp(-\lambda t^{\nu})$ and a survival function $S_0(t) = \exp(-\lambda t^{\nu})$ at time $0 \leq t < \infty$, where $\lambda > 0$ is a positive inverse scale (rate) parameter, and $\nu > 0$ is a positive shape parameter. This follows the proportional hazards parametrization of the Weibull distribution in the NICE technical support document on survival analysis, where λ is referred to as a scale parameter [105].

marginal B vs. C treatment effect and its variance are estimated through a Cox proportional hazards regression of outcome on treatment. These estimates make up the only information on aggregate outcomes available to the analyst.

The simulation study examines five factors in a fully factorial arrangement with $3 \times 3 \times 3 \times 2 \times 3 = 162$ scenarios to explore the interaction between factors. The simulation scenarios are defined by varying the values of the following parameters, which are inspired by applications of MAIC and STC in NICE technology appraisals:

- The number of patients in the AC trial, $N \in \{150, 300, 600\}$ under a 1:1 active intervention vs. control allocation ratio. The sample sizes correspond to typical values for a Phase III RCT [109] and for trials included in applications of MAIC and STC submitted to HTA authorities [30].

- The strength of the association between the prognostic variables and the outcome, $\beta_{1,k} \in \{-\ln(0.67), -\ln(0.5), -\ln(0.33)\}$ (moderate, strong and very strong prognostic variable effect), where k indexes a given covariate. These regression coefficients correspond to fixing the conditional hazard ratios for the effect of each prognostic variable at approximately 1.5, 2 and approximately 3, respectively.

- The strength of interaction of the effect modifiers with treatment, $\beta_{2,k} \in \{-\ln(0.67), -\ln(0.5), -\ln(0.33)\}$ (moderate, strong and very strong interaction effect), where k indexes a given effect modifier.

- The level of correlation between covariates, $\text{cor}(x_{n,k}, x_{n,l}) \in \{0, 0.35\}$ (no correlation and moderate correlation), for subject n and covariates $k \neq l$.

- The degree of covariate imbalance.[5] For both trials, each covariate k follows a normal marginal distribution. For the BC trial, we fix $x_{n,k} \sim \text{Normal}(0.6, 0.2^2)$, for subject n. For the AC trial, the normal distributions have mean μ_k, such that $x_{n,k} \sim \text{Normal}(\mu_k, 0.2^2)$, varying $\mu_k \in \{0.45, 0.3, 0.15\}$. This yields strong, moderate and poor covariate overlap, respectively, corresponding to average percentage reductions in ESS across scenarios of 19%, 53% and 79%. These percentage reductions in ESS are representative of the range encountered in NICE TAs (see below).

Each active intervention has a very strong conditional treatment effect $\beta_z = \ln(0.25)$ at baseline (when the effect modifiers are zero) versus the common comparator. The covariates may represent comorbidities, which are associated with shorter survival and, in the case of the effect modifiers, which interact with treatment to render it less effective. Figure 3 shows

5 Due to the simulation study design, where the covariate distributions are symmetric, covariate *balance* is a proxy for covariate *overlap* in this parameter setting. Imbalance refers to the difference in covariate distributions across studies, as measured by the difference in (standardized) average covariate values. Overlap describes the degree of similarity in the covariate ranges across studies — there is complete overlap if the ranges are the same. In real scenarios, lack of complete overlap does not necessarily imply imbalance (and vice versa). Imbalances in effect modifiers across studies bias the standard indirect comparison, motivating the use of population adjustment. Lack of complete overlap hinders the use of population adjustment, as the covariate data may be too limited to make any conclusions in the regions of non-overlap.

the Weibull-distributed survival curves for patients under the active treatment (A and B) with varying levels of the covariates. Figure 4 shows the Weibull-distributed survival curves for subjects under the common comparator (C). In Figures 3 and 4, the strength of each prognostic term and each effect-modifying interaction is moderate.

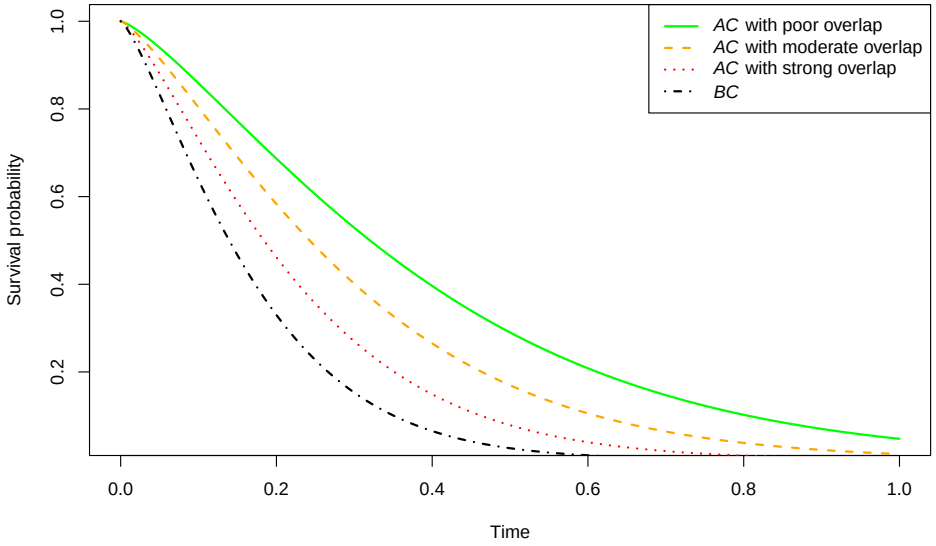


Figure 3: Weibull-distributed curves used to simulate survival times for subjects under the active treatment for different trial populations. The covariates are associated with shorter survival and, in the case of the effect modifiers, interact with treatment to render it less effective. As the mean values of the AC covariates decrease, overlap decreases.

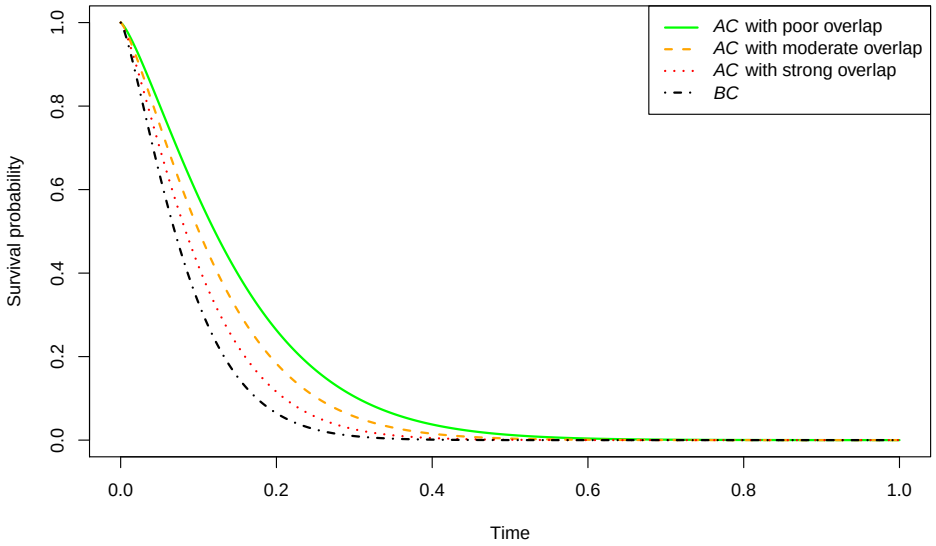


Figure 4: Weibull-distributed curves used to simulate survival times for subjects under the common comparator for different trial populations.

The simulation study meets the shared effect modifier assumption [18], i.e., active treatments A and B have the same set of effect modifiers with respect to the common comparator, and the interaction effects $\beta_{2,k}$ of each effect modifier k are identical for both treatments.

The varying degrees of covariate overlap are inspired by applications of MAIC in technology appraisals submitted to NICE. Only 13 of the 27 appraisals carrying out a MAIC have effective sample sizes available, albeit some appraisals contain multiple comparisons for different endpoints. In most applications, weighting considerably reduces the effective sample size from the original AC sample size. The median percentage reduction is 58% (range: 7.9% to 94.1%; interquartile range: 42.2% to 74.2%). The final effective sample sizes are also representative of those in the technology appraisals, which are also small (median: 80; range: 4.8 to 639; interquartile range: 37 to 174).

3.1.3 *Estimands*

The estimand of interest is the marginal A vs. B treatment effect in the BC population. The treatment coefficient $\beta_z = \ln(0.25)$ is identical for both A vs. C and B vs. C, and the shared effect modifier assumption holds in the simulation study. Hence, the true conditional effect for A vs. B in the BC population is zero. Because the true unit-level treatment effects are zero for all subjects, the true marginal treatment effect for A vs. B in the BC population is zero ($\Delta_{12}^{(2)} = 0$), which implies a "null" simulation setup in terms of this contrast, and marginal and conditional estimands for A vs. B in the BC population coincide by design.

Note that the true marginal effect for A vs. B in the BC population is a composite of that for A vs. C and that for B vs. C, both of which are non-null. These are the same and cancel out. For reference, the true values of the marginal treatment effect in the BC population for the active treatments vs. the common comparator ($\Delta_{10}^{(2)}$ and $\Delta_{20}^{(2)}$) are provided in Table 1. These have been computed as follows. Two potential cohorts of 500,000 subjects are simulated, with the BC covariate distribution and the outcome-generating mechanism in subsection 3.1.2. One cohort is under the active treatment and the other is under the common comparator. The number of simulated subjects is sufficiently large to minimize sampling variability. The two cohorts are concatenated and simple univariable Cox regressions are fitted, regressing the simulated survival times on an indicator variable denoting treatment status. The treatment coefficient of each regression estimates the expected difference in the potential outcomes on the log hazard ratio scale, and serves as the log of the true marginal hazard ratio for the two interventions under consideration. This is because the survival times have been generated according to the true data-generating mechanism, where the true conditional effects are explicit, and which uses the correct conditional model by definition. Due to the non-collapsibility of the hazard ratio, this simulation-based approach has been adopted in previous research to determine the true marginal effect [84, 110, 111].

Interaction effect ($\beta_{2,k}$)	Main covariate effect ($\beta_{1,k}$)	True marginal treatment effects ($\Delta_{10}^{(2)}$, $\Delta_{20}^{(2)}$)
-ln(0.67)	$-\ln(0.67)$	-0.69
-ln(0.67)	$-\ln(0.5)$	-0.67
-ln(0.67)	$-\ln(0.33)$	-0.64
-ln(0.5)	$-\ln(0.67)$	-0.43
-ln(0.5)	$-\ln(0.5)$	-0.42
-ln(0.5)	$-\ln(0.33)$	-0.41
-ln(0.33)	$-\ln(0.67)$	-0.07
-ln(0.33)	$-\ln(0.5)$	-0.08
-ln(0.33)	$-\ln(0.33)$	-0.09

Table 1: True marginal log hazard ratios for the active treatments versus the common comparator corresponding to different simulation settings. The covariates are assumed to be uncorrelated.

3.1.4 *Methods*

Each simulated dataset is analyzed using the following methods:

- Matching-adjusted indirect comparison, as originally proposed by Signorovitch et al. [10], where covariates are balanced for active treatment and control arms combined and weights are estimated using the method of moments. To avoid further reductions in effective sample size and precision, only the effect modifiers are balanced. A weighted Cox proportional hazards model is fitted to the IPD using the `R` package `survival` [112]. Standard errors for the A vs. C treatment effect are computed using a robust sandwich estimator [10, 68] by setting `robust=TRUE` in `coxph`. Given the often arbitrary factors driving selection into different trials, the data-generating mechanism in subsection 3.1.2 does not specify a trial assignment model. Nevertheless, the logistic regression model for estimating the weights is considered the "best-case" model because the "right" subset of covariates is selected as effect modifiers. The estimated weights are adequate for bias removal because the balancing property [113–116] holds with respect to the effect modifier means. Namely, conditional on the weights, all effect modifier means are balanced between the two trials, and one can potentially achieve unbiased estimation of treatment effects in the BC population due to conditional exchangeability over trial assignment.

- Simulated treatment comparison: a Cox proportional hazards regression on survival time is fitted to the IPD, with the IPD effect modifiers centered at the BC mean values. The outcome regression is correctly specified. We include all of the covariates in the regression but only center the effect modifiers.

- The Bucher method [7] gives the standard indirect comparison. We know that this will be biased as it does not adjust for the bias induced by the imbalance in effect modifiers.

In all methods, the variances of the within-trial relative effects are summed to estimate the variance of the A vs. B treatment effect, $\hat{V}(\hat{\Delta}_{12}^{(2)})$. Confidence intervals are constructed using normal distributions: $\hat{\Delta}_{12}^{(2)} \pm 1.96\sqrt{\hat{V}(\hat{\Delta}_{12}^{(2)})}$, assuming relatively large N.

3.1.5 *Performance measures*

We generate and analyze 1,000 Monte Carlo replicates of trial data per simulation scenario. Let $\hat{\Delta}_{12,q}^{(2)}$ denote the estimator for the q-th Monte Carlo replicate and let $\mathbb{E}(\hat{\Delta}_{12}^{(2)})$ denote its mean across the 1,000 simulations. Based on a test run of the method and simulation scenario with the highest long-run variability (MAIC under Scenario 109), we assume that $\text{SD}(\hat{\Delta}_{12}^{(2)}) \leq 0.45$ and that, conservatively, the variance across simulations of the estimated treatment effect is always less than approximately 0.2. Given that the Monte Carlo standard error (MCSE) of the bias is equal to $\sqrt{\text{Var}(\hat{\Delta}_{12}^{(2)})/N_{sim}}$, where N_{sim} is the number of simulations, it is at most 0.014 under 1,000 simulations. We consider the degree of precision provided by the MCSE, which quantifies the simulation uncertainty, under 1,000 repetitions to be acceptable in relation to the size of the effects. If the empirical coverage rate of the methods is 95%, $N_{sim} = 1000$ implies that the MCSE of the coverage is $\left(\sqrt{(95 \times 5)/1000}\right) = 0.69\%$, with the worst-case MCSE being 1.58% under 50% coverage. We also consider this degree of precision to be acceptable. Hence, the simulation study is conducted under $N_{sim} = 1000$.

The following criteria are considered jointly to assess the methods' performances. MCSEs are estimated for each performance metric in order to quantify the simulation uncertainty due to using a finite number of simulation replicates.

- To assess aim 1, we compute the **bias** in the estimated treatment effect

$$\mathbb{E}(\hat{\Delta}_{12}^{(2)} - \Delta_{12}^{(2)}) = \frac{1}{1000}\sum_{q=1}^{1000}\hat{\Delta}_{12,q}^{(2)} - \Delta_{12}^{(2)}.$$

  As $\Delta_{12}^{(2)} = 0$, the bias is equal to the average estimated treatment effect across the simulations. The MCSE of the bias is estimated as $\sqrt{\frac{1}{1000\times 999}\sum_{q=1}^{1000}(\hat{\Delta}_{12,q}^{(2)} - \mathbb{E}(\hat{\Delta}_{12}^{(2)}))}$.

- To assess aim 2, we calculate the **variability ratio** of the treatment effect estimate, defined [117] as the ratio of the average model standard error and the observed standard deviation of the treatment effect estimates (empirical standard error):

$$\text{VR}\left(\hat{\Delta}_{12}^{(2)}\right) = \frac{\frac{1}{1000}\sum_{q=1}^{1000}\sqrt{\hat{V}(\hat{\Delta}_{12,q}^{(2)})}}{\sqrt{\frac{1}{999}\sum_{q=1}^{1000}(\hat{\Delta}_{12,q}^{(2)} - (\hat{\Delta}_{12}^{(2)}))^2}}. \quad (6)$$

  VR being greater than (or smaller) than one suggests that, on average, standard errors overestimate (or underestimate) the variability of the treatment effect estimate. It is important to note that this metric assumes that the correct estimand and corresponding variance are being targeted. A variability ratio of one is of little use if this is not the case, e.g. if both the model standard errors and the empirical standard errors are taken over

estimates targeting the wrong estimand. The MCSE of the variability ratio is approximated as:

$$\sqrt{\frac{\frac{1}{1000}\sum_{q=1}^{1000}\left(\sqrt{\hat{V}(\hat{\Delta}_{12,q}^{(2)})}-\mathbb{E}\left(\sqrt{\hat{V}(\hat{\Delta}_{12}^{(2)})}\right)\right)^2}{999\times \text{ESE}(\hat{\Delta}_{12}^{(2)})^2}+\frac{\left(\frac{1}{1000}\sum_{q=1}^{1000}\sqrt{\hat{V}(\hat{\Delta}_{12,q}^{(2)})}\right)^2}{2\times 999\times \text{ESE}(\hat{\Delta}_{12}^{(2)})^2}},$$

where $\text{ESE}(\hat{\Delta}_{12}^{(2)})$ is the estimated empirical standard error, which is the denominator in Equation 6.

- Aim 3 is assessed using the **coverage** of confidence intervals, estimated as the proportion of times that the true treatment effect is enclosed in the $(100\times(1-\alpha))\%$ confidence interval of the estimated treatment effect, where $\alpha = 0.05$ is the nominal significance level. The MCSE of the coverage is computed as $\sqrt{\frac{\text{Cover}(\hat{\Delta}_{12}^{(2)})\times(1-\text{Cover}(\hat{\Delta}_{12}^{(2)}))}{1000}}$, where $\text{Cover}(\hat{\Delta}_{12}^{(2)})$ is the estimated coverage percentage.
- We use **empirical standard error** (ESE) to assess aim 4 as it measures the precision or long-run variability of the treatment effect estimate. The ESE is defined above, as the denominator in Equation 6. The MCSE of the empirical standard error is estimated as $\frac{\text{ESE}(\hat{\Delta}_{12}^{(2)})}{\sqrt{2\times 999}}$.
- The **mean square error** (MSE) of the estimated treatment effect

$$\text{MSE}(\hat{\Delta}_{12}^{(2)}) = \mathbb{E}\left[(\hat{\Delta}_{12}^{(2)}-\Delta_{12}^{(2)})^2\right] = \frac{1}{1000}\sum_{q=1}^{1000}(\hat{\Delta}_{12,q}^{(2)}-\Delta_{12}^{(2)})^2,$$

provides a summary value of overall accuracy (efficiency), integrating elements of bias (aim 1) and variability (aim 4). The Monte Carlo standard error of the MSE is computed as $\sqrt{\frac{\sum_{q=1}^{1000}\left[(\hat{\Delta}_{12,q}^{(2)}-\Delta_{12}^{(2)})^2-\text{MSE}(\hat{\Delta}_{12}^{(2)})\right]^2}{1000\times 999}}$, where $\text{MSE}(\hat{\Delta}_{12}^{(2)})$ is the estimated mean square error.

3.2 RESULTS OF THE SIMULATION STUDY

The performance measures across all 162 simulation scenarios are illustrated in Figures 5 to 9 using nested loop plots [118], which arrange all scenarios into a lexicographical order, looping through nested factors. In the nested sequence of loops, we consider first the parameters with the largest perceived influence on the performance metric. Notice that this order is considered on a case-by-case basis for each performance measure. Given the large number of simulation scenarios, depiction of Monte Carlo standard errors, quantifying the simulation uncertainty, is difficult. The performance metrics for each scenario and the Monte Carlo standard errors of each performance metric are reported in Supplementary Appendix C. In MAIC, 1 of 162,000 weighted regressions had a separation issue, i.e., there is a total lack of covariate overlap (Scenario 115, with $N = 150$). Results for this replicate were discarded. The outcome regressions converged for all replicates in STC and the Bucher method.

3.2.1 *Unbiasedness of treatment effect*

The impact of the bias will depend on the uncertainty in the estimated treatment effect [119, 120], measured by the empirical standard error. To assess such impact, we consider standardizing the biases [120] by computing these as a percentage of the empirical standard error. In a review of missing data methods, Schafer and Graham [119] consider bias to be troublesome under 1,000 simulations if its absolute size is greater than about one half of the estimate's empirical standard error, i.e., the standardized bias has magnitude greater than 50%. Under this rule of thumb, MAIC does not produce problematic biases in any of the simulation scenarios. On the other hand, STC and the Bucher method generate problematic biases in 71 of 162 scenarios, and in 147 of 162 scenarios, respectively. The biases in MAIC do not appear to have any practical significance, as they do not degrade coverage and efficiency.

Figure 5 shows the bias for the methods across all scenarios. MAIC is the least biased method, followed by STC and the Bucher method. In the scenarios considered in this simulation study, STC produces negative bias when the interaction effects are moderate and positive bias when they are very strong. In addition, biases vary more widely when prognostic effects are larger. When interaction effects are weaker, stronger prognostic effects shift the bias negatively. Note that all methods perform the same unadjusted analysis (i.e., a simple regression of outcome on treatment) to estimate the marginal treatment effect of B versus C. Because the BC study is a relatively large RCT, this comparison is unbiased with respect to the true marginal log hazard ratios in the BC population, reported in Table 1. Therefore, any bias in the A vs. B comparison should arise from bias in the A vs. C comparison, for which marginal treatment effects are non-null. The degree of systematic bias in STC arises from a mismatch between the conditional estimates produced for A versus C and the corresponding marginal estimands that should be targeted. This mismatch is a result of the non-collapsibility of the (log) hazard ratio (see Section 2.5).

In some cases, e.g. under very strong prognostic variable effects and moderate effect-modifying interactions, STC even has increased bias compared to the Bucher method. In other scenarios, e.g. where there are strong effect-modifying interactions and moderate or strong prognostic variable effects, STC estimates are virtually unbiased. This is because, in these scenarios, the conditional and marginal estimands for A vs. C are almost identical and the non-collapsibility of the measure of effect does not induce bias. It is worth noting that conclusions arising from the interpretation of these patterns or other trends in Figure 5 for STC are by-products of non-collapsibility. Any generalization should be cautious, as the divergence between conditional and marginal estimands is simply due to a mathematical property of the (log) hazard ratio.

As expected, the strength of interaction effects is an important driver of bias in the Bucher method and the incurred bias increases with greater covariate imbalance. This is because the more substantial the imbalance in effect modifiers and the greater their interaction with treatment, the larger the bias of the unadjusted comparison. The impact of these factors on the bias appears to be slightly reduced when prognostic effects are stronger and contribute more

"explanatory power" to the outcome. Varying the number of patients in the AC trial does not seem to have any discernible impact on the bias for any method. Biases in MAIC seem to be unaffected when varying the degree of covariate imbalance/overlap.

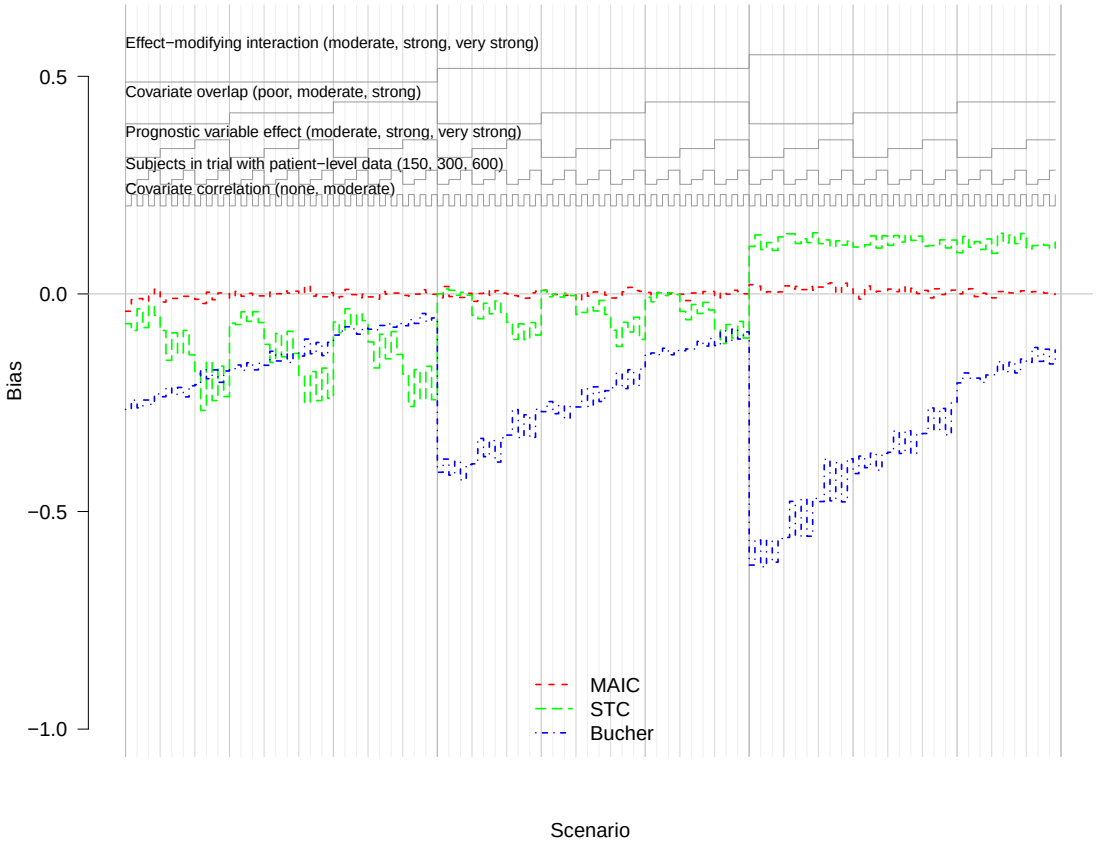


Figure 5: Bias across all simulation scenarios. The nested loop plot arranges all 162 scenarios into a lexicographical order, looping through nested factors. In the nested sequence of loops, we consider first the parameters with the largest perceived influence on the performance metric.

3.2.2 *Unbiasedness of variance of treatment effect*

In the Bucher method, the variability ratio is close to one under the vast majority of simulation scenarios (Figure 6). This suggests that standard error estimates for the methods are unbiased, i.e., that the model standard errors coincide with the empirical standard errors. In STC, variability ratios are generally close to one under $N = 300$ and $N = 600$, and any bias in the estimated variances appears to be negligible. However, the variability ratios decrease when the AC sample size is small ($N = 150$). In these scenarios, there is some underestimation of variability by the model standard errors. It is important to recall that this metric assumes that the correct estimand and corresponding variance are being targeted. This is not the case in our application of STC, in the sense that both model standard errors and empirical standard errors are taken over an incompatible indirect treatment comparison. MAIC standard

errors underestimate variability when $N = 150$, and also when covariate overlap is poor, in which case underestimation under $N = 150$ is exacerbated. Under the smallest sample size and poor covariate overlap, variability ratios are often below 0.9, with model standard errors clearly underestimating the empirical standard errors. This is likely due to the robust sandwich estimator used to derive the standard errors. In the literature, this has exhibited an underestimation of variability in small samples [121, 122]. The understated uncertainty is problematic for inference, and will also be propagated through the cost-effectiveness analysis, potentially leading to inappropriate decision-making [23].

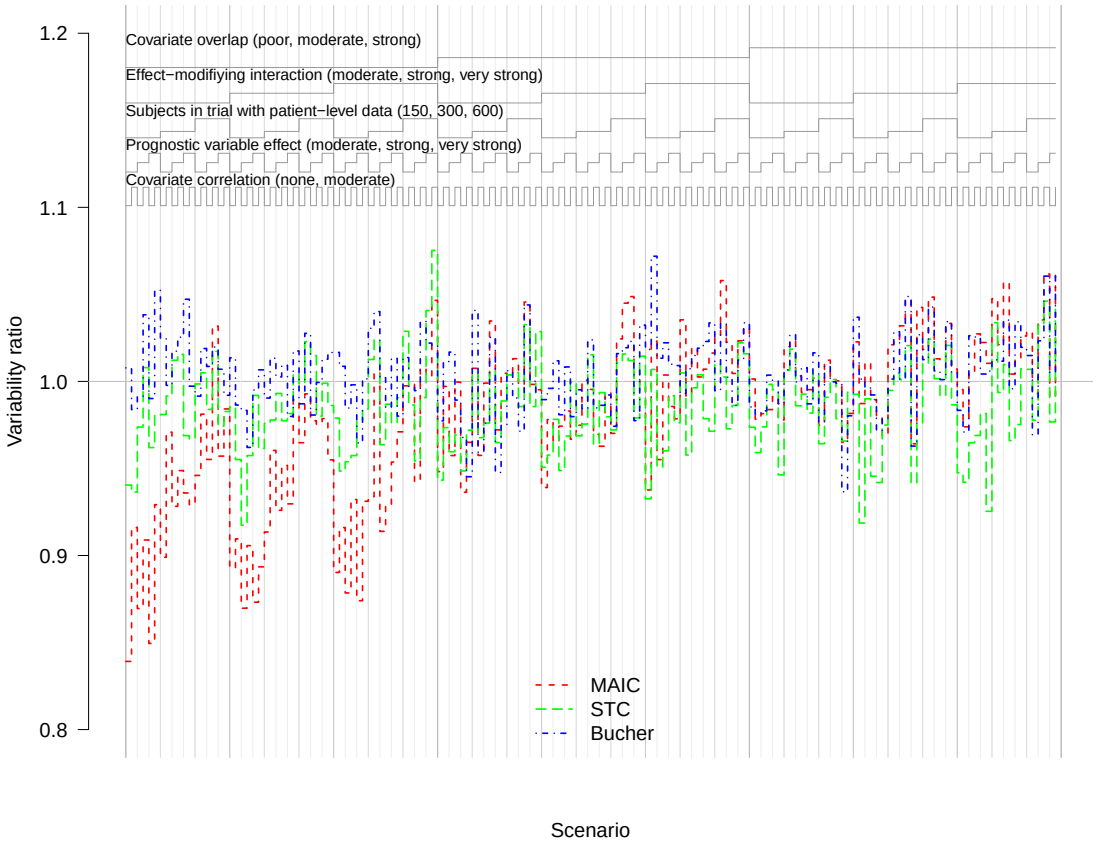


Figure 6: Variability ratio across all simulation scenarios.

3.2.3 *Randomization validity*

From a frequentist viewpoint [123], 95% confidence intervals are randomization-valid if these are guaranteed to include the true treatment effect 95% of the time. This means that the empirical coverage rate should be approximately equal to the nominal coverage rate, in this case 0.95 for 95% confidence intervals, to obtain appropriate type I error rates for testing a "no effect" null hypothesis. Theoretically, the empirical coverage rate is statistically significantly different to 0.95 if, roughly, it is less than 0.9365 or more than 0.9635, assuming 1,000 independent simulations per scenario. These values differ by approximately two standard

errors from the nominal coverage rate. When randomization validity cannot be attained, one would at least expect the interval estimates to be confidence-valid, i.e., the 95% confidence intervals include the true treatment effect *at least* 95% of the time.

In general, empirical coverage rates for MAIC do not overestimate the advertised nominal coverage rate. Only 4 of 162 scenarios have a rate above 0.9635. On the other hand, empirical coverage rates are significantly below the nominal coverage rate when the AC sample size is low ($N = 150$) and under poor covariate overlap. With $N = 150$, 24 of 54 coverage rates are below 0.9365. When covariate overlap is poor, 38 of 54 coverage rates are below 0.9365 — 18 of these under $N = 150$. When there is both poor overlap and a low AC sample size, coverage rates for MAIC are inappropriate: these may even fall below 90%, i.e., at least double the nominal rate of error. Poor coverage rates are a decomposition of both the bias and the standard error used to compute the width of the confidence intervals. It is not bias that degrades the coverage rates for this method but the standard error underestimation mentioned in subsection 3.2.2. Poor coverage is induced by the standard errors used in the construction of the confidence intervals.

Figure 7 shows the empirical coverage rates for the methods across all scenarios. Undercoverage is a pervasive problem in STC, for which 126 of 162 scenarios have empirical coverage rates below 0.9365. In the case of STC, undercoverage can be attributed to the bias induced by the non-collapsibility of the log hazard ratio, discussed in subsection 2.5. The coverage rates drop under the most important determinants of bias, e.g. moderate effect-modifying interactions and very strong prognostic variable effects. Under these conditions, the bias of STC is high enough to shift the coverage rates negatively, pulling these below 80% in some scenarios.

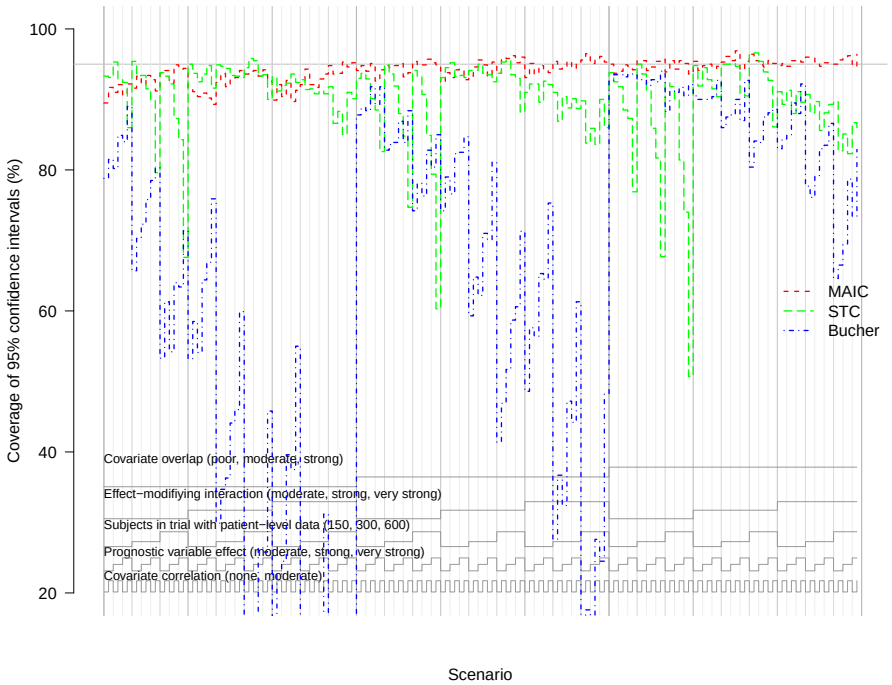


Figure 7: Empirical coverage percentage of 95% confidence intervals across all simulation scenarios.

Confidence intervals from the Bucher method are not confidence-valid for virtually all scenarios. Coverage rates deteriorate markedly under the most important determinants of bias. When there is greater imbalance between the covariates and when interaction effects are stronger, the induced bias is larger and coverage rates are degraded. Under very strong interactions with treatment, empirical coverage may drop below 50%. Therefore, the Bucher method will incorrectly detect significant results a large proportion of times in these scenarios. Such overconfidence will lead to very high type I error rates for testing a "no effect" null hypothesis.

3.2.4 *Precision and efficiency*

Several trends are revealed upon visual inspection of the empirical standard error across scenarios (Figure 8). As expected, the ESE decreases for all methods (i.e., the estimate is more precise) as the number of subjects in the AC trial increases. The strengths of interaction effects and of prognostic variable effects appear to have a negligible impact on the precision of population adjustment methods. The degree of covariate overlap has an important influence on the ESE and population adjustment methods incur losses of precision when covariate overlap is poor. When overlap is poor, there exists a subpopulation in BC that does not overlap with the AC population. Therefore, inferences in this subpopulation rely largely on extrapolation. Outcome regression methods such as STC require greater extrapolation when the covariate overlap is poorer [18]. In reweighting methods such as MAIC, extrapolation is not even possible. When covariate overlap is poor, observations in the AC patient-level data (those that are not covered by the range of the effect modifiers in the BC population) are assigned very low weights (low odds of enrolment in BC vs. AC). On the other hand, the relatively small number of units in the overlapping region of the covariate space are assigned very large weights, dominating the reweighted sample. These extreme weights lead to large reductions in ESS and to the deterioration of precision and efficiency.

In MAIC, the presence of correlation mitigates the effect of decreasing covariate overlap on a consistent basis. This is due to the correlation increasing the overlap between the joint covariate distributions of AC and BC, lessening the reduction in effective sample size and providing greater stability to the estimates. ESE for the Bucher method does not vary across different degrees of covariate overlap, as these are not considered by the method, and overprecise estimates are produced.

Contrary to ESE, MSE also takes into account the true value of the estimand as it incorporates the bias. Hence, main drivers of bias and ESE are generally key properties for MSE. Figure 9 is inspected in order to explore patterns in the mean square error. Estimates are less accurate for MAIC when prognostic variable effects are stronger, AC sample sizes are smaller and covariate overlap is poorer. As bias is negligible for MAIC, precision is the driver of accuracy. On the contrary, as the Bucher method is systematically biased and overprecise, the driver of accuracy is bias. Poor accuracy in STC is also driven by bias, particularly under low sample sizes and strong prognostic variable effects. STC was consistently less accurate than MAIC, with larger mean square errors in all simulation scenarios. In some cases where the STC

bias was strong, e.g. very strong prognostic variable effects and moderate effect-modifying interactions, STC even increased the MSE compared to the Bucher method.

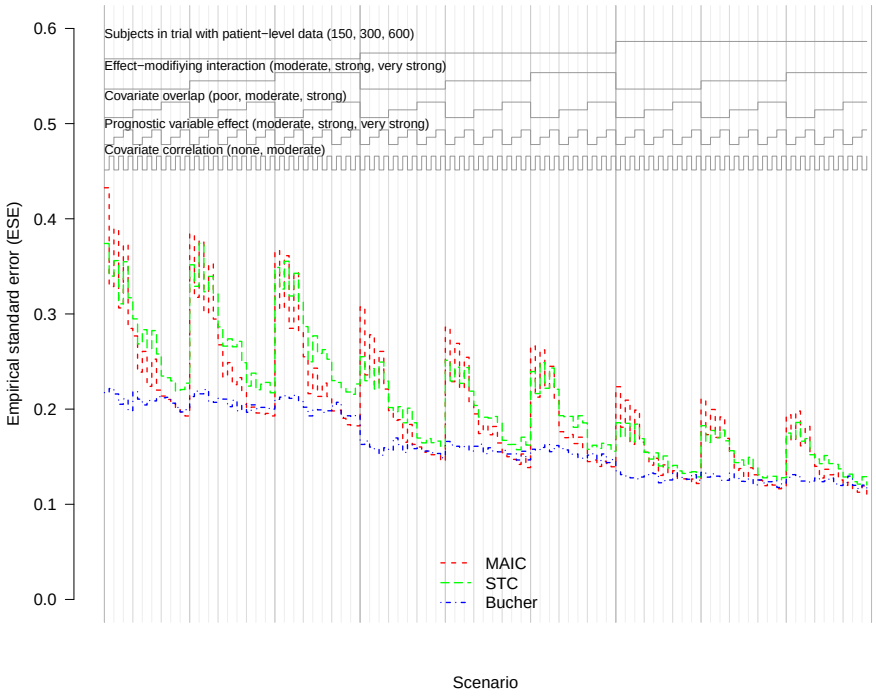


Figure 8: Empirical standard error across all simulation scenarios.

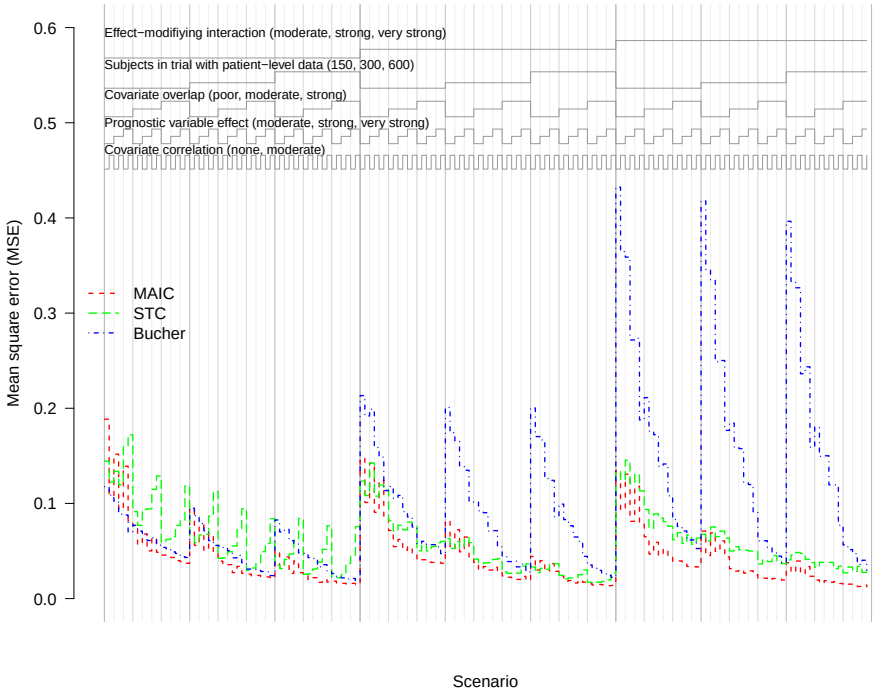


Figure 9: Mean square error across all simulation scenarios.

In accordance with the trends observed for the ESE, the MSE is also very sensitive to the value of N and decreases for all methods as N increases. We highlight that the number of subjects in the BC trial (not varied in this simulation study) is a less important performance driver than the number of subjects in AC; while it contributes to sampling variability, the reweighting or regressions are performed in the AC patient-level data.

3.3 DISCUSSION OF SIMULATION STUDY RESULTS

In this section, I discuss the implications of, and recommendations for, performing population adjustment, based on the simulation study. Finally, I highlight potential limitations of the simulation study, primarily relating to the extrapolation of its results to practical guidance. We have seen in Section 3.2 that the typical use of STC produces systematic bias as a result of the non-collapsibility of the log hazard ratio. The estimate $\hat{\Delta}_{10}^{(2)}$ targets a conditional treatment effect that is incompatible with the estimate $\hat{\Delta}_{20}^{(2)}$. This leads to bias in estimating the marginal treatment effect for A vs B, despite all assumptions for population adjustment being met. Given the clear inadequacy of STC in this setting, we focus on MAIC as a population adjustment method.

3.3.1 *Summary of findings*

BIAS-VARIANCE TRADE-OFFS Before performing population adjustment, it is important to assess the magnitude of the bias induced by effect modifier imbalances. Such bias depends on the degree of covariate imbalance and on the strength of interaction effects, i.e., the effect modifier status of the covariates. The combination of these two factors determines the level of bias reduction that would be achieved with population adjustment.

Inevitably, due to bias-variance trade-offs, the increase in variability that we are willing to accept with population adjustment depends on the magnitude of the bias that would be corrected. Such variability is largely driven by the degree of covariate overlap and by the AC sample size. Hence, while the potential extent of bias correction increases with greater covariate imbalance, so does the potential imprecision of the treatment effect estimate (assuming that the imbalance is accompanied by poor overlap).

In our simulation study, under no failures of assumptions, this trade-off always favours the bias correction offered by MAIC over the precision of the Bucher method, implying that the reductions in ESS based on unstable weights are worth it, even under stronger covariate overlap. Across scenarios, the relative accuracy of MAIC with respect to that of the Bucher method improves under greater degrees of covariate imbalance and poorer overlap. It is worth noting that, even in scenarios where the Bucher method is relatively accurate, it is still flawed in the context of inference and decision-making due to overprecision and undercoverage.

The magnitude of the bias that would be corrected with population adjustment also depends on the strength of interaction effects, i.e., the effect modifier status of the covariates. In the simulation study, the lowest effect-modifying interaction coefficient was $-\ln(0.67) = 0.4$.

Despite the relatively low magnitude of bias induced in this setting, MAIC was consistently more efficient than the Bucher method. Larger interaction effects warrant greater bias reduction but do not degrade the precision of the population-adjusted estimate. Hence, the relative accuracy of MAIC with respect to the Bucher method improves further as the effect-modifying coefficients increase.

VARIANCE ESTIMATION IN MAIC MAIC was generally randomization-valid, except in situations with poor covariate overlap and small sample sizes, where robust sandwich standard errors underestimated empirical estimates of the standard error and, consequently, there was undercoverage. MAIC exhibited variability ratios below 0.9 in scenarios with the smallest sample size and poor covariate overlap. In these scenarios, confidence intervals were narrow, achieving coverage rates which were statistically significantly below 95% and sometimes dropping below 90%. As mentioned in subsection 3.2.2, this is probably due to the robust sandwich estimator used to derive the standard errors, which has previously underestimated variability in small samples in simulation studies [121, 122]. It is worth noting that this estimator is based on large-sample (asymptotic) arguments and infinite populations. Therefore, it is not surprising that performance is poor under the smallest effective sample sizes, which occur where the AC trial sample size is small and covariate overlap is poor. Where effective sample sizes are small, confidence intervals derived from robust sandwich variance estimators should be interpreted cautiously, as these may understate uncertainty and this underestimation will be propagated through the cost-effectiveness analysis, potentially leading to inappropriate decision-making.

This robust variance estimator is easy to use by analysts performing MAIC (and propensity score weighting, in general) because it is computationally efficient and is typically implemented in standard routines in statistical computing software such as `R`. For instance, in `R`, by setting `robust=TRUE` in the `coxph` function, built in the `survival` package [112] for survival analysis, or using the `sandwich` package [124] for the treatment coefficient of a weighted generalized linear model. It is worth noting that these readily available implementations assume that the weights are fixed or known and do not account for the uncertainty in the estimation of the weights.

In principle, one could circumvent this issue by using the bootstrap to obtain the variance and confidence intervals of the A vs. C treatment effect, as in the simulation study by Petto et al. [25] or in the article by Sikirica et al. [75]. Bootstrap methods are beneficial because they can account for the variability of the estimated weights and are straightforward to implement, potentially providing unbiased variance estimators with a large number of resamples. However, bootstrapping is orders of magnitude more expensive computationally than applying the closed-form sandwich variance estimator. In addition, bootstrap resampling procedures are inherently random and exhibit some seed-dependence, which is only mitigated by increasing the number of resamples and computational demand. It is necessary to compare different approaches to variance estimation and assess whether implementations of the bootstrap can compete with the robust sandwich estimator.

Another alternative to variance estimation is the development of closed-form robust sandwich estimators that properly account for the uncertainty in estimating the propensity score logistic regression for the weights. These have been explicitly derived for accurate variance estimation in the causal inference literature [125–128], but not for MAIC.

3.3.2 *Implications for practice*

JUSTIFICATION OF EFFECT MODIFIER STATUS In the simulation study, we know that population adjustment is required as we set the cross-trial imbalances between covariates and have specified some of these as effect modifiers. Most applications of population adjustment present evidence of the former, e.g. through tables of baseline characteristics with covariate means and proportions ("Table 1" in a RCT publication). However, quantitative evidence justifying the effect modifier status of the selected covariates is rarely brought forward. Presenting this type of supporting evidence is very important when justifying the use of population adjustment.

Typically, the selection of effect modifiers is supported by clinical expert opinion. However, clinical expert judgment and subject-matter knowledge are fallible when determining effect modifier status because: (1) the therapies being evaluated are often novel; and (2) effect modifier status is scale-specific — clinical experts may not have the mathematical intuition to assess whether covariates are effect modifiers on the linear predictor scale (as opposed to the natural outcome scale).

Therefore, applications of population adjustment often balance all available covariates on the grounds of expert opinion. This is probably because the clinical experts cannot rule out bias-inducing interactions with treatment for any of the baseline characteristics. Almost invariably, the level of covariate overlap and precision will decrease as a larger number of covariates are accounted for in the analysis. Presenting quantitative evidence along with clinical expert opinion would help establish whether adjustment is necessary for each covariate [129].

As proposed by Phillippo et al. [9], we encourage the analyst to fit regression models with interaction terms to the IPD for an exploratory assessment of effect modifier status. One possible strategy is to consider each potential effect modifier one-at-a-time by adding the corresponding interaction term to the main (treatment) effect model [77]. Then, the interaction coefficient can be multiplied by the difference in effect modifier means to gauge the level of induced bias [18]. This analysis should be purely exploratory, since individual trials are typically underpowered for interaction testing [130, 131]. The dichotomization or categorization of continuous variables, the poor representation of a variable, e.g. a limited age range, and incorrectly assuming linearity may dilute interactions further.

Meta-analyses of multiple trials, involving the same outcome and similar treatments and conditions, provide greater power to detect interactions, particularly using IPD [131, 132]. With unavailable IPD, it may still be possible to conduct an IPD meta-analysis if the owners of the data are willing to provide the interaction effects [133], or one may conduct an ALD meta-analysis if covariate-treatment interactions are included in the clinical trial reports [130]. In any case, the identification of effect modifiers is in essence observational [134, 135], and

requires much more evidence than demonstrating a main treatment effect [136]. Therefore, it may be reasonable to balance a variable if there is a strong biological rationale for effect modification, even if the interaction is statistically weak, e.g. the P-value is large and the null hypothesis of interaction is not rejected [136].

NUANCES IN THE INTERPRETATION OF RESULTS It is worth noting that the conclusions of this simulation study are dependent on the outcome and model type. We have considered survival outcomes and the Cox proportional hazards model, as these are the most prevalent outcome type and modeling framework in MAIC and STC applications. However, further simulation studies are required with alternative outcome types and models. For example, exploratory simulations with binary outcomes and logistic regression have found that the performance of MAIC is more affected by low sample sizes and poor covariate overlap than seen for survival outcomes. This is likely due to logistic regression being less efficient [137] and more prone to small-sample bias [138] than Cox regression.

Furthermore, we have only considered and adjusted for two effect modifiers that induce bias in the same direction, i.e., the effect modifiers in a given study have the same means, the cross-trial differences in means are in the same direction, and the interaction effects are in the same direction. In real applications of population adjustment, it is not uncommon to see more than 10 covariates being balanced [30]. As this simulation study considered percentage reductions in effective sample size for MAIC that are representative of scenarios encountered in NICE TAs, real applications will likely have imbalances for each individual covariate that are smaller than those considered in this study. In addition, the means for the effect modifiers within a given study will differ, with the mean differences across studies and/or the effect-modifying interactions potentially being in opposite directions. Therefore, the induced biases could cancel out but, then again, this is not directly testable in a practical scenario.

3.3.3 *Limitations*

POTENTIAL FAILURES IN ASSUMPTIONS Most importantly, all the assumptions required for indirect treatment comparisons and valid population adjustment hold, by design, in the simulation study. While the simulation study provides proof-of-principle for the methods, it does not inform how robust these are to failures in assumptions. The assumptions are hard to meet and most of them are not directly testable. It is important that researchers are aware of these, as their violation may lead to biased estimates of the treatment effect. In practice, we will never come across an idealistic scenario in which all assumptions perfectly hold. Therefore, researchers should exercise caution when interpreting the results of population-adjusted analyses. These should not be taken directly at face value, but only as tools to simplify a complex reality.

Firstly, MAIC, STC and the Bucher method rely on trials AC and BC being internally valid, implying appropriate designs, the absence of non-compliance, appropriate randomization and reasonably large sample sizes. Secondly, all indirect treatment comparisons (standard or

population-adjusted) rely on consistency under parallel studies, i.e., potential outcomes are homogeneous for a given treatment regardless of the study assigned to a subject. For instance, treatment C should be administered in the same setting in both trials, or differences in the nature of treatment should not change its effect. This means that MAIC and STC cannot account for cross-trial differences that are perfectly confounded with the nature of treatments, e.g. treatment administration or dosing formulation. MAIC and STC can only account for differences in the characteristics of the trial populations.

In practice, the additional assumptions made by MAIC and STC may be problematic. Firstly, it assumed that one accounts for all effect modifiers.[6] By design, the simulation study assumes that complete information is available for both trials and that all effect modifiers are included. In practice, this assumption is hard to meet — it is difficult to ascertain the effect modifier status of covariates, particularly for new treatments with limited prior empirical evidence and clinical domain knowledge. Hence, the analyst may select the effect modifiers incorrectly. In addition, information on some effect modifiers could be unmeasured or unpublished for one of the trials. The incorrect omission of effect modifiers leads to the wrong specification of the trial assignment logistic regression model in MAIC, and of the outcome regression in STC. Relative effects will no longer be conditionally constant across trials and this will lead MAIC and STC to produce biased estimates.

In the simulation study, we know the correct data-generating mechanism, and are aware of which covariates are purely prognostic variables and which covariates are effect modifiers. This is something that one cannot typically ascertain in practice. Exploratory simulations show that the relative precision and accuracy of MAIC deteriorate, with respect to STC and the Bucher method, if we treat all four covariates as effect modifiers. This is due to the loss of effective sample size and inflation of the standard error due to the overspecification of effect modifiers.

Alternatively, it is more burdensome to specify the outcome regression model for STC than the propensity score model for MAIC; the outcome regression requires specifying both prognostic and interaction terms, while the trial assignment model in MAIC only requires the specification of effect modifiers. The relative precision and accuracy of STC deteriorate if the terms corresponding to the purely prognostic covariates are not included in the outcome regression. Nevertheless, this does not alter the conclusions of the simulation study: the other terms in the outcome regression already account for a considerable portion of the variability of the outcome and relative effects have very similar accuracy in any case.

Another assumption made by MAIC and STC, that holds in this simulation study, is that there is some overlap between the ranges of the selected covariates in AC and BC. In population

6 In the anchored scenario, we are interested in a comparison of *relative* outcomes or effects, not *absolute* outcomes. Hence, an anchored comparison only requires conditioning on the effect modifiers, the covariates that explain the heterogeneity of the A vs. C treatment effect. This assumption is denoted the *conditional constancy of relative effects* by Phillippo et al. [9, 18], i.e., given the selected effect-modifying covariates, the marginal A vs. C treatment effect is constant across the AC and BC populations. There are analogous formulations of this assumption [17, 59–61, 139], such as the conditional ignorability, unconfoundedness or exchangeability of trial assignment for such treatment effect, i.e., trial selection is conditionally independent of the treatment effect, given the selected effect modifiers. One can consider that being in population AC or population BC does not carry any information about the marginal A vs C treatment effect, once we condition on the treatment effect modifiers. This means that after adjusting for these effect modifiers, treatment effect heterogeneity and trial assignment are conditionally independent.

adjustment methods, the indirect comparison is performed in the BC population. This implies that the ranges of the covariates in the BC population should be covered by their respective ranges in the AC trial. In practice, this assumption may break down if the inclusion/exclusion criteria of AC and BC are inconsistent. When there is no overlap, weighting methods like MAIC are unable to extrapolate beyond the AC population, and may not even produce an estimate. However, STC can extrapolate beyond the covariate space observed in the AC patient-level data, using the the linearity assumption or other appropriate assumptions about the input space. Note that the validity of the extrapolation depends on accurately capturing the true covariate-outcome relationships. We view extrapolation as a desirable property because poor overlap, with small effective sample sizes and large percentage reductions in effective sample size, is a pervasive issue in HTA [30].

MAIC and STC make certain assumptions about the joint distribution of covariates in BC. Where no correlation information is available for the BC study, both methods seem to assume that the joint BC covariate distribution is the product of the published marginal distributions. The implicit assumptions are, in fact, more nuanced. In MAIC, as stated in the NICE Decision Support Unit Technical Support Document [18], "when covariate correlations are not available from the (BC) population, and therefore cannot be balanced by inclusion in the weighting model, they are assumed to be equal to the correlations amongst covariates in the pseudo-population formed by weighting the (AC) population." In the typical usage of STC, the correlations between the BC covariates are assumed to be equal to the correlations between covariates in the AC study. In the "covariate simulation" approach to STC, discussed in Section 2.4, this assumption is also made, albeit more explicitly, if the correlation structure observed in the AC IPD is used to simulate the covariates.

Indirect treatment comparisons are typically conducted on the linear predictor scale [18], upon which the treatment effect is assumed to be additive. We have assumed that the effect modifiers have been defined on the linear predictor scale and are additive on this scale. In the simulation study, it is known that effect modification is linear on the log hazard ratio scale. A central component of population-adjusted indirect comparisons is the specification of a model that is typically parametric. That is the propensity score model for the weights in MAIC or the outcome regression in STC. Parametric modeling assumptions may not be appropriate in real applications, where there is a danger of model misspecification. This is more evident in a regression adjustment method like STC, where an explicit outcome regression is formulated. The parametric model depends on functional form assumptions that will be violated if the covariate-outcome relationships are not correctly captured.

Even though the logistic regression model for the weights in MAIC does not make reference to the outcome, MAIC is also susceptible to model misspecification bias, albeit in a more implicit form. The model for estimating the weights in the simulation study is the best-case model because the right subset of covariates has been selected as effect modifiers and the balancing property holds for the weights with respect to the effect modifier means, as mentioned in subsection 3.1.4. In practice, the model can lead to a biased estimate if effect modifiers are omitted. Also, scale conflicts may arise if effect modification status, which is scale-specific,

has been justified on the wrong scale, e.g. when treatment effect modification is specified as linear but is non-linear or multiplicative, e.g. age in cardiovascular disease treatments. Note that, in practice, we find that it may be more difficult to specify a correct parametric model for the outcome than an approximately correct parametric model for the trial assignment weights.

3.4 CONCLUDING REMARKS

In the performance measures I have considered, MAIC was the least biased and most accurate method. I therefore recommend its use for survival outcomes, provided that its assumptions are reasonable. MAIC was generally randomization-valid, except in situations with poor covariate overlap and small sample sizes, where standard errors underestimated variability and there was undercoverage. STC was systematically biased because it targets a conditional treatment effect for A vs. C. This effect was incompatible in the indirect comparison due to the non-collapsibility of the log hazard ratio. The bias induced by STC could have considerable impact on decision making and policy, and could lead to perverse decisions and subsequent misuse of resources. Therefore, STC should be avoided in settings with a non-collapsible measure of effect. The Bucher method is systematically biased and overprecise when there are imbalances in effect modifiers and interaction effects that induce bias in the treatment effect.

An important objective, that I develop in the next chapter, is proposing an alternative formulation to outcome regression that estimates a marginal treatment effect for A vs. C. A crucial additional step, missing from the current implementation, is to integrate or average the conditional effect estimates over the BC covariates. Then, outcome regression could potentially obtain a marginal treatment effect estimate that is comparable to the marginal B vs. C estimate published in the BC study. This would avoid the bias caused by incompatibility in the indirect comparison and provide inference for the marginal treatment effect for A vs. B in the BC population.

4

CHAPTER 4: MARGINALIZATION OF REGRESSION-ADJUSTED TREATMENT EFFECTS: NOVEL METHODOLOGIES

The crucial element that has been missing from the application of outcome regression is the marginalization of treatment effect estimates. When adjusting for covariates, one must integrate or average the conditional estimate over the relevant joint covariate distribution to recover a marginal treatment effect that is compatible in the indirect comparison. In this chapter, I develop several methods to accomplish this and present these methods in detail.

Section 4.1 presents the context for the use of outcome regression in population-adjusted indirect comparisons, and some assumptions. Section 4.2 outlines the data structure/requirements for the methods. Section 4.3 introduces a covariate simulation step, which is necessary to approximate the joint covariate distribution of the BC population. Then, several methodologies are proposed for marginalizing the conditional effect estimates produced by the conventional covariate-adjusted outcome regression. Section 4.4 describes a marginalization method based on parametric G-computation or model-based standardization, often applied in observational studies in epidemiology and medical research where treatment assignment is non-random. Section 4.5 adapts parametric G-computation to a Bayesian statistical framework, which explicitly accounts for relevant sources of uncertainty, allows for the incorporation of prior evidence (e.g. expert opinion), and naturally integrates the analysis into a probabilistic framework, typically required for HTA. In Section 4.6, Bayesian parametric G-computation is extended to a novel general-purpose method based on the ideas underlying multiple imputation. This method is termed *multiple imputation marginalization* (MIM) and is applicable to a wide range of models, including parametric survival models. Section 4.7 clarifies how to combine the effect estimates in an indirect treatment comparison. Section 4.8 provides some insight for setting the number of resamples/syntheses in each method. Finally, Section 4.9 provides some brief concluding remarks.

Part of the research in this chapter is condensed in the article “Parametric G-computation for Compatible Indirect Treatment Comparisons with Limited Individual Patient Data” (Remiro-Azócar et al., 2021), and in the working paper “Marginalization of Regression-Adjusted Treatment Effects in Indirect Comparisons with Limited Patient-Level Data” (Remiro-Azócar et al., 2021).[1]

1 The former has been submitted to Research Synthesis Methods and is available at: https://arxiv.org/abs/2108.12208. The latter is available at: https://arxiv.org/abs/2008.05951

4.1 INTRODUCTION

4.1.1 *The need for outcome regression approaches*

Matching-adjusted indirect comparison (MAIC) is the most commonly used population-adjusted indirect comparison method [30]. In Chapter 2, we learned that "matching-adjusted" is a misnomer, as the indirect comparison is actually "weighting-adjusted", with the population adjustment based on propensity score weighting. A logistic regression is used to model the trial assignment odds conditional on a selected set of baseline covariates. The weights estimated by the model represent the "trial selection" odds, i.e., the odds of being enrolled in the BC trial as opposed to being enrolled in the AC trial. These are balancing scores that, when applied to the AC IPD, form a pseudo-population that has balanced covariate moments with respect to the BC population. The weights are often applied to a weighted simple regression to estimate the marginal treatment effect for A vs. C in the BC population. However, MAIC does not explicitly require an outcome model. The development of outcome regression methods, which estimate an outcome-generating mechanism given treatment and the baseline covariates, is appealing for several reasons:

(1) **Statistical precision and efficiency.** Outcome regression tends to give more precise estimates than weighting. Weighting is particularly inefficient and unstable where covariate overlap is poor and effective sample sizes are small [37, 125, 140–142], as it is sensitive to inordinate influence by extreme weights. Outcome regression can extrapolate the association between outcome and covariates where overlap is insufficient, while weighting methods cannot extrapolate beyond the observed covariate space in the AC IPD. Valid extrapolation, using the linearity assumption or other appropriate assumptions about the input space, requires accurately capturing the true covariate-outcome relationships.

(2) **Different modeling assumptions.** While MAIC relies on a correctly specified model for the conditional odds of trial assignment given the covariates, outcome regression methods rely on a correctly specified model for the conditional expectation of the outcome given treatment and the covariates. In my experience, identifying the variables that affect outcome is more straightforward than identifying the factors that drive trial assignment in the context of population-adjusted indirect comparisons. This is not typically the case in the standard use of propensity score weighting in observational studies, where one identifies the factors that drive treatment (as opposed to trial) assignment. Nevertheless, in our scenario, the factors driving selection into different RCTs are often arbitrary [143, 144]. Researchers may benefit from the use of distinct modeling approaches with different assumptions, as these can yield different results, especially if there is a violation of assumptions.

(3) **Flexibility.** Researchers could use augmented or doubly robust methods [53, 145–147] that combine the model for the expectation of the outcome with the trial assignment odds model. These are attractive due to their increased robustness to model misspecification: consistent estimation only requires the correct specification of either of the two models, not necessarily

both [145, 148]. Even with the reduced misspecification risk, they tend to have improved precision and efficiency with respect to the standard weighting estimators [149].

4.1.2 *Some assumptions*

Besides the differences in (typically parametric) model specification, weighting methods such as MAIC and outcome regression methods such as those discussed in this chapter mostly require the same set of assumptions. An in-depth non-technical description of these is detailed in Supplementary Appendix A. The assumptions include:

1. Internal validity of the AC and BC trials, e.g. appropriate randomization and sufficient sample sizes so that the treatment groups are comparable, no interference, negligible measurement error or missing data, the absence of non-compliance, etc.

2. Consistency under parallel studies such that both trials have identical control treatments, sufficiently similar study designs and outcome measure definitions, and have been conducted in care settings with a high degree of similarity.

3. Accounting for all effect modifiers of treatment A vs. C in the adjustment. This assumption is called the conditional constancy of the A vs. C marginal treatment effect [18], and requires that a sufficiently rich set of baseline covariates has been measured for the AC study and is available in the BC study publication. Another advantage of outcome regression with respect to weighting is that, by being less sensitive to overlap issues, it allows for the inclusion of larger numbers of effect modifiers. This makes it easier to satisfy the conditional constancy of relative effects.

4. Overlap between the covariate distributions in AC and BC. More specifically, that the ranges of the selected covariates in the AC trial cover their respective moments in the BC population. The overlap assumption (often referred to as “positivity”) can be overcome in outcome regression if one is willing to rely on model extrapolation, assuming correct model specification [53].

5. Correct specification of the BC population. Namely, that it is appropriately represented by the information available to the analyst, that does not have access to patient-level data from the BC study.

Most assumptions are causal and untestable, with their justification typically requiring prior substantive knowledge [136]. Nevertheless, we shall assume that they hold throughout the rest of the thesis. We have already discussed potential failures of assumptions in Section 3.3.3, in the context of our first simulation study, but also do so in Section 5.3.3, in the context of our second simulation study, and in Supplementary Appendix A.

4.2 DATA STRUCTURE

The outlined methodologies have the following data requirements. For the AC trial IPD, let $\mathcal{D}_{AC} = (\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})$. Here, $\boldsymbol{x}$ is a matrix of baseline characteristics (covariates), e.g. age, gender, comorbidities, baseline severity, of size $N \times K$, where N is the number of subjects in the trial and K is the number of available covariates. For each subject $n = 1, \ldots, N$, a row vector $\boldsymbol{x}_n$ of K covariates is recorded. As per Section 2.2, it is assumed that all available baseline characteristics are prognostic of the outcome and that a subset of these, $\boldsymbol{x}^{(EM)} \subseteq \boldsymbol{x}$, is selected as effect modifiers on the linear predictor scale, with a row vector $\boldsymbol{x}_n^{(EM)}$ recorded for each subject. We let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ represent a vector of outcomes and $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ is a treatment indicator ($z_n = 1$ if subject n is under treatment A and $z_n = 0$ if under C). For simplicity, we shall assume that there are no missing values in $\mathcal{D}_{AC}$. As outlined in subsection 5.3.2, the outcome regression methodologies can be readily adapted to address this issue, particularly under a Bayesian framework, but this is an area for future research.

We let $\mathcal{D}_{BC} = [\boldsymbol{\theta}, \boldsymbol{\rho}, \hat{\Delta}_{20}^{(2)}, \hat{V}(\hat{\Delta}_{20}^{(2)})]$ denote the information available for the BC study. No individual-level information on covariates, treatment or outcomes is available. Here, $\boldsymbol{\theta}$ represents a vector of published covariate summaries, e.g. proportions or means. For ease of exposition, we shall assume that these are available for all K covariates (otherwise, one would take the intersection of the available covariates), and that the selected effect modifiers are also available such that $\boldsymbol{\theta}^{(EM)} \subseteq \boldsymbol{\theta}$. An estimate $\hat{\Delta}_{20}^{(2)}$ of the B vs. C treatment effect in the BC population, and an estimate of its variance $\hat{V}(\hat{\Delta}_{20}^{(2)})$, either published directly or derived from crude aggregate outcomes in the literature, are also available. Note that these are not used in the adjustment mechanism but are ultimately required to perform inference for the indirect comparison in the BC population.

Finally, we let the symbol $\boldsymbol{\rho}$ stand for the dependence structure of the BC covariates. Under certain assumptions about representativeness, this can be retrieved from the AC trial, e.g. through the observed pairwise correlations, or from external data sources such as registries. This information, together with the published covariate summary statistics, is required to characterize the joint covariate distribution of the BC population. A pseudo-population of N^* subjects is simulated from this joint distribution, such that $\boldsymbol{x}^*$ denotes a matrix of baseline covariates of dimensions $N^* \times K$, with a row vector $\boldsymbol{x}_i^*$ of K covariates simulated for each subject $i = 1, \ldots N^*$. Notice that the value of N^* does not necessarily have to correspond to the actual sample size of the BC study; however, the simulated cohort must be sufficiently large so that the sampling distribution is stabilized, minimizing sampling variability. Again, a subset of the simulated covariates, $\boldsymbol{x}^{*(EM)} \subseteq \boldsymbol{x}^*$, makes up the treatment effect modifiers on the linear predictor scale, with a row vector $\boldsymbol{x}_i^{*(EM)}$ for each subject $i = 1, \ldots, N^*$. In this chapter, the asterisk superscript represents unobserved quantities that have been constructed in the BC population.

The outcome regression approaches discussed in this chapter estimate treatment effects with respect to a hypothetical pseudo-population for the BC study. Before outlining the specific

outcome regression methods, we explain how to generate values for the individual-level covariates x^* for the BC population using Monte Carlo simulation.

4.3 INDIVIDUAL-LEVEL COVARIATE SIMULATION

Ideally, the BC population should be characterized by the full joint distribution of covariates. However, the restriction of limited IPD makes it unlikely that the joint distribution of the BC covariates is available. Where there are not many covariates and these are binary, this is sometimes available as a cross-tabulation. However, most of the time we need to approximate the joint distribution appropriately. This is important to avoid bias arising from the incomplete specification of the BC population. The published summary values $\boldsymbol{\theta}$ and the correlation structure $\boldsymbol{\rho}$ are combined, making certain parametric assumptions about the marginal distributional forms, to infer the joint distribution of the BC covariates and construct an appropriate pseudo-population for inferences. The proposed approaches allow the analyst to bring in some prior knowledge or evidence to inform the potential distributions of the covariates. However, it is worth noting that we cannot give a general recipe for this step, which requires context-specific knowledge that is likely not available from the observed data in the trials.

Firstly, the marginal distributions for each covariate are specified. The mean and, if applicable, the standard deviation of the marginals are sourced from the BC report to match the published summary statistics. As the true marginal distributional forms are not known, some parametric assumptions are required. For instance, if it is reasonable to assume that sampling variability for a continuous covariate can be described using a normal distribution, and the covariate's mean and standard deviation are published in the BC report, we can assume that it is marginally normally distributed. Hence, we can also select the family for the marginal distribution using the theoretical validity of the candidate distributions alongside the IPD. For example, the marginal distribution of duration of prior treatment at baseline could be modeled as a log-normal or Gamma distribution as these distributions are right-skewed and bounded to the left by zero. Truncated distributions can be used to resemble the inclusion/exclusion criteria for continuous covariates in the BC trial, e.g. age limits, and avoid deterministic overlap violations.

Secondly, the correlations between covariates are specified. We suggest two possible data-generating model structures for this purpose: (1) simulating the covariates from a multivariate Gaussian copula [38, 106]; or (2) factorizing the joint distribution of the covariates into the product of marginal and conditional distributions. The former approach is perhaps more general-purpose. The latter is more flexible, defining separate models for each variable, but its specification can be daunting where there are many covariates and interdependencies are complex.

Any multivariate joint distribution can be decomposed in terms of univariate marginal distribution functions and a dependence structure [150]. A Gaussian copula "couples" the marginal distribution functions for each covariate to a multivariate Gaussian distribution function. The main appeal of a copula is that the correlation structure of the covariates and the marginal distribution for each covariate can be modeled separately. We may use the pairwise correlation

structure observed in the AC patient-level data as the dependence structure, while keeping the marginal distributions inferred from the BC summary values and the IPD. Note that the term "Gaussian" does not refer to the marginal distributions of the covariates but to the correlation structure. While the Gaussian copula is sufficiently flexible for most modeling purposes, more complex copula types (e.g. Clayton, Gumbel, Frank) may provide different and more customizable correlation structures [106].

Alternatively, we can account for the correlations by factorizing the joint distribution of covariates in terms of marginal and conditional densities. This strategy is common in implementations of sequential conditional algorithms for parametric multiple imputation [151, 152]. For instance, consider two baseline characteristics: age, which is a continuous variable, and the presence of a comorbidity c, which is dichotomous. We can factorize the joint distribution of the covariates such that $p(age, c) = p(c \mid age)p(age)$.

In this scenario, we draw age_i for subject i from a suitable marginal distribution, e.g. a normal, with the mean and standard deviation sourced from the published BC summaries or official life tables. The mean π_i^c of c (the conditional proportion of the comorbidity) given the age, can be modeled through a regression: $\pi_i^c = g^{-1}(\alpha_0^c + \alpha_1^c(age_i - \overline{age}))$, with $c_i \sim$ Bernoulli(π_i^c) where $g(\cdot)$ is an appropriate link function. Here, the coefficients α_0^c and α_1^c represent respectively the overall proportion of comorbidity c in the BC population (marginalizing out the age), and the correlation level between comorbidity c and (the centered version of) age. The former coefficient can be directly sourced from the published BC summaries, whereas the latter could be derived from pairwise correlations observed in the AC IPD or from external sources, e.g. clinical expert opinion, registries or administrative data, applying the selection criteria of the BC trial to subset the data. Figure 10 provides an example of a similar probabilistic structure with three covariates: age and the presence of two comorbidities, c and d. In this example, the distribution of the covariates is factorized such that $p(age, c, d) = p(d \mid c, age)p(c \mid age)p(age)$.

It is important to acknowledge that this "covariate simulation" step arises due to a suboptimal scenario, where patient-level data on covariates are unavailable for the BC study. Ideally, this should be freely available or, at least, disclosed by the sponsor company. Raw patient-level data are always the preferred input for statistical inference, allowing for the testing of assumptions [153]. The underlying reasons for unavailable IPD are diverse and span across a range of issues. Perhaps the most sensitive of these is privacy, with the General Data Protection Regulation [154] ratified by the European Union in 2018 recognizing data concerning health as a special category of data with specific protection safeguards and disclosure regulations.

We note that, if the main hindrance to the availability of IPD is privacy, the manufacturer itself could facilitate statistical inference by using the IPD to create fully artificial covariate datasets [155]. The release of such datasets would not involve a violation of privacy or confidentiality and would avoid the need for the "covariate simulation" step. Alternatively, Bonofiglio et al. [156] have recently proposed a framework where access to covariate correlation summaries is made possible through distributed computing. It is unclear whether access to such framework would be granted to a competitor submitting evidence for reimbursement to HTA bodies, albeit the summaries could be reported in clinical trial publications.

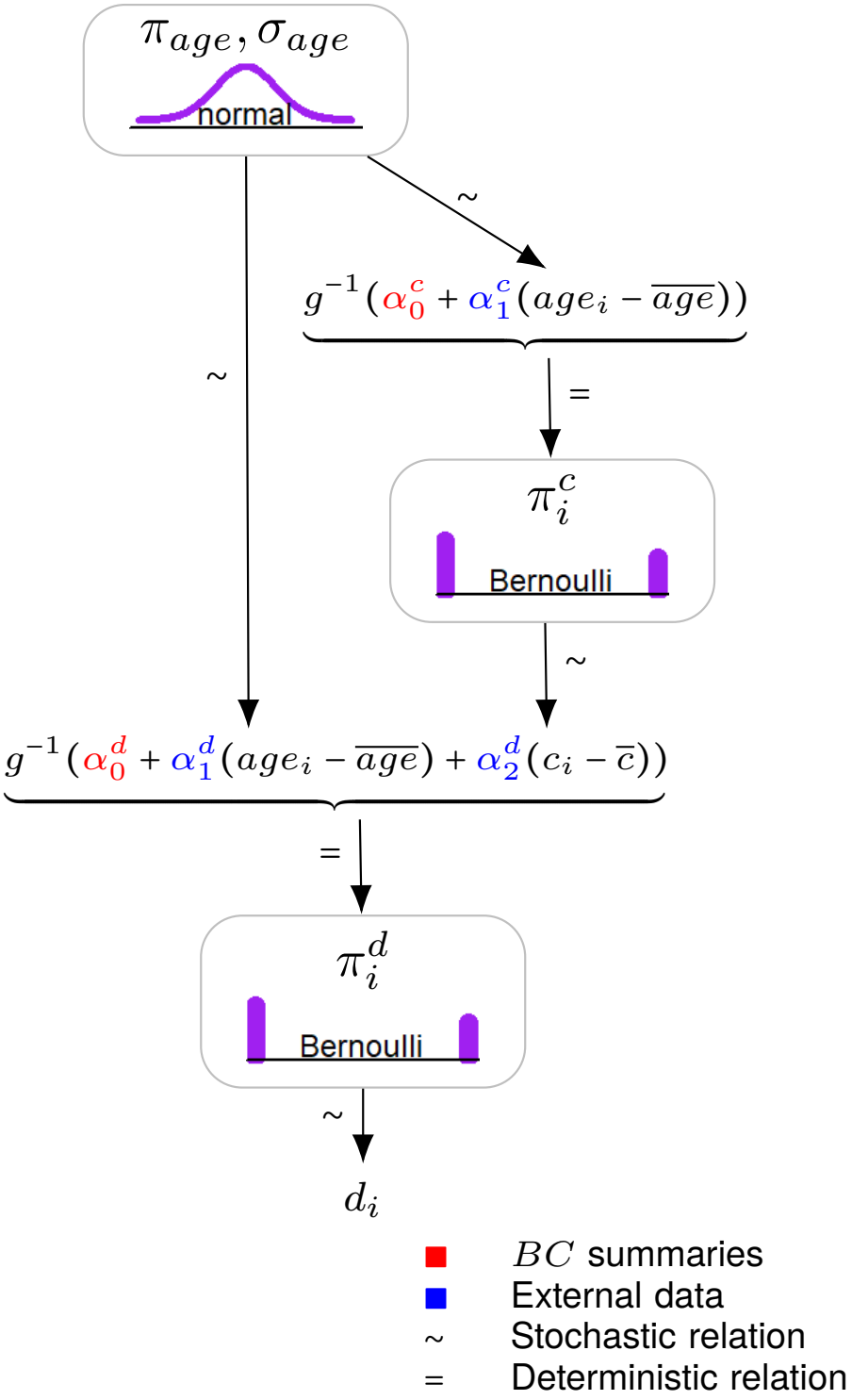


Figure 10: An example of individual-level Monte Carlo covariate simulation where the joint distribution of three baseline characteristics, age, comorbidity c and comorbidity d, is factorized into the product of marginal and conditional distributions, such that $p(age, c, d) = p(d \mid c, age)p(c \mid age)p(age)$. The joint distribution is valid because the conditional distributions defining the covariates are compatible: we start with a marginal distribution for age and construct the joint distribution by modeling each additional covariate, one-by-one, conditionally on the covariates that have already been simulated. This diagram adopts the convention of Kruschke [157].

4.4 MARGINALIZATION VIA PARAMETRIC G-COMPUTATION

The crucial element that has been missing from the typical application of outcome regression is the marginalization of the A vs. C treatment effect estimate. When adjusting for covariates, one must integrate or average the conditional estimate over the joint BC covariate distribution to recover a marginal treatment effect that is compatible in the indirect comparison. Parametric G-computation [42, 43, 158–160] is an established method for marginalizing regression-adjusted conditional estimates. The literature on population-adjusted indirect comparisons has been developed separately to G-computation, despite the close relationships between the methodologies. We build a new link between the two in the next paragraphs.

Succinctly, G-computation in this context consists of: (1) predicting the conditional outcome expectations under treatments A and C for each subject in the BC population; (2) averaging the predictions to produce marginal outcome means on the natural scale; and (3) back-transforming the averages to the linear predictor scale, contrasting the linear predictions to estimate the marginal A vs. C treatment effect in the BC population. This marginal effect is compatible in the indirect treatment comparison. This procedure is a form of standardization, a technique which has been performed for decades in epidemiology, e.g. when computing standardized mortality ratios [146]. Parametric G-computation is often called model-based standardization [44–46] because a parametric model is used to predict the conditional outcome expectations under each treatment. When the covariates and outcome are discrete, the estimation of the conditional expectations could be non-parametric, in which case G-computation is numerically identical to crude direct post-stratification [16].

G-computation marginalizes the conditional estimates by separating the regression modeling outlined for STC in Section 2.4 from the estimation of the marginal treatment effect for A vs. C. Firstly, a regression model of the observed outcome $\boldsymbol{y}$ on the covariates $\boldsymbol{x}$ and treatment z is fitted to the AC IPD:

$$g(\mu_n) = \beta_0 + \boldsymbol{x}_n\boldsymbol{\beta_1} + \left(\beta_z + \boldsymbol{x}_n^{(\boldsymbol{EM})}\boldsymbol{\beta_2}\right)\mathbb{1}(z_n = 1). \tag{7}$$

Again, μ_n is the expected outcome of subject n on the natural scale, $g(\cdot)$ is an appropriate invertible canonical link function, β_0 is the intercept, $\boldsymbol{\beta_1}$ is a vector of regression coefficients for the prognostic variables, $\boldsymbol{\beta_2}$ is a vector of interaction coefficients for the effect modifiers and β_z represents a conditional A vs. C treatment effect. Contrary to Equation 4, this regression model is not centered on the mean BC covariates for reasons we shall explain shortly. In the context of G-computation, the regression model in Equation 7 is often called the "Q-model".

Having fitted the Q-model, the regression coefficients are treated as nuisance parameters. The parameters are applied to the simulated covariates $\boldsymbol{x}^*$ to predict hypothetical outcomes for each subject under both possible treatments. Namely, a pair of predicted outcomes, also called *potential* outcomes [62], under A and under C, is generated for each subject. Because G-computation has been developed within the counterfactual framework for causal inference [42], we refer to these outcomes as counterfactual outcomes. In our description, these are

known as counterfactual to denote what outcomes might have been observed had subjects in a different population, in which the A vs. C trial was not conducted, received treatment.

Parametric G-computation typically relies on maximum-likelihood estimation to fit the regression model in Equation 7. In this case, the methodology proceeds as follows. We denote the maximum-likelihood estimate of the regression parameters as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}, \hat{\beta}_z)$. Leaving the simulated covariates $\boldsymbol{x}^*$ at their set values, we fix the treatment values, indicated by a vector $\boldsymbol{z}^* = (z_1^*, z_2^*, \ldots, z_{N^*}^*)$, for all N^*. By plugging treatment A into the maximum-likelihood regression fit for each simulated individual, we predict the marginal outcome mean, on the natural scale, when all subjects are under treatment A:

$$\hat{\mu}_1(\boldsymbol{x}^*) = \int_{\boldsymbol{x}^*} g^{-1}(\hat{\beta}_0 + \boldsymbol{x}^* \hat{\boldsymbol{\beta}}_{\mathbf{1}} + \hat{\beta}_z + \boldsymbol{x}^{*(\boldsymbol{EM})} \hat{\boldsymbol{\beta}}_{\mathbf{2}}) p(\boldsymbol{x}^*) d\boldsymbol{x}^* \tag{8}$$

$$\approx \frac{1}{N^*} \sum_{i=1}^{N^*} g^{-1}(\hat{\beta}_0 + \boldsymbol{x}_i^* \hat{\boldsymbol{\beta}}_{\mathbf{1}} + \hat{\beta}_z + \boldsymbol{x}_i^{*(\boldsymbol{EM})} \hat{\boldsymbol{\beta}}_{\mathbf{2}}). \tag{9}$$

Equation 8 follows from the law of total expectation, such that the (marginal) expected outcome is equal to the expected value of the conditional expected outcome, given the predictors. The joint probability density function for the BC covariates is denoted $p(\boldsymbol{x}^*)$. This could be replaced by a probability mass function if the covariates are discrete or by a mixture density if there is a combination of discrete and continuous covariates. Replacing the integral by the summation in Equation 9 follows from using the empirical joint distribution of the simulated covariates as a non-parametric estimator of the density $p(\boldsymbol{x}^*)$ [50].

Similar to above, by plugging treatment C into the regression fit for every simulated observation, we predict the marginal outcome mean in the hypothetical scenario in which all units are under treatment C:

$$\hat{\mu}_0(\boldsymbol{x}^*) = \int_{\boldsymbol{x}^*} g^{-1}(\hat{\beta}_0 + \boldsymbol{x}^* \hat{\boldsymbol{\beta}}_{\mathbf{1}}) p(\boldsymbol{x}^*) d\boldsymbol{x}^* \tag{10}$$

$$\approx \frac{1}{N^*} \sum_{i=1}^{N^*} g^{-1}(\hat{\beta}_0 + \boldsymbol{x}_i^* \hat{\boldsymbol{\beta}}_{\mathbf{1}}). \tag{11}$$

To estimate the marginal or population-average treatment effect for A vs. C in the linear predictor scale, one back-transforms to this scale the average predictions, taken over all subjects on the natural outcome scale, and calculates the difference between the average linear predictions:

$$\hat{\Delta}_{10}^{(2)} = g(\hat{\mu}_1) - g(\hat{\mu}_0), \tag{12}$$

where we have removed the dependence on $\boldsymbol{x}^*$ for simplicity in the notation. If the outcome model in Equation 7 is correctly specified, the estimators for the marginal outcome means on the natural scale should be consistent with respect to convergence to their true value, and so should the marginal treatment effect estimate $\hat{\Delta}_{10}^{(2)} \to \Delta_{10}^{(2)}$.

For illustrative purposes, consider a logistic regression for binary outcomes. In this case, $\hat{\mu}_1$ is the average of the individual counterfactual probabilities predicted by the regression when all participants are assigned to treatment A. Similarly, $\hat{\mu}_0$ is the average probability when everyone is assigned to treatment C. The inverse link function $g^{-1}(\cdot)$ would be the inverse logit function $\text{expit}(\cdot) = \exp(\cdot)/[1+\exp(\cdot)]$, and the average predictions in the probability scale could be substituted into Equation 12 and transformed to the log-odds ratio scale, using the logit link function.

More interpretable summary measures of the marginal contrast, e.g. odds ratios, relative risks or risk differences, can also be produced by manipulating the average natural outcome means differently than in Equation 12, mapping these to other scales. For instance, a marginal odds ratio can be estimated as $\exp[g(\hat{\mu}_1)]/\exp[g(\hat{\mu}_0)] = \frac{\hat{\mu}_1/(1-\hat{\mu}_1)}{\hat{\mu}_0/(1-\hat{\mu}_0)}$, where $g(\cdot)$ denotes the logit link function. Clinicians and epidemiologists often criticize the (log) odds ratio as a summary measure of effect, suggesting that other measures such as the relative risk are more relevant for clinical decision-making and causal inference [53, 161–163]. Nevertheless, the standard scale commonly used for performing indirect treatment comparisons is the log-odds ratio scale [6, 7, 9] and this linear predictor scale is used to define effect modification, which is scale-specific [18]. Hence, we assume that the marginal log-odds ratio is the relative effect measure of interest.

Note that the estimated absolute outcomes $\hat{\mu}_1$ and $\hat{\mu}_0$, e.g. the average outcome probabilities under each treatment in the case of logistic regression, are sometimes desirable in health economic models without any further processing [164]. In addition, these could be useful in unanchored comparisons, where there is no common comparator group included in the analysis, e.g. if the competitor trial is an RCT without a common control or a single-arm trial evaluating the effect of treatment B alone. In the unanchored case, absolute outcome means are compared across studies as opposed to relative effects. As mentioned in Chapter 2, unanchored comparisons make very strong assumptions which are largely considered impossible to meet (absolute effects are conditionally constant as opposed to relative effects being conditionally constant) [9, 18].

4.4.1 *Cox proportional hazards regression*

The most popular outcome types in applications of population-adjusted indirect comparisons are survival or time-to-event outcomes (e.g. overall or progression-free survival), and the most prevalent measure of effect is the (log) hazard ratio [30]. Therefore, developing G-computation approaches where the nuisance model is a Cox proportional hazards regression is important and useful to practitioners. In this setting, $\hat{\Delta}_{10}^{(2)}$ and $\hat{\Delta}_{20}^{(2)}$ should target marginal log hazard ratios for indirect treatment comparisons in the linear predictor scale. Something to bear in mind is that, even if Cox models are very frequently used in evidence synthesis for time-to-event data, health economic modelers typically use parametric survival models for extrapolation purposes. In Section 4.6, we develop a novel general-purpose methodology that can be used in scenarios where the outcome regression of interest is a parametric survival model.

This subsection builds on previous research by Stitleman et al. [165], Sjölander [166, 167] and Lambert [168], who have developed approaches for regression-based standardization or G-computation with Cox proportional hazards models. These approaches use standardization to adjust for measured confounders in an observational study, allowing one to obtain standardized survival curves and marginal hazard ratio estimates over the covariate distribution observed in the sample. I extend the approaches to the context of population-adjusted indirect comparisons. In this scenario, one must marginalize over the population of an external trial (the BC study) to perform an indirect comparison in such population. While model-based standardization with Cox regressions is relatively common in epidemiological research, it has not yet been applied to population-adjusted indirect comparisons in HTA, neither in health technology appraisals nor in peer-reviewed publications.

Consider that a Cox proportional hazards model has first been fitted, conditional on covariates which follow the functional form in the linear predictor of Equation 7. For the generalized linear model, we were interested in the average counterfactual outcome predictions in the natural scale. With Cox regression, the average counterfactual survival probabilities are of interest. We proceed similarly as in Equations 8-12. Leaving the simulated covariates $\boldsymbol{x}^*$ at their set values, we fix the value of treatment at z_i^* for all $i = 1, \ldots, N^*$. By plugging treatment A into the Cox regression fit for each simulated unit, we compute the expected marginal survival probability when all subjects are under treatment A [165]:

$$\hat{P}(T_1 > t) = \frac{1}{N^*} \sum_{i=1}^{N^*} \hat{S}_i^{(1)}(t \mid \boldsymbol{x}_i^*) \tag{13}$$

$$= \frac{1}{N^*} \sum_{i=1}^{N^*} \exp[-\hat{H}_0(t))]^{\exp(\hat{\beta}_0 + \boldsymbol{x}_i^* \hat{\boldsymbol{\beta}}_1 + \hat{\beta}_z + \boldsymbol{x}_i^{*(EM)} \hat{\boldsymbol{\beta}}_2)}. \tag{14}$$

Above, t denotes a particular time point and T_1 denotes a potential counterfactual event time under treatment A, such that $\hat{P}(T_1 > t)$ is the mean treatment-specific probability of surviving beyond t. In Equation 13, $\hat{S}_i^{(1)}(t \mid \boldsymbol{x}_i^*)$ denotes an estimate of the counterfactual survival probability under treatment A at time t for simulated subject i with covariates $\boldsymbol{x}_i^*$. Equation 14 follows from expressing the survival function in terms of $\hat{H}_0(t)$, an estimate of the baseline cumulative hazard function at time t, raised to the power of the exponentiated linear predictor term. Estimates of the baseline cumulative hazard are easily obtained from Cox regressions fitted with the standard survival analysis software packages.

Similarly, the expected marginal survival probability when all simulated subjects are under treatment C is given by:

$$\hat{P}(T_0 > t) = \frac{1}{N^*} \sum_{i=1}^{N^*} \hat{S}_i^{(0)}(t \mid \boldsymbol{x}_i^*) \tag{15}$$

$$= \frac{1}{N^*} \sum_{i=1}^{N^*} \exp[-\hat{H}_0(t)]^{\exp(\hat{\beta}_0 + \boldsymbol{x}_i^* \hat{\boldsymbol{\beta}}_1)}, \tag{16}$$

where T_0 denotes a potential counterfactual event time under treatment C, and $\hat{S}_i^{(0)}(t \mid \boldsymbol{x}_i^*)$ denotes the estimated counterfactual survival probability under treatment C at time t for subject i with simulated covariates $\boldsymbol{x}_i^*$. The marginal hazard at time t for treatment $z^* \in \{0,1\}$ can be expressed as the negative logarithm of the survival probability, $-\ln[\hat{P}(T_{z^*} > t)]$. Therefore, the estimate of the marginal log hazard ratio for A vs. C in the BC population at time t is:

$$\hat{\Delta}_{10,t}^{(2)} = \ln\{-\ln[\hat{P}(T_1 > t)]\} - \ln\{-\ln[\hat{P}(T_0 > t)]\}, \tag{17}$$

where $\hat{P}(T_1 > t)$ and $\hat{P}(T_0 > t)$ are obtained using Equations 13 and 14, and Equations 15 and 16, respectively.

The Cox regression assumes that the true marginal log hazard ratio is independent of time due to the proportional hazards assumption. However, as pointed out by Varadhan et al., [169] the estimate $\hat{\Delta}_{10,t}^{(2)}$ in Equation 17 may vary across different values of t. We have to set t to a specific time point, or alternatively, to estimate the marginal hazard ratio over a set of time points and display the estimates graphically. When selecting a value of t, bear in mind that, in Equation 17, the marginal log hazard ratio estimate is undefined at t for which $\hat{P}(T_{z^*} > t) = 1$ for treatment $z^* \in \{0,1\}$ [165]. A simulation procedure for marginalizing estimates of conditional hazard ratios has recently been proposed by Daniel et al. [50]. This approach should avoid these issues by averaging the marginal log hazard ratio over a set time frame, but adapting the methodology to the current setting is beyond the scope of this chapter.

One can manipulate the expected marginal survival probabilities differently than in Equation 17 to produce estimates of the marginal risk difference (the additive difference in survival probabilities) or the marginal log relative risk at a particular time point [165]. These effect measures are more easily interpreted. However, indirect treatment comparisons with survival outcomes are typically performed in the log hazard ratio scale [6], and this linear predictor scale is used to define effect modification, which is scale-specific [18]. Therefore, the marginal log hazard ratio is the relative effect measure of interest.

4.4.2 *Model fitting and selection*

The regression in Equation 7 will be our working model from now onward:

$$g(\mu_n) = \beta_0 + \boldsymbol{x}_n\boldsymbol{\beta_1} + \left(\beta_z + \boldsymbol{x}_n^{(EM)}\boldsymbol{\beta_2}\right)\mathbb{1}(z_n = 1).$$

Therefore, we briefly discuss some good practices for model fitting and model selection. Time and care should be taken to perform these exercises and fit an appropriate regression.

The inclusion of all imbalanced effect modifiers in Equation 7 is required for unbiased estimation of both the marginal and conditional A vs. C treatment effects in the BC population [170]. As discussed in Section 2.4 for STC, a strong fit of the regression model, evaluated by model checking criteria such as the residual deviance and information criteria, may increase precision. Hence, we could select the model with the lowest information criterion conditional on including all effect modifiers [170]. Model checking criteria should not guide causal decisions

on effect modifier status, which should be defined prior to fitting the outcome model. As effect-modifying covariates are likely to be good predictors of outcome, the inclusion of appropriate effect modifiers should provide an acceptable fit. In addition, note that any model comparison criteria will only provide information about the observed AC data and therefore tell just part of the story [171]. We have no information on the fit of the selected model to the BC patient-level data.

At this point, the readers may be wondering why the outcome regression is different to the model fitted in Section 2.4. In the conventional outcome regression described in 2.4 and by the NICE technical support document for STC [18], the IPD covariates are centered by plugging in the mean BC covariate values. In the Q-model required for G-computation, outlined in this section, the covariates are not centered and the regression fit is used to make predictions for the simulated covariates. The underlying reason for this has been described for generalized linear models with non-linear link functions, such as logistic or Poisson regression [172–174]. On the natural scale, averaging the individual outcome predictions made at the centered covariates of the sample does not consistently estimate the marginal mean response for the centered sample. In the words of Bartlett [172], "prediction at the mean" value of the baseline covariates for a treatment group does not result in the "marginal mean" under such treatment. Similarly, in the words of Qu and Luo [173], the "mean at mean covariates" of the study sample is generally not equivalent to the marginal response over the subjects in the sample. The former results in a conditional estimate whereas the latter produces a marginal population-level estimate, of interest in our scenario.

We have postulated a single outcome model for all subjects in the AC IPD, which includes the necessary treatment-covariate interaction terms to capture effect modification over the covariates. Nevertheless, another possible strategy is to fit two outcome models separately for each treatment group in the randomized trial, i.e., to fit one regression to the patients under treatment A and then another regression among the patients under C, then predicting the conditional outcome expectations and averaging these out on the entire simulated pseudo-population. This is perhaps a more "objective" approach to covariate adjustment, as the model-fitting is performed independently of reference to a conditional treatment effect (in this case, the fitted regressions do not have a treatment coefficient), but obviates the estimation of treatment-by-covariate interactions [125, 175]. Throughout this chapter, we consider the nuisance model in Equation 7 to be a parametric regression. Alternatively, non-parametric estimators of the conditional expectation may be less susceptible to model misspecification. We discuss the potential application of these methods in Section 6.3.

4.4.3 *Variance estimation*

From a frequentist perspective, it is not easy to derive analytically a closed-form expression for the standard error of the marginal A vs. C treatment effect with non-linear outcome models. Deriving the asymptotic distribution is not straightforward as the estimate is a non-linear function of each of the components of $\hat{\beta}$. When using maximum-likelihood estimation to fit the outcome

model, standard errors and interval estimates can be obtained using resampling-based methods such as the traditional non-parametric bootstrap [176] or the m-out-of-n bootstrap [169]. In our bootstrap implementation, we only resample the IPD of the AC trial due to patient-level data limitations for the BC study. The standard error would be estimated as the sample standard deviation of the resampled marginal treatment effect estimates. Assuming that the sample size N is reasonably large, we can appeal to the asymptotic normality of the marginal treatment effect and construct Wald-type normal distribution-based confidence intervals. Alternatively, one can construct interval estimates using the relevant quantiles of the bootstrapped treatment effect estimates, without necessarily assuming normality. This avoids relying on the adequacy of the asymptotic normal approximation, an approximation which will be inappropriate where the true model likelihood is distinctly non-normal [177], and may allow for the more principled propagation of uncertainty.

An alternative to bootstrapping for statistical inference is to simulate the parameters of the multivariable regression in Equation 7 from the asymptotic multivariate normal distribution with means set to the maximum-likelihood estimator and with the corresponding variance-covariance matrix, iterate over Equations 8-12 and compute the sample variance. This parametric simulation approach is less computationally intensive than bootstrap resampling. It has the same reliance on random numbers and may offer similar performance [178]. It is equivalent to approximating the posterior distribution of the regression parameters, assuming constant non-informative priors and a large enough sample size. Again, this large-sample formulation relies on the adequacy of the asymptotic normal approximation.

4.5 BAYESIAN PARAMETRIC G-COMPUTATION

A Bayesian approach to parametric G-computation may be beneficial for several reasons. Firstly, the maximum-likelihood estimates of the outcome regression coefficients may be unstable where the sample size N of the AC IPD is small, the data are sparse or the covariates are highly correlated, e.g. due to finite-sample bias or variance inflation. This leads to poor frequentist properties in terms of precision. A Bayesian approach with default shrinkage priors, i.e., priors specifying a low likelihood of a very large effect, can reduce variance, stabilize the estimates and improve their accuracy in these cases [158].

Secondly, we can use external data and/or contextual information on the prognostic effect and effect-modifying strength of covariates, e.g. from covariate model parameters reported in the literature, to construct informative prior distributions for $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, respectively, and skeptical priors (i.e., priors with mean zero, where the variance is chosen so that the probability of a large effect is relatively low) for the conditional treatment effect β_z, if necessary. Where meaningful prior knowledge cannot be leveraged, one can specify generic default priors instead. For instance, it is unlikely in practice that conditional odds ratios are outside the range $0.1 - 10$. Therefore, we could use a null-centered normal prior with standard deviation 1.15, which is equivalent to just over 95% of the prior mass being between 0.1 and 10. As mentioned earlier, this “weakly informative” contextual knowledge may result in shrinkage that improves accuracy

with respect to maximum-likelihood estimators [158]. Finally, it is simpler to account naturally for issues in the AC IPD such as missing data and measurement error within a Bayesian formulation [179, 180].

In the generalized linear modeling context, consider that we use Bayesian methods to fit the outcome regression model in Equation 7. The difference between Bayesian G-computation and its maximum-likelihood counterpart is in the estimated distribution of the predicted outcomes. The Bayesian approach also marginalizes, integrates or standardizes over the joint posterior distribution of the conditional nuisance parameters of the outcome regression, as well as the joint covariate distribution $p(\boldsymbol{x}^*)$. Following Keil et al. [158], Rubin [181] and Saarela et al. [182], we draw a vector of size N^* of predicted counterfactual outcomes $\boldsymbol{y}^*_{z^*}$ under each set intervention $z^* \in \{0,1\}$ from its posterior predictive distribution under the specific treatment. This is defined as $p(\boldsymbol{y}^*_{z^*} \mid \mathcal{D}_{AC}) = \int_{\boldsymbol{\beta}} p(\boldsymbol{y}^*_{z^*} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \mathcal{D}_{AC}) d\boldsymbol{\beta}$, where $p(\boldsymbol{\beta} \mid \mathcal{D}_{AC})$ is the posterior distribution of the outcome regression coefficients $\boldsymbol{\beta}$, which encode the predictor-outcome relationships observed in the AC trial IPD. The posterior predictive distribution [158] is given by:

$$p(\boldsymbol{y}^*_{z^*} \mid \mathcal{D}_{AC}) = \int_{\boldsymbol{x}^*} p(\boldsymbol{y}^* \mid \boldsymbol{z}^*, \boldsymbol{x}^*, \mathcal{D}_{AC}) p(\boldsymbol{x}^* \mid \mathcal{D}_{AC}) d\boldsymbol{x}^* \tag{18}$$

$$= \int_{\boldsymbol{x}^*} \int_{\boldsymbol{\beta}} p(\boldsymbol{y}^* \mid \boldsymbol{z}^*, \boldsymbol{x}^*, \boldsymbol{\beta}) p(\boldsymbol{x}^* \mid \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \mathcal{D}_{AC}) d\boldsymbol{\beta} d\boldsymbol{x}^*. \tag{19}$$

As noted by Keil et al. [158], the posterior predictive distribution $p(\boldsymbol{y}^*_{z^*} \mid \mathcal{D}_{AC})$ is a function only of the observed data $\mathcal{D}_{AC}$, the joint probability density function $p(\boldsymbol{x}^*)$ of the simulated BC pseudo-population, which is independent of $\boldsymbol{\beta}$, the set treatment values $\boldsymbol{z}^*$, and the prior distribution $p(\boldsymbol{\beta})$ of the regression coefficients.

In practice, the integrals in Equations 18 and 19 can be approximated numerically, using full Bayesian estimation via Markov chain Monte Carlo (MCMC) sampling. This is carried out as follows. As per the maximum-likelihood procedure, we leave the simulated covariates at their set values and fix the value of treatment to create two counterfactual datasets: one where all simulated subjects are under treatment A and another where all simulated subjects are under treatment C. The outcome regression model in Equation 7 is fitted to the original AC IPD with the treatment actually received. From this model, conditional parameter estimates are drawn from their posterior distribution $p(\boldsymbol{\beta} \mid \mathcal{D}_{AC})$, given the observed patient-level data and some suitably defined prior $p(\boldsymbol{\beta})$.

It is relatively straightforward to integrate the model-fitting and outcome prediction within a single Bayesian computation module using efficient simulation-based sampling methods such as MCMC. Assuming convergence of the MCMC algorithm, we form realizations of the parameters $\{\hat{\boldsymbol{\beta}}^{(l)} = (\hat{\beta}_0^{(l)}, \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(l)}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{(l)}, \hat{\beta}_z^{(l)}) : l = 1, 2, \ldots, L\}$, where L is the number of MCMC draws after convergence and l indexes each specific draw. Again, these conditional coefficients are nuisance parameters, not of direct interest in our scenario. Nevertheless, the samples are used to extract draws of the conditional expectations for each simulated subject i (the posterior draws of the linear predictor transformed by the inverse link function) from their posterior

distribution. The l-th draw of the conditional expectation for simulated subject i set to treatment A is:

$$\hat{\mu}_{1,i}^{(l)} = g^{-1}(\hat{\beta}_0^{(l)} + \boldsymbol{x}_i^* \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(l)} + \hat{\beta}_z^{(l)} + \boldsymbol{x}_i^{*(\boldsymbol{EM})} \hat{\boldsymbol{\beta}}_{\mathbf{2}}^{(l)}). \tag{20}$$

Similarly, the l-th draw of the conditional expectation for simulated subject i under treatment C is:

$$\hat{\mu}_{0,i}^{(l)} = g^{-1}(\hat{\beta}_0^{(l)} + \boldsymbol{x}_i^* \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{(l)}). \tag{21}$$

The conditional expectations drawn from Equations 20 and 21 are used to impute the individual-level outcomes $\{y_{1,i}^{*(l)} : i = 1,\ldots,N^*; l = 1,2,\ldots,L\}$ under treatment A and $\{y_{0,i}^{*(l)} : i = 1,\ldots,N^*; l = 1,2,\ldots,L\}$ under treatment C, as independent draws from their posterior predictive distribution at each iteration of the MCMC chain. For instance, if the outcome model is a normal linear regression with a Gaussian likelihood, one multiplies the simulated covariates and the set treatment z_i^* for each subject i by the l-th random draw of the posterior distribution of the regression coefficients, given the observed IPD and some suitably defined prior, to form draws of the conditional expectation $\hat{\mu}_{z^*,i}^{(l)}$ (which is equivalent to the linear predictor because the link function is the identity link in linear regression). Then each predicted outcome $y_{z^*,i}^{*(l)}$ would be drawn from a normal distribution with mean equal to $\hat{\mu}_{z^*,i}^{(l)}$ and standard deviation equal to the corresponding posterior draw of the error standard deviation. With a logistic regression as the outcome model, one would impute values of a binary response $y_{z^*,i}^{*(l)}$ by random sampling from a Bernoulli distribution with mean equal to the expected conditional probability $\hat{\mu}_{z^*,i}^{(l)}$.

Producing draws from the posterior predictive distribution of outcomes is fairly simple using dedicated Bayesian software such as `BUGS` [183], `JAGS` [184] or `Stan` [185], where the outcome regression and prediction can be implemented simultaneously in the same module. Over the L MCMC draws, these programs typically return a $L \times N^*$ matrix of simulations from the posterior predictive distribution of outcomes. The l-th row of this matrix is a vector of outcome predictions of size N^* using the corresponding draw of the regression coefficients from their posterior distribution. We can estimate the marginal treatment effect for A vs. C in the BC population by: (1) averaging out the imputed outcome predictions in each draw over the simulated subjects, i.e., over the columns, to produce the marginal outcome means on the natural scale; and (2) taking the difference in the sample means under each treatment in a suitably transformed scale. Namely, for the l-th draw, the A vs. C marginal treatment effect is:

$$\hat{\Delta}_{10}^{(2,l)} = g\left(\frac{1}{N^*}\sum_{i=1}^{N^*} y_{1,i}^{*(l)}\right) - g\left(\frac{1}{N^*}\sum_{i=1}^{N^*} y_{0,i}^{*(l)}\right). \tag{22}$$

The average, variance and interval estimates of the marginal treatment effect can be derived empirically from draws of the posterior density, i.e., by taking the sample mean, variance and the relevant percentiles over the L draws, which approximate the posterior distribution of the marginal treatment effect. The computational expense of the Bayesian approach to G-computation is expected to be similar to that of the maximum-likelihood version, given that the latter typically requires bootstrapping for uncertainty quantification. Computational cost can

be reduced by adopting approximate Bayesian inference methods such as integrated nested Laplace approximation (INLA) [186] instead of MCMC sampling to draw from the posterior predictive distribution of outcomes.

Note that Equation 22 is the Bayesian version of Equation 12. Other parameters of interest can be obtained, e.g. the risk difference by using the identity link function in this equation, but these are typically not of direct relevance in our scenario. Again, where the contrast between two different interventions is not of primary interest, the absolute outcome draws from their posterior predictive distribution under each treatment may be relevant. The average, variance and interval estimates of the absolute outcomes can be derived empirically over the L draws. An argument in favor of a Bayesian approach is that, once the simulations have been conducted, one can obtain a full characterization of uncertainty on any scale of interest.

In the Cox regression scenario described in subsection 4.4.1, Bayesian G-computation would follow a similar approach, and would involve drawing the marginal survival probabilities under each treatment from their posterior predictive distribution.

4.6 MULTIPLE IMPUTATION MARGINALIZATION

We now develop a general-purpose marginalization procedure labeled multiple imputation marginalization (MIM) because it contains many similarities to multiple imputation. This procedure might be useful where the effect measure of interest cannot be readily summarized in terms of predicted outcomes and G-computation cannot be easily applied. An example scenario where this is the case is when the outcome model is a parametric survival regression. Parametric survival distributions, e.g. exponential, Weibull, Gompertz, log-logistic, log-normal and generalized gamma, are commonly used in health economic evaluations to extrapolate published Kaplan-Meier survival curves from the clinical trial follow-up period to a lifetime horizon [105, 187–190]. As well as permitting survival extrapolation, these may allow for non-proportional and time-varying hazards — for instance, the treatment coefficient of (a parametrization of) the Weibull, log-logistic or log-normal models is interpreted as an "acceleration factor" as opposed to a hazard ratio [105]. The area under the extrapolation is used to estimate the mean survival benefit of an intervention in cost-effectiveness analyses, typically in terms of life years or quality-adjusted life years. Parametric survival models are particularly of interest in oncology health technology appraisals.

Consider that the outcome model of interest is a parametric survival model. In this scenario, an anchored regression-adjusted indirect comparison would be conducted as follows: (1) a univariable parametric survival regression of outcome on treatment group is fitted to the BC trial data (the subject-level data is typically reconstructed from digitized Kaplan-Meier curves, e.g. using the algorithm by Guyot et al. [191]); (2) a multivariable covariate-adjusted parametric survival model (of the same family as the model in Step 1) is fitted to the AC trial data with treatment group as a covariate; (3) the coefficients of the covariate-adjusted regression are marginalized to derive a marginal treatment effect estimate for A vs. C in the BC population (with the location or rate coefficient and, potentially, ancillary coefficients such as shapes being

treated as nuisance parameters); and (4) this relative treatment effect is applied to the survival curve of common comparator C in the BC study to yield a survival curve for treatment A in the BC population. The marginalization procedure in Step 3 cannot be readily conducted in this scenario using parametric G-computation because the treatment effect measure is difficult to summarize in terms of the potential outcomes. This motivates the development of a general-purpose framework such as MIM.

Conceptually, MIM splits the population adjustment into two separate stages: (1) the generation (*synthesis*) of synthetic datasets; and (2) the *analysis* of the generated datasets. The synthesis is completely separated from the analysis — only after the synthesis has been completed is the marginal effect of treatment on the outcome estimated. This is analogous to the separation between design and analysis in propensity score methods, between imputation and analysis in multiple imputation, or between fitting (and predicting outcomes with) the Q-model and estimating the marginal treatment effect in G-computation.

Similarly to Bayesian G-computation, MIM sits naturally within a Bayesian framework in integrating different sources of evidence to fully characterize probabilistic relationships among a set of relevant variables, using a simulation approach. A more detailed explanation of each module is provided below. Figure 11 displays a Bayesian directed acyclic graph (DAG) summarizing the general MIM structure and the links between the modules. In this graphical representation, the nodes represent the variables of the model (constants are denoted as squares and stochastic nodes are circular); single arrows indicate probabilistic relationships and double arrows indicate logical functions. The plate notation indicates repeated analyses. We return to Figure 11 and provide further explanations for the notation throughout this section. For consistency with the rest of the chapter, MIM is presented within a generalized linear modeling formulation. Nevertheless, its formal integration in a unified parametric survival analysis framework for HTA, which contains many particularities, is a necessary and important piece of currently ongoing research.

4.6.1 *Generation of synthetic datasets: a missing data problem*

The first stage, synthetic data generation, consists of two steps. Initially, the *first-stage regression* builds a model to capture the relationship between the outcome $\boldsymbol{y}$ and the covariates $\boldsymbol{x}$ and treatment z in the observed IPD. In the *outcome prediction* step, we generate predicted outcomes $\boldsymbol{y}^*$ for A and C in the BC population by drawing from the posterior predictive distribution of outcomes, given the observed predictor-outcome relationships in the AC trial IPD, the simulated covariates $\boldsymbol{x}^*$ and the set treatment.

These steps are identical to those described for the Bayesian G-computation procedure in Section 4.5. Interestingly, Bayesian G-computation follows closely the basic principles of multiple imputation [48]. This is a simulation technique where missing data points are replaced with a set of plausible values conditional on some pre-specified imputation mechanism. Multiple imputation can be regarded as a fundamentally Bayesian operation [48, 171, 192], as the imputed outcomes are drawn from the posterior predictive distribution of observed outcomes.

Our problem can be conceptualized as a missing data problem, where the individual-level outcomes for treatments A and C in the BC population are treated as systematically missing data under a complete case analysis [193]. Namely, we only observe the outcomes for the subjects in the AC trial, with the outcomes experienced in the BC population "missing". The imputation mechanism would be the statistical model in Equation 7 relating the outcomes $\boldsymbol{y}$ to the predictors $(\boldsymbol{x}, \boldsymbol{z})$. This dependence structure is estimated using the original IPD and used to construct the posterior predictive distribution of outcomes.

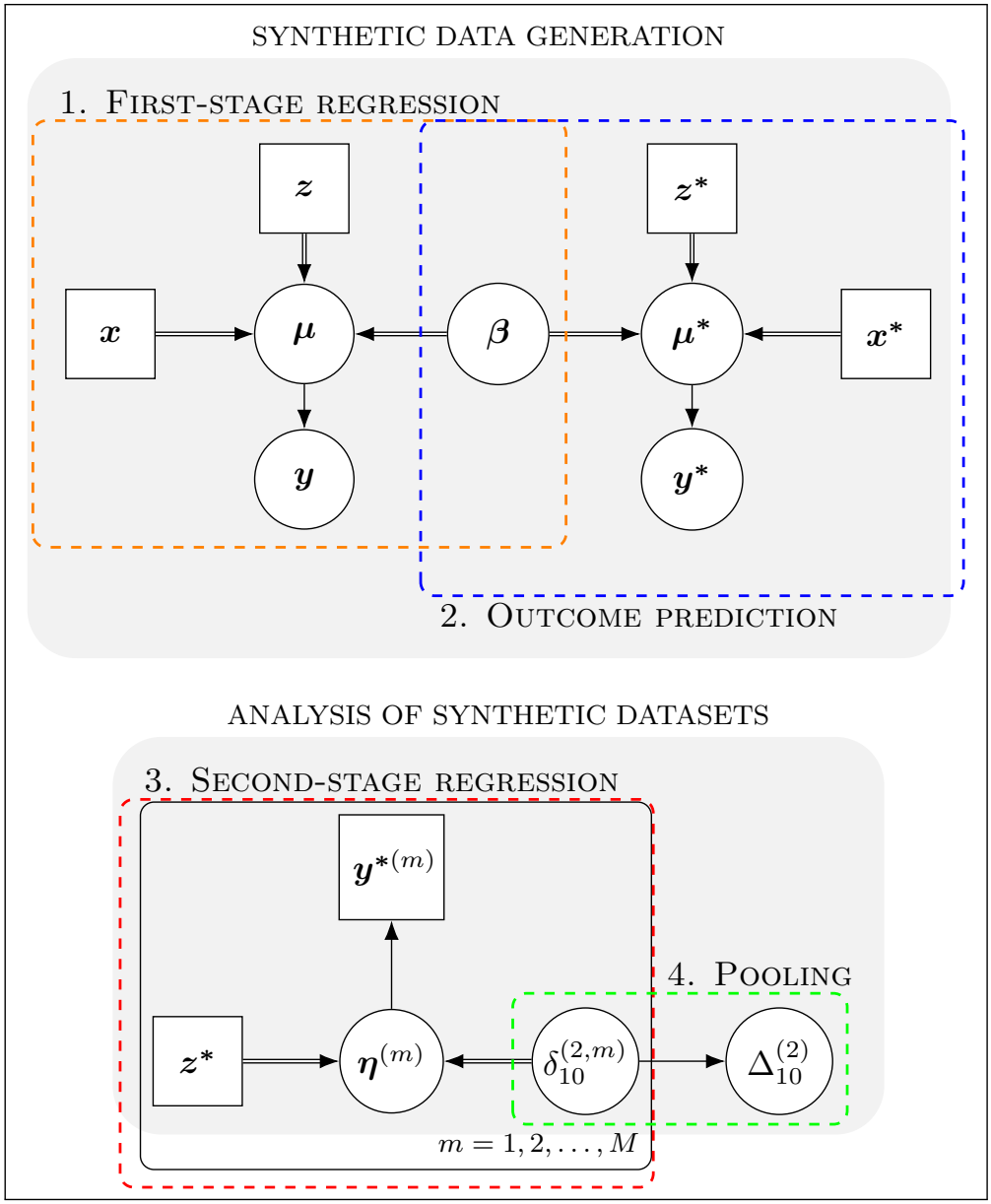


Figure 11: A Bayesian directed acyclic graph representing multiple imputation marginalization (MIM) and accounting for its two main stages: (1) synthetic data generation; and (2) the analysis of synthetic datasets. Square nodes represent constant variables, circular nodes indicate stochastic variables, single arrows denote stochastic dependence, double arrows indicate logical relationships and the plate notation indicates repeated analyses. The difference between MIM and Bayesian G-computation is that MIM requires specifying a marginal structural model for each synthesis, the second-stage regression, in the analysis stage. The results of these regressions are then pooled across all syntheses.

Practically, we may frame Bayesian G-computation as conducting L hypothetical trials comparing A vs. C in the BC population. Extending the parallel with the missing data literature, the outcome-generation process in these trials is based on the assumption of a missing-at-random mechanism. Namely, the missing relative outcomes for A vs. C in the BC population are conditionally exchangeable with those observed in the AC population (conditioning on the predictors that have been adjusted for). Therefore, this missing-at-random assumption is analogous to the conditional constancy of relative effects, which is untestable using the available data alone.

There is one conceptual difference between the synthesis stage of MIM and parametric G-computation, which arises in order to facilitate the presentation of MIM. Instead of considering two counterfactual datasets with N^* subjects each (one under treatment A and the other under treatment C), each synthesis considers a single dataset with N^* individuals that maintains the original treatment allocation ratio of the AC trial. Treatment in all synthetic datasets will be fixed to $\boldsymbol{z}^* = (z_1^*, \ldots, z_{N^*}^*)$, a vector of size N^*. In practice, this different conceptualization will not make a difference provided that N^* is reasonably large (different values of N^* are explored in Supplementary Appendix D, in the context of a simulation study in Chapter 5). This is because, in the synthetic samples, we "enforce" the randomization of individuals into A and C by simulating the covariates for active treatment and control arms combined.

As per Bayesian G-computation, the synthesis stage can be performed using MCMC. Iterating over the L converged draws of the MCMC algorithm, one generates $M \leq L$ synthetic datasets, $\mathcal{D}_{AC}^* = \{\mathcal{D}_{AC}^{*(m)} : m = 1, 2 \ldots, M\}$, where $\mathcal{D}_{AC}^{*(m)} = (\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^{*(m)})$. Here, $\boldsymbol{x}^*$ is a $N^* \times K$ matrix of individual-level BC covariates, drawn from their approximate joint distribution as per Section 4.3, and $\boldsymbol{z}^*$ is the assigned treatment in the syntheses, as previously described. Each $\boldsymbol{y}^{*(m)}$ is a vector of predicted outcomes of size N^*. We fill in $\boldsymbol{y}^{*(m)}$ by drawing from its posterior predictive distribution. In line with the multiple imputation framework, these draws are repeated independently M times to create M completed syntheses, with the posterior samples making up the imputed datasets. In standard multiple imputation, it is not uncommon to release as little as 5 imputed datasets [48, 194]. However, MIM is likely to require a larger value of M as it imputes an entire dataset as opposed to a relatively small proportion of missing values, i.e., the "fraction of missing information" is 1.

Within a survival analysis framework, the fitted first-stage regression would be used to predict survival times in the simulated pseudo-population for the BC study. One would assume that censoring is non-informative, which is an assumption made, in any case, by the Cox proportional hazards regression and the standard parametric survival models. Namely, one would not attempt to simulate censoring according to any given distribution, or to mimic any particular censoring pattern (essentially, assuming that all times are uncensored in the simulated data structure).

4.6.2 *Analysis of synthetic datasets*

In the second stage, the analysis of synthetic datasets, we seek inferences about the marginal A vs. C treatment effect in the BC population, $\Delta_{10}^{(2)}$, given the synthesized outcomes. This will ultimately be compared with the treatment effect for B vs. C, to produce a marginal treatment effect for A vs. B in the BC population. The analysis stage consists of another two steps. In the *second-stage regression* step, we regress each predicted outcome $\boldsymbol{y}^{*(m)}$ on the treatment indicator z^* alone, to generate estimates of the marginal A vs. C treatment effect in each synthesis. This second-stage regression is effectively a marginal structural model of outcome on treatment [195], and adds some computational expense with respect to Bayesian parametric G-computation. In the *pooling* step, the treatment effect estimates and their variances are combined across all M syntheses to produce an estimate of the average marginal treatment effect in the BC population.

In a typical problem involving multiple imputation, the imputation (i.e., synthesis) and analysis stages may be performed simultaneously in a joint model [171]. However, this is problematic in MIM as the dependent variable $\boldsymbol{y}^*$ of the analysis is completely synthesized. Consider the Bayesian DAG in Figure 11. In a joint model, the predicted outcomes are a *collider* variable and block the only path between the first and the second module, i.e., information from the directed arrows "collides" at the node. Due to these non-trivial joint modeling issues, we have considered the data synthesis and analysis stages as separate units in a two-stage modular framework. The analysis stage conditions on the response variable predicted by the synthesis stage, treating it as observed data.

4.6.2.1 *Second-stage regression*

We fit M second-stage regressions of predicted outcomes $\boldsymbol{y}^*$ on the treatment $\boldsymbol{z}^*$. Identical analyses are performed on each $\boldsymbol{y}^{*(m)}$ ($\boldsymbol{z}^*$ is fixed), such that for $m = 1, 2, \ldots, M$:

$$g(\eta_i^{(m)}) = \delta_0^{(m)} + \delta_{10}^{(m)} z_i^*, \tag{23}$$

where $\eta_i^{(m)}$ is the expected outcome on the natural scale of subject i in the m-th synthesis, the coefficient $\delta_0^{(m)}$ is an intercept term and $\delta_{10}^{(m)}$ denotes the marginal A vs. C treatment effect in the m-th synthesis. There is some non-trivial computational complexity to performing a Bayesian fit in this step. This would embed a nested simulation scheme. Namely, if we draw M samples $\{\boldsymbol{y}^{*(m)} : m = 1, 2, \ldots M\}$ in the synthesis stage, a further number of samples, say R, of the treatment effect $\{\hat{\delta}_{10}^{(m,r)} : m = 1, 2 \ldots M; r = 1, 2, \ldots R\}$ would be drawn for each of these realizations separately. This structure is likely to be unfeasible in terms of running time and we choose to prioritize computational efficiency.

Using maximum-likelihood estimation, we generate a point estimate $\hat{\delta}_{10}^{(m)}$ of the marginal treatment effect and a measure of its variance $\hat{v}^{(m)}$ in each synthesis $\boldsymbol{y}^{*(m)}$. The model is relatively simple as we have enforced randomization in the trial by simulating covariates for both arms jointly. Hence, this step emulates the unadjusted analysis of an RCT. A marginal

treatment effect estimate is produced because a simple regression of outcome on treatment alone is performed (covariate adjustment was performed by the first-stage regression, i.e., the Q-model), with the fitted coefficient $\hat{\delta}_{10}^{(m)}$ estimating a relative effect between subjects that, on expectation, have the same distribution of covariates [91]. Assigned treatment was already included as a predictor in the first-stage regression. Hence, the second-stage regression is more restrictive and therefore "congenial" (i.e., compatible for unbiased estimation) with the synthesis stage [192].

4.6.2.2 *Pooling*

We now combine the M point estimates of the A vs. C treatment effect and their variances to generate a posterior distribution for the A vs. C marginal treatment effect, in the BC population. Due to the two-stage structure of MIM, it is necessary to pool the estimates across the analyses to estimate this effect. The analysis of a single synthesis accounts for two sources of uncertainty: (1) the uncertainty in the regression coefficients used to generate the predicted outcomes; and (2) prediction error or random individual variation. However, it will produce attenuated measures of variability for the average A vs. C treatment effect. We must account for a third source of variation to produce valid statistical inference: the uncertainty due to the data being synthesized. This is incorporated by pooling across multiple syntheses, a question shared with the domain of statistical disclosure limitation [196–202].

In statistical disclosure limitation, data agencies mitigate the risk of identity disclosure by releasing multiple *fully synthetic* datasets, i.e., datasets that only contain simulated values, in lieu of the original confidential data of real survey respondents. Raghunathan et al. [196] describe full synthesis as a two-step process: (1) construct multiple synthetic populations by repeatedly drawing from the posterior predictive distribution, conditional on a model fitted to the original data; and (2) draw random samples from each synthetic population and release these synthetic samples to the public. In practice, as indicated by Reiter and Raghunathan [201], it is not a requirement to generate the populations, but only to generate values for the synthetic samples. Once the samples are released, the analyst seeks inferences based on the synthetic data alone.

MIM is analogous to this problem, albeit there are some differences. In MIM, the analyst also acts as the synthesizer of data, and there is no "original data" on outcomes as such – the AC trial has not been conducted in the BC population. In any case, values for the samples are generated in the synthesis stage by repeatedly drawing from the posterior predictive distribution of outcomes. This is conditional on the predictor-outcome relationships indexed by the model fitted to the AC IPD, and on the simulated covariates.

The target for inference in this step is the marginal A vs. C treatment effect conditional on the synthetic outcomes (and treatment), i.e., we seek to construct the posterior distribution $p(\Delta_{10}^{(2)} \mid \boldsymbol{y}^*, \boldsymbol{z}^*)$. Following Raab et al. [202], we view each $\boldsymbol{y}^{*(m)}$ as a random sample from $p(\boldsymbol{y}^* \mid \hat{\boldsymbol{\beta}}^{(m)}, \boldsymbol{x}^*, \boldsymbol{z}^*)$, where $\hat{\boldsymbol{\beta}}^{(m)}$ is sampled from its posterior $p(\boldsymbol{\beta} \mid \mathcal{D}_{AC})$. Hence, the "true" marginal treatment effect $\delta_{10}^{(m)}$ for the m-th synthesis, corresponding to $\hat{\boldsymbol{\beta}}^{(m)}$, can be defined as

a function of this sample. In each second-stage regression in Equation 23, this is the treatment effect estimated by $\hat{\delta}_{10}^{(m)}$.

Therefore, following Raghunathan et al. [196], the estimators $\{\hat{\delta}_{10}^{(m)}, \hat{v}^{(m)}; m = 1, 2, \ldots, M\}$ from the second-stage regressions are treated as "data", and are used to construct an approximation to the posterior density $p(\Delta_{10}^{(2)} \mid \boldsymbol{y}^*, \boldsymbol{z}^*)$. This density is assumed to be approximately normal and is parametrized by its first two moments: the mean μ_Δ, and the variance σ_Δ^2. To derive the conditional distribution $p(\mu_\Delta, \sigma_\Delta^2 \mid \boldsymbol{y}^*, \boldsymbol{z}^*)$ of these moments given the syntheses, the estimators $\{\hat{\delta}_{10}^{(m)}, \hat{v}^{(m)}; m = 1, 2, \ldots, M\}$, where $\hat{v}^{(m)}$ is the point estimate of the variance in the m-th second-stage regression, are treated as sufficient summaries of the syntheses, and μ_Δ and σ_Δ^2 are treated as parameters. Then, the posterior distribution $p(\Delta_{10}^{(2)} \mid \boldsymbol{y}^*, \boldsymbol{z}^*)$ is constructed as:

$$p(\Delta_{10}^{(2)} \mid \boldsymbol{y}^*, \boldsymbol{z}^*) = \int_{\mu_\Delta, \sigma_\Delta^2} p(\Delta_{10}^{(2)} \mid \mu_\Delta, \sigma_\Delta^2) p(\mu_\Delta, \sigma_\Delta^2 \mid \boldsymbol{y}^*, \boldsymbol{z}^*) d(\mu_\Delta, \sigma_\Delta^2). \tag{24}$$

We have two options to approximate the posterior distribution. The first involves direct Monte Carlo simulation and the second uses a simple normal approximation. In analogy with the theory of multiple imputation [48], both approaches require the following quantities for inference:

$$\bar{\delta}_{10} = \sum_{m=1}^{M} \hat{\delta}_{10}^{(m)} / M, \tag{25}$$

$$\bar{v} = \sum_{m=1}^{M} \hat{v}^{(m)} / M, \tag{26}$$

$$b = \sum_{m=1}^{M} (\hat{\delta}_{10}^{(m)} - \bar{\delta}_{10})^2 / (M - 1), \tag{27}$$

where $\bar{\delta}_{10}$ is the average of the treatment effect point estimates across the M syntheses, $\bar{v}$ is the average of the point estimates of the variance (the "within" variance), and b is the sample variance of the point estimates (the "between" variance). These quantities are computed using the point estimates from the second-stage regressions.

In many applications, the target estimand is non-scalar and has multiple components. For instance, this is the case where the outcome model is a multivariate regression (i.e., with multiple dependent variables) of correlated outcomes with treatment. This scenario typically evaluates surrogate endpoints and involves combining correlated treatment effects corresponding to multiple outcomes [203]. The inferential framework for pooling outlined in this section is extended to multi-component estimands in Supplementary Appendix B.

POOLING VIA POSTERIOR SIMULATION After deriving the quantities in Equations 25, 26 and 27, the posterior in Equation 24 is approximated by direct Monte Carlo simulation. Firstly, one draws μ_Δ and σ_Δ^2 from their posterior distributions, conditional on the syntheses. These distributions are derived by Raghunathan et al. [196]. We draw values of μ_Δ from a normal distribution:

$$p(\mu_\Delta \mid \boldsymbol{y}^*, \boldsymbol{z}^*) \sim \mathrm{N}(\bar{\delta}_{10}, \bar{v}/M), \tag{28}$$

We draw values of σ_Δ^2 from a chi-squared distribution with $M-1$ degrees of freedom:

$$p\left((M-1)b/(\sigma_\Delta^2+\bar{v}) \mid \boldsymbol{y}^*, \boldsymbol{z}^*\right) \sim \chi^2_{M-1}. \tag{29}$$

Given draws of μ_Δ and σ_Δ^2, we draw values of $\Delta_{10}^{(2)}$ from a t-distribution with $M-1$ degrees of freedom [196]:

$$p(\Delta_{10}^{(2)} \mid \mu_\Delta, \sigma_\Delta^2) \sim t_{M-1}\left(\mu_\Delta, (1+1/M)\sigma_\Delta^2\right), \tag{30}$$

where the σ_Δ^2/M term in the variance is necessary as an adjustment for there being a finite number of syntheses; as $M \to \infty$, the variance tends to σ_Δ^2.

By performing a large number of simulations, we are estimating the posterior distribution in Equation 24 by approximating the integral of the posterior in Equation 30 with respect to the posteriors in Equations 28 and 29 [196]. Hence, the resulting draws of $\Delta_{10}^{(2)}$ are samples from the posterior distribution $p(\Delta_{10}^{(2)} \mid \boldsymbol{y}^*, \boldsymbol{z}^*)$ in Equation 24. We can take the expectation over the posterior draws to produce a point estimate $\hat{\Delta}_{10}^{(2)}$ of the marginal A vs. C treatment effect, in the BC population. An estimate of its variance $\hat{V}(\hat{\Delta}_{10}^{(2)})$ can be directly computed from the draws of the posterior density. Uncertainty measures such as 95% interval estimates can be calculated from the corresponding empirical quantiles.

The posterior distributions in Equations 28, 29 and 30 have been derived under certain normality assumptions, which are adequate for reasonably large sample sizes, where the relevant sample sizes are both the size of the AC trial and the size N^* of the synthetic datasets. Another assumption is that priors for the parameters in this step are diffuse, i.e., non-informative in the range where the posteriors have support from the data [196].

POOLING VIA COMBINING RULES A simple alternative to direct Monte Carlo simulation is to use a basic normal approximation to the posterior density in Equation 24, such that the sampling distribution in Equation 30 is a Normal as opposed to a t-distribution. The posterior mean is the average of the treatment effect point estimates across the M syntheses. The simple combining rule for the variance arises from using $b-\bar{v}$ to estimate σ_Δ^2, which is equivalent to setting σ_Δ^2 at its approximate posterior mean in Equation 29 [200]. Again, the b/M term is necessary as an adjustment for there being a finite number of syntheses.

Consequently, point estimates for the A vs. C treatment effect and its variance can be derived using the following plug-in estimators:

$$\hat{\Delta}_{10}^{(2)} = \bar{\delta}_{10}, \tag{31}$$

$$\hat{V}\left(\hat{\Delta}_{10}^{(2)}\right) = (1+1/M)b - \bar{v}. \tag{32}$$

Interval estimates can be approximated using a normal distribution, e.g. taking ± 1.96 times the square root of the variance computed in Equation 32 [196]. A heavier-tailed t-distribution with $\nu_f = (M-1)(1+\bar{v}/[(1+1/M)b)]^2$ degrees of freedom has also been proposed, as normal distributions may produce excessively narrow intervals and undercoverage when M is

more modest [198]. Note that the combining rules in Equations 31 and 32 are only appropriate for reasonably large M. The choice of M is discussed in Section 4.8.

4.7 INDIRECT TREATMENT COMPARISON

The estimated marginal treatment effect for A vs. C is typically compared with that for B vs. C to estimate the marginal treatment effect for A vs. B in the BC population. This is the indirect treatment comparison in the BC population performed in Equation 2.

There is some flexibility in this step. Bayesian G-computation and the indirect comparison can be performed in one step under an MCMC approach. Similarly, so can the MIM pooling stage and the indirect comparison. In these cases, the estimation of $\Delta_{20}^{(2)}$ would be integrated within the estimation or simulation of the posterior of $\Delta_{10}^{(2)}$, under suitable priors, and a posterior distribution for $\Delta_{12}^{(2)}$ would be generated. This would require inputting as data the available aggregate outcomes for each treatment group in the published BC study, or reconstructing subject-level data from these outcomes. For binary outcomes, event counts from the cells of a 2×2 contingency table would be required to estimate probabilities of the binary outcome as the incidence proportion for each treatment (dividing the number of subjects with the binary outcome in a treatment group by the total number of subjects in the group), to then estimate a marginal log-odds ratio for B vs. C. For survival outcomes, one can input patient-level data (with outcome times and event indicators for each subject) reconstructed from digitized Kaplan-Meier curves, e.g. using the algorithm by Guyot et al. [191].

The advantage of this approach is that it directly generates a full posterior distribution for $\Delta_{12}^{(2)}$. Hence, its output is perfectly compatible with a probabilistic cost-effectiveness model. Samples of the posterior are directly incorporated into the decision analysis, so that the relevant economic measures can be evaluated for each sample without further distributional assumptions [6]. If necessary, we can take the expectation over the draws of the posterior density to produce a point estimate $\hat{\Delta}_{12}^{(2)}$ of the marginal A vs. B treatment effect, in the BC population. Variance and interval estimates are derived empirically from the draws.

Alternatively, we can perform the G-computation and indirect comparison, or the MIM pooling and indirect comparison, in two steps. Irrespective of the selected inferential framework, point estimates $\hat{\Delta}_{10}^{(2)}$ and $\hat{\Delta}_{20}^{(2)}$ can be directly substituted in Equation 2. As the associated variance estimates $\hat{V}(\hat{\Delta}_{10}^{(2)})$ and $\hat{V}(\hat{\Delta}_{20}^{(2)})$ are statistically independent, these are summed to estimate the variance of the A vs. B treatment effect:

$$\hat{V}(\hat{\Delta}_{12}^{(2)}) = \hat{V}(\hat{\Delta}_{10}^{(2)}) + \hat{V}(\hat{\Delta}_{20}^{(2)}). \tag{33}$$

With relatively large M and sample sizes, interval estimates can be constructed using normal distributions, $\hat{\Delta}_{12}^{(2)} \pm 1.96\sqrt{\hat{V}(\hat{\Delta}_{12}^{(2)})}$. This two-step strategy is simpler and easier to apply but sub-optimal in terms of integration with probabilistic sensitivity analysis, although one could perform forward Monte Carlo simulation from a normal distribution with mean $\hat{\Delta}_{12}^{(2)}$ and variance $\hat{V}(\hat{\Delta}_{12}^{(2)})$. Ultimately, it is the distribution of $\Delta_{12}^{(2)}$ that is relevant for HTA purposes.

4.8 NUMBER OF RESAMPLES OR SYNTHETIC DATASETS

When performing parametric G-computation with maximum-likelihood estimation, the choice of the number of bootstrap resamples is important. Similarly, when performing Bayesian parametric G-computation, the number L of converged MCMC draws is important, as is the number $M \leq L$ of syntheses in MIM. Given recent advances in computing power, we encourage setting these values as large as possible, in order to maximize the precision and accuracy of the treatment effect estimator, and to minimize the Monte Carlo error in the estimate. A sensible strategy is to increase the number of bootstrap resamples/syntheses until repeated analyses across different random seeds give similar results, within a specified degree of accuracy.

In MIM, MCMC simulation is used in the synthesis stage. The value of M is likely to be a fraction of the total number of iterations/posterior samples required for convergence. As computation time is driven by the synthesis stage, increasing M provides more precise and efficient estimation [198, 204] of the treatment effect at little cost in the analysis stage. In the context of statistical disclosure limitation, it is not uncommon to set the number of syntheses as low as $M = 10$ [205]. This is because the original survey data may involve several hundreds of subjects and variables. Releasing a large number of syntheses with a large number of subjects may not be practical, placing undue demands on the analyst, e.g. in terms of storage costs and processing needs. MIM has been developed with much smaller numbers of subjects and covariates in mind, in a context in which the data synthesizer and analyst are the same entity.

For MIM, an inconvenience of the expressions in Equation 29 and Equation 32 is that these may produce negative variances. When the posterior in Equation 29 generates a negative value of σ^2_Δ, i.e., when $\frac{(M-1)b}{\chi^*} < v$ (where χ^* is the draw from the posterior in Equation 29), the variance of the posterior distribution in Equation 30 is negative. Similarly, Equation 32 produces a negative variance when $(1 + 1/M)b < \bar{v}$. This is because the formulations have been derived using method-of-moments approximations, where estimates are not necessarily constrained to fall in the parameter space. Negative variances are unlikely to occur if M and the size of the synthetic datasets are relatively large. This is due to lower variability in σ^2_Δ and $\hat{V}\left(\hat{\Delta}^{(2)}_{10}\right)$ [199]: $\bar{v}$ decreases with larger syntheses and b is less variable with larger M [198]. Reasonable values of M are likely to depend on the specific setting, e.g. the size of the AC trial and the properties of the outcome model type, and we discuss these in the context of a simulation study in subsection 5.1.4.

4.9 CONCLUDING REMARKS

In this chapter, I have developed several methods to marginalize the conditional covariate-adjusted treatment effect estimates produced by the conventional outcome regression. Firstly, I have proposed a marginalization method based on parametric G-computation or model-based standardization. In addition, I have introduced a novel general-purpose method based on the ideas underlying multiple imputation, which I have termed multiple imputation marginalization, and is applicable to a wide range of models, including parametric survival models.

Both parametric G-computation and multiple imputation marginalization can be viewed as extensions to the conventional STC, with all methods making use of the same outcome model (albeit, in the conventional STC, this is centered). The novel methodologies are outcome regression approaches, thereby capable of extrapolation, that target marginal treatment effects. They do so by separating the covariate adjustment regression model from the evaluation of the marginal treatment effect of interest. The conditional parameters of the regression are viewed as nuisance parameters, not directly relevant to the research question. The methods can be implemented in a Bayesian statistical framework, which explicitly accounts for relevant sources of uncertainty, allows for the incorporation of prior evidence (e.g. expert opinion), and naturally integrates the analysis into a probabilistic framework.

Finally, an advantage of the outcome modeling approaches proposed in this chapter is that they produce estimates for both conditional and marginal estimands, as the conditional estimates are standardized into marginals. On the other hand, propensity score weighting-based methods such as MAIC are restricted to marginal inference; the marginal estimates cannot be directly “expanded” into conditionals.

5

CHAPTER 5: MARGINALIZATION OF REGRESSION-ADJUSTED TREATMENT EFFECTS: A SIMULATION STUDY

In this chapter, I carry out a simulation study to benchmark the performance of the novel "marginalized" outcome regression methods I have proposed in Chapter 4. The simulations are useful to provide proof-of-principle for the new methods, ensuring they are viable in the scenarios for which they were designed. The simulation study is also useful for comparative evaluation with existing methods such as MAIC and STC. The simulations investigate settings with binary outcomes and continuous covariates, with the log-odds ratio as the measure of effect. Binary outcomes such as response to treatment or the occurrence of an adverse event are relatively common in applications of population-adjusted indirect comparisons, particularly in oncology technology appraisals.

Section 5.1 outlines the simulation study design and execution. Section 5.2 describes the results from the simulation study. I present an extended discussion of my findings in Section 5.3. Finally, Section 5.4 provides some brief concluding remarks. Part of the research in this chapter is condensed in the article "Parametric G-computation for Compatible Indirect Treatment Comparisons with Limited Individual Patient Data" (Remiro-Azócar et al., 2021), and in the working paper "Marginalization of Regression-Adjusted Treatment Effects in Indirect Comparisons with Limited Patient-Level Data" (Remiro-Azócar et al., 2021).[1]

5.1 SIMULATION STUDY DESIGN

5.1.1 *Aims*

The simulation study in Chapter 3 evaluated the performance of existing approaches to population-adjusted indirect comparisons. The objectives of the simulation study in this chapter are to evaluate the statistical properties of the novel approaches to outcome regression and to compare the performance of these methods with that of the existing approaches. A range of scenarios, that may be encountered in practice, is considered.

1 The former has been submitted to Research Synthesis Methods and is available at: https://arxiv.org/abs/2108.12208. The latter is available at: https://arxiv.org/abs/2008.05951

We evaluate each estimator on the basis of the following finite-sample frequentist characteristics [102]: (1) unbiasedness; (2) variance unbiasedness; (3) randomization validity;[2] and (4) precision. The selected performance measures assess these criteria specifically (see 5.1.5). The simulation study is reported following the ADEMP (Aims, Data-generating mechanisms, Estimands, Methods, Performance measures) structure [102]. All simulations and analyses were performed using R software version 3.6.3 [103]. The design of the simulation study is similar to that presented in Chapter 3, but features binary outcomes instead of survival outcomes, with a logistic regression outcome model as opposed to a Cox model, and a different treatment allocation ratio in the trials.[3] Example R code implementing the methods on a simulated example is provided in Supplementary Appendix F.

5.1.2 *Data-generating mechanisms*

We consider binary outcomes using the log-odds ratio as the measure of effect. The binary outcome may be response to treatment or the occurrence of an adverse event.

For trials AC and BC, outcome y_n for subject n is simulated from a Bernoulli distribution with probabilities of success generated from logistic regression, such that:

$$\text{logit}[p(y_n \mid \boldsymbol{x}_n, z_n)] = \beta_0 + \boldsymbol{x}_n\boldsymbol{\beta_1} + (\beta_z + \boldsymbol{x}_n^{(EM)}\boldsymbol{\beta_2})\mathbb{1}(z_n = 1),$$

using the notation of the AC trial data. Four correlated continuous covariates $\boldsymbol{x}_n$ are generated per subject by simulating from a multivariate normal distribution with pre-specified variable means and covariance matrix [206]. Two of the covariates are purely prognostic variables; the other two ($\boldsymbol{x}_n^{(EM)}$) are effect modifiers, modifying the effect of both treatments A and B versus C on the log-odds ratio scale, and prognostic variables. The strength of the association between the prognostic variables and the outcome is set to $\beta_{1,k} = -\ln(0.5)$, where k indexes a given covariate. This regression coefficient fixes the conditional odds ratio for the effect of each prognostic variable on the odds of outcome at 2, indicating a strong prognostic effect. The strength of interaction of the effect modifiers is set to $\beta_{2,k} = -\ln(0.67)$, where k indexes a given effect modifier. This fixes the conditional odds ratio for the interaction effect on the odds of the outcome at approximately 1.5. Both active treatments have the same effect modifiers with respect to the common comparator and identical interaction coefficients for each. Therefore, the shared effect modifier assumption [18] holds in the simulation study by design. Pairwise Pearson correlation coefficients between the covariates are set to 0.2, indicating a moderate level of positive correlation.

The binary outcome represents the occurrence of an adverse event. Each active intervention has a very strong conditional treatment effect $\beta_z = \ln(0.17)$ at baseline (when the effect modifiers are zero) versus the common comparator. Such relative effect is associated with

2 In a sufficiently large number of repetitions, $(100 \times (1-\alpha))$% interval estimates based on normal distributions should contain the true value $(100 \times (1-\alpha))$% of the time, for a nominal significance level α.

3 The files required to run the simulations are available at http://github.com/remiroazocar/marginalized_indirect_comparisons_simstudy.

a "major" reduction of serious adverse events in a classification of extent categories by the German national HTA agency [207]. The covariates may represent comorbidities, which are associated with greater rates of the adverse event and, in the case of the effect modifiers, which interact with treatment to render it less effective. The intercept $\beta_0 = -0.62$ is set to fix the baseline event percentage at 35% (under treatment C, when the values of the covariates are zero).

The number of subjects in the BC trial is 600, under a 2:1 active treatment vs. control allocation ratio. For the BC trial, the individual-level covariates and outcomes are aggregated to obtain summaries. The continuous covariates are summarized as means and standard deviations, which would be available to the analyst in the published study in a table of baseline characteristics ("Table 1" of the RCT publication). The binary outcomes are summarized as overall event counts, e.g. from the cells of a 2×2 contingency table. Typically, the published study only provides this aggregate information to the analyst.

The simulation study investigates two factors in an arrangement with nine scenarios, thus exploring the interaction between these factors. We have selected the factors because they had the largest perceived influence on the performance metrics of the simulation study in Chapter 3. The simulation scenarios are defined by the values of the following parameters:

- The number of subjects in the AC trial, $N \in \{200, 400, 600\}$ under a 2:1 active intervention vs. control allocation ratio. The sample sizes correspond to typical values for a Phase III RCT [109] and for trials included in applications of MAIC submitted to HTA authorities [30].

- The degree of covariate imbalance. For both trials, each covariate k follows a normal marginal distribution with mean μ_k and standard deviation σ_k, such that $x_{i,k} \sim \text{Normal}(\mu_k, \sigma_k^2)$ for subject i. For the BC trial, we fix $\mu_k = 0.6$. For the AC trial, we vary the means of the marginal normal distributions such that $\mu_k \in \{0.45, 0.3, 0.15\}$. The standard deviation of each marginal distribution is fixed at $\sigma_k = 0.4$ for both trials. This setup corresponds to standardized differences [208] or Cohen effect size indices [209] (the difference in means in units of the pooled standard deviation) of 0.375, 0.75 and 1.125, respectively. This yields strong, moderate and poor covariate overlap; with overlap between the univariate marginal distributions of 85%, 71% and 57%, respectively, with $N = 600$. To compute the overlap percentages, we have followed a derivation by Cohen [209] for normally-distributed populations with equal size and equal variance. Note that the percentage overlap between the multivariate joint covariate distributions of each study is substantially lower. The strong, moderate and poor covariate overlap scenarios correspond to average percentage reductions in effective sample size of 22%, 60% and 85%, respectively. These percentage reductions are representative of the range encountered in NICE technology appraisals [30, 91], as discussed in subsection 3.1.3.

5.1.3 *Estimands*

The estimand of interest is the marginal log-odds ratio for A vs. B in the BC population. The treatment coefficient $\beta_z = \ln(0.17)$ is the same for both A vs. C and B vs. C, and the shared effect modifier assumption holds in the simulation study. Subtracting the treatment coefficient for A vs. C by that for B vs. C yields a true conditional treatment effect of zero for A vs. B in the BC population. Therefore, the true conditional treatment effect for A vs. B in the BC population is zero. As the true subject-level conditional effects are zero for all units, the true marginal log-odds ratio in the BC population is zero ($\Delta_{12}^{(2)} = 0$). This implies a null hypothesis-like simulation setup of no treatment effect for A vs. B, and marginal and conditional estimands in the BC population coincide by design.

Note that the true marginal effect for A vs. B in the BC population is a composite of that for A vs. C and that for B vs. C, both of which are non-null. These are the same and cancel out. For reference, the true marginal log-odds ratio in the BC population for the active treatments vs. the common comparator ($\Delta_{10}^{(2)}$ and $\Delta_{20}^{(2)}$) is computed as -1.15. This has been calculated by simulating two potential cohorts of 500,000 subjects, with the BC covariate distribution and the outcome-generating mechanism in subsection 5.1.2. One cohort is under the active treatment and the other is under the common comparator. The number of simulated subjects is sufficiently large to minimize sampling variability. The two cohorts are concatenated and a simple logistic regression is fitted, regressing the simulated binary outcomes on an indicator variable for treatment assignment. The treatment coefficient estimates the average difference in the potential outcomes on the log-odds ratio scale, and serves as the log of the true marginal odds ratio for the two interventions under consideration. Due to the non-collapsibility of the odds ratio and as per subsection 3.1.3, this simulation-based approach is necessary to determine the true marginal effect for A vs. C and B vs. C.

All methods compared in the simulation study perform the same unadjusted analysis (i.e., a simple regression of outcome on treatment) to estimate the marginal treatment effect of B versus C. Because the BC study is a relatively large RCT, this comparison should be unbiased with respect to the true marginal log-odds ratio in BC. Therefore, any bias in the A vs. B comparison should arise from bias in the A vs. C comparison, for which marginal and conditional relative treatment effects are non-null.

5.1.4 *Methods*

5.1.4.1 *Matching-adjusted indirect comparison*

Matching-adjusted indirect comparison (MAIC) is implemented using the original method of moments formulation presented by Signorovitch et al. [10, 18, 71, 91]. To avoid further reductions in effective sample size and precision, only the effect modifiers are included in the weighting model. A weighted logistic regression is fitted to the AC IPD and standard errors for the A vs. C marginal treatment effect are computed by resampling via the ordinary non-

parametric bootstrap with replacement [74, 75], with 1,000 resamples of each simulated dataset. The average marginal log-odds ratio for A vs. C is calculated as the mean across the 1,000 bootstrap resamples. Its corresponding standard error is the sample standard deviation across the resamples. Note that the standard version of MAIC [10, 18, 71, 91] uses a robust sandwich estimator for variance estimation [68, 210, 211] that accounts for the heteroskedasticity or correlation induced by the weighting. Nevertheless, this has understated variability under small effective sample sizes in the simulation study of Chapter 3, and most software implementations of the estimator treat the weights as fixed quantities. The bootstrap approach should account for the uncertainty in estimating the weights from the data.

In our implementation of MAIC, we only balance the covariate means and balance these for active treatment and control arms combined. Other approaches have been proposed, such as balancing the covariates separately for active treatment and common comparator arms [25, 28], or balancing terms of higher order than means, e.g. by including squared covariates in the weight estimation to balance variances. The former approach is discouraged because it may break randomization in the IPD, distorting the balance between treatment arms A and C on covariates that are not accounted for in the weighting, and potentially compromising the internal validity of the within-study estimate. The latter approach may increase finite-sample bias [69] and has performed poorly in recent simulation studies, in terms of both bias and precision, where covariate variances differ across studies [25, 27, 34, 164].

Given the often arbitrary factors driving selection into different trials, the data-generating mechanism in subsection 5.1.2 does not specify a trial assignment model. As per subsection 3.1.4, the logistic regression model for estimating the weights is the best-case model because it selects the right subset of covariates as effect modifiers and the balancing property holds for the weights with respect to the effect modifier means.

In a test simulation scenario with $N = 200$, bootstrapped MAIC has a running time of approximately 2.7 seconds per simulated dataset, using an Intel Core i7-8650 CPU (1.90 GHz) processor. Computation time increases linearly with the number of bootstrap resamples.

5.1.4.2 *Conventional simulated treatment comparison*

The conventional version of simulated treatment comparison (STC), as described by HTA guidance and recommendations [18], is implemented. A covariate-adjusted logistic regression is fitted to the IPD using maximum-likelihood estimation. The outcome regression is correctly specified. All covariates are accounted for in the regression but only the treatment effect modifiers are centered at their mean BC values, and interaction terms are only included for the effect modifiers. The log-odds ratio estimate for A vs. C is the treatment coefficient of the centered multivariable regression, with its standard error quantifying the standard deviation of the treatment effect.

In a test simulation scenario with $N = 200$, the conventional STC has a running time of 0.02 seconds per simulated dataset.

5.1.4.3 *Maximum-likelihood parametric G-computation*

We consider two implementations of parametric G-computation. In the first implementation, we use maximum-likelihood estimation to fit the multivariable outcome regression. The Q-model is correctly specified. We construct the joint distribution of the four BC covariates by simulating these from a multivariate Gaussian copula. This uses normally-distributed marginals with the BC means and standard deviations, and the pairwise linear correlations of the AC IPD. $N^* = 1000$ subject profiles are simulated for the BC pseudo-population, a value high enough to minimize sampling variability and provide an adequate degree of precision. Outcomes in the BC population are predicted by plugging the simulated covariates into the maximum-likelihood fit. The procedure is resampled using the ordinary non-parametric bootstrap with replacement, with 1,000 resamples of each simulated dataset. Increasing further the number of resamples produces minimal gains in estimation precision and accuracy, with the Monte Carlo error across different random seeds remaining relatively insensitive to these increases. The average marginal log-odds ratio for A vs. C is calculated as the mean across the 1,000 bootstrap resamples. Its corresponding standard error is the sample standard deviation across the resamples.

In a test simulation scenario with $N = 200$, parametric G-computation using maximum-likelihood estimation has a running time of approximately 3.5 seconds per replicate. Computation time increases linearly with the number of bootstrap resamples.

5.1.4.4 *Bayesian parametric G-computation*

In the second implementation of parametric G-computation, we use MCMC simulation to fit the outcome regression. This is implemented using the package `rstanarm` [212], a high-level appendage to the `rstan` package [213], the `R` interface for `Stan` [185]. Again, the Q-model is correctly specified. The joint distribution of the BC covariates is constructed by simulating $N^* = 1000$ subjects from a multivariate Gaussian copula, with normally-distributed marginals with the BC means and standard deviations, and the pairwise linear correlations of the AC IPD. Predicted outcomes for the simulated covariates are drawn from their posterior predictive distribution.

We use the default independent "weakly informative" priors for the logistic regression intercept and predictor coefficients, i.e., the likelihood dominates under a reasonably large amount of data and the prior strongly influences the posterior if the data are weak [214]. These are normally-distributed priors centered at mean 0. The scale of the normal prior distribution for the intercept is 1. The scale parameter of the normal priors for the other coefficients is 2.5, rescaled in terms of the standard deviation of the predictor in question. This places most of the prior mass in the range of plausible effects, discarding coefficient values that are implausibly strong, e.g. log-odds ratios over 3 (corresponding, approximately, to odds ratios over 20). This provides some regularization and helps stabilize computation. Alternative prior specifications are considered to check that we are not incorporating any unintended information into the models through the priors. Results are robust to the definitions of the prior distributions.

We run two Markov chains with 4,000 total iterations per chain. These include 2,000 warmup/burn-in iterations for each chain that are not used for posterior inference. This gives a total of 4,000 iterations for performing the analysis. Approximate mixing of the chains was attained, with all within-chain relative to between-chain statistics (R-hat) below 1.1 [215]. Satisfactory convergence was confirmed by the inspection of trace plots and the assessment of diagnostics such as the effective sample size and the Gelman-Rubin convergence diagnostic (potential scale reduction factor) [215]. The average marginal treatment effect for A vs. C is estimated taking the sample mean of the marginal log-odds ratio across the 4,000 MCMC iterations. The corresponding standard error is estimated using the sample standard deviation of the posterior draws of the marginal log-odds ratio.

In a test simulation scenario with $N = 200$, Bayesian parametric G-computation has a running time of approximately 4.2 seconds per replicate. Computation time increases linearly with the total number of MCMC iterations.

5.1.4.5 *Multiple imputation marginalization*

In the synthesis stage (first-stage regression and outcome prediction), we follow the MCMC procedure outlined for Bayesian G-computation to generate the syntheses, using identical prior specifications. We run two MCMC chains using `rstanarm` with 4,000 iterations per chain, where the burn-in is of 2,000 iterations. The convergence and mixing of the chains are satisfactory. The first-stage logistic regression is correctly specified. The joint distribution of the BC covariates is constructed by simulating from a multivariate Gaussian copula, with normally-distributed marginals with the BC means and standard deviations, and the pairwise linear correlations of the AC IPD. Predicted outcomes for the simulated covariates are drawn from their posterior predictive distribution as for Bayesian G-computation.

The MCMC chains are thinned every 4 iterations to use $M = 4000/4 = 1000$ syntheses in the analysis stage. Each synthesis is of size $N^* = 1000$, while keeping the same treatment allocation ratio of the original AC trial. We consider the selected value of M to provide an adequate degree of precision. In a test simulation scenario ($N = 200$), $M = 1000$ is high enough to minimize the Monte Carlo noise in the treatment effect estimate, such that the Monte Carlo error across different random seeds is small with respect to the uncertainty in the estimator (estimates are approximately within 0.01 across seeds). We explore varying the value of N^* in Supplementary Appendix D. We consider a two-step approach to pooling and the indirect comparison. In this formulation, the combining rules in Equations 31 and 32 are used to pool the point estimates of the second stage regressions and to estimate the marginal log-odds ratio for A vs. C in the BC population. Under $M = 1000$, variance estimates for the marginal A vs. C treatment effect are never negative under any scenario.

In a test simulation scenario with $N = 200$, MIM has a running time of approximately 7.9 seconds per replicate. Assuming that the total number of MCMC iterations is fixed, computation time increases linearly with the number of syntheses M.

5.1.4.6 *Indirect treatment comparison*

For all methods, the marginal log-odds ratio for B vs. C is estimated directly from the event counts, and its standard error is computed using the delta method [216]. The marginal log-odds ratio estimate for A vs. B and its standard error are obtained by combining the within-study point estimates, as per Section 4.7 (using Equation 2 to compare point estimates and Equation 33 to sum the point estimates of the variance). Wald-type 95% interval estimates are constructed for the marginal A vs. B treatment effect using normal distributions.

In Bayesian G-computation, we have used a two-step approach for: (1) the population-adjusted analysis of the AC trial (estimation of the marginal effect for A vs. C); and (2) the indirect treatment comparison (estimation of the marginal effect for A vs. B). We also consider integrating the two in one stage, using MCMC sampling. In this case, for estimation of the marginal log-odds ratio for B vs. C, the true underlying event rates/proportions for the treatments are given non-informative Jeffreys Beta$(0.5, 0.5)$ priors. The number of events in each arm is sampled from two independent Binomial likelihoods, parametrized by the aforementioned event probabilities and the total number of subjects in each arm. Means and variances for the marginal A vs. B treatment effect are obtained empirically from the posterior samples, with interval estimates calculated from the quantiles of the posterior distribution.

Similarly, in MIM, we have used a two-step approach for: (1) pooling (estimation of the average marginal effect for A vs. C); and (2) the indirect treatment comparison (estimation of the marginal effect for A vs. B). We consider using posterior simulation (Equations 28-30) to pool the point estimates from the second-stage regressions, integrating the pooling and the indirect comparison within a single Bayesian computation module. In this case, we apply MCMC sampling using the aforementioned prior specifications for the event rates in the BC study.

While the Bayesian inferential frameworks might be convenient for parametric G-computation and for MIM in the context of probabilistic sensitivity analysis, the selected inferential framework has little bearing on computation time and on the results of this simulation study, both in terms of a single case study and of the long-run frequentist statistical properties of the methods. Integrating the indirect treatment comparison step within a Bayesian module leads to virtually identical performance measures than the two-step approaches. Therefore, results are not reported.

5.1.5 *Performance measures*

We generate and analyze 2,000 Monte Carlo replicates of trial data per simulation scenario. Recall that in our implementations of MAIC, G-computation (both versions) and MIM, a large number of bootstrap resamples, MCMC draws or syntheses are performed for each of the 2,000 replicates. For instance, the analysis for one simulation scenario using Bayesian G-computation contains 4,000 MCMC draws (after burn-in) times 2,000 simulation replicates, which equals a total of 8 million posterior draws. Based on the method and simulation scenario with the

highest long-run variability (the conventional STC with $N = 200$ and poor covariate overlap), we consider the degree of precision provided by the Monte Carlo standard errors under 2,000 replicates to be acceptable in relation to the size of the effects.[4]

We evaluate the performance of the outcome regression methods and MAIC on the basis of the following criteria: (1) bias; (2) variability ratio; (3) empirical coverage rate of the interval estimates; (4) empirical standard error (ESE); and (5) mean square error (MSE). These criteria are explicitly defined in Chapter 3, albeit note that there are 2,000 simulation replicates (not 1,000) in the current simulation study.

With respect to the simulation study aims in subsection 5.1.1, the bias in the estimated treatment effect assesses aim 1. This is equivalent to the average estimated treatment effect across simulations because the true treatment effect $\Delta_{12}^{(2)} = 0$. The variability ratio evaluates aim 2. This represents the ratio of the average model standard error and the sample standard deviation of the treatment effect estimates (the empirical standard error) [117]. Coverage targets aim 3, and is estimated as the proportion of simulated datasets for which the true treatment effect is contained within the nominal $(100 \times (1 - \alpha))\%$ interval estimate of the estimated treatment effect. In this simulation study, $\alpha = 0.05$ is the nominal significance level. The empirical standard error is the standard deviation of the treatment effect estimates across the 2,000 simulated datasets. Therefore, it measures precision or long-run variability, and evaluates aim 4. The mean square error is equivalent to the average of the squared bias plus the variance across the 2,000 simulated datasets. Therefore, it is a summary value of overall accuracy (efficiency), that accounts for both bias (aim 1) and variability (aim 4).

5.2 RESULTS OF THE SIMULATION STUDY

Performance metrics for all simulation scenarios are displayed in Figure 12, Figure 13 and Figure 14. Figure 12 displays the results for the three data-generating mechanisms under $N = 200$. Figure 13 presents the results for the three scenarios with $N = 400$. Figure 14 depicts the results for the three scenarios with $N = 600$. From top to bottom, each figure considers the scenario with strong overlap first, followed by the moderate and poor overlap scenarios. For each scenario, there is a box plot of the point estimates of the A vs. B marginal treatment effect across the 2,000 simulated datasets. Below, is a summary tabulation of the performance measures for each method. Each performance measure is followed by its Monte Carlo standard error, presented in parentheses, which quantifies the simulation uncertainty.

In the figures, ATE is the average marginal treatment effect estimate for A vs. B across the simulated datasets (this is equal to the bias as the true effect is zero). LCI is the average

4 Conservatively, we assume that $\text{SD}(\hat{\Delta}_{12}^{(2)}) \leq 1.71$ and that the variance across simulations of the estimated treatment effect is always less than 2.92. Given that the MCSE of the bias is equal to $\sqrt{\text{Var}(\hat{\Delta}_{12}^{(2)})/N_{sim}}$, where $N_{sim} = 2000$ is the number of simulations, it is at most 0.038 under 2,000 simulations. We consider the degree of precision provided by the MCSE of the bias to be acceptable in relation to the size of the effects. If the empirical coverage rate of the methods is 95%, $N_{sim} = 2000$ implies that the MCSE of the coverage is $\left(\sqrt{(95 \times 5)/2000}\right) = 0.49\%$, with the worst-case MCSE being 1.12% under 50% coverage. We also consider this degree of precision to be acceptable. Hence, the simulation study is conducted under $N_{sim} = 2000$.

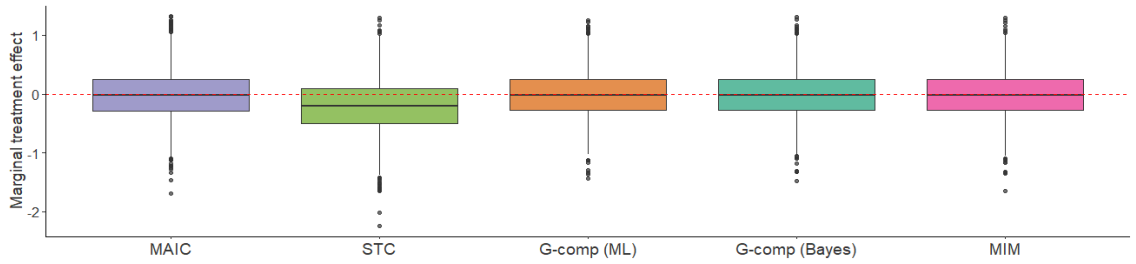
lower bound of the 95% interval estimate. UCI is the average upper bound of the 95% interval estimate. VR, ESE and MSE are the variability ratio, empirical standard error and mean square error, respectively. Cov is the empirical coverage rate of the 95% interval estimates. G-comp (ML) stands for the maximum-likelihood version of parametric G-computation and G-comp (Bayes) denotes its Bayesian counterpart using MCMC estimation.

In MIM, no simulation replicates produce negative variances and ad hoc truncation is not required. Weight estimation cannot be performed for 4 of the 18,000 replicates in MAIC, where there are no feasible weighting solutions. This issue occurs in the most extreme scenario, corresponding to $N = 200$ and poor covariate overlap. Feasible weighting solutions do not exist due to separation problems, i.e., there is a total lack of covariate overlap. Because MAIC is incapable of producing an estimate in these cases, the affected replicates are discarded altogether (the scenario in question analyzes 1,996 simulated datasets for MAIC). This phenomenon has also been observed in a recent simulation study [34], where MAIC cannot generate an estimate in scenarios with small sample sizes and poor overlap.

UNBIASEDNESS OF TREATMENT EFFECT ESTIMATES The impact of the bias largely depends on the uncertainty in the estimated treatment effect, quantified by the empirical standard error. We compute standardized biases (bias as a percentage of the empirical standard error). With $N = 200$, MAIC has standardized biases of magnitude 11.3% and 16.1% under moderate and poor covariate overlap, respectively. Otherwise, the magnitude of the standardized bias is below 10%. Similarly, under $N = 200$, the maximum-likelihood version of parametric G-computation has standardized biases of magnitude 13.3% and 24.8% in the scenarios with moderate and poor overlap, respectively. In all other scenarios, the standardized bias has magnitude below 10%. For Bayesian parametric G-computation, standardized biases never have a magnitude above 10% and troublesome biases are not produced in any of the simulation scenarios. The maximum absolute value of the standardized bias is 9.7% in a scenario with $N = 200$ and moderate covariate overlap. In MIM, no standardized biases are larger than 10% in either direction, and the maximum absolute value is 9.4% in a simulation scenario with $N = 200$ and moderate overlap.

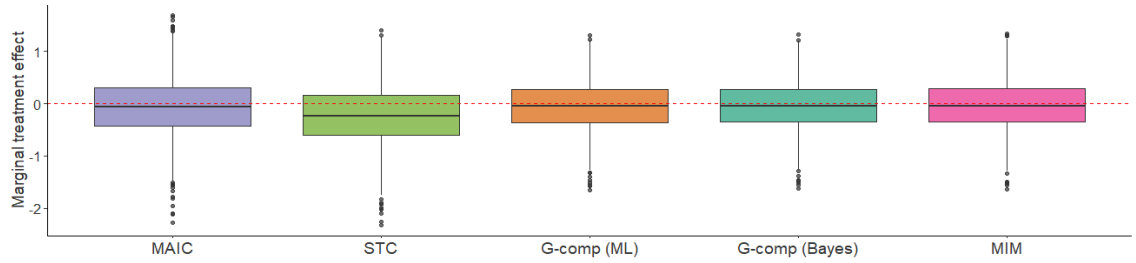
To evaluate whether the bias in MAIC and parametric G-computation has any practical significance, we investigate whether the coverage rates are degraded by it. Coverage is not affected for maximum-likelihood parametric G-computation, where empirical coverage rates for all simulation scenarios are very close to the nominal coverage rate, 0.95 for 95% interval estimates. In the case of MAIC, there is discernible undercoverage in the scenario with $N = 200$ and poor covariate overlap (empirical coverage rate of 0.916). This is the scenario with the lowest effective sample size after weighting. Hence, the results are probably related to small-sample bias [217] in the weighted logistic regression [138]. This bias for MAIC was not observed in the more extreme scenarios of the simulation study in Chapter 3, which considered survival outcomes and the Cox proportional hazards regression as the outcome model. In absolute terms, the bias of MAIC is greater than that of MIM and both versions of parametric G-computation where the number of patients in the AC trial is small ($N = 200$) or covariate

overlap is poor. In fact, when both of these apply, the bias of MAIC is important (-0.144). Otherwise, with $N = 400$ or greater, and moderate or strong overlap, the aforementioned methods produce similarly low levels of bias.



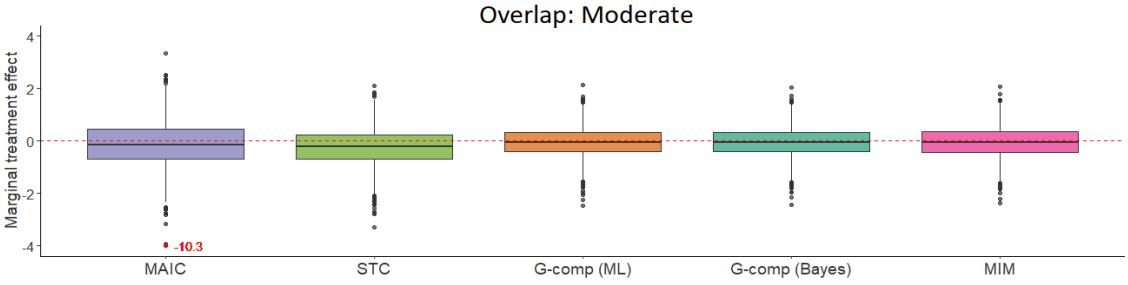
Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.020 (0.009)	-0.819 (0.009)	0.779 (0.009)	1.011 (0.016)	0.950 (0.005)	0.404 (0.006)	0.163 (0.006)
STC	-0.206 (0.010)	-1.086 (0.011)	0.674 (0.009)	0.994 (0.016)	0.940 (0.005)	0.452 (0.007)	0.246 (0.009)
G-comp (ML)	-0.015 (0.008)	-0.772 (0.009)	0.742 (0.008)	1.020 (0.016)	0.958 (0.005)	0.379 (0.006)	0.143 (0.005)
G-comp (Bayes)	-0.015 (0.008)	-0.762 (0.009)	0.733 (0.008)	1.010 (0.016)	0.954 (0.005)	0.378 (0.006)	0.143 (0.005)
MIM	-0.018 (0.009)	-0.739 (0.009)	0.703 (0.008)	0.960 (0.015)	0.950 (0.005)	0.383 (0.006)	0.147 (0.005)

Overlap: Strong



Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.061 (0.012)	-1.087 (0.013)	0.965 (0.012)	0.967 (0.015)	0.938 (0.005)	0.541 (0.009)	0.297 (0.010)
STC	-0.241 (0.012)	-1.287 (0.014)	0.805 (0.011)	0.956 (0.015)	0.938 (0.005)	0.558 (0.009)	0.370 (0.012)
G-comp (ML)	-0.042 (0.010)	-0.929 (0.011)	0.844 (0.010)	0.985 (0.016)	0.946 (0.005)	0.459 (0.007)	0.212 (0.007)
G-comp (Bayes)	-0.044 (0.010)	-0.906 (0.010)	0.818 (0.010)	0.964 (0.015)	0.942 (0.005)	0.456 (0.007)	0.210 (0.007)
MIM	-0.043 (0.010)	-0.881 (0.011)	0.795 (0.010)	0.930 (0.015)	0.936 (0.005)	0.460 (0.007)	0.213 (0.007)

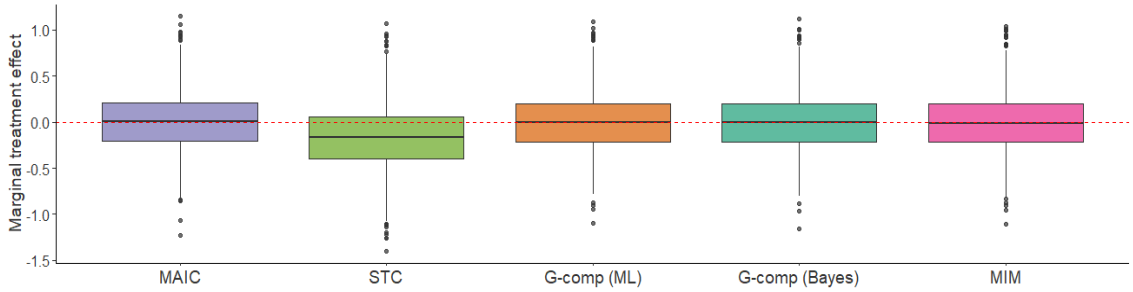
Overlap: Moderate



Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.144 (0.020)	-2.116 (0.312)	1.827 (0.302)	1.122 (0.175)	0.916 (0.006)	0.896 (0.014)	0.824 (0.059)
STC	-0.235 (0.016)	-1.546 (0.017)	1.076 (0.015)	0.940 (0.015)	0.933 (0.006)	0.712 (0.011)	0.562 (0.020)
G-comp (ML)	-0.044 (0.013)	-1.162 (0.013)	1.073 (0.013)	0.981 (0.016)	0.942 (0.005)	0.581 (0.009)	0.340 (0.011)
G-comp (Bayes)	-0.045 (0.013)	-1.109 (0.013)	1.019 (0.013)	0.945 (0.015)	0.930 (0.006)	0.574 (0.009)	0.332 (0.011)
MIM	-0.049 (0.013)	-1.094 (0.013)	0.997 (0.013)	0.921 (0.015)	0.923 (0.006)	0.579 (0.009)	0.338 (0.011)

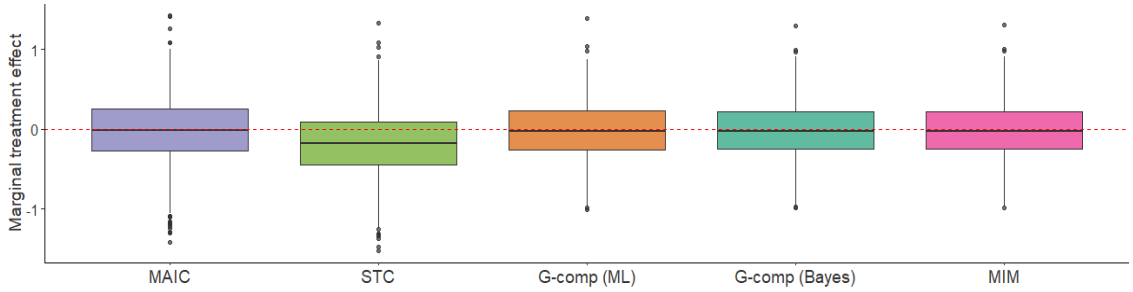
Overlap: Poor

Figure 12: Point estimates and performance metrics across all methods for each simulation scenario with $N = 200$. The model standard error for the MAIC outlier in the poor overlap scenario has an inordinate influence on the variability ratio; removing it reduces the variability ratio to 0.980 (0.019).



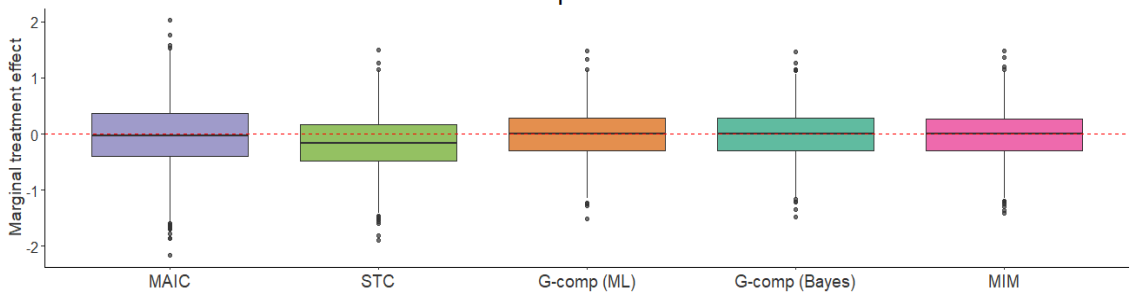
Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	0.004 (0.007)	-0.611 (0.007)	0.619 (0.007)	0.996 (0.016)	0.952 (0.005)	0.315 (0.005)	0.099 (0.003)
STC	-0.164 (0.008)	-0.824 (0.008)	0.496 (0.007)	0.971 (0.015)	0.918 (0.006)	0.347 (0.005)	0.147 (0.005)
G-comp (ML)	-0.003 (0.007)	-0.591 (0.007)	0.584 (0.007)	0.982 (0.016)	0.946 (0.005)	0.305 (0.005)	0.093 (0.003)
G-comp (Bayes)	-0.003 (0.007)	-0.607 (0.007)	0.601 (0.007)	1.013 (0.016)	0.952 (0.005)	0.304 (0.005)	0.093 (0.003)
MIM	-0.006 (0.007)	-0.577 (0.007)	0.565 (0.007)	0.943 (0.015)	0.936 (0.005)	0.309 (0.005)	0.095 (0.003)

Overlap: Strong



Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.014 (0.009)	-0.778 (0.009)	0.751 (0.009)	0.980 (0.016)	0.942 (0.005)	0.398 (0.006)	0.159 (0.005)
STC	-0.181 (0.009)	-0.945 (0.009)	0.583 (0.009)	0.975 (0.015)	0.926 (0.006)	0.400 (0.006)	0.193 (0.006)
G-comp (ML)	-0.017 (0.008)	-0.684 (0.008)	0.650 (0.008)	0.994 (0.016)	0.947 (0.005)	0.342 (0.005)	0.117 (0.004)
G-comp (Bayes)	-0.018 (0.008)	-0.695 (0.008)	0.659 (0.008)	1.013 (0.016)	0.954 (0.005)	0.341 (0.005)	0.117 (0.004)
MIM	-0.020 (0.008)	-0.667 (0.008)	0.627 (0.008)	0.954 (0.015)	0.930 (0.006)	0.346 (0.005)	0.120 (0.004)

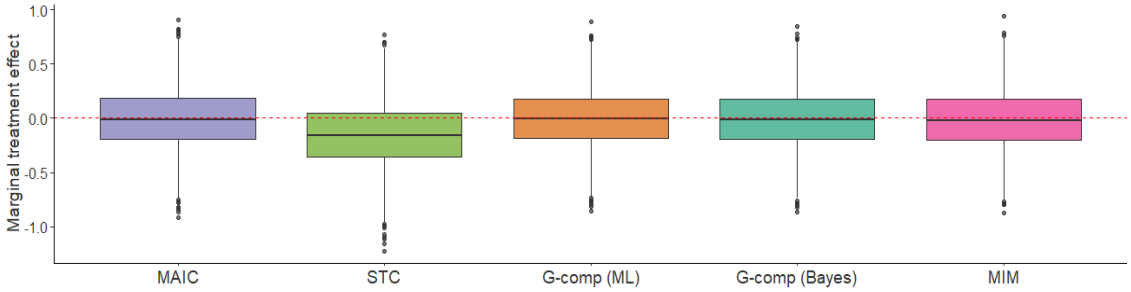
Overlap: Moderate



Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.030 (0.013)	-1.153 (0.013)	1.092 (0.013)	0.985 (0.016)	0.930 (0.006)	0.582 (0.009)	0.339 (0.011)
STC	-0.166 (0.011)	-1.098 (0.011)	0.765 (0.010)	1.001 (0.016)	0.942 (0.005)	0.475 (0.008)	0.253 (0.008)
G-comp (ML)	-0.010 (0.009)	-0.821 (0.009)	0.800 (0.009)	1.015 (0.016)	0.959 (0.004)	0.408 (0.006)	0.166 (0.005)
G-comp (Bayes)	-0.012 (0.009)	-0.820 (0.009)	0.797 (0.009)	1.019 (0.016)	0.959 (0.004)	0.405 (0.006)	0.164 (0.005)
MIM	-0.015 (0.009)	-0.799 (0.009)	0.768 (0.009)	0.982 (0.016)	0.950 (0.005)	0.407 (0.006)	0.166 (0.005)

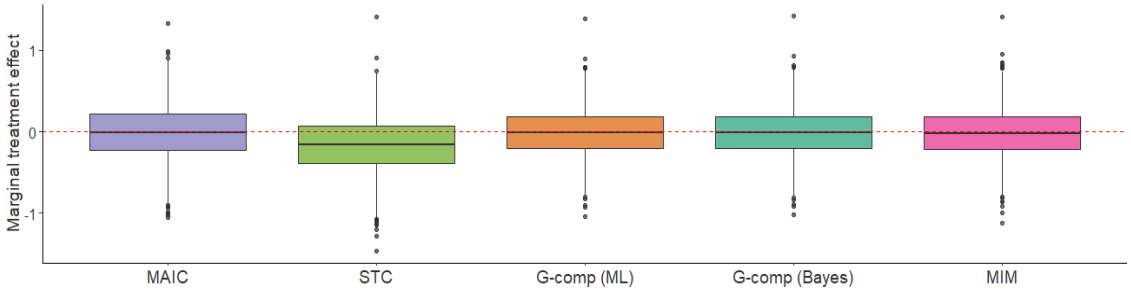
Overlap: Poor

Figure 13: Point estimates and performance metrics across all methods for each simulation scenario with $N = 400$.



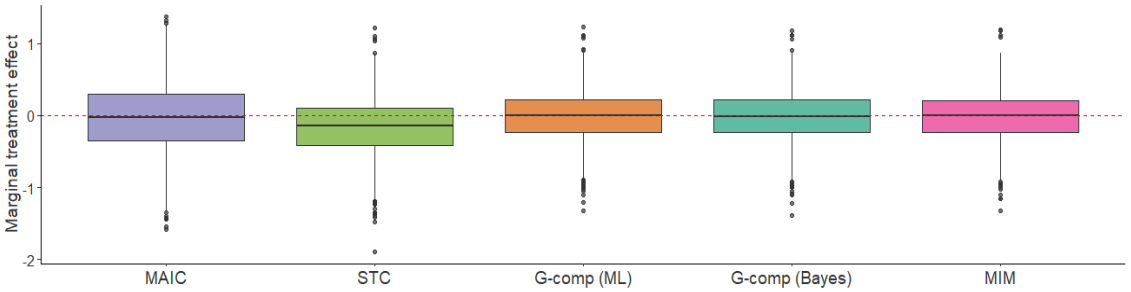
Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.006 (0.006)	-0.549 (0.006)	0.537 (0.006)	0.981 (0.016)	0.942 (0.005)	0.283 (0.004)	0.080 (0.003)
STC	-0.163 (0.007)	-0.739 (0.007)	0.413 (0.007)	0.980 (0.016)	0.909 (0.006)	0.300 (0.005)	0.117 (0.004)
G-comp (ML)	-0.011 (0.006)	-0.533 (0.006)	0.512 (0.006)	0.990 (0.016)	0.950 (0.005)	0.269 (0.004)	0.073 (0.002)
G-comp (Bayes)	-0.011 (0.006)	-0.558 (0.006)	0.536 (0.006)	1.035 (0.016)	0.960 (0.004)	0.269 (0.004)	0.073 (0.002)
MIM	-0.016 (0.006)	-0.526 (0.006)	0.494 (0.006)	0.950 (0.015)	0.941 (0.005)	0.274 (0.004)	0.075 (0.002)

Overlap: Strong



Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.004 (0.007)	-0.660 (0.007)	0.652 (0.007)	1.018 (0.016)	0.954 (0.005)	0.329 (0.005)	0.108 (0.003)
STC	-0.163 (0.007)	-0.816 (0.008)	0.489 (0.007)	1.004 (0.016)	0.932 (0.006)	0.332 (0.005)	0.137 (0.004)
G-comp (ML)	-0.008 (0.007)	-0.588 (0.006)	0.571 (0.007)	1.013 (0.016)	0.956 (0.005)	0.292 (0.005)	0.085 (0.003)
G-comp (Bayes)	-0.009 (0.007)	-0.608 (0.006)	0.590 (0.007)	1.052 (0.017)	0.962 (0.004)	0.291 (0.005)	0.085 (0.003)
MIM	-0.013 (0.007)	-0.579 (0.007)	0.553 (0.007)	0.972 (0.015)	0.946 (0.005)	0.297 (0.005)	0.088 (0.003)

Overlap: Moderate



Method	ATE	LCI	UCI	VR	Cov	ESE	MSE
MAIC	-0.027 (0.011)	-0.973 (0.011)	0.918 (0.011)	1.008 (0.016)	0.945 (0.005)	0.479 (0.008)	0.230 (0.007)
STC	-0.164 (0.009)	-0.946 (0.009)	0.619 (0.009)	1.018 (0.016)	0.942 (0.005)	0.392 (0.006)	0.181 (0.006)
G-comp (ML)	-0.012 (0.008)	-0.700 (0.008)	0.677 (0.008)	1.032 (0.016)	0.957 (0.005)	0.340 (0.005)	0.116 (0.004)
G-comp (Bayes)	-0.013 (0.008)	-0.712 (0.008)	0.687 (0.008)	1.050 (0.017)	0.961 (0.004)	0.340 (0.005)	0.116 (0.004)
MIM	-0.014 (0.008)	-0.685 (0.008)	0.657 (0.008)	0.998 (0.016)	0.952 (0.005)	0.343 (0.005)	0.118 (0.004)

Overlap: Poor

Figure 14: Point estimates and performance metrics across all methods for each simulation scenario with $N = 600$.

STC generates problematic negative biases in all nine scenarios considered in this simulation study, with a standardized bias of magnitude greater than 50% in all cases. STC consistently produces the highest bias of all methods, and the magnitude of this bias appears to increase under the smallest sample size ($N = 200$). Recall that all methods perform the same unadjusted analysis for the B versus C comparison, which should be unbiased with respect to the true marginal log-odds ratio in BC (-1.15). The systematic bias in STC is due to the divergence of the conditional estimates produced for A versus C from the corresponding marginal estimand that should be targeted. This is a result of the non-collapsibility of the (log) odds ratio.

UNBIASEDNESS OF VARIANCE ESTIMATES In MAIC, the variability ratio of treatment effect estimates is close to one under all simulation scenarios except one. That is the scenario with $N = 200$ and poor covariate overlap, where the variability ratio is 1.122. This high value is attributed to the undue influence of an outlier (as seen in the box plot of point estimates) on the average model standard error. Once the outlier is removed, the variability ratio decreases to 0.98, just outside from being within Monte Carlo error of one but not statistically significantly different. This suggests very little bias in the standard error estimates in this scenario, i.e., that the model standard errors tend to coincide with the empirical standard error. In the simulation study of Chapter 3, robust sandwich standard errors underestimated the variability of estimates in MAIC under small sample sizes and poor covariate overlap. The non-parametric bootstrap seems to provide more conservative variance estimation in these extreme settings.

In STC, variability ratios are generally close to one with $N = 400$ and $N = 600$. Any bias in the estimated variances appears to be negligible, although there is a slight decrease in the variability ratios when the AC sample size is small ($N = 200$). Recall that this metric assumes that the correct estimand and corresponding variance are being targeted. However, in our application of STC, both model standard errors and empirical standard errors are taken over an incompatible indirect treatment comparison.

In maximum-likelihood parametric G-computation, variability ratios are generally very close to one. In Bayesian parametric G-computation, variability ratios are generally close to one but are slightly above it in some scenarios with $N = 600$ (1.05 and 1.052 with moderate and poor covariate overlap, respectively). This suggests some overestimation of the empirical standard error by the model standard errors. On the other hand, MIM displays some underestimation of variability by the model standard errors. This is more pronounced under the smallest sample size, with variability ratios of 0.93 and 0.921 for moderate and poor overlap, respectively. The underestimation is likely due to the normality assumptions used to derive the model standard errors — the posterior distribution of $\Delta_{10}^{(2)}$ is assumed to be normal in the derivation of the combining rules, and low sample sizes may break the normality assumption.

RANDOMIZATION VALIDITY The empirical coverage rate should be approximately equal to the nominal coverage rate, in this case 0.95 for 95% interval estimates, to obtain appropriate type I error rates for testing a “no effect” null hypothesis. Theoretically, the empirical coverage rate is statistically significantly different to 0.95 if, roughly, it is less than 0.94 or more than 0.96,

assuming 2,000 independent simulations per scenario. These values differ by approximately two standard errors from the nominal coverage rate. Poor coverage rates are a decomposition of both the bias and the standard error used to compute the Wald-type interval estimates. In the simulation scenarios, none of the methods lead to overly conservative inferences but there are some issues with undercoverage.

Empirical coverage rates for MAIC are significantly different from the advertised nominal coverage rate in three scenarios. In the three, the coverage rate is below 0.94 (empirical coverage rates of 0.938, 0.93 and 0.916). The last two of these occur in scenarios with poor covariate overlap, with the latter corresponding to the smallest effective sample size after weighting ($N = 200$). This is the scenario which integrates the two most important determinants of small-sample bias, in which MAIC has exhibited discernible bias. In this case, undercoverage is bias-induced. On the other hand, in our previous simulation study in Chapter 3, undercoverage was induced by the robust sandwich variance underestimating standard errors.

In the conventional version of STC, coverage rates are degraded by the bias induced by the non-collapsibility of the log-odds ratio. Almost invariably, there is undercoverage. Interestingly, the empirical coverage does not markedly deteriorate — coverage percentages never fall below 90%, i.e., never at least double the nominal rate of error. In general, both versions of parametric G-computation exhibit appropriate coverage. Only one scenario provides rates below 0.94 (Bayesian G-computation with $N = 200$ and poor overlap, with an empirical coverage rate of 0.93). No scenarios have empirical coverage above 0.96. Coverage rates for the maximum-likelihood implementation are always appropriate, with most empirical coverage percentages within Monte Carlo error of 95%.

In MIM, coverage rates generally exhibit some underestimation of the advertised nominal coverage rate. Empirical coverage rates are significantly below the nominal rate in four scenarios (empirical coverage rates of 0.936, 0.936, 0.93 and 0.923). Again, the most inappropriate of these (0.923) occurs where there is poor covariate overlap and the AC sample size is low ($N = 200$). In this scenario, it is not bias that degrades the coverage rate for MIM. Poor coverage is induced by the standard errors used to construct the Wald-type interval estimates, which underestimate variability.

PRECISION AND EFFICIENCY MIM and both versions of parametric G-computation have reduced empirical standard errors compared to MAIC across all scenarios. Interestingly, conventional STC is even less precise than MAIC in most scenarios (all the scenarios with moderate or strong overlap, where reductions in effective sample size after weighting are tolerable). Several trends are revealed upon comparison of the ESEs, and upon visual inspection of the spread of the point estimates in the box plots. As expected, the ESE increases for all methods (i.e., estimates are less precise) as the number of subjects in the AC trial is lower. The decrease in precision is more substantial for MAIC than for the outcome regression methods.

The degree of covariate overlap has an important influence on the ESE and population adjustment methods incur losses of precision when covariate overlap is poor. Again, this loss of precision is more substantial for MAIC than for the outcome regression approaches. Where overlap is poor, there exists a subpopulation in BC that does not overlap with the AC population. Therefore, inferences in this subpopulation rely largely on extrapolation. Outcome regression approaches require greater extrapolation when the covariate overlap is weaker, thereby incurring a loss of precision.

Where covariate overlap is strong, MIM and both versions of parametric G-computation display very similar ESEs than MAIC. As mentioned earlier, conventional STC offers even lower precision than MAIC in these cases. To illustrate this, consider the scenario with $N = 200$ and moderate overlap, where MAIC is expected to have a low effective sample size after weighting and perform comparatively worse than outcome regression. Even in this scenario, MAIC appears to be more precise (empirical standard error of 0.541) than conventional STC (empirical standard error of 0.558). As overlap decreases, precision is reduced more markedly for MAIC compared to the outcome regression methods. Under poor overlap, MAIC considerably increases the ESE compared to the conventional STC.

In MAIC, extrapolation is not even possible. Where covariate overlap is poor, the observations in the AC IPD that are not covered by the ranges of the selected covariates in BC are assigned weights that are very close to zero. The relatively small number of individuals in the overlapping region of the covariate space are assigned inflated weights, dominating the reweighted sample. These extreme weights lead to large reductions in ESS and affect very negatively the precision of estimates.

Similar to the trends observed for the ESE, the MSE is also very sensitive to the value of N and to the level of covariate overlap. The MSE decreases for all methods as N and the level of overlap increase. The accuracy of MAIC and the marginalized outcome regression methods is comparable when the AC sample size is high or covariate overlap is strong. As the AC sample size and overlap decrease, the relative accuracy of MAIC with respect to MIM and both approaches to parametric G-computation is markedly reduced. Accuracy for the conventional version of STC is comparatively poor and this is driven by bias.

Where covariate overlap is strong or moderate, the marginalized outcome regression methods have the highest accuracy, followed by MAIC and STC. Where overlap is poor, the marginalized outcome regression methods are considerably more accurate than MAIC, with much smaller mean square errors. MAIC also provides less accurate estimates than STC in terms of mean square error. The variability of estimates under MAIC increases considerably in these scenarios. The precision is sufficiently poor to offset the loss of bias with respect to STC.

5.3 DISCUSSION OF SIMULATION STUDY RESULTS

5.3.1 *Summary of findings*

The marginalized outcome regression methods and MAIC can yield unbiased estimates of the marginal A vs. B treatment effect in the BC population. Conventional STC targets a conditional treatment effect for A vs. C that is incompatible in the indirect comparison. Bias is produced because the log-odds ratio is a non-collapsible measure of effect. Across all scenarios, MIM and both versions of parametric G-computation largely eliminate the bias induced by effect modifier imbalances. There is some negative bias in MAIC and the marginalized outcome regression methods where the sample size N is small. In the case of MAIC, this is problematic where covariate overlap is poor. The bias for MAIC was not observed in our previous simulation study in Chapter 3. In that study, the model of interest was a Cox proportional hazards regression with survival outcomes. The difference in results is likely due to logistic regression being more prone to small-sample or sparse data bias [217] than Cox regression [138].

As for precision, the marginalized outcome regression approaches have reduced variability compared to MAIC. The superior precision is demonstrated by their lower empirical standard errors across all scenarios. Because the methods are generally unbiased, precision is the driver of comparative accuracy. The simulation study confirms that, under correct model specification, parametric G-computation and MIM have lower mean square errors than weighting and are therefore more efficient. The differences in performance are exacerbated where covariate overlap is poor and sample sizes are low. In these cases, the effective sample size after weighting is small, and this leads to inflated variances and wider interval estimates for MAIC. Specific bias-variance trade-offs will depend on the outcome model of interest, e.g. the logistic regression with binary outcomes setup in this study is less efficient than the Cox regression with survival outcomes [137] explored in our previous simulation study (Chapter 3).

The performance measures for Bayesian G-computation and MIM are very similar, as they use the same MCMC procedure to fit the outcome model and to predict outcomes. In terms of bias, precision and efficiency, there is no particular reason to believe that the performance metrics for Bayesian G-computation or MIM are superior one to the other. In terms of variance estimation and coverage, the performance measures for both parametric G-computation approaches are superior to those of MIM. MIM exhibits undercoverage in some scenarios due to the model standard errors underestimating variability. Generally speaking, coverage rates for the interval estimates are more appropriate for the parametric G-computation methods than for MIM. Where the outcome regression is a generalized linear model, parametric G-computation is easier to implement and has lesser potential complications. In any case, MIM could be useful with different outcome model types. Further research on this method is required, and will focus on developing alternative variance estimators that avoid negative variances and are more conservative.

For the conventional STC, outcome regression may have decreased precision relative to MAIC, as dictated by the empirical standard errors. On the other hand, the marginalized

outcome regression methods are more precise than both MAIC and conventional STC. From a frequentist perspective, the standard error of the estimator of a conditional log-odds ratio for A vs. C, targeted by conventional STC, is larger than the standard error of a regression-adjusted estimate of the marginal log-odds ratio for A vs. C, produced by G-computation and MIM. This precision comparison likely lacks relevance, because one is comparing estimators that target different estimands. Nevertheless, it supports previous findings on non-collapsible measures of effect when adjusting for prognostic covariates [44, 50]. When we marginalize and compare estimators targeting like-for-like marginal estimands, we find that outcome regression is no longer detrimental for precision and efficiency compared to weighting.

In our previous simulation study in Chapter 3 [91], we evaluated MAIC using a robust sandwich variance estimator. This underestimated variability and produced narrow interval estimates under small effective sample sizes. In the simulation study in this chapter, the bootstrap procedure provides more conservative variance estimation compared to the sandwich estimator in the more extreme settings. This implies that the bootstrap approach should be preferred for statistical inference where there are violations of the overlap assumption and small sample sizes.

5.3.2 *Implications for practice*

EXTRAPOLATION CAPABILITIES AND PRECISION CONSIDERATIONS We expect the conclusions in the previous paragraphs to be robust. The number of simulated datasets per scenario (2,000) is large enough so that the outlined performance differences are not due to chance. Nevertheless, we now clarify some aspects of the conclusions that are more nuanced.

In real applications, the effective sample sizes and percentage reductions in effective sample size may be lower and higher, respectively, than those considered in this simulation study [30]. In these situations, covariate overlap is poor and this leads to a high loss of precision in MAIC. The marginalized outcome regression methods should be considered because they are substantially more statistically efficient. This is particularly the case where the outcome model is a logistic regression, more prone to small-sample bias [34, 138], imprecision [137], and inefficiency [137] than other models, e.g. the Cox regression. In addition, where sample sizes are small and the number of covariates is large, feasible weighting solutions may not exist for MAIC due to separation problems [35], as observed in one of the scenarios of this simulation study ($N = 200$ with poor overlap) and, notably, in another recent simulation study [34]. An advantage of outcome regression is that it can be applied in these settings. MAIC cannot extrapolate beyond the covariate space observed in the IPD. Therefore, it cannot overcome the failure of assumptions that is the lack of covariate overlap and is incapable of producing an estimate.

Moreover, we note that MAIC requires accounting for all effect modifiers (balanced and imbalanced), as excluding balanced covariates from the weighting procedure does not ensure balance after the weighting. On the other hand, outcome regression methods do not necessarily

require the inclusion of the effect modifiers that are in balance. This may mitigate losses of precision further, particularly where the number of potential effect modifiers is large.

With limited overlap, outcome regression methods can use the linearity assumption to extrapolate beyond the AC population, provided the true relationship between the covariates and the outcome is adequately captured. We view this as a desirable attribute because poor overlap, with small effective sample sizes and large percentage reductions in effective sample size, is a pervasive issue in health technology appraisals [30]. Nevertheless, where overlap is more considerable, one may wish to restrict inferences to the region of overlap and avoid relying on a model for extrapolation outside this region (reducing the dependence on assumptions that are inherently untestable) [218, 219].

Note that the model extrapolation uncertainty is not reflected in the interval estimates for the outcome regression approaches and that some consider weighting approaches to give a "more honest reflection of the overall uncertainty" [146]. The gain in efficiency produced by outcome regression must be counterbalanced against the potential for model misspecification bias. Weighting methods are often perceived to rely on less demanding parametric assumptions, yet model misspecification is an issue for both methods as we discuss later in this section.

It is worth noting that we have used the standard MAIC formulation proposed by Signorovitch et al. [10, 18, 71, 91] and that our conclusions are based on this approach. Nevertheless, MAIC is a rapidly developing methodology with novel implementations. An alternative formulation based on entropy balancing has been recently presented [25, 28, 70, 71]. This approach is similar to the original version with a subtle modification to the weight estimation procedure. While it has some interesting computational properties, Phillippo et al. [71] have recently shown that the standard method of moments and entropy balancing produce weights that are mathematically equivalent (up to optimization error or a normalizing constant). Drawing from Zubizarreta [220], Jackson et al. [35] propose a distinct weight estimation procedure that satisfies the conventional method of moments and maximizes the effective sample size. A larger effective sample size translates into minimizing the variance of the weights, with more stable weights producing a gain in precision at the expense of introducing some bias.

A potential extension to MAIC could involve estimating the treatment mechanism in the AC trial as well as the trial assignment mechanism. In a randomized trial, the treatment assignment mechanism is known — the true conditional probability of treatment among the randomized subjects is known and, in expectation, independent of the covariates. Nevertheless, modeling this probability, e.g. using a parametric model, is beneficial to control for random (chance) imbalances in baseline covariates between study arms. In other contexts, this has improved the precision and efficiency of propensity score estimators [125, 221–225].

BAYESIAN MODULARITY The marginalized outcome regression methods, particularly the Bayesian approaches, can be readily adapted to address missing values in the AC IPD. As seen in subsection 4.6.1, Bayesian G-computation and the synthesis stage of MIM follow very closely the principles of multiple imputation, which is also, arguably, a fundamentally Bayesian operation. Missing covariates and outcomes in the IPD could be imputed in each MCMC

iteration, accounting naturally for the uncertainty in the missing data. Addressing "missingness" in the BC study is not possible without access to the patient-level data.

Throughout the thesis, we have made certain assumptions about the covariate distribution in the BC population. We have treated the covariate moments $\boldsymbol{\theta}$ and the correlation information $\boldsymbol{\rho}$ as fixed. The Bayesian frameworks could be extended to account for this additional layer of uncertainty, in the specification of $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$ and also in the selected marginal distribution forms for BC. Bayesian regression approaches can also account for other issues such as measurement error in the IPD [179, 180]. Bayesian model averaging can be incorporated to capture structural or model uncertainty [226]. By drawing outcome predictions under various models, complex relationships in the patient-level data may be reproduced more accurately, offering some protection against parametric model misspecification.

In the Bayesian procedures, both "hard" (e.g. the results of a meta-analysis) and "soft" (e.g. clinical rationale from experts) evidence can be used to form the prior distributions for the conditional prognostic and interaction effects. The specification of the parametric outcome model requires "dichotomizing" whether a variable is an effect modifier or not, i.e., in statistical terms, specifying whether interactions with treatment do or do not exist. Bayesian shrinkage methods allow interactions to be "half in, half out" of the model [227–229]. For instance, one can specify skeptical or regularization prior distributions for the interaction effects, over all potential candidate effect modifiers. In the words of Simon and Freedman [229], this "encourages the quantification of prior belief about the size of interactions that may exist. Rather than forcing the investigator to adopt one of two extreme positions regarding interactions, it provides for the specification of intermediate positions".

5.3.3 *Limitations of the methods and simulation study*

METHOD LIMITATIONS Care must be taken where sample sizes are small in population-adjusted indirect comparisons. Low sample sizes cause substantial issues for the accuracy of MAIC due to unstable weights. Also, MIM assumes that the posterior distribution of $\Delta_{10}^{(2)}$ is approximately normal, and low sample sizes may break this normality assumption. As the sponsor company is directly responsible for setting the value of N, the AC trial should be as large as possible to maximize precision and accuracy. The sample size requirements for indirect comparisons, and more generally for economic evaluation, are considerably larger than those required to demonstrate an effect for the main clinical outcome in a single RCT. However, trials are usually powered for the main clinical comparison, even if there is a prospective indirect, potentially adjusted, comparison down the line. Ideally, if the manufacturer intends to use standard or population-adjusted indirect comparisons for reimbursement purposes, its clinical study should be powered for the relevant methods.

Note that sponsors tend to run multiple RCTs instead of one larger RCT for marketing authorization applications. If there are many different IPD RCTs, it is necessary to fit the covariate-adjusted regression to each patient-level dataset and marginalize against the BC pseudo-population in G-computation and MIM. Similarly, one would apply MAIC to each study

individually, reweighting each patient-level dataset against the BC study report. Then, a meta-analysis of effect measure estimates can be performed in the same population using the marginalized or weighted results from the IPD studies and the original effect estimate published in the ALD study.

The population adjustment methods outlined in this thesis are only applicable to pairwise indirect comparisons, and not easily generalizable to larger network structures of treatments and studies. This is because the methods have been developed in the two-study scenario seen in this simulation study, very common in HTA submissions, where there is one AC study with IPD and another BC study with ALD. In this very sparse network, indirect comparisons are vulnerable to bias induced by effect modifier imbalances. In larger networks, multiple pairwise comparisons do not necessarily generate a consistent set of relative effect estimates for all treatments. This is because the comparisons must be undertaken in the ALD populations.

Another issue is that the ALD population(s) may not correspond precisely to the target population for the decision. Marginal estimands in different populations may not match if there are differences in the distribution of effect modifiers. This is a problem of external validity: if populations are non-exchangeable, an internally valid estimate for the marginal estimand in one population is not necessarily unbiased for the marginal estimand in the other(s) [230, 231]. To address this, one suggestion would be for the decision-maker to define a target population for a specific disease into which all manufacturers should conduct their indirect comparisons. The outcome regression approaches discussed in Chapter 4 could be applied to produce marginal effects in any target population. The target could be represented by the joint covariate distribution of a registry, cohort study or some other observational dataset, and one would marginalize over this distribution. Similarly, MAIC can reweight the IPD with respect to a different population than that of the BC study.

METHOD ASSUMPTIONS The comments about potential failures of assumptions in Section 3.3.3 are still relevant, with additional concerns being raised in the next few paragraphs. Population-adjusted indirect comparisons mostly depend on the same assumptions (detailed in Supplementary Appendix A) including: (i) internal validity of the AC and BC trials, (ii) consistency under parallel studies, (iii) accounting for all effect modifiers of treatment A vs. C in the adjustment (i.e., the conditional constancy of the A vs. C marginal treatment effect or the conditional ignorability, unconfoundedness or exchangeability of trial assignment/selection for such treatment effect), (iv) that there is overlap between the covariate distributions in AC and BC (more specifically, that the ranges of the selected covariates in the AC trial cover some of their respective ranges in the BC population), (v) that the joint covariate distribution of the BC population has been correctly specified, (vi) and parametric modeling assumptions.

Assumptions (i) and (ii) are made by any indirect treatment comparison or meta-analysis. The other, largely untestable, assumptions are unique to population-adjusted analyses and their violation may lead to bias. The most crucial assumptions underlying population-adjusted indirect comparisons relate to the correct specification of the trial assignment logistic regression

(in the case of MAIC), and of the covariate-adjusted outcome regression (in the case of conventional STC, parametric G-computation and MIM).

In practice, there will be model misspecification if there is incomplete information on effect modifiers for one or both of the trials. Conditional exchangeability ("no omitted effect modifiers") is a fundamental assumption for all methods. However, it is not directly testable with the available data due to the lack of additional individual-level outcome information for the BC study [61]. In collaboration with clinical experts, the most plausible effect modifiers should be selected for the base-case analysis. Nevertheless, the effect modifier status of covariates is difficult to ascertain, particularly for novel treatments with limited prior empirical evidence and clinical domain knowledge [136]. Therefore, we will never be completely certain that all effect modifiers have been accounted for, or of the validity of the population adjustment.

Consequently, sensitivity analyses are warranted under alternative model specifications to explore the dependence of inferences on the model and the robustness of results [232–234]. In the context of "generalizability", Nguyen et al. [233] have recently developed an approach for sensitivity analysis. This is applicable where potential effect modifiers are measured only in the AC trial but not in the BC study, given some assumptions about the missing effect modifiers. Dahabreh [175] proposes a bias function strategy for sensitivity analyses, which does not require individual-level information on unobserved effect modifiers. Further research should adapt this recent work to our "limited patient-level data" setup.

Parametric modeling assumptions will not hold under incorrect model specification, e.g. in the outcome regression methods, if only linear relationships are considered and the selected covariates have non-linear interactions with treatment on the linear predictor scale. This simulation study only considers a best-case scenario with correct parametric model specification. To predict the outcomes, we use the logistic regression model implied by the data-generating mechanism. Similarly, the model for estimating the weights is the best-case model in MAIC because the right subset of covariates has been selected as effect modifiers and the balancing property holds for the weights with respect to the effect modifier means, as mentioned in subsections 3.1.4 and 5.1.4.1.[5] Also, effect modification has been correctly specified as linear, but scale conflicts would arise if effect modification status, which is scale-specific, had been justified on the wrong scale, e.g. if the true treatment effect modification were non-linear or multiplicative, e.g. age in cardiovascular disease treatments.

In real applications, these modeling assumptions are difficult to hold because, unlike in simulations, the correct specification is unknown, particularly where there are a large number of covariates and complex relationships exist between them. The simulation study presented in this chapter demonstrates proof-of-concept for the outcome regression methods and for MAIC, but does not investigate how robust the methods are to failures in assumptions. Future

5 The MAIC implementation is optimal in terms of precision and accuracy because the trial assignment model only balances the two covariates that interact with treatment. Nevertheless, these are not the only two covariates that are associated with trial assignment. Consider balancing the full set of covariates that predict trial assignment (a total of four covariates, including the two predictors with only main effects in the data-generating outcome model). Variance would be increased without improving the potential for bias reduction in the BC population. The behavior of MAIC would be more unstable because of weaker overlap. More extreme weights would be produced, and finite-sample or "chance" overlap violations would be more likely, particularly with small AC sample sizes.

simulation studies should explore performance in scenarios where assumptions are violated, in order to draw more accurate conclusions with respect to practical applications and limitations.

The general-purpose nature of the methods presented in Chapter 4 may provide some degree of robustness against model misspecification because the covariate-adjusted outcome model does not necessarily need to be parametric. Non-parametric regression techniques or other data-adaptive estimation approaches can be used to detect (higher-order) interactions, product terms and non-linear relationships, offering more flexible functions to predict the conditional outcome expectations. These may enhance the likelihood of correct model specification with respect to parametric regressions, but are prone to overfitting, particularly with small sample sizes. They can also minimize "data snooping" problems (e.g. the analyst selecting the model specification or the effect modifiers on the basis of statistically significant treatment effects), specially when there are no clear hypotheses about effect modification ex ante.

SPECIFICATION OF THE BC POPULATION Population-adjusted indirect comparisons make certain assumptions to approximate the joint distribution of covariates in the BC trial, but these assumptions differ slightly. In MAIC, as stated in the NICE Decision Support Unit technical support document [18], "when covariate correlations are not available from the (BC) population, and therefore cannot be balanced by inclusion in the weighting model, they are assumed to be equal to the correlations amongst covariates in the pseudo-population formed by weighting the (AC) population." In the conventional version of STC, the correlations between the BC covariates are assumed to be equal to the correlations between covariates in the AC trial.

In the marginalization methods proposed in Chapter 4 (parametric G-computation and MIM), more explicit and stringent distributional assumptions are made in the "covariate simulation" step. The methods assume the joint distribution of the BC covariates is specified correctly, by the combination of the specified marginal distributions and correlation structure. In the simulation study, we have assumed that the pairwise correlations of the covariates and the parametric forms of their marginal distributions are identical across trials. It is important to assess the robustness of the methods to failures in these distributional assumptions.

Note that the covariate distributional assumptions could be relaxed or verified empirically if trial publications included more complete summary statistics, e.g. information on the covariates' correlation structure or their observed marginal distributions, as opposed to simple summary tables of means/proportions and standard deviations. This information would allow us to approximate the full joint distribution of the BC covariates more accurately and reduce the risk of misspecifying the BC population. We have decided to mimic the AC pairwise correlations as, in principle, the relationships between covariates should be similar across trials.

5.4 CONCLUDING REMARKS

In Chapter 2, I established that the traditional regression adjustment approach in population-adjusted indirect comparisons targets a conditional treatment effect for A vs. C. In Chapter

3, I showed empirically that this effect is incompatible in the indirect treatment comparison, producing biased estimation where the measure of effect is non-collapsible. In addition, this effect is not of interest in our scenario because, as discussed in the next chapter, we seek marginal effects for policy decisions at the population level. In Chapter 4, I proposed several approaches for marginalizing the conditional estimates produced by covariate-adjusted regressions. The procedures are applicable to a wide range of outcome models and target marginal treatment effects for A vs. C that have no compatibility issues in the indirect treatment comparison.

In this chapter, I have demonstrated that the novel marginalized outcome regression approaches achieve greater precision than MAIC and are unbiased under no failures of assumptions. Hence, the development of these approaches is appealing and impactful. As observed in the simulation study in this chapter, these methodologies are more efficient than weighting, providing more stable estimators. We can now capitalize on the advantages offered by outcome regression with respect to weighting in our scenario, e.g. extrapolation capabilities and increased statistical precision. Furthermore, I have shown that the marginalized regression-adjusted estimates provide greater statistical precision than the conditional estimates produced by the conventional version of STC. While this precision comparison is irrelevant, because it is made for estimators of different estimands, it supports previous research on non-collapsible measures of effect [44, 50].

Marginal and conditional effects are regularly conflated in the literature on population-adjusted indirect comparisons, with many simulation studies comparing the bias, precision and efficiency of estimators of different effect measures. The implications of this conflation are widely misunderstood but must be acknowledged in order to provide meaningful comparisons of methods. I have built on previous research conducted by the original authors of STC, who have also suggested the use of a preliminary "covariate simulation" step [19, 81]. Nevertheless, up until now, there was no consensus on how to marginalize the conditional effect estimates. For instance, in Chapter 2, I discouraged the "covariate simulation" approach when attempting to average on the linear predictor scale [91]. Averaging on the linear predictor scale, i.e., computing the conditional linear prediction under each treatment for every simulated subject and averaging the linear predictions across all subjects, then calculating the difference between the average predictions, reduces to the conventional version of STC (i.e., to "plugging in" the mean BC covariate values). It is equivalent to averaging "predictions at the mean" [172] or estimating the "mean at mean covariates" [173] (as discussed in subsection 4.4.3), hence producing conditional effect estimates for A vs. C, as opposed to marginal estimates. I hope to have established some clarity.

The presented marginalization methods have been developed in a very specific context, common in HTA, where access to patient-level data is limited and an indirect comparison is required. However, their principles are applicable to estimate marginal treatment effects in other situations. For instance, in scenarios which require marginalizing out regression-adjusted estimates over the study sample in which they have been computed. Alternatively, the frameworks can be used to transport the results of a randomized experiment to any other

external target population; not necessarily that of the BC trial. In both cases, the required assumptions are weaker than those required for population-adjusted indirect comparisons.

6

CHAPTER 6: CONCLUDING REMARKS

In this final chapter, I provide a succinct summary of the thesis' contributions (Section 6.1), referring back to the aims set out in Chapter 1. In Section 6.2, I address the last of these objectives, providing clarifications on what the target of the analysis, i.e. the estimand, should be for population-adjusted indirect comparisons. Part of the research in this section is included in the commentary "Target estimands for population-adjusted indirect comparisons" (Remiro-Azócar, 2021).[1] Finally, Section 6.3 suggests themes worth exploring for ongoing and future research.

6.1 CONTRIBUTIONS OF THE THESIS

We return to the objectives of the thesis, set out in Chapter 1. These are:

- To review methods currently used for population-adjusted indirect comparisons, evaluating and comparing their statistical performance through comprehensive simulation studies;
- To develop novel outcome modeling methodologies that improve the performance of the existing population adjustment methods and can be embedded within a Bayesian framework;
- To influence practice by making recommendations on the way and circumstances in which population-adjusted indirect comparisons should be applied;
- To provide clarifications on what the target of the analysis, i.e. the estimand, should be for population-adjusted indirect comparisons, given that these are used to inform reimbursement decisions at the population level in HTA.

I have addressed the first research objective by: (1) providing a detailed review of the methods currently used for population-adjusted indirect comparisons in HTA in Chapter 2; and (2) by conducting comprehensive simulation studies in Chapter 3 and Chapter 5. The studies identified issues with the typical approach to outcome regression. Namely, the treatment coefficient of the multivariable regression used for covariate adjustment produces a conditional

1 The article has been submitted to Statistics in Medicine and is available at: https://arxiv.org/abs/2112.08023.

estimate, which is incompatible in the indirect comparison and not relevant for reimbursement decisions at the population level.

I have addressed the second objective by proposing several approaches to marginalization in Chapter 4, and by evaluating their statistical performance with respect to that of the existing methods for population adjustment in Chapter 5. The novel methodologies are applicable for the most common types of outcome models in evidence synthesis in HTA, generalized linear models and survival models, and can accommodate a Bayesian statistical framework. The methods integrate or average out the covariate-adjusted conditional model over the BC covariate distribution to produce marginal treatment effect estimates in the BC population. The recovered marginal estimates allow for compatible indirect treatment comparisons and are appropriate for population-based inference. I find that, when adjusting for covariate differences across populations, (marginalized) outcome regression is unbiased under no failures of assumptions, and more precise and efficient than weighting in estimating marginal treatment effects.

The results of the simulation studies in Chapter 3 and 5 can be translated into practice, thereby addressing the third objective of the thesis. For instance, I have established that MAIC incurs more substantial precision losses than outcome modeling where covariate overlap is poor and effective sample sizes after weighting are small. This suggests that MAIC should be avoided in these scenarios and outcome modeling preferred. In addition, I have shown that, when the effect measure of interest is the log-odds ratio, regression-adjusted estimates of the marginal effect are more precise and efficient than the original conditional estimates produced by conventional outcome regression. This has implications for practice, where conditional estimates are often preferred on the grounds of precision and efficiency [164]. Another example of an implication for practice is that the robust sandwich variance estimator in MAIC underestimates variability where effective sample sizes are small. In these cases, the non-parametric bootstrap should be preferred for uncertainty quantification.

The final objective is addressed in the next section. Some disagreement remains on what the target estimand should be for population-adjusted indirect treatment comparisons. The debate is of central importance for policy-makers and applied practitioners in HTA. The debate is also of crucial importance for this thesis, where I have established a clear preference for marginal estimands as the inferential target in population-adjusted indirect comparisons.

6.2 TARGET ESTIMANDS FOR POPULATION-ADJUSTED INDIRECT COMPARISONS

I recently participated in a very interesting discussion with Phillippo, Dias, Ades and Welton [39, 164], in response to their research article titled “Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study” [34]. The original article presents an extensive simulation study evaluating the statistical performance of different population adjustment methods in the context of anchored indirect treatment comparisons. Three methods are investigated: MAIC, STC, and a novel method recently proposed by the authors called multilevel network meta-regression (ML-NMR) [38].

In a recent editorial, co-authored with Anna Heath and Gianluca Baio [39], I highlight that the different methodologies target distinct measures of effect (as per Section 2.5). MAIC is based on propensity score weighting and targets a marginal treatment effect. STC and ML-NMR are outcome modeling-based methods, with effects estimated by the treatment coefficient of a multivariable regression. As we have seen in this thesis, the typical implementation of STC targets a conditional treatment effect that, almost invariably, is incompatible in a pairwise indirect comparison, producing bias for non-collapsible measures of effect [39, 91]. ML-NMR, developed by the authors of the simulation study [38], extends outcome modeling to handle larger networks of treatments and studies. It targets a conditional treatment effect without the estimand compatibility issues of STC. In my original response [39], I remark that the appropriateness of each methodology depends on the preferred inferential target, and that one should carefully consider whether a marginal or conditional treatment effect is of interest in a population-adjusted indirect comparison.

In their reply to my editorial, Phillippo et al. demonstrate that ML-NMR can potentially be used to target marginal treatment effects [164]. Therefore, the method could support inference at the individual level and at the population level. This extension is a very relevant and impactful development for evidence synthesis, which will help overcome many limitations of pairwise indirect treatment comparisons in the estimation of marginal effects. Nevertheless, disagreement remains on what the target estimand for HTA should be. In their response, Phillippo et al. comprehensively endorse the use of conditional treatment effect estimates to inform decision-making at the population level [164]. However, HTA agencies make reimbursement decisions at the population level. Therefore, I believe that estimates of the marginal treatment effect are necessary. Settling this debate is of central importance to offer a conclusion for policy-makers and applied practitioners in the field.

6.2.1 *Target estimands in randomized controlled trials*

The objective of population-adjusted indirect comparisons (and, more generally, of evidence synthesis in HTA) is to emulate the analysis that would have been performed in an ideal head-to-head RCT, directly comparing the drugs of interest. There has been much relevant debate over what the target estimand of an RCT should be. Note that, by “ideal”, I mean that the hypothetical RCT should have high internal validity, but also high external validity (this shall be discussed in Section 6.2.3) [235, 236].

Phillippo et al. use the following arguments to select the conditional estimand as the most appropriate inferential target for decision-makers in population-adjusted indirect comparisons [164]:

1. Conditional estimands account for differences in the distribution of prognostic covariates between groups but “marginal estimands do not account for known population characteristics”.

2. Conditional estimands have a “population-average” interpretation if treatment-by-covariate interactions are excluded from the analysis.

3. Conditional estimands are “a more efficient choice” than marginal estimands.

I examine these points, which follow from common misunderstandings, on a case-by-case basis. All of the arguments are based on properties that are inherent to *estimators* (the method of analysis), not *estimands* (the target of the analysis). Point 1 conflates the terms “marginal” and “unadjusted”. Nevertheless, estimates of the marginal effect need not be crude or unadjusted and may also be covariate-adjusted [50]. Points 2 and 3 generalize conclusions based on covariate adjustment with linear regression, and do not apply to non-linear models with non-collapsible measures of effect. With respect to the second point, the population-level interpretation of conditional estimates follows from collapsibility and does not necessarily hold for the underlying conditional estimands. For non-collapsible effect measures, neither conditional estimates nor estimands have a population-level interpretation. Concerning the third point, estimators of marginal effects tend to be more precise and efficient than estimators of conditional effects where the measure of effect is non-collapsible. In any case, precision and efficiency comparisons are inconsequential for estimators targeting distinct estimands.

6.2.1.1 *Marginal is not synonymous with unadjusted*

RCT analyses often adjust for one or more baseline covariates to correct for empirical confounding caused by chance imbalances between treatment groups. Covariate-adjusted analyses have many advantages over unadjusted analyses. Incorporating the prognostic information can result in a more efficient use of data and, as stated by Phillippo et al. [164], the adjusted analysis is “the recommended analysis that would be undertaken in the ideal (RCT) evidence scenario” in the trials literature [40, 237–239].

Nevertheless, the term “marginal” is not interchangeable with “unadjusted”. Marginal is often interpreted as unadjusted and conditional as adjusted. However, the distinction between marginal and conditional describes the estimand, and that between adjusted and unadjusted relates to the estimator. It is true that unadjusted estimates of the marginal effect ignore any information on the distribution of prognostic covariates in the sample. Therefore, these cannot directly compensate for any lack of balance between treatment groups. However, estimates of the marginal effect can also be covariate-adjusted. In fact, covariate-adjusted marginal estimates are regularly used in the RCTs literature to correct for chance imbalances in baseline covariates and to improve precision [240–242].

Indeed, one can adjust for covariates using the outcome model and then average or standardize over a specific population to estimate marginal (but covariate-adjusted) effects that do compensate for lack of balance [50, 243, 244]. For instance, in their reply, Phillippo et al. [164] illustrate how the conditional effect estimates produced by ML-NMR could be marginalized where the outcome regression is a generalized linear model, by integration over the joint covariate distribution in the target population. The resulting covariate-adjusted estimate of

the marginal effect would fully exploit the covariate information and account for imbalances in baseline characteristics.

After all, "population-adjusted" indirect comparisons are "covariate-adjusted" indirect comparisons. MAIC uses weighting to produce a covariate-adjusted estimate of the marginal effect that is combined with an unadjusted marginal estimate in a pairwise indirect comparison. Similarly, STC can be adapted using G-computation or model-based standardization [159], as outlined in Chapter 4 of this thesis, so that a covariate-adjusted marginal estimate is produced that has no compatibility issues in the indirect treatment comparison [245, 246]. One can estimate the marginal mean outcomes by treatment group, based on the fitted outcome regression. Then, the model-based predictions can be averaged over the joint covariate distribution of the target trial (that with aggregate-level data) and contrasted to produce an estimate of the marginal treatment effect that accounts for baseline covariates [245, 246].

In summary, I do not call for crude unadjusted analyses over analyses adjusted for measured covariates [39]. Covariate-adjusted analyses may target marginal or conditional effects, and the debate is about covariate-adjusted estimates of the marginal estimand versus covariate-adjusted estimates of the conditional estimand.

ON THE PREFERENCES OF REGULATORY AGENCIES Phillippo et al. state that "it is recommended practice to include prespecified prognostic factors in the analysis model" [164]. Indeed, covariate adjustment is strongly advised by regulatory authorities such as the United States Food and Drug Administration (FDA) [247, 248] and the European Medicines Agency (EMA) [249] when approving new treatments. Nevertheless, this does not mean that the conditional estimand is the primary focus of such agencies. Recent FDA guidance seems to encourage the use of covariate-adjusted estimates of the marginal treatment effect as opposed to covariate-adjusted estimates of the conditional effect [247]. The addendum to the International Council for Harmonisation E9 guidelines, Statistical Principles for Clinical Trials [250], which introduces the estimands framework and has been adopted by the FDA and the EMA, discusses "population-level summary measures" of outcome as the primary target of inference in RCTs. This suggests a preference for marginal estimands (see 6.2.1.2).

6.2.1.2 *On the population-average interpretation of conditional estimands*

Phillippo et al. argue that conditional estimands may have a "population-average" or an "individual-level" interpretation [164]. For instance, covariate adjustment through the "analysis of covariance" (ANCOVA) linear model specifies main effects but excludes treatment-by-covariate interactions. Therefore, the conditional treatment effect is assumed identical across covariate levels and is, therefore, not specific to subgroup membership. Conversely, with the inclusion of interaction terms, the conditional treatment effect is believed to differ across covariate values and the estimate may have an "individual-level" interpretation. It is argued that the "population-average" estimate of the conditional effect can be used to make population-level decisions [164].

I make two important remarks. Firstly, the homogeneity or uniformity of the conditional treatment effect estimate follows from statistical modeling assumptions, which may not be plausible. While the estimator may assume constancy across all patients, the true subgroup effects are not necessarily constant. If the constancy assumption does not hold, the ANCOVA conditional estimand no longer has a population-average interpretation, and corresponds to an ambiguous weighted average of subgroup-specific estimands [251]. In any case, the preference for marginal or conditional estimands as inferential targets should not depend on the estimator. Conditional estimands can be well defined as subgroup-specific or individual-level effects, regardless of modeling assumptions. On the other hand, the population-level interpretation of the conditional estimate depends on such implicit assumptions.

Secondly, the population-level interpretation of the conditional estimate in the case of no interaction relies on the measure of effect being collapsible. An effect measure is collapsible when the marginal measure can be expressed as a weighted average of the subgroup-specific conditional measures [93, 99, 252, 253]. Mean differences in linear regression are collapsible across covariates. Without interaction terms, one assumes that there is no effect modification on the mean difference scale, such that mean differences are the same in every subgroup. In this case, the subgroup-specific mean difference estimates are also equal to the marginal mean difference estimate. The “population-average” interpretation of any conditional estimate simply reflects that it coincides with the marginal estimate due to collapsibility. Therefore, making a distinction [164] between “population-average” conditional and “population-average” marginal estimands is not necessary.

The situation is even more nuanced where the measure of effect is non-collapsible, as is the case for the (log) odds ratio or the (log) hazard ratio. Unlike the mean difference, these are non-collapsible because the marginal measure cannot represent a weighted average of the individual- or subgroup-level conditional measures, even in the absence of confounding [88, 98, 254]. Consider that a conditional log-odds ratio estimate is derived from the treatment coefficient of a main effects logistic regression, assuming homogeneity. This estimate, which is equal to the constant subgroup-specific effect estimates, cannot have a population-level interpretation [88, 96, 253, 255]. This is a mathematical phenomenon that pertains to numeric properties of the measure of effect [93, 99]. For the odds ratio, it is consequence of a special case of Jensen’s inequality [254].

THE TRANSPORTABILITY OF EFFECT MEASURES A general empirical observation is that conditional effects are more generalizable or transportable than marginal effects across different populations [164], because marginal estimands may change across different marginal covariate distributions. It is worth noting that this is currently an area of debate. Another intuition is that conditional effects are less transportable because the estimand is dependent on the selected adjustment model. There may be many conditional estimands for a given population, one for every possible combination of baseline covariates and model specification. Conditioning on different covariate sets leads to different conditional estimands, with their estimates not being comparable across studies [50, 256]. On the other hand, marginal estimands can be

clearly defined without reference to a particular adjustment model. Pearl and Bareinboim [139] claim that marginal effects are more transportable than conditional effects, providing a mathematical proof. In the words of Daniel et al. [50], this result highlights that "any measured covariate can be adjusted for in the analysis, and then marginalized over according to any desired reference distribution, resulting in a marginal estimand that is just as transportable as any conditional estimand".

6.2.1.3 *Efficiency considerations*

The main argument by Phillippo et al. is that the conditional estimand is more efficient than the marginal estimand in the hypothetical evidence scenario described by the ideal RCT [164]. It is claimed that the conditional is a more appropriate target estimand for health policy because it provides "more efficient decision-making".

It is true that for linear regression with maximum-likelihood estimation and continuous outcomes, a covariate-adjusted estimate of the conditional treatment effect should have a lower standard error than the unadjusted estimate of the marginal effect. The decrease in standard error is greater when the correlation between the baseline covariate(s) and outcome is strong, leading to a reduction in residual variance [257]. Nevertheless, this is not the case when working with non-collapsible effect measures such as odds ratios in logistic regression with binary outcomes [44, 96, 257, 258], or hazard ratios in Cox proportional hazards regression with survival outcomes [259, 260]. These are two of the most widely used parameters, statistical models and outcome types in evidence synthesis in HTA [30].

In these cases, adjusted estimates of the conditional estimand tend to have reduced precision and efficiency with respect to unadjusted estimates of the marginal estimand. For odds and hazard ratios, the covariate-adjusted maximum-likelihood estimator of a conditional effect has a standard error at least as large as the unadjusted maximum-likelihood estimator of a marginal estimand, in the ideal RCT [50, 258, 259]. In fact, for non-collapsible effect measures, it is marginalized covariate-adjusted estimates that tend to have lower standard errors than both the original covariate-adjusted estimates of the conditional [44, 50] and the unadjusted estimates of the marginal [240–242]. In addition, covariate-adjusted estimates of the marginal odds ratio seem to be less susceptible to small-sample and sparse-data bias than covariate-adjusted estimates of the conditional odds ratio [261].

Pursuing greater precision and efficiency would make the covariate-adjusted marginal estimates more attractive than the original conditional estimates for non-collapsible effect measures in the ideal RCT. However, these comparisons are arguably inconsequential, because they are made for estimators targeting different estimands [50, 242]. One has to suitably define the target estimand before performing a "like-for-like" comparison of different estimators. When adjusted and unadjusted estimates of the marginal estimand are compared in the ideal RCT, covariate adjustment does increase precision by leveraging the prognostic information accounting for unexplained variation in the outcome.

A small but important aside: in contrast to the "ideal RCT" scenario, covariate adjustment increases the variance of indirect treatment comparisons [91]. This is a desirable feature

because standard unadjusted approaches such as the Bucher method [7] ignore cross-trial differences in covariates that influence the outcome. Therefore, as seen in Chapter 3, they may produce overly precise estimates and undercoverage [58, 91]. Covariate-adjusted indirect comparisons account for the additional uncertainty produced by covariate differences. A reduction in precision is natural and necessary, and a function of the "distance" between the covariate distributions; we are trying to learn about a treatment effect in a different study than that in which it was originally observed. The simulation study in Chapter 5 shows that standardized regression-adjusted (and in some cases, weighting-adjusted) estimates of the marginal (log) odds ratio are also more precise than regression-adjusted estimates of the conditional (log) odds ratio in this context. These results are expected to hold for non-collapsible effect measures in general.

I emphasize that the estimand of interest should be tailored to the scientific question that is being addressed. Namely, the choice of estimand should determine the estimator, and not vice versa. It makes sense to proceed sequentially, first determining the estimand that best answers the decision problem, and then using a method or analytic approach that is well suited for estimating it from the clinical trial data. Statistical efficiency should not drive the choice of the estimand. On the other hand, the estimand, unambiguously selected on the basis of relevance to decision-making, should drive the choice of the most statistically efficient estimator. This is because efficiency is a property inherent to estimators, not estimands.

6.2.2 *Target estimands in health technology assessment*

The development of pharmaceuticals is a multi-stage process, and HTA generally takes place late in this process. Subsection 6.2.1 has described factors that are relevant in preparation for the drug licensing stage. In this stage, the efficacy of a new medical technology is typically evaluated versus placebo or standard of care in an RCT. This trial may provide evidence supporting the regulatory approval of the drug by agencies such as the FDA and the EMA. In this setting, power considerations to test the "no effect" null hypothesis may also deserve attention.[2] Indirect treatment comparisons are not typically applied for hypothesis testing or to obtain regulatory approval. These are highly underpowered in scenarios commonly met in practice [24, 101, 263] and often lead to a conclusion of "no clinical benefit" [264].

I now set aside the long-standing debate about target estimands in RCTs and focus on the decision problem at hand. Following regulatory approval, a pharmaceutical product can be submitted to HTA agencies worldwide, e.g. NICE in England and Wales, which formulate recommendations on whether health care technologies should be publicly funded by national health care systems. The population adjustment methodologies evaluated and developed in this thesis aim to quantify treatment efficacy or effectiveness in this scenario. Nevertheless, the demonstration of efficacy and/or effectiveness is necessary but not sufficient for HTA agencies.

2 Covariate adjustment tends to produce increased power to detect a non-null treatment effect [238, 257, 258]. Power comparisons between marginal and conditional estimands are relevant because both share the same null value [50, 262], e.g. under the null hypothesis of no treatment effect, both marginal and conditional odds ratios are equal to one.

The resulting treatment effect estimates are used as inputs to health economic evaluations, e.g. cost-effectiveness analyses comparing two or more competitor treatments. HTA bodies typically make reimbursement decisions on the basis of these evaluations.

For instance, NICE operates an appraisal process in which companies submit evidence on both relative clinical and cost effectiveness. In this process, it is the mean cost and effectiveness at the population level that are relevant [265]. Evidence synthesis methods inform the mean treatment benefit in such analyses, where the main quantity of interest tends to be the incremental cost-effectiveness ratio (ICER), which is a population-level measure. For differential cost-effectiveness inferences and policies based on subject-level covariates, individual-level ICERs have been proposed [266]. However, these are currently of secondary interest in policy and may be problematic [267]. Within centralized health care systems such as the National Health Service, planning entities and providers use the appraisal process to set population-level policies such as quality measures and clinical guidelines, and to select treatments for the population of patients within their remit. In this context, the marginal treatment effect is a more relevant target than the conditional effect.

Conditional estimands would be of greater relevance in clinical practice or personalized/precision medicine, from the perspective of physicians making treatment decisions for individual patients. This is particularly the case if there is treatment effect heterogeneity and this is accounted for by the inclusion of treatment-by-covariate interactions. As advocated by Hauck et al. [40], conditional treatment effects "come as close as possible to the clinically most relevant subject-specific measure of effect". For instance, physicians may be interested in how effective treatment is conditional on the age, gender and/or medical history of a particular patient, and may not desire to average over these characteristics. Indeed, marginal estimands make little sense in the context of clinical care and are not applicable in the context of decision-making for individual patients.

If health care providers and reimbursement agencies were to make decisions at such level of granularity, conditional estimands would be of greater interest than marginal estimands. However, the research questions made by bodies such as NICE investigate how the average effect of an intervention impacts outcomes at the population level, and are used to make broad policy decisions and recommendations. In the health decision sciences, conditional treatment effects could also be of interest for individual-level microsimulation models that simulate the impact of interventions or policies on individual trajectories, which may be averaged out to estimate an overall population-level ICER.

Finally, a very important consideration in health economic evaluation is uncertainty quantification [268], where the assessment of parameter uncertainty is a central component [269, 270]. The conflation of marginal and conditional estimands is an issue for both collapsible and non-collapsible effect measures, because estimators targeting different estimands will produce different variance estimates. These variances quantify parameter uncertainty in cost-effectiveness analyses. Marginal and conditional estimates will quantify parameter uncertainty differently, and conflating these will lead to the incorrect propagation of uncertainty to the wider health

economic decision model. This is particularly dangerous for probabilistic sensitivity analysis [271].

6.2.3 *External validity*

6.2.3.1 *Established population-adjusted indirect comparison methods*

In subsection 6.2.2, I established why marginal estimands should be preferred as inferential targets for population-level reimbursement decisions in HTA. In 6.2.1.2, I stated that there is only one well-defined marginal estimand for a specific population. Nevertheless, marginal estimands can change if the population definition is modified. As population-adjusted indirect comparisons assume treatment effect heterogeneity, different populations with different effect modifier distributions will have marginal estimands of distinct magnitudes. In addition, as I discuss in this section, most "population-adjusted" indirect comparisons refer to "sample-adjusted" indirect comparisons. This clouds the discussion of estimands further.

The methods explored in this thesis have been developed in the context of pairwise comparisons in a two-study scenario, where there is one "index" study with IPD and another "comparator" study with unavailable IPD and only published aggregate-level data. A distinctive feature of the methodologies is that, due to patient-level data limitations, the methods contrast treatments in the comparator study (BC) sample, defined by the summary moments of baseline characteristics in "Table 1" of the publication. Inferences can only be interpreted within this sample-specific context, which imposes constraints on the marginal estimand that is targeted.

As currently conceptualized, the methods imply that the comparator study sample on which inferences are made is exactly the study's target population. Alternatively, the assumption is that the study sample is a random sample, i.e., representative, of such population, ignoring sampling variability in the patients' baseline characteristics and assuming that no random error attributable to such exists. In reality, the subjects of the comparator study have been sampled from a, typically more diverse, target population of eligible patients, defined by the trial's inclusion and exclusion criteria.

Random sampling is seldom feasible in recruitment strategies for trial participants. For instance, individuals with health-seeking behaviors are more likely to enroll in the trial. Candidates who meet the trial eligibility criteria may not be invited to participate [272]. Conversely, invited study-eligible individuals may not provide informed consent, an ethical necessity for enrollment, and choose not to participate. In summary, it is rarely the case that a study sample is a random sample of the target population of the study, because trial participation is subject to convenience sampling, and volunteerism or self-selection [272, 273]. This is a problem of external validity with respect to the target population of the trial.

A more important limitation is that, even if the comparator study sample is representative of the study's target population, such target population may be systematically different to the target population of policy interest [273, 274], i.e., the group of patients who will receive the intervention in routine clinical practice. The populations targeted by clinical trials tend to be more

narrowly defined and less heterogeneous in composition and health status,[3] to maximize power in efficacy and safety testing, and to enhance statistical precision and efficiency [277–279]. In addition, the comparator study may have been conducted in many separate geographical regions, different to that of relevance for HTA decision-making. Payers reimburse treatments at the local market level and HTA decisions are likely to concern local patient populations [280]. These are questions of external validity with respect to the target population for the decision.

In terms of estimands, let's consider three distinct marginal estimands. MAIC and our implementations of parametric G-computation and MIM would target a sample-average marginal estimand in the comparator study. However, this may not coincide with the population-average marginal estimand for the target population of the study. In turn, this may not match the population-average marginal estimand for the relevant target population required for HTA decision-making. If the samples or populations are non-exchangeable, an internally valid estimate for the marginal estimand in one sample/population is not necessarily unbiased for the marginal estimand in the others [230, 231]. The relationships between the different samples and populations are displayed in Figure 15. This diagram has been inspired by Degtiar and Rose [281]. Moving horizontally, pairwise methods such as MAIC are limited to transporting inferences from the index study sample to the comparator study sample.

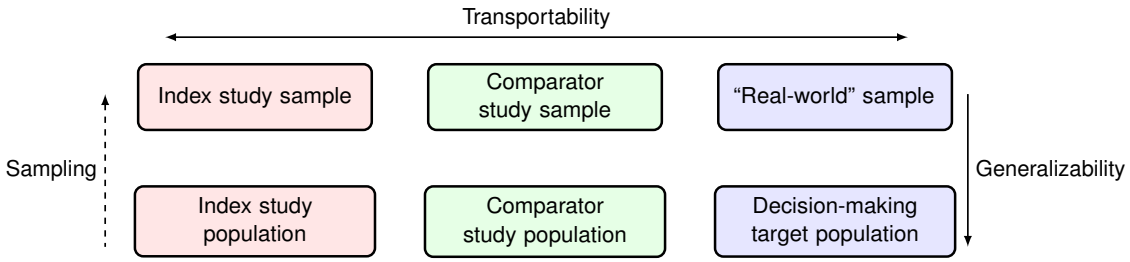


Figure 15: External validity addresses whether inferences can be extended beyond specific samples. Researchers make a distinction between generalizability and transportability. Generalizability entails generalizing the findings from an RCT to the population from which the trial participants were drawn, i.e., the RCT sample is a proper subset of the trial-eligible population. Transportability involves translating inferences to an external target sample or population.

6.2.3.2 *ML-NMR: new directions for evidence synthesis?*

ML-NMR presents abundant opportunities for evidence synthesis. Following the "marginalization" extension by Phillippo et al. [164], it allows for the estimation of marginal estimands in any of the study samples included in the meta-analysis. For instance, one could produce estimates in the most heterogeneous study, which may be more representative of the target population for HTA decision-making. Alternatively, one could produce estimates in the most homogeneous study to avoid overlap violations. Better yet, inferences may not be necessarily restricted to one of the studies included in the meta-regression. The target sample or population could be

3 An exception are so-called "pragmatic", "effectiveness" or "practical" trials [275, 276]. These are large-scale multicenter trials carried out in "real-world" settings. Their patients tend to be more representative of decision-making target populations because the trials have broad eligibility criteria and design elements that promote enrollment of a wide range of participants. However, this may come at the expense of lower adherence and higher drop-out rates.

generated from an external data source. Presuming that this contains the covariates that have been adjusted for in the studies, inferences would be transported to the target of interest.

The target could be defined by HTA policy-makers for the specific disease under study. It could be characterized by the joint covariate distribution observed in an observational sample or in secondary health data sources such as disease registries, cohort studies and insurance claims databases. Such administrative datasets have high cross-sectional richness, and are typically larger, less selected, and more representative of target populations of interest than the participants recruited into trials [282–284]. Electronic health records from hospital systems are also valuable tools to define appropriate "real-world" targets [285]. These are compelling due to several aspects, including reasonably large sample sizes and a high degree of clinical detail, specificity and breadth [286, 287]. HTA bodies such as NICE are increasingly using "real-world data" to identify representative populations and inform or update assessments of effectiveness and cost-effectiveness [288]. These covariate data are likely to be available at the time of the HTA appraisal process.

The sponsor of the index study, submitting evidence to HTA bodies, could use pairwise methods to weight or standardize its results to the external target source provided by decision-makers. However, IPD are unavailable for the comparator trial, for both the manufacturer submitting the evidence and the HTA agency assessing the evidence. Hence, the submitting company cannot weight or standardize the comparator study results to the external target. Therefore, it is challenging for the pairwise methods explored in this thesis, as currently conceptualized, to facilitate an indirect comparison in this target.

Finally, we should ask ourselves to what substantive population are indirect treatment comparisons supposed to apply to. The main premise of population-adjusted indirect comparisons is to provide equipoise, removing bias due to covariate differences across studies and comparing effect estimates in the same population. However, because treatment effects are assumed heterogeneous, any discussion of equipoise needs to be framed within the question: "equipoise to whom?". Equipoise in a particular study sample does not guarantee equipoise in the target population for decision-making. Any claim about equipoise is ill-formed without reference to the desired target population for inference. The same applies to any claim about representativeness. Unfortunately, most applications of population-adjusted indirect comparisons do not explicitly describe the target population of the analysis, as found by a recent review of HTA appraisals [30].

6.3 RECOMMENDATIONS FOR FUTURE WORK

Ongoing and future research will focus on the following themes to address limitations of the current work. Many of the points discussed in subsections 5.3.2, 5.3.3 and 6.2.3 are complementary to these ideas:

INTEGRATION OF MARGINALIZATION METHODS AND SURVIVAL ANALYSIS IN A UNIFIED HTA FRAMEWORK Further research will consider extensions of this work investigating

marginalization methods for parametric survival models. These extensions are important because most applications of population-adjusted indirect comparisons are in oncology and usually require parametric survival analysis [30]. However, current survival analysis applications of MAIC and STC in academic papers, submissions for reimbursement and simulation studies, focus exclusively on the Cox proportional hazards regression as the outcome model of interest.

DECOMPOSITION OF THE ASSUMPTIONS AND POTENTIAL SOURCES OF BIAS IN POPULATION-ADJUSTED INDIRECT COMPARISONS Given the large number of assumptions made by population-adjusted indirect comparisons, future simulation studies should assess the robustness of the methods to failures in assumptions under different degrees of data availability and model misspecification. All the assumptions required for valid population-adjusted indirect comparisons hold, by design, in the simulation studies in Chapter 3 and Chapter 5. Nevertheless, these assumptions are hard to meet and most of them are not directly testable.

UNANCHORED COMPARISONS While the thesis has focused on anchored indirect comparisons, most applications of population adjustment in HTA are in the unanchored setting [30], both in published studies and in health technology appraisals. We stress that RCTs deliver the gold standard for evidence on efficacy and that unanchored comparisons make very strong assumptions which are largely considered impossible to meet (absolute effects are conditionally constant as opposed to relative effects being conditionally constant) [9, 18]. Unanchored comparisons effectively assume that absolute outcomes can be predicted from the covariates, a heroic assumption. However, the number of unanchored comparisons is likely to continue growing as regulators such as the United States Food and Drug Administration and the European Medicines Agency are, increasingly, and particularly in oncology, approving new treatments on the basis of observational or single-armed evidence, or disconnected networks with no common comparator [289, 290]. As pharmaceutical companies use this type of evidence to an increasing extent to obtain accelerated or conditional regulatory approval, reimbursement agencies will, in turn, be increasingly asked to evaluate interventions where only this type of evidence is available. Therefore, further examinations of the performance of population adjustment methods must be performed in the unanchored setting.

DOUBLY-ROBUST ESTIMATION In this thesis, we have compared two types of "singly-robust" estimators: weighting-based methods such as MAIC, and outcome modeling-based methods. The former specify a model for the trial assignment mechanism and the latter specify a model for treatment effect heterogeneity. Both are said to be singly-robust because they rely on a single nuisance function. In the outcome modeling approaches, the nuisance function is the covariate-adjusted outcome regression. Its parameters are not of interest per se, but are necessary to estimate the marginal treatment effect for A vs. C. Similarly, in MAIC, the trial assignment logistic regression for the weights is not of interest per se, but is necessary to estimate the marginal treatment effect for A vs. C.

There is a third class of estimators: "doubly-robust" methods that combine the weighting model with the outcome model [145, 221, 291–295], thereby providing two opportunities to specify the correct model and for valid inference [37]. Typically, the methods are consistent if either the trial assignment model or the outcome model is correctly estimated, but not necessarily both [145]. In addition, they are potentially more efficient than singly-robust weighting estimators [147, 149]. While, in general doubly robust estimators should be less prone to model misspecification, they may amplify bias and imprecision where the two models are misspecified [148, 296]. To my knowledge, doubly robust methods have not been developed for the setting described in this thesis, and remain a topic for further investigation. These approaches are attractive because, in our context, we often have limited data and inadequate background knowledge on the drivers of the trial assignment process and, also, of treatment effect heterogeneity.

NON-PARAMETRIC TECHNIQUES For the outcome regression methods, we have focused on fitting finite-dimensional, a priori specified, parametric models in the AC study to predict outcomes in the BC population. For MAIC, we have assumed correct specification of a parametric trial assignment model. These assumptions may be unreasonable with weak background theory or information. The general-purpose nature of the proposed outcome regression approaches may offer some degree of protection against model misspecification. In outcome regression, one can estimate the Q-model (the first-stage regression, in the case of MIM) non-parametrically to avoid heavily relying on correct parametric model specification. Flexible algorithmic frameworks such as Bayesian additive regression trees [297–299] are well-suited to capture complex functional forms, e.g. interactions, non-linear and higher-order relationships, potentially being less susceptible than parametric regressions to model misspecification. Bayesian additive regression trees are implemented in many R packages, typically require little parameter tuning, and have exhibited excellent performance in predictions and causal inference applications [298, 300].

MACHINE LEARNING For both weighting and outcome modeling approaches, estimation of the nuisance model can be viewed as a prediction problem for which data-adaptive or machine learning methods can be used, e.g. in G-computation, these would be used to predict the potential outcomes. Techniques such as super learning [301, 302], a generalized ensemble methodology that uses cross-validation to combine multiple candidate prediction algorithms (e.g. splines, random forests, etc.) into a single predictive function, have outperformed the traditional parametric models under model misspecification in epidemiology applications [301, 303, 304].

Currently, the use of machine learning methods in our scenario is hindered by two issues: (1) machine learning algorithms have rich data needs; because our scenario is data-poor, they are prone to overfitting; and (2) limited theoretical justification for valid statistical inference (e.g. standard errors and interval estimates) when data-adaptive methods are used to estimate singly-robust nuisance models [305–307]. The use of the non-parametric bootstrap to obtain

standard errors is not supported by theoretical results and is, generally, not valid [308]. Typically, one does not know the sampling distribution of the marginal treatment effect estimators, which is likely irregular. Hence, use of the singly-robust estimators discussed in this thesis in conjunction with machine learning may be problematic.

On the other hand, the use of doubly-robust estimators together with machine learning may provide consistent variance estimators and valid interval estimates, due to a particular orthogonality property some possess [309]. A promising doubly-robust framework that can be integrated with flexible data-adaptive learners is targeted maximum-likelihood estimation [291, 293]. This has been used in conjunction with superlearning [303, 310] and cross-fit estimation [311, 312] to control for confounding in observational studies.

EXTENSION TO LARGER NETWORKS As mentioned in Section 6.2, a novel outcome regression method named multilevel network meta-regression (ML-NMR) has recently been introduced [38, 313]. ML-NMR generalizes IPD network meta-regression [314] to include aggregate-level data, reducing to this method when IPD are available for all studies.

ML-NMR is an outcome regression approach, with the outcome model of interest being identical to that of parametric G-computation and MIM. While the methods share the same assumptions in the two-study scenario, ML-NMR generalizes the regression to handle larger networks. Like Bayesian G-computation and MIM, ML-NMR has been developed under a Bayesian framework and estimates the outcome model using MCMC. It also makes parametric assumptions to characterize the marginal covariate distributions in BC and reconstructs the joint covariate distribution using a copula. The methods average over the BC population in different ways; Bayesian G-computation and MIM simulate individual-level covariates from their approximate joint distribution and ML-NMR uses numerical integration over the approximate joint distribution (quasi-Monte Carlo methods).

ML-NMR is a timely addition. It is applicable in treatment networks of any size with the aforementioned two-study scenario as a special case. This is important because a recent review [30] finds that 56% of NICE technology appraisals include larger networks, where the standard pairwise population-adjusted indirect comparisons cannot be readily applied.

In its original publication [38], ML-NMR targets a conditional treatment effect (avoiding the compatibility issues of conventional STC), because the effect estimate is derived from the treatment coefficient of a covariate-adjusted multivariable regression. However, ML-NMR can directly calculate marginalization integrals akin to those required for Bayesian G-computation and MIM (Equations 18 and 19 in Chapter 4). Phillippo et al. have recently demonstrated that ML-NMR can be adapted to target marginal treatment effects [164]. The population-adjusted indirect comparisons explored in this thesis target marginal estimands that are specific to the BC study. These may not be directly relevant for HTA decision-making. On the other hand, ML-NMR can potentially estimate marginal effects in any target population, presenting novel and exciting opportunities for evidence synthesis.

EXTENSION TO OBSERVATIONAL STUDIES Throughout the thesis, we have assumed that the IPD study is a randomized trial. The proposed population adjustment methodologies could be extended to the situation where the trial with IPD is an observational study. In the outcome regression approaches, one would have to include all confounders of the treatment-outcome relationship in the outcome model. In MAIC, a potential extension involves estimating the treatment assignment mechanism in the observational study as well as the trial assignment mechanism. The conditional probability of treatment among the study subjects, given the confounders, could be modeled using a parametric model. Then, additional reweighting by the inverse probability of treatment would be required.

In this scenario, one must overcome the limited internal validity of the study design. Because treatment assignment is non-random, additional assumptions would be required, e.g. conditional exchangeability within the study arms ("no unmeasured confounding") and the associated overlap/positivity condition [315, 316]. These assumptions are similar to those discussed in subsection 4.1.2 but would be expected to hold across treatment arms in the IPD study in addition to across study populations.

CASE STUDIES AND SOFTWARE TOOLS In the context of this research, IPD are difficult to obtain in practice. Therefore, a case study demonstrating the application of the methods is missing. The proposed methodologies should be applied to real examples in order to influence applied practice. This is a key priority for future research. Eventually, the research in this thesis should be implemented in suitable software tools with a number of worked examples. For instance, an R package would help practitioners and analysts to apply the methods in submissions for reimbursement.

SUPPLEMENTARY APPENDIX A: METHOD ASSUMPTIONS

Indirect comparisons of treatments seek to mimic the analysis that would be conducted in a head-to-head RCT and to recover the causal effect of treatment on the clinical outcome of interest. In our particular case, the term "causal" also alludes to the need to control for effect modification, an inherently causal concept. Hence, we base our discussion of assumptions on the ideas underlying the Neyman-Rubin Model for causal treatment effects. This was originally suggested by Neyman [317] in experiments with randomization-based inference, with extensions to observational studies later introduced by Rubin [62, 181, 318]. The central concept of this general framework is that of a potential outcomes approach to causal inference. Note that this discussion is non-technical and detailed theory and notation based on potential outcomes are not presented.

MAIC and the outcome regression approaches discussed in this dissertation share the following assumptions, required to make valid causal inferences in the BC population: (1) internal validity; (2) consistency under parallel studies; (3) conditional strong ignorability of trial assignment for the A vs. C treatment effect (this requires both the conditional constancy of relative effects and overlap/positivity across the covariate distributions); (4) correct specification of the BC population; and (5) (typically linear/parametric) modeling assumptions. The first two assumptions are made by any indirect treatment comparison or meta-analysis. Assumptions that are not specific to indirect treatment comparisons, e.g. those that are specific to the type of regression model used, such as proportional hazards or non-informative censoring for a Cox regression, are not discussed.

Note that the following assumptions can only guarantee a valid indirect comparison if the within-trial relative effects target compatible estimands of the same type. The majority of RCTs publish an estimate for B vs. C that targets a marginal treatment effect (any published conditional treatment effect is likely incompatible with that for A vs. C). Therefore, population adjustment methods should target a marginal treatment effect for A vs. C. If a comparison of conditional treatment effects is performed, these would have to be adjusted across identical sets of covariates, using the same model specification.

Those studying the generalizability of treatment effects often make a distinction between sample-average and population-average marginal effects [17, 59–61]. Typically, another implicit assumption made by population-adjusted indirect comparisons is that the marginal treatment effects estimated in the BC sample, as described by its published covariate moments in the case of the A vs. C treatment effect, coincide with those that would be estimated in the target population of the trial. Namely, either the study sample on which inferences are made is the study target population, or it is a simple random sample (i.e., representative) of such population, ignoring sampling variability.

Furthermore, when referring to "effect modifiers", we describe the covariates modifying the treatment effect measure for A vs. C in the linear predictor scale. We select the effect modifiers of treatment A with respect to C (as opposed to the effect modifiers of treatment B with respect to C), because we have to adjust for these to perform the indirect comparison in the BC population, implicitly assumed to be the target population. If we had IPD for the BC study and ALD for the AC study, we would have to account for the covariates that modify the effect of treatment B vs. C, in order to perform the comparison in the AC population.

INTERNAL VALIDITY

The first set of assumptions relates to the internal validity of the AC and BC trials. The trials are internally valid under the following structural assumptions, which are necessary for causal inference:

- Stable unit treatment value assignment (SUTVA). This assumption implies that: (1) the treatment of a given subject does not affect the potential outcomes of other individuals (non-interference) [319, 320]; and (2) there is only one version of each treatment (treatment-variation irrelevance) [56], implying that the treatment is comparable across units [321]. The first condition is questionable, for example, in a vaccine trial, where the outcome of an individual (i.e., developing the flu) depends on the vaccination status of others because of herd immunity. The second condition is questionable if there are differences among versions of treatment, e.g. in the delivery mechanism, that are relevant to the outcome of interest.
- Strongly ignorable treatment assignment. Ignorability implies that treatment assignment is independent of the potential outcomes [322]. Ignorability can be conditional on the observed baseline covariates or unconditional. Conditional ignorability is strong when there is positivity or overlap [323] i.e., any subject has a positive probability of being assigned to either treatment group given the baseline covariates.

The SUTVA assumption is met by appropriate study design [318]. By design, the conditions of positivity [324] and ignorability [325], whether this is plain or conditional on baseline characteristics, are met by randomized trials. The random allocation of treatment ensures that, on expectation, there are no systematic differences in the distribution of (measured and unmeasured) baseline covariates between treatment groups, i.e., there is covariate balance [84, 326]. Note that balance is a large sample property. In small samples, one may still observe modest residual differences in baseline characteristics. As formulated by Senn [327], in a RCT, over all the randomizations the groups are balanced, but for a particular randomization they may be unbalanced. Therefore, the internal validity assumptions are met if the AC and BC studies are appropriately designed trials with appropriate randomization and reasonably large sample sizes. Finally, we have assumed that internal validity in each trial is not compromised by other issues, such that there is negligible measurement error or missing data, the absence of non-compliance, etc.

In the unanchored case with single-arm studies, strongly ignorable treatment assignment is not required as there is no common comparator arm. However, unanchored comparisons are subject to the additional assumptions and biases of non-randomized study designs, which are often stronger [328].

CONSISTENCY UNDER PARALLEL STUDIES

Consistency under parallel studies [61] is the cross-trial version of the second condition of SUTVA (treatment-variation irrelevance). This assumption implies that potential outcomes for an individual under a given treatment are homogeneous regardless of the study assigned to the individual. For instance, treatment C should be administered in the same setting in both trials, or differences in the nature of treatment, e.g. in the clinical protocol or delivery mechanism, should not change its effect. In there are non-negligible differences in the versions of treatment, for instance, if treatment C is accompanied by adherence counseling in one of the trials, while such counseling is absent in the other, this assumption could be invalid.

Consistency under parallel studies means that population adjustment methods cannot adjust for cross-trial differences related to the nature of treatments, e.g. treatment administration, switching, dosing formulation, titration or co-treatments. Differences of this type are perfectly confounded with treatment [18], and MAIC and the outcome regression methods can only adjust for differences in trial population characteristics. This assumption is required to perform any valid indirect comparison across studies.

CONDITIONALLY STRONG IGNORABILITY OF TRIAL ASSIGNMENT

Strongly ignorable trial assignment (specifically, assignment to the AC trial), conditional on the selected covariates, is the primary assumption underlying population-adjusted indirect comparisons and is required for unbiased estimation of $\Delta_{10}^{(2)}$. This is akin to the strongly ignorable sample or trial assignment assumption [61] commonly used in the generalizability, transportability or external validity literature [17, 59–61]. This literature seeks to calibrate relative treatment effects obtained from a RCT into a, more diverse, target population. In MAIC and the discussed outcome regression methods, the indirect comparison is performed in the BC population, and the A vs. C treatment effect is transported to the BC population. (Conditionally) strong ignorability consists of two assumptions: (conditional) ignorability and overlap (or positivity). Note that, even though strong ignorability has been proposed in the context of propensity score modeling, it is also a crucial assumption for the causal interpretation of outcome regression results in the BC population.

Conditional ignorability

There are many ways to articulate this assumption. One can consider that trial assignment/selection is conditionally ignorable, unconfounded or exchangeable for the A vs. C treatment effect (the potential A vs. C relative outcomes), i.e., conditionally independent of the treatment effect, given the selected effect modifiers. This means that after adjusting for these effect modifiers, treatment effect heterogeneity and trial assignment are conditionally independent. The NICE technical support document [9, 18] describes this assumption as the conditional constancy of relative effects across populations (namely, given the selected effect-modifying covariates, the A vs. C treatment effect is constant across populations).

MAIC will only meet conditional ignorability if *all* (observed or unobserved) effect modifiers are accounted for, regardless of whether these are balanced before the weighting. Excluding balanced covariates from the weighting procedure does not ensure balance after the weighting. The outcome regression methods meet conditional ignorability if all *imbalanced* effect modifiers are accounted for in the covariate-adjusted regression model (in the case of multiple imputation marginalization, that is the first-stage regression).

This is a demanding assumption in practice, which is also untestable. On one hand, it is tied to the measure used to define the treatment effects and effect modifiers. Most crucially, ignorability is hard to meet because it requires complete information on all treatment effect modifiers to be measured and available across trials AC and BC, and for all effect modifiers to be accounted for by the analyst. Firstly, it is conceivable that information on some effect modifiers is unavailable or unpublished in one or both studies. Secondly, the analyst may select the effect modifiers incorrectly. It is generally difficult to ascertain the effect modifier status of variables, particularly for new treatments with limited prior empirical evidence and clinical domain knowledge. We can never eliminate the possibility that this assumption is broken, as we cannot guarantee that there are no unobserved or unmeasured effect modifiers. Nevertheless, the careful selection of effect modifiers [136] from the observed baseline covariates is within the investigator's control and can provide some protection. Overspecification of effect modifiers should not bias the comparison but may inflate standard errors and lead to a subsequent loss of precision [18].

Effect modifier status is often determined by carrying out subgroup analyses in the IPD, or by examining statistical covariate-treatment interactions in outcome regressions fitted to the IPD [57, 272, 329]. The latter is a preferred approach and the former is typically discouraged [136]. Non-parametric tree-based regressions have recently been used for this purpose. These are appealing because they are data-driven and can detect interactions without pre-specifying which candidate variables to include in the model [330, 331]. Nevertheless, all statistical approaches are hindered by the lack of power of individual RCTs to identify interactions [91].

In the unanchored case, there is no common comparator group included in the analysis. Therefore, estimates are based on a comparison of within-trial absolute outcomes from single treatment arms, obtained from single-arm studies or individual arms of observational studies or RCTs, not on a comparison of within-trial relative effects. In this scenario, trial assignment

is ignorable if it is conditionally independent of the potential absolute outcomes given the covariates accounted for in the adjustment mechanism.

Here, we cannot draw a distinction between the predictors of outcome that are not treatment-specific (prognostic variables), i.e., associated with outcomes on follow-up regardless of the treatment provided, and factors that are associated with the outcome under a specific intervention because of interaction with treatment (effect modifiers). In fact, treatment effect modification cannot be quantified and is ill-defined. This is because its definition is reliant on contrasting outcomes between two groups, and there is no reference control group to define a relative treatment effect in the IPD trial.[4] The effect of purely prognostic variables and variables that are predictive of response to a specific treatment is conflated and cannot be disentangled. Therefore, unanchored comparisons rely on the assumption of no systematic cross-trial differences in predictors of the absolute outcome under treatment A, regardless of whether they have a purely prognostic role or are predictors of the response to treatment. The population adjustment methods only meet ignorability in the unanchored case if all variables that are prognostic of outcome under treatment A are balanced. In the unanchored case, ignorable trial assignment is equivalent to the conditional constancy of absolute effects described in the literature [9, 18].

Overlap

Conditional ignorability of trial assignment is strong if there is positivity or overlap, i.e., if every subject in the BC population has a positive probability of being assigned to the AC trial given the covariates accounted for in the adjustment mechanism. This implies that the ranges of the covariates in the BC population are covered by their respective ranges in the AC trial. This assumption may pose a problem if the inclusion/exclusion criteria of AC and BC are inconsistent. For instance, consider a situation where age is selected as an effect modifier and the age ranges of trial AC and trial BC are 60-70 and 40-70, respectively. There exists a subpopulation (age 40-60) in BC that does not overlap with the AC population. Hence, the AC study provides no evidence about the treatment effect and treatment effect modification in the excluded age group, and the A vs. C treatment effect estimate may be biased in the full comparator trial population (ages 40-70).

In such cases, reweighting methods like MAIC cannot extrapolate beyond the observed covariate space in the AC IPD, as there are no subjects to reweight. Where overlap is insufficient, outcome regression methods can extrapolate beyond the AC population, using the linearity assumption or other appropriate assumptions about the input space. However, valid extrapolation requires accurately capturing the true relationship between the covariates and the

4 If the IPD study is a single-arm trial, it is not possible to determine whether the outcomes are due to strong prognostic indicators or to the treatment itself and its interaction with effect modifiers, in excess of prognostic impact. Treatment effect modifiers cannot be identified because single-arm trials do not provide information about outcomes in a control group not receiving the intervention. Even if the index trial is an RCT comparing treatments A and C, effect modification for the relative effect of A vs C does not translate across studies if there is not a common comparator group. For instance, if the comparator study contrasts treatments B and control D, a covariate that modifies the effect of active intervention A with respect to C is not necessarily an effect modifier with respect to D.

outcome. Conversely, the exclusion of patients enrolled in AC from the BC population, e.g. if the AC population is more diverse, does not necessarily violate the overlap assumption. This is because these methods deliver estimates in the BC population. Hence, adjustment in this scenario is an interpolation as opposed to an extrapolation of the observed AC data. In this scenario, $\hat{\Delta}_{10}^{(2)}$ may be unbiased because the BC population is covered within that of AC. In MAIC, the excluded subpopulation will receive very low weights (low odds of enrolment in BC vs. AC), while the included subpopulation receives high weights and dominates the reweighted sample. These extreme weights lead to large reductions in ESS and to the deterioration of precision and efficiency. Removing observations from the AC patient-level data, so that inclusion/exclusion criteria are consistent, explicitly lowers the AC sample size and may degrade precision further. Of course, when there is no interpolation or extrapolation overlap whatsoever, MAIC cannot generate population-adjusted estimate for the treatment effect, as a feasible weighting solution does not exist due to separation problems [35].

If IPD were available for the BC study, the overlap assumption could be easily checked by visualizing the ranges of the selected covariates and their empirical distributions. However, this is challenging in our setup without further distributional assumptions due to patient-level data limitations for BC.

SPECIFICATION OF THE JOINT COVARIATE DISTRIBUTION IN BC

Population-adjusted indirect comparisons make certain assumptions to approximate the joint distribution of covariates in the BC trial. The restriction of limited IPD makes it unlikely that such joint distribution is available. Summary statistics for the marginal distributions are typically published instead. Where no correlation information is available for the BC study, MAIC and the conventional centered version of STC seem to assume that the joint BC covariate distribution is the product of the published marginal distributions. The implicit assumptions are, in fact, more nuanced and differ slightly between methods.

In MAIC, as stated in the NICE Decision Support Unit [18], "when covariate correlations are not available from the (BC) population, and therefore cannot be balanced by inclusion in the weighting model, they are assumed to be equal to the correlations amongst covariates in the pseudo-population formed by weighting the (AC) population." In typical usage, MAIC only balances the marginal distributions of the selected baseline covariates, not the multidimensional joint covariate distributions, due to the lack of published correlation data for BC. In the typical usage of STC (i.e., the "plug-in" approach to the method in Section 2.4), the assumption differs slightly. The correlations between the BC covariates are assumed to be equal to the correlations between covariates in the AC study.

In the novel outcome regression methods discussed in Chapter 4, (parametric G-computation and MIM), more explicit distributional assumptions are made to characterize the BC population in the "covariate simulation" step. The methods assume the joint distribution of the BC covariates is specified correctly, by the combination of the specified marginal distributions and correlation structure. In the simulation study in Chapter 5, pseudo-populations are constructed

under certain parametric assumptions. We have assumed that the pairwise correlations of the covariates and the parametric forms of their marginal distributions are identical across trials, because the correlation structure observed in the AC IPD is used in the "covariate simulation" step. These assumptions cannot be verified empirically as we have no information on the covariates' correlation structure and true marginal distributions. Information on correlations or on the joint distribution of covariates for BC is rarely published, but could be requested. We have decided to mimic the AC pairwise correlations as, in principle, the relationships between covariates should be similar across trials.

MODELING ASSUMPTIONS

Indirect treatment comparisons are typically conducted on the linear predictor scale [18], upon which the treatment effect is assumed to be additive for all indirect comparisons. In the main text, the anchored population-adjusted indirect comparisons have additionally assumed that the effect modifiers have been defined on the linear predictor scale and are additive on this scale, but the linearity assumption is not always appropriate. Hence, all population adjustment methods are subject to scale conflicts or to bias if effect modification status, which is scale-specific, has been justified on the wrong scale, e.g. when treatment effect modification is specified as linear but is non-linear or multiplicative, e.g. age in cardiovascular disease treatments.

This form of model misspecification is more evident in the outcome modeling approaches, where an explicit outcome regression is formulated. The parametric model depends on functional form assumptions that will be violated if the relationship between the covariates and the outcome is not captured correctly, in which case the methods may be biased. Even though the logistic regression model for the weights in MAIC does not make reference to the outcome, the method is also susceptible to model misspecification bias, albeit in a more implicit form. The model for estimating the weights is approximately correct in the simulation studies because the right subset of covariates has been selected as effect modifiers and the balancing property holds for the weights, as mentioned in subsection 3.1.4. In practice, the model will be incorrectly specified if this is not the case, potentially leading to a biased estimate. Note that, in practice, we find that it may be more difficult to specify a correct parametric model for the outcome than an approximately correct parametric model for the trial assignment weights.

CONCLUDING REMARKS

In practice, some of the assumptions above may be hard to meet. If these are violated, the resulting treatment effect may be biased. Hence, it is important to assess the robustness of the methods to failures of assumptions and under different degrees of model misspecification in future simulation studies.

SUPPLEMENTARY APPENDIX B: EXTENSION OF MULTIPLE IMPUTATION MARGINALIZATION TO MULTI-COMPONENT ESTIMANDS

In many applications, the target estimand is non-scalar and has multiple components. For instance, this is the case where the outcome model is a multivariate regression (i.e., with multiple dependent variables) of correlated outcomes with treatment. This scenario typically evaluates surrogate endpoints and involves combining correlated treatment effects corresponding to multiple outcomes [203]. With non-scalar or multivariate estimands, the pooling stage must propagate the covariance or correlation structure of treatment effects through the analysis. Therefore, the inferential framework outlined in subsection 4.6.2.2 is extended by Reiter [332]. The column vector of treatment effect point estimates for the m-th synthesis is denoted $\hat{\boldsymbol{\xi}}^{(m)}$ and has $J \geq 2$ components. The estimated $J \times J$ covariance matrix of treatment effect estimates for the m-th synthesis is denoted $\hat{\boldsymbol{v}}^{(m)}$. Analogous to Equations 25, 26 and 27, the following multivariate quantities are required for inference:

$$\bar{\boldsymbol{\xi}} = \sum_{m=1}^{M} \hat{\boldsymbol{\xi}}^{(m)} / M, \tag{34}$$

$$\bar{\boldsymbol{v}} = \sum_{m=1}^{M} \hat{\boldsymbol{v}}^{(m)} / M, \tag{35}$$

$$\boldsymbol{b} = \sum_{m=1}^{M} (\hat{\boldsymbol{\xi}}^{(m)} - \bar{\boldsymbol{\xi}})(\hat{\boldsymbol{\xi}}^{(m)} - \bar{\boldsymbol{\xi}})^{\top} / (M-1). \tag{36}$$

Here, $\bar{\boldsymbol{\xi}}$ is a vector of size J of average treatment effect point estimates across the M syntheses, $\bar{\boldsymbol{v}}$ is a $J \times J$ matrix of the average estimated covariance matrices, and $\boldsymbol{b}$ is the $J \times J$ sample covariance matrix of the treatment effect point estimates.

The target estimands for inference are the average marginal treatment effects in the BC population, denoted by vector $\boldsymbol{\Xi}^{(2)}$. The posterior distribution $p(\boldsymbol{\Xi}^{(2)} \mid \boldsymbol{y}^*, \boldsymbol{z}^*)$ is assumed to be approximately multivariate normal and is constructed as:

$$p(\boldsymbol{\Xi}^{(2)} \mid \boldsymbol{y}^*, \boldsymbol{z}^*) = \int_{\boldsymbol{\mu}_{\boldsymbol{\Xi}}, \boldsymbol{\Sigma}_{\boldsymbol{\Xi}}} p(\boldsymbol{\Xi}^{(2)} \mid \boldsymbol{\mu}_{\boldsymbol{\Xi}}, \boldsymbol{\Sigma}_{\boldsymbol{\Xi}}) p(\boldsymbol{\mu}_{\boldsymbol{\Xi}}, \boldsymbol{\Sigma}_{\boldsymbol{\Xi}} \mid \boldsymbol{y}^*, \boldsymbol{z}^*) d(\boldsymbol{\mu}_{\boldsymbol{\Xi}}, \boldsymbol{\Sigma}_{\boldsymbol{\Xi}}), \tag{37}$$

with the posterior density parametrized by two moments: a vector of means $\boldsymbol{\mu}_{\boldsymbol{\Xi}}$ and a $J \times J$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\Xi}}$. After deriving the quantities in Equations 34, 35, 36, the posterior in Equation 37 is approximated by the following distributions (by analogy to Equations 28-30):

$$p(\boldsymbol{\mu}_{\boldsymbol{\Xi}} \mid \boldsymbol{y}^*, \boldsymbol{z}^*) \sim \mathrm{N}(\bar{\boldsymbol{\xi}}, \bar{\boldsymbol{v}} / M), \tag{38}$$

$$p\left((M-1)\boldsymbol{b}/(\boldsymbol{\Sigma}_{\Xi}+\bar{\boldsymbol{v}}) \mid \boldsymbol{y}^{*}, z^{*}\right) \sim \text{Wishart}_{M-1}, \tag{39}$$

$$p(\Xi^{(2)} \mid \boldsymbol{\mu}_{\Xi}, \boldsymbol{\Sigma}_{\Xi}) \sim t_{M-1}\left(\boldsymbol{\mu}_{\Xi}, (1+1/M)\boldsymbol{\Sigma}_{\Xi}\right). \tag{40}$$

Note that the division in the left-hand side of Equation 39 is an element-wise (Hadamard) division. One can approximate the integral of the posterior in Equation 40 with respect to the posteriors in Equations 38 and 39 via simulation. However, it is considerably simpler to use a multivariate normal approximation to the posterior density in Equation 37, with means $\bar{\boldsymbol{\xi}}$ and covariance $(1+1/M)\boldsymbol{b}-\bar{\boldsymbol{v}}$, such that the sampling distribution in Equation 40 is normal. This yields the following combining rules [332], used to derive point estimates, $\hat{\Xi}^{(2)}$ and $\widehat{Cov}(\hat{\Xi}^{(2)})$, for the average marginal treatment effects in the BC population and their covariance matrix, respectively:

$$\hat{\Xi}^{(2)} = \bar{\boldsymbol{\xi}}, \tag{41}$$

$$\widehat{Cov}(\hat{\Xi}^{(2)}) = (1+1/M)\boldsymbol{b}-\bar{\boldsymbol{v}}. \tag{42}$$

The plug-in estimators in Equations 41 and 42 are valid for reasonably large M.

SUPPLEMENTARY APPENDIX C: CHAPTER 3 SIMULATION STUDY

SIMULATION STUDY SCENARIO SETTINGS

In Table 2, parameter values for each simulation scenario are presented.

Table 2: Parameter values for the simulation study scenarios.

Scenario	Number of subjects in *AC*	Prognostic effect	Interaction effect	Mean of *AC* covariates	Covariate correlation
1	150	0.40	0.40	0.45	0.00
2	300	0.40	0.40	0.45	0.00
3	600	0.40	0.40	0.45	0.00
4	150	0.69	0.40	0.45	0.00
5	300	0.69	0.40	0.45	0.00
6	600	0.69	0.40	0.45	0.00
7	150	1.11	0.40	0.45	0.00
8	300	1.11	0.40	0.45	0.00
9	600	1.11	0.40	0.45	0.00
10	150	0.40	0.69	0.45	0.00
11	300	0.40	0.69	0.45	0.00
12	600	0.40	0.69	0.45	0.00
13	150	0.69	0.69	0.45	0.00
14	300	0.69	0.69	0.45	0.00
15	600	0.69	0.69	0.45	0.00
16	150	1.11	0.69	0.45	0.00
17	300	1.11	0.69	0.45	0.00
18	600	1.11	0.69	0.45	0.00
19	150	0.40	1.11	0.45	0.00
20	300	0.40	1.11	0.45	0.00
21	600	0.40	1.11	0.45	0.00
22	150	0.69	1.11	0.45	0.00
23	300	0.69	1.11	0.45	0.00
24	600	0.69	1.11	0.45	0.00
25	150	1.11	1.11	0.45	0.00
26	300	1.11	1.11	0.45	0.00
27	600	1.11	1.11	0.45	0.00
28	150	0.40	0.40	0.45	0.35
29	300	0.40	0.40	0.45	0.35
30	600	0.40	0.40	0.45	0.35
31	150	0.69	0.40	0.45	0.35
32	300	0.69	0.40	0.45	0.35
33	600	0.69	0.40	0.45	0.35
34	150	1.11	0.40	0.45	0.35
35	300	1.11	0.40	0.45	0.35
36	600	1.11	0.40	0.45	0.35
37	150	0.40	0.69	0.45	0.35
38	300	0.40	0.69	0.45	0.35
39	600	0.40	0.69	0.45	0.35
40	150	0.69	0.69	0.45	0.35
41	300	0.69	0.69	0.45	0.35
42	600	0.69	0.69	0.45	0.35
43	150	1.11	0.69	0.45	0.35

Table 2: Parameter values for the simulation study scenarios. *(continued)*

Scenario	Number of subjects in *AC*	Prognostic effect	Interaction effect	Mean of *AC* covariates	Covariate correlation
44	300	1.11	0.69	0.45	0.35
45	600	1.11	0.69	0.45	0.35
46	150	0.40	1.11	0.45	0.35
47	300	0.40	1.11	0.45	0.35
48	600	0.40	1.11	0.45	0.35
49	150	0.69	1.11	0.45	0.35
50	300	0.69	1.11	0.45	0.35
51	600	0.69	1.11	0.45	0.35
52	150	1.11	1.11	0.45	0.35
53	300	1.11	1.11	0.45	0.35
54	600	1.11	1.11	0.45	0.35
55	150	0.40	0.40	0.30	0.00
56	300	0.40	0.40	0.30	0.00
57	600	0.40	0.40	0.30	0.00
58	150	0.69	0.40	0.30	0.00
59	300	0.69	0.40	0.30	0.00
60	600	0.69	0.40	0.30	0.00
61	150	1.11	0.40	0.30	0.00
62	300	1.11	0.40	0.30	0.00
63	600	1.11	0.40	0.30	0.00
64	150	0.40	0.69	0.30	0.00
65	300	0.40	0.69	0.30	0.00
66	600	0.40	0.69	0.30	0.00
67	150	0.69	0.69	0.30	0.00
68	300	0.69	0.69	0.30	0.00
69	600	0.69	0.69	0.30	0.00
70	150	1.11	0.69	0.30	0.00
71	300	1.11	0.69	0.30	0.00
72	600	1.11	0.69	0.30	0.00
73	150	0.40	1.11	0.30	0.00
74	300	0.40	1.11	0.30	0.00
75	600	0.40	1.11	0.30	0.00
76	150	0.69	1.11	0.30	0.00
77	300	0.69	1.11	0.30	0.00
78	600	0.69	1.11	0.30	0.00
79	150	1.11	1.11	0.30	0.00
80	300	1.11	1.11	0.30	0.00
81	600	1.11	1.11	0.30	0.00
82	150	0.40	0.40	0.30	0.35
83	300	0.40	0.40	0.30	0.35
84	600	0.40	0.40	0.30	0.35
85	150	0.69	0.40	0.30	0.35
86	300	0.69	0.40	0.30	0.35
87	600	0.69	0.40	0.30	0.35
88	150	1.11	0.40	0.30	0.35
89	300	1.11	0.40	0.30	0.35
90	600	1.11	0.40	0.30	0.35
91	150	0.40	0.69	0.30	0.35
92	300	0.40	0.69	0.30	0.35
93	600	0.40	0.69	0.30	0.35
94	150	0.69	0.69	0.30	0.35
95	300	0.69	0.69	0.30	0.35
96	600	0.69	0.69	0.30	0.35
97	150	1.11	0.69	0.30	0.35
98	300	1.11	0.69	0.30	0.35
99	600	1.11	0.69	0.30	0.35

Table 2: Parameter values for the simulation study scenarios. *(continued)*

Scenario	Number of subjects in *AC*	Prognostic effect	Interaction effect	Mean of *AC* covariates	Covariate correlation
100	150	0.40	1.11	0.30	0.35
101	300	0.40	1.11	0.30	0.35
102	600	0.40	1.11	0.30	0.35
103	150	0.69	1.11	0.30	0.35
104	300	0.69	1.11	0.30	0.35
105	600	0.69	1.11	0.30	0.35
106	150	1.11	1.11	0.30	0.35
107	300	1.11	1.11	0.30	0.35
108	600	1.11	1.11	0.30	0.35
109	150	0.40	0.40	0.15	0.00
110	300	0.40	0.40	0.15	0.00
111	600	0.40	0.40	0.15	0.00
112	150	0.69	0.40	0.15	0.00
113	300	0.69	0.40	0.15	0.00
114	600	0.69	0.40	0.15	0.00
115	150	1.11	0.40	0.15	0.00
116	300	1.11	0.40	0.15	0.00
117	600	1.11	0.40	0.15	0.00
118	150	0.40	0.69	0.15	0.00
119	300	0.40	0.69	0.15	0.00
120	600	0.40	0.69	0.15	0.00
121	150	0.69	0.69	0.15	0.00
122	300	0.69	0.69	0.15	0.00
123	600	0.69	0.69	0.15	0.00
124	150	1.11	0.69	0.15	0.00
125	300	1.11	0.69	0.15	0.00
126	600	1.11	0.69	0.15	0.00
127	150	0.40	1.11	0.15	0.00
128	300	0.40	1.11	0.15	0.00
129	600	0.40	1.11	0.15	0.00
130	150	0.69	1.11	0.15	0.00
131	300	0.69	1.11	0.15	0.00
132	600	0.69	1.11	0.15	0.00
133	150	1.11	1.11	0.15	0.00
134	300	1.11	1.11	0.15	0.00
135	600	1.11	1.11	0.15	0.00
136	150	0.40	0.40	0.15	0.35
137	300	0.40	0.40	0.15	0.35
138	600	0.40	0.40	0.15	0.35
139	150	0.69	0.40	0.15	0.35
140	300	0.69	0.40	0.15	0.35
141	600	0.69	0.40	0.15	0.35
142	150	1.11	0.40	0.15	0.35
143	300	1.11	0.40	0.15	0.35
144	600	1.11	0.40	0.15	0.35
145	150	0.40	0.69	0.15	0.35
146	300	0.40	0.69	0.15	0.35
147	600	0.40	0.69	0.15	0.35
148	150	0.69	0.69	0.15	0.35
149	300	0.69	0.69	0.15	0.35
150	600	0.69	0.69	0.15	0.35
151	150	1.11	0.69	0.15	0.35
152	300	1.11	0.69	0.15	0.35
153	600	1.11	0.69	0.15	0.35
154	150	0.40	1.11	0.15	0.35
155	300	0.40	1.11	0.15	0.35

Table 2: Parameter values for the simulation study scenarios. *(continued)*

Scenario	Number of subjects in *AC*	Prognostic effect	Interaction effect	Mean of *AC* covariates	Covariate correlation
156	600	0.40	1.11	0.15	0.35
157	150	0.69	1.11	0.15	0.35
158	300	0.69	1.11	0.15	0.35
159	600	0.69	1.11	0.15	0.35
160	150	1.11	1.11	0.15	0.35
161	300	1.11	1.11	0.15	0.35
162	600	1.11	1.11	0.15	0.35

SIMULATION STUDY RESULTS

Table 3 displays the key performance measures associated with each of the simulated scenarios. Biases are displayed in red when their absolute size is greater than one half of the treatment effect estimate's empirical standard error. Coverage rates are presented in red when these are statistically significantly different to 0.95; namely, if the rate is less than 0.9365 or more than 0.9635. Monte Carlo standard errors for each measure are presented in parentheses.

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison.

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
1	MAIC	-0.010 (0.007)	-0.429 (0.007)	0.410 (0.007)	0.046 (0.002)	0.170 (0.007)	0.950 (0.007)	1.001 (0.022)	0.214 (0.005)
1	STC	-0.065 (0.007)	-0.513 (0.008)	0.383 (0.007)	0.059 (0.003)	0.195 (0.007)	0.938 (0.008)	0.974 (0.022)	0.235 (0.005)
1	Bucher	-0.095 (0.007)	-0.504 (0.007)	0.315 (0.006)	0.054 (0.002)	0.187 (0.007)	0.927 (0.008)	0.985 (0.022)	0.212 (0.005)
2	MAIC	0.010 (0.005)	-0.310 (0.005)	0.331 (0.005)	0.026 (0.001)	0.129 (0.005)	0.955 (0.007)	1.020 (0.023)	0.160 (0.004)
2	STC	-0.034 (0.005)	-0.369 (0.005)	0.300 (0.005)	0.030 (0.001)	0.140 (0.005)	0.950 (0.007)	1.006 (0.023)	0.170 (0.004)
2	Bucher	-0.075 (0.005)	-0.390 (0.005)	0.239 (0.005)	0.031 (0.001)	0.141 (0.005)	0.928 (0.008)	1.007 (0.023)	0.159 (0.004)
3	MAIC	-0.009 (0.004)	-0.267 (0.004)	0.250 (0.004)	0.018 (0.001)	0.108 (0.004)	0.939 (0.008)	0.975 (0.022)	0.135 (0.003)
3	STC	-0.046 (0.004)	-0.312 (0.004)	0.220 (0.004)	0.022 (0.001)	0.120 (0.004)	0.924 (0.008)	0.964 (0.022)	0.141 (0.003)
3	Bucher	-0.092 (0.004)	-0.347 (0.004)	0.162 (0.004)	0.026 (0.001)	0.129 (0.004)	0.883 (0.010)	0.977 (0.022)	0.133 (0.003)
4	MAIC	-0.007 (0.007)	-0.407 (0.007)	0.393 (0.006)	0.043 (0.002)	0.163 (0.007)	0.940 (0.008)	0.981 (0.022)	0.208 (0.005)
4	STC	-0.111 (0.007)	-0.549 (0.007)	0.328 (0.007)	0.065 (0.003)	0.200 (0.007)	0.918 (0.009)	0.974 (0.022)	0.230 (0.005)
4	Bucher	-0.081 (0.007)	-0.478 (0.007)	0.316 (0.006)	0.049 (0.002)	0.176 (0.007)	0.935 (0.008)	0.983 (0.022)	0.206 (0.005)
5	MAIC	0.006 (0.005)	-0.301 (0.005)	0.312 (0.005)	0.024 (0.001)	0.122 (0.005)	0.941 (0.007)	1.003 (0.022)	0.156 (0.003)
5	STC	-0.093 (0.005)	-0.419 (0.005)	0.234 (0.005)	0.037 (0.002)	0.152 (0.005)	0.913 (0.009)	0.986 (0.022)	0.169 (0.004)
5	Bucher	-0.072 (0.005)	-0.378 (0.005)	0.233 (0.005)	0.030 (0.001)	0.136 (0.005)	0.920 (0.009)	1.000 (0.022)	0.156 (0.003)
6	MAIC	0.002 (0.004)	-0.247 (0.004)	0.251 (0.004)	0.016 (0.001)	0.100 (0.004)	0.943 (0.007)	1.001 (0.022)	0.127 (0.003)
6	STC	-0.087 (0.004)	-0.346 (0.004)	0.173 (0.004)	0.025 (0.001)	0.127 (0.004)	0.901 (0.009)	1.001 (0.022)	0.132 (0.003)
6	Bucher	-0.076 (0.004)	-0.324 (0.004)	0.172 (0.004)	0.022 (0.001)	0.119 (0.004)	0.910 (0.009)	1.000 (0.022)	0.127 (0.003)
7	MAIC	-0.003 (0.006)	-0.386 (0.006)	0.379 (0.006)	0.039 (0.002)	0.157 (0.006)	0.947 (0.007)	0.991 (0.022)	0.197 (0.004)
7	STC	-0.184 (0.007)	-0.614 (0.007)	0.246 (0.007)	0.082 (0.004)	0.233 (0.007)	0.873 (0.011)	0.995 (0.022)	0.221 (0.005)
7	Bucher	-0.073 (0.006)	-0.459 (0.006)	0.313 (0.006)	0.044 (0.002)	0.165 (0.006)	0.936 (0.008)	1.000 (0.022)	0.197 (0.004)
8	MAIC	0.000 (0.005)	-0.295 (0.005)	0.294 (0.005)	0.023 (0.001)	0.122 (0.005)	0.954 (0.007)	0.995 (0.022)	0.151 (0.003)
8	STC	-0.174 (0.005)	-0.494 (0.005)	0.146 (0.005)	0.058 (0.002)	0.197 (0.005)	0.817 (0.012)	0.982 (0.022)	0.166 (0.004)
8	Bucher	-0.068 (0.005)	-0.364 (0.005)	0.229 (0.005)	0.028 (0.001)	0.133 (0.005)	0.925 (0.008)	0.987 (0.022)	0.153 (0.003)
9	MAIC	0.003 (0.004)	-0.237 (0.004)	0.242 (0.004)	0.016 (0.001)	0.102 (0.004)	0.951 (0.007)	0.965 (0.022)	0.127 (0.003)
9	STC	-0.166 (0.004)	-0.420 (0.004)	0.088 (0.004)	0.046 (0.002)	0.179 (0.004)	0.740 (0.014)	0.966 (0.022)	0.134 (0.003)
9	Bucher	-0.063 (0.004)	-0.303 (0.004)	0.178 (0.004)	0.021 (0.001)	0.116 (0.004)	0.905 (0.009)	0.937 (0.021)	0.131 (0.003)
10	MAIC	0.003 (0.006)	-0.399 (0.006)	0.406 (0.006)	0.040 (0.002)	0.160 (0.006)	0.954 (0.007)	1.023 (0.023)	0.201 (0.004)
10	STC	-0.007 (0.007)	-0.443 (0.007)	0.429 (0.007)	0.050 (0.002)	0.179 (0.007)	0.948 (0.007)	0.993 (0.022)	0.224 (0.005)
10	Bucher	-0.138 (0.006)	-0.538 (0.006)	0.261 (0.006)	0.058 (0.003)	0.192 (0.006)	0.905 (0.009)	1.037 (0.023)	0.197 (0.004)
11	MAIC	0.003 (0.005)	-0.305 (0.005)	0.312 (0.005)	0.024 (0.001)	0.123 (0.005)	0.953 (0.007)	1.021 (0.023)	0.154 (0.003)
11	STC	0.003 (0.005)	-0.323 (0.005)	0.329 (0.005)	0.028 (0.001)	0.134 (0.005)	0.950 (0.007)	0.995 (0.022)	0.167 (0.004)
11	Bucher	-0.139 (0.005)	-0.447 (0.005)	0.168 (0.005)	0.044 (0.002)	0.172 (0.005)	0.860 (0.011)	1.001 (0.022)	0.157 (0.004)
12	MAIC	0.000 (0.004)	-0.250 (0.004)	0.250 (0.004)	0.016 (0.001)	0.101 (0.004)	0.964 (0.006)	1.018 (0.023)	0.125 (0.003)
12	STC	0.003 (0.004)	-0.257 (0.004)	0.262 (0.004)	0.017 (0.001)	0.106 (0.004)	0.964 (0.006)	1.010 (0.023)	0.131 (0.003)
12	Bucher	-0.140 (0.004)	-0.389 (0.004)	0.110 (0.004)	0.035 (0.001)	0.156 (0.004)	0.804 (0.013)	1.013 (0.023)	0.126 (0.003)
13	MAIC	-0.001 (0.006)	-0.389 (0.006)	0.387 (0.006)	0.038 (0.002)	0.155 (0.006)	0.948 (0.007)	1.011 (0.023)	0.196 (0.004)
13	STC	-0.040 (0.007)	-0.468 (0.007)	0.387 (0.007)	0.050 (0.002)	0.178 (0.007)	0.945 (0.007)	0.990 (0.022)	0.220 (0.005)
13	Bucher	-0.131 (0.006)	-0.522 (0.007)	0.261 (0.006)	0.057 (0.003)	0.190 (0.006)	0.900 (0.009)	0.996 (0.022)	0.200 (0.004)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
14	MAIC	0.002 (0.005)	-0.296 (0.005)	0.300 (0.005)	0.022 (0.001)	0.117 (0.005)	0.955 (0.007)	1.032 (0.023)	0.147 (0.003)
14	STC	-0.024 (0.005)	-0.344 (0.005)	0.295 (0.005)	0.027 (0.001)	0.131 (0.005)	0.951 (0.007)	1.001 (0.022)	0.163 (0.004)
14	Bucher	-0.127 (0.005)	-0.427 (0.005)	0.174 (0.005)	0.039 (0.002)	0.162 (0.005)	0.878 (0.010)	1.001 (0.022)	0.153 (0.003)
15	MAIC	0.005 (0.004)	-0.237 (0.004)	0.248 (0.004)	0.015 (0.001)	0.097 (0.004)	0.949 (0.007)	1.013 (0.023)	0.122 (0.003)
15	STC	-0.019 (0.004)	-0.274 (0.004)	0.235 (0.004)	0.017 (0.001)	0.106 (0.004)	0.954 (0.007)	1.002 (0.022)	0.130 (0.003)
15	Bucher	-0.120 (0.004)	-0.364 (0.004)	0.124 (0.004)	0.030 (0.001)	0.141 (0.004)	0.839 (0.012)	1.006 (0.023)	0.124 (0.003)
16	MAIC	-0.010 (0.006)	-0.385 (0.006)	0.366 (0.006)	0.038 (0.002)	0.155 (0.006)	0.944 (0.007)	0.980 (0.022)	0.195 (0.004)
16	STC	-0.082 (0.007)	-0.503 (0.007)	0.339 (0.007)	0.059 (0.003)	0.190 (0.007)	0.915 (0.009)	0.942 (0.021)	0.228 (0.005)
16	Bucher	-0.119 (0.006)	-0.501 (0.006)	0.264 (0.006)	0.054 (0.002)	0.185 (0.006)	0.904 (0.009)	0.972 (0.022)	0.201 (0.004)
17	MAIC	-0.001 (0.005)	-0.289 (0.005)	0.288 (0.005)	0.023 (0.001)	0.123 (0.005)	0.942 (0.007)	0.964 (0.022)	0.153 (0.003)
17	STC	-0.070 (0.005)	-0.385 (0.005)	0.244 (0.005)	0.034 (0.002)	0.146 (0.005)	0.908 (0.009)	0.942 (0.021)	0.170 (0.004)
17	Bucher	-0.109 (0.005)	-0.403 (0.005)	0.184 (0.005)	0.036 (0.001)	0.154 (0.005)	0.870 (0.011)	0.963 (0.022)	0.156 (0.003)
18	MAIC	0.005 (0.004)	-0.230 (0.004)	0.241 (0.004)	0.014 (0.001)	0.093 (0.004)	0.955 (0.007)	1.033 (0.023)	0.116 (0.003)
18	STC	-0.063 (0.004)	-0.313 (0.004)	0.187 (0.004)	0.020 (0.001)	0.113 (0.004)	0.927 (0.008)	1.021 (0.023)	0.125 (0.003)
18	Bucher	-0.104 (0.004)	-0.342 (0.004)	0.135 (0.004)	0.025 (0.001)	0.128 (0.004)	0.881 (0.010)	1.034 (0.023)	0.118 (0.003)
19	MAIC	0.000 (0.006)	-0.386 (0.006)	0.387 (0.006)	0.039 (0.002)	0.157 (0.006)	0.948 (0.007)	0.995 (0.022)	0.198 (0.004)
19	STC	0.095 (0.007)	-0.332 (0.007)	0.522 (0.007)	0.062 (0.003)	0.198 (0.007)	0.911 (0.009)	0.947 (0.021)	0.230 (0.005)
19	Bucher	-0.204 (0.006)	-0.596 (0.007)	0.187 (0.006)	0.083 (0.003)	0.239 (0.006)	0.830 (0.012)	0.983 (0.022)	0.203 (0.005)
20	MAIC	0.012 (0.005)	-0.285 (0.005)	0.309 (0.005)	0.021 (0.001)	0.116 (0.005)	0.960 (0.006)	1.048 (0.023)	0.145 (0.003)
20	STC	0.107 (0.005)	-0.212 (0.005)	0.427 (0.005)	0.036 (0.002)	0.154 (0.005)	0.907 (0.009)	1.034 (0.023)	0.158 (0.004)
20	Bucher	-0.194 (0.005)	-0.495 (0.005)	0.107 (0.005)	0.061 (0.002)	0.207 (0.005)	0.777 (0.013)	1.012 (0.023)	0.152 (0.003)
21	MAIC	0.002 (0.004)	-0.240 (0.004)	0.243 (0.004)	0.014 (0.001)	0.095 (0.004)	0.950 (0.007)	1.029 (0.023)	0.120 (0.003)
21	STC	0.102 (0.004)	-0.152 (0.004)	0.356 (0.004)	0.027 (0.001)	0.135 (0.004)	0.897 (0.010)	1.007 (0.023)	0.129 (0.003)
21	Bucher	-0.203 (0.004)	-0.448 (0.004)	0.041 (0.004)	0.056 (0.002)	0.208 (0.004)	0.646 (0.015)	1.015 (0.023)	0.123 (0.003)
22	MAIC	-0.009 (0.006)	-0.384 (0.006)	0.366 (0.006)	0.036 (0.002)	0.152 (0.006)	0.947 (0.007)	1.006 (0.023)	0.190 (0.004)
22	STC	0.093 (0.007)	-0.327 (0.007)	0.513 (0.007)	0.058 (0.003)	0.191 (0.007)	0.913 (0.009)	0.965 (0.022)	0.222 (0.005)
22	Bucher	-0.186 (0.006)	-0.571 (0.006)	0.199 (0.006)	0.071 (0.003)	0.220 (0.006)	0.849 (0.011)	1.026 (0.023)	0.191 (0.004)
23	MAIC	0.007 (0.004)	-0.282 (0.004)	0.297 (0.004)	0.020 (0.001)	0.111 (0.004)	0.953 (0.007)	1.057 (0.024)	0.140 (0.003)
23	STC	0.117 (0.005)	-0.197 (0.005)	0.431 (0.005)	0.039 (0.002)	0.161 (0.005)	0.897 (0.010)	1.010 (0.023)	0.159 (0.004)
23	Bucher	-0.179 (0.005)	-0.475 (0.005)	0.117 (0.005)	0.053 (0.002)	0.194 (0.005)	0.792 (0.013)	1.035 (0.023)	0.146 (0.003)
24	MAIC	0.005 (0.004)	-0.231 (0.004)	0.241 (0.004)	0.014 (0.001)	0.093 (0.004)	0.958 (0.006)	1.035 (0.023)	0.116 (0.003)
24	STC	0.116 (0.004)	-0.135 (0.004)	0.366 (0.004)	0.029 (0.001)	0.139 (0.004)	0.851 (0.011)	1.032 (0.023)	0.124 (0.003)
24	Bucher	-0.182 (0.004)	-0.422 (0.004)	0.059 (0.004)	0.047 (0.002)	0.188 (0.004)	0.695 (0.015)	1.023 (0.023)	0.120 (0.003)
25	MAIC	0.004 (0.006)	-0.362 (0.006)	0.371 (0.006)	0.033 (0.001)	0.145 (0.006)	0.956 (0.006)	1.022 (0.023)	0.183 (0.004)
25	STC	0.103 (0.007)	-0.311 (0.007)	0.518 (0.007)	0.057 (0.003)	0.191 (0.007)	0.911 (0.009)	0.982 (0.022)	0.216 (0.005)
25	Bucher	-0.157 (0.006)	-0.536 (0.006)	0.221 (0.006)	0.062 (0.002)	0.203 (0.006)	0.880 (0.010)	1.004 (0.022)	0.192 (0.004)
26	MAIC	0.005 (0.005)	-0.278 (0.005)	0.289 (0.004)	0.020 (0.001)	0.113 (0.005)	0.952 (0.007)	1.013 (0.023)	0.143 (0.003)
26	STC	0.114 (0.005)	-0.196 (0.005)	0.425 (0.005)	0.039 (0.002)	0.162 (0.005)	0.881 (0.010)	0.975 (0.022)	0.163 (0.004)
26	Bucher	-0.155 (0.005)	-0.446 (0.005)	0.137 (0.005)	0.045 (0.002)	0.176 (0.005)	0.835 (0.012)	1.034 (0.023)	0.144 (0.003)
27	MAIC	-0.001 (0.004)	-0.233 (0.004)	0.230 (0.004)	0.014 (0.001)	0.094 (0.004)	0.947 (0.007)	0.991 (0.022)	0.119 (0.003)
27	STC	0.104 (0.004)	-0.143 (0.004)	0.351 (0.004)	0.028 (0.001)	0.134 (0.004)	0.867 (0.011)	0.977 (0.022)	0.129 (0.003)
27	Bucher	-0.161 (0.004)	-0.398 (0.004)	0.076 (0.004)	0.040 (0.001)	0.171 (0.004)	0.734 (0.014)	1.005 (0.022)	0.120 (0.003)
28	MAIC	-0.006 (0.007)	-0.409 (0.007)	0.397 (0.006)	0.044 (0.002)	0.167 (0.007)	0.937 (0.008)	0.978 (0.022)	0.210 (0.005)
28	STC	-0.089 (0.007)	-0.527 (0.008)	0.349 (0.007)	0.062 (0.003)	0.200 (0.007)	0.925 (0.008)	0.959 (0.021)	0.233 (0.005)
28	Bucher	-0.085 (0.007)	-0.494 (0.007)	0.323 (0.007)	0.052 (0.002)	0.181 (0.007)	0.935 (0.008)	0.981 (0.022)	0.212 (0.005)
29	MAIC	-0.002 (0.005)	-0.312 (0.005)	0.309 (0.005)	0.024 (0.001)	0.124 (0.005)	0.955 (0.007)	1.023 (0.023)	0.155 (0.003)
29	STC	-0.071 (0.005)	-0.400 (0.005)	0.257 (0.005)	0.032 (0.001)	0.144 (0.005)	0.936 (0.008)	1.019 (0.023)	0.164 (0.004)
29	Bucher	-0.081 (0.005)	-0.394 (0.005)	0.233 (0.005)	0.031 (0.001)	0.141 (0.005)	0.927 (0.008)	1.027 (0.023)	0.156 (0.003)
30	MAIC	0.000 (0.004)	-0.253 (0.004)	0.252 (0.004)	0.016 (0.001)	0.100 (0.004)	0.955 (0.007)	1.012 (0.023)	0.127 (0.003)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
30	STC	-0.061 (0.004)	-0.324 (0.004)	0.201 (0.004)	0.022 (0.001)	0.116 (0.004)	0.918 (0.009)	0.991 (0.022)	0.135 (0.003)
30	Bucher	-0.079 (0.004)	-0.333 (0.004)	0.176 (0.004)	0.023 (0.001)	0.121 (0.004)	0.912 (0.009)	1.009 (0.023)	0.129 (0.003)
31	MAIC	-0.012 (0.006)	-0.397 (0.006)	0.373 (0.006)	0.040 (0.002)	0.159 (0.006)	0.949 (0.007)	0.984 (0.022)	0.199 (0.004)
31	STC	-0.170 (0.007)	-0.599 (0.007)	0.258 (0.007)	0.077 (0.003)	0.224 (0.007)	0.885 (0.010)	0.998 (0.022)	0.219 (0.005)
31	Bucher	-0.082 (0.006)	-0.479 (0.007)	0.314 (0.006)	0.047 (0.002)	0.172 (0.006)	0.932 (0.008)	1.004 (0.022)	0.202 (0.005)
32	MAIC	0.000 (0.005)	-0.297 (0.005)	0.296 (0.005)	0.023 (0.001)	0.121 (0.005)	0.946 (0.007)	0.994 (0.022)	0.152 (0.003)
32	STC	-0.149 (0.005)	-0.469 (0.005)	0.171 (0.005)	0.049 (0.002)	0.184 (0.005)	0.859 (0.011)	0.993 (0.022)	0.165 (0.004)
32	Bucher	-0.069 (0.005)	-0.373 (0.005)	0.236 (0.005)	0.028 (0.001)	0.133 (0.005)	0.928 (0.008)	1.008 (0.023)	0.154 (0.003)
33	MAIC	0.002 (0.004)	-0.241 (0.004)	0.244 (0.004)	0.015 (0.001)	0.100 (0.004)	0.955 (0.007)	1.000 (0.022)	0.124 (0.003)
33	STC	-0.139 (0.004)	-0.395 (0.004)	0.117 (0.004)	0.037 (0.001)	0.160 (0.004)	0.814 (0.012)	0.979 (0.022)	0.134 (0.003)
33	Bucher	-0.068 (0.004)	-0.315 (0.004)	0.179 (0.004)	0.021 (0.001)	0.116 (0.004)	0.917 (0.009)	0.999 (0.022)	0.126 (0.003)
34	MAIC	-0.005 (0.006)	-0.374 (0.006)	0.364 (0.006)	0.037 (0.002)	0.153 (0.006)	0.942 (0.007)	0.978 (0.022)	0.193 (0.004)
34	STC	-0.259 (0.007)	-0.680 (0.007)	0.163 (0.007)	0.118 (0.005)	0.283 (0.007)	0.769 (0.013)	0.946 (0.021)	0.227 (0.005)
34	Bucher	-0.059 (0.006)	-0.445 (0.006)	0.327 (0.006)	0.043 (0.002)	0.163 (0.006)	0.936 (0.008)	0.987 (0.022)	0.199 (0.004)
35	MAIC	0.009 (0.005)	-0.276 (0.005)	0.295 (0.005)	0.021 (0.001)	0.118 (0.005)	0.957 (0.006)	1.002 (0.022)	0.146 (0.003)
35	STC	-0.241 (0.005)	-0.557 (0.005)	0.074 (0.005)	0.084 (0.003)	0.248 (0.005)	0.677 (0.015)	1.004 (0.022)	0.160 (0.004)
35	Bucher	-0.045 (0.005)	-0.342 (0.005)	0.252 (0.005)	0.024 (0.001)	0.126 (0.005)	0.940 (0.008)	1.016 (0.023)	0.149 (0.003)
36	MAIC	-0.001 (0.004)	-0.235 (0.004)	0.234 (0.004)	0.015 (0.001)	0.097 (0.004)	0.936 (0.008)	0.982 (0.022)	0.122 (0.003)
36	STC	-0.243 (0.004)	-0.494 (0.004)	0.008 (0.004)	0.076 (0.002)	0.246 (0.004)	0.507 (0.016)	0.994 (0.022)	0.129 (0.003)
36	Bucher	-0.055 (0.004)	-0.296 (0.004)	0.186 (0.004)	0.019 (0.001)	0.110 (0.004)	0.921 (0.009)	0.980 (0.022)	0.125 (0.003)
37	MAIC	-0.005 (0.006)	-0.396 (0.007)	0.386 (0.006)	0.041 (0.002)	0.157 (0.006)	0.941 (0.007)	0.987 (0.022)	0.202 (0.005)
37	STC	-0.018 (0.008)	-0.446 (0.008)	0.410 (0.007)	0.057 (0.003)	0.186 (0.008)	0.919 (0.009)	0.919 (0.021)	0.238 (0.005)
37	Bucher	-0.136 (0.006)	-0.537 (0.007)	0.266 (0.006)	0.060 (0.003)	0.194 (0.006)	0.900 (0.009)	1.001 (0.022)	0.204 (0.005)
38	MAIC	0.003 (0.005)	-0.298 (0.005)	0.304 (0.005)	0.022 (0.001)	0.120 (0.005)	0.961 (0.006)	1.024 (0.023)	0.150 (0.003)
38	STC	-0.004 (0.005)	-0.324 (0.005)	0.317 (0.005)	0.026 (0.001)	0.130 (0.005)	0.950 (0.007)	1.005 (0.022)	0.163 (0.004)
38	Bucher	-0.125 (0.005)	-0.433 (0.005)	0.183 (0.005)	0.039 (0.002)	0.161 (0.005)	0.875 (0.010)	1.029 (0.023)	0.153 (0.003)
39	MAIC	-0.001 (0.004)	-0.247 (0.004)	0.244 (0.004)	0.014 (0.001)	0.095 (0.004)	0.955 (0.007)	1.048 (0.023)	0.120 (0.003)
39	STC	-0.002 (0.004)	-0.258 (0.004)	0.255 (0.004)	0.016 (0.001)	0.103 (0.004)	0.966 (0.006)	1.024 (0.023)	0.128 (0.003)
39	Bucher	-0.130 (0.004)	-0.380 (0.004)	0.120 (0.004)	0.032 (0.001)	0.148 (0.004)	0.841 (0.012)	1.043 (0.023)	0.122 (0.003)
40	MAIC	-0.015 (0.006)	-0.393 (0.006)	0.362 (0.006)	0.038 (0.002)	0.157 (0.006)	0.950 (0.007)	0.990 (0.022)	0.195 (0.004)
40	STC	-0.059 (0.007)	-0.479 (0.007)	0.360 (0.007)	0.055 (0.002)	0.187 (0.007)	0.928 (0.008)	0.946 (0.021)	0.226 (0.005)
40	Bucher	-0.126 (0.006)	-0.519 (0.007)	0.266 (0.006)	0.057 (0.002)	0.189 (0.006)	0.900 (0.009)	0.992 (0.022)	0.202 (0.005)
41	MAIC	0.000 (0.004)	-0.291 (0.005)	0.291 (0.004)	0.020 (0.001)	0.114 (0.004)	0.969 (0.005)	1.047 (0.023)	0.142 (0.003)
41	STC	-0.045 (0.005)	-0.360 (0.005)	0.270 (0.005)	0.027 (0.001)	0.129 (0.005)	0.947 (0.007)	1.021 (0.023)	0.157 (0.004)
41	Bucher	-0.109 (0.005)	-0.410 (0.005)	0.192 (0.005)	0.033 (0.001)	0.147 (0.005)	0.900 (0.009)	1.049 (0.023)	0.147 (0.003)
42	MAIC	0.006 (0.004)	-0.233 (0.004)	0.244 (0.004)	0.014 (0.001)	0.096 (0.004)	0.946 (0.007)	1.012 (0.023)	0.120 (0.003)
42	STC	-0.037 (0.004)	-0.288 (0.004)	0.215 (0.004)	0.018 (0.001)	0.109 (0.004)	0.939 (0.008)	0.992 (0.022)	0.129 (0.003)
42	Bucher	-0.103 (0.004)	-0.347 (0.004)	0.142 (0.004)	0.026 (0.001)	0.131 (0.004)	0.875 (0.010)	1.001 (0.022)	0.125 (0.003)
43	MAIC	0.004 (0.006)	-0.362 (0.006)	0.370 (0.006)	0.037 (0.002)	0.153 (0.006)	0.947 (0.007)	0.969 (0.022)	0.193 (0.004)
43	STC	-0.114 (0.007)	-0.529 (0.007)	0.301 (0.007)	0.060 (0.003)	0.192 (0.007)	0.904 (0.009)	0.975 (0.022)	0.217 (0.005)
43	Bucher	-0.085 (0.006)	-0.469 (0.006)	0.299 (0.006)	0.047 (0.002)	0.173 (0.006)	0.927 (0.008)	0.979 (0.022)	0.200 (0.004)
44	MAIC	0.008 (0.004)	-0.275 (0.004)	0.292 (0.004)	0.019 (0.001)	0.111 (0.004)	0.954 (0.007)	1.042 (0.023)	0.139 (0.003)
44	STC	-0.102 (0.005)	-0.412 (0.005)	0.209 (0.005)	0.037 (0.002)	0.153 (0.005)	0.905 (0.009)	0.976 (0.022)	0.162 (0.004)
44	Bucher	-0.081 (0.005)	-0.376 (0.005)	0.214 (0.005)	0.028 (0.001)	0.133 (0.005)	0.927 (0.008)	1.027 (0.023)	0.147 (0.003)
45	MAIC	0.000 (0.004)	-0.232 (0.004)	0.233 (0.004)	0.014 (0.001)	0.095 (0.004)	0.951 (0.007)	1.007 (0.023)	0.118 (0.003)
45	STC	-0.101 (0.004)	-0.349 (0.004)	0.146 (0.004)	0.027 (0.001)	0.132 (0.004)	0.861 (0.011)	0.987 (0.022)	0.128 (0.003)
45	Bucher	-0.087 (0.004)	-0.327 (0.004)	0.152 (0.004)	0.022 (0.001)	0.122 (0.004)	0.886 (0.010)	1.005 (0.022)	0.122 (0.003)
46	MAIC	0.006 (0.006)	-0.374 (0.006)	0.386 (0.006)	0.040 (0.002)	0.160 (0.006)	0.951 (0.007)	0.974 (0.022)	0.199 (0.004)
46	STC	0.132 (0.007)	-0.289 (0.007)	0.553 (0.007)	0.069 (0.003)	0.213 (0.007)	0.893 (0.010)	0.942 (0.021)	0.228 (0.005)
46	Bucher	-0.182 (0.007)	-0.576 (0.007)	0.213 (0.006)	0.076 (0.003)	0.223 (0.007)	0.844 (0.011)	0.971 (0.022)	0.207 (0.005)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
47	MAIC	-0.006 (0.005)	-0.299 (0.005)	0.287 (0.005)	0.021 (0.001)	0.118 (0.005)	0.960 (0.006)	1.021 (0.023)	0.147 (0.003)
47	STC	0.120 (0.005)	-0.196 (0.005)	0.436 (0.005)	0.041 (0.002)	0.163 (0.005)	0.881 (0.010)	0.994 (0.022)	0.162 (0.004)
47	Bucher	-0.193 (0.005)	-0.497 (0.005)	0.110 (0.005)	0.061 (0.002)	0.210 (0.005)	0.761 (0.013)	1.012 (0.023)	0.153 (0.003)
48	MAIC	-0.001 (0.004)	-0.240 (0.004)	0.239 (0.004)	0.015 (0.001)	0.099 (0.004)	0.956 (0.006)	0.993 (0.022)	0.123 (0.003)
48	STC	0.126 (0.004)	-0.126 (0.004)	0.379 (0.004)	0.033 (0.001)	0.150 (0.004)	0.827 (0.012)	0.978 (0.022)	0.132 (0.003)
48	Bucher	-0.186 (0.004)	-0.433 (0.004)	0.060 (0.004)	0.051 (0.002)	0.195 (0.004)	0.665 (0.015)	0.969 (0.022)	0.130 (0.003)
49	MAIC	0.005 (0.006)	-0.364 (0.006)	0.375 (0.006)	0.034 (0.002)	0.146 (0.006)	0.955 (0.007)	1.027 (0.023)	0.184 (0.004)
49	STC	0.139 (0.007)	-0.275 (0.007)	0.553 (0.007)	0.067 (0.003)	0.208 (0.007)	0.880 (0.010)	0.969 (0.022)	0.218 (0.005)
49	Bucher	-0.156 (0.006)	-0.544 (0.006)	0.232 (0.006)	0.062 (0.002)	0.201 (0.006)	0.895 (0.010)	1.026 (0.023)	0.193 (0.004)
50	MAIC	0.002 (0.005)	-0.284 (0.005)	0.288 (0.005)	0.021 (0.001)	0.117 (0.005)	0.947 (0.007)	1.004 (0.022)	0.145 (0.003)
50	STC	0.135 (0.005)	-0.176 (0.005)	0.446 (0.005)	0.045 (0.002)	0.172 (0.005)	0.856 (0.011)	0.966 (0.022)	0.164 (0.004)
50	Bucher	-0.157 (0.005)	-0.455 (0.005)	0.141 (0.005)	0.048 (0.002)	0.182 (0.005)	0.830 (0.012)	0.996 (0.022)	0.153 (0.003)
51	MAIC	0.007 (0.004)	-0.227 (0.004)	0.242 (0.004)	0.013 (0.001)	0.090 (0.004)	0.962 (0.006)	1.062 (0.024)	0.113 (0.003)
51	STC	0.139 (0.004)	-0.110 (0.004)	0.387 (0.004)	0.034 (0.001)	0.153 (0.004)	0.823 (0.012)	1.046 (0.023)	0.121 (0.003)
51	Bucher	-0.150 (0.004)	-0.392 (0.004)	0.092 (0.004)	0.036 (0.001)	0.161 (0.004)	0.787 (0.013)	1.060 (0.024)	0.117 (0.003)
52	MAIC	0.005 (0.006)	-0.356 (0.006)	0.366 (0.006)	0.033 (0.001)	0.147 (0.006)	0.953 (0.007)	1.010 (0.023)	0.182 (0.004)
52	STC	0.111 (0.007)	-0.300 (0.007)	0.522 (0.007)	0.064 (0.003)	0.202 (0.007)	0.896 (0.010)	0.925 (0.021)	0.226 (0.005)
52	Bucher	-0.123 (0.006)	-0.505 (0.006)	0.258 (0.006)	0.053 (0.002)	0.185 (0.006)	0.922 (0.008)	1.006 (0.023)	0.193 (0.004)
53	MAIC	0.002 (0.004)	-0.278 (0.004)	0.282 (0.004)	0.019 (0.001)	0.109 (0.004)	0.951 (0.007)	1.026 (0.023)	0.139 (0.003)
53	STC	0.112 (0.005)	-0.196 (0.005)	0.419 (0.005)	0.037 (0.002)	0.154 (0.005)	0.893 (0.010)	1.007 (0.023)	0.156 (0.003)
53	Bucher	-0.127 (0.005)	-0.421 (0.005)	0.167 (0.005)	0.038 (0.002)	0.158 (0.005)	0.866 (0.011)	1.020 (0.023)	0.147 (0.003)
54	MAIC	0.000 (0.004)	-0.230 (0.004)	0.230 (0.003)	0.012 (0.001)	0.089 (0.004)	0.964 (0.006)	1.061 (0.024)	0.111 (0.002)
54	STC	0.119 (0.004)	-0.126 (0.004)	0.365 (0.004)	0.029 (0.001)	0.138 (0.004)	0.860 (0.011)	1.041 (0.023)	0.120 (0.003)
54	Bucher	-0.129 (0.004)	-0.367 (0.004)	0.110 (0.004)	0.030 (0.001)	0.143 (0.004)	0.829 (0.012)	1.062 (0.024)	0.114 (0.003)
55	MAIC	-0.010 (0.009)	-0.524 (0.009)	0.505 (0.009)	0.077 (0.003)	0.224 (0.009)	0.941 (0.007)	0.948 (0.022)	0.277 (0.006)
55	STC	-0.067 (0.009)	-0.612 (0.010)	0.478 (0.009)	0.091 (0.004)	0.242 (0.009)	0.930 (0.008)	0.943 (0.021)	0.295 (0.007)
55	Bucher	-0.172 (0.007)	-0.590 (0.007)	0.247 (0.007)	0.077 (0.003)	0.222 (0.007)	0.878 (0.010)	0.977 (0.022)	0.218 (0.005)
56	MAIC	-0.005 (0.006)	-0.397 (0.006)	0.387 (0.006)	0.039 (0.002)	0.159 (0.006)	0.948 (0.007)	1.008 (0.023)	0.198 (0.004)
56	STC	-0.041 (0.006)	-0.438 (0.006)	0.356 (0.006)	0.042 (0.002)	0.164 (0.006)	0.949 (0.007)	1.004 (0.023)	0.202 (0.005)
56	Bucher	-0.175 (0.005)	-0.496 (0.005)	0.145 (0.005)	0.055 (0.002)	0.196 (0.005)	0.828 (0.012)	1.041 (0.023)	0.157 (0.004)
57	MAIC	-0.006 (0.005)	-0.310 (0.005)	0.299 (0.005)	0.024 (0.001)	0.123 (0.005)	0.954 (0.007)	1.003 (0.023)	0.155 (0.003)
57	STC	-0.041 (0.005)	-0.345 (0.005)	0.263 (0.005)	0.026 (0.001)	0.127 (0.005)	0.941 (0.007)	1.004 (0.022)	0.155 (0.003)
57	Bucher	-0.174 (0.004)	-0.433 (0.004)	0.084 (0.004)	0.048 (0.002)	0.186 (0.004)	0.742 (0.014)	1.006 (0.023)	0.131 (0.003)
58	MAIC	-0.008 (0.008)	-0.497 (0.009)	0.481 (0.008)	0.068 (0.003)	0.208 (0.008)	0.940 (0.008)	0.957 (0.022)	0.261 (0.006)
58	STC	-0.116 (0.009)	-0.649 (0.009)	0.417 (0.009)	0.094 (0.004)	0.244 (0.009)	0.929 (0.008)	0.960 (0.022)	0.283 (0.006)
58	Bucher	-0.164 (0.006)	-0.571 (0.007)	0.243 (0.006)	0.069 (0.003)	0.209 (0.006)	0.884 (0.010)	1.017 (0.023)	0.204 (0.005)
59	MAIC	0.000 (0.006)	-0.370 (0.006)	0.370 (0.006)	0.036 (0.002)	0.150 (0.006)	0.948 (0.007)	0.999 (0.023)	0.189 (0.004)
59	STC	-0.091 (0.006)	-0.478 (0.007)	0.297 (0.006)	0.049 (0.002)	0.175 (0.006)	0.919 (0.009)	0.976 (0.022)	0.203 (0.005)
59	Bucher	-0.154 (0.005)	-0.465 (0.005)	0.157 (0.005)	0.050 (0.002)	0.183 (0.005)	0.839 (0.012)	0.976 (0.022)	0.163 (0.004)
60	MAIC	0.002 (0.005)	-0.288 (0.005)	0.292 (0.005)	0.022 (0.001)	0.117 (0.005)	0.949 (0.007)	0.998 (0.022)	0.148 (0.003)
60	STC	-0.086 (0.005)	-0.383 (0.005)	0.212 (0.005)	0.031 (0.001)	0.141 (0.005)	0.907 (0.009)	0.986 (0.022)	0.154 (0.003)
60	Bucher	-0.158 (0.004)	-0.409 (0.004)	0.094 (0.004)	0.042 (0.002)	0.171 (0.004)	0.763 (0.013)	0.971 (0.022)	0.132 (0.003)
61	MAIC	0.000 (0.008)	-0.464 (0.008)	0.463 (0.008)	0.064 (0.003)	0.201 (0.008)	0.926 (0.008)	0.936 (0.021)	0.253 (0.006)
61	STC	-0.187 (0.009)	-0.712 (0.009)	0.338 (0.009)	0.115 (0.005)	0.268 (0.009)	0.885 (0.010)	0.949 (0.021)	0.283 (0.006)
61	Bucher	-0.143 (0.007)	-0.538 (0.007)	0.253 (0.006)	0.064 (0.003)	0.202 (0.007)	0.893 (0.010)	0.968 (0.022)	0.209 (0.005)
62	MAIC	-0.005 (0.006)	-0.356 (0.006)	0.345 (0.006)	0.034 (0.001)	0.148 (0.006)	0.932 (0.008)	0.972 (0.022)	0.184 (0.004)
62	STC	-0.178 (0.006)	-0.559 (0.006)	0.202 (0.006)	0.072 (0.003)	0.215 (0.006)	0.848 (0.011)	0.965 (0.022)	0.201 (0.004)
62	Bucher	-0.137 (0.005)	-0.440 (0.005)	0.165 (0.005)	0.045 (0.002)	0.172 (0.005)	0.839 (0.012)	0.948 (0.021)	0.163 (0.004)
63	MAIC	-0.007 (0.004)	-0.283 (0.004)	0.270 (0.005)	0.020 (0.001)	0.113 (0.004)	0.952 (0.007)	0.998 (0.022)	0.141 (0.003)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
63	STC	-0.170 (0.005)	-0.462 (0.005)	0.121 (0.005)	0.052 (0.002)	0.187 (0.005)	0.794 (0.013)	0.986 (0.022)	0.151 (0.003)
63	Bucher	-0.142 (0.004)	-0.386 (0.004)	0.103 (0.004)	0.036 (0.001)	0.156 (0.004)	0.804 (0.013)	1.000 (0.022)	0.125 (0.003)
64	MAIC	0.008 (0.008)	-0.484 (0.009)	0.501 (0.009)	0.072 (0.003)	0.210 (0.008)	0.924 (0.008)	0.939 (0.021)	0.268 (0.006)
64	STC	0.008 (0.009)	-0.525 (0.009)	0.542 (0.009)	0.082 (0.004)	0.226 (0.009)	0.931 (0.008)	0.951 (0.021)	0.286 (0.006)
64	Bucher	-0.270 (0.007)	-0.680 (0.007)	0.139 (0.006)	0.118 (0.004)	0.289 (0.007)	0.742 (0.014)	0.990 (0.022)	0.211 (0.005)
65	MAIC	-0.001 (0.006)	-0.374 (0.006)	0.371 (0.006)	0.038 (0.002)	0.155 (0.006)	0.933 (0.008)	0.975 (0.022)	0.195 (0.004)
65	STC	0.000 (0.006)	-0.387 (0.007)	0.387 (0.006)	0.041 (0.002)	0.162 (0.006)	0.941 (0.007)	0.969 (0.022)	0.204 (0.005)
65	Bucher	-0.275 (0.005)	-0.589 (0.005)	0.039 (0.005)	0.102 (0.003)	0.282 (0.005)	0.593 (0.016)	0.995 (0.022)	0.161 (0.004)
66	MAIC	-0.003 (0.005)	-0.297 (0.005)	0.291 (0.005)	0.024 (0.001)	0.123 (0.005)	0.937 (0.008)	0.970 (0.022)	0.155 (0.003)
66	STC	0.002 (0.005)	-0.296 (0.005)	0.300 (0.005)	0.024 (0.001)	0.124 (0.005)	0.937 (0.008)	0.972 (0.022)	0.156 (0.003)
66	Bucher	-0.286 (0.004)	-0.539 (0.004)	-0.032 (0.004)	0.099 (0.003)	0.287 (0.004)	0.414 (0.016)	0.974 (0.022)	0.133 (0.003)
67	MAIC	-0.016 (0.008)	-0.487 (0.008)	0.455 (0.008)	0.062 (0.003)	0.197 (0.008)	0.931 (0.008)	0.965 (0.022)	0.249 (0.006)
67	STC	-0.047 (0.009)	-0.572 (0.009)	0.478 (0.009)	0.077 (0.003)	0.222 (0.009)	0.952 (0.007)	0.978 (0.022)	0.274 (0.006)
67	Bucher	-0.260 (0.006)	-0.661 (0.007)	0.142 (0.006)	0.108 (0.004)	0.278 (0.006)	0.766 (0.013)	1.012 (0.023)	0.203 (0.005)
68	MAIC	-0.005 (0.006)	-0.363 (0.006)	0.353 (0.006)	0.034 (0.002)	0.147 (0.006)	0.956 (0.006)	0.995 (0.022)	0.183 (0.004)
68	STC	-0.030 (0.006)	-0.411 (0.006)	0.351 (0.006)	0.038 (0.002)	0.154 (0.006)	0.950 (0.007)	1.015 (0.023)	0.191 (0.004)
68	Bucher	-0.258 (0.005)	-0.566 (0.005)	0.049 (0.005)	0.090 (0.003)	0.265 (0.005)	0.622 (0.015)	1.024 (0.023)	0.153 (0.003)
69	MAIC	0.006 (0.004)	-0.276 (0.004)	0.288 (0.004)	0.019 (0.001)	0.109 (0.004)	0.958 (0.006)	1.045 (0.023)	0.138 (0.003)
69	STC	-0.015 (0.005)	-0.307 (0.005)	0.277 (0.005)	0.022 (0.001)	0.118 (0.005)	0.956 (0.006)	1.016 (0.023)	0.147 (0.003)
69	Bucher	-0.247 (0.004)	-0.495 (0.004)	0.001 (0.004)	0.077 (0.002)	0.249 (0.004)	0.517 (0.016)	1.019 (0.023)	0.124 (0.003)
70	MAIC	-0.009 (0.007)	-0.458 (0.007)	0.440 (0.008)	0.054 (0.002)	0.186 (0.007)	0.943 (0.007)	0.983 (0.022)	0.233 (0.005)
70	STC	-0.083 (0.009)	-0.598 (0.009)	0.431 (0.008)	0.080 (0.004)	0.224 (0.009)	0.934 (0.008)	0.969 (0.022)	0.271 (0.006)
70	Bucher	-0.216 (0.006)	-0.607 (0.006)	0.176 (0.006)	0.086 (0.003)	0.244 (0.006)	0.825 (0.012)	1.008 (0.023)	0.198 (0.004)
71	MAIC	0.005 (0.006)	-0.338 (0.006)	0.348 (0.006)	0.033 (0.001)	0.144 (0.006)	0.933 (0.008)	0.963 (0.022)	0.182 (0.004)
71	STC	-0.065 (0.006)	-0.440 (0.006)	0.309 (0.006)	0.041 (0.002)	0.164 (0.006)	0.935 (0.008)	0.994 (0.022)	0.192 (0.004)
71	Bucher	-0.218 (0.005)	-0.518 (0.005)	0.083 (0.005)	0.071 (0.002)	0.229 (0.005)	0.701 (0.014)	0.986 (0.022)	0.155 (0.003)
72	MAIC	0.008 (0.004)	-0.264 (0.004)	0.279 (0.004)	0.019 (0.001)	0.111 (0.004)	0.952 (0.007)	1.002 (0.022)	0.138 (0.003)
72	STC	-0.054 (0.005)	-0.340 (0.005)	0.233 (0.005)	0.025 (0.001)	0.125 (0.005)	0.930 (0.008)	0.979 (0.022)	0.149 (0.003)
72	Bucher	-0.210 (0.004)	-0.453 (0.004)	0.032 (0.004)	0.060 (0.002)	0.214 (0.004)	0.606 (0.015)	0.977 (0.022)	0.127 (0.003)
73	MAIC	0.001 (0.008)	-0.468 (0.008)	0.469 (0.008)	0.065 (0.003)	0.205 (0.008)	0.931 (0.008)	0.938 (0.021)	0.255 (0.006)
73	STC	0.108 (0.009)	-0.417 (0.009)	0.632 (0.009)	0.094 (0.004)	0.246 (0.009)	0.910 (0.009)	0.933 (0.021)	0.287 (0.006)
73	Bucher	-0.413 (0.006)	-0.816 (0.007)	-0.010 (0.006)	0.211 (0.006)	0.414 (0.006)	0.486 (0.016)	1.018 (0.023)	0.202 (0.005)
74	MAIC	0.018 (0.006)	-0.339 (0.006)	0.376 (0.006)	0.031 (0.002)	0.140 (0.006)	0.954 (0.007)	1.035 (0.023)	0.176 (0.004)
74	STC	0.113 (0.006)	-0.268 (0.006)	0.493 (0.006)	0.050 (0.002)	0.176 (0.006)	0.910 (0.009)	1.005 (0.023)	0.193 (0.004)
74	Bucher	-0.399 (0.005)	-0.708 (0.005)	-0.091 (0.005)	0.184 (0.004)	0.400 (0.005)	0.276 (0.014)	0.993 (0.022)	0.158 (0.004)
75	MAIC	0.006 (0.004)	-0.275 (0.004)	0.287 (0.004)	0.020 (0.001)	0.112 (0.004)	0.956 (0.006)	1.025 (0.023)	0.140 (0.003)
75	STC	0.105 (0.005)	-0.187 (0.005)	0.398 (0.005)	0.033 (0.001)	0.148 (0.005)	0.898 (0.010)	1.010 (0.023)	0.148 (0.003)
75	Bucher	-0.405 (0.004)	-0.654 (0.004)	-0.156 (0.004)	0.180 (0.003)	0.405 (0.004)	0.110 (0.010)	0.995 (0.022)	0.128 (0.003)
76	MAIC	0.011 (0.008)	-0.444 (0.008)	0.466 (0.008)	0.059 (0.003)	0.193 (0.008)	0.936 (0.008)	0.955 (0.022)	0.243 (0.005)
76	STC	0.111 (0.009)	-0.405 (0.009)	0.628 (0.009)	0.089 (0.004)	0.237 (0.009)	0.922 (0.008)	0.951 (0.021)	0.277 (0.006)
76	Bucher	-0.364 (0.006)	-0.760 (0.006)	0.033 (0.006)	0.172 (0.005)	0.369 (0.006)	0.564 (0.016)	1.013 (0.023)	0.200 (0.004)
77	MAIC	0.016 (0.005)	-0.330 (0.006)	0.363 (0.005)	0.030 (0.001)	0.139 (0.005)	0.947 (0.007)	1.019 (0.023)	0.173 (0.004)
77	STC	0.119 (0.006)	-0.257 (0.006)	0.495 (0.006)	0.051 (0.002)	0.182 (0.006)	0.904 (0.009)	0.996 (0.022)	0.192 (0.004)
77	Bucher	-0.367 (0.005)	-0.670 (0.005)	-0.063 (0.005)	0.158 (0.004)	0.367 (0.005)	0.324 (0.015)	1.015 (0.023)	0.153 (0.003)
78	MAIC	0.013 (0.004)	-0.261 (0.004)	0.287 (0.004)	0.019 (0.001)	0.109 (0.004)	0.951 (0.007)	1.023 (0.023)	0.137 (0.003)
78	STC	0.123 (0.005)	-0.165 (0.005)	0.411 (0.005)	0.038 (0.002)	0.157 (0.005)	0.859 (0.011)	0.973 (0.022)	0.151 (0.003)
78	Bucher	-0.367 (0.004)	-0.612 (0.004)	-0.122 (0.004)	0.150 (0.003)	0.367 (0.004)	0.158 (0.012)	1.002 (0.022)	0.125 (0.003)
79	MAIC	0.008 (0.007)	-0.431 (0.007)	0.447 (0.007)	0.052 (0.002)	0.182 (0.007)	0.942 (0.007)	0.986 (0.022)	0.227 (0.005)
79	STC	0.110 (0.008)	-0.400 (0.008)	0.620 (0.008)	0.081 (0.004)	0.226 (0.008)	0.926 (0.008)	0.992 (0.022)	0.263 (0.006)
79	Bucher	-0.321 (0.006)	-0.710 (0.006)	0.068 (0.006)	0.141 (0.004)	0.328 (0.006)	0.645 (0.015)	1.010 (0.023)	0.197 (0.004)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
80	MAIC	0.012 (0.005)	-0.325 (0.005)	0.348 (0.005)	0.029 (0.001)	0.137 (0.005)	0.944 (0.007)	1.007 (0.023)	0.170 (0.004)
80	STC	0.113 (0.006)	-0.257 (0.006)	0.483 (0.006)	0.050 (0.002)	0.177 (0.006)	0.901 (0.009)	0.979 (0.022)	0.193 (0.004)
80	Bucher	-0.313 (0.005)	-0.611 (0.005)	-0.015 (0.005)	0.120 (0.003)	0.314 (0.005)	0.442 (0.016)	1.023 (0.023)	0.149 (0.003)
81	MAIC	0.003 (0.004)	-0.264 (0.004)	0.270 (0.004)	0.018 (0.001)	0.108 (0.004)	0.961 (0.006)	1.023 (0.023)	0.133 (0.003)
81	STC	0.106 (0.004)	-0.179 (0.004)	0.390 (0.005)	0.031 (0.001)	0.143 (0.004)	0.900 (0.009)	1.022 (0.023)	0.142 (0.003)
81	Bucher	-0.324 (0.004)	-0.564 (0.004)	-0.083 (0.004)	0.120 (0.003)	0.324 (0.004)	0.245 (0.014)	1.012 (0.023)	0.121 (0.003)
82	MAIC	0.006 (0.008)	-0.461 (0.008)	0.473 (0.007)	0.057 (0.003)	0.192 (0.008)	0.948 (0.007)	0.997 (0.022)	0.239 (0.005)
82	STC	-0.070 (0.008)	-0.583 (0.009)	0.442 (0.008)	0.077 (0.004)	0.218 (0.008)	0.937 (0.008)	0.973 (0.022)	0.268 (0.006)
82	Bucher	-0.163 (0.007)	-0.581 (0.007)	0.254 (0.006)	0.071 (0.003)	0.214 (0.007)	0.888 (0.010)	1.011 (0.023)	0.210 (0.005)
83	MAIC	0.001 (0.006)	-0.352 (0.006)	0.354 (0.006)	0.035 (0.002)	0.150 (0.006)	0.945 (0.007)	0.958 (0.022)	0.188 (0.004)
83	STC	-0.063 (0.006)	-0.437 (0.006)	0.311 (0.006)	0.043 (0.002)	0.166 (0.006)	0.936 (0.008)	0.968 (0.022)	0.197 (0.004)
83	Bucher	-0.158 (0.005)	-0.477 (0.005)	0.161 (0.005)	0.054 (0.002)	0.191 (0.005)	0.836 (0.012)	0.959 (0.021)	0.170 (0.004)
84	MAIC	-0.005 (0.004)	-0.285 (0.004)	0.276 (0.004)	0.020 (0.001)	0.111 (0.004)	0.948 (0.007)	1.013 (0.023)	0.141 (0.003)
84	STC	-0.066 (0.005)	-0.356 (0.005)	0.225 (0.005)	0.026 (0.001)	0.130 (0.005)	0.932 (0.008)	1.005 (0.022)	0.147 (0.003)
84	Bucher	-0.159 (0.004)	-0.417 (0.004)	0.098 (0.004)	0.043 (0.002)	0.175 (0.004)	0.781 (0.013)	0.990 (0.022)	0.133 (0.003)
85	MAIC	0.004 (0.007)	-0.434 (0.007)	0.443 (0.007)	0.050 (0.003)	0.175 (0.007)	0.951 (0.007)	1.000 (0.023)	0.224 (0.005)
85	STC	-0.158 (0.008)	-0.658 (0.009)	0.343 (0.008)	0.094 (0.004)	0.244 (0.008)	0.901 (0.009)	0.969 (0.022)	0.264 (0.006)
85	Bucher	-0.132 (0.007)	-0.537 (0.007)	0.274 (0.006)	0.061 (0.003)	0.197 (0.007)	0.918 (0.009)	0.989 (0.022)	0.209 (0.005)
86	MAIC	0.007 (0.005)	-0.328 (0.005)	0.342 (0.005)	0.027 (0.001)	0.132 (0.005)	0.951 (0.007)	1.035 (0.023)	0.165 (0.004)
86	STC	-0.145 (0.006)	-0.510 (0.006)	0.221 (0.006)	0.055 (0.002)	0.188 (0.006)	0.880 (0.010)	1.006 (0.023)	0.186 (0.004)
86	Bucher	-0.138 (0.005)	-0.448 (0.005)	0.173 (0.005)	0.043 (0.002)	0.169 (0.005)	0.869 (0.011)	1.025 (0.023)	0.154 (0.003)
87	MAIC	0.006 (0.004)	-0.261 (0.004)	0.273 (0.004)	0.017 (0.001)	0.104 (0.004)	0.957 (0.006)	1.046 (0.023)	0.130 (0.003)
87	STC	-0.137 (0.004)	-0.420 (0.004)	0.147 (0.004)	0.038 (0.001)	0.162 (0.004)	0.849 (0.011)	1.033 (0.023)	0.140 (0.003)
87	Bucher	-0.136 (0.004)	-0.386 (0.004)	0.115 (0.004)	0.033 (0.001)	0.152 (0.004)	0.828 (0.012)	1.044 (0.023)	0.123 (0.003)
88	MAIC	0.019 (0.007)	-0.397 (0.007)	0.435 (0.007)	0.049 (0.002)	0.179 (0.007)	0.942 (0.007)	0.965 (0.022)	0.220 (0.005)
88	STC	-0.250 (0.008)	-0.741 (0.008)	0.242 (0.008)	0.129 (0.005)	0.292 (0.008)	0.826 (0.012)	0.972 (0.022)	0.258 (0.006)
88	Bucher	-0.104 (0.007)	-0.499 (0.007)	0.291 (0.007)	0.056 (0.003)	0.188 (0.007)	0.918 (0.009)	0.945 (0.021)	0.213 (0.005)
89	MAIC	-0.001 (0.005)	-0.322 (0.005)	0.320 (0.005)	0.027 (0.001)	0.128 (0.005)	0.938 (0.008)	1.005 (0.023)	0.163 (0.004)
89	STC	-0.245 (0.006)	-0.605 (0.006)	0.115 (0.006)	0.095 (0.003)	0.261 (0.006)	0.747 (0.014)	0.989 (0.022)	0.186 (0.004)
89	Bucher	-0.112 (0.005)	-0.414 (0.005)	0.191 (0.005)	0.038 (0.002)	0.157 (0.005)	0.884 (0.010)	0.975 (0.022)	0.158 (0.004)
90	MAIC	0.003 (0.004)	-0.254 (0.004)	0.260 (0.004)	0.017 (0.001)	0.106 (0.004)	0.949 (0.007)	0.995 (0.022)	0.132 (0.003)
90	STC	-0.240 (0.004)	-0.518 (0.004)	0.039 (0.004)	0.077 (0.002)	0.244 (0.004)	0.603 (0.015)	1.029 (0.023)	0.138 (0.003)
90	Bucher	-0.111 (0.004)	-0.356 (0.004)	0.133 (0.004)	0.028 (0.001)	0.136 (0.004)	0.850 (0.011)	0.994 (0.022)	0.126 (0.003)
91	MAIC	-0.002 (0.007)	-0.450 (0.008)	0.447 (0.007)	0.055 (0.003)	0.187 (0.007)	0.938 (0.008)	0.978 (0.022)	0.234 (0.005)
91	STC	-0.007 (0.008)	-0.506 (0.009)	0.493 (0.008)	0.071 (0.003)	0.213 (0.008)	0.945 (0.007)	0.958 (0.022)	0.266 (0.006)
91	Bucher	-0.247 (0.007)	-0.658 (0.007)	0.163 (0.006)	0.105 (0.004)	0.272 (0.007)	0.790 (0.013)	0.996 (0.022)	0.210 (0.005)
92	MAIC	-0.005 (0.006)	-0.347 (0.006)	0.337 (0.005)	0.030 (0.001)	0.138 (0.006)	0.946 (0.007)	1.001 (0.022)	0.174 (0.004)
92	STC	-0.007 (0.006)	-0.374 (0.006)	0.359 (0.006)	0.037 (0.002)	0.153 (0.006)	0.941 (0.007)	0.976 (0.022)	0.192 (0.004)
92	Bucher	-0.256 (0.005)	-0.571 (0.005)	0.058 (0.005)	0.092 (0.003)	0.262 (0.005)	0.648 (0.015)	0.984 (0.022)	0.163 (0.004)
93	MAIC	-0.002 (0.004)	-0.274 (0.004)	0.270 (0.004)	0.018 (0.001)	0.107 (0.004)	0.959 (0.006)	1.024 (0.023)	0.136 (0.003)
93	STC	-0.004 (0.005)	-0.289 (0.005)	0.281 (0.005)	0.021 (0.001)	0.114 (0.005)	0.953 (0.007)	1.011 (0.023)	0.144 (0.003)
93	Bucher	-0.258 (0.004)	-0.512 (0.004)	-0.004 (0.004)	0.083 (0.002)	0.261 (0.004)	0.469 (0.016)	1.016 (0.023)	0.128 (0.003)
94	MAIC	0.003 (0.007)	-0.428 (0.007)	0.434 (0.007)	0.051 (0.002)	0.180 (0.007)	0.942 (0.007)	0.974 (0.022)	0.226 (0.005)
94	STC	-0.043 (0.008)	-0.536 (0.009)	0.451 (0.008)	0.072 (0.004)	0.211 (0.008)	0.932 (0.008)	0.949 (0.021)	0.265 (0.006)
94	Bucher	-0.218 (0.007)	-0.620 (0.007)	0.184 (0.006)	0.091 (0.004)	0.249 (0.007)	0.823 (0.012)	0.981 (0.022)	0.209 (0.005)
95	MAIC	0.005 (0.005)	-0.324 (0.006)	0.333 (0.005)	0.030 (0.001)	0.138 (0.005)	0.949 (0.007)	0.970 (0.022)	0.173 (0.004)
95	STC	-0.039 (0.006)	-0.399 (0.006)	0.321 (0.006)	0.038 (0.002)	0.153 (0.006)	0.937 (0.008)	0.964 (0.022)	0.191 (0.004)
95	Bucher	-0.213 (0.005)	-0.521 (0.005)	0.094 (0.005)	0.071 (0.002)	0.226 (0.005)	0.710 (0.014)	0.985 (0.022)	0.159 (0.004)
96	MAIC	-0.003 (0.004)	-0.266 (0.004)	0.261 (0.004)	0.016 (0.001)	0.103 (0.004)	0.962 (0.006)	1.049 (0.023)	0.128 (0.003)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
96	STC	-0.047 (0.004)	-0.327 (0.004)	0.233 (0.004)	0.022 (0.001)	0.118 (0.004)	0.935 (0.008)	1.012 (0.023)	0.141 (0.003)
96	Bucher	-0.222 (0.004)	-0.471 (0.004)	0.026 (0.004)	0.065 (0.002)	0.225 (0.004)	0.587 (0.016)	1.023 (0.023)	0.124 (0.003)
97	MAIC	-0.008 (0.007)	-0.423 (0.007)	0.406 (0.007)	0.048 (0.002)	0.172 (0.007)	0.928 (0.008)	0.967 (0.022)	0.219 (0.005)
97	STC	-0.121 (0.008)	-0.608 (0.008)	0.367 (0.008)	0.076 (0.004)	0.220 (0.008)	0.931 (0.008)	1.000 (0.022)	0.249 (0.006)
97	Bucher	-0.183 (0.006)	-0.576 (0.007)	0.211 (0.006)	0.075 (0.003)	0.222 (0.006)	0.847 (0.011)	0.980 (0.022)	0.205 (0.005)
98	MAIC	0.015 (0.005)	-0.305 (0.005)	0.335 (0.005)	0.027 (0.001)	0.133 (0.005)	0.953 (0.007)	0.994 (0.022)	0.164 (0.004)
98	STC	-0.088 (0.006)	-0.443 (0.006)	0.268 (0.006)	0.042 (0.002)	0.162 (0.006)	0.925 (0.008)	0.980 (0.022)	0.185 (0.004)
98	Bucher	-0.174 (0.005)	-0.476 (0.005)	0.127 (0.005)	0.054 (0.002)	0.194 (0.005)	0.811 (0.012)	0.992 (0.022)	0.155 (0.003)
99	MAIC	0.006 (0.004)	-0.250 (0.004)	0.262 (0.004)	0.017 (0.001)	0.105 (0.004)	0.960 (0.006)	0.997 (0.022)	0.131 (0.003)
99	STC	-0.105 (0.004)	-0.380 (0.004)	0.171 (0.004)	0.030 (0.001)	0.140 (0.004)	0.882 (0.010)	1.014 (0.023)	0.138 (0.003)
99	Bucher	-0.173 (0.004)	-0.416 (0.004)	0.071 (0.004)	0.044 (0.001)	0.180 (0.004)	0.713 (0.014)	1.032 (0.023)	0.120 (0.003)
100	MAIC	-0.012 (0.007)	-0.445 (0.007)	0.422 (0.007)	0.047 (0.002)	0.171 (0.007)	0.951 (0.007)	1.023 (0.023)	0.216 (0.005)
100	STC	0.108 (0.008)	-0.385 (0.008)	0.600 (0.008)	0.074 (0.003)	0.214 (0.008)	0.922 (0.008)	1.007 (0.023)	0.250 (0.006)
100	Bucher	-0.373 (0.006)	-0.778 (0.006)	0.033 (0.006)	0.176 (0.005)	0.376 (0.006)	0.576 (0.016)	1.072 (0.024)	0.193 (0.004)
101	MAIC	0.004 (0.005)	-0.326 (0.005)	0.335 (0.005)	0.029 (0.001)	0.136 (0.005)	0.943 (0.007)	0.992 (0.022)	0.170 (0.004)
101	STC	0.133 (0.006)	-0.227 (0.006)	0.494 (0.006)	0.055 (0.002)	0.188 (0.006)	0.869 (0.011)	0.958 (0.021)	0.192 (0.004)
101	Bucher	-0.366 (0.005)	-0.676 (0.005)	-0.056 (0.005)	0.159 (0.004)	0.367 (0.005)	0.367 (0.015)	1.007 (0.023)	0.157 (0.004)
102	MAIC	0.008 (0.004)	-0.257 (0.004)	0.272 (0.004)	0.016 (0.001)	0.102 (0.004)	0.965 (0.006)	1.058 (0.024)	0.128 (0.003)
102	STC	0.133 (0.005)	-0.147 (0.005)	0.414 (0.005)	0.038 (0.001)	0.160 (0.005)	0.838 (0.012)	1.003 (0.022)	0.143 (0.003)
102	Bucher	-0.366 (0.004)	-0.617 (0.004)	-0.116 (0.004)	0.149 (0.003)	0.366 (0.004)	0.176 (0.012)	1.033 (0.023)	0.124 (0.003)
103	MAIC	0.005 (0.007)	-0.414 (0.007)	0.425 (0.007)	0.045 (0.002)	0.168 (0.007)	0.948 (0.007)	1.004 (0.023)	0.213 (0.005)
103	STC	0.134 (0.008)	-0.352 (0.008)	0.621 (0.008)	0.085 (0.004)	0.230 (0.008)	0.896 (0.010)	0.960 (0.022)	0.259 (0.006)
103	Bucher	-0.315 (0.006)	-0.714 (0.006)	0.084 (0.006)	0.139 (0.004)	0.327 (0.006)	0.653 (0.015)	1.022 (0.023)	0.199 (0.004)
104	MAIC	0.002 (0.005)	-0.320 (0.005)	0.324 (0.005)	0.027 (0.001)	0.131 (0.005)	0.954 (0.007)	1.003 (0.022)	0.164 (0.004)
104	STC	0.135 (0.006)	-0.221 (0.006)	0.491 (0.006)	0.051 (0.002)	0.181 (0.006)	0.888 (0.010)	1.004 (0.022)	0.181 (0.004)
104	Bucher	-0.314 (0.005)	-0.620 (0.005)	-0.009 (0.005)	0.122 (0.003)	0.316 (0.005)	0.472 (0.016)	1.019 (0.023)	0.153 (0.003)
105	MAIC	0.001 (0.004)	-0.257 (0.004)	0.259 (0.004)	0.017 (0.001)	0.106 (0.004)	0.958 (0.006)	1.005 (0.022)	0.131 (0.003)
105	STC	0.133 (0.005)	-0.143 (0.004)	0.409 (0.005)	0.038 (0.001)	0.160 (0.005)	0.836 (0.012)	0.986 (0.022)	0.143 (0.003)
105	Bucher	-0.321 (0.004)	-0.568 (0.004)	-0.075 (0.004)	0.119 (0.003)	0.322 (0.004)	0.276 (0.014)	0.986 (0.022)	0.128 (0.003)
106	MAIC	-0.009 (0.007)	-0.418 (0.007)	0.400 (0.007)	0.045 (0.002)	0.169 (0.007)	0.938 (0.008)	0.979 (0.022)	0.213 (0.005)
106	STC	0.111 (0.008)	-0.372 (0.008)	0.594 (0.008)	0.076 (0.003)	0.220 (0.008)	0.921 (0.009)	0.976 (0.022)	0.252 (0.006)
106	Bucher	-0.262 (0.006)	-0.654 (0.006)	0.130 (0.006)	0.108 (0.004)	0.279 (0.006)	0.753 (0.014)	1.009 (0.023)	0.198 (0.004)
107	MAIC	-0.001 (0.005)	-0.317 (0.005)	0.314 (0.005)	0.025 (0.001)	0.127 (0.005)	0.953 (0.007)	1.016 (0.023)	0.159 (0.004)
107	STC	0.124 (0.006)	-0.229 (0.006)	0.478 (0.006)	0.050 (0.002)	0.179 (0.006)	0.887 (0.010)	0.971 (0.022)	0.186 (0.004)
107	Bucher	-0.263 (0.005)	-0.563 (0.005)	0.037 (0.005)	0.091 (0.003)	0.268 (0.005)	0.613 (0.015)	1.034 (0.023)	0.148 (0.003)
108	MAIC	0.009 (0.004)	-0.245 (0.004)	0.262 (0.004)	0.016 (0.001)	0.100 (0.004)	0.953 (0.007)	1.031 (0.023)	0.126 (0.003)
108	STC	0.124 (0.004)	-0.149 (0.004)	0.397 (0.004)	0.034 (0.001)	0.149 (0.004)	0.860 (0.011)	1.016 (0.023)	0.137 (0.003)
108	Bucher	-0.251 (0.004)	-0.493 (0.004)	-0.008 (0.004)	0.077 (0.002)	0.252 (0.004)	0.482 (0.016)	1.034 (0.023)	0.120 (0.003)
109	MAIC	-0.040 (0.014)	-0.751 (0.015)	0.672 (0.014)	0.189 (0.009)	0.344 (0.014)	0.895 (0.010)	0.839 (0.020)	0.433 (0.010)
109	STC	-0.068 (0.012)	-0.758 (0.012)	0.621 (0.012)	0.144 (0.007)	0.300 (0.012)	0.933 (0.008)	0.940 (0.021)	0.374 (0.008)
109	Bucher	-0.265 (0.007)	-0.695 (0.007)	0.164 (0.007)	0.118 (0.004)	0.285 (0.007)	0.788 (0.013)	1.008 (0.023)	0.217 (0.005)
110	MAIC	-0.011 (0.010)	-0.552 (0.010)	0.531 (0.010)	0.095 (0.004)	0.248 (0.010)	0.916 (0.009)	0.899 (0.021)	0.307 (0.007)
110	STC	-0.034 (0.008)	-0.525 (0.008)	0.456 (0.008)	0.066 (0.003)	0.208 (0.008)	0.954 (0.007)	0.981 (0.022)	0.255 (0.006)
110	Bucher	-0.262 (0.005)	-0.589 (0.005)	0.065 (0.005)	0.095 (0.003)	0.269 (0.005)	0.657 (0.015)	1.025 (0.023)	0.163 (0.004)
111	MAIC	0.000 (0.007)	-0.415 (0.007)	0.414 (0.007)	0.050 (0.002)	0.181 (0.007)	0.937 (0.008)	0.946 (0.021)	0.224 (0.005)
111	STC	-0.027 (0.006)	-0.391 (0.006)	0.336 (0.006)	0.035 (0.002)	0.149 (0.006)	0.938 (0.008)	0.999 (0.022)	0.186 (0.004)
111	Bucher	-0.254 (0.004)	-0.516 (0.004)	0.009 (0.004)	0.083 (0.002)	0.257 (0.004)	0.533 (0.016)	0.992 (0.022)	0.135 (0.003)
112	MAIC	-0.019 (0.012)	-0.682 (0.013)	0.645 (0.013)	0.152 (0.007)	0.310 (0.012)	0.910 (0.009)	0.870 (0.020)	0.389 (0.009)
112	STC	-0.084 (0.011)	-0.764 (0.012)	0.596 (0.011)	0.134 (0.006)	0.294 (0.011)	0.953 (0.007)	0.974 (0.022)	0.356 (0.008)
112	Bucher	-0.236 (0.007)	-0.656 (0.007)	0.183 (0.007)	0.102 (0.004)	0.265 (0.007)	0.802 (0.013)	0.991 (0.022)	0.216 (0.005)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
113	MAIC	-0.010 (0.009)	-0.516 (0.009)	0.496 (0.009)	0.077 (0.003)	0.223 (0.009)	0.921 (0.009)	0.928 (0.021)	0.278 (0.006)
113	STC	-0.089 (0.008)	-0.571 (0.008)	0.392 (0.008)	0.067 (0.003)	0.205 (0.008)	0.934 (0.008)	1.012 (0.023)	0.243 (0.005)
113	Bucher	-0.231 (0.005)	-0.550 (0.005)	0.089 (0.005)	0.079 (0.003)	0.240 (0.005)	0.723 (0.014)	1.017 (0.023)	0.160 (0.004)
114	MAIC	-0.005 (0.007)	-0.398 (0.007)	0.387 (0.007)	0.044 (0.002)	0.166 (0.007)	0.923 (0.008)	0.955 (0.022)	0.209 (0.005)
114	STC	-0.084 (0.006)	-0.441 (0.006)	0.273 (0.006)	0.041 (0.002)	0.162 (0.006)	0.922 (0.008)	0.984 (0.022)	0.185 (0.004)
114	Bucher	-0.237 (0.004)	-0.493 (0.004)	0.020 (0.004)	0.073 (0.002)	0.240 (0.004)	0.542 (0.016)	1.009 (0.023)	0.130 (0.003)
115	MAIC	-0.012 (0.012)	-0.633 (0.012)	0.609 (0.013)	0.139 (0.007)	0.296 (0.012)	0.906 (0.009)	0.849 (0.020)	0.373 (0.008)
115	STC	-0.183 (0.011)	-0.852 (0.012)	0.486 (0.011)	0.159 (0.007)	0.319 (0.011)	0.924 (0.008)	0.962 (0.022)	0.355 (0.008)
115	Bucher	-0.209 (0.007)	-0.617 (0.007)	0.200 (0.006)	0.088 (0.003)	0.244 (0.007)	0.840 (0.012)	0.990 (0.022)	0.211 (0.005)
116	MAIC	0.004 (0.008)	-0.474 (0.008)	0.482 (0.009)	0.068 (0.003)	0.212 (0.008)	0.928 (0.008)	0.936 (0.021)	0.261 (0.006)
116	STC	-0.156 (0.008)	-0.630 (0.008)	0.317 (0.008)	0.087 (0.004)	0.240 (0.008)	0.899 (0.010)	0.969 (0.022)	0.249 (0.006)
116	Bucher	-0.195 (0.005)	-0.506 (0.005)	0.116 (0.005)	0.061 (0.002)	0.207 (0.005)	0.785 (0.013)	1.047 (0.023)	0.152 (0.003)
117	MAIC	0.002 (0.006)	-0.371 (0.006)	0.376 (0.007)	0.040 (0.002)	0.162 (0.006)	0.943 (0.007)	0.957 (0.022)	0.199 (0.004)
117	STC	-0.166 (0.006)	-0.518 (0.006)	0.186 (0.006)	0.061 (0.002)	0.202 (0.006)	0.843 (0.012)	0.974 (0.022)	0.184 (0.004)
117	Bucher	-0.203 (0.004)	-0.453 (0.004)	0.047 (0.004)	0.057 (0.002)	0.209 (0.004)	0.634 (0.015)	1.007 (0.023)	0.127 (0.003)
118	MAIC	-0.009 (0.012)	-0.681 (0.013)	0.663 (0.013)	0.147 (0.007)	0.306 (0.012)	0.916 (0.009)	0.893 (0.021)	0.384 (0.009)
118	STC	0.000 (0.011)	-0.679 (0.011)	0.679 (0.011)	0.124 (0.006)	0.282 (0.011)	0.950 (0.007)	0.985 (0.022)	0.352 (0.008)
118	Bucher	-0.410 (0.007)	-0.833 (0.007)	0.014 (0.006)	0.213 (0.006)	0.412 (0.007)	0.532 (0.016)	1.014 (0.023)	0.213 (0.005)
119	MAIC	-0.006 (0.009)	-0.518 (0.009)	0.506 (0.010)	0.082 (0.004)	0.231 (0.009)	0.924 (0.008)	0.913 (0.021)	0.286 (0.006)
119	STC	0.008 (0.008)	-0.472 (0.008)	0.489 (0.008)	0.063 (0.003)	0.201 (0.008)	0.949 (0.007)	0.976 (0.022)	0.251 (0.006)
119	Bucher	-0.416 (0.005)	-0.739 (0.005)	-0.094 (0.005)	0.201 (0.005)	0.417 (0.005)	0.299 (0.014)	0.990 (0.022)	0.166 (0.004)
120	MAIC	-0.008 (0.007)	-0.406 (0.007)	0.390 (0.007)	0.044 (0.002)	0.165 (0.007)	0.936 (0.008)	0.965 (0.022)	0.211 (0.005)
120	STC	0.004 (0.006)	-0.354 (0.006)	0.361 (0.006)	0.033 (0.002)	0.144 (0.006)	0.945 (0.007)	0.999 (0.022)	0.183 (0.004)
120	Bucher	-0.427 (0.004)	-0.686 (0.004)	-0.168 (0.004)	0.200 (0.004)	0.427 (0.004)	0.091 (0.009)	0.987 (0.022)	0.134 (0.003)
121	MAIC	-0.017 (0.012)	-0.660 (0.013)	0.626 (0.012)	0.142 (0.006)	0.303 (0.012)	0.904 (0.009)	0.870 (0.020)	0.377 (0.008)
121	STC	-0.049 (0.012)	-0.722 (0.012)	0.624 (0.012)	0.142 (0.007)	0.299 (0.012)	0.925 (0.008)	0.917 (0.021)	0.374 (0.008)
121	Bucher	-0.390 (0.007)	-0.807 (0.007)	0.026 (0.007)	0.199 (0.006)	0.395 (0.007)	0.541 (0.016)	0.983 (0.022)	0.216 (0.005)
122	MAIC	0.007 (0.009)	-0.481 (0.009)	0.496 (0.009)	0.072 (0.003)	0.214 (0.009)	0.918 (0.009)	0.926 (0.021)	0.269 (0.006)
122	STC	-0.021 (0.008)	-0.497 (0.008)	0.454 (0.008)	0.060 (0.003)	0.195 (0.008)	0.944 (0.007)	0.993 (0.022)	0.244 (0.005)
122	Bucher	-0.374 (0.005)	-0.690 (0.005)	-0.058 (0.005)	0.166 (0.004)	0.375 (0.005)	0.363 (0.015)	1.003 (0.022)	0.161 (0.004)
123	MAIC	0.008 (0.006)	-0.375 (0.006)	0.390 (0.007)	0.040 (0.002)	0.161 (0.006)	0.941 (0.007)	0.977 (0.022)	0.200 (0.004)
123	STC	-0.016 (0.006)	-0.369 (0.006)	0.337 (0.006)	0.032 (0.001)	0.142 (0.006)	0.955 (0.007)	1.015 (0.023)	0.178 (0.004)
123	Bucher	-0.386 (0.004)	-0.640 (0.004)	-0.133 (0.004)	0.167 (0.003)	0.387 (0.004)	0.151 (0.011)	0.981 (0.022)	0.132 (0.003)
124	MAIC	0.000 (0.011)	-0.604 (0.012)	0.603 (0.012)	0.124 (0.006)	0.281 (0.011)	0.902 (0.009)	0.873 (0.020)	0.353 (0.008)
124	STC	-0.065 (0.011)	-0.725 (0.011)	0.595 (0.011)	0.119 (0.005)	0.278 (0.011)	0.948 (0.007)	0.992 (0.022)	0.339 (0.008)
124	Bucher	-0.324 (0.007)	-0.731 (0.007)	0.082 (0.006)	0.149 (0.005)	0.335 (0.007)	0.669 (0.015)	0.995 (0.022)	0.209 (0.005)
125	MAIC	-0.006 (0.008)	-0.470 (0.008)	0.457 (0.009)	0.065 (0.003)	0.203 (0.008)	0.931 (0.008)	0.930 (0.021)	0.254 (0.006)
125	STC	-0.069 (0.008)	-0.538 (0.008)	0.400 (0.008)	0.064 (0.003)	0.201 (0.008)	0.933 (0.008)	0.982 (0.022)	0.244 (0.005)
125	Bucher	-0.330 (0.005)	-0.640 (0.005)	-0.020 (0.005)	0.135 (0.004)	0.332 (0.005)	0.460 (0.016)	0.980 (0.022)	0.161 (0.004)
126	MAIC	0.003 (0.006)	-0.365 (0.006)	0.370 (0.006)	0.037 (0.002)	0.153 (0.006)	0.942 (0.007)	0.979 (0.022)	0.192 (0.004)
126	STC	-0.066 (0.006)	-0.414 (0.006)	0.283 (0.006)	0.036 (0.002)	0.152 (0.006)	0.935 (0.008)	0.999 (0.022)	0.178 (0.004)
126	Bucher	-0.330 (0.004)	-0.578 (0.004)	-0.081 (0.004)	0.124 (0.003)	0.330 (0.004)	0.258 (0.014)	1.013 (0.023)	0.125 (0.003)
127	MAIC	0.021 (0.012)	-0.618 (0.012)	0.661 (0.012)	0.135 (0.006)	0.296 (0.012)	0.899 (0.010)	0.890 (0.021)	0.367 (0.008)
127	STC	0.109 (0.011)	-0.560 (0.011)	0.779 (0.011)	0.134 (0.006)	0.294 (0.011)	0.932 (0.008)	0.979 (0.022)	0.349 (0.008)
127	Bucher	-0.623 (0.007)	-1.042 (0.007)	-0.204 (0.006)	0.432 (0.009)	0.623 (0.007)	0.146 (0.011)	1.017 (0.023)	0.210 (0.005)
128	MAIC	0.003 (0.008)	-0.484 (0.008)	0.491 (0.009)	0.071 (0.003)	0.213 (0.008)	0.921 (0.009)	0.934 (0.021)	0.266 (0.006)
128	STC	0.102 (0.008)	-0.374 (0.008)	0.577 (0.008)	0.068 (0.003)	0.208 (0.008)	0.928 (0.008)	1.013 (0.023)	0.239 (0.005)
128	Bucher	-0.627 (0.005)	-0.946 (0.005)	-0.308 (0.005)	0.418 (0.006)	0.627 (0.005)	0.018 (0.004)	1.031 (0.023)	0.158 (0.004)
129	MAIC	0.003 (0.006)	-0.376 (0.006)	0.383 (0.006)	0.038 (0.002)	0.153 (0.006)	0.948 (0.007)	0.997 (0.023)	0.194 (0.004)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
129	STC	0.100 (0.006)	-0.252 (0.006)	0.453 (0.006)	0.041 (0.002)	0.160 (0.006)	0.918 (0.009)	1.029 (0.023)	0.175 (0.004)
129	Bucher	-0.616 (0.004)	-0.871 (0.004)	-0.361 (0.004)	0.396 (0.005)	0.616 (0.004)	0.002 (0.001)	1.014 (0.023)	0.128 (0.003)
130	MAIC	0.018 (0.011)	-0.603 (0.012)	0.640 (0.012)	0.131 (0.006)	0.287 (0.011)	0.904 (0.009)	0.878 (0.020)	0.361 (0.008)
130	STC	0.138 (0.011)	-0.526 (0.011)	0.803 (0.011)	0.145 (0.007)	0.301 (0.011)	0.920 (0.009)	0.954 (0.021)	0.355 (0.008)
130	Bucher	-0.560 (0.007)	-0.973 (0.007)	-0.146 (0.007)	0.359 (0.008)	0.560 (0.007)	0.245 (0.014)	0.985 (0.022)	0.214 (0.005)
131	MAIC	0.017 (0.008)	-0.452 (0.008)	0.487 (0.009)	0.069 (0.003)	0.210 (0.008)	0.921 (0.009)	0.914 (0.021)	0.262 (0.006)
131	STC	0.116 (0.008)	-0.354 (0.008)	0.586 (0.008)	0.075 (0.004)	0.217 (0.008)	0.915 (0.009)	0.964 (0.022)	0.249 (0.006)
131	Bucher	-0.556 (0.005)	-0.870 (0.005)	-0.242 (0.005)	0.335 (0.006)	0.556 (0.005)	0.057 (0.007)	1.002 (0.022)	0.160 (0.004)
132	MAIC	0.017 (0.006)	-0.349 (0.006)	0.383 (0.006)	0.039 (0.002)	0.162 (0.006)	0.937 (0.008)	0.943 (0.021)	0.198 (0.004)
132	STC	0.118 (0.006)	-0.230 (0.006)	0.466 (0.006)	0.048 (0.002)	0.178 (0.006)	0.883 (0.010)	0.954 (0.021)	0.186 (0.004)
132	Bucher	-0.557 (0.004)	-0.808 (0.004)	-0.306 (0.004)	0.327 (0.005)	0.557 (0.004)	0.006 (0.002)	0.995 (0.022)	0.129 (0.003)
133	MAIC	0.016 (0.011)	-0.572 (0.011)	0.604 (0.011)	0.118 (0.005)	0.277 (0.011)	0.897 (0.010)	0.874 (0.020)	0.343 (0.008)
133	STC	0.116 (0.011)	-0.541 (0.011)	0.772 (0.011)	0.131 (0.006)	0.289 (0.011)	0.936 (0.008)	0.977 (0.022)	0.343 (0.008)
133	Bucher	-0.477 (0.007)	-0.883 (0.007)	-0.072 (0.007)	0.274 (0.007)	0.479 (0.007)	0.375 (0.015)	0.965 (0.022)	0.214 (0.005)
134	MAIC	0.026 (0.008)	-0.432 (0.008)	0.484 (0.008)	0.061 (0.003)	0.199 (0.008)	0.929 (0.008)	0.954 (0.022)	0.245 (0.005)
134	STC	0.109 (0.008)	-0.357 (0.008)	0.574 (0.008)	0.071 (0.003)	0.213 (0.008)	0.912 (0.009)	0.976 (0.022)	0.243 (0.005)
134	Bucher	-0.475 (0.005)	-0.784 (0.005)	-0.167 (0.005)	0.250 (0.005)	0.476 (0.005)	0.139 (0.011)	1.013 (0.023)	0.155 (0.003)
135	MAIC	0.025 (0.006)	-0.333 (0.006)	0.382 (0.006)	0.034 (0.002)	0.146 (0.006)	0.946 (0.007)	1.003 (0.023)	0.182 (0.004)
135	STC	0.112 (0.005)	-0.232 (0.005)	0.457 (0.005)	0.041 (0.002)	0.161 (0.005)	0.910 (0.009)	1.041 (0.023)	0.169 (0.004)
135	Bucher	-0.478 (0.004)	-0.725 (0.004)	-0.231 (0.004)	0.244 (0.004)	0.478 (0.004)	0.031 (0.005)	1.022 (0.023)	0.123 (0.003)
136	MAIC	-0.013 (0.010)	-0.603 (0.011)	0.577 (0.011)	0.108 (0.005)	0.261 (0.010)	0.917 (0.009)	0.916 (0.021)	0.329 (0.007)
136	STC	-0.084 (0.011)	-0.708 (0.011)	0.540 (0.011)	0.122 (0.006)	0.277 (0.011)	0.931 (0.008)	0.936 (0.021)	0.340 (0.008)
136	Bucher	-0.244 (0.007)	-0.671 (0.007)	0.183 (0.007)	0.109 (0.004)	0.276 (0.007)	0.815 (0.012)	0.984 (0.022)	0.222 (0.005)
137	MAIC	-0.018 (0.007)	-0.467 (0.008)	0.431 (0.008)	0.056 (0.003)	0.187 (0.007)	0.933 (0.008)	0.971 (0.022)	0.236 (0.005)
137	STC	-0.077 (0.007)	-0.524 (0.007)	0.369 (0.007)	0.059 (0.003)	0.190 (0.007)	0.934 (0.008)	0.992 (0.022)	0.230 (0.005)
137	Bucher	-0.244 (0.005)	-0.569 (0.005)	0.082 (0.005)	0.087 (0.003)	0.253 (0.005)	0.703 (0.014)	0.998 (0.022)	0.166 (0.004)
138	MAIC	0.011 (0.006)	-0.337 (0.006)	0.359 (0.006)	0.033 (0.001)	0.145 (0.006)	0.941 (0.007)	0.981 (0.022)	0.181 (0.004)
138	STC	-0.049 (0.005)	-0.386 (0.005)	0.288 (0.005)	0.032 (0.001)	0.141 (0.005)	0.933 (0.008)	1.005 (0.023)	0.171 (0.004)
138	Bucher	-0.230 (0.004)	-0.492 (0.004)	0.032 (0.004)	0.070 (0.002)	0.235 (0.004)	0.613 (0.015)	1.019 (0.023)	0.131 (0.003)
139	MAIC	-0.012 (0.010)	-0.558 (0.010)	0.534 (0.010)	0.094 (0.004)	0.245 (0.010)	0.921 (0.009)	0.909 (0.021)	0.306 (0.007)
139	STC	-0.152 (0.010)	-0.766 (0.010)	0.461 (0.010)	0.120 (0.005)	0.277 (0.010)	0.928 (0.008)	1.008 (0.023)	0.311 (0.007)
139	Bucher	-0.216 (0.006)	-0.633 (0.007)	0.201 (0.006)	0.089 (0.003)	0.245 (0.006)	0.849 (0.011)	1.038 (0.023)	0.205 (0.005)
140	MAIC	-0.009 (0.007)	-0.428 (0.007)	0.410 (0.007)	0.051 (0.002)	0.180 (0.007)	0.934 (0.008)	0.949 (0.021)	0.225 (0.005)
140	STC	-0.137 (0.007)	-0.577 (0.007)	0.303 (0.007)	0.068 (0.003)	0.207 (0.007)	0.915 (0.009)	1.016 (0.023)	0.221 (0.005)
140	Bucher	-0.215 (0.005)	-0.533 (0.005)	0.103 (0.005)	0.071 (0.002)	0.228 (0.005)	0.758 (0.014)	1.025 (0.023)	0.158 (0.004)
141	MAIC	-0.009 (0.005)	-0.339 (0.005)	0.320 (0.005)	0.027 (0.001)	0.130 (0.005)	0.949 (0.007)	1.032 (0.023)	0.163 (0.004)
141	STC	-0.140 (0.005)	-0.470 (0.005)	0.190 (0.005)	0.047 (0.002)	0.177 (0.005)	0.873 (0.011)	1.017 (0.023)	0.165 (0.004)
141	Bucher	-0.211 (0.004)	-0.466 (0.004)	0.044 (0.004)	0.061 (0.002)	0.216 (0.004)	0.640 (0.015)	1.020 (0.023)	0.128 (0.003)
142	MAIC	-0.022 (0.009)	-0.541 (0.009)	0.497 (0.010)	0.081 (0.004)	0.228 (0.009)	0.922 (0.008)	0.929 (0.021)	0.285 (0.006)
142	STC	-0.268 (0.010)	-0.877 (0.010)	0.341 (0.010)	0.172 (0.007)	0.335 (0.010)	0.860 (0.011)	0.981 (0.022)	0.317 (0.007)
142	Bucher	-0.175 (0.006)	-0.584 (0.006)	0.235 (0.006)	0.070 (0.003)	0.213 (0.006)	0.890 (0.010)	1.052 (0.024)	0.198 (0.004)
143	MAIC	-0.013 (0.007)	-0.416 (0.007)	0.389 (0.007)	0.049 (0.002)	0.177 (0.007)	0.912 (0.009)	0.929 (0.021)	0.221 (0.005)
143	STC	-0.245 (0.007)	-0.680 (0.007)	0.190 (0.007)	0.113 (0.004)	0.280 (0.007)	0.796 (0.013)	0.967 (0.022)	0.229 (0.005)
143	Bucher	-0.175 (0.005)	-0.486 (0.005)	0.137 (0.005)	0.056 (0.002)	0.196 (0.005)	0.803 (0.013)	0.997 (0.022)	0.159 (0.004)
144	MAIC	0.001 (0.005)	-0.318 (0.005)	0.320 (0.005)	0.027 (0.001)	0.133 (0.005)	0.944 (0.007)	0.984 (0.022)	0.165 (0.004)
144	STC	-0.236 (0.005)	-0.562 (0.005)	0.089 (0.005)	0.084 (0.003)	0.249 (0.005)	0.676 (0.015)	0.983 (0.022)	0.169 (0.004)
144	Bucher	-0.177 (0.004)	-0.427 (0.004)	0.073 (0.004)	0.048 (0.002)	0.187 (0.004)	0.714 (0.014)	0.992 (0.022)	0.129 (0.003)
145	MAIC	0.017 (0.010)	-0.549 (0.010)	0.583 (0.010)	0.101 (0.005)	0.253 (0.010)	0.911 (0.009)	0.910 (0.021)	0.317 (0.007)
145	STC	0.011 (0.010)	-0.605 (0.011)	0.627 (0.010)	0.108 (0.005)	0.262 (0.010)	0.937 (0.008)	0.955 (0.021)	0.329 (0.007)
145	Bucher	-0.380 (0.007)	-0.803 (0.007)	0.044 (0.007)	0.192 (0.006)	0.384 (0.007)	0.585 (0.016)	0.986 (0.022)	0.219 (0.005)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
146	MAIC	-0.008 (0.007)	-0.439 (0.007)	0.423 (0.007)	0.052 (0.002)	0.181 (0.007)	0.937 (0.008)	0.960 (0.022)	0.229 (0.005)
146	STC	-0.003 (0.007)	-0.443 (0.007)	0.437 (0.007)	0.053 (0.002)	0.183 (0.007)	0.947 (0.007)	0.978 (0.022)	0.229 (0.005)
146	Bucher	-0.385 (0.005)	-0.707 (0.005)	-0.062 (0.005)	0.174 (0.004)	0.385 (0.005)	0.347 (0.015)	1.013 (0.023)	0.162 (0.004)
147	MAIC	0.001 (0.005)	-0.336 (0.005)	0.338 (0.006)	0.030 (0.001)	0.139 (0.005)	0.940 (0.008)	0.993 (0.022)	0.173 (0.004)
147	STC	-0.003 (0.005)	-0.333 (0.005)	0.328 (0.005)	0.027 (0.001)	0.132 (0.005)	0.958 (0.006)	1.023 (0.023)	0.165 (0.004)
147	Bucher	-0.392 (0.004)	-0.651 (0.004)	-0.133 (0.004)	0.170 (0.003)	0.392 (0.004)	0.141 (0.011)	1.028 (0.023)	0.129 (0.003)
148	MAIC	0.001 (0.010)	-0.534 (0.010)	0.536 (0.010)	0.091 (0.004)	0.238 (0.010)	0.909 (0.009)	0.906 (0.021)	0.301 (0.007)
148	STC	-0.057 (0.010)	-0.664 (0.011)	0.550 (0.010)	0.108 (0.005)	0.260 (0.010)	0.936 (0.008)	0.957 (0.022)	0.323 (0.007)
148	Bucher	-0.332 (0.007)	-0.748 (0.007)	0.084 (0.007)	0.159 (0.005)	0.342 (0.007)	0.644 (0.015)	0.962 (0.022)	0.221 (0.005)
149	MAIC	-0.003 (0.007)	-0.417 (0.007)	0.410 (0.007)	0.049 (0.002)	0.178 (0.007)	0.930 (0.008)	0.952 (0.022)	0.222 (0.005)
149	STC	-0.046 (0.007)	-0.481 (0.007)	0.389 (0.007)	0.054 (0.002)	0.183 (0.007)	0.934 (0.008)	0.977 (0.022)	0.227 (0.005)
149	Bucher	-0.337 (0.005)	-0.653 (0.005)	-0.021 (0.005)	0.139 (0.004)	0.338 (0.005)	0.441 (0.016)	1.009 (0.023)	0.160 (0.004)
150	MAIC	0.002 (0.005)	-0.324 (0.005)	0.327 (0.006)	0.029 (0.001)	0.135 (0.005)	0.933 (0.008)	0.975 (0.022)	0.170 (0.004)
150	STC	-0.043 (0.005)	-0.369 (0.005)	0.283 (0.005)	0.031 (0.001)	0.137 (0.005)	0.933 (0.008)	0.980 (0.022)	0.170 (0.004)
150	Bucher	-0.331 (0.004)	-0.584 (0.004)	-0.077 (0.004)	0.126 (0.003)	0.331 (0.004)	0.264 (0.014)	0.999 (0.022)	0.130 (0.003)
151	MAIC	-0.004 (0.009)	-0.520 (0.009)	0.512 (0.010)	0.087 (0.004)	0.236 (0.009)	0.893 (0.010)	0.893 (0.020)	0.295 (0.007)
151	STC	-0.105 (0.010)	-0.709 (0.010)	0.500 (0.010)	0.114 (0.005)	0.267 (0.010)	0.916 (0.009)	0.961 (0.022)	0.321 (0.007)
151	Bucher	-0.266 (0.007)	-0.674 (0.007)	0.143 (0.006)	0.113 (0.004)	0.283 (0.007)	0.759 (0.014)	1.007 (0.023)	0.207 (0.005)
152	MAIC	-0.010 (0.007)	-0.413 (0.007)	0.393 (0.007)	0.043 (0.002)	0.168 (0.007)	0.941 (0.007)	0.989 (0.022)	0.208 (0.005)
152	STC	-0.105 (0.007)	-0.536 (0.007)	0.326 (0.007)	0.059 (0.002)	0.195 (0.007)	0.931 (0.008)	1.004 (0.023)	0.219 (0.005)
152	Bucher	-0.278 (0.005)	-0.589 (0.005)	0.033 (0.005)	0.101 (0.003)	0.280 (0.005)	0.599 (0.015)	1.017 (0.023)	0.156 (0.003)
153	MAIC	0.007 (0.005)	-0.310 (0.005)	0.323 (0.006)	0.029 (0.001)	0.136 (0.005)	0.926 (0.008)	0.955 (0.021)	0.169 (0.004)
153	STC	-0.095 (0.005)	-0.417 (0.005)	0.228 (0.005)	0.037 (0.002)	0.152 (0.005)	0.900 (0.009)	0.986 (0.022)	0.167 (0.004)
153	Bucher	-0.265 (0.004)	-0.515 (0.004)	-0.016 (0.004)	0.086 (0.002)	0.267 (0.004)	0.458 (0.016)	1.017 (0.023)	0.125 (0.003)
154	MAIC	0.018 (0.010)	-0.530 (0.010)	0.566 (0.010)	0.093 (0.005)	0.239 (0.010)	0.921 (0.009)	0.916 (0.021)	0.305 (0.007)
154	STC	0.136 (0.010)	-0.471 (0.010)	0.742 (0.011)	0.125 (0.006)	0.278 (0.010)	0.912 (0.009)	0.949 (0.021)	0.326 (0.007)
154	Bucher	-0.565 (0.007)	-0.986 (0.007)	-0.145 (0.006)	0.365 (0.008)	0.566 (0.007)	0.226 (0.013)	1.009 (0.023)	0.213 (0.005)
155	MAIC	0.005 (0.007)	-0.414 (0.007)	0.424 (0.007)	0.045 (0.002)	0.170 (0.007)	0.943 (0.007)	1.005 (0.023)	0.213 (0.005)
155	STC	0.118 (0.007)	-0.316 (0.007)	0.552 (0.007)	0.061 (0.003)	0.198 (0.007)	0.914 (0.009)	1.024 (0.023)	0.216 (0.005)
155	Bucher	-0.566 (0.005)	-0.886 (0.005)	-0.247 (0.005)	0.345 (0.006)	0.566 (0.005)	0.059 (0.007)	1.040 (0.023)	0.157 (0.004)
156	MAIC	0.005 (0.005)	-0.322 (0.005)	0.331 (0.005)	0.028 (0.001)	0.133 (0.005)	0.936 (0.008)	0.999 (0.022)	0.167 (0.004)
156	STC	0.131 (0.005)	-0.195 (0.005)	0.457 (0.005)	0.046 (0.002)	0.171 (0.005)	0.866 (0.011)	0.987 (0.022)	0.169 (0.004)
156	Bucher	-0.562 (0.004)	-0.818 (0.004)	-0.306 (0.004)	0.333 (0.005)	0.562 (0.004)	0.014 (0.004)	0.997 (0.022)	0.131 (0.003)
157	MAIC	0.011 (0.009)	-0.509 (0.009)	0.532 (0.009)	0.081 (0.004)	0.228 (0.009)	0.927 (0.008)	0.932 (0.021)	0.285 (0.006)
157	STC	0.138 (0.010)	-0.460 (0.010)	0.736 (0.010)	0.121 (0.006)	0.278 (0.010)	0.923 (0.008)	0.957 (0.022)	0.319 (0.007)
157	Bucher	-0.477 (0.007)	-0.890 (0.007)	-0.063 (0.006)	0.272 (0.007)	0.478 (0.007)	0.396 (0.015)	0.998 (0.022)	0.211 (0.005)
158	MAIC	0.010 (0.007)	-0.395 (0.007)	0.415 (0.007)	0.050 (0.002)	0.179 (0.007)	0.917 (0.009)	0.929 (0.021)	0.223 (0.005)
158	STC	0.126 (0.007)	-0.303 (0.007)	0.556 (0.007)	0.065 (0.003)	0.206 (0.007)	0.908 (0.009)	0.987 (0.022)	0.222 (0.005)
158	Bucher	-0.471 (0.005)	-0.786 (0.005)	-0.157 (0.005)	0.249 (0.005)	0.471 (0.005)	0.153 (0.011)	0.985 (0.022)	0.163 (0.004)
159	MAIC	0.009 (0.005)	-0.310 (0.005)	0.327 (0.005)	0.026 (0.001)	0.132 (0.005)	0.953 (0.007)	1.008 (0.023)	0.161 (0.004)
159	STC	0.141 (0.005)	-0.182 (0.005)	0.463 (0.005)	0.047 (0.002)	0.176 (0.005)	0.850 (0.011)	0.991 (0.022)	0.166 (0.004)
159	Bucher	-0.470 (0.004)	-0.722 (0.004)	-0.218 (0.004)	0.236 (0.004)	0.470 (0.004)	0.043 (0.006)	1.034 (0.023)	0.124 (0.003)
160	MAIC	0.024 (0.009)	-0.491 (0.009)	0.538 (0.009)	0.080 (0.003)	0.228 (0.009)	0.913 (0.009)	0.931 (0.021)	0.282 (0.006)
160	STC	0.124 (0.010)	-0.479 (0.010)	0.727 (0.010)	0.113 (0.005)	0.268 (0.010)	0.924 (0.008)	0.984 (0.022)	0.313 (0.007)
160	Bucher	-0.380 (0.007)	-0.788 (0.007)	0.028 (0.006)	0.188 (0.005)	0.386 (0.007)	0.550 (0.016)	1.001 (0.022)	0.208 (0.005)
161	MAIC	0.001 (0.007)	-0.398 (0.007)	0.400 (0.007)	0.044 (0.002)	0.169 (0.007)	0.936 (0.008)	0.971 (0.022)	0.210 (0.005)
161	STC	0.122 (0.007)	-0.307 (0.007)	0.551 (0.007)	0.064 (0.003)	0.200 (0.007)	0.908 (0.009)	0.991 (0.022)	0.221 (0.005)
161	Bucher	-0.388 (0.005)	-0.699 (0.005)	-0.077 (0.005)	0.177 (0.004)	0.389 (0.005)	0.312 (0.015)	0.980 (0.022)	0.162 (0.004)
162	MAIC	0.012 (0.005)	-0.302 (0.005)	0.325 (0.005)	0.023 (0.001)	0.123 (0.005)	0.952 (0.007)	1.047 (0.024)	0.153 (0.003)

Table 3: Performance metrics for each method and simulation scenario. Monte Carlo standard errors for each measure are presented in parentheses. ATE: average estimated marginal treatment effect for A vs. B (is equal to the bias as the true effect is zero); LCI: average lower bound of the 95 percent confidence interval; UCI: average upper bound of the 95 percent confidence interval; MSE: mean square error; MAE: mean absolute error; Cover: coverage rate of the 95 percent confidence intervals; VR: variability ratio; ESE: empirical standard error; MAIC: matching-adjusted indirect comparison; STC: simulated treatment comparison. *(continued)*

Scenario	Method	ATE	LCI	UCI	MSE	MAE	Cover	VR	ESE
162	STC	0.122 (0.005)	-0.199 (0.005)	0.442 (0.005)	0.038 (0.002)	0.157 (0.005)	0.901 (0.009)	1.075 (0.024)	0.152 (0.003)
162	Bucher	-0.379 (0.004)	-0.628 (0.004)	-0.130 (0.004)	0.159 (0.003)	0.379 (0.004)	0.147 (0.011)	1.020 (0.023)	0.124 (0.003)

SUPPLEMENTARY APPENDIX D: SYNTHESIS SIZE IN MULTIPLE IMPUTATION MARGINALIZATION

In the simulation study in Chapter 5, we set $N^* = 1000$ for multiple imputation marginalization and adopted the same allocation ratio of the AC trial. Nevertheless, it is not clear what the sample size of each hypothetical trial should be. We now explore varying N^* under the simulation scenario with $N = 400$ and poor overlap, such that $N^* \in \{200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000\}$, while keeping the treatment allocation ratio as that of the original AC trial. Table 4 presents a summary of performance metrics for multiple imputation marginalization under different values of N^*. Monte Carlo standard errors for each performance measure are presented in parentheses. The percentage of simulation replicates that produce negative variance estimates for the marginal A vs. C treatment effect are also reported. There are hardly any changes to the bias, variability ratio and coverage rates when increasing the synthesis size above its original value of $N^* = 1000$. The lack of sensitivity to further increases of N^* suggests that any potential performance gains do not offset the computational cost of the change. Conversely, there is less stability with smaller syntheses. Variance estimates seem to underestimate variability, coverage rates are more conservative, and there is some risk of negative variance estimates ($N^* = 200$).

In multiple imputation marginalization, the inferential framework in the analysis stage depends on reasonably large values of N^* — the posterior distributions used for pooling (Equations 28, 29 and 30) have been derived under certain normality assumptions, where the size of the synthetic datasets is relevant. Therefore, one would expect better inferences with higher values of N^*.

Synthesis size, N^*	Bias	Variability ratio	Coverage rate	% $\hat{V}(\hat{\Delta}_{10}^{(2)}) < 0$
200	-0.034 (0.010)	0.919 (0.015)	0.927 (0.006)	0.5
400	-0.018 (0.009)	0.962 (0.015)	0.943 (0.005)	0
600	-0.016 (0.009)	0.962 (0.015)	0.942 (0.005)	0
800	-0.015 (0.009)	0.977 (0.016)	0.945 (0.005)	0
1000	-0.014 (0.009)	0.981 (0.016)	0.949 (0.005)	0
1200	-0.013 (0.009)	0.984 (0.016)	0.950 (0.005)	0
1400	-0.013 (0.009)	0.983 (0.016)	0.950 (0.005)	0
1600	-0.011 (0.009)	0.984 (0.016)	0.952 (0.005)	0
1800	-0.012 (0.009)	0.988 (0.016)	0.951 (0.005)	0
2000	-0.011 (0.009)	0.985 (0.016)	0.950 (0.005)	0

Table 4: Simulation results for multiple imputation marginalization varying the synthesis size N^*.

SUPPLEMENTARY APPENDIX E: CHAPTER 3 EXAMPLE CODE

Example R code implementing MAIC, the conventional version of STC and the Bucher method on a simulated example is provided in this appendix. The code and data are available at https://github.com/remiroazocar/population_adjustment_simstudy in the `Example` subdirectory. Full code for the simulation study in Chapter 3 is available in the online repository. The simulation study and the provided example use survival outcomes, with a Cox proportional hazards regression as the outcome model of interest in the analysis.

MATCHING-ADJUSTED INDIRECT COMPARISON

```
library("survival") # required for weighted Cox regression

AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data

N <- nrow(AC.IPD) # number of subjects in AC
X.EM <- AC.IPD[,c("X1","X2")] # AC effect modifiers
bar.X.EM.BC <- BC.ALD[,c("mean_X1", "mean_X2")] # BC effect modifier means
K.EM <- ncol(X.EM) # number of effect modifiers

# center the AC effect modifiers on the BC means
for (k in 1:K.EM) {
  X.EM[,k] <- X.EM[,k] - bar.X.EM.BC[,k]
}

# objective function to be minimized for weight estimation
Q <- function(alpha, X.EM) {
  return(sum(exp(X.EM %*% alpha)))
}

alpha <- rep(1,K.EM) # arbitrary starting point for the optimiser
# objective function minimized using BFGS
Q.min <- optim(fn=Q, X.EM=as.matrix(X.EM), par=alpha, method="BFGS")
hat.alpha <- Q.min$par # finite solution is the logistic regression parameters
log.hat.w <- rep(0, N)
for (k in 1:K.EM) {
  log.hat.w <- log.hat.w + hat.alpha[k]*X.EM[,k]
}
hat.w <- exp(log.hat.w) # estimated weights
aess <- sum(hat.w)^2/sum(hat.w^2) # approximate effective sample size
```

```
# fit weighted Cox proportional hazards model using robust=TRUE for robust variance
outcome.fit <- coxph(Surv(time, status)~trt, robust=TRUE, weights=hat.w,data=AC.IPD
    )

# fitted treatment coefficient is relative effect for A vs. C
hat.Delta.AC <- summary(outcome.fit)$coef[1]
hat.var.Delta.AC <- vcov(outcome.fit)[[1]] # estimated variance for A vs. C
hat.Delta.BC <- with(BC.ALD, logHR_B) # B vs. C
hat.var.Delta.BC <- with(BC.ALD, var_logHR_B)
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC # A vs. B
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

CONVENTIONAL SIMULATED TREATMENT COMPARISON

```
library("survival") # required for standard Cox regression

AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data

# fit regression of outcome on the baseline characteristics and treatment
# effect modifiers are centered at the mean BC values
# purely prognostic variables are included but not centered
outcome.fit <- coxph(Surv(time, status)~X3+X4+trt*I(X1-BC.ALD$mean_X1)+trt*I(X2-BC.
    ALD$mean_X2),
                     data=AC.IPD)

# estimated treatment coefficient is relative effect for A vs. C
hat.Delta.AC <- coef(outcome.fit)["trt"]
hat.var.Delta.AC <- vcov(outcome.fit)["trt", "trt"] # estimated variance for A vs.
    C
hat.Delta.BC <- with(BC.ALD, logHR_B) # B vs. C
hat.var.Delta.BC <- with(BC.ALD, var_logHR_B)
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC # A vs. B
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

BUCHER METHOD

```
library("survival") # required for standard Cox regression

AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data

# simple regression of outcome on treatment
outcome.fit <- coxph(Surv(time, status)~trt, data=AC.IPD)

# fitted treatment coefficient is relative effect for A vs. C
hat.Delta.AC <- coef(outcome.fit)["trt"]
hat.var.Delta.AC <- vcov(outcome.fit)["trt", "trt"] # estimated variance for A vs.
    C
hat.Delta.BC <- with(BC.ALD, logHR_B) # B vs. C
hat.var.Delta.BC <- with(BC.ALD, var_logHR_B)
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC # A vs. B
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

SUPPLEMENTARY APPENDIX F: CHAPTER 5 EXAMPLE CODE

Example `R` code implementing MAIC, the conventional STC, maximum-likelihood parametric G-computation, Bayesian parametric G-computation and MIM on a simulated dataset is provided below. The code and data are available at https://github.com/remiroazocar/marginalized_indirect_comparisons_simstudy in the `Example` subdirectory. Full code for implementing the simulation study in Chapter 5 is available in the online repository.

The simulation study and the provided example use binary outcomes and a logistic regression outcome model. Nevertheless, all methods are general-purpose frameworks that, under a generalized linear modeling formulation, can be easily adapted to different outcome models, outcome types, and scalar measures of treatment effect. The code below can be altered by changing the link function in the outcome model. For instance: (1) for a normal linear regression, by setting `family=gaussian` in the arguments to the `glm` (or `stanglm`) function, such that the link is the identity function (for the weighted outcome model, in the case of MAIC, and for the first- and second-stage regressions, in the case of MIM); (2) for a Gamma regression, set `family=Gamma`, and, for parametric G-computation, transform the predicted marginal outcome means to the linear predictor scale using the "negative inverse" link ($g(\mu) = -\mu^{-1}$, for outcome mean μ); (3) for a Poisson regression, set `family=poisson`, and, for parametric G-computation, transform the marginal outcome means to the linear predictor scale using the log link ($g(\mu) = \ln(\mu)$); and (4) for an inverse Gaussian regression, set `family=inverse.gaussian`, and, for parametric G-computation, transform the marginal outcome means to the linear predictor scale using the "inverse squared" link ($g(\mu) = \mu^{-2}$).

At the end of this appendix, I provide `R` code implementing maximum-likelihood parametric G-computation on a simulated example with survival outcomes and Cox regression as the outcome model.

MATCHING-ADJUSTED INDIRECT COMPARISON

```
library("boot") # for non-parametric bootstrap

AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data

set.seed(555) # set seed for reproducibility

# objective function to be minimized for standard method of moments
Q <- function(alpha, X.EM) {
  return(sum(exp(X.EM %*% alpha)))
}

# function to be bootstrapped
maic.boot <- function(data, indices) {
  dat <- data[indices,] # AC bootstrap sample
  N <- nrow(dat) # number of subjects in sample
  x.EM <- dat[,c("X1","X2")] # AC effect modifiers
  # BC effect modifier means, assumed fixed
  theta <- BC.ALD[c("mean.X1", "mean.X2")]
  K.EM <- ncol(x.EM) # number of effect modifiers
  # center the AC effect modifiers on the BC means
  x.EM$X1 <- x.EM$X1 - theta$mean.X1
  x.EM$X2 <- x.EM$X2 - theta$mean.X2
  # MAIC weight estimation using method of moments
  alpha <- rep(1,K.EM) # arbitrary starting point for the optimizer
  # objective function minimized using BFGS
  Q.min <- optim(fn=Q, X.EM=as.matrix(x.EM), par=alpha, method="BFGS")
  # finite solution is the logistic regression parameters
  hat.alpha <- Q.min$par
  log.hat.w <- rep(0, N)
  for (k in 1:K.EM) {
    log.hat.w <- log.hat.w + hat.alpha[k]*x.EM[,k]
  }
  hat.w <- exp(log.hat.w) # estimated weights
  # fit weighted logistic regression model using glm
  outcome.fit <- glm(y~trt, family="quasibinomial", weights=hat.w,
                     data=dat)
  # fitted treatment coefficient is marginal effect for A vs. C
  hat.Delta.AC <- coef(outcome.fit)["trt"]
  return(hat.Delta.AC)
}

# non-parametric bootstrap with 1000 resamples
boot.object <- boot::boot(data=AC.IPD, statistic=maic.boot, R=1000)
```

```
# bootstrap mean of marginal A vs. C treatment effect estimate
hat.Delta.AC <- mean(boot.object$t)
# bootstrap variance of A vs. C treatment effect estimate
hat.var.Delta.AC <- var(boot.object$t)
# B vs. C marginal treatment effect from reported event counts
hat.Delta.BC <- with(BC.ALD, log(y.B.sum*(N.C-y.C.sum)/
                                 (y.C.sum*(N.B-y.B.sum))))
# B vs. C marginal effect variance using the delta method
hat.var.Delta.BC <- with(BC.ALD, 1/y.C.sum+1/(N.C-y.C.sum)+
                         1/y.B.sum+1/(N.B-y.B.sum))
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC # A vs. B
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

CONVENTIONAL SIMULATED TREATMENT COMPARISON

```
AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data

# fit regression model of outcome on treatment and covariates
# IPD effect modifiers centered at the mean BC values
# purely prognostic variables are included but not centered
outcome.model <- glm(y~X3+X4+trt*I(X1-BC.ALD$mean.X1)+
                       trt*I(X2-BC.ALD$mean.X2),
                     data=AC.IPD, family=binomial)
# fitted treatment coefficient is relative A vs. C conditional effect
hat.Delta.AC <- coef(outcome.model)["trt"]
# estimated variance for A vs. C from model fit
hat.var.Delta.AC <- vcov(outcome.model)["trt", "trt"]
# B vs. C marginal treatment effect estimated from reported event counts
hat.Delta.BC <- with(BC.ALD, log(y.B.sum*(N.C-y.C.sum)/
                                 (y.C.sum*(N.B-y.B.sum))))
# B vs. C marginal treatment effect variance using the delta method
hat.var.Delta.BC <- with(BC.ALD, 1/y.C.sum+1/(N.C-y.C.sum)+
                         1/y.B.sum+1/(N.B-y.B.sum))
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC # A vs. B
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

MAXIMUM-LIKELIHOOD PARAMETRIC G-COMPUTATION

```
library("copula") # for simulating BC covariates from Gaussian copula
library("boot") # for non-parametric bootstrap

AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data

set.seed(555) # set seed for reproducibility

# matrix of pairwise correlations between IPD covariates
rho <- cor(AC.IPD[,c("X1","X2","X3","X4")])
#  covariate simulation for BC trial using copula package
cop <- normalCopula(param=c(rho[1,2],rho[1,3],rho[1,4],rho[2,3],
                            rho[2,4],rho[3,4]),
                    dim=4, dispstr="un") # AC IPD pairwise correlations
# sample covariates from approximate joint distribution using copula
mvd <- mvdc(copula=cop, margins=c("norm", "norm", # Gaussian marginals
                                  "norm", "norm"),
            # BC covariate means and standard deviations
            paramMargins=list(list(mean=BC.ALD$mean.X1, sd=BC.ALD$sd.X1),
                              list(mean=BC.ALD$mean.X2, sd=BC.ALD$sd.X2),
                              list(mean=BC.ALD$mean.X3, sd=BC.ALD$sd.X3),
                              list(mean=BC.ALD$mean.X4, sd=BC.ALD$sd.X4)))
# simulated BC pseudo-population of size 1000
x_star <- as.data.frame(rMvdc(1000, mvd))
colnames(x_star) <- c("X1", "X2", "X3", "X4")
# this function will be bootstrapped
gcomp.ml <- function(data, indices) {
  dat = data[indices,]
  # outcome logistic regression fitted to IPD using maximum likelihood
  outcome.model <- glm(y~X3+X4+trt*X1+trt*X2, data=dat, family=binomial)
  # counterfactual datasets
  data.trtA <- data.trtC <- x_star
  # intervene on treatment while keeping set covariates fixed
  data.trtA$trt <- 1 # dataset where everyone receives treatment A
  data.trtC$trt <- 0 # dataset where all observations receive C
  # predict counterfactual event probs, conditional on treatment/covariates
  hat.mu.A.i <- predict(outcome.model, type="response", newdata=data.trtA)
  hat.mu.C.i <- predict(outcome.model, type="response", newdata=data.trtC)
  hat.mu.A <- mean(hat.mu.A.i) # (marginal) mean probability prediction under A
  hat.mu.C <- mean(hat.mu.C.i) # (marginal) mean probability prediction under C
  # marginal A vs. C log-odds ratio (mean difference in expected log-odds)
  # estimated by transforming from probability to linear predictor scale
  hat.Delta.AC <- log(hat.mu.A/(1-hat.mu.A)) - log(hat.mu.C/(1-hat.mu.C))
  # hat.Delta.AC <- qlogis(hat.mu.A) - qlogis(hat.mu.C)
```

```
  return(hat.Delta.AC)
}
# non-parametric bootstrap with 1000 resamples
boot.object <- boot::boot(data=AC.IPD, statistic=gcomp.ml, R=1000)
# bootstrap mean of marginal A vs. C treatment effect estimate
hat.Delta.AC <- mean(boot.object$t)
# bootstrap variance of A vs. C treatment effect estimate
hat.var.Delta.AC <- var(boot.object$t)
# marginal log-odds ratio for B vs. C from reported event counts
hat.Delta.BC <- with(BC.ALD,log(y.B.sum*(N.C-y.C.sum)/
                                (y.C.sum*(N.B-y.B.sum))))
# variance of B vs. C using delta method
hat.var.Delta.BC <- with(BC.ALD,1/y.C.sum+1/(N.C-y.C.sum)+
                           1/y.B.sum+1/(N.B-y.B.sum))
# marginal treatment effect for A vs. B
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC
# variance for A vs. B
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

BAYESIAN PARAMETRIC G-COMPUTATION

```
library("copula") # for simulating BC covariates from Gaussian copula
# for outcome regression and drawing outcomes from posterior predictive dist.
library("rstanarm")

AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data

set.seed(555) # set seed for reproducibility

# matrix of pairwise correlations between IPD covariates
rho <- cor(AC.IPD[,c("X1","X2","X3","X4")])
#  covariate simulation for BC trial using copula package
cop <- normalCopula(param=c(rho[1,2],rho[1,3],rho[1,4],rho[2,3],
                            rho[2,4],rho[3,4]),
                    dim=4, dispstr="un") # AC IPD pairwise correlations
# sample covariates from approximate joint distribution using copula
mvd <- mvdc(copula=cop, margins=c("norm", "norm", # Gaussian marginals
                                  "norm", "norm"),
            # BC covariate means and standard deviations
            paramMargins=list(list(mean=BC.ALD$mean.X1, sd=BC.ALD$sd.X1),
                              list(mean=BC.ALD$mean.X2, sd=BC.ALD$sd.X2),
```

```
                              list(mean=BC.ALD$mean.X3, sd=BC.ALD$sd.X3),
                              list(mean=BC.ALD$mean.X4, sd=BC.ALD$sd.X4)))
# simulated BC pseudo-population of size 1000
x_star <- as.data.frame(rMvdc(1000, mvd))
colnames(x_star) <- c("X1", "X2", "X3", "X4")
# outcome logistic regression fitted to IPD using MCMC (Stan)
outcome.model <- stan_glm(y~X3+X4+trt*X1+trt*X2, data=AC.IPD,
                          family=binomial, algorithm="sampling",
                          iter=4000, warmup=2000, chains=2)
# counterfactual datasets
data.trtA <- data.trtC <- x_star
# intervene on treatment while keeping set covariates fixed
data.trtA$trt <- 1 # dataset where everyone receives treatment A
data.trtC$trt <- 0 # dataset where all observations receive C
# draw binary responses from posterior predictive distribution
# matrix of posterior predictive draws under A
y.star.A <- posterior_predict(outcome.model, newdata=data.trtA)
# matrix of posterior predictive draws under C
y.star.C <- posterior_predict(outcome.model, newdata=data.trtC)
# compute marginal log-odds ratio for A vs. C for each MCMC sample
# by transforming from probability to linear predictor scale
hat.delta.AC <- qlogis(rowMeans(y.star.A)) - qlogis(rowMeans(y.star.C))
hat.Delta.AC <- mean(hat.delta.AC) # average over samples
hat.var.Delta.AC <- var(hat.delta.AC) # sample variance
# B vs. C from reported aggregate event counts in contingency table
hat.Delta.BC <- with(BC.ALD, log(y.B.sum*(N.C-y.C.sum)/
                                   (y.C.sum*(N.B-y.B.sum))))
# B vs. C variance using the delta method
hat.var.Delta.BC <- with(BC.ALD, 1/y.C.sum+1/(N.C-y.C.sum)+
                           1/y.B.sum+1/(N.B-y.B.sum))
# marginal treatment effect for A vs. B
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC
# A vs. B variance
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

MULTIPLE IMPUTATION MARGINALIZATION

```
library("copula") # for simulating BC covariates from Gaussian copula
library("rstanarm") # for MCMC posterior sampling in data synthesis stage

AC.IPD <- read.csv("Example/AC_IPD.csv") # load AC patient-level data
BC.ALD <- read.csv("Example/BC_ALD.csv") # load BC aggregate-level data
```

```
set.seed(555) # set seed for reproducibility

# hyper-parameter settings
M <- 1000 # number of syntheses used in analysis stage
N_star <- 1000 # size of syntheses or simulated BC pseudo-populations
alloc <- 2/3 # 2:1 A:C allocation ratio in synthesis
# MCMC info
n.chains <- 2 # number of Markov chains for MCMC
warmup <- 2000 # discarded warmup/burn-in iterations per chain
iters <- 4000 # total iterations per chain (including warmup)

## SYNTHESIS STAGE (as per Bayesian G-computation) ##
# matrix of pairwise correlations between IPD covariates
rho <- cor(AC.IPD[,c("X1","X2","X3","X4")])
#  covariate simulation for BC trial using copula package
cop <- normalCopula(param=c(rho[1,2],rho[1,3],rho[1,4],rho[2,3],
                            rho[2,4],rho[3,4]),
                    dim=4, dispstr="un") # AC IPD pairwise correlations
# sample covariates from approximate joint distribution using copula
mvd <- mvdc(copula=cop, margins=c("norm", "norm", # Gaussian marginals
                                  "norm", "norm"),
            # BC covariate means and standard deviations
            paramMargins=list(list(mean=BC.ALD$mean.X1, sd=BC.ALD$sd.X1),
                              list(mean=BC.ALD$mean.X2, sd=BC.ALD$sd.X2),
                              list(mean=BC.ALD$mean.X3, sd=BC.ALD$sd.X3),
                              list(mean=BC.ALD$mean.X4, sd=BC.ALD$sd.X4)))
# simulated BC pseudo-population of size N_star
x_star <- as.data.frame(rMvdc(N_star, mvd))
colnames(x_star) <- c("X1", "X2", "X3", "X4")
# first-stage logistic regression fitted to IPD using MCMC (Stan)
outcome.model <- stan_glm(y~X3+X4+trt*X1+trt*X2,
                          data=AC.IPD, family=binomial,
                          algorithm="sampling", iter=iters,
                          warmup=warmup, chains=n.chains,
                          # thin to use M independent samples in analysis
                          thin=(n.chains*(iters-warmup))/M)
# tratment assignment in synthesis
N_active <- round(N_star*alloc) # number of patients in synthesis under A
N_control <- N_star - N_active # number of patients in synthesis under C
trt_star <- c(rep(1,N_active), rep(0,N_control))
x_star$trt <- trt_star
# draw binary outcomes from posterior predictive distribution
y_star <- posterior_predict(outcome.model, newdata=x_star)

## ANALYSIS stage ##
# second-stage regression (marginal structural model) on each synthesis
```

```
reg2.fits <- lapply(1:M, function(m) glm(y_star[m,]~trt_star,
                                         family=binomial))
# treatment coefficient is marginal effect for A vs. C in m-th synthesis
hats_delta_AC <- unlist(lapply(reg2.fits,
                               function(fit) coef(fit)["trt_star"][[1]]))
# point estimates of variance for A vs. C
hats_v <- unlist(lapply(reg2.fits,
                        function(fit) vcov(fit)["trt_star", "trt_star"]))
# quantities originally defined by Rubin (1987) for multiple imputation
bar_delta_AC <- mean(hats_delta_AC) # average of point estimates
bar_v <- mean(hats_v) # within variance (average of point estimates of variance)
# between variance (sample variance of point estimates)
b <- var(hats_delta_AC)
# pooling + indirect comparison (combining rules)
# average of point estimates is the marginal effect for A vs. C
hat.Delta.AC <- bar_delta_AC
# variance combining rule for A vs. C
hat.var.Delta.AC <- (1+(1/M))*b-bar_v
# B vs. C from reported aggregate event counts in contingency table
hat.Delta.BC <- with(BC.ALD, log(y.B.sum*(N.C-y.C.sum)/
                                   (y.C.sum*(N.B-y.B.sum))))
# B vs. C variance using the delta method
hat.var.Delta.BC <- with(BC.ALD, 1/y.C.sum+1/(N.C-y.C.sum)+
                           1/y.B.sum+1/(N.B-y.B.sum))
# marginal treatment effect for A vs. B
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC
# A vs. B variance
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

COX REGRESSION: MAXIMUM-LIKELIHOOD PARAMETRIC G-COMPUTATION

Below, I provide example R code implementing parametric G-computation with survival outcomes and Cox regression as the outcome model. We use maximum-likelihood estimation to fit the multivariable Cox regression, then predicting the outcomes on the BC population. Variance estimation for the marginal A vs. C treatment effect is performed by resampling via the ordinary non-parametric bootstrap with replacement.

Parametric Bayesian G-computation would follow a similar approach, and would involve drawing the marginal survival probabilities under each treatment from their posterior predictive distribution. Implementing Bayesian parametric G-computation in the Cox regression scenario is a research priority.

```
library("survival") # to fit Cox proportional hazards regression
```

```
library("copula") # for simulating BC covariates from Gaussian copula
library("boot") # for non-parametric bootstrap

AC.IPD <- read.csv("Example/Survival/AC_IPD_survival.csv") # load AC patient-level
    data
BC.ALD <- read.csv("Example/Survival/BC_ALD_survival.csv") # load BC aggregate-
    level data

set.seed(555) # set seed for reproducibility

# matrix of pairwise correlations between IPD covariates
rho <- cor(AC.IPD[,c("X1","X2","X3","X4")])
#  covariate simulation for BC trial using copula package
cop <- normalCopula(param=c(rho[1,2],rho[1,3],rho[1,4],rho[2,3],
                            rho[2,4],rho[3,4]),
                    dim=4, dispstr="un") # AC IPD pairwise correlations
# sample covariates from approximate joint distribution using copula
mvd <- mvdc(copula=cop, margins=c("norm", "norm", # Gaussian marginals
                                  "norm", "norm"),
            # BC covariate means and standard deviations
            paramMargins=list(list(mean=BC.ALD$mean.X1, sd=BC.ALD$sd.X1),
                              list(mean=BC.ALD$mean.X2, sd=BC.ALD$sd.X2),
                              list(mean=BC.ALD$mean.X3, sd=BC.ALD$sd.X3),
                              list(mean=BC.ALD$mean.X4, sd=BC.ALD$sd.X4)))
# simulated BC pseudo-population of size 1000
x_star <- as.data.frame(rMvdc(1000, mvd))
colnames(x_star) <- c("X1", "X2", "X3", "X4")

# function to be resampled by non-parametric bootstrap
gcomp.ml <- function(data, indices) {
  dat = data[indices,]
  # outcome Cox regression model fitted to IPD using maximum likelihood
  outcome.model <- coxph(Surv(time, status)~trt*X1+trt*X2+X3+X4, data=dat)
  # event time selected for unit 50 (random selection)
  unit.time <- 50
  # estimated cumulative baseline hazard
  hat.H0 <- basehaz(outcome.model)[unit.time,1]
  # counterfactual datasets (two hypothetical worlds)
  data.trtA <- data.trtC <- x_star
  # intervene on treatment while keeping set covariates fixed
  data.trtA$trt <- 1 # dataset where everyone receives treatment A
  data.trtC$trt <- 0 # dataset where all observations receive C
  # linear predictor where everyone receives treatment A
  LP.A <- with(outcome.model, x_star$X1*(coefficients["X1"] + coefficients["trt:X1
      "]) +
                 x_star$X2*(coefficients["X2"] + coefficients["trt:X2"]) +
                 x_star$X3*coefficients["X3"] + x_star$X4*coefficients["X4"] +
```

```
                coefficients["trt"])
  # linear predictor where all observations receive treatment C
  LP.C <- with(outcome.model, x_star$X1*coefficients["X1"] + x_star$X2*coefficients
      ["X2"] +
                x_star$X3*coefficients["X3"] + x_star$X4*coefficients["X4"])
  # predict individual survival probabilities, conditional on treatment/covariates
  hat.S.A.i <- exp(-hat.H0)^exp(LP.A)
  hat.S.C.i <- exp(-hat.H0)^exp(LP.C)
  # mean survival probability prediction under each treatment
  hat.P.A <- mean(hat.S.A.i)
  hat.P.C <- mean(hat.S.C.i)
  # estimate marginal A vs. B log hazard ratio (mean difference in expected log
      hazard)
  # by transforming from survival probability to linear predictor scale
  hat.Delta.AC <- log(-log(hat.P.A)) - log(-log(hat.P.C))
  return(hat.Delta.AC)
}
# non-parametric bootstrap with 1000 resamples (ignore warnings)
boot.object <- boot::boot(data=AC.IPD, statistic=gcomp.ml, R=1000)
# bootstrap mean of marginal A vs. C treatment effect estimate
hat.Delta.AC <- mean(boot.object$t)
# bootstrap variance of A vs. C treatment effect estimate
hat.var.Delta.AC <- var(boot.object$t)
# marginal log hazard ratio for B vs. C reported in BC article
hat.Delta.BC <- BC.ALD$logHR_B
# variance of B vs. C in aggregate outcomes in published article
hat.var.Delta.BC <- BC.ALD$var_logHR_B
# marginal treatment effect for A vs. B
hat.Delta.AB <- hat.Delta.AC - hat.Delta.BC
# variance for A vs. B
hat.var.Delta.AB <- hat.var.Delta.AC + hat.var.Delta.BC
# construct Wald-type normal distribution-based confidence interval
uci.Delta.AB <- hat.Delta.AB + qnorm(0.975)*sqrt(hat.var.Delta.AB)
lci.Delta.AB <- hat.Delta.AB + qnorm(0.025)*sqrt(hat.var.Delta.AB)
```

BIBLIOGRAPHY

[1] Rick A Vreman, Huseyin Naci, Wim G Goettsch, Aukje K Mantel-Teeuwisse, Sebastian G Schneeweiss, Hubert GM Leufkens, and Aaron S Kesselheim. Decision making under uncertainty: comparing regulatory and health technology assessment reviews of medicines in the united states and europe. *Clinical Pharmacology & Therapeutics*, 108(2):350–357, 2020.

[2] AM Glenny, DG Altman, F Song, C Sakarovitch, JJ Deeks, R D'amico, M Bradburn, and AJ Eastwood. Indirect comparisons of competing interventions. 2005.

[3] Robert Temple and Susan S Ellenberg. Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: ethical and scientific issues. *Annals of internal medicine*, 133(6):455–463, 2000.

[4] John E Paul and Paul Trueman. 'fourth hurdle reviews', nice, and database applications. *Pharmacoepidemiology and drug safety*, 10(5):429–438, 2001.

[5] Alex Sutton, AE Ades, Nicola Cooper, and Keith Abrams. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics*, 26(9):753–767, 2008.

[6] Sofia Dias, Alex J Sutton, AE Ades, and Nicky J Welton. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617, 2013.

[7] Heiner C Bucher, Gordon H Guyatt, Lauren E Griffith, and Stephen D Walter. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of clinical epidemiology*, 50(6):683–691, 1997.

[8] Lesley A Stewart and Jayne F Tierney. To ipd or not to ipd? advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the health professions*, 25(1):76–97, 2002.

[9] David M Phillippo, Anthony E Ades, Sofia Dias, Stephen Palmer, Keith R Abrams, and Nicky J Welton. Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical Decision Making*, 38(2):200–211, 2018.

[10] James E Signorovitch, Eric Q Wu, P Yu Andrew, Charles M Gerrits, Evan Kantor, Yanjun Bao, Shiraz R Gupta, and Parvez M Mulani. Comparative effectiveness without head-to-head trials. *Pharmacoeconomics*, 28(10):935–945, 2010.

[11] James Signorovitch, M Haim Erder, Jipan Xie, Vanja Sikirica, Mei Lu, Paul S Hodgkins, and Eric Q Wu. Comparative effectiveness research using matching-adjusted indirect comparison: an application to treatment with guanfacine extended release or atomoxetine in children with attention-deficit/hyperactivity disorder and comorbid oppositional defiant disorder. *pharmacoepidemiology and drug safety*, 21:130–137, 2012.

[12] James E Signorovitch, Vanja Sikirica, M Haim Erder, Jipan Xie, Mei Lu, Paul S Hodgkins, Keith A Betts, and Eric Q Wu. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value in Health*, 15(6):940–947, 2012.

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

[14] J Jaime Caro and K Jack Ishak. No head-to-head trial? simulate the missing arms. *Pharmacoeconomics*, 28(10):957–967, 2010.

[15] Zhiwei Zhang. Covariate-adjusted putative placebo analysis in active-controlled clinical trials. *Statistics in Biopharmaceutical Research*, 1(3):279–290, 2009.

[16] Olli S Miettinen. Standardization of risk ratios. *American Journal of Epidemiology*, 96(6):383–388, 1972.

[17] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

[18] David Phillippo, Tony Ades, Sofia Dias, Stephen Palmer, Keith R Abrams, and Nicky Welton. Nice dsu technical support document 18: methods for population-adjusted indirect comparisons in submissions to nice. 2016.

[19] K Jack Ishak, Irina Proskorovsky, and Agnes Benedict. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics*, 33(6):537–549, 2015.

[20] John W Stevens, Christine Fletcher, Gerald Downey, and Anthea Sutton. A review of methods for comparing treatments evaluated in studies that form disconnected networks of evidence. *Research synthesis methods*, 9(2):148–162, 2018.

[21] H Thom, SM Jugl, E Palaka, and S Jawla. Matching adjusted indirect comparisons to assess comparative effectiveness of therapies: usage in scientific literature and health technology appraisals. *Value in Health*, 19(3):A100–A101, 2016.

[22] Gianluca Baio. *Bayesian methods in health economics*. CRC Press, 2012.

[23] Karl Claxton, Mark Sculpher, Chris McCabe, Andrew Briggs, Ron Akehurst, Martin Buxton, John Brazier, and Tony O'Hagan. Probabilistic sensitivity analysis for nice technology assessment: not an optional extra. *Health economics*, 14(4):339–347, 2005.

[24] Sarah Kühnast, Julia Schiffner-Rohe, Jörg Rahnenführer, and Friedhelm Leverkus. Evaluation of adjusted and unadjusted indirect comparison methods in benefit assessment. *Methods of information in medicine*, 56(03):261–267, 2017.

[25] Helmut Petto, Zbigniew Kadziola, Alan Brnabic, Daniel Saure, and Mark Belger. Alternative weighting approaches for anchored matching-adjusted indirect comparisons via a common comparator. *Value in Health*, 22(1):85–91, 2019.

[26] David Cheng, Rajeev Ayyagari, and James Signorovitch. The statistical performance of matching-adjusted indirect comparisons. *arXiv preprint arXiv:1910.06449*, 2019.

[27] Anthony James Hatswell, Nick Freemantle, and Gianluca Baio. The effects of model misspecification in unanchored matching-adjusted indirect comparison (maic): Results of a simulation study. *Value in Health*, 2020.

[28] M Belger, A Brnabic, Z Kadziola, H Petto, and D Faries. Inclusion of multiple studies in matching adjusted indirect comparisons (maic). *Value in Health*, 18(3):A33, 2015.

[29] Joy Leahy and Cathal Walsh. Assessing the impact of a matching-adjusted indirect comparison in a bayesian network meta-analysis. *Research synthesis methods*, 2019.

[30] David M Phillippo, Sofia Dias, Ahmed Elsada, AE Ades, and Nicky J Welton. Population adjustment methods for indirect comparisons: A review of national institute for health and care excellence technology appraisals. *International journal of technology assessment in health care*, pages 1–8, 2019.

[31] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[32] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

[33] Keisuke Hirano and Guido W Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278, 2001.

[34] David M Phillippo, Sofia Dias, AE Ades, and Nicky J Welton. Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. *Statistics in Medicine*, 2020.

[35] Dan Jackson, Kirsty Rhodes, and Mario Ouwens. Alternative weighting schemes when performing matching-adjusted indirect comparisons. *Research Synthesis Methods*, 2020.

[36] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[37] Romain Neugebauer and Mark van der Laan. Why prefer double robust estimators in causal inference? *Journal of statistical planning and inference*, 129(1-2):405–426, 2005.

[38] David M Phillippo, Sofia Dias, AE Ades, Mark Belger, Alan Brnabic, Alexander Schacht, Daniel Saure, Zbigniew Kadziola, and Nicky J Welton. Multilevel network meta-regression for population-adjusted treatment comparisons. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2020.

[39] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. Conflating marginal and conditional treatment effects: Comments on "assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study". *Statistics in Medicine*, 40(11):2753–2758, 2021.

[40] Walter W Hauck, Sharon Anderson, and Sue M Marcus. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled clinical trials*, 19(3):249–256, 1998.

[41] Issa J Dahabreh, Lucia C Petito, Sarah E Robertson, Miguel A Hernán, and Jon A Steingrimsson. Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, 31(3):334–344, 2020.

[42] James Robins. 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.

[43] James Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of chronic diseases*, 40:139S–161S, 1987.

[44] Kelly L Moore and Mark J van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in medicine*, 28(1):39–64, 2009.

[45] Peter C Austin. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of clinical epidemiology*, 63(1):2–6, 2010.

[46] Michael Rosenblum and Mark J Van Der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The international journal of biostatistics*, 6(1), 2010.

[47] Zhiwei Zhang. Estimating a marginal causal odds ratio subject to confounding. *Communications in Statistics-Theory and methods*, 38(3):309–321, 2008.

[48] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[49] James M Robins, Steven D Mark, and Whitney K Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

[50] Rhian Daniel, Jingjing Zhang, and Daniel Farewell. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 63(3):528–557, 2021.

[51] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[52] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

[53] Miguel A Hernán and James M Robins. Causal inference: what if, 2020.

[54] Antonio Remiro-Azócar. Target estimands for population-adjusted indirect comparisons. *arXiv preprint arXiv:2112.08023*, 2021.

[55] Charles F Manski. Meta-analysis for medical decisions. 2019.

[56] Tyler J VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009.

[57] Kenneth J Rothman, Sander Greenland, and Alexander M Walker. Concepts of interaction. *American journal of epidemiology*, 112(4):467–470, 1980.

[58] Fujian Song, Douglas G Altman, Anne-Marie Glenny, and Jonathan J Deeks. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *Bmj*, 326(7387):472, 2003.

[59] Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1):107–115, 2010.

[60] Holger L Kern, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, 9(1):103–127, 2016.

[61] Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778, 2015.

[62] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[63] Sofia Dias, Nicky J Welton, Alex J Sutton, Deborah M Caldwell, Guobing Lu, and AE23804508 Ades. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Medical Decision Making*, 33(5):641–656, 2013.

[64] Areti Angeliki Veroniki, Sharon E Straus, Charlene Soobiah, Meghan J Elliott, and Andrea C Tricco. A scoping review of indirect comparison methods and applications using individual patient data. *BMC medical research methodology*, 16(1):47, 2016.

[65] K Ndirangu, V Tongbram, and D Shah. Trends in the use of matching-adjusted indirect comparisons in published literature and nice technology assessments: A systematic review. *Value in Health*, 19(3):A99–A100, 2016.

[66] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[67] Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 75(1):45–49, 2004.

[68] Halbert White et al. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica*, 48(4):817–838, 1980.

[69] Frank Windmeijer. A finite sample correction for the variance of linear efficient two-step gmm estimators. *Journal of econometrics*, 126(1):25–51, 2005.

[70] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[71] David M Phillippo, Sofia Dias, AE Ades, and Nicky J Welton. Equivalence of entropy balancing and the method of moments for matching-adjusted indirect comparison. *Research Synthesis Methods*, 2020.

[72] Jihane Aouni, Nadia Gaudel-Dedieu, and Bernard Sebastien. Matching-adjusted indirect comparisons: Application to time-to-event data. *Statistics in Medicine*, 2020.

[73] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[74] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[75] Vanja Sikirica, Robert L Findling, James Signorovitch, M Haim Erder, Ryan Dammerman, Paul Hodgkins, Mei Lu, Jipan Xie, and Eric Q Wu. Comparative efficacy of guanfacine extended release versus atomoxetine for the treatment of attention-deficit/hyperactivity disorder in children and adolescents: applying matching-adjusted indirect comparison methodology. *CNS drugs*, 27(11):943–953, 2013.

[76] Ewout W Steyerberg et al. *Clinical prediction models*. Springer, 2019.

[77] Frank E Harrell and James C Slaughter. *Biostatistics for biomedical research*. 2016.

[78] Stephen Senn, Erika Graf, and Angelika Caputo. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in medicine*, 26(30):5529–5544, 2007.

[79] Sabine E Grimm, Nigel Armstrong, Bram LT Ramaekers, Xavier Pouwels, Shona Lang, Svenja Petersohn, Rob Riemsma, Gillian Worthy, Lisa Stirk, Janine Ross, et al. Nivolumab for treating metastatic or unresectable urothelial cancer: an evidence review group perspective of a nice single technology appraisal. *Pharmacoeconomics*, 37(5):655–667, 2019.

[80] Shijie Ren, Hazel Squires, Emma Hock, Eva Kaltenthaler, Andrew Rawdin, and Constantine Alifrangis. Pembrolizumab for locally advanced or metastatic urothelial cancer where cisplatin is unsuitable: An evidence review group perspective of a nice single technology appraisal. *PharmacoEconomics*, 37(9):1073–1080, 2019.

[81] KJ Ishak, M Rael, H Phatak, C Masseria, and T Lanitis. Simulated treatment comparison of time-to-event (and other non-linear) outcomes. *Value in Health*, 18(7):A719, 2015.

[82] Marshall M Joffe, Thomas R Ten Have, Harold I Feldman, and Stephen E Kimmel. Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4):272–279, 2004.

[83] PR Rosenbaum, T Colton, and P Armitage. Encyclopedia of biostatistics. 1998.

[84] Peter C Austin. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, 32(16):2837–2849, 2013.

[85] Peter C Austin. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in medicine*, 29(20):2137–2148, 2010.

[86] Peter C Austin, Andrea Manca, Merrick Zwarenstein, David N Juurlink, and Matthew B Stanbrook. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of clinical epidemiology*, 63(2):142–153, 2010.

[87] John M Neuhaus, John D Kalbfleisch, and Walter W Hauck. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review/Revue Internationale de Statistique*, pages 25–35, 1991.

[88] Sander Greenland. Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology*, 125(5):761–768, 1987.

[89] Alan E Hubbard, Jennifer Ahern, Nancy L Fleischer, Mark Van der Laan, Sheri A Satariano, Nicholas Jewell, Tim Bruckner, and William A Satariano. To gee or not to gee: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, pages 467–474, 2010.

[90] Miguel Angel Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271, 2004.

[91] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. Methods for population adjustment with limited access to individual patient data: A review and simulation study. *Research synthesis methods*, 12(6):750–775, 2021.

[92] Peter C Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in medicine*, 33(7):1242–1258, 2014.

[93] Sander Greenland, James M Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statistical science*, pages 29–46, 1999.

[94] Holly Janes, Francesca Dominici, and Scott Zeger. On quantifying the magnitude of confounding. *Biostatistics*, 11(3):572–582, 2010.

[95] Torben Martinussen and Stijn Vansteelandt. On collapsibility and confounding bias in cox and aalen regression models. *Lifetime data analysis*, 19(3):279–296, 2013.

[96] Mitchell H Gail, S Wieand, and Steven Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3):431–444, 1984.

[97] Olli S Miettinen and E Francis Cook. Confounding: essence and detection. *American journal of epidemiology*, 114(4):593–603, 1981.

[98] Sander Greenland and Hal Morgenstern. Confounding in health research. *Annual review of public health*, 22(1):189–212, 2001.

[99] Sander Greenland and Judea Pearl. Adjustments and their consequences—collapsibility analysis using graphical models. *International Statistical Review*, 79(3):401–426, 2011.

[100] Jay S Kaufman. Marginalia: comparing adjusted effect measures. *Epidemiology*, 21(4):490–493, 2010.

[101] Dorothea Weber, Katrin Jensen, and Meinhard Kieser. Comparison of methods for estimating therapy effects by indirect comparisons: A simulation study. *Medical Decision Making*, 40(5):644–654, 2020.

[102] Tim P Morris, Ian R White, and Michael J Crowther. Using simulation studies to evaluate statistical methods. *Statistics in medicine*, 38(11):2074–2102, 2019.

[103] R Core Team et al. R: A language and environment for statistical computing. 2013.

[104] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

[105] Nicholas R Latimer. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Medical Decision Making*, 33(6):743–754, 2013.

[106] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[107] Michal Abrahamowicz, Roxane du Berger, Daniel Krewski, Richard Burnett, Gillian Bartlett, Robyn M Tamblyn, and Karen Leffondré. Bias due to aggregation of individual covariates in the cox regression model. *American journal of epidemiology*, 160(7):696–706, 2004.

[108] Richard P. Brent. An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14(4):422–425, 1971.

[109] Kenneth Stanley. Design of randomized controlled trials. *Circulation*, 115(9):1164–1169, 2007.

[110] Peter C Austin. Variance estimation when using inverse probability of treatment weighting (iptw) with survival analysis. *Statistics in medicine*, 35(30):5642–5655, 2016.

[111] Catherine R Lesko and Bryan Lau. Bias due to confounders for the exposure-competing risk relationship. *Epidemiology (Cambridge, Mass.)*, 28(1):20, 2017.

[112] Terry M Therneau and Patricia M Grambsch. The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer, 2000.

[113] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[114] Ingeborg Waernbaum. Propensity score model specification for estimation of average treatment effects. *Journal of Statistical Planning and Inference*, 140(7):1948–1956, 2010.

[115] Zhong Zhao. Sensitivity of propensity score methods to the specifications. *Economics Letters*, 98(3):309–319, 2008.

[116] Donald B Rubin and Neal Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585, 2000.

[117] Clémence Leyrat, Agnès Caille, Allan Donner, and Bruno Giraudeau. Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Statistics in medicine*, 33(20):3556–3575, 2014.

[118] Gerta Rücker and Guido Schwarzer. Presenting simulation results in a nested loop plot. *BMC medical research methodology*, 14(1):129, 2014.

[119] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[120] Andrea Burton, Douglas G Altman, Patrick Royston, and Roger L Holder. The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292, 2006.

[121] Göran Kauermann and Raymond J Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396, 2001.

[122] Michael P Fay and Barry I Graubard. Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*, 57(4):1198–1206, 2001.

[123] Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

[124] Achim Zeileis. Econometric computing with hc and hac covariance matrix estimators. 2004.

[125] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[126] Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Katie R Mollan, Paul E Sax, Eric S Daar, Adaora A Adimora, Joseph J Eron, and Michael J Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A,(Statistics in Society)*, 181(4):1193, 2018.

[127] Fan Li et al. Propensity score weighting for causal inference with multiple treatments. *Annals of Applied Statistics*, 13(4):2389–2415, 2019.

[128] Huzhang Mao, Liang Li, and Tom Greene. Propensity score weighting analysis and treatment effect discovery. *Statistical methods in medical research*, 28(8):2439–2454, 2019.

[129] Federico Ricciardi, Silvia Liverani, and Gianluca Baio. Dirichlet process mixture models for regression discontinuity designs. *arXiv preprint arXiv:2003.11862*, 2020.

[130] DJ Fisher, AJ Copas, JF Tierney, and MKB Parmar. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. *Journal of clinical epidemiology*, 64(9):949–967, 2011.

[131] David J Fisher, James R Carpenter, Tim P Morris, Suzanne C Freeman, and Jayne F Tierney. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *bmj*, 356:j573, 2017.

[132] Jayne F Tierney, Claire Vale, Richard Riley, Catrin Tudur Smith, Lesley Stewart, Mike Clarke, and Maroeska Rovers. Individual participant data (ipd) meta-analyses of randomised controlled trials: guidance on their use. *PLoS Med*, 12(7):e1001855, 2015.

[133] Sofia Dias, Anthony E Ades, Nicky J Welton, Jeroen P Jansen, and Alexander J Sutton. *Network meta-analysis for decision-making*. John Wiley & Sons, 2018.

[134] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2011.

[135] Sofia Dias, Alex J Sutton, Nicky J Welton, and AE Ades. Nice dsu technical support document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011.

[136] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. Effect modification in anchored indirect treatment comparisons: Comments on "matching-adjusted indirect comparisons: Application to time-to-event data". *Statistics in Medicine*, 2021.

[137] Isabella Annesi, Thierry Moreau, and Joseph Lellouch. Efficiency of the logistic regression and cox proportional hazards models in longitudinal studies. *Statistics in medicine*, 8(12):1515–1521, 1989.

[138] Eric Vittinghoff and Charles E McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology*, 165(6):710–718, 2007.

[139] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, pages 579–595, 2014.

[140] Mark J Van der Laan, MJ Laan, and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

[141] Ashkan Ertefaie and David A Stephens. Comparing approaches to causal inference for longitudinal data: inverse probability weighting versus propensity scores. *The international journal of biostatistics*, 6(2), 2010.

[142] Rhian M Daniel, SN Cousens, BL De Stavola, Michael G Kenward, and JAC Sterne. Methods for dealing with time-dependent confounding. *Statistics in medicine*, 32(9):1584–1618, 2013.

[143] Premnath Shenoy and Anand Harugeri. Elderly patients' participation in clinical trials. *Perspectives in clinical research*, 6(4):184, 2015.

[144] Safi U Khan, Muhammad Zia Khan, Charumathi Raghu Subramanian, Haris Riaz, Muhammad U Khan, Ahmad Naeem Lone, Muhammad Shahzeb Khan, Eve-Marie

Benson, Mohamad Alkhouli, Michael J Blaha, et al. Participation of women and older participants in randomized clinical trials of lipid-lowering therapies: a systematic review. *JAMA network open*, 3(5):e205202–e205202, 2020.

[145] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[146] Stijn Vansteelandt and Niels Keiding. Invited commentary: G-computation–lost in translation? *American journal of epidemiology*, 173(7):739–742, 2011.

[147] Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

[148] Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

[149] Zhiqiang Tan. Comment: Understanding or, ps and dr. *Statistical Science*, 22(4):560–568, 2007.

[150] M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[151] Patrick Royston. Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241, 2004.

[152] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[153] An-Wen Chan, Fujian Song, Andrew Vickers, Tom Jefferson, Kay Dickersin, Peter C Gøtzsche, Harlan M Krumholz, Davina Ghersi, and H Bart Van Der Worp. Increasing value and reducing waste: addressing inaccessible research. *The Lancet*, 383(9913):257–266, 2014.

[154] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[155] Beata Nowok, Gillian M Raab, Chris Dibben, et al. synthpop: Bespoke creation of synthetic data in r. *J Stat Softw*, 74(11):1–26, 2016.

[156] Federico Bonofiglio, Martin Schumacher, and Harald Binder. Recovery of original individual person data (ipd) inferences from empirical ipd summaries only: Applications to distributed computing under disclosure constraints. *Statistics in Medicine*.

[157] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[158] Alexander P Keil, Eric J Daza, Stephanie M Engel, Jessie P Buckley, and Jessie K Edwards. A bayesian approach to the g-formula. *Statistical methods in medical research*, 27(10):3183–3204, 2018.

[159] Jonathan M Snowden, Sherri Rose, and Kathleen M Mortimer. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*, 173(7):731–738, 2011.

[160] Aolin Wang, Roch A Nianogo, and Onyebuchi A Arah. G-computation of average treatment effects on the treated and the untreated. *BMC medical research methodology*, 17(1):1–5, 2017.

[161] David L Sackett, Jonathan J Deeks, and Doughs G Altman. Down with odds ratios! *BMJ Evidence-Based Medicine*, 1(6):164, 1996.

[162] Robert G Newcombe. A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine*, 25(24):4235–4240, 2006.

[163] Edna Schechtman. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—which of these should we use? *Value in health*, 5(5):431–436, 2002.

[164] David M Phillippo, Sofia Dias, Anthony E Ades, and Nicky J Welton. Target estimands for efficient decision making: Response to comments on “assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study”. *Statistics in Medicine*, 40(11):2759–2763, 2021.

[165] Ori M Stitelman, C William Wester, Victor De Gruttola, and Mark J van der Laan. Targeted maximum likelihood estimation of effect modification parameters in survival analysis. *The international journal of biostatistics*, 7(1), 2011.

[166] Arvid Sjölander. Regression standardization with the r package stdreg. *European journal of epidemiology*, 31(6):563–574, 2016.

[167] Arvid Sjölander. Estimation of causal effect measures with the r-package stdreg. *European journal of epidemiology*, 33(9):847–858, 2018.

[168] Paul Lambert. Stpm2_standsurv: Stata module to obtain standardized survival curves after fitting an stpm2 survival model. 2018.

[169] Ravi Varadhan, Nicholas C Henderson, and Carlos O Weiss. Cross-design synthesis for extending the applicability of trial evidence when treatment effect is heterogeneous: Part i. methodology. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 2(3-4):112–126, 2016.

[170] Zhiwei Zhang, Lei Nie, Guoxing Soon, and Zonghui Hu. New methods for treatment effect calibration, with applications to non-inferiority trials. *Biometrics*, 72(1):20–29, 2016.

[171] Andrea Gabrio, Alexina J Mason, and Gianluca Baio. A full bayesian model to handle structural ones and missingness in economic evaluations from individual-level data. *Statistics in medicine*, 38(8):1399–1420, 2019.

[172] Jonathan W Bartlett. Covariate adjustment and estimation of mean response in randomised trials. *Pharmaceutical statistics*, 17(5):648–666, 2018.

[173] Yongming Qu and Junxiang Luo. Estimation of group means when adjusting for covariates in generalized linear models. *Pharmaceutical statistics*, 14(1):56–62, 2015.

[174] Peter W Lane and John A Nelder. Analysis of covariance and standardization as instances of prediction. *Biometrics*, pages 613–621, 1982.

[175] Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, 2020.

[176] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[177] Donald B Rubin and Nathaniel Schenker. Logit-based interval estimation for binomial data using the jeffreys prior. *Sociological methodology*, pages 131–144, 1987.

[178] Odd O Aalen, Vernon T Farewell, Daniela De Angelis, Nicholas E Day, and O Nöel Gill. A markov model for hiv disease progression including the effect of hiv diagnosis and treatment: application to aids prediction in england and wales. *Statistics in medicine*, 16(19):2191–2210, 1997.

[179] Alexander P Keil, Julie L Daniels, and Irva Hertz-Picciotto. Autism spectrum disorder, flea and tick medication, and adjustments for exposure misclassification: the charge (childhood autism risks from genetics and environment) case–control study. *Environmental Health*, 13(1):1–10, 2014.

[180] Maria Josefsson and Michael J Daniels. Bayesian semi-parametric g-computation for causal inference in a cohort study with mnar dropout and death. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(2):398–414, 2021.

[181] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[182] Olli Saarela, Elja Arjas, David A Stephens, and Erica EM Moodie. Predictive bayesian inference and dynamic treatment regimes. *Biometrical Journal*, 57(6):941–958, 2015.

[183] David Lunn, Chris Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. *The BUGS book: A practical introduction to Bayesian analysis*. CRC press, 2012.

[184] Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

[185] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[186] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

[187] Gianluca Baio. survhe: Survival analysis for health economic evaluation and cost-effectiveness modeling. *Journal of Statistical Software*, 95(1):1–47, 2020.

[188] Ilse van Oostrum, Mario Ouwens, Antonio Remiro-Azócar, Gianluca Baio, Maarten J Postma, Erik Buskens, and Bart Heeg. Comparison of parametric survival extrapolation approaches incorporating general population mortality for adequate health technology assessment of new oncology drugs. *Value in Health*, 2021.

[189] P Mohr, J Larkin, VF Paly, A Remiro Azocar, G Baio, M Kurt, A Amadi, JI Rizzo, HM Johnson, A Moshyk, et al. 1105p estimating long-term survivorship in patients with advanced melanoma treated with immune-checkpoint inhibitors: Analyses from the phase iii checkmate 067 trial. *Annals of Oncology*, 31:S747, 2020.

[190] V Paly, P Mohr, J Larkin, M Middleton, JH Youn, A Remiro-Azocar, G Baio, A Moshyk, S Kotapati, M Hamilton, et al. Pcn193 assessing the impact of modeling non-disease-related mortality on long-term survivorship rates in previously untreated advanced melanoma: A case study from checkmate 067. *Value in Health*, 24:S55, 2021.

[191] Patricia Guyot, AE Ades, Mario JNM Ouwens, and Nicky J Welton. Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, 12(1):9, 2012.

[192] Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558, 1994.

[193] Selene Leon, Anastasios A Tsiatis, and Marie Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.

[194] Joseph L Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

[195] Tosiya Sato and Yutaka Matsuyama. Marginal structural models as a tool for standardization. *Epidemiology*, pages 680–686, 2003.

[196] Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1, 2003.

[197] Donald B Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

[198] Jerome P Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531, 2002.

[199] Jerome P Reiter. Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):185–205, 2005.

[200] Yajuan Si and Jerome P Reiter. A comparison of posterior simulation and inference by combining rules for multiple imputation. *Journal of Statistical Theory and Practice*, 5(2):335–347, 2011.

[201] Jerome P Reiter and Trivellore E Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471, 2007.

[202] Gillian M Raab, Beata Nowok, and Chris Dibben. Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7(3):67–97, 2016.

[203] S Bujkiewicz, F Achana, T Papanikos, R Riley, and K Abrams. Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. *NICE DSU technical support document*, 20, 2019.

[204] Jerome P Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.

[205] Trivellore E Raghunathan. Synthetic data. *Annual Review of Statistics and Its Application*, 8, 2020.

[206] Brian D Ripley. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

[207] Guido Skipka, Beate Wieseler, Thomas Kaiser, Stefanie Thomas, Ralf Bender, Jürgen Windeler, and Stefan Lange. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biometrical Journal*, 58(1):43–58, 2016.

[208] Bernhard K Flury and Hans Riedwyl. Standard distance in univariate and multivariate analysis. *The American Statistician*, 40(3):249–251, 1986.

[209] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[210] Peter J Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press, 1967.

[211] Jennifer Hill and Jerome P Reiter. Interval estimation for treatment effects using propensity score matching. *Statistics in medicine*, 25(13):2230–2256, 2006.

[212] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via stan. *R package version*, 2(4):1758, 2018.

[213] Stan Developent Team et al. Rstan: the r interface to stan. *R package version*, 2(1):522, 2016.

[214] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*, 2(4):1360–1383, 2008.

[215] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[216] J Martin Bland and Douglas G Altman. The odds ratio. *Bmj*, 320(7247):1468, 2000.

[217] Sander Greenland, Mohammad Ali Mansournia, and Douglas G Altman. Sparse data bias: a problem hiding in plain sight. *bmj*, 352, 2016.

[218] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

[219] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

[220] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

[221] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[222] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[223] Elizabeth J Williamson, Andrew Forbes, and Ian R White. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in medicine*, 33(5):721–737, 2014.

[224] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[225] Fan Li, Ashley L Buchanan, and Stephen R Cole. Generalizing trial evidence to target populations in non-nested designs: Applications to aids clinical trials. *arXiv preprint arXiv:2103.04907*, 2021.

[226] David Madigan and Adrian E Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

[227] Dennis O Dixon and Richard Simon. Bayesian subset analysis. *Biometrics*, pages 871–881, 1991.

[228] David J Spiegelhalter, Laurence S Freedman, and Mahesh KB Parmar. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157(3):357–387, 1994.

[229] Richard Simon and Laurence S Freedman. Bayesian design and analysis of two x two factorial clinical trials. *Biometrics*, pages 456–464, 1997.

[230] Daniel Westreich, Jessie K Edwards, Catherine R Lesko, Stephen R Cole, and Elizabeth A Stuart. Target validity and the hierarchy of study designs. *American journal of epidemiology*, 188(2):438–443, 2019.

[231] Kosuke Imai, Gary King, and Elizabeth A Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2):481–502, 2008.

[232] Suzie Cro, Tim P Morris, Michael G Kenward, and James R Carpenter. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Statistics in Medicine*.

[233] Trang Quynh Nguyen, Cyrus Ebnesajjad, Stephen R Cole, and Elizabeth A Stuart. Sensitivity analysis for an unobserved moderator in rct-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, pages 225–247, 2017.

[234] Issa J Dahabreh and Miguel A Hernán. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, 34(8):719–722, 2019.

[235] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, 2018.

[236] Alexander Breskin, Daniel Westreich, Stephen R Cole, and Jessie K Edwards. Using bounds to compare the strength of exchangeability assumptions for internal and external validity. *American journal of epidemiology*, 188(7):1355–1360, 2019.

[237] SJ Senn. Covariate imbalance and random allocation in clinical trials. *Statistics in medicine*, 8(4):467–475, 1989.

[238] Adrián V Hernández, Ewout W Steyerberg, and J Dik F Habbema. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of clinical epidemiology*, 57(5):454–460, 2004.

[239] Brennan C Kahan, Vipul Jairath, Caroline J Doré, and Tim P Morris. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*, 15(1):1–7, 2014.

[240] Elizabeth Colantuoni and Michael Rosenblum. Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in medicine*, 34(18):2602–2617, 2015.

[241] Iván Díaz, Elizabeth Colantuoni, and Michael Rosenblum. Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics*, 72(2):422–431, 2016.

[242] Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Xiaomin Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677, 2008.

[243] Kelly L Moore, Romain Neugebauer, Thamban Valappil, and Mark J van der Laan. Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Statistics in medicine*, 30(19):2389–2408, 2011.

[244] Fei Jiang, Lu Tian, Haoda Fu, Takahiro Hasegawa, and LJ Wei. Robust alternatives to ancova for estimating the treatment effect via a randomized comparative study. *Journal of the American Statistical Association*, 114(528):1854–1864, 2019.

[245] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. Parametric g-computation for compatible indirect treatment comparisons with limited individual patient data. *arXiv preprint arXiv:2108.12208*, 2021.

[246] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio. Marginalization of regression-adjusted treatment effects in indirect comparisons with limited patient-level data. *arXiv preprint arXiv:2008.05951*, 2020.

[247] FDA. Adjusting for covariates in randomized clinical trials for drugs and biological products. draft guidance for industry, 2021. Accessed: 25th November 2021.

[248] FDA. Covid-19: Developing drugs and biological products for treatment or prevention. guidance for industry, 2021. Accessed: 25th November 2021.

[249] EMA. Guideline on adjustment for baseline covariates in clinical trials, 2015. Accessed: 25th November 2021.

[250] EMA. Ich e9 (r1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, 2020. Accessed: 25th November 2021.

[251] Tymon Słoczyński. Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *The Review of Economics and Statistics*, pages 1–27, 2020.

[252] Alice S Whittemore. Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):328–340, 1978.

[253] Anders Huitfeldt, Mats J Stensrud, and Etsuji Suzuki. On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging themes in epidemiology*, 16(1):1–5, 2019.

[254] Menglan Pang, Jay S Kaufman, and Robert W Platt. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical methods in medical research*, 25(5):1925–1937, 2016.

[255] Walter W Hauck, John M Neuhaus, John D Kalbfleisch, and Sharon Anderson. A consequence of omitted covariates when estimating odds ratios. *Journal of Clinical Epidemiology*, 44(1):77–81, 1991.

[256] Edwin P Martens, Wiebe R Pestman, and Olaf H Klungel. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a monte carlo study (pn/a) by austin pc et al. *Statistics in medicine*, 26:3208–3210, 2007.

[257] Susan F Assmann, Stuart J Pocock, Laura E Enos, and Linda E Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

[258] Laurence D Robinson and Nicholas P Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 227–240, 1991.

[259] Ian Ford, John Norrie, and Susan Ahmadi. Model inconsistency, illustrated by the cox proportional hazards model. *Statistics in medicine*, 14(8):735–746, 1995.

[260] Theodore Karrison and Masha Kocherginsky. Restricted mean survival time: Does covariate adjustment improve precision in randomized clinical trials? *Clinical Trials*, 15(2):178–188, 2018.

[261] Rachael K Ross, Stephen R Cole, and David B Richardson. Decreased susceptibility of marginal odds ratios to finite-sample bias. *Epidemiology*, 32(5):648–652, 2021.

[262] MH Gail, Wai-Yuan Tan, and Steven Piantadosi. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75(1):57–64, 1988.

[263] Edward J Mills, Isabella Ghement, Christopher O'Regan, and Kristian Thorlund. Estimating the power of indirect comparisons: a simulation study. *PloS one*, 6(1):e16237, 2011.

[264] Jörg Ruof, Friedrich Wilhelm Schwartz, J-Matthias Schulenburg, and Charalabos-Markos Dintsios. Early benefit assessment (eba) in germany: analysing decisions 18 months after introducing the new amnog legislation. *The European Journal of Health Economics*, 15(6):577–589, 2014.

[265] Simon G Thompson and Julie A Barber. How should cost data in pragmatic randomised trials be analysed? *Bmj*, 320(7243):1197–1200, 2000.

[266] Andrew R Willan. On the probability of cost-effectiveness using data from randomized clinical trials. *BMC medical research methodology*, 1(1):8, 2001.

[267] Anthony O'Hagan and John W Stevens. The probability of cost-effectiveness. *BMC Medical Research Methodology*, 2(1):5, 2002.

[268] Christopher H Jackson, Simon G Thompson, and Linda D Sharples. Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(2):383–404, 2009.

[269] Andrew H Briggs. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics*, 17(5):479–500, 2000.

[270] Christopher H Jackson, Linda D Sharples, and Simon G Thompson. Structural and parameter uncertainty in bayesian cost-effectiveness models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):233–253, 2010.

[271] Susan C Griffin, Karl P Claxton, Stephen J Palmer, and Mark J Sculpher. Dangerous omissions: the consequences of ignoring decision uncertainty. *Health economics*, 20(2):212–224, 2011.

[272] Issa J Dahabreh, Rodney Hayward, and David M Kent. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International journal of epidemiology*, 45(6):2184–2193, 2016.

[273] Noel S Weiss. Generalizing from the results of randomized studies of treatment: Can non-randomized studies be of help? *European journal of epidemiology*, 34(8):715–718, 2019.

[274] Elizabeth Tipton. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266, 2013.

[275] Tjeerd-Pieter van Staa, Lisa Dyson, Gerard McCann, Shivani Padmanabhan, Rabah Belatri, Ben Goldacre, Jackie Cassell, Munir Pirmohamed, David Torgerson, Sarah Ronaldson, et al. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health technology assessment*, 18(43):1–146, 2014.

[276] Niteesh K Choudhry. Randomized, controlled trials in health insurance systems. *New England Journal of Medicine*, 377(10):957–964, 2017.

[277] Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

[278] Peter M Rothwell. Commentary: External validity of results of randomized trials: disentangling a complex concept. *International journal of epidemiology*, 39(1):94–96, 2010.

[279] Joel B Greenhouse, Eloise E Kaizar, Kelly Kelleher, Howard Seltman, and William Gardner. Generalizing from clinical trial data: a case study. the risk of suicidality among pediatric antidepressant users. *Statistics in medicine*, 27(11):1801–1813, 2008.

[280] Michael Happich, Alan Brnabic, Douglas Faries, Keith Abrams, Katherine B Winfree, Allicia Girvan, Pall Jonsson, Joseph Johnston, Mark Belger, and IMI GetReal Work Package 1. Reweighting randomized controlled trial evidence to better reflect real life–a case study of the innovative medicines initiative. *Clinical Pharmacology & Therapeutics*, 108(4):817–825, 2020.

[281] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *arXiv preprint arXiv:2102.11904*, 2021.

[282] Jacqueline Corrigan-Curay, Leonard Sacks, and Janet Woodcock. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *Jama*, 320(9):867–868, 2018.

[283] Til Stürmer, Tiansheng Wang, Yvonne M Golightly, Alex Keil, Jennifer L Lund, and Michele Jonsson Funk. Methodological considerations when analysing and interpreting real-world data. *Rheumatology*, 59(1):14–25, 2020.

[284] Cynthia J Girman, Mary E Ritchey, Wei Zhou, and Nancy A Dreyer. Considerations in characterizing real-world data relevance and quality for regulatory purposes: a commentary. *Pharmacoepidemiology and drug safety*, 28(4):439, 2019.

[285] Scott D Ramsey, Blythe J Adamson, Xiaoliang Wang, Danielle Bargo, Shrujal S Baxi, Shuhag Ghosh, and Neal J Meropol. Using electronic health record data to identify comparator populations for comparative effectiveness research. *Journal of Medical Economics*, 23(12):1618–1622, 2020.

[286] Gillis Carrigan, Samuel Whipple, William B Capra, Michael D Taylor, Jeffrey S Brown, Michael Lu, Brandon Arnieri, Ryan Copping, and Kenneth J Rothman. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. *Clinical Pharmacology & Therapeutics*, 107(2):369–377, 2020.

[287] Ian Chau, Dung T Le, Patrick A Ott, Beata Korytowsky, Hannah Le, T Kim Le, Ying Zhang, Teresa Sanchez, Gregory A Maglinte, Melissa Laurie, et al. Developing real-world comparators for clinical trials in chemotherapy-refractory patients with gastric cancer or gastroesophageal junction cancer. *Gastric Cancer*, 23(1):133–141, 2020.

[288] Ashley Jaksa, James Wu, Páll Jónsson, Hans-Georg Eichler, Sarah Vititoe, and Nicolle M Gatto. Organized structure of real-world evidence best practices: moving from fragmen-

ted recommendations to comprehensive guidance. *Journal of Comparative Effectiveness Research*, 10(9):711–731, 2021.

[289] Anthony J Hatswell, Gianluca Baio, Jesse A Berlin, Alar Irs, and Nick Freemantle. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of ema and fda approvals 1999–2014. *BMJ open*, 6(6), 2016.

[290] Julia A Beaver, Lynn J Howie, Lorraine Pelosof, Tamy Kim, Jinzhong Liu, Kirsten B Goldberg, Rajeshwari Sridhara, Gideon M Blumenthal, Ann T Farrell, Patricia Keegan, et al. A 25-year experience of us food and drug administration accelerated approval of malignant hematology and oncology drugs and biologics: a review. *JAMA oncology*, 4(6):849–856, 2018.

[291] Kara E Rudolph and Mark J van der Laan. Robust estimation of encouragement-design intervention effects transported across sites. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(5):1509, 2017.

[292] Mark J van der Laan and Susan Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1), 2012.

[293] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

[294] Andrea Rotnitzky and Stijn Vansteelandt. Double-robust methods. In *Handbook of missing data methodology*, pages 185–212. CRC Press, 2014.

[295] Rhian M Daniel. Double robustness. *Wiley StatsRef: Statistics Reference Online*, pages 1–14, 2014.

[296] Ingeborg Waernbaum and Laura Pazzagli. Model misspecification and bias for inverse probability weighting and doubly robust estimators. *arXiv preprint arXiv:1711.09388*, 2017.

[297] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

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

[299] P Richard Hahn, Jared S Murray, Carlos M Carvalho, et al. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[300] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone, et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

[301] Sherri Rose. Mortality risk score prediction in an elderly population using machine learning. *American journal of epidemiology*, 177(5):443–452, 2013.

[302] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[303] Megan S Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185(1):65–73, 2017.

[304] Romain Pirracchio, Maya L Petersen, and Mark Van Der Laan. Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology*, 181(2):108–119, 2015.

[305] Aad van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, pages 679–686, 2014.

[306] Ashley I Naimi and Edward H Kennedy. Nonparametric double robustness. *arXiv preprint arXiv:1711.07137*, 2017.

[307] Edward H Kennedy and Sivaraman Balakrishnan. Discussion of" data-driven confounder selection via markov and bayesian networks" by jenny h\" aggstr\" om. *arXiv preprint arXiv:1710.11566*, 2017.

[308] Peter J Bickel, Friedrich Götze, and Willem R van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. In *Selected works of Willem van Zwet*, pages 267–297. Springer, 2012.

[309] Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2020.

[310] Miguel Angel Luque-Fernandez, Michael Schomaker, Bernard Rachet, and Mireille E Schnitzer. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in medicine*, 37(16):2530–2546, 2018.

[311] Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

[312] Paul N Zivich and Alexander Breskin. Machine learning for causal inference: on the use of cross-fit estimators. *Epidemiology*, 32(3):393–401, 2021.

[313] David M Phillippo. *Calibration of treatment effects in network meta-analysis using individual patient data*. PhD thesis, University of Bristol, Bristol, UK, 2019.

[314] Jesse A Berlin, Jill Santanna, Christopher H Schmid, Lynda A Szczech, and Harold I Feldman. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Statistics in medicine*, 21(3):371–387, 2002.

[315] R Faria, M Hernandez Alava, A Manca, and A Wailoo. Nice dsu technical support document 17: the use of observational data to inform estimates of treatment effectiveness for technology appraisal: methods for comparative individual patient data. *Sheffield: NICE Decision Support Unit*, 2015.

[316] James M Robins and Miguel A Hernán. Estimation of the causal effects of time-varying exposures. *Longitudinal data analysis*, 553:599, 2009.

[317] Jerzy S Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9.(translated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

[318] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

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

[320] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[321] Tyler J VanderWeele and Miguel A Hernan. Causal inference under multiple versions of treatment. *Journal of causal inference*, 1(1):1, 2013.

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

[323] Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

[324] Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

[325] Sander Greenland and James M Robins. Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, 6(1):4, 2009.

[326] Sander Greenland. Randomization, statistics, and causal inference. *Epidemiology*, pages 421–429, 1990.

[327] Stephen Senn. Testing for baseline balance in clinical trials. *Statistics in medicine*, 13(17):1715–1726, 1994.

[328] Jonathan J Deeks, Jac Dinnes, Roberto D'Amico, Amanda J Sowden, Charlotte Sakarovitch, Fujian Song, Mark Petticrew, DG Altman, et al. Evaluating non-randomised intervention studies. *Health technology assessment (Winchester, England)*, 7(27):iii–173, 2003.

[329] Issa J Dahabreh, Thomas A Trikalinos, David M Kent, and Christopher H Schmid. Heterogeneity of treatment effects. *Methods in comparative effectiveness research*, page 227, 2017.

[330] Jon Arni Steingrimsson and Jiabei Yang. Subgroup identification using covariate-adjusted interaction trees. *Statistics in medicine*, 38(21):3974–3984, 2019.

[331] Jiabei Yang, Issa J Dahabreh, and Jon A Steingrimsson. Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics*, 2021.

[332] Jerome P Reiter. Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131(2):365–377, 2005.