Aus dem Institut für Medizinische Biometrie und Informatik
Universitätsklinik Heidelberg
Abteilung Medizinische Biometrie
Geschäftsführender Direktor: Prof. Dr. sc. hum. Meinhard Kieser

Applying Matching Procedures in the Generation and Synthesis of Evidence

Inauguraldissertation
zur Erlangung des
„Doctor scientiarum humanarum“

an der Medizinischen Fakultät Heidelberg
der Ruprecht-Karls-Universität

vorgelegt von
Dorothea Weber
aus
Malsch

2020

Dekan:	Prof. Dr. med. Hans Georg Kräusslich
Doktorvater:	Prof. Dr. sc. hum. Meinhard Kieser

Contents

Abbreviations and Symbols

AB	trial comparing treatment A and treatment B - here a direct comparison
AC	trial comparing treatment A and treatment C - here an indirect comparison
AgD	aggregated data
AIS	acute ischemic stroke
AML	acute myeloid leukemia
ASPECTS	Alberta Stroke Program Early CT score
BFGS	Broyden-Fletcher-Goldfarb-Shanno
CB	trial comparing treatment C and treatment B - here a direct comparison
CI	confidence interval
Cov	coverage of the confidence interval
CS	conscious sedation
CT	computed tomography
EM	effect modifier

EP	endpoint
ESS	effective sample size
H_0	null hypothesis
H_1	alternative hypothesis
HTA	health technology assessment
HR	hazard ratio
IPD	individual patient data
IQWiG	Institute for Quality and Efficiency in Healthcare
KEEP SIMPLEST	KEep Evaluating Protocol Simplification In Managing Periinterventional Light Sedation for Endovascular Stroke Treatment
LL	lower left corner
LR	lower right corner
MAIC	Matching Adjusted Indirect Comparison
mRS	modified Rankin Scale
NICE	National Institute for Health and Care Excellence
NIHSS	National Institutes of Health Stroke Scale
OR	odds ratio
RCT	randomized controlled trial
RMSE	root mean squared error
sd	standard deviation
SIESTA	Sedation vs. Intubation for Endovascular Stroke Treatment

SOP	system operating procedure
STC	simulated treatment comparison
TTE	time-to-event
UL	upper left corner
UR	upper right corner
Var	variance
α	significance level
b	number of resampling steps
$b*$	binary variable
β_*	regression coefficient for variable $*$, in case of β_0 it describes the model intercept
β	vector containing the regression coefficients
$\hat{\beta}$	vector containing the estimates of the regression coefficients
$Bin(\cdot)$	probability mass function of the binomial distribution
c	continuous variable
δ	effect estimate comparing two treatments
$e(\cdot)$	propensity score function
$g(\cdot)$	link function
λ	baseline hazard rate of the Weibull distribution
l	lower limit of the confidence interval
M	number of matching partners

M_{max}	maximal number of matching partners
mr	matching rate
$\overline{mr}$	mean matching rate
$max.time$	maximal follow-up time for time to event endpoints
μ	mean of a probability distribution (e.g. normal distribution)
n_*	sample size, the index indicates the trial or group
N_*	number of trials, the index indicates the considered treatment comparison
$\mathcal{N}$	probability mass function of the normal distribution
ν	shape parameter of the Weibull distribution
Φ	probability distribution function of the standard normal distribution
π	(binary) response/event rate
$P(\cdot)$	probability function
s	number of patients recruited for the interim analysis, referred to as time point of interim analysis
σ	standard deviation of a probability distribution (e.g. normal distribution)
T	trial, here $T = 0$ corresponds to trial AB and $T = 1$ corresponds to trial CB
t	treatment arm
tr	treatment variable
τ	tolerance in iterative matching procedure (maximal difference to 1:1 matching rate)

$\boldsymbol{\omega}$	vector of weights (ω_i the weight for patient i)
X	data matrix containing the patient characteristics
$\boldsymbol{x}_i$	vector containing the patient characteristics of patient i
$\mathbf{x}_{ij}$	value of patient characteristic j of patient i
$\overline{x}$	vector containing the mean patient characteristics of a trial
Y	vector of outcome variable
y_i	outcome for patient i
Z	group variable
z_i	group assignment for patient i

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

In clinical research, blinded randomized controlled trials (RCTs) are the gold-standard for evaluating the efficacy of a medical intervention. The random allocation of patients to treatment and control group ensures the comparability of patient cohorts. However, there are situations that do not achieve comparability despite a randomized trial design was chosen, e.g., in oncological trials where group sizes are small or, patient cohorts are heterogeneous (Harrison, 2016; Gan et al., 2010). An alternative design to an RCT is a larger single-arm trial comparing the treatment effect to a predefined value, which often fails since they ignore variety in patient characteristics influencing the treatment effect. Beside unsuccessful randomization, it is not always feasible to randomize a clinical trial due to ethical concerns or practical reasons (Frakt, 2015; Faraoni and Schaefer, 2016; Goodman et al., 2017), but circumstances may allow for an observational trial. If furthermore, it is not practicable to observe the treatment and the control group at the same time, an observational single-arm study might remain the only option. Single-arm studies have the disadvantage that a direct comparison to placebo or the standard therapy is not possible. When historical individual patient data (IPD) for the control group recruited within an earlier study or a registry is available, an alternative strategy would be to use this external control group for comparison. A naïve approach, which does not adjust for confounders, to compare patient groups of dif-

ferent trials, may lead to severe bias due to potential differences in patient characteristics. The lack of comparability can be addressed by matching procedures that aim to balance the patient groups concerning chosen matching variables. So far, matching is applied after all patients are recruited, and the study database is closed. While performing the matching procedure, one may observe that some patients cannot find an appropriate matching partner and thus are dropped from subsequent analysis. This failure to identify suitable matching partners may cause the power to decrease. In practice, it cannot be expected to match all patients in the control group to an intervention patient when recruiting just the same number of patients as in the control group. Thus, published prospective matched case-control trials, defined as a prospective single-arm study compared with an external control group under the usage of a matching approach, prespecified an additional percentage of intervention patients. Or, for example, the trial of Charpentier et al. (2001) applied an algorithm that directly tries to find a matching partner for an intervention patient, which results in recruiting 30% more treated patients than the number of control patients. In case a lower or higher number of patients can be matched to one of the controls than expected, the sample size would be too small or more patients than needed are recruited. The fraction of patients matched to one of the controls, the matching rate, is, therefore, an important statistical measure of such trials aimed to be as high as possible. Additionally, if the number of control patients allows to identify more than one suitable matching partner per treated patient, power can be increased. So far, the designs for prospective matched case control trials are limited which offers scope for development.

Until now, the methods for matching are used for the generation of evidence. The second part of this work covers a situation in evidence synthesis, where the comparability of patient groups might not be given. In medical practice, physicians frequently face circumstances where various therapy options exist. It would be desirable that all these therapies were previously compared at once in one or several clinical trials. However, multi-arm trials are seldom available, and two-arm trials were conducted comparing just a subset of all possible therapies, whereas some comparisons are covered by multiple studies. In situations where a so-called head-to-head comparison (a trial directly comparing two treatments) is missing, the question arises whether and how reliable and valid conclusions on the choice of the best treatment option can be drawn without initiating a new trial. In recent years, the so-called

indirect comparisons attracted considerable attention (Signorovitch et al., 2013; Nash et al., 2018; Veroniki et al., 2016). In particular, indirect comparisons are of increasing interest in the field of health technology assessments (HTAs) (IQWiG, 2017, 2019; Phillippo et al., 2018). For early benefit assessment in the framework of HTAs, the valid comparator treatment is predefined, and frequently, there is a lack of direct comparisons with this valid comparator (Kühnast et al., 2017). Imprudently combining the results from different trials to get an estimate for the unavailable comparison of interest can cause severe bias due to cross-trial differences, such as differences in effect modifier distributions or worse baseline disease status of patients in one of the trials, which may mean that the treatment is more or less effective (Signorovitch et al., 2010). Additionally, published results in the form of aggregated data (AgD) are usually employed, because access to IPD is seldom available for all relevant studies. In case IPD is available, its usage may increase the reliability of the results and may reduce the uncertainty in treatment effects compared to situations where only AgD is available. Indirect comparisons taking the potential imbalance between trials into account are called adjusted indirect comparisons. The method of Bucher (Bucher et al., 1997) and the Matching Adjusted Indirect Comparison (MAIC) (Signorovitch et al., 2010) address this setting of an anchored adjusted indirect comparison. There are published simulation studies in the context of indirect comparisons which show unsatisfactory performance in terms of power, meaning it is hard to demonstrate an existing treatment effect by an indirect comparison (Mills et al., 2011; Kühnast et al., 2017). Furthermore, the sample size needed for an indirect comparison is always higher than for the underlying direct comparison (Snapinn and Jiang, 2011). Despite there are simulation results on indirect comparisons available, their performance is not sufficiently studied in situations where effect modifications are present, assumptions of the methods for indirect comparisons are violated, or when cross-trial differences exist, such as differences in patient population or different confounder adjustment of regression models for evaluating the treatment effect. Moreover, the power provided by the sample size calculation for the head-to-head trials may have a substantial impact on the power of the indirect comparison which is not sufficiently quantified, yet. This is of particular interest for investigators when designing a head-to-head trial, which is already planned to be included in a later indirect comparison. To examine those situations, simulation studies

covering a variety of practically relevant scenarios are needed (Phillippo et al., 2018; Song et al., 2009; Glenny et al., 2005; Petto et al., 2019).

The major weakness of established methods for indirect comparisons is the limitation for considering only one study per treatment comparison. However, often more studies and, therefore, more evidence is available, which should all be considered. Mainly because not using all available evidence may introduce additional bias when selecting one of the various available trials for the indirect comparison. So far, Belger et al. (2015) presented some solutions for the use of multiple studies in a frequentist framework at a conference, and some are stated in the guideline of National Institute for Health and Care Excellence (NICE) (Phillippo et al., 2016); Leahy and Walsh (2019) evaluated the situation of multiple IPD trials in a Bayesian framework. But, no published paper that includes a systematic comparison of such methods in a frequentist setting is available. Especially for numerous IPD and AgD trials, there is no recommendation. Therefore, methods for indirect comparisons need further development to enhance power and reduce bias under the usage of all available evidence and they need to be compared in practically relevant scenarios.

1.2 Objectives and Structure of the Present Work

The overall objective of this thesis is to examine matching procedures for the generation and synthesis of evidence in clinically relevant situations and to further develop the existing methods. The part about generation of evidence pursue the aim to develop methods that take the study-specific matching rate already in the planning stage into account, e.g., in the form of a sample size recalculation for a prospective matched case-control trial. The synthesis of evidence aims in answering the question whether and in which settings indirect comparisons produce valid treatment effects under the usage of matching procedures. Throughout this thesis, a simulation study is designed with the purpose to compare the method of Bucher (Bucher et al., 1997) and the MAIC (Signorovitch et al., 2010) in a wide range of practically relevant scenarios where assumptions are violated, and cross-trial differences exist. As a second aim of this simulation study, the influence of the planned power for the corresponding head-to-head trials on the power of the indirect comparison is examined. Furthermore this

thesis has the objective of refining those methods to include multiple studies in indirect comparisons.

The presented thesis is structured as follows. The second chapter focuses on the methods, Section 2.1 explains the methods for the generation of evidence, including new approaches for considering the matching rate already in the planning stage. The second part (Section 2.2) covers the tools for the synthesis of evidence. The results in Chapter 3 are split according to the simulation studies. Each section includes the data simulation process, the evaluation measures, the simulation scenarios, and the simulation results. One of the new approaches for the generation of evidence is applied to a real data example which is included in the results chapter (Section 3.1.3). The discussion is again structured according to the two parts, generation and synthesis of evidence. It contains a discussion of the results, its contribution to research, as well as limitations and directions for further research. Appendix A includes additional tables, Appendix B supplementary methodological background, and Appendix C implementations of essential functions in R.

Chapter 2

Methods

The methods described in this chapter are divided into methods applied for the generation of evidence (Section 2.1) and tools needed for the synthesis of evidence (Section 2.2).

2.1 Generation of Evidence

When imbalances in important patient characteristics between study arms are observed matching procedures can address this issue, Optimal Matching or Propensity Score Matching are frequently applied methods (Gu and Rosenbaum, 1993; Rosenbaum and Rubin, 1985). Throughout this thesis, the evaluation of developed approaches which use a matching procedure for the generation of evidence is based on the propensity score method by Rosenbaum and Rubin (Rosenbaum and Rubin, 1983, 1984; Rubin and Thomas, 1996) which is explained before the details of the new approaches are given. Nevertheless, the new approaches can be combined with any matching algorithm and are not limited to the propensity score.

In nonrandomized studies, a direct comparison of treatment and control group may give misleading results because of systematic differences between groups. Matching approaches aim to find appropriate pairs of treated and control patients, which can be used for a more reasonable comparison. It is assumed that there are $n_{treated}$ patients in the intervention group and $n_{control}$ patients in the control group. In natural settings $n_{control} \geqslant n_{treated}$ is assumed. Propensity score matching purposes to minimize the influence of observed and considered

baseline characteristics on the treatment effect (Austin, 2011). The propensity score $e(X)$ is a function depending on the given (relevant) confounders X such that the conditional distribution of being assigned to the treated study arm ($Z = 1$) and the control arm ($Z = 0$) is the same (Austin, 2011). Assuming that there are $n_{control} + n_{treated}$ patients included, the propensity score is defined as the conditional probability of being assigned to the treatment group, given the vector of considered confounders $\boldsymbol{x}_i$

$$e(\boldsymbol{x}_i) = P(Z_i = 1 \mid \boldsymbol{x}_i) \qquad i = 1, \ldots, n_{control} + n_{treated},$$

where $Z_i \in \{0, 1\}$ represents the group assignment. Here independence is assumed which means

$$P(Z_1, \ldots, Z_{n_{control}+n_{treated}} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_{control}+n_{treated}}) = \prod_{i=1}^{n_{control}+n_{treated}} e(\boldsymbol{x}_i)^{z_i}(1 - e(\boldsymbol{x}_i))^{1-z_i}.$$

The propensity score function is unknown in the case of nonrandomized studies. To generate an estimate of the propensity score function, for example, a logistic regression model with treatment status as outcome variable and the relevant baseline characteristics as covariates can be considered (Rosenbaum and Rubin, 1984)

$$logit(z_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \tag{2.1}$$

This logistic model provides the propensity scores; in other words, the probability of being assigned to the treatment group. To form pairs, the treatment and control patients are matched according to the logit of the estimated propensity score

$$\ln \frac{e(\boldsymbol{x}_i)}{1 - e(\boldsymbol{x}_i)}$$

by using some caliper width of these estimates (Rosenbaum and Rubin, 1985). The caliper width is a predefined amount, which is the maximal difference in propensity scores between treated and control patients. The choice of the caliper width follows a bias-variance trade-off; small calipers reduce bias but also may reduce sample size, which increases variance of the estimated treatment effect. Austin recommends a caliper width of 0.2 of the standard deviation of the logit of the propensity score (Austin, 2011). According to Austin et al. (2007), variables that influence the outcome or both treatment assignment or selection and outcome should be considered in the propensity score estimation.

Beside finding matched pairs, the propensity score can be used for adjustment in the analysis of observational data to reduce potential bias in the estimated treatment effect (Rosenbaum and Rubin, 1985).

Assumed is a situation where an RCT is not feasible. Therefore, a prospective matched case-control trial design is intended. This means the control group already exist because it was part of an earlier randomized controlled trial, and the treatment group will be newly recruited. One essential part of the planning stage of a clinical trial is the sample size calculation, which is usually based on 80% to 90% power and 5% type I error rate. The sample size calculation has the aim of securing that a clinically relevant treatment effect can be detected. In a prospective matched case-control trial, the sample size calculation is not straightforward.

In the following, two methods for sample size recalculation in prospective matched case-control trials are introduced. It is assumed that a historical control group already exists. Section 2.1.1 covers the situation of a historical control, which was part of an earlier study, including a sample size which was based on a sample size calculation. Subsequently, Section 2.1.2 comprises a procedure for a large historical control group, for example a registry.

For the first setting, the sample size in the already recruited study arm is fixed and was usually based on a clinically relevant effect considered in the sample size calculation. Therefore, the objective is to find an appropriate matching partner for as many patients of the control group as possible. The matching rate is unknown and as in most practical situations likely to be less than 100%. This means that recruiting just the same number of patients as in the external control arm will not result in a situation where a matching partner is found for all patients in the control group. A possible step to address this uncertainty concerning the matching rate could be to include an interim analysis. This interim analysis is set after a predefined number of treated patients are recruited and used to estimate the actual matching rate. At the same time, information about the primary endpoint is not needed. The results of the interim analysis can then be used to recalculate the sample size for the treatment group, which makes it possible to find a matching partner to all control patients in the final analysis. Thus one receives an adaptive matched case-control design.

A first approach might be to use all available treated and control patients for estimating the matching rate at interim analysis. The sample size recalculation is then based on this estimate of the matching rate. In the following, this strategy will be referred to as the *naïve method*. In practice, a potential overestimation of the matching rate may occur when all patients are used at interim analysis. In consequence, a smaller number of patients than necessary is recruited after interim analysis and therefore, a lower matching rate is achieved at the final analysis. To avoid this overestimation, the naïve method is refined.

Comment: Parts of the following Chapter (Section 2.1.1) have already been published in Weber et al. (2019). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

2.1.1 Resampling CI Method

A resampling approach and two propensity score matching steps are the core parts of the proposed adaptive design for recalculating the sample size in a prospective matched case-control trial.

The adaptive design includes two matching steps. At the interim analysis, the matching rate is determined and is used for recalculation of the sample size needed to reach a high matching rate for the final analysis. The matching procedure at interim analysis is solely used for calculating the matching rate, and pairs are not fixed for the final analysis. For estimating the treatment effect in the final analysis, the second matching step produces the final 1:1 matches.

The control study arm includes $n_{control}$ patients. Initially, the number of treated patients is set to $n_{treated} = n_{control}$. A number of $n_{treated,interim}$ patients are recruited in the treatment group, which is a predefined proportion of $n_{treated}$. Conducting the matching step using all $n_{control}$ controls may lead to an overestimation in the matching rate. To avoid this potential overestimation, equally sized groups are used for the matching procedure at interim analysis. To achieve equal-sized groups, a sample of $n_{control,interim}$ is taken from the $n_{control}$ controls without replacement with $n_{control,interim} = n_{treated,interim}$. The sampled controls and all

$n_{treated,interim}$ treated patients are used to perform the matching step and to calculate the matching rate. To avoid bias due to the random sampling, the resampling and the matching step, including the calculation of the matching rate (mr) at the interim analysis, are repeated b times.

The mean resampling matching rate $\overline{mr}$ is calculated by

$$\overline{mr} = \frac{1}{b} \sum_{j=1}^{b} mr_j$$

The lower limit of the $100 \cdot (1 - \alpha_{CI})$ % confidence interval (CI) is given by

$$l_{mr} = \overline{mr} - \Phi(1 - \alpha_{CI}) \cdot \sqrt{\overline{mr} * (1 - \overline{mr})/n_{control,matched}} \tag{2.2}$$

with the maximal number of $n_{control,matched}$ matched pairs and $\Phi()$ the probability distribution function of the standard normal distribution. The lower limit of the confidence interval is then used for recalculating the sample size.

The total number of patients needed in the treated group is estimated by

$$n_{treated,final} = \frac{n_{control}}{l_{mr}}. \tag{2.3}$$

In practice, the maximal number of $n_{treated}$ may be limited due to practical or temporal reasons. This number $n_{treated,max}$ is fixed beforehand and leads to the following final number of patients in the treated group

$$n_{treated,final} = \min \left\{ \frac{n_{control}}{l_{mr}}, n_{treated,max} \right\}. \tag{2.4}$$

In the following, this approach is called *resampling CI method*. The pseudocode of the steps are given in Table 1.

Besides the lower limit of the $100 \cdot (1 - \alpha_{CI})$% CI, there are other values that could be used for the recalculation of sample size. One could use the mean resampling matching rate directly or a quantile of the distribution of the resampling matching rates. Using $\overline{mr}$ directly may overestimate the true matching rate. Quantiles are independent of the number of patients in the control group. But, in trials with a large control arm, a higher diversity of patients may be represented; therefore, one would expect to observe a higher matching rate. Hence, taking the number of control patients into account has the advantage of a smaller confidence

interval for a larger number of control patients. Consequently, the proposed definition of the $100 \cdot (1 - \alpha_{CI})\%$ CI is used hereafter. The resampling CI method is here combined with propensity score matching, but the fundamental idea of this method can be combined with any matching algorithm.

Table 1: *Steps of the resampling CI method at interim analysis (adapted from Weber et al. (2019)).*

Given entities:
b the number of resampling steps.
$n_{control}$ the number of control patients in already recruited study arm.
$n_{treated,interim}$ the number of treated patients at interim analysis.
$n_{treated,max}$ the maximal number of treated patients if applicable.

---

1. Repeat (a) - (d) b times:
   (a) Sample $n_{control,interim} = n_{treated,interim}$ patients without replacement out of the control group.
   (b) Calculate propensity scores for sampled control patients and treated patients.
   (c) Conduct a 1:1 matching according to the logit of the propensity scores.
   (d) Calculate the matching rate mr.
2. Calculate the mean matching rate $\overline{mr}$ of the b matching rates calculated in step 1.
3. Calculate the lower limit of the $100 \cdot (1 - \alpha_{CI})\%$ confidence interval using formula (2.2).
4. Calculate the total number of treated patients needed for analysis as in formula (2.3 or 2.4).

So far, the available historical data have contained a limited number of patients. When a large historical control exist, it might be possible to find more than one matching partner to the majority of control patients and the resampling CI method may not give satisfying results. Including more controls may enhance power or one can reduce the number of treated patients. Therefore, the question of how many control patients could be matched per intervention patient arises. The procedure introduced in Section 2.1.2 address this setting of a large

control group and makes it possible to iteratively determine the number of matching partners under the trial-specific matching rate.

Comment: Parts of the following Chapter (Section 2.1.2) find application in the Matched Threshold Crossing Design by Krisam et al. (2020), which is already submitted. The part of the manuscript describing the iterative determination of the number of matching partners has been written by myself but may contain comments and corrections from the co-authors.

2.1.2 Iterative Matching Procedure

In the proposed design, a sample size calculation determines the number of treated patients $n_{treated}$ for a balanced design. Thereof, an initial number of $n_{treated,interim}$ treated patients are recruited (a fixed number or a proportion of $n_{treated}$) and a large data set including $n_{control}$ control patients is available. As before, a matching procedure is conducted twice in the considered adaptive design. In the interim analysis, the matching determines the number of matching partners M by an iterative process (compare Table 2); furthermore, the matching rate is calculated at interim analysis to extrapolate the matching rate for recalculating the sample size in the treated group. In the final analysis, the matching is performed to find the fixed number of matching partners M for the treated patients. It is obvious that higher numbers of matching partners M will lead to a more powerful trial. However, when increasing M the matching rate may decrease because it gets more challenging to find suitable matching partners. The statistical analysis is only based on the matched patients; therefore the matching rate should also be sufficiently high to avoid a power loss of the trial. Thus, the interim analysis aims to select a suitable number of matched controls M, which also guarantee an adequately high matching rate. The iterative procedure at interim analysis starts by a 1:1 propensity score matching and is followed by calculating the corresponding matching rate. In the next step, a 1:2 propensity score matching ($M = 2$) is performed and the matching rate is calculated, respectively. The number of matching partners M is increased as long as the matching rate is equal or higher than the 1:1 matching rate minus a predefined tolerance criterion τ. This tolerance parameter τ defines the maximally tolerated deviation from the 1:1 matching rate. If e.g., $\tau = 0$, the 1:2 matching rate is not allowed to be smaller than the 1:1 matching rate, otherwise M will be set to 1. Choosing a value of $\tau = 0$ ensures that the maximum number of treated patients is included into the

analysis, but may ignore that a larger control group could be built; this corresponds to the most conservative approach. If $\tau = 0.05$ is predefined and suppose the 1:1 matching rate is 0.95 at the interim analysis, so the iterative procedure will increase M as long as the calculated 1:M matching rate does not fall below $0.95 - 0.05 = 0.9$. The pseudocode for the iterative process is given in Table 2.

Table 2: *Pseudocode for the setup of the iterative matching procedure (Krisam et al. (2020)).*

---

step 1:
  $M = 1$
  perform 1:1 propensity score matching
  calculate matching rate $mr_{1:1}$
  set $M = 2$
step 2:
  perform 1: M propensity score matching
  calculate matching rate $mr_{1:M}$
    if $(mr_{1:1} - \tau\ ) \leqslant mr_{1:M}$
      increase M to $M + 1$ and perform step 2
    else
      stop

For the number of patients in the treatment group, the following holds $n_{treated,interim} \leqslant n_{treated}$. To ensure that M suitable matching partners per intervention patient can still be found in the final analysis, the maximum number of control patients per intervention patient $M_{\max}$ needs to be predefined. To determine an estimate of $M_{\max}$, the number of patients needed for a balanced design complying 80% power and 5% type I error rate can be used for estimating the number of treated patients in the trial. This number $n_{treated,planned}$ can be calculated by established sample size formulas depending on the outcome and the trial design. Using these quantities, the estimate of $M_{\max}$ is given by

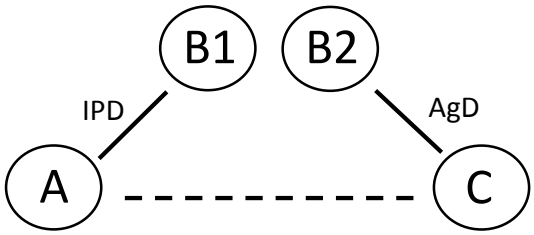
$$M_{\max} = \left\lfloor \frac{n_{control}}{n_{treated,planned}} \right\rfloor .$$

The iterative matching procedure can be combined with two-stage designs including an interim analysis allowing for a sample size recalculation, a stop for futility, or a stop for efficacy based on the treatment effect observed at the interim analysis, for example, the matched threshold crossing design (Krisam et al., 2018). The test statistics of the two stages (interim and final analysis, respectively) in such designs need to be independent in order to ensure type I error rate control. Therefore, one has to note that in case of such designs, the controls matched to patients at interim analysis are not reassigned to keep the independence. On the other hand, if the interim analysis is just used for estimating the matching rate and determine the number of matching partners, the final matching partners can be found at the final analysis. Assigning the matching partners at final analysis may have the advantage that the optimal partners are found under consideration of all patients considered in the analysis.

2.2 Synthesis of Evidence

The situation considered throughout the part about the synthesis of evidence in this thesis is the following: two treatments A and C are compared to a common comparator B in head-to-head trials. For trial A versus B (AB), IPD is available, for the trial C versus B (CB), only AgD is accessible from published results (see Figure 1).

Figure 1: *The plot shows the situation of the indirect comparison A versus C considered in this simulation study. To illustrate that cross-trial differences may exist, treatment B is described as B1 for the individual patient data (IPD) trial and B2 for the aggregated data (AgD) trial (Weber et al., 2020a).*



The aim is to demonstrate a treatment effect between treatment A and C (AC); this is called an indirect comparison. This thesis deals with a frequentist setting, so all the methods are explained accordingly.

In the current Section 2.2 the established methods for indirect comparisons (MAIC and the method of Bucher) are explained, followed by subsections about meta-analysis and the approximate adjustment which are needed for the ways to include multiple studies in indirect comparisons (Subsection 2.2.4).

2.2.1 Indirect Comparisons

There are several established methods addressing the situation of an indirect comparison, such that the method of Bucher (Bucher et al., 1997), MAIC (Signorovitch et al., 2010, 2012), simulated treatment comparison (STC) (Caro and Ishak, 2010; Ishak et al., 2015a), cross-design synthesis (Droitcour et al., 1993), or likelihood reweighting methods (Nie et al., 2013). This thesis focus on the widely used and accepted method of Bucher and the MAIC; the latter includes a matching step. Later on, the methods will be compared and extended for the use of multiple studies.

2.2.1.1 Method of Bucher

To perform the method of Bucher, the treatment effects and the corresponding variances on aggregated data level of the two studies AB and CB are sufficient; IPD is not needed. The method of Bucher preserves the within-study randomization since the treatment effects are calculated for each trial separately. This calculation takes the randomization within the respective trial into consideration. Furthermore, a common comparator is needed for calculating this indirect comparison; assuming head-to-head trials AB and CB are available treatment B is the common comparator between trials AB and CB. The assumptions made are comparable study populations for essential effect modifiers. If the treatment effect differs according to another variable, this variable is an effect modifier, whereas the presence of a confounder variable introduces bias in the estimated treatment effect.

Following the method of Bucher, the effect estimate δ_{AC} for the indirect comparison AC is

given by

$$\delta_{AC} = \delta_{AB} - \delta_{CB}, \tag{2.5}$$

with δ_{AB} and δ_{CB} denoting the respective effect estimate reported in trial AB and CB. The variance of the indirect effect estimate δ_{AC} is given by

$$Var(\delta_{AC}) = Var(\delta_{AB}) + Var(\delta_{CB}).$$

Insufficient comparability of studies according to important effect modifiers leads to a violation of assumptions for the method of Bucher. MAIC addresses this issue of differing patient populations by a matching procedure. However, IPD needs to be available for one trial and AgD for the other trial to conduct an indirect comparison by MAIC. When IPD is available for both trials, a propensity score matching or outcome regression approach might be more appropriate for estimating a treatment effect (Phillippo et al., 2016).

2.2.1.2 MAIC

Without loss of generality, it is assumed that IPD is available for the comparison AB and solely AgD for the trial comparing CB. The MAIC approach addresses this situation and makes use of the IPD data. The aim is to match the IPD to the AgD of the other trial to reach balance in summary measures of the baseline characteristics of the two trials. The matching procedure selects a weight ω_i for each patient in trial AB (IPD available), which follows the idea of propensity score matching (Rosenbaum and Rubin, 1983, 1984). The variable T denotes the trial, here $T = 0$ corresponds to trial AB and $T = 1$ to trial CB, respectively. The outcome is described by Y and X includes the baseline covariates, which will be considered within the matching procedure. In the case of $T = 1$ there are no individual values available, but the means or proportions of baseline covariates $\overline{\boldsymbol{x}}_{CB}$ and the outcome $\overline{y}_{CB}$ are observed. Considering these quantities, an estimate of the effect for trial AC is given by

$$\delta_{AC} = \frac{\sum_{i=1}^{n} y_i(1-t_i)\omega_i}{\sum_{i=1}^{n}(1-t_i)\omega_i} - \overline{y}_{CB} \tag{2.6}$$

with n the number of patients in AB and CB together ($n = n_{AB} + n_{CB}$, with n_{AB} the number of patients in trial AB and n_{CB} the number of patients in trial CB, respectively), t_i the trial affiliation of patient i ($t_i = 0$ for AB, $t_i = 1$ for CB), and $\omega_i = \frac{P(T_i=1|\boldsymbol{x}_i)}{P(T_i=0|\boldsymbol{x}_i)}$ the weight for

patient i (for $i = 1, \ldots, n_{AB}$), which is the odds between being a patient in trial AB versus belonging to trial CB given the baseline covariates $\boldsymbol{x}_i$. These weights achieve that patients who better fit in the CB than the AB trial (according to the observed baseline characteristics) will be up-weighted to balance between the trials. Before calculating the estimated treatment effect, one has to estimate the weight ω_i. As frequently in propensity score matching, it is assumed that the weights for the correspondence to one of the trials AB or CB follow a logistic regression model

$$\omega_i = \exp(\alpha + \boldsymbol{x}_i'\boldsymbol{\beta}),$$

where x_i includes the baseline covariates of patient i (for $i = 1, \ldots, n_{AB}$). Because IPD of baseline characteristics is only available for one of the trials, the maximum likelihood method cannot be applied. Instead, the method of moments addresses this setting and allows to estimate $\hat{\boldsymbol{\beta}}$ for the coefficients $\boldsymbol{\beta}$ (details see Appendix B). In order to determine the estimates $\hat{\boldsymbol{\beta}}$, the following equation needs to be solved

$$0 = \frac{\sum_{i:t_i=0}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{CB}) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{i:t_i=0} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

where $t_i = 0$ the trial affiliation of patient i to trial AB for $i = 1, \ldots, n_{AB}$. The optimization with respect to $\boldsymbol{\beta}$ is done using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970). The effect estimate $\hat{\delta}_{AC}$ is determined by using the before calculated weights $\hat{\boldsymbol{\omega}}$ and plug them into Equation (2.6). Hence, the estimate $\hat{\delta}_{AC}$ of δ_{AC} based on $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\delta}_{AC} = \frac{\sum_{i:t_i=0} y_i \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{\sum_{i:t_i=0} \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})} - \overline{y}_{CB}.$$

The standard errors of $\hat{\delta}_{AC}$ are calculated using a robust sandwich estimator (Huber, 1967; White, 1980).

Weighting patients in the calculation of a treatment effect to balance populations reduces the effective sample size (ESS). For MAIC, the ESS is calculated to measure the differences in baseline characteristics between the trials. The set of estimated weights $\hat{\boldsymbol{\omega}}$ contains the information about these differences. The ESS is calculated as follows

$$n_{effective} = \frac{(\sum_{i=1}^{n_{AB}} \hat{\omega}_i)^2}{\sum_{i=1}^{n_{AB}} \hat{\omega}_i^2},$$

where n_{AB} describes the sample size of trial AB and $\hat{\omega}_i$ the estimated weight for patient i (Phillippo et al., 2018).

In practice, it is likely that two or even more studies compare the same treatments. Combining evidence from different studies is done by meta-analysis.

2.2.2 Meta-analysis

Before technically a meta-analysis is performed, an intensive literature search according to clear rules is conducted, which identifies several independent studies comparing the same treatments. This is called a systematic review which aims to find, summarize, and rate the quality of all available evidence. Then one may be interested in the common treatment effect over the chosen studies. The results in terms of treatment effects and the corresponding variance of individual studies can be combined by mathematical considerations, which are then called a meta-analysis (Borenstein et al., 2011). Trials are not merely combined to increase sample size; methods for meta-analysis adjust for differences in sample size, variability in treatment effects and heterogeneity between trials. Appropriately conducted meta-analyses make a more objective evaluation of treatment effects possible (Egger and Smith, 1997). Facing a frequentist setting, the most commonly used approaches are the fixed and random-effects models. The fixed-effects model assumes that the trials are based on the same (true) treatment effect, which means they are comparable in terms of the target population and the definition of dependent and independent variables. The only source of error in the treatment effect is, therefore, within the study. The inverse of the study variance then provides the weights for the included studies. That those treatment effects are the same is a strong assumption which may be violated in practice even if the studies are similar enough to fulfill the literature search criteria. The random-effects model allows for such variations by assuming a distribution for the true effect size. The weights are again calculated by the inverse of the variance, but two sources of error apply in this setting, the within-study variance and the between-study variance, which are included in the weight calculation (Sutton et al., 2000).

2.2.3 Approximate Adjustment

In some of the approaches which will be introduced in Section 2.2.4, trials are included in several indirect comparisons. To correct for this multiple use, the approximate adjustment, according to Rücker et al. (2017) is applied. This method was originally proposed for the inclusion of multi-arm trials in a generic inverse variance meta-analysis to avoid unit-of-

analysis errors. The idea is to increase the standard error of a comparison when using the full data set several times instead of splitting the data. In the setting of an indirect comparison having several AgD studies, one could split IPD to avoid multiple uses, which addresses a situation closely related to Rücker et al. (2017). The standard deviation of each indirect comparison AC_{ij} ($sd(\delta_{AC_{ij}})$) is adjusted by

$$sd(\delta_{AC_{ij}}) = N_{CB} \cdot sd(\delta_{AB_i}) + N_{AB} \cdot sd(\delta_{CB_j})$$

with N_{AB} denoting the number of AB trials, N_{CB} the number of CB trials, $i = 1, \ldots, N_{AB}$, and $j = 1, \ldots, N_{CB}$.

Until now, the methods for adjusted indirect comparisons are designed to include one study per direct comparison. The following section consists of the proposed solutions for incorporating multiple studies within MAIC under a frequentist setting. Methods like IPD meta-analysis cover situations where IPD is available on both sides. The same holds true for cases where only AgD can be used (Sutton et al., 2000; Stewart et al., 2012).

Comment: Parts of the following Chapter (Section 2.2.4) are already included in the submitted manuscript Weber et al. (2020b). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

2.2.4 Inclusion of Multiple Studies in Indirect Comparisons

Different situations of multiple studies may occur. There could be one IPD trial and multiple AgD trials (A.), multiple IPD trials and one AgD trial (B.), or multiple IPD and AgD trials (C.). The underlying situation includes IPD for the AB trial(s) and AgD for the CB trial(s).

A. **One IPD trial (AB) and multiple AgD trials (CB)**

A.1. Pool AgD:

- The AgD trials are pooled by a meta-analysis. The meta-analysis results (treatment effect and the corresponding standard error) are used for the indirect comparison.

- A weighted average of the aggregated data according to the standard error of the relevant treatment effect is calculated.
- The IPD within the MAIC is matched to the weighted average of the aggregated data.
- The indirect comparison is conducted using the results of the steps before.

A.2. All indirect comparisons:

- All indirect comparisons are conducted separately, applying the variance correction (see the adjusted standard deviation in Section 2.2.3).
- A meta-analysis is used to combine the effect estimates calculated by the indirect comparisons.

B. **Multiple IPD trials (AB) and one AgD trial (CB)**

B.1. Pool IPD:

- The IPD trials are pooled into one data set.
- One indirect comparison is conducted using the pooled IPD data.

B.2. All indirect comparisons:

- All indirect comparisons are conducted separately applying the variance correction (see the adjusted standard deviation in Section 2.2.3).
- A meta-analysis is used to combine the effect estimates calculated by the indirect comparisons.

C. **Multiple IPD trials (AB) and multiple AgD trials (CB)**

C.1. Pool IPD, pool AgD:

- The IPD trials are pooled into one data set.

- The AgD trials are pooled by a meta-analysis. The meta-analysis results (treatment effect and the corresponding standard error) are used for the indirect comparison.
- A weighted average of the aggregated data according to the standard error of the relevant treatment effect is calculated.
- The IPD within the MAIC is matched to the weighted average of the aggregated data.
- One indirect comparison is conducted using the results of the steps before.

C.2. Pool IPD:

- The IPD trials are pooled into one data set.
- The AgD trials are considered separately.
- All indirect comparisons are conducted using the pooled IPD data set and applying the variance correction (see the adjusted standard deviation in Section 2.2.3).
- A meta-analysis is used to combine the effect estimates calculated by the indirect comparisons.

C.3. All indirect comparisons:

- All indirect comparisons are conducted separately applying the variance correction (see the adjusted standard deviation in Section 2.2.3).
- A meta-analysis is used to combine the effect estimates calculated by the indirect comparisons.

When IPD trials are pooled and a common effect estimate is calculated, one needs to take the clustered structure into account. For example, this can be done by including a random intercept for the trial in a mixed-effects regression model for estimating the treatment effect over all the IPD trials.

Chapter 3

Results

In this chapter, the methods introduced in Chapter 2 are evaluated using simulation studies. For each simulation study, the data generating process, the simulation scenarios, and the corresponding results are presented.

Comment: Parts of the following Section 3.1 have already been published in Weber et al. (2019). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

3.1 Generation of Evidence - Resampling CI Method

3.1.1 Data Simulation

For the evaluation of the resampling CI method and its comparison to the naïve approach, a simulation study using 10,000 runs is performed. The distribution parameters for the involved baseline variables are chosen motivated by a clinical example (Section 3.1.3), which deals with patients suffering from acute cerebral infarction. Simplifications like distribution assumptions for baseline variables and fewer variables within the matching procedure were made within the simulation study.

The outcome variable is assumed to be binary indicating some favourable event. The corresponding hypotheses are formulated in terms of rates

$$\begin{aligned} \mathrm{H}_0 &: \ \pi_{control} \geqslant \pi_{treated} \\ \mathrm{H}_1 &: \ \pi_{control} < \pi_{treated}, \end{aligned}$$

with $\pi_{control}$ and $\pi_{treated}$ are the true event rates in the control and the treatment group, respectively.

The simulated data include three binary variables (X_1, X_2, and the group variable Z), one categorical variable (X_4), and two continuous variables (X_3, X_5). The variables are used to simulate group assignment and the outcome variable (Y), as well as they are considered within the matching procedure.

First, two binary (X_1, X_2) and one continuous variable (X_3) are independently sampled, which describe, for example, gender, diabetes (yes/no), and age. The binary variables are assumed to follow a binomial distribution; the continuous variable is sampled out of a normal distribution.

$$\begin{aligned} X_1 &\sim Bin(1; 0.5) \\ X_2 &\sim Bin(1; 0.2) \\ X_3 &\sim \mathcal{N}(70; 15) \end{aligned}$$

The assignment to the treatment or the control group depends on the variables X_1 and X_3. In the subsequent step, the group variable is simulated based on a logistic regression model using the baseline variables X_1 and X_3 as covariates. In the following, the group variable is considered as Z.

$$Z = logit\left(\frac{P(Z=1)}{1-P(Z=1)}\right) = -0.6 + 0.35X_1 - 0.01X_3$$

Based on the group allocation, two additional variables, X_4 and X_5, are simulated. The variable X_4 is an ordinal variable assumed to follow a binomial distribution with ten levels which may represent the Alberta Stroke Program Early CT score (ASPECTS) here. The ASPECTS is a tool for detecting early ischemic changes on non-contrast computed tomography (CT) scans (Barber et al., 2000). The variable X_5 follows a normal distribution and

describes here the National Institutes of Health Stroke Scale (NIHSS). The NIHSS is a tool to assess stroke severity (Lyden et al., 2001). These two additional variables are sampled out of different distributions (according to the group).

$$X_{4,control} \sim Bin(10; 0.8)$$
$$X_{4,treated} \sim Bin(10; 0.75)$$
$$X_{5,control} \sim \mathcal{N}(17; 5)$$
$$X_{5,treated} \sim \mathcal{N}(16; 4)$$

Under the alternative hypothesis the outcome is then sampled out of a logistic regression model using variables X_4 and Z as covariates

$$Y_{H_1} = logit\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = -0.5 + Z + 0.2X_4.$$

In this outcome model, the coefficient for the group variable Z is chosen to be $\beta_{Z,outcome} = 1$.

Under the null hypothesis treatment and control are assumed to perform equally, so the outcome is assumed to follow a binomial distribution:

$$Y_{H_0} \sim Bin(1; 0.5).$$

Intending to simplify the simulation study, the baseline variables X_1 to X_5 are assumed to be independent. However, in practice, correlations are likely to occur and should be taken into account when selecting the matching variables. By using a logistic regression model for the group allocation and sampling clinical variables out of different distributions for treatment and control patients, differences between the groups are included, which can be balanced by the matching procedure.

The propensity score estimation is done by using a logistic regression model for the group as outcome variable (Z) and the baseline variables X_2, X_3, and X_5 as covariates. This model includes baseline variable X_2, which was not used for group assignment. Considering X_2 in the propensity score model leads to misspecification. However, the true model is usually not known, and therefore, this setting avoids to get over-optimistic results in the simulation study.

A set of confidence levels is considered with $\alpha_{CI} \in \{0.01, 0.05, 0.1\}$, and hence the resampling CI method is evaluated for 99%, 95%, and 90% confidence intervals in this simulation study.

When all patients are recruited according to the recalculated sample size, the null-hypothesis is tested by the McNemar test for paired data. An approach for paired data is used to account for the matched design. In practice, an alternative option to adjust for matching variables and additional confounders may a generalized mixed effects model considering the matching ID as random effect.

Fixed Time Point - Varying Number of Control Patients

The number of patients in the control group $n_{control}$ and a fixed fraction t of patients for the interim analysis are needed to start a prospective observational trial including an adaptive matching approach. To evaluate the new approach by its power, type I error rate, matching rate, and sample size, the *time point* for the interim analysis needs to be fixed. It is set to $s = 0.5 \cdot n_{control}$, with

$$n_{control} \in \{25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300\},$$

because this might be the most intuitive *time point*. In a fixed design, a number of 142 patients per group would have been needed to show the simulated treatment effect with a power of 80% at a type I error rate of 5%. Underpowered, less than 142 patients per group, as well as overpowered scenarios, more than 142 patients per group, are investigated. Underpowered situations may occur when the expected effect in the existing trial, where the control group of the prospective observational trial is taken from, was higher than expected in the new trial. When the existing trial was planned based on a smaller expected effect or included multiple primary hypothesis, this may result in a greater control group than needed for a fixed design and leads to an overpowered scenario. At the interim analysis, the matching rate on $b = 200$ resampling sets of size $n_{treated,interim}$ is calculated. By using the simulated data as described in Section 3.1.1 and performing the steps in Table 1, the proposed method is compared with the naïve approach. The properties of the two approaches are evaluated according to the matching rate at the final analysis, the recruited sample size $n_{treated,final}$, as well as the type I error rate and power at final analysis. To calculate the power, the number of correct test decisions under the alternative hypothesis is counted. To assess the type I error rate, the rate of rejected hypotheses under the null hypothesis is determined.

Time Point of Interim Analysis

Now, the aim is to assess the *time point* for the interim analysis. A fixed number of patients $n_{control}$ in the control group for a prospective observational trial is given. The confidence level for the resampling CI method is set to 99%, corresponding to an alpha-level of $\alpha_{CI} = 0.01$. The *time points* s considered for the interim analysis are

$$s \in \left\{ \left\{ \frac{1}{10}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{9}{10} \right\} \cdot n_{\text{control}} \right\}.$$

For recalculation of the sample size, the resampling CI method, as well as the naïve approach, is used. The conclusion and recommendation will be based on the evaluation of the matching rate, the recruited sample size $n_{treated}$, as well as the type I error rate and power at the final analysis.

The time point of interim analysis may be influenced by the size of the control group. Therefore, a small ($n_{\text{control}} = 50$), medium ($n_{\text{control}} = 150$), and a large ($n_{\text{control}} = 500$) sample size in the control group is considered. To obtain 80% power within each scenario, the regression coefficient for the group variable $\beta_{Z,outcome}$ varies between the considered sample sizes:

$$\begin{aligned} \text{small: } & \beta_{Z,outcome} = 2 \\ \text{medium: } & \beta_{Z,outcome} = 1 \\ \text{large: } & \beta_{Z,outcome} = 0.55 \end{aligned}$$

Due to problems in finding matching partners if fewer than 15 patients are included in the matching procedure at the interim analysis step, if a small sample size in the control group is assumed, the considered time points of interim analysis start at $\frac{1}{3} \cdot n_{\text{control}}$ and for the medium sample size at $\frac{1}{4} \cdot n_{\text{control}}$.

All simulations in Section 3.1.2 were done in R version 3.4.3 under the usage of the packages Matching (by using the function Match) and boot (by using the function inv.logit) (R Core Team, 2017; Sekhon, 2011; Canty and Ripley, 2017).

3.1.2 Simulation Results

First, the results varying the number of control patients under the usage of a fixed time point for the interim analysis are described and visualized. The subsequent paragraph deals

with the results for the time point of interim analysis, which is evaluated for both considered methods.

Fixed Time Point - Varying Number of Control Patients

The main evaluation measure in this setting is the matching rate. The matching rates depending on the number of patients in the control group are plotted in Figure 2 upper right corner (UR). Comparing the matching rate curves of the naïve approach at interim and final analysis, it can be observed that for all scenarios, the matching rate at interim analysis is higher than the matching rate for the final analysis. This means that the matching rate at interim analysis overestimates the true matching rate and results in undersized recruitment of patients for the final analysis. The consequence of a lower matching rate at the final analysis, caused by an overestimation of the matching rate at the interim analysis, is a loss in power (Figure 2 upper left corner (UL)).

The proposed method counteracts this overestimation by using equal sample sizes for the matching procedure at the interim analysis, which leads more likely to an underestimation of the true matching rate. Therefore, more treated patients are recruited (Figure 2 lower right corner (LR)), which achieves a higher matching rate at the final analysis. Hence, a higher number of matched pairs are included in the final analysis, which increases power.

When considering the naïve method, a dependency between the number of patients in the control group and the matching rate is observed: the matching rate increases with the number of patients in the control group. In contrast, for the resampling CI method the matching rate at final analysis stays on a constant level. The mean matching rate is around 92% for $\alpha_{CI} = 0.01$ applying the resampling CI method, whereas the mean matching rate of the naïve approach ($\alpha_{CI} = 0.01$) lies between 79 - 86% (Figure 2 UR and Table 3).

In this simulation study, the fixed design, based on 80% power, 5% type I error rate, and the Chi-squared test, would have required a number of $n = 142$ patients per group. For the proposed design, 80% power is reached for $n_{control} \approx 150$. Thus, the always intended power of 80% is achieved, requiring only slightly more patients in the control group than would have been needed in a fixed randomized design (Figure 2 UL). The number of required control patients is higher, because the observed matching rate is lower than 100% (Figure 2 UR).

In all scenarios, as well as for both considered methods, the type I error rate is approximately 5% (between 4.37% and 5.72%). As expected, a difference between the two methods according to the type I error is not observed (Figure 2 lower left corner (LL)).

Varying the confidence level within the resampling CI method results in small differences in the mean lower CI limit of the matching rate. The mean lower CI limit of the matching rate at interim analysis increases slightly for increasing α_{CI} or decreasing the confidence level, respectively (Table 3). This increase leads to a slightly lower number of recruited patients for lower confidence levels. For $\alpha_{CI} = 0.05$, the mean recruited sample size in the treatment group is around 4 patients higher than for $\alpha_{CI} = 0.1$, and for $\alpha_{CI} = 0.01$ is for another 8 patients higher, for details, see Table 4.

Time Point of Interim Analysis

The focus of this section is the results for a medium sample size in the control arm ($n_{\text{control}} = 150$), as the simulations for small and large sample sizes show comparable results.

Using the naïve method, for early time points of the interim analysis, a matching rate close to 100% is observed, but in the final analysis, it is less than 85% (Figure 3 UR). Even for later time points, the matching rate is below 90%, and as a consequence, the power is less than 80% for all considered time points (Figure 3 UL). The total sample size is lowest for the early time point (Figure 3 LR) because the matching rate at interim analysis is highest for this time point and indicates the lowest power value. As expected, the type 1 error rate is around 5% (Figure 3 LL).

The resampling CI method uses equal-sized groups at the interim analysis. When performing an early interim analysis, the matching rate is poor, and a high number of patients need to be additionally recruited for the final analysis. Comparing the matching rate at the final analysis between the different time points, the gain in the matching rate and, therefore, in power is small when performing an early interim analysis. With an increasing number of patients at interim analysis, the matching rate seems to converge and the observed changes are tiny (in the matching rate) when increasing the number of patients in the control group used at interim analysis above 50% (Figure 3 UR). Taking also the recruited sample size into account, it seems that a time point between $\frac{1}{2}$ and $\frac{2}{3}$ of the control patients is a good choice

as a trade-off between matching rate and sample size (Figure 3 LR). The matching rate lies between 90.7% and 93.4% for all considered time points. In all scenarios, the achieved power is around 80% and the type I error rate around 5% (Figure 3 UL, 3 LL).

For small sample sizes ($n_{control} = 50$) in the control group, it is observed that a later interim analysis could be a good choice, because sample size decreases and matching rate as well as power do not decrease in a considerable amount. Using only 50% of the control patients at the interim analysis in small trials leads to a low absolute number of patients, which underestimates the matching rate. For large sample sizes, an earlier time point seems to be the right choice since the matching rate converges already for early time points of the interim analysis. For large sample sizes, it is observed that a smaller absolute number of control patients leads to a good estimate of the matching rate. The results are shown in Appendix A.1 (Figure 11 and 12).

Table 3: *Mean matching rate/mean lower CI limit of the matching rate at interim and final analysis for the naïve approach and the resampling CI method for different numbers of patients in the control group. The resampling CI method is applied for different confidence levels ($\alpha_{CI} \in \{0.01, 0.05, 0.1\}$) (adapted from Weber et al. (2019)).*

$n_{control}$	Naïve		99%-CI		95%-CI		90%-CI	
	interim	final	interim	final	interim	final	interim	final
50	0.89	0.79	0.49	0.92	0.54	0.91	0.56	0.90
75	0.93	0.81	0.58	0.92	0.62	0.91	0.64	0.90
100	0.95	0.82	0.64	0.92	0.67	0.91	0.68	0.91
125	0.96	0.83	0.67	0.92	0.70	0.91	0.71	0.91
150	0.97	0.84	0.70	0.92	0.72	0.91	0.73	0.91
175	0.98	0.84	0.72	0.92	0.74	0.91	0.75	0.91
200	0.98	0.84	0.73	0.92	0.75	0.91	0.76	0.91
225	0.98	0.85	0.74	0.92	0.76	0.91	0.77	0.91
250	0.99	0.85	0.76	0.92	0.77	0.91	0.78	0.91
275	0.99	0.85	0.76	0.92	0.78	0.91	0.79	0.91
300	0.99	0.86	0.77	0.92	0.79	0.91	0.80	0.91

Table 4: *Mean total number of recruited patients in the treatment group for the naïve approach and the resampling CI method for different numbers of patients in the control group. The resampling CI method is applied for different confidence levels ($\alpha_{CI} \in \{0.01, 0.05, 0.1\}$) (adapted from Weber et al. (2019)).*

$n_{control}$	Naïve	99%-CI	95%-CI	90%-CI
50	57.20	103.05	94.11	89.97
75	81.88	130.51	122.83	119.10
100	106.24	158.13	150.99	147.45
125	130.73	187.46	180.42	176.88
150	155.34	215.80	208.85	205.32
175	179.98	245.10	238.09	234.52
200	204.84	273.86	266.81	263.19
225	229.42	303.28	296.11	292.42
250	254.39	331.75	324.53	320.80
275	279.19	361.06	353.69	349.90
300	304.00	389.50	382.07	378.21

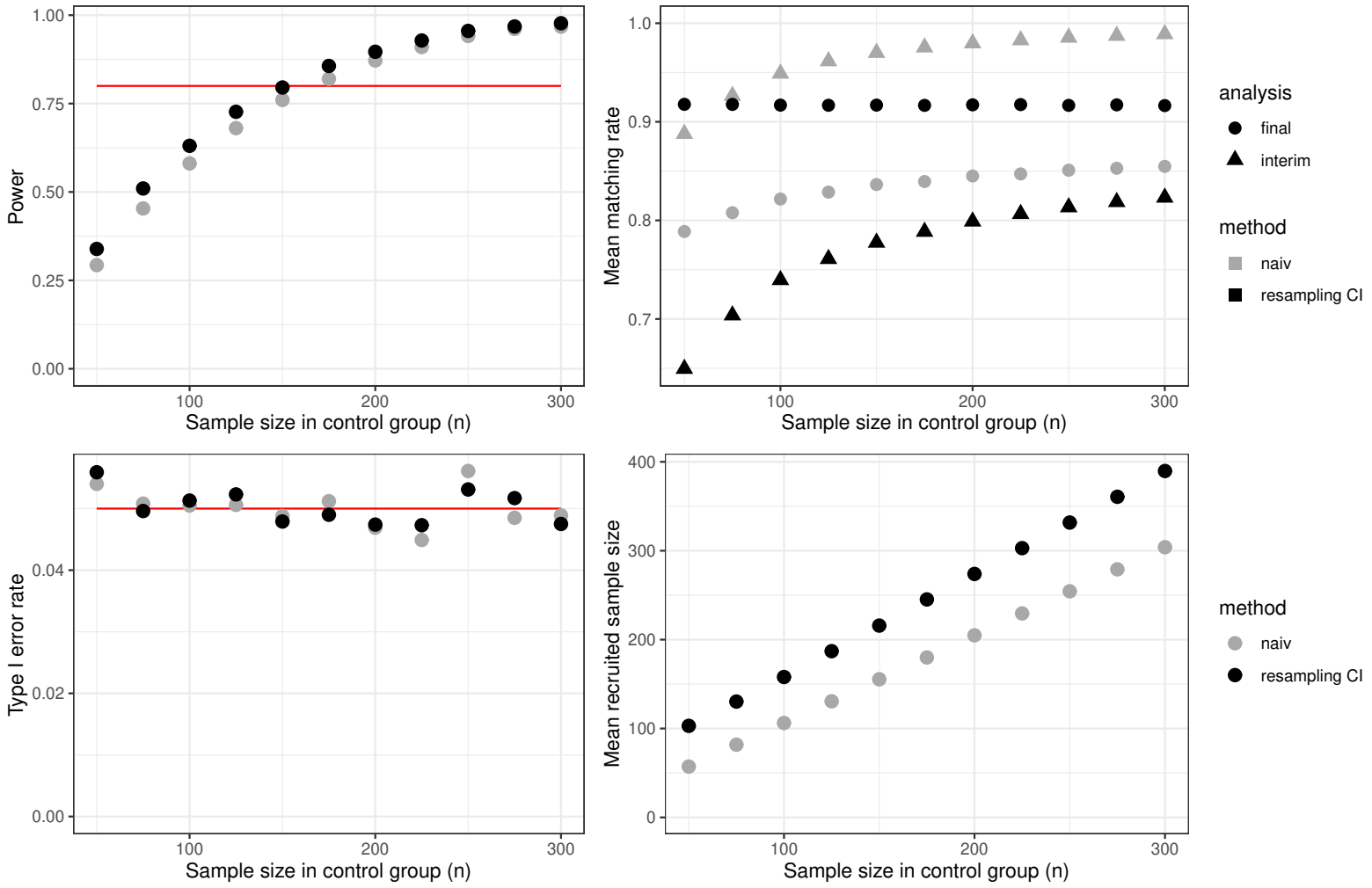


Figure 2: *Power, type I error rate, mean matching rate, and mean sample size in treated group for different sample sizes in the control group. Time point of interim analysis is* $\frac{1}{2} \cdot n_{control}$ *(adapted from Weber et al. (2019)).*

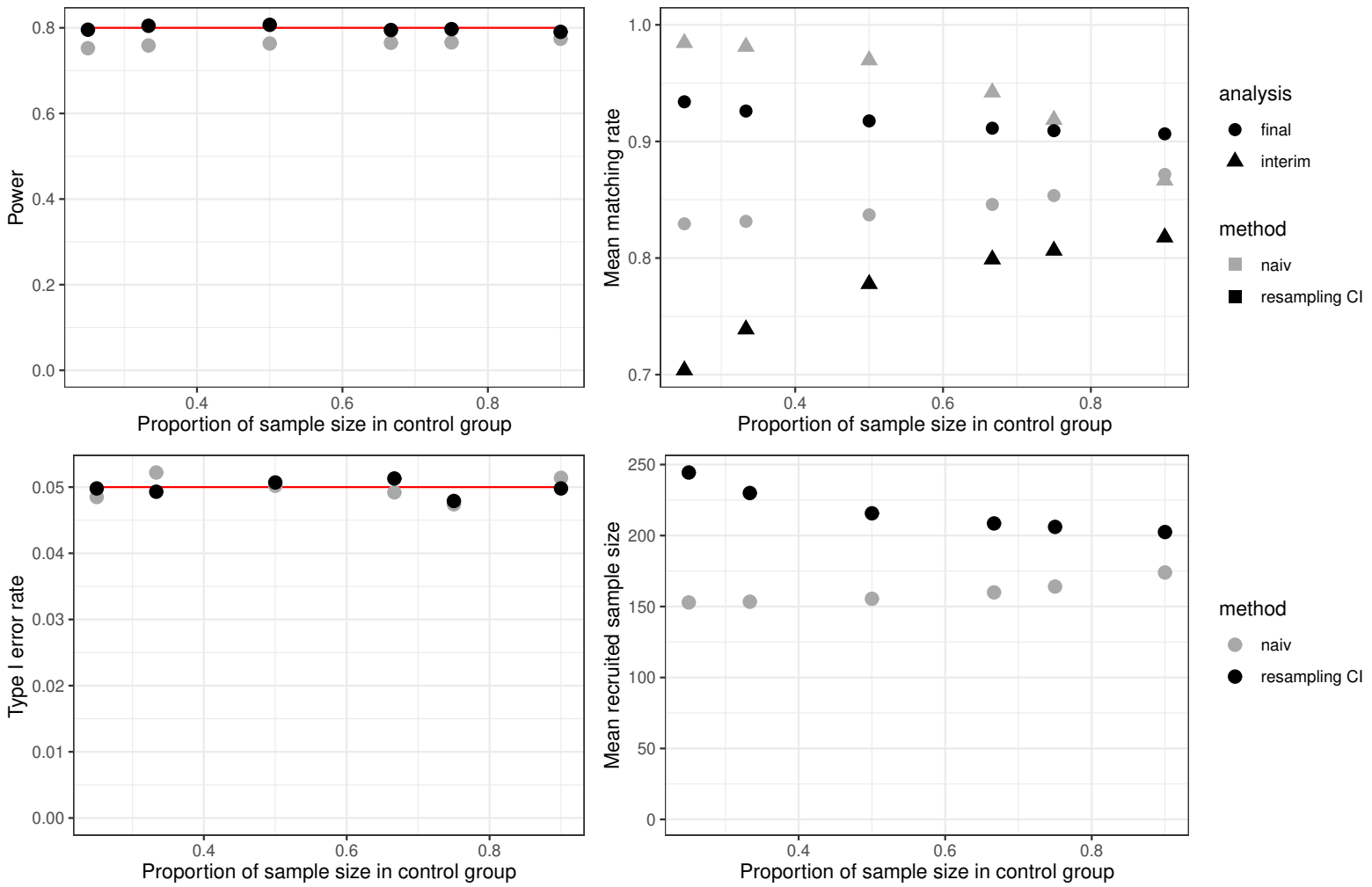


Figure 3: *Power, type I error rate, mean matching rate, and mean sample size in treated group for different time points of the interim analysis ($n_{control} = 150$) (adapted from Weber et al. (2019)).*

Comment: Parts of the following Section 3.1.3 have already been published in Schönenberger et al. (2019). The statistical analysis of this trial has been carried out by myself and the manuscript part about the statistical analysis has been written by myself, but may contain comments and corrections from the co-authors.

The following section 3.1.3 covers the application of the resampling CI method to a real data set, the KEep Evaluating Protocol Simplification In Managing Periinterventional Light Sedation for Endovascular Stroke Treatment (KEEP SIMPLEST) trial.

3.1.3 Real Data Example

The KEEP SIMPLEST trial (Schönenberger et al., 2019) had the objective to compare different aspects of peri-interventional management in patients with acute cerebral infarction treated according to a new system operating procedure (SOP) with patients having been randomized into the conscious sedation (CS) group of the SIESTA trial (Sedation vs. Intubation for Endovascular Stroke Treatment) (Schönenberger et al., 2016). The aim is to evaluate if the new SOP yields benefit in terms of patient outcome due to smooth, fast, clear, and effective processes compared to the early implementation stages of conscious sedation. The primary endpoint was early neurological improvement at 24 hours measured by the NIHSS.

The CS group of the Sedation vs. Intubation for Endovascular Stroke Treatment (SIESTA) trial includes 77 patients, but four were excluded before the matching analysis due to missing values in at least one of the matching variables. The study protocol of the KEEP SIMPLEST trial intended an interim analysis after 50 patients to estimate the matching rate and to perform a possible recalculation of the sample size. The actually recruited number of treated patients at interim analysis was 51. The conducted simulation study gave promising results using the resampling CI method. Therefore, the resampling CI method is used for the recalculation of the sample size using 200 resampling steps. Additionally, for reasons of comparability, the analysis is done using the naïve method, but this was not part of the medical publication (Schönenberger et al., 2019).

Within the matching procedure, four baseline variables were considered: Age, NIHSS on admission, premorbid modified Rankin Scale (mRS), and the ASPECTS score. The propensity score was estimated by a logistic regression model for group affiliation. For the matching on

the propensity score, a caliper width of 0.2 of the standard deviation of the propensity score was used. The interim analysis results in the following matching rates:

$$mr_{naiv} = 0.607$$

$$\overline{mr} = 0.461$$

The same pattern as in the simulation study is observed, the matching rate at interim analysis using the naïve method is substantially higher than the matching rate of the resampling CI method.

The extrapolation of $\overline{mr} = 0.461$ (resampling CI method) leads to a total sample size of 161. The evaluable KEEP SIMPLEST data set consists of 154 patients with complete data, seven of 161 included patients who had a missing ASPECTS score. The ASPECTS score is considered as matching variable; therefore, those patients who had a missing ASPECTS score were excluded from the trial. The matching procedure using the resampling CI method reached a matching rate of 94.5% in the final analysis; hence 69 pairs were found and analyzed. The naïve approach would result in a total sample size of 122. In case just 122 patients would have been used for the second matching procedure in the treated group, this would have resulted in 63 matched pairs and hence a matching rate of 86.3%. Thus, in the KEEP SIMPLEST trial, the resampling CI method achieves an 8.2% higher matching rate in the final analysis compared to the naïve approach.

The patient characteristics addressed in the propensity score matching are given in Table 5 for the SIESTA and the KEEP SIMPLEST trial before and after matching. One can recognize that patient characteristics are balanced after matching.

The KEEP SIMPLEST trial could not find a difference in early neurological improvement (NIHSS after 24 hours) and mRS at three months. Differences in secondary endpoints like the door-to-recanalization time with mean time in minutes of 128.6 (sd=69.47) versus 156.8 minutes (sd=75.91), mean duration of mechanical thrombectomy of 92.01 minutes (sd=52) versus 131.9 (sd=64.03), door-to-first angiographic image with mean time in minutes of 51.61 (sd=31.7) versus 64.23 minutes (sd=21.53), and computed tomography-to-first angiographic image time with a mean of 31.61 minutes (sd=20.6) versus 44.61 minutes (sd=19.3) were shorter in the group treated under the new SOP.

Further details and results of the KEEP SIMPLEST trial can be found in Schönenberger et al. (2019).

Table 5: *Patient characteristics of the SIESTA and the KEEP SIMPLEST trial before and after matching.*

	Before Matching		After Matching	
	SIESTA (n=73)	KEEP SIMPLEST (n=154)	SIESTA (n=69)	KEEP SIMPLEST (n=69)
Age, mean (sd)	71.1 (14.9)	76.1 (11.0)	72 (14.7)	72.7 (12.4)
NIHSS on admission, mean (sd)	17.4 (3.7)	14.3 (7.7)	17.2 (3.7)	16.9 (6.8)
premorbid mRS, n(%)				
0	36 (49.3)	48 (31.2)	32 (46.4)	32 (46.4)
1	18 (24.7)	40 (26.0)	18 (26.1)	17 (24.6)
2	13 (17.8)	22 (14.3)	13 (18.8)	15 (21.7)
> 2	6 (8.2)	44 (28.6)	6 (8.7)	5 (7.2)
ASPECTS, n(%)				
10-8	42 (57.5)	103 (66.9)	42 (60.8)	39 (56.5)
7-6	23 (31.5)	34 (22.1)	21 (30.4)	21 (30.4)
<6	8 (11.0)	17 (11.0)	6 (8.6)	9 (13)
Median (Q1-Q3)	8 (6-9)	9 (7-10)	8 (7-9)	8 (6.25-9)

Comment: Parts of the following Section 3.2 find application in the Matched Threshold Crossing Design by Krisam et al. (2020), which is already submitted. The part of the manuscript describing the iterative determination of the number of matching partners has been written by myself but may contain comments and corrections from the co-authors.

3.2 Generation of Evidence - Iterative Matching Procedure

Simulations investigate the characteristics of the iterative matching procedure. The simulations involve 10,000 runs for each data scenario which are described in Section 3.2.1. The maximal number of matching partners is set to $M = 10$, and a caliper of 0.2 is used within the propensity score matching. The tolerance is set to $\tau \in \{0, 0.05, 0.1\}$.

3.2.1 Data Simulation

The data for the iterative matching procedure is motivated by refractory acute myeloid leukemia (AML) patients. It is assumed that there exists a cohort of 1000 refractory AML patients treated with the current standard of care. To investigate a novel treatment for AML a single-arm phase II trial including 25 patients (at interim analysis) using the historical data as the control group is planned. The response rate under standard treatment is assumed to be $\pi_{control} = 0.3$; for the new therapy, a response rate of $\pi_{treated} = 0.5$ is assumed. The data consist of three baseline variables, two binary variables that may represent the prevalence of high-risk cytogenetics and the presence of a FLT3 mutation, as well as one continuous representing the patients' age. Two data scenarios are considered. First, patient characteristics in the treatment and control group are assumed to follow the same distributions. Age follows a normal distribution with $\mu = 55$ and $\sigma = 15$. The contingency table of the binary variables given in Table 6 is assumed for both cohorts.

Differences in baseline characteristics characterize the second scenario. In the control group the normally distributed age variable has $\mu = 60$ and $\sigma = 5$, whereas in the treatment group $\mu = 55$ and $\sigma = 15$. For the binary variable, the contingency tables given in Table 7 are assumed.

The response variable Y can be modelled by a logistic regression model conditioned on both binary and the continuous baseline variable, as well as including the treatment affiliation with

Table 6: *Contingency table of the binary variables (high-risk cytogenetics and FLT3 mutation) for both cohorts in case patient characteristics in the treatment and control group follow the same distributions.*

		Cytogenetics	
		yes	no
FLT3	yes	0.02	0.18
	no	0.32	0.48

Table 7: *Contingency table of the binary variables (high-risk cytogenetics and FLT3 mutation) for both cohorts in case patient characteristics in the treatment and control group have different distribution parameters.*

		Cytogenetics			
		Control		Treatment	
		yes	no	yes	no
FLT3	yes	0.04	0.2	0.02	0.18
	no	0.24	0.52	0.32	0.48

log odds ratio $\delta = \log\left(\frac{\pi_{treated}(1-\pi_{control})}{\pi_{control}(1-\pi_{treated})}\right) = \log\left(\frac{0.3}{0.7}\right)$

$$logit(y) = \beta_0 + \delta x_{tr} + \beta_{FLT3} x_{FLT3} + \beta_{Cyto} x_{Cyto} + \beta_{Age} x_{Age}.$$

The coefficients of this model are set to $\beta_0 = 2$, $\beta_{FLT3} = -0.2$, $\beta_{Cyto} = -0.5$, and $\beta_{Age} = -0.05$. Because the response distribution depends on these three baseline variables, all of them are considered within the matching procedure.

3.2.2 Simulation Results

For equal patient characteristic distributions in the control and the treatment group, most of the simulation runs reach $M = 10$ matching partners for $\tau = 0.1$. Decreasing τ leads to a shift towards less matching partners. The mean matching rate, its standard deviation (sd), and the distribution of the number of matching partners is given in Table 8 for equal patient characteristics. The mean matching rate is close to 1 in case $\tau = 0$ and all numbers of matching partners and decreases with increasing τ (which is not surprising because of the

design of the tolerance criterion). For $\tau = 0.05$ it is observed that the mean matching rate is similar for $M \in [2, 9]$ and is higher for $M = 1$ and $M = 10$. In case $\tau = 0.1$ the matching rate is around 92% for $M \in [3, 9]$; for $M = 10$ a matching rate of 97.7% is observed. When the predefined maximal number of matching partners is not reached, the mean matching rate reflects approximately the tolerance criterion τ.

In case the populations differ in regard to distribution parameters of baseline characteristics (9), the highest matching rate $mr = 0.867$ is reached for $M = 10$ and $\tau = 0$. Using higher tolerance values results in a lower matching rate, even if the maximal number of matching partners is reached. The maximal number of matching partners can be reached for less than 3% for $\tau = 0$, 10% and 25% for $\tau = 0.05$ and $\tau = 0.1$, respectively. The mean matching rate for $\tau = 0$ and $M \in [1, 9]$ is between 76% and 81%. For $\tau = 0.05$ it is around 75% and for $\tau = 0.1$ around 70%, respectively. When the number of matching partners is compared to the scenario with equal populations, there is a shift towards smaller numbers of matching partners, which is stronger for $\tau = 0$ compared to higher values of τ.

The distribution of matching partners is displayed in Figure 4 for both data scenarios and the considered tolerance values.

Table 8: *Mean and standard deviation of the matching rate split by the number of matching partners (M). The columns entitled n include the number of simulation runs ending with this number of matching partners. Values are given for a tolerance of* $\tau \in (0, 0.05, 0.1)$. *Populations of the control and intervention group are sampled from equal distributions.*

	$\tau = 0$			$\tau = 0.05$			$\tau = 0.1$		
M	mean	sd	n	mean	sd	n	mean	sd	n
1	0.997	0.011	380	1.000	0.000	6			0
2	0.998	0.008	424	0.963	0.021	50			0
3	0.999	0.006	453	0.962	0.016	85	0.914	0.028	7
4	0.999	0.007	479	0.962	0.013	126	0.925	0.014	24
5	0.998	0.009	510	0.961	0.010	160	0.921	0.017	38
6	0.999	0.008	481	0.961	0.011	184	0.921	0.016	64
7	0.998	0.009	542	0.961	0.013	290	0.925	0.016	104
8	0.999	0.008	615	0.961	0.013	323	0.923	0.013	113
9	0.999	0.007	600	0.961	0.011	450	0.922	0.014	193
10	0.999	0.007	5516	0.985	0.021	8326	0.977	0.029	9457

Table 9: *Mean and standard deviation of the matching rate split by the number of matching partners (M). The columns entitled n include the number of simulation runs ending with this number of matching partners. Values are given for a tolerance of* $\tau \in (0, 0.05, 0.1)$. *Populations of the control and intervention group are sampled from different distributions.*

	$\tau = 0$			$\tau = 0.05$			$\tau = 0.1$		
M	mean	sd	n	mean	sd	n	mean	sd	n
1	0.763	0.106	6395	0.762	0.099	2416	0.770	0.093	604
2	0.781	0.116	1493	0.730	0.106	2306	0.700	0.098	1506
3	0.787	0.121	745	0.739	0.109	1323	0.696	0.103	1284
4	0.787	0.130	425	0.739	0.116	972	0.696	0.108	1104
5	0.807	0.138	271	0.736	0.121	648	0.699	0.106	824
6	0.814	0.132	160	0.748	0.121	482	0.698	0.114	756
7	0.833	0.133	111	0.750	0.134	378	0.699	0.121	602
8	0.837	0.152	98	0.740	0.135	295	0.705	0.119	509
9	0.810	0.155	78	0.757	0.131	226	0.697	0.127	423
10	0.867	0.145	224	0.777	0.154	954	0.722	0.146	2388

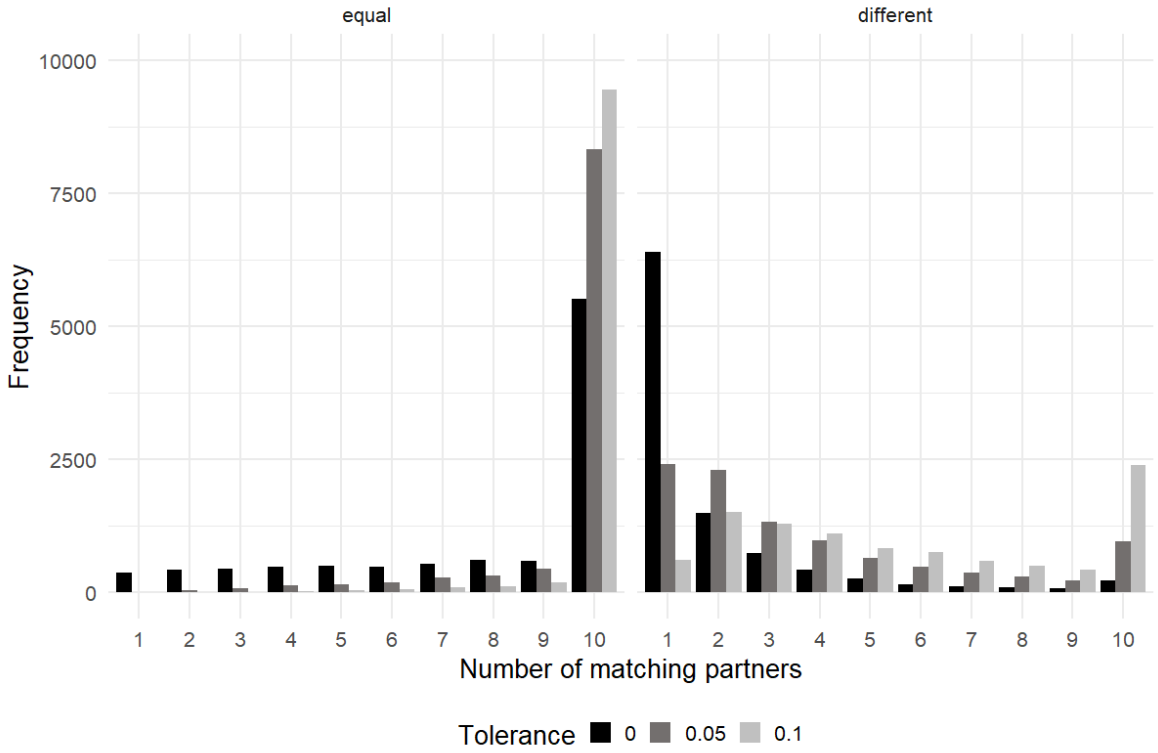


Figure 4: *The plot shows the distribution of the number of matching partners (10,000 simulation runs) for equal and different patient populations including tolerance values of* $\tau \in (0, 0.05, 0.1)$.

Comment: Parts of the following Section 3.3 are already included in the accepted manuscript Weber et al. (2020a). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

3.3 Synthesis of Evidence - Method Comparison

A simulation study covering a wide range of practically relevant scenarios in medical context is performed. The aim is to investigate the method of Bucher and the MAIC for indirect comparisons considering time-to-event (TTE) and binary endpoints. The evaluated methods and scenarios are transferable to continuous endpoints. The statistical properties of the methods are assessed and compared. The main evaluation measures used in the comparison are the bias in the estimated therapy effects, root mean squared error (RMSE), coverage, type I error rates, and power. The simulation comprises $n_{sim} = 10,000$ runs for each scenario. In case the method of Bucher is applied one has to assume that no differences between trial AB and CB are observed with respect to effect modifiers. Nonetheless, this assumption needs to be evaluated for each situation in practice. All simulations were done using R version 3.3.3 (R Core Team, 2017) using the packages corpcor (Schafer et al., 2017) and survival (Therneau, 2015).

3.3.1 Data Simulation

The simulation setting comprises two studies; one study covers the comparison between treatment A and B (AB) and another study compares treatment C versus B (CB). It is assumed that a study comparing A versus C is not available (Figure 1). The true treatment effects of treatment AB, BC, and AC are expressed on the log hazard ratio (HR) scale for a time-to-event setting and on the log odds ratio (OR) scale for a binary endpoint, respectively. In the data generating process of the full data set, beyond the treatment variable, one continuous and three binary variables are involved. A covariance matrix for the error term is considered which leads to correlations between the four variables (Table 10). Subsequently, the term "similar populations" refers to the situation where data for the trials AB and CB follows the same distribution, when divergence in the distribution parameters are assumed they are called "different". The event and the censoring times are sampled from a Weibull distribution ($\lambda_{event} = 0.0002$, $\nu_{event} = 1.8$, $\lambda_{censoring} = 0.00012$, $\nu_{censoring} = 2$, $max.time =$

100) for the generation of the TTE endpoint. The endpoint is then generated by a Cox proportional hazard model using the simulated event times, event statuses, and the patient characteristics as covariates. The binary endpoint is generated by a logistic regression model that considers the patient characteristics as covariates. The included covariates are called confounders. The outcome generation model with the link function $g(\cdot)$ looks as follows

$$g(y_i) = \beta_0 + \beta_{tr} x_{tr,i} + \beta_{b1} x_{b1,i} + \beta_{b2} x_{b2,i} + \beta_{b3} x_{b3,i} + \beta_c x_{c,i}$$

where $b1$, $b2$, and $b3$ refer to the binary variables, c to the continuous variable, and tr indicates the treatment variable. For TTE endpoints the link function between the log-hazard function and covariates is assumed to be linear and for binary endpoints the logit-link is used $g(y_i) = logit(y_i) = \log \frac{y_i}{1-y_i}$. The values for log HR and log OR of the confounders in the models with respect to the assumed true treatment effect (Table 12) are given in Table 11. In addition, some of the simulation scenarios cover an interaction term between a binary variable and the treatment assignment. In this case, the variable is called effect modifier. The inclusion of the interaction between binary variable 1 (x_{b1}) and treatment (binary variable 1 is an effect modifier for treatment) in the outcome generation model is shown in the following equation:

$$g(y_i) = \beta_0 + \beta_{tr} x_{tr,i} + \beta_{tr \cdot b1} x_{tr,i} \cdot x_{b1,i} + \beta_{b1} x_{b1,i} + \beta_{b2} x_{b2,i} + \beta_{b3} x_{b3,i} + \beta_c x_{c,i}$$

In Table 11 the corresponding log HR and log OR for the interaction term can be found. Note, that in case the interaction term is only included in one of the trials, the shared effect modifier assumption is violated which is assumed for the method of Bucher. The simulation study is limited to the described clinically inspired data because the aim is not to examine the influence of the number of confounders or distributions of patient characteristics itself, but rather the violation of assumptions and occurrence of cross-trial differences. The population in trial CB defines the target population where the treatment effect between A and C is estimated for. The true effect size of the trial AC is simulated as high, moderate, low, and no effect, the exact values for HRs and ORs are given in Table 12. The classification of effect sizes in terms of log HRs for TTE endpoints is done according to Skipka et al. (2015); for the ease of comparability, the log ORs for binary endpoints are set to comparable values. This classification of treatment effects is traced back to the benefit assessment of new drugs, which aims to test whether a new drug achieves an added benefit compared to the current

standard of practice. Sample size calculations for balanced designs in trials AB and CB are based on established formulas (Schoenfeld, 1983; Cohen, 1988) assuming the effects given in Table 12, 5% type I error rate, and 80% power.

Table 10: *Patient characteristics and covariance matrices used for the data generating process for the two trials (AB and CB) (adapted from Weber et al. (2020a)).*

	population		
	similar	different	
Variable	AB/CB	AB_1	CB_2
continuous, mean (sd)	55 (15)	55 (15)	65 (10)
binary 1 ($x_{b1} = 1$), %	0.7	0.7	0.5
binary 2 ($x_{b2} = 1$), %	0.8	0.8	0.6
binary 3 ($x_{b3} = 1$), %	0.4	0.4	0.45

Covariance matrix for AB/CB similar and AB_1:

$$\begin{pmatrix} 225 & 0.25 & 0.05 & 0.01 \\ 0.25 & 0.2 & 0.01 & 0 \\ 0.05 & 0.01 & 0.15 & 0.05 \\ 0.01 & 0 & 0.05 & 0.1 \end{pmatrix}$$

Covariance matrix for CB_2:

$$\begin{pmatrix} 100 & 0 & 0.05 & 0.01 \\ 0 & 0.25 & -0.01 & 0 \\ 0.05 & -0.01 & 0.1 & 0.05 \\ 0.01 & 0 & 0.05 & 0.15 \end{pmatrix}$$

3.3.2 Evaluation Measures

The main evaluation measures to assess the performance of the two methods are the bias of the effect estimate (the difference to the true treatment effect as given in Table 12), the RMSE, the power, the type I error rate, and the two-sided 95% CI coverage. Where the calculated CIs for the effect estimates in the regression models are based on a normal approximation.

Table 11: *Regression coefficients in terms of log hazard ratio for Cox proportional hazards models and log odds ratio for logistic regression models considered for simulation of outcomes in trials AB and CB (adapted from Weber et al. (2020a)).*

	Time to event	Binary
Variable	log HR	log OR
continuous	-0.0051	0.06
binary 1 ($b1 = 1$)	-0.2	-1.76
binary 2 ($b2 = 1$)	0.18	1.26
binary 3 ($b3 = 1$)	-0.14	-0.2
interaction:		
treatment and binary 1 (=1)	0.02	0.04

The power is assessed by the proportion of simulation runs where 0 (no effect) is not included in the two-sided 95% CI of the effect estimate of the indirect comparison AC when in fact an effect exists. The power is divided into the categories high, moderate, and low effect. In case of no effect, one is interested in the type I error rate which is based on the proportion of simulation runs where again 0 (no effect) is not covered by the two-sided 95% CI of the effect estimate for the indirect comparison AC. The bias is calculated by

$$bias = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\delta}_{AC,i} - \delta_{AC} \tag{3.1}$$

and the RMSE is given by

$$RMSE = \sqrt{\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\delta}_{AC,i} - \delta_{AC})^2} \tag{3.2}$$

where n_{sim} the number of simulation runs and δ_{AC} denotes the true treatment effect and $\hat{\delta}_{AC,i}$ its estimate in simulation run i (Morris et al., 2019). The aim is to minimize bias and RMSE, whereas power ought to reach high values, the type I error rate should be around 5%, and the CI coverage around 95%. All evaluation measures are calculated for the indirect comparison AC and correspond to the main treatment effect even if an interaction term is included in the regression model, because its effect cannot be assessed by the marginal effect (Norton et al., 2004).

Table 12: *True effect sizes in terms of log odds ratios for binary endpoints including the binary event rates* (π_1, π_2) *and log hazard ratios for time-to-event endpoints for different effect classes. The column AC is the difference between the effects in columns AB and CB (adapted from Weber et al. (2020a)).*

	Time to event			Binary		
	AC	AB	CB	AC	AB (π_1, π_2)	CB (π_1, π_2)
high	-0.69	-0.91	-0.22	-0.48	-0.7 (0.45, 0.62)	-0.22 (0.45, 0.51)
moderate	-0.22	-0.44	-0.22	-0.23	-0.45 (0.45, 0.56)	-0.22 (0.45, 0.51)
low	-0.05	-0.27	-0.22	-0.06	-0.28 (0.45, 0.52)	-0.22 (0.45, 0.51)
no	0	-0.22	-0.22	0	-0.22 (0.45, 0.51)	-0.22 (0.45, 0.51)

3.3.3 Simulation Scenarios

The definition of simulation scenarios aims to introduce cross-trial differences between the trials AB and CB. They are characterized by the following four aspects

- similar or different distributions of patient characteristics (proportions of categorical variables, mean and variance in distributions of continuous variables substantially differ), but note that the cut-off between similar and different (as in Table 10) depends on the variable and the objective of the comparison,

- inclusion of effect modification (interaction term between a binary variable and treatment),

- similar or different confounder variables included in data generating process (trial CB does not include the variable binary 3 (x_{b3})),

- differences in the presence of the interaction (trial AB does not include an interaction term).

The simulation scenarios are given in Table 13 where the distribution of patient characteristics is considered within each scenario.

Additionally, the influence of the true power considered in the sample size calculation of the head-to-head trials (AB and CB) on the performance, in particular, the power of the indirect comparison (AC) is investigated. The sample size calculations are based on established formulas (Schoenfeld, 1983; Cohen, 1988) by assuming the effects given in Table 12, a type I error rate of 5%, and a power of 80%, 90%, 95%, or 99%. The scenarios are again evaluated using characteristics such as the power, type I error rate, 95% CI coverage, and bias of the effect estimate in the indirect comparison.

All scenarios for the indirect comparisons are analyzed for a binary and a TTE endpoint.

Table 13: *Considered simulation scenarios (adapted from Weber et al. (2020a)).*

Scenario	Population		Confounders		Interactions	
	Similar	Different	Similar	Different	Similar	Different
I	x		x			
		x	x			
II	x		x		x	
		x	x		x	
III	x			x	x	
		x		x	x	
IV	x		x			x
		x	x			x
V	x			x		x
		x		x		x

3.3.4 Simulation Results

This section includes a paragraph for each of the evaluation measures introduced in Section 3.3.2, followed by a paragraph focusing on the influence of the planned power, which is considered in the sample size calculation of the individual trials, to the indirect comparison.

For the method comparison, the underlying assumptions for the sample size calculation for trials AB and CB evaluated under scenarios I to V are 80% power, 5% type I error rate,

and the treatment effect or proportions given in Table 12. Three different settings for the calculation of direct effect estimates are considered:

1. The regression models for trials AB and CB are adjusted for all relevant effect modifiers (in terms of an interaction) and confounders.

2. The regression models in trials AB and CB are only adjusted for confounders (effect modifiers are treated as confounders).

3. The regression models for trials CB are not adjusted for effect modifiers or confounders.

The scenarios are evaluated in all three settings if procurable. When effect modification is present, MAIC is applied twice: First, considering all confounders and effect modifiers as matching variables; and second, considering only effect modifiers as matching variables.

Each paragraph covers the results for TTE and binary endpoints. Initially, the results for setting 1 are described, when differences between the settings are observed they are mentioned in the corresponding paragraph. The detailed results of the different scenarios and endpoints based on the described evaluation measures when regression models (trials AB and CB) are adjusted for all effect modifiers and confounders (setting 1) can be found in Tables 14 to 18. All other results are given in Appendix A.2 (Tables 27 to 35). The differences in distributions of variables considered in the MAIC procedure influence the ESS. That means the ESS ($n_{effective}$) is independent of interactions or adjustment of regression models. Hence, ESS differs when the assumed set of effect modifiers differs, but the results are comparable for all considered settings considered for calculating the direct treatment effect (see Table 19 and Appendix A.2 Tables 36 to 39). A performance summary of the methods over all scenarios is given in Table 20 for TTE endpoints and in Table 21 for binary endpoints.

Power

In scenario I (Table 14), the method of Bucher and MAIC produce equal results, but adjusting MAIC for all confounders leads to a loss in power when patient characteristics differ. Scenario II is characterized by an interaction that makes MAIC reach higher power values in case there are differences in the confounder and effect modifier distributions for TTE endpoints (Table 15). When adjusting MAIC only for the effect modifiers, power is slightly decreasing.

For binary endpoints and when characteristics are similar, MAIC results in higher power values. Additionally, a small increase is observed when MAIC adjusts for effect modifiers only. In case confounder overlap differs (scenario III, Table 16), similar results as for scenario II are observed. In scenario IV (Table 17), the power values are relatively high and are comparable for both methods when populations are similar. When population distributions differ, adjusting MAIC for all confounders leads to power loss compared to only adjusting for effect modifiers. In case the effect modification is not considered within the regression models (setting 2, Appendix Tables 27 to 30), scenarios, where the effect modification is present in both trials, give better results in terms of power. When only unadjusted effect estimates are available for CB (setting 3, Appendix Tables 31 to 35), power decreases for scenarios where effect modification is only present in CB trials.

Type I Error Rate

Type I error rates are around 5% in scenario I for MAIC and the method of Bucher, as well as for similar and different confounder distributions and both endpoints (Table 14). In scenario II endpoints differ, TTE endpoints stay round 5%, whereas binary endpoints lead to a type I error rate around 10% (Table 15). In case confounder overlap differs (scenario III, Table 16), type I error rate is still around 5%, only MAIC shows higher values for binary endpoints. However, a clear type I error rate inflation is observed if effect modification is only present in trial CB (scenario IV and V, Tables 17 and 18). When populations are similar, the methods perform equally, but in case of differences MAIC leads to lower type I error rates, which are still highly inflated. For binary endpoints, MAIC leads to inflated type I error rates in all scenarios where effect modification is present. When the effect modification is not considered in the estimation of effects in direct evidence (setting 2, Appendix Tables 27 to 30), type I error rate is controlled for scenarios with effect modification in both trials (scenarios II and III).

Coverage

Coverage is observed to be between 90% and 95% in scenario I focusing on similar cohorts (Table 14). High treatment effects and different population distributions have a coverage below 90%, whereas for smaller treatment effects, the coverage is above 90%. When an effect

modifier is present (scenario II, Table 15), MAIC reaches coverage over 90% whereas Bucher leads to values lower than 90% for binary endpoints. Considering a TTE endpoint, results are the same between methods. When additionally confounders differ between trials in the binary case, MAIC reaches a higher coverage, whereas for TTE endpoints both methods lead to similar results. In scenario IV and V when population distributions differ MAIC reaches higher power values for both endpoints, but those values are below 85% (Tables 17 and 18).

Bias and RMSE

In scenario I bias and RMSE are the same when patient cohorts are similar, but in case of differences for MAIC higher bias and RMSE are observed for both endpoints (Table 14). For both endpoints, the bias and the RMSE are slightly higher for MAIC when all confounders are considered in the matching step of MAIC in scenario I. When effect modifiers are present and regression models for effect estimation are adjusted for those effect modifiers, Bucher and MAIC give similar results for TTE endpoints (Table 15). For the binary endpoint, MAIC results in lower bias and RMSE. In scenario III (Table 16), confounder overlap differs, only small differences to scenario II are recognized. Scenarios IV and V show the lowest bias and RMSE values for TTE endpoints (Tables 17 and 18)). In those scenarios, differences between Bucher and MAIC are negligible for both endpoints. When effect modification is not considered in the regression models for effect estimates of AB and CB, this leads to even slightly smaller bias and RMSE (setting 2, Appendix Tables 27 to 30). When CB is not adjusted for any confounder, bias and RMSE increase (setting 3, Appendix Tables 31 to 35).

Table 14: *Simulation results for scenario I (setting 1). The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α) (adapted from Weber et al. (2020a)).*

			Bucher				MAIC			
EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.385	0.919	-0.090	0.263	0.383	0.918	-0.090	0.265
TTE	similar	moderate	0.206	0.945	0.032	0.166	0.203	0.946	0.032	0.166
TTE	similar	low	0.066	0.943	0.039	0.136	0.066	0.944	0.039	0.136
TTE	similar	no	0.050	0.950	0.005	0.122	0.050	0.950	0.005	0.122
TTE	diff	high	0.430	0.913	-0.091	0.248	0.338	0.870	-0.060	0.397
TTE	diff	moderate	0.227	0.941	0.031	0.160	0.168	0.935	0.039	0.230
TTE	diff	low	0.066	0.945	0.036	0.129	0.067	0.942	0.041	0.172
TTE	diff	no	0.052	0.948	0.007	0.118	0.053	0.947	0.012	0.152
binary	similar	high	0.487	0.912	-0.228	0.438	0.485	0.913	-0.228	0.439
binary	similar	moderate	0.202	0.948	-0.004	0.212	0.201	0.948	-0.004	0.212
binary	similar	low	0.065	0.951	-0.001	0.154	0.066	0.951	-0.001	0.154
binary	similar	no	0.052	0.948	0.000	0.142	0.052	0.948	0.000	0.142
binary	diff	high	0.475	0.907	-0.241	0.457	0.346	0.881	-0.303	0.655
binary	diff	moderate	0.193	0.947	-0.001	0.214	0.140	0.940	-0.004	0.306
binary	diff	low	0.066	0.954	0.000	0.153	0.067	0.946	-0.002	0.207
binary	diff	no	0.050	0.950	0.003	0.142	0.050	0.950	0.002	0.183

Table 15: *Simulation results for scenario II (setting 1). The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC (adapted from Weber et al. (2020a)).*

			Bucher				MAIC - all confounders				MAIC - only effect modifiers			
EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.061	0.814	0.519	0.834	0.062	0.813	0.520	0.837	0.061	0.814	0.519	0.834
TTE	similar	moderate	0.049	0.846	0.230	0.341	0.047	0.846	0.230	0.341	0.049	0.846	0.230	0.341
TTE	similar	low	0.053	0.932	0.076	0.191	0.054	0.932	0.076	0.191	0.053	0.932	0.076	0.191
TTE	similar	no	0.054	0.946	-0.007	0.155	0.053	0.947	-0.007	0.155	0.054	0.946	-0.007	0.155
TTE	diff	high	0.069	0.841	0.584	1.037	0.123	0.816	0.750	1.478	0.081	0.830	0.596	1.071
TTE	diff	moderate	0.052	0.867	0.248	0.387	0.065	0.878	0.284	0.508	0.055	0.865	0.248	0.388
TTE	diff	low	0.056	0.934	0.077	0.211	0.060	0.931	0.088	0.271	0.056	0.934	0.077	0.211
TTE	diff	no	0.052	0.948	-0.008	0.169	0.061	0.939	-0.002	0.213	0.053	0.947	-0.008	0.169
binary	similar	high	0.048	0.876	0.463	0.757	0.288	0.944	-0.081	0.413	0.287	0.944	-0.080	0.412
binary	similar	moderate	0.052	0.897	0.214	0.402	0.210	0.941	-0.054	0.255	0.210	0.942	-0.054	0.255
binary	similar	low	0.054	0.946	0.038	0.252	0.160	0.902	-0.127	0.237	0.160	0.902	-0.127	0.237
binary	similar	no	0.049	0.951	-0.018	0.226	0.123	0.877	-0.150	0.241	0.123	0.877	-0.150	0.241
binary	diff	high	0.049	0.901	0.448	1.002	0.222	0.918	-0.119	0.602	0.228	0.944	-0.048	0.443
binary	diff	moderate	0.053	0.918	0.209	0.455	0.154	0.932	-0.059	0.336	0.176	0.945	-0.039	0.265
binary	diff	low	0.048	0.952	0.038	0.281	0.122	0.917	-0.127	0.271	0.138	0.914	-0.115	0.234
binary	diff	no	0.053	0.947	-0.019	0.255	0.105	0.895	-0.148	0.267	0.116	0.884	-0.141	0.239

Table 16: *Simulation results for scenario III (setting 1). The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC (adapted from Weber et al. (2020a)).*

			Bucher				MAIC - all confounders				MAIC - only effect modifiers			
EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.059	0.811	0.496	0.760	0.060	0.810	0.497	0.765	0.059	0.811	0.496	0.760
TTE	similar	moderate	0.053	0.837	0.228	0.337	0.052	0.837	0.228	0.337	0.053	0.837	0.228	0.337
TTE	similar	low	0.049	0.934	0.074	0.185	0.050	0.934	0.074	0.185	0.049	0.934	0.074	0.185
TTE	similar	no	0.053	0.947	-0.009	0.152	0.053	0.947	-0.009	0.152	0.053	0.947	-0.009	0.152
TTE	diff	high	0.069	0.841	0.555	0.961	0.112	0.821	0.697	1.347	0.073	0.836	0.559	0.969
TTE	diff	moderate	0.055	0.863	0.237	0.376	0.072	0.878	0.262	0.476	0.056	0.864	0.237	0.376
TTE	diff	low	0.051	0.935	0.08	0.207	0.053	0.933	0.090	0.260	0.051	0.935	0.080	0.208
TTE	diff	no	0.054	0.946	-0.005	0.167	0.058	0.942	0.003	0.208	0.054	0.946	-0.005	0.167
binary	similar	high	0.048	0.875	0.470	0.833	0.296	0.944	-0.091	0.416	0.295	0.943	-0.091	0.416
binary	similar	moderate	0.049	0.905	0.208	0.401	0.220	0.941	-0.065	0.258	0.220	0.940	-0.065	0.258
binary	similar	low	0.047	0.947	0.046	0.252	0.147	0.906	-0.125	0.234	0.148	0.906	-0.125	0.233
binary	similar	no	0.047	0.953	-0.020	0.227	0.122	0.878	-0.152	0.243	0.123	0.877	-0.152	0.243
binary	diff	high	0.050	0.895	0.481	1.055	0.228	0.924	-0.123	0.591	0.238	0.947	-0.051	0.437
binary	diff	moderate	0.051	0.914	0.217	0.462	0.160	0.940	-0.062	0.329	0.178	0.946	-0.037	0.263
binary	diff	low	0.051	0.947	0.039	0.285	0.128	0.913	-0.128	0.270	0.141	0.910	-0.117	0.235
binary	diff	no	0.050	0.950	-0.019	0.255	0.111	0.889	-0.152	0.269	0.116	0.884	-0.142	0.241

Table 17: *Simulation results for scenario IV. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC (adapted from Weber et al. (2020a)).*

			Bucher				MAIC - all confounders				MAIC - only effect modifiers			
EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.627	0.801	-0.209	0.286	0.625	0.802	-0.209	0.287	0.627	0.801	-0.209	0.286
TTE	similar	moderate	0.655	0.786	-0.157	0.205	0.653	0.786	-0.156	0.205	0.655	0.786	-0.157	0.205
TTE	similar	low	0.551	0.664	-0.182	0.214	0.549	0.665	-0.182	0.214	0.551	0.664	-0.182	0.214
TTE	similar	no	0.489	0.511	-0.225	0.249	0.488	0.512	-0.225	0.249	0.489	0.511	-0.225	0.249
TTE	diff	high	0.685	0.777	-0.214	0.278	0.505	0.770	-0.191	0.357	0.635	0.779	-0.212	0.289
TTE	diff	moderate	0.693	0.764	-0.157	0.202	0.481	0.825	-0.152	0.234	0.646	0.783	-0.156	0.207
TTE	diff	low	0.579	0.646	-0.184	0.214	0.414	0.750	-0.179	0.227	0.542	0.665	-0.183	0.215
TTE	diff	no	0.510	0.490	-0.227	0.250	0.380	0.620	-0.224	0.258	0.483	0.517	-0.227	0.251
binary	similar	high	0.712	0.766	-0.491	0.628	0.712	0.768	-0.491	0.629	0.712	0.766	-0.491	0.628
binary	similar	moderate	0.524	0.808	-0.262	0.361	0.524	0.809	-0.262	0.361	0.524	0.808	-0.262	0.361
binary	similar	low	0.362	0.740	-0.259	0.327	0.361	0.739	-0.259	0.327	0.362	0.740	-0.259	0.327
binary	similar	no	0.278	0.722	-0.258	0.320	0.278	0.722	-0.258	0.320	0.278	0.722	-0.258	0.320
binary	diff	high	0.709	0.762	-0.508	0.646	0.505	0.787	-0.563	0.816	0.657	0.776	-0.518	0.675
binary	diff	moderate	0.523	0.812	-0.265	0.362	0.353	0.857	-0.271	0.425	0.480	0.820	-0.266	0.372
binary	diff	low	0.371	0.742	-0.259	0.325	0.271	0.806	-0.259	0.353	0.347	0.756	-0.258	0.329
binary	diff	no	0.284	0.716	-0.260	0.321	0.223	0.777	-0.261	0.343	0.270	0.730	-0.261	0.325

Table 18: *Simulation results for scenario V (setting 1). The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC (adapted from Weber et al. (2020a)).*

			Bucher				MAIC - all confounders				MAIC - only effect modifiers			
EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.655	0.794	-0.207	0.277	0.650	0.791	-0.208	0.277	0.655	0.794	-0.207	0.277
TTE	similar	moderate	0.677	0.778	-0.156	0.202	0.674	0.777	-0.156	0.203	0.677	0.778	-0.156	0.202
TTE	similar	low	0.555	0.655	-0.181	0.212	0.555	0.656	-0.181	0.212	0.555	0.655	-0.181	0.212
TTE	similar	no	0.484	0.516	-0.224	0.248	0.483	0.517	-0.224	0.248	0.484	0.516	-0.224	0.248
TTE	diff	high	0.706	0.772	-0.208	0.27	0.531	0.782	-0.190	0.334	0.647	0.778	-0.206	0.280
TTE	diff	moderate	0.712	0.766	-0.156	0.198	0.500	0.826	-0.150	0.225	0.662	0.783	-0.155	0.202
TTE	diff	low	0.589	0.634	-0.183	0.212	0.433	0.734	-0.179	0.224	0.557	0.656	-0.183	0.214
TTE	diff	no	0.514	0.486	-0.225	0.248	0.388	0.612	-0.223	0.255	0.483	0.517	-0.225	0.249
binary	similar	high	0.722	0.762	-0.498	0.632	0.722	0.762	-0.498	0.632	0.712	0.766	-0.491	0.628
binary	similar	moderate	0.519	0.812	-0.265	0.360	0.519	0.812	-0.265	0.360	0.524	0.808	-0.262	0.361
binary	similar	low	0.367	0.742	-0.260	0.326	0.367	0.742	-0.260	0.326	0.362	0.740	-0.259	0.327
binary	similar	no	0.290	0.710	-0.263	0.325	0.290	0.710	-0.263	0.325	0.278	0.722	-0.258	0.320
binary	diff	high	0.707	0.768	-0.497	0.637	0.644	0.782	-0.502	0.661	0.657	0.776	-0.518	0.675
binary	diff	moderate	0.521	0.809	-0.263	0.361	0.485	0.821	-0.264	0.371	0.480	0.820	-0.266	0.372
binary	diff	low	0.370	0.734	-0.261	0.327	0.348	0.750	-0.261	0.331	0.347	0.756	-0.258	0.329
binary	diff	no	0.285	0.715	-0.262	0.322	0.270	0.730	-0.262	0.325	0.270	0.730	-0.261	0.325

Effective Sample Size

When population distributions are similar, the ESS for the MAIC procedure is equal to the actual sample size which is considered in the method of Bucher. Differences in patient characteristics between trials and when all confounders are considered as matching variables in MAIC cause the ESS to reach only half of the actual sample size. Whereas the ESS is solely reduced by 15% to 20% when MAIC is adjusted for relevant effect modifiers only.

Table 19: *Mean and standard deviation (sd) of Effective Sample Size (ESS) for scenario I considering a time-to-event (TTE) as well as binary endpoint (adapted from Weber et al. (2020a)).*

Endpoint	Population	Effect	Sample Size	mean ESS	sd ESS
TTE	similar	high	94	92.883	1.571
TTE	similar	moderate	396	394.731	1.783
TTE	similar	low	1044	1042.307	2.376
TTE	similar	no	1578	1575.991	2.829
TTE	different	high	94	38.927	6.735
TTE	different	moderate	396	161.756	15.513
TTE	different	low	1044	423.204	28.605
TTE	different	no	1578	639.000	37.511
binary	similar	high	192	190.883	1.574
binary	similar	moderate	648	646.692	1.848
binary	similar	low	1600	1598.248	2.438
binary	similar	no	2176	2174.037	2.799
binary	different	high	192	79.788	10.188
binary	different	moderate	648	264.013	20.993
binary	different	low	1600	646.443	37.148
binary	different	no	2176	878.080	44.440

Influence of Planned Power of Direct Comparisons

For the independent trials AB and CB, the power used for calculating the sample size is varied to investigate the influence on the power of the indirect comparison. Trials are powered

at 80%, 90%, 95%, and 99% including combinations of this values. In simulations it was observed, that the power of the indirect comparison increases with increasing power of the head-to-head comparisons. Higher treatment effects in the indirect comparison gain more power by increasing the power in head-to-head trials (see Figure 5). However, even in case both trials are powered at 99%, the power of the indirect comparison is less than 60% when all method assumptions are met and the treatment effect is assumed to be high (see Appendix A.2 Table 40). Fixing the power in the AgD trial (CB) to 80% and increasing power of the IPD trial AB also demonstrates that there is an increase in power of the indirect comparison, but it is still clearly below the aspired 80%. The corresponding results are plotted in Figure 6 and detailed power values are included in Table 41. The type I error rate remains at around 5% for all power scenarios. The results for bias of the effect estimate, RMSE, and the coverage of the 95% CI are comparable for all power scenarios; these measures are already discussed in the paragraphs above.

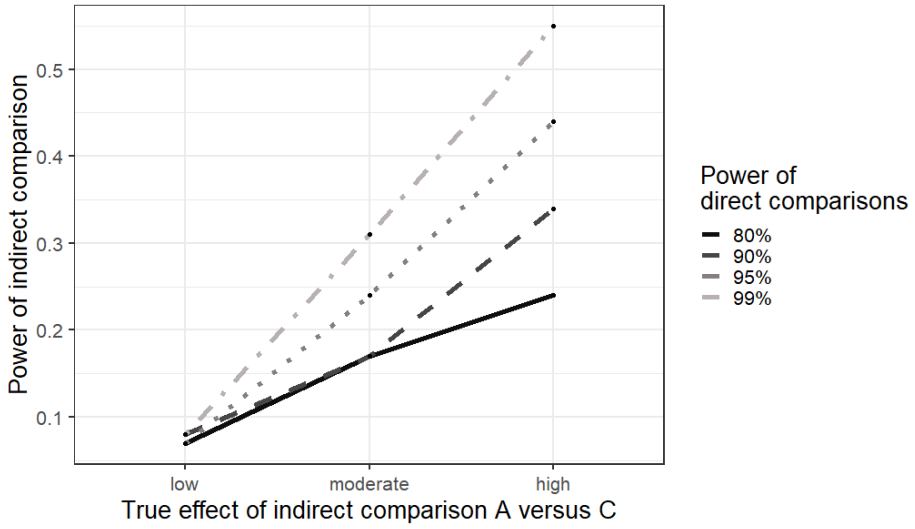


Figure 5: *The plot shows the power depending on the true effect of the indirect comparison (A versus C) for balanced groups and different power scenarios for the direct trials (A versus B and C versus B).*

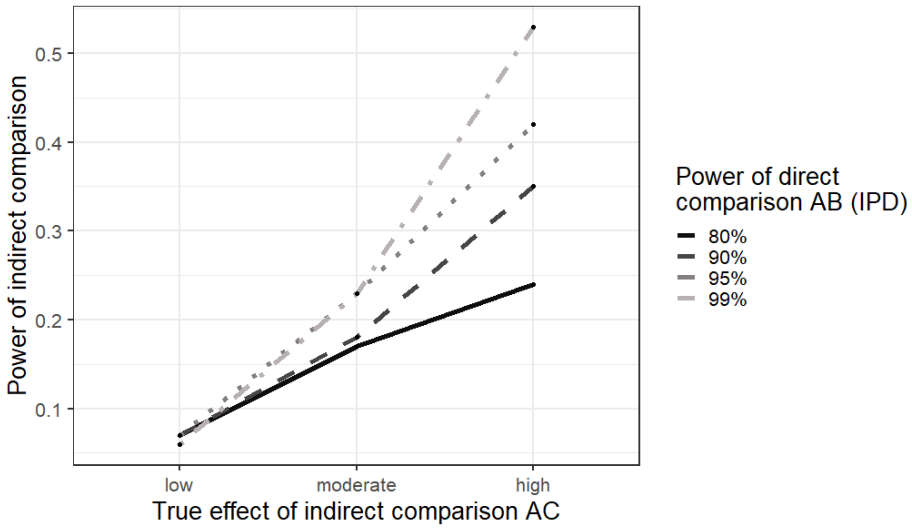


Figure 6: *The plot shows the power depending on the true effect of the indirect comparison A versus C (AC) for balanced groups and different power scenarios for the direct trials (A versus B and C versus B). Fixing the power in the AgD trial (CB) to 80% (adapted from Weber et al. (2020a)).*

Table 20: *Summary of the method comparison. An overview of situations where the method of Bucher outperforms the MAIC procedure and vice versa with regard to the considered simulation scenarios and a time-to-event (TTE) endpoint is presented. 1 signifies that both methods perform equally, 2 the method of Bucher works best, and 3 MAIC outperforms Bucher (adapted from Weber et al. (2020a)).*

Theoretical situation					Analysis deviations	Results					But, note
Scenario	Endpoint	Population	Confounder	Interaction	from simulated situation	Power	Cov	Type I error	Bias/RMSE	Sample Size	
I	TTE	similar	similar	no		1	1	1	1	1	
	TTE	different	similar	no		2	2	1	2	2	maximal power < 40%
	TTE	similar	similar	no	CB not adjusted for confounders	1	1	1	1	1	
	TTE	different	similar	no		2	2	1	2	2	
II	TTE	similar	similar	yes - similar		1	1	1	1	1	
	TTE	different	similar	yes - similar		3	3	2	2	2	
	TTE	similar	similar	yes - similar	CB not adjusted for confounders	1	1	1	1	1	maximal power < 30%
	TTE	different	similar	yes - similar		3	3	2	2	2	
	TTE	similar	similar	yes - similar	Effect modification not adjusted	1	1	1	1	1	
	TTE	different	similar	yes - similar		2	2	2	2	2	
III	TTE	similar	different	yes - similar		1	1	1	1	1	
	TTE	different	different	yes - similar		3	1	2	1	2	
	TTE	similar	different	yes - similar	CB not adjusted for confounders	1	1	1	1	1	maximal power < 15%
	TTE	different	different	yes - similar		3	3	3	2	2	
	TTE	similar	different	yes - similar	Effect modification not adjusted	1	1	1	1	1	
	TTE	different	different	yes - similar		2	2	2	2	2	
IV	TTE	similar	similar	yes - different		1	1	1	1	1	
	TTE	different	similar	yes - different		2	3	3	1	2	
	TTE	similar	similar	yes - different	CB not adjusted for confounders	1	1	1	1	1	maximal power < 80%
	TTE	different	similar	yes - different		2	3	3	1	2	highly inflated type I error rate
	TTE	similar	similar	yes - different	Effect modification not adjusted	1	1	1	1	1	
	TTE	different	similar	yes - different		2	3	3	1	2	
V	TTE	similar	different	yes - different		1	1	1	1	1	
	TTE	different	different	yes - different		2	3	3	1	2	
	TTE	similar	different	yes - different	CB not adjusted for confounders	1	1	1	1	1	maximal power < 80%
	TTE	different	different	yes - different		2	3	3	1	2	highly inflated type I error rate
	TTE	similar	different	yes - different	Effect modification not adjusted	1	1	1	1	1	
	TTE	different	different	yes - different		2	3	3	1	2	

Table 21: *Summary of the method comparison. An overview of situations where the method of Bucher outperforms the MAIC procedure and vice versa with regard to the considered simulation scenarios for a binary endpoint is presented. 1 signifies that both methods perform equally, 2 the method of Bucher works best, and 3 MAIC outperforms Bucher (adapted from Weber et al. (2020a)).*

Theoretical situation					Analysis deviations	Results					But, note
Scenario	Endpoint	Population	Confounder	Interaction	from simulated situation	Power	Cov	Type I error	Bias/RMSE	Sample Size	
I	Binary	similar	similar	no		1	1	1	1	1	
	Binary	different	similar	no		2	2	1	2	2	maximal power < 50%
	Binary	similar	similar	no	CB not adjusted for confounders	1	1	1	1	1	
	Binary	different	similar	no		2	2	1	2	2	
II	Binary	similar	similar	yes - similar		3	3	2	3	1	
	Binary	different	similar	yes - similar		3	3	2	3	2	
	Binary	similar	similar	yes - similar	CB not adjusted for confounders	3	3	3	3	1	maximal power < 35%
	Binary	different	similar	yes - similar		3	3	3	3	2	
	Binary	similar	similar	yes - similar	Effect modification not adjusted	1	1	1	1	1	
	Binary	different	similar	yes - similar		2	2	2	2	2	
III	Binary	similar	different	yes - similar		3	3	2	3	1	
	Binary	different	different	yes - similar		3	3	2	3	2	
	Binary	similar	different	yes - similar	CB not adjusted for confounders	3	3	3	3	1	maximal power < 30%
	Binary	different	different	yes - similar		3	3	3	3	2	
	Binary	similar	different	yes - similar	Effect modification not adjusted	1	1	1	1	1	
	Binary	different	different	yes - similar		2	1	3	1	2	
IV	Binary	similar	similar	yes - different		1	1	1	1	1	
	Binary	different	similar	yes - different		2	3	3	1	2	
	Binary	similar	similar	yes - different	CB not adjusted for confounders	1	1	1	1	1	maximal power < 80%
	Binary	different	similar	yes - different		2	3	3	2	2	highly inflated type I error rate
	Binary	similar	similar	yes - different	Effect modification not adjusted	1	1	1	1	1	
	Binary	different	similar	yes - different		2	3	3	2	2	
V	Binary	similar	different	yes - different		1	1	1	1	1	
	Binary	different	different	yes - different		2	3	3	2	2	
	Binary	similar	different	yes - different	CB not adjusted for confounders	1	1	1	1	1	maximal power < 80%
	Binary	different	different	yes - different		2	3	3	2	2	highly inflated type I error rate
	Binary	similar	different	yes - different	Effect modification not adjusted	1	1	1	1	1	
	Binary	different	different	yes - different		2	3	3	2	2	

The results of the method comparison show unsatisfactory values for power as well as bias and RMSE. The natural progression is including more studies, and therefore increasing the sample size, in the indirect comparison. Methods for this situation are inspected in Section 3.4.

Comment: Parts of the following Section 3.4 are already included in the submitted manuscript Weber et al. (2020b). The manuscript has been written by myself but may contain comments and corrections from the co-authors.

3.4 Synthesis of Evidence - Multiple Studies

A simulation study evaluates the methods described in Section 2.1 for including multiple studies within the adjusted indirect comparison by MAIC for TTE endpoints. It aims to investigate the methods for practically relevant scenarios in a medical context. The method of Bucher is additionally applied in all the considered simulation settings for comparison. Statistical performance measures of the methods are assessed and compared such as the bias of the estimated therapy effects, RMSE, power, coverage of the 95% CI, and type I error rate. Due to the high computation time and memory capacity required, the number of simulation runs is limited to $n_{sim} = 2,000$ for each scenario. To evaluate the error introduced by the simulation, the Monte Carlo standard errors of the bias are calculated. The method of Bucher assumes that there are no differences between trials AB and CB with respect to effect modifiers. Nevertheless, this assumption needs to be evaluated for each situation in practice. Simulations were done using R version 3.5.1 (R Core Team, 2017) using the packages meta (Balduzzi et al., 2019) and survival (Therneau, 2015).

3.4.1 Data Simulation

The data generating process is similar to Section 3.3.1. The key points are described below, and details about the formulas can be found in Section 3.3.1.

The data for trials AB and CB contains one continuous and three binary variables (Table 10). A covariance matrix is specified to include a random error term (Table 10) or rather correlations between the variables. The true treatment effects of treatment comparisons AB, CB, and AC (defined as the difference between AB and CB) are expressed on the log HR scale.

For generating the TTE outcomes, the event time and the censoring time are sampled from a Weibull distribution ($\lambda_{event} = 0.0002$, $\nu_{event} = 1.8$, $\lambda_{censoring} = 0.00012$, $\nu_{censoring} = 2$, max. time=100). Utilizing those times and the baseline characteristics a Cox proportional hazard model is applied to generate the event status. The values for log HRs of the covariates in the models are given in Table 22. Some of the simulation scenarios comprise an interaction term between a binary baseline variable and treatment. This baseline variable is then called effect modifier. The corresponding log HRs for the interaction term are also specified in Table 22, where the settings of a positive and a negative interaction term is considered. The simulation study is limited to the described data set because its purpose is not to examine the influence of patient characteristics itself. The true effect size of indirect comparison AC is simulated as *high, moderate, low,* and *no effect*. This classification of effect sizes follows the recommendations in Skipka et al. (2015) and refers to the benefit assessment of new drugs. The treatment effect in terms of log HR is assumed to follow a normal distribution to introduce some variability in treatment effects between trials comparing the same treatments. The mean values for the treatment effects are given in Table 23, which are combined with a small ($\sigma = 0.2$) or a large ($\sigma = 0.4$) variance value. The desired power of 80%, 5% type I error rate (two-sided), and the mean log HRs for the head-to-head treatment effect (see Table 23) are assumed for the sample size calculations, which are based on established formulas (Schoenfeld, 1983). The sample sizes for trial AB and CB according to the different treatment effects of the indirect comparison AC are given in Table 24. The underlying group allocation ratio is 1:1 for the treatment and control group.

Table 22: *Regression coefficients in terms of log hazard ratios for Cox-regression models considered for simulation of outcomes (adapted from Weber et al. (2020b)).*

Variable	log HR
continuous	-0.0051
binary 1 ($x_{b1} = 1$)	-0.4
binary 2 ($x_{b2} = 1$)	1.122
binary 3 ($x_{b3} = 1$)	-0.2
interaction:	
treatment and binary 1 (=1)	$\pm$0.02

Table 23: *Mean values of a normal distribution for log hazard ratios for time-to-event endpoints for different effect classes. Where the AC column is the difference between AB and CB (adapted from Weber et al. (2020b)).*

	AC	AB	CB
high	-0.69	-0.91	-0.22
moderate	-0.22	-0.44	-0.22
low	-0.05	-0.27	-0.22
no	0	-0.22	-0.22

Table 24: *Sample size in individual trials comparing treatment A and B (AB), comparing treatment C and B (CB), depending on the treatment effects in the indirect comparison AC. The sample sizes are based on sample size calculations (adapted from Weber et al. (2020b)).*

Treatment effect	Sample Size	
AC	AB	CB
high	94	1622
moderate	388	1622
low	1078	1622
no	1622	1622

3.4.2 Evaluation Measures

The performance of the approaches for the inclusion of multiple studies are evaluated by the bias of the effect estimate (the difference to the true treatment effect as given in Table 23, see Formula 3.1), the RMSE (see Formula 3.2), the power, the type I error rate, and the two-sided 95% CI coverage. The calculation of CIs corresponding to the effect estimate in the regression model relies on a normal approximation. A detailed explanation of the evaluation measures and its calculation can be found in Section 3.3.2. Due to a lower number of simulation runs, the Monte Carlo standard errors of the bias are calculated for each considered scenario

$$RMSE = \sqrt{\frac{1}{n_{sim}(n_{sim}-1)} \sum_{i=1}^{n_{sim}} (\hat{\delta}_{AC,i} - \overline{\delta}_{AC})^2} \tag{3.3}$$

where n_{sim} the number of simulation runs and $\hat{\delta}_{AC,i}$ denotes the estimate of the true treatment effect in simulation run i and $\overline{\delta}_{AC,i}$ the mean estimate over all simulation runs (Morris et al., 2019).

3.4.3 Simulation Scenarios

The methods for inclusion of multiple studies are applied to 2, 4, and 10 trials for the direct comparison. If numerous studies are available for both trials AB and CB, the number of trials is assumed to be equal. The matching variables, which are aimed to be balanced, are the variables *binary1*, *binary2*, and the *continuous* one. This implies the misspecification of the matching model since MAIC should only balance for essential effect modifiers. All baseline variables are considered as covariates in the regression models for estimating the treatment effects. Some of the approaches involve a synthesis of treatment effects by meta-analysis which is implemented using fixed and random effects models. Furthermore, the influence of the magnitude of the variance ($\sigma = 0.2$ versus $\sigma = 0.4$) in the treatment effects is evaluated.

To investigate the advantage of combining multiple studies within the MAIC procedure, the simulation scenarios include situations where assumptions for the methods of indirect comparisons are violated. The simulation scenarios are described in Table 25, all scenarios are evaluated for all the described approaches.

Table 25: *Simulation scenarios (adapted from Weber et al. (2020b)).*

Scenario	Population		Interaction	
	similar	different	yes	no
I	x			x
II		x		x
III	x		x	
IV		x	x	

3.4.4 Simulation Results

This results section is split according to the classification of approaches in Section 2.2.4, within each of the three paragraphs the different evaluation measures are discussed. Detailed

results for the described evaluation measures of the different scenarios are documented in Appendix Section A.3 (Tables 42 to 53). Over all scenarios, the ESS in MAIC is less than half of the original sample size. For multiple IPD (see Table 26 at the end of this section) the results on ESS are given for illustration; all other scenarios lead to comparable ESS. Simulations showed similar results for positive and negative log HRs for the interaction term; therefore, this section is limited to the results for the negative log HR.

If a meta-analysis is involved in an approach, the results described and discussed in detail are based on the random-effects model. For selected scenarios, the fixed-effects model is implemented additionally. In strategies where the meta-analysis is conducted to combine the treatment effects of several indirect comparisons differences between those two models are observed. Fixed-effects meta-analysis results in slightly higher power (Appendix Tables 54 to 59). However, the use and interpretation of a fixed-effects model for the combination of indirect evidence should be done carefully, and its suitability needs to be checked for each situation individually.

Results for all simulation scenarios are described and discussed for a variance of $\sigma = 0.2$ in the treatment effect distribution. Moreover, scenarios I and II are evaluated for a variance of $\sigma = 0.4$ in the treatment effect distribution (see Appendix Table 60 to 65). It is observed that increasing the variance component when simulating the treatment effects for individual trials leads to higher bias and RMSE in the effect estimate of the indirect comparison. The power is slightly lower, and type I error rates are comparable to those of the lower variance scenarios.

The Monte Carlo standard errors for the bias of the indirect effect estimates are all smaller than 0.02, in most cases even lower than 0.01. Detailed results are included in the Tables in Appendix Section A.3.

In the following the results apply for the method of Bucher and MAIC; otherwise, differences are further described.

Figure 7, 8, 9, and 10 assume a high treatment effect and demonstrated the power, type I error rate, bias, and RMSE of the scenarios and methods depending on the number of trials.

One IPD trial (AC) and multiple AgD trials (CB)

Either the AgD trials are combined by a meta-analysis (approach A.1), or all indirect comparisons are conducted (approach A.2).

Power

In scenarios I, II, and III it is apparent that approach A.1 results in higher power values compared to approach A.2. For high treatment effects, the desired power of at least 80% is achieved for approach A.1 when patient characteristics are similar. Differences in baseline characteristics between trials lead to a loss in power for MAIC under approach A.1, whereas approach A.2 leads to the same results for both methods which are close to the results for similar cohorts. Smaller differences between approach A.1 and A.2 are observed in scenario IV. Applying MAIC under consideration of approach A.2 when effect modification is present and patient characteristics differ leads to higher power values compared to approach A.1. Increasing the number of studies leads to a power increase which is observed to be stronger for smaller treatment effects (Appendix Table 42 and 48).

Type I error

Type I error rate is about 6% for approach A.2 and around 10% for approach A.1 under the conditions of scenarios I and II (Appendix Table 42). In scenario III where effect modification is present an inflated type I error rate is observed for both methods which increases by the number of studies (Appendix Table 48). When additionally patient cohorts differ, a substantially increased type I error rate is documented even for a small number of studies (Appendix Table 48).

Coverage

Scenarios I and II have a coverage of around 95% for approach A.2, whereas approach A.1 results in values less than 90% (Appendix Table 42). Approach A.1 leads to lower coverage compared to approach A.2 in scenarios III and IV, values are observed to be less than 95% for both approaches, but some scenarios are close to 95% (Appendix Table 48).

Bias/RMSE

Bias and RMSE increase for higher treatment effects. In scenario I bias and the RMSE are slightly higher for approach A.2. compared to A.1. The same results are observed for scenarios I and II for the method of Bucher. Applying MAIC the two approaches A.1 and A.2 give equal results. Bias and RMSE are higher for scenarios that include effect modification (scenario III and IV). In scenario III RMSE is higher for approach A.2, whereas the bias is lower or comparable to approach A.1. Scenario IV results in less biased effect estimates and lower RMSE for approach A.2 when treatment effects are high, but the reversed results are observed for smaller treatment effects (Appendix Table 43 and 49).

Multiple IPD trials (AC) and one AgD trial (CB)

Either the IPD trials are pooled before the indirect comparison (approach B.1), or all indirect comparisons are conducted and effect estimates are pooled afterward (approach B.2).

Power

It is observed that the power increases by the number of trials. The higher the treatment effect, the fewer studies are needed to reach reasonable power regions (above 80%). Apparently, for the defined low treatment effect, even a high number of IPD trials cannot assure the desired power. Scenarios I and II give the same results for the method of Bucher under both approaches (Appendix Table 44). The approaches perform equally for MAIC when patient cohorts are similar, but approach B.2 reaches better power values in scenario II. In case of scenario IV approach B.2 results in high power values compared to approach B.1, whereas in scenario III for MAIC the approach B.1 leads to higher power values, especially for a small number of studies (Appendix Table 50).

Type I error

Type I error rate is around 5% for scenarios I and II (Appendix Table 44). Focusing on scenarios III and IV the type I error rate is considerably increased with values between 10% and 26% (Appendix Table 50). For MAIC, approach B.1 leads to lower type I error rates compared to B.2 in scenario IV.

Coverage

In scenario I and II, the coverage is around 95% and it is observed to be independent of the number of studies. For scenarios III the coverage is above 80% and closer to 95% for higher treatment effects. When differences between trials occur (scenario IV) the coverage is decreasing. For MAIC, the approach B.2 performs better in scenario IV when treatment effects are high or medium, for low and no effects approach B.1 leads to higher coverage (Appendix Table 44 and 50).

Bias/RMSE

The bias and the RMSE values rise with increasing treatment effect. However, a higher number of IPD trials leads to more precise estimates, meaning a lower bias and RMSE are observed. Scenarios I and II show equal results for both approaches. The MAIC shows better performance in terms of bias and the RMSE when applying approach B.2 in scenario IV, whereas in scenario II, the performance is comparable (Appendix Table 45 and 51). The presence of an effect modifier leads to higher bias, but the RMSE is not considerably increased.

Multiple IPD trials (AC) and multiple AgD trials (CB)

When various studies cover both head-to-head comparisons, either the AgD trials are combined by a meta-analysis and IPD is pooled (C.1), only IPD trials are pooled to conduct all indirect comparisons (approach C.2), or all possible indirect comparisons are conducted (approach C.3).

Power

Likewise the results above, the simulations show that power increases by the number of trials. All considered approaches (approaches C.1, C.2, and C.3) give comparable results in terms of power in scenarios I and III. In scenario II, approaches C.1, C.2, and C.3 perform similarly for the method of Bucher. Applying MAIC lower power values are observed under approaches C.1 and C.2 compared to C.3 as well as compared to the method of Bucher. Power is lower in scenarios III and IV compared to scenarios I and II. In scenario II and IV MAIC reaches the highest power when applying approach C.3 (Appendix Table 46 and 52).

Type I error

In scenarios I and II, type I error rate is around 5% for all approaches. Scenarios III and IV lead to clearly increased type I error rates which are increasing with the number of studies (Appendix Table 46 and 52). But it is noticeable that under approach C.1 and C.2, MAIC shows lower type I error rates (Appendix Table 52).

Coverage

The coverage is around 95% for scenarios I and II and smaller for scenarios III and IV (Appendix Table 46 and 52). All methods perform equally in the situation of scenario III. In scenario IV, applying MAIC and approach C.3 shows the best coverage for high and medium treatment effects, whereas approaches C.1 and C.2 give better results in terms of coverage for smaller treatment effects (Appendix Table 52).

Bias/RMSE

Bias and RMSE decrease with an increasing number of studies. For the method of Bucher, the three approaches lead to similar results in all scenarios (Appendix Table 47 and 53). For MAIC, scenarios I and III lead to consistent results between the strategies. Scenarios II and IV show higher RMSE values for MAIC compared to Bucher for approaches C.1 and C.2, whereas approach C.3 leads to equal results (Appendix Table 47 and 53).

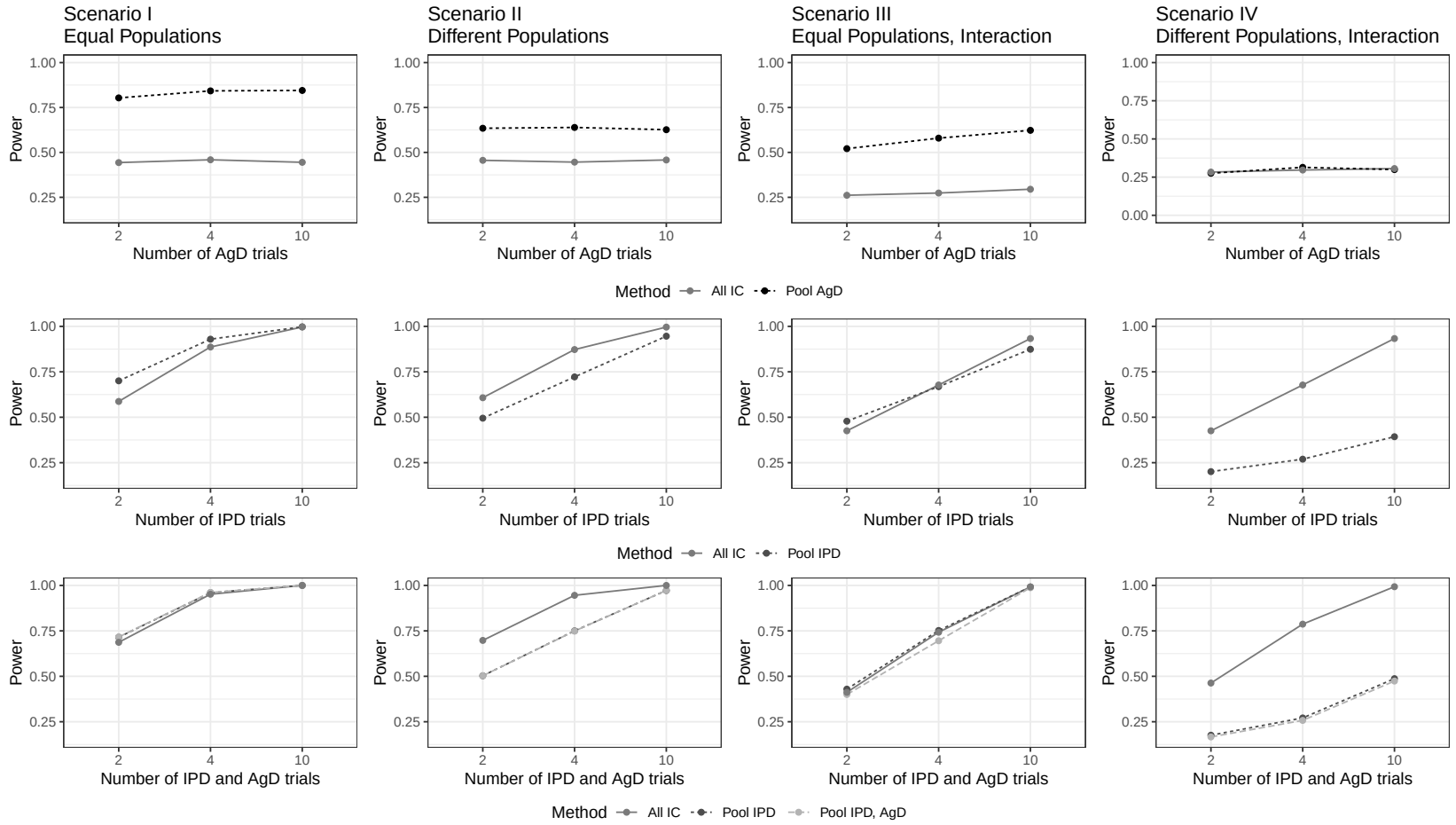


Figure 7: *The plots show the power depending on the number of trials considered for estimating the indirect treatment effect (here for high treatment effect) for all approaches and scenarios when using the MAIC approach for indirect comparisons. The first row shows the results for one IPD trial and multiple AgD trials, the second for one AgD trial and multiple IPD trials, and the third row for multiple IPD and AgD trials (adapted from Weber et al. (2020b)).*

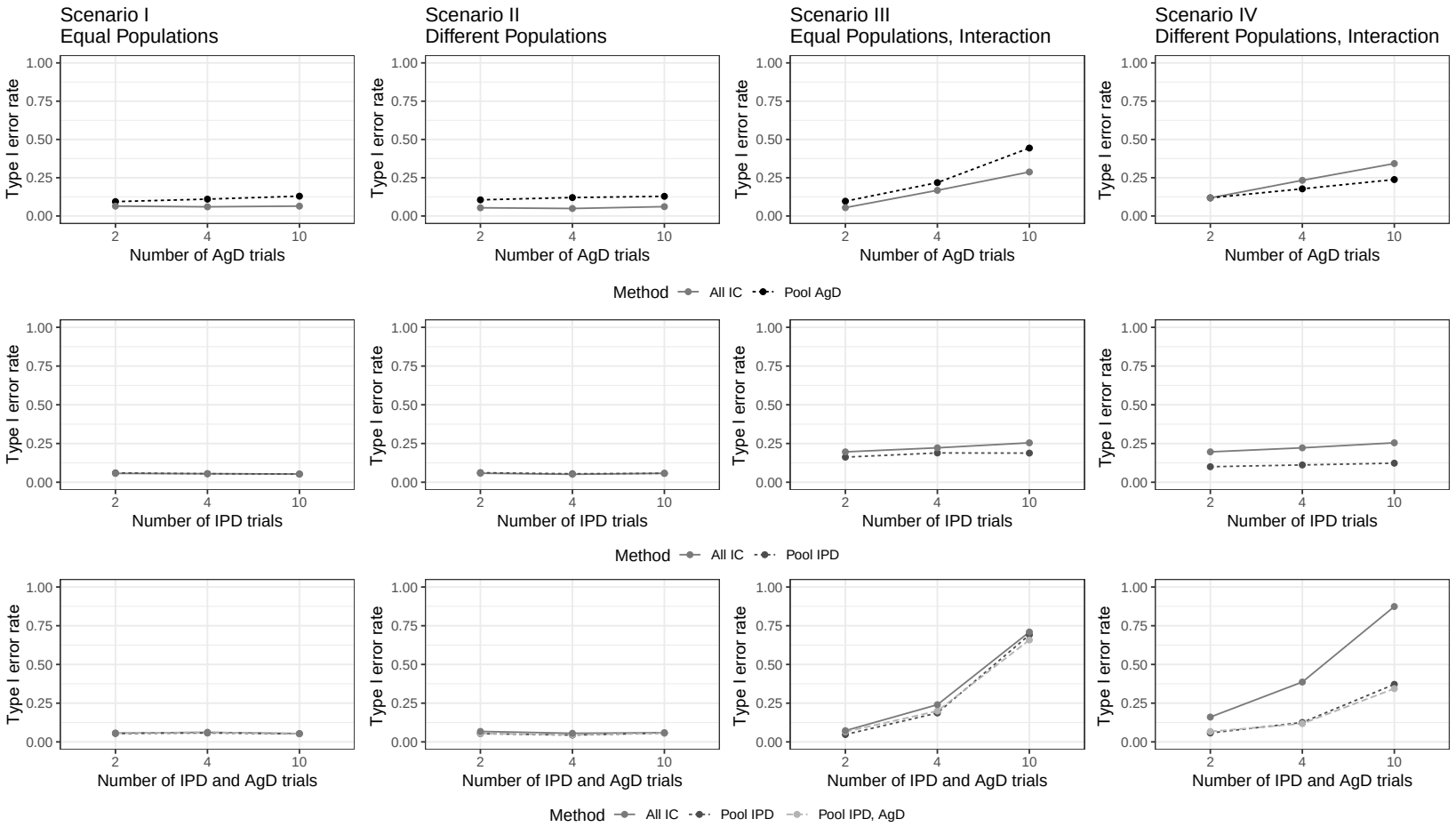


Figure 8: *The plots show the type I error rate depending on the number of trials considered for estimating the indirect treatment effect (here for high treatment effect) for all approaches and scenarios when using the MAIC approach for indirect comparisons. The first row shows the results for one IPD trial and multiple AgD trials, the second for one AgD trial and multiple IPD trials, and the third row for multiple IPD and AgD trials (adapted from Weber et al. (2020b)).*

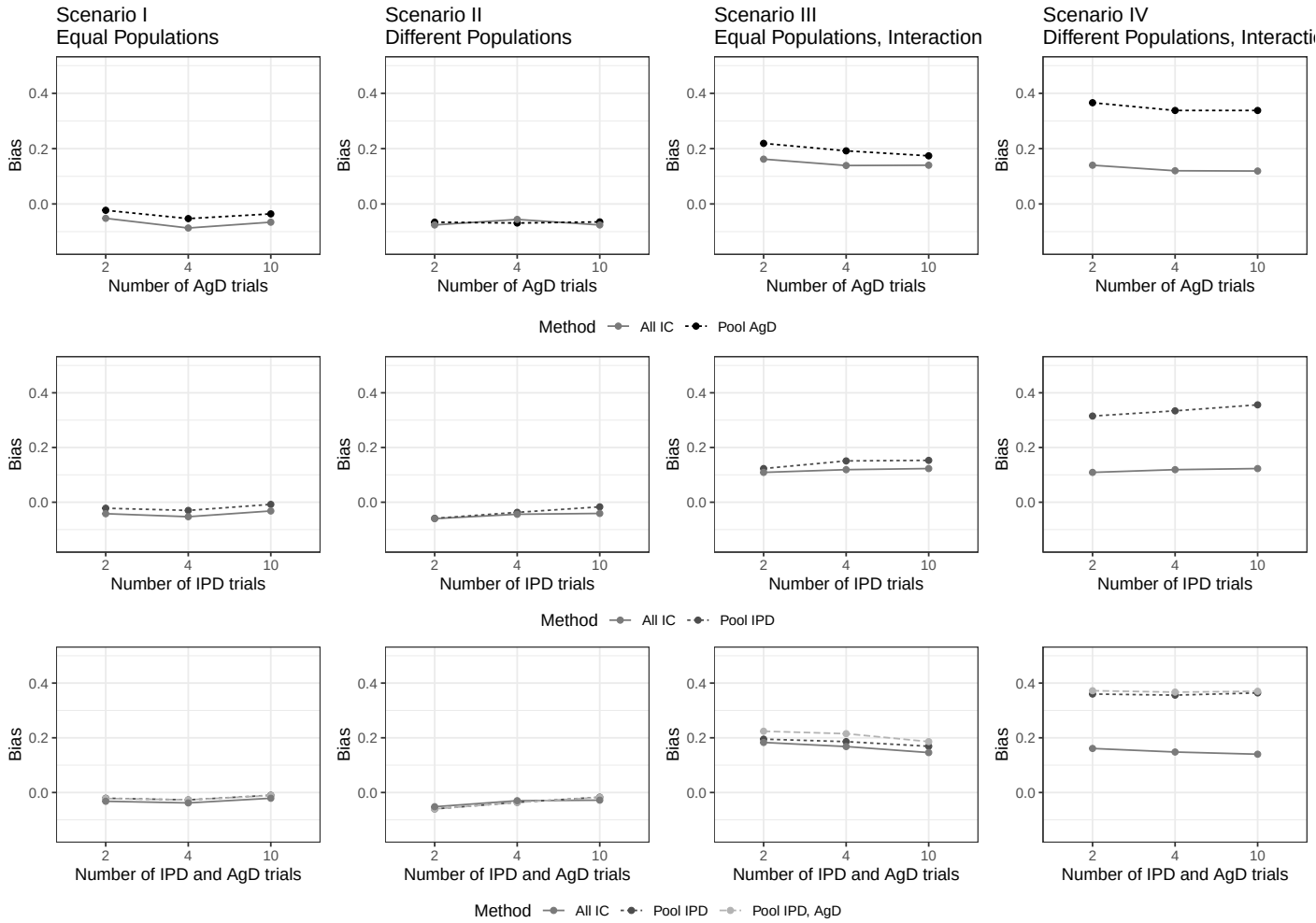


Figure 9: *The plots show the bias depending on the number of trials considered for estimating the indirect treatment effect (here for high treatment effect) for all approaches and scenarios when using the MAIC approach for indirect comparisons. The first row shows the results for one IPD trial and multiple AgD trials, the second for one AgD trial and multiple IPD trials, and the third row for multiple IPD and AgD trials.*

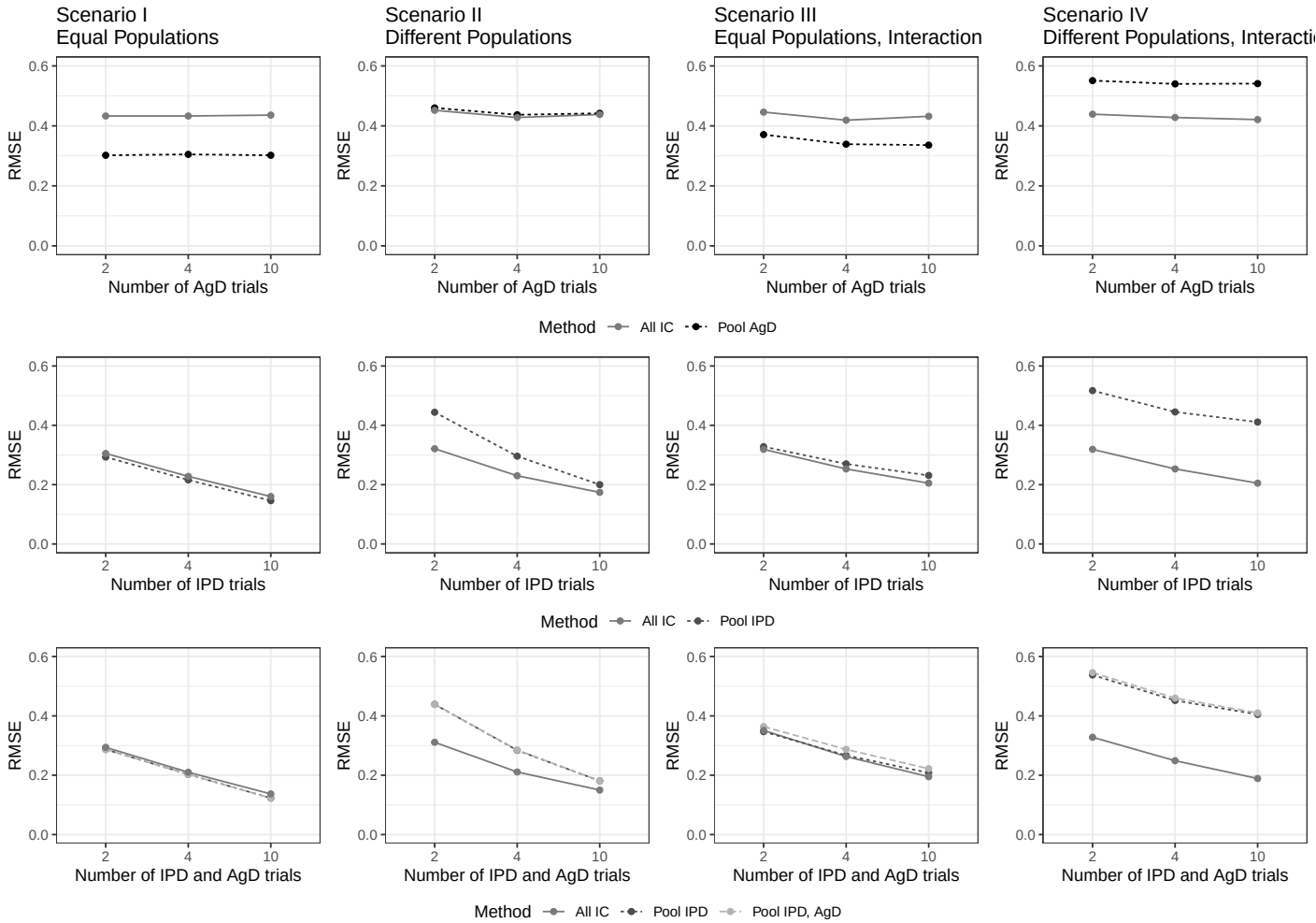


Figure 10: *The plots show the Root mean squared error (RMSE) depending on the number of trials considered for estimating the indirect treatment effect (here for high treatment effect) for all approaches and scenarios when using the MAIC approach for indirect comparisons. The first row shows the results for one IPD trial and multiple AgD trials, the second for one AgD trial and multiple IPD trials, and the third row for multiple IPD and AgD trials.*

Table 26: *Mean and standard deviation (sd) of the Effective Sample Size (ESS) of the MAIC procedure for the scenarios characterized by differences in patient characteristics (Scenarios II and IV). The actual number of patients in the trial is given in the column Sample Size. As an example, the situation including two IPD studies which are pooled for the MAIC is shown here (adapted from Weber et al. (2020b)).*

Effect	Sample Size	mean ESS	sd ESS
high	188	81.47	11.08
medium	776	329.20	30.68
low	2156	908.63	70.65
no	3244	1368.38	100.73

Chapter 4

Discussion

Comment: Parts of the following Chapter 4 have already been published or submitted in one of the manuscripts Weber et al. (2019), Krisam et al. (2020), Weber et al. (2020a), and Weber et al. (2020b). (Parts of) the manuscripts have been written by myself, but may contain comments and corrections from the co-authors.

This chapter discusses the methods developed and applied in this thesis; a summary of its contribution to research is given. Moreover, limitations and directions for future research topics are presented. This is provided separately for the two main topics of this thesis 1. Generation of Evidence (Section 4.1) and 2. Synthesis of Evidence (Section 4.2).

4.1 Generation of Evidence

4.1.1 Discussion and Contributions to Research

In situations where intervention and control group cannot be carried out at the same time; a prospective matched case-control trial may be a good alternative to an RCT achieving comparable study groups. This thesis presents two ways for recalculating the sample size at an interim analysis step, the naïve and the resampling CI method (Section 2.1.1). Simulations showed that the naïve method might severely overestimate the matching rate at the interim analysis. The consequence is a low matching rate at the final analysis and, therefore, low power what makes it hard to detect the treatment effect. The resampling CI method

avoids this overestimation and produces a better estimate for the matching rate. So, a higher matching rate can be achieved at the final analysis and this is related to an ascent in power. Even a greater sample size is needed, it is still reasonable and therefore, applying the resampling CI approach is an efficient alternative to the naïve approach in terms of matching rate and power. Increasing the confidence level showed only a small influence on the sample size in the treatment group and matching rate at the final analysis. Nevertheless, increasing the confidence level leads to a higher sample size and increases the matching rate. A clear advantage over a predefined higher proportion of recruited patients, is the flexibility and adaption to the trial-specific situation. In contrast to the algorithm applied in Charpentier et al. (2001), the resampling CI method is less time-consuming and complex since a matching step is conducted twice instead matching each patient individually. The implementation in practice is straightforward and leads to a manageable additional effort, which makes the method relevant for applicants of prospective matched case-control trials. The application to the KEEP SIMPLEST trial data confirmed the simulation results that the resampling CI method leads to good results for the matching rate in contrast to the naïve approach. At the same time, the balance of patient groups is achieved (Schönenberger et al., 2019).

The time point of interim analysis needs to be fixed beforehand. As a trade-off between matching rate, power, and sample size, the simulations lead to the recommendation to use a proportion of $\frac{1}{2}$ to $\frac{2}{3}$ of the number of patients in the control group at the interim analysis. It seems that obtaining a reasonable estimate for the matching rate depends more on the absolute number of patients at the interim analysis than on the relative number of control patients. Giving a recommendation for an absolute number of patients needed for the interim analysis independent of the trial size is complicated because this number would need to fit trials with very small and also large sample sizes. Consequently, this number would be limited by the small sample size or would not be applicable for small trials. Therefore, the recommendation is based on the proportion of control patients.

Presumed there is a big data set of historical controls the iterative matching procedure (Section 2.1) supplies a framework to determine an appropriate number of matching partners per intervention patient, because the resampling CI method might fail in this situation. The iterative matching procedure aims to potentially increase power due to a bigger sized control

group while still the main reason of applying a matching, the differences in patient cohorts, is addressed. By choosing the tolerance criterion τ, the procedure allows deciding whether it is more important that all treated patients find at least one matching partner ($\tau = 0$) or to increase the number of patients in the control group ($\tau > 0$). One should find a trade-off between the matching rate and the number of matching partners because a decision to one or the other direction may counteract the aim of finding a suitable control group and increase power. The number M of matching partners is highly dependent on the overlap of populations according to the considered matching variables. Nevertheless, this iterative procedure is flexible, user-friendly, and can be implemented, for example, in a two-stage framework which includes a scheduled interim analysis.

4.1.2 Limitations and Directions for Future Research

A limitation of the resampling CI method, as well as the iterative matching procedure are their particular application area.

For the resampling CI method, a relatively high number of intervention patients need to be included, which perhaps causes problems for some applications. Therefore, ethical concerns may limit the implementation of an adaptive matched case-control trial. Another limitation of matching approaches in general is the fact that the maximal sample size per group is limited by the number of patients in the control group, which leads to power restrictions. In practice, the trial could be underpowered from scratch due to a small number of control patients.

The simulation study demonstrates that the resampling CI approach is a powerful technique to reach a reasonable matching rate and a high power at the final analysis, but they are based on only one single model including different types of covariates. Even though more complex models are not evaluated in the simulations, higher model complexity is not expected to strongly influence the performance of the approaches when model convergence is guaranteed. A higher degree of misspecification of the propensity score model would lead to a lower matching rate. However, this would be the case for both discussed methods, the naïve and the resampling CI approach. More simulation scenarios are needed to give a detailed assessment of the amount this misspecification influences the matching rate and the power.

The developed iterative procedure for situations where a large control group exists and a $1 : M$ matching design is pursued, is evaluated for two different scenarios of overlap between matching variables only. To quantify the influence of this overlap to the number of matching partners chosen by the iterative procedure, a broader range of scenarios is needed, which might be a topic for future research. Furthermore, the simulations do not cover an effect estimation to quantify power and type I error, because this highly depends on the type of endpoint and the study design. Krisam et al. (2020) includes the iterative procedure in the Matched Threshold Crossing Design, but further research is needed to investigate the characteristics of the procedure combined with other study designs.

4.2 Synthesis of Evidence

4.2.1 Discussion and Contributions to Research

In the field of evidence synthesis, indirect comparisons allow for estimation of therapy effects when direct evidence is not available. To identify the possible underlying differences between the trials constitutes an important step before conducting an indirect comparison. Based on this investigations, a careful decision on the method for the indirect comparison should be made to avoid bias. A checklist on how to transparently document the present situation and to put the results into context was proposed by Kiefer et al. (2015). The simulation study contrasting methods for indirect comparisons observes that indirect effect estimates have wide confidence intervals in scenarios commonly met in practice. Scenarios, where the assumptions of the methods (see Section 2.2) hold true, perform better, but the behavior is far from good performance in terms of power. This observation is in agreement with other publications that include simulation studies covering indirect comparisons (Mills et al., 2011; Kühnast et al., 2017). Facing the results, the fact that indirect comparisons in early benefit assessment mostly lead to the conclusion that there is no additional benefit is not surprising (Ruof et al., 2014). Even though a benefit would actually exist, low power and wide confidence intervals may cause that it will hardly be shown by the indirect comparison. The results demonstrate that there are situations where the method of Bucher performs better than MAIC and vice versa. Discrepancies from underlying method assumptions induce biased effect estimates, even though in some scenarios MAIC results in less biased estimates. The presence of an effect

modification comes to superiority of the MAIC over the method of Bucher. However, there are also situations where MAIC leads to higher bias and less power compared to the method of Bucher, one reason might be the one-arm weighting when models are already adjusted for all influencing confounders. Or it is caused by adjusting matching models in MAIC for confounders which are not effect modifiers, the weighting seems to result in more biased effects for the indirect comparison. These results are in line with the observations of Kühnast et al. (2017), although the underlying sample sizes are chosen differently. In practice, it may not be given, that the models are adjusted for all relevant confounders and effect modifiers, moreover this assumption cannot be checked and increases bias and RMSE. Differences in the set of confounders between trials lead to similar results. Whereas, when the overlap of effect modifiers differs, type I error rate is inflated and for binary endpoints bias and RMSE are also higher compared to scenarios with complete overlap. Ignoring the effect modification was observed to give better results which might be due to the fact that the evaluation focus on the marginal effect of the treatment. The interaction term itself is not evaluated because it cannot be assessed by the marginal effect of the interaction (Norton et al., 2004), and the effect modification is not chosen to be extremely large. These results substantiate the request by Leahy and Walsh (2019) that a strong justification for the assumed effect modifiers in MAIC needs to be given. To summarize, in the case of similar patient characteristics and adjusted effect estimates, the method of Bucher has the advantage of preserving the within-study randomization. However, if effect modification is present in one or both trials as well as differences with respect to effect modifiers and adjustment of regression models is observed, MAIC provides less biased effect estimates and higher coverage.

In practice, a trial investigator may already have in mind to use a particular study for a later indirect comparison. In case the individual IPD (or AgD) trial is planned at higher power levels, a higher power is achieved for the indirect comparison. Situations where all method assumptions are met gain more power than situations with differences in patient characteristics, in the adjustment of effect estimates, or in the presence of effect modification. So, more precise estimates in the head-to-head comparisons lead to a higher power in the indirect comparison. Therefore, with a view to the later indirect comparison, it might worth it to invest more sample size into the head-to-head trial because power can be influenced.

Another attempt to achieve more precise estimates is to include more trials and therefore, a higher sample size in the indirect comparison. Moreover, when indirect evidence is needed, one should always consult all disposable information to avoid an additional bias due to the choice of the study. The simulation study carried out in this thesis observes that a higher power can be achieved by using more studies. Whereas Mills et al. (2011) report that power is still lower than 20% when using multiple studies within the method of Bucher for the considered scenarios. Certainly, the magnitude of the gain in efficiency depends on type (AgD or IPD) and the number of included studies as well as on the presence of effect modification and differences in patient characteristic distributions. For situations including one IPD study and multiple AgD studies, increasing the number of studies is not sharply enhancing power. When assuming a higher variance in the treatment effect distribution simulations lead to higher bias. Applying fixed or random-effects meta-analysis results are comparable in the considered scenarios. Nevertheless, it is necessary to evaluate the underlying situation to decide whether the strong assumptions for fixed effects meta-analysis are fulfilled. The availability of multiple IPD studies increases power by increasing the number of available trials, which is a direct consequence of the higher sample size. Reasonable power regions, usually above 80%, are observed to be reached for moderate and high treatment effects. For lower treatment effects, even ten IPD studies do not increase power noticeably. In case that both multiple IPD and multiple AgD trials are used the power increases with the number of trials. Bias and RMSE are rising with the treatment effect because the planned sample size for detecting higher treatment effects is smaller which causes higher standard errors. The use of multiple IPD studies can reduce bias and RMSE which is independent of the number of AgD studies. Furthermore, power results can be enhanced by using more IPD trials while increasing the number of AgD does not show an effect of the same extent. However, it is important to note that some scenarios showed an increase in type I error rates by increasing the number of studies. Differences in patient characteristics lead to lower power (in comparison to similar patient characteristics) in the MAIC procedure for some approaches. This may be due to the misspecification in the matching model, meaning that not only effect modifiers are included in the matching model. Including also confounders in MAIC is a realistic scenario because, in practice, it is usually unknown which variables are effect modifiers and need to be considered in MAIC. The simulations indicate that when using the MAIC due to differences

in patient populations and the presence of effect modifications, an approach conducting all indirect comparisons leads to better results. However, when characteristics are equal or only one IPD trial is included, the methods which pool the treatment effects or data before conducting the indirect comparison perform better. Using the Rücker et al. (2017) adjustment for standard errors leads to higher bias, which is potentially a sign of conservatism.

In summary, the proposed methods allow enhancing quality (in terms of power, coverage, bias, and RMSE) of indirect comparison by including all available evidence within an indirect comparison, but underlying assumptions need to be addressed and considered in the choice of methods and the interpretation of results.

4.2.2 Limitations and Directions for Future Research

Indirect comparisons using MAIC always suffer from the issue of defining a target population, because IPD will be shifted towards the AgD trial, which means the AgD trial defines the target population. The method of Bucher just assumes that the target population is the same in both trials, which is a strong assumption and likely to be violated in practice. By combining multiple indirect comparisons, this issue is reinforced because the treatment effect is calculated for an average population that may not at all be relevant or observable in practice. Conducting meta-analysis to calculate an average effect estimate over different studies before the indirect comparison is weighting the data on patient-level is the counterintuitive direction because the treatment effect is already summarized. Further simulations are needed to investigate the question of target populations, which should be based on different types of simulation scenarios than considered in this thesis; this applies for the method comparison as well as the inclusion of multiple studies.

A limitation of both simulation studies is that just the method of Bucher and the MAIC are included because they are most often used, which is probably the case since they are accepted in the field of HTA. Nevertheless, there are other methods available for indirect comparisons, such as simulated treatment comparisons (Caro and Ishak, 2010; Ishak et al., 2015a,b), cross-design synthesis (Droitcour et al., 1993), or likelihood reweighting methods (Nie et al., 2013) for which properties are currently not sufficiently examined (Kühnast et al., 2017). But note, that the STC by Ishak et al. (2015b) is included in the method comparison

by Weber et al. (2020a), where similar patters for MAIC and STC are found but bias and RMSE are observed to be higher for STC in most scenarios. Further research could compare the less frequently used methods with the method of Bucher and MAIC to get more insights and to give more detailed recommendations for users. Additionally, these methods are not evaluated regarding the use of multiple studies, which also needs further investigations and development.

One strength of the simulation study comparing the methods for adjusted indirect comparisons is the variety of clinically relevant scenarios, including confounders, correlations, interactions (effect modification), adjustment of regression models, and overlap of effect modifiers and confounders, which are evaluated and compared within this work. Nevertheless, the following limitations apply to the simulation study: One clinically inspired data set is considered only, it is assumed that the interaction terms have the same sign, and treatment effect modifiers are known; additionally, for MAIC the overlap is good enough to expect matching to work well. Also, the simulations evaluating the methods for incorporating multiple studies in indirect comparisons cover clinically relevant scenarios, including situations where assumptions of methods for indirect comparisons are violated. Nevertheless, just one underlying data example and one set of covariates in each comparison were considered, which are the same between head-to-head trials. Moreover, the effect modification (in terms of an interaction with treatment) was limited to one variable, and the Cox models were adjusted for all covariates (except the interaction term). Further research is needed to explore the performance among other regression models for estimating the treatment effect which is not adjusted for all the effect modifiers, different overlap of effect modifiers, and various scenarios of misspecification in the matching model. Application to other endpoints is also needed to give complete guidance.

The sample size of all simulated trials is based on a sample size calculation for the assumed effects making the results realistic and transferable to real trials. The treatment effects are chosen according to official recommendations for the classification of effects in benefit assessment (Skipka et al., 2015), which makes the scenarios practically relevant. When multiple studies are used, it would be interesting if results change in case trials have substantially different sample sizes.

For the incorporation of multiple studies, the results concerning the evaluation measures need to be interpreted with caution, since the mean of the effect estimate distribution is used as a reference value to calculate the evaluation measures which may not be a good choice, especially for a small number of studies. When the same trial is used in multiple indirect comparisons, the approximate adjustment according to Rücker et al. (2017) was transferred, but further research is needed to validate and extend this adjustment in the underlying situation, as well as to examine its potential conservatism.

4.3 Conclusion

In conclusion, matching procedures are useful tools that find application in many application areas. The power of matched case-control trials can be enhanced by using an estimate of the trial-specific matching rate for sample size recalculation at the interim analysis. Depending on the size of the external study arm and the aim of the current investigations, different strategies should be followed. For small historical data, the developed resampling approach showed good properties, whereas, for large control groups an iterative procedure to determine the number of matching partners is the better option. In evidence synthesis, matching procedures cannot be used straight forward because issues with the definition of the target population, low power, and potential bias may arise. A key finding of this thesis is that matching variables in MAIC need to be chosen carefully, because confounders, which do not modify the effect, considerably influence the precision of the indirect comparison. Addressing the low power, investigations for the inclusion of multiple studies in indirect comparisons are made, which identifies promising scenarios, but a clear recommendation on how to include various studies in MAIC cannot be given. Moreover, a careful interpretation is needed when results of indirect comparisons are discussed and they cannot replace direct evidence by RCTs.

Chapter 5

Summary

5.1 Summary (English)

The gold standard for clinical studies are blinded randomized trials, but such a design is not always feasible due to ethical or practical reasons. Using an external historical control group out of an earlier conducted trial or registry might be an option. When using historical controls, one often faces the situation of non-comparable study populations. Matching procedures may help to build balanced samples for comparison. In this thesis an adaptive matched case-control trial design is established, which allows for a sample size recalculation at a planned interim analysis with the goal to enhance the matching rate at final analysis. The recalculation is based on the lower confidence interval limit of the matching rate observed at interim analysis. The newly developed resampling CI method estimates the 1:1 matching rate using a bootstrap like procedure (without replacement) and equal-sized groups for matching at interim. A naïve approach would be to use all patients for estimating the matching rate and directly reflect this value for recalculating the sample size. The new approach shows good performance in terms of power and type I error rate but needs more newly recruited patients than the naïve approach. Additionally, investigations for the time point of interim analysis are done. Simulations result in a number of $\frac{1}{2}$ to $\frac{2}{3}$ of the control patients, however, it seems that the time point is more depending on the actual number of patients used for matching than on the proportion. However, if the historical control group is large and for example only

a small phase II trial is feasible the before described method might not be a good choice. Rather, each intervention patient may find more than one matching partner. Therefore, an iterative procedure to determining the number of matching partners is developed. The idea is an interim analysis, which includes an iterative increase in the number of matching partners and a parallel calculation of the matching rate. The number will be increased as long as the 1:M matching rate is higher than the 1:1 matching rate including a potential tolerance. The 1:M matching rate at interim analysis can then be used for recalculating the sample size. This procedure is easy to implement and can be combined with many study designs, such as two-stage designs. One has to note that the number of matching partners highly depends on the overlap of patient populations, meaning a small overlap leads to a low number of matching partners and vice versa. To conclude, by involving the trial-specific matching rate in the sample size recalculation one is able to enhance power in a matched case-control trial. Not only in the generation of evidence unbalanced patient cohorts arise, but also in evidence synthesis this poses a problem. A common situation in evidence synthesis is an indirect comparison, where the comparison of interest, assume treatment A versus C, is not examined in a direct comparison. But there are trials comparing A with treatment B and another trial comparing C and B. using those trials to calculate a treatment effect for A versus C is called indirect comparison. It is likely that the independent trials AB and CB do not have the same underlying population. A special case, where individual patient data is available for one of the trials is assumed. Then a matching-like procedure can help to balance the cohorts, this method is called matching adjusted indirect comparison which is not sufficiently examined, yet. Another widely used method for indirect comparisons is the method of Bucher. A method comparison between those two methods is conducted for clinically relevant scenarios where assumptions of the methods are violated. Simulations lead to the conjecture that indirect comparisons are considerably underpowered. The method of Bucher and the matching adjusted indirect comparison show similar performance in scenarios without cross-trial differences. The matching approach leads to higher coverage and power when populations differ, effect modifiers are present, and regression models are not sufficiently adjusted. But matching confounders which do not modify the effect leads to increased bias. Until now, indirect comparisons are applied using one study per treatment comparison because the matching adjusted indirect comparison is designed for this setting. Nevertheless,

it is likely that there are two or even more studies comparing the same treatments. When synthesizing evidence, one should always aim to include all appropriate evidence. Therefore, approaches to include multiple studies in indirect comparisons are introduced and compared. All include a step for combining treatment effects and one for calculating indirect treatment effects. The main difference between the approaches is the order of those two steps. An increasing number of studies can enhance power to desired regions above 80%, but it was not possible to identify one best performing method over all considered scenarios. In conclusion, when applying matching procedures in evidence synthesis the underlying situation needs to be checked carefully, and matching variables need to be chosen carefully because adjusting for confounders influences the precision of the indirect comparison.

5.2 Zusammenfassung (Deutsch)

Der Goldstandard im Rahmen klinischer Studien ist eine doppelt-verblindete, randomisierte Studie. Es gibt jedoch Situationen, die ein solches Design aus ethischen oder praktischen Gründen nicht zulassen. Eine Möglichkeit dennoch einen Vergleich durchzuführen, ist die Verwendung einer externen Kontrollgruppe, diese kann aus einer bereits durchgeführten Studie oder auch aus einem Register stammen. Zieht man eine historische Kontrollgruppe heran, so sind die Studienpopulationen oft nicht vergleichbar. Matchingverfahren können dazu beitragen trotzdem einen Vergleich mit balancierten Patientengruppen durchführen zu können. In dieser Arbeit wird ein adaptives gematchtes Fall-Kontroll-Studiendesign entwickelt, welches eine Zwischenauswertung mit Fallzahlrekalkulation vorsieht. Die Rekalkulation der Fallzahl basiert auf der unteren Grenze des Konfidenzintervalls der Matchingrate bei Zwischenauswertung. Der Schätzer der Matchingrate wird in diesem neuen Ansatz mittels einer dem Bootstrap-ähnlichen Methode auf der Basis von gleich großen Gruppen bestimmt. Dieser Ansatz zeigt gute Ergebnisse in Bezug auf Power und dem Fehler erster Art im Vergleich zu der Herangehensweise unter Verwendung aller Kontrollpatienten bei der Zwischenauswertung. Zudem wurde der Zeitpunkt der Zwischenauswertung untersucht. Simulationen kamen zu dem Ergebnis, dass eine Zwischenauswertung nach Rekrutierung von $\frac{1}{2}$ bis $\frac{2}{3}$ der Zahl an Kontrollpatienten eine gute Wahl ist. Jedoch scheint es, als wäre der Zeitpunkt mehr von der absoluten Zahl an Patienten als vom Anteil abhängig. Ist die Kontrollgruppe jedoch sehr groß ist die zuvor besprochene Methode nicht die beste Wahl. Vielmehr ist es möglich mehr als nur einen Kontrollpatienten pro Interventionspatienten zu matchen. Um diese Anzahl an Matchingpartnern zu bestimmen kann das entwickelte iterative Verfahren verwendet werden. Die Idee basiert auf einer Erhöhung der Matchingpartner und gleichzeitiger Bestimmung der Matchingrate während einer Zwischenauswertung. Die Anzahl wird solange die 1:M Matchingrate oberhalb der 1:1 Matchingrate abzüglich einer definierten Toleranz liegt, erhöht. Die entsprechende 1:M Matchingrate kann dann noch für eine Fallzahlrekalkulation genutzt werden. Das iterative Verfahren ist einfach zu implementieren und lässt eine Kombination mit vielen Studiendesigns zu. Es muss beachtet werden, dass die Anzahl der Matchingpartner maßgeblich von der Überlappung der Patientenkollektive abhängt, d.h. liegt nur wenig Überlappung vor können dementsprechend nur wenige Matchingpartner gefunden werden und

umgekehrt. Daraus folgt, dass die studienspezifische Matchingrate für eine Rekalkulation der Fallzahl von Nutzen sein kann, um die Power zu erhöhen. Nicht nur bei der Generierung von Evidenz treten unbalancierte Populationen auf, auch in der Synthese von Evidenz spielt dieses Problem eine Rolle. Eine häufige Situation der Evidenzsynthese ist ein indirekter Vergleich. Dieser ist von Interesse, wenn zwei Medikamente, angenommen A und C, nicht in einem direkten Vergleich gegenübergestellt wurden, jedoch sind Vergleiche der Medikamente A und B sowie C und B verfügbar. Es ist denkbar, dass die Studien AB und CB unterschiedliche Patientenpopulationen betrachten. Liegen in der Situation eines indirekten Vergleichs von einer Studie individuelle Patientendaten vor, kann dieses Ungleichgewicht von Matchingverfahren adressiert werden. Neben der Methode von Bucher ist der Matching Adjusted Indirect Comparison eine etablierte Methode für indirekte Vergleiche, dieser ist jedoch noch nicht hinreichend für praxisrelevante Situationen untersucht. Um diese beiden Methoden gegenüberzustellen und näher in praxisrelevanten Situationen (z.B. verletzte Annahmen) zu untersuchen wird eine Simulationsstudie durchgeführt. Simulationen führen zu der Vermutung, dass indirekte Vergleiche deutlich zu wenig Power haben, um einen Effekt nachweisen zu können. Wenn sich die Studienpopulationen unterscheiden, Effektmodifikation auftritt und Regressionsmodelle nicht ausreichend adjustiert sind, erreicht man mit dem Matching-Ansatz eine höhere Konfidenzintervallüberdeckung sowie Power. Wird allerdings für Kovariaten gematcht, die den Effekt nicht beeinflussen, so führt das zu einer höheren Verzerrung, die Matchingvariablen sollten deshalb mit Bedacht gewählt werden. Bisher wurde nur eine Studie pro direktem Vergleich verwendet, da Matching Adjusted Indirect Comparisons für diese Situation entwickelt wurden. Es ist jedoch wahrscheinlich, dass mehrere Studien den gleichen Vergleich durchgeführt haben. Kombiniert man Evidenz, sollten immer alle verfügbaren Studien verwendet werden. Um eine Vielzahl von Studien in indirekte Vergleiche einbinden zu können, werden verschiedene Ansätze entwickelt und verglichen. Diese beinhalten alle einen Schritt für die Synthese der Evidenz und einen für den indirekten Vergleich. Der Hauptunterschied zwischen den Ansätzen ist die Reihenfolge dieser beiden Schritte. Berücksichtigt man eine höhere Zahl an Studien kann die gewünschte Power von 80% erreicht werden. Es war allerdings nicht möglich eine Methode zu identifizieren, die in allen betrachteten Szenarien die besten Ergebnisse erzielt. Folglich muss die zugrunde liegende Situation sorgfältig analysiert werden, sowie die Methodik und die Matchingvariablen mit Bedacht gewählt werden.

Bibliography

Austin, P. (2011).

Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies.

Pharm Stat, 10:150–161.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007).

A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study.

Statistics in Medicine, 26:734–753.

Balduzzi, S., Rücker, G., and Schwarzer, G. (2019).

How to perform a meta-analysis with r: a practical tutorial.

Evidence-Based Mental Health, 22(4):153–160.

Barber, P., Demchuk, A., Zhang, J., and Buchan, A. (2000).

Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy.

The Lancet, 355(9216):1670 – 1674.

Belger, M., Brnabic, A., Kadziola, Z., Petto, H., and Faries, D. (2015).

Inclusion Of Multiple Studies In Matching Adjusted Indirect Comparisons (MAIC).

Value in Health, 8:18:A33.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2011).

Introduction to Meta-Analysis.

John Wiley & Sons, Chichester, U.K.

Broyden, C. G. (1970).

The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations.

IMA Journal of Applied Mathematics, 6(1):76–90.

Bucher, H. C., Guyatt, G. H., Griffith, L. E., and Walter, S. D. (1997).

The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials.

Journal of Clinical Epidemiology, 50(6):683 – 691.

Canty, A. and Ripley, B. (2017).

boot: Bootstrap R (S-Plus) Functions.

R package version 1.3-20.

Caro, J. and Ishak, K. (2010).

No Head-to-Head Trial? Simulate the Missing Arms.

Pharmacoeconomics, 28:957–967.

Charpentier, P., Bogardus, S., and Inouye, S. (2001).

An algorithm for prospective individual matching in a non-randomized clinical trial.

Journal of Clinical Epidemiology, 54(11):1166 – 1173.

Cohen, J. (1988).

Statistical Power Analysis for the Behavioral Sciences.

Lawrence Erlbaum Associates, Hillsdale, NJ.

Droitcour, J., Silberman, G., and Chelimsky, E. (1993).

Cross-Design Synthesis: A New Form of Meta-Analysis for Combining Results from Randomized Clinical Trials and Medical-Practice Databases.

International Journal of Technology Assessment in Health Care, 9(3):440–449.

Egger, M. and Smith, G. D. (1997).

Meta-analysis: Potentials and promise.

BMJ, 315(7119):1371–1374.

Faraoni, D. and Schaefer, S. (2016).
Randomized controlled trials vs. observational studies: why not just live together?
BMC Anesthesiology, 16(1).

Frakt, A. (2015).
An observational study goes where randomized clinical trials have not.
Jama, 313(11):1091 – 1092.

Gan, H., Grothey, A., Pond, G., Moore, M., Siu, L., and Sargent, D. (2010).
Randomized phase ii trials: Inevitable or inadvisable?
Journal of Clinical Oncology, 28(15):2641–2647.

Glenny, A., Altman, D., Song, F., Sakarovitch, C., and Deeks, J. (2005).
Indirect comparisons of competing interventions.
Health Technol Assess, 9(26).

Goodman, S., Schneeweiss, S., and Baiocchi, M. (2017).
Using design thinking to differentiate useful from misleading evidence in observational research.
JAMA, 317(7):705–707.

Gu, X. and Rosenbaum, P. (1993).
Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms.
Journal of Computational and Graphical Statistics, 2(4):405–420.

Harrison, R. (2016).
Phase ii and phase iii failures: 2013-2015.
Nature Reviews Drug Discovery, 15:817–818.

Huber, P. J. (1967).
The behavior of maximum likelihood estimates under nonstandard conditions.
In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233, Berkeley, Calif. University of California Press.

IQWiG (2017).

General Methods 5.0.

`https://www.iqwig.de/en/methods/methods-paper.3020.html`.

Accessed: 02.03.2020.

IQWiG (2019).

IQWiG's results on AMNOG at a glance.

`https://www.iqwig.de/en/press/amnog-at-a-glance.7723.html`.

Accessed: 02.03.2020.

Ishak, K., Proskorovsky, I., and Benedict, A. (2015a).

Simulation and Matching-Based Approaches for Indirect Comparison of Treatments.

PharmacoEconomics, 33:537–549.

Ishak, K., Rael, M., Phatak, H., Masseria, C., and Lanitis, T. (2015b).

Simulated Treatment Comparison of Time-To-Event (And Other Non-Linear) Outcomes.

Value in Health, 18(7):719.

Kiefer, C., Sturtz, S., and Bender, R. (2015).

Indirect comparisons and network meta-analyses: Estimation of effects in the absence of head-to-head trials – part 22 of a series on evaluation of scientific publications.

Dtsch Arztebl Int, 112(7119):803 – 8.

Krisam, J., Weber, D., Schlenk, R., and Kieser, M. (2018).

Matched-Threshold-Crossing (MTC): a novel trial design to enhance single-arm phase II trials by including matched control patients.

Presented at the GMDS, doi: 10.3205/18gmds033.

Kühnast, S., Schiffner-Rohe, J., Rahnenführer, J., and Leverkus, F. (2017).

Evaluation of Adjusted and Unadjusted Indirect Comparison Methods in Benefit Assessment.

Methods Inf Med, 58(1):43–58.

Leahy, J. and Walsh, C. (2019).

Assessing the impact of a matching adjusted indirect comparison in bayesian network meta analysis.

Research Synthesis Methods, 10:546 – 568.

Lyden, P., Lu, M., Levine, S., Brott, T., and Broderick, J. (2001).

NINDS rtPA Stroke Study Group. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity.

Stroke, 32:1310–1317.

Mills, E. J., Ghement, I., O'Regan, C., and Thorlund, K. (2011).

Estimating the Power of Indirect Comparisons: A Simulation Study.

PLOS ONE, 6:1–8.

Morris, T. P., White, I. R., and Crowther, M. J. (2019).

Using simulation studies to evaluate statistical methods.

Statistics in Medicine, 38(11):2074–2102.

Nash, P., McInnes, I. B., Mease, P. J., Thom, H., Hunger, M., Karabis, A., Gandhi, K., Mpofu, S., and Jugl, S. M. (2018).

Secukinumab Versus Adalimumab for Psoriatic Arthritis: Comparative Effectiveness up to 48 Weeks Using a Matching-Adjusted Indirect Comparison.

Rheumatol Ther., 5(1):99–122.

Nie, L., Zhang, Z., Rubin, D., and Chu, J. (2013).

Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference.

Ann. Appl. Stat., 7(3):1796–1813.

Norton, E., Wang, H., and Ai, C. (2004).

Computing interaction effects and standard errors in logit and probit models.

Stata Journal, 4(2):154 – 167.

Petto, H., Kadziola, Z., Brnabic, A., Saure, D., and Belger, M. (2019).

Alternative Weighting Approaches for Anchored Matching-Adjusted Indirect Comparisons via a Common Comparator.

Value in Health, 22(1):85 – 91.

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., and Welton, N. J. (2016). NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE.
`http://nicedsu.org.uk/technical-support-documents/population-adjusted-indirect-comparisons-maic-and-stc/`.
Accessed: 02.03.2020.

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., and Welton, N. J. (2018). Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. *Medical Decision Making*, 38(2):200–211.

R Core Team (2017).
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria.

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

Rosenbaum, P. and Rubin, D. (1984).
Reducing bias in observational studies using subclassification on the propensity score.
Journal of the American Statistical Association, 79(1):33–38.

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

Rubin, D. and Thomas, N. (1996).
Matching Using Estimated Propensity Scores: Relating Theory to Practice.
Biometrics, 52(1):516–524.

Rücker, G., Cates, C. J., and Schwarzer, G. (2017).

Methods for including information from multi-arm trials in pairwise meta-analysis.
Research Synthesis Methods, 8:392 – 403.

Ruof, J., Schwartz, F. W., Schulenburg, J.-M., and Dintsios, C.-M. (2014).
Early benefit assessment (EBA) in Germany: analysing decisions 18 months after introducing the new AMNOG legislation.
Eur J Health Econ., 15(6):577–589.

Schafer, J., Opgen-Rhein, R., Zuber, V., Ahdesmaki, M., Silva, A. P. D., and Strimmer., K. (2017).
corpcor: Efficient Estimation of Covariance and (Partial) Correlation.
`https://CRAN.R-project.org/package=corpcor`.
R package version 1.6.9.

Schoenfeld, D. (1983).
Sample-Size Formula for the Proportional-Hazards Regression Model.
Biometrics, 39(2):499–503.

Schönenberger, S., Uhlmann, L., Hacke, W., et al. (2016).
Effect of conscious sedation vs general anesthesia on early neurological improvement among patients with ischemic stroke undergoing endovascular thrombectomy: A randomized clinical trial.
JAMA, 316(19):1986–1996.

Schönenberger, S., Weber, D., Ungerer, M. N., et al. (2019).
The KEEP SIMPLEST Study: Improving In-House Delays and Periinterventional Management in Stroke Thrombectomy—A Matched Pair Analysis.
Neurocritical Care, 31. doi: 10.1007/s12028-018-00667-3.

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

Signorovitch, J., Sikirica, V., Erder, M. H., Xie, J., Lu, M., Hodgkins, P. S., Betts, K. A., and Wu, E. Q. (2012).
Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research.
Value Health, 15(6):940–947.

Signorovitch, J., Swallow, E., Kantor, E., Wang, X., Klimovsky, J., Haas, T., Devine, B., and Metrakos, P. (2013).
Everolimus and sunitinib for advanced pancreatic neuroendocrine tumors: a matching-adjusted indirect comparison.
Exp Hematol Oncol., 2(1):32.

Signorovitch, J. E., Wu, E. Q., Yu, A. P., Gerrits, C. M., Kantor, E., Bao, Y., Gupta, S. R., and Mulani, P. M. (2010).
Comparative Effectiveness Without Head-to-Head Trials.
PharmacoEconomics, 28(10):935–945.

Skipka, G., Wieseler, B., Kaiser, T., Thomas, S., Bender, R., Windeler, J., and Lange, S. (2015).
Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs.
Biometrical Journal, 56(3):261–267.

Snapinn, S. and Jiang, Q. (2011).
Indirect comparisons in the comparative efficacy and non-inferiority settings.
Pharmaceutical Statistics, 10(5):420–426.

Song, F., Loke, Y. K., Walsh, T., Glenny, A.-M., Eastwood, A. J., and Altman, D. G. (2009).
Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews.
BMJ, 338. doi: 10.1136/bmj.b1147.

Stewart, G. B., Altman, D. G., Askie, L. M., Duley, L., Simmonds, M. C., and Stewart, L. A. (2012).

Statistical Analysis of Individual Participant Data Meta-Analyses: A Comparison of Methods and Recommendations for Practice.

PLOS ONE, 7(10):1–8.

Sutton, A., Abrams, K., Jones, D., Sheldon, T., and Song, F. (2000).

Methods for Meta-Analysis in Medical Research.

John Wiley & Sons, Chichester, U.K.

Therneau, T. M. (2015).

A Package for Survival Analysis in S. R package version 2.38.

`https://CRAN.R-project.org/package=survival`.

Veroniki, A., Straus, S., and Soobiah, C. e. a. (2016).

A scoping review of indirect comparison methods and applications using individual patient data.

BMC Med Res Methodol, 16(47). doi: 10.1186/s12874-016-0146-y.

Weber, D., Uhlmann, L., Schönenberger, S., and Kieser, M. (2019).

Adaptive propensity score procedure improves matching in prospective observational trials.

BMC Med Res Methodol 19(150). doi:10.1186/s12874-019-0763-3.

White, H. (1980).

A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.

Econometrica, 48(4):817–838.

Publications

Methodological publications

Weber D., Uhlmann L., Schönenberger S., Kieser M. (2019) Adaptive propensity score procedure improves matching in prospective observational trials. *BMC Medical Research Methodology*, 19:150.

Weber D., Jensen K., Kieser M. (2020a) Comparison of Methods for Estimating Therapy Effects by Indirect Comparisons: A Simulation Study. *Medical Decision Making* (accepted)

Weber D., Jensen K., Kieser M. (2020b) Incorporation of Multiple Studies in Indirect Comparisons. (submitted)

Krisam J., **Weber D.**, Schlenk R., Kieser M (2020) Enhancing single-arm phase II trials by inclusion of matched control patients. *Statistics in Biopharmaceutical Research* (submitted)

Project-based publications

El Shafie R. A., Celik A., **Weber D.**, Schmitt D., Lang K., König L., Bernhardt D., Höne S., Forster T., von Nettelbladt B., Adeberg S., Debus J., Rieken S. (2020): A matched-pair analysis comparing stereotactic radiosurgery with whole-brain radiotherapy for patients with multiple brain metastases. *Journal of Neuro-Oncology*, 1-12.
doi: 10.1007/s11060-020-03447-2.

Bäumer A., **Weber D.**, Staufer S., Pretzl B., Körner G., Wang Y. (2020): Tooth loss in aggressive periodontitis: results 25 years after active periodontal therapy in a private practice, *Journal of Clinical Periodontology*, 47: 223-232. doi: 10.1111/jcpe.13225

Schwindling F. S., Hilgenfeld T., **Weber D.**, Kosinski M. A., Rammelsberg P., Tasaka A. (2019): In-vitro diagnostic accuracy of low-dose CBCT for evaluation of peri-implant bone lesions. *Clinical Oral Implants Research*, 30: 1200-1208. doi: 10.1111/clr.13533.

Sauer C., Weis J., Faller H., Junne F., Hönig K., Bergelt C., Hornemann B., Stein B., Teufel M., Goerling U., Erim Y., Geiser F., Niecke A., Senf B., **Weber D.**, Maatouk I. (2019): Impact of social support on psychosocial symptoms and quality of life in cancer patients: results of a multilevel model approach from a longitudinal multicenter study. *Acta Oncologica*, 58(9):1298-1306. doi: 10.1080/0284186X.2019.1631471.

Hilgenfeld T., Juerchott A., Deisenhofer U. K., **Weber D.**, Rues S., Rammelsberg P., Heiland S., Bendszus M., Schwindling F. S. (2019): In vivo accuracy of tooth surface reconstruction based on CBCT and dental MRI - A clinical pilot study. *Clinical Oral Implants Research*, 30: 920-927. doi: 10.1111/clr.13498.

Maharani C., Kes M., Frihandini Afief D., **Weber D.**, Marx M., Loukanova S. (2019): Primary care physicians' satisfaction after health care reform: a cross-sectional study from two cities in Central Java, Indonesia. *BMC Health Services Research*, 19: 290.
doi: 10.1186/s12913-019-4121-2.

Wollny A., Altiner A., Brand T., Garbe K., Kamradt M., Kaufmann-Kolle P., Leyh M., Poß-Doering R., Szecsenyi J., Uhlmann L., Voss A., **Weber D.**, Wensing M., Löffler C. (2019): Converting habits of antibiotic use for respiratory tract infections in German primary care - study protocol of the cluster-randomized controlled CHANGE-3 trial. *Trials*, 20(1):103. doi: 10.1186/s13063-019-3209-7.

El Sayed N., Baeumer A., El Sayed S., Wieland L., **Weber D.**, Eickholz P., Pretzl B. (2019): Twenty years later: Oral health-related quality of life and standard of treatment in patients with chronic periodontitis. *Journal of Periodontology*, 90(4):323-330.
doi: 10.1002/JPER.18-0417.

Weber D., Koller M., Theuns D., Yap S., Kühne M., Sticherling C., Reichlin T., Szili-Torok T., Osswald S., Schaer B. (2019): Predicting defibrillator benefit in patients with cardiac resynchronization therapy: a competing risk study, Heart Rhythm. 16(7):1057-1064. doi: 10.1016/j.hrthm.2019.01.033.

El Shafie R. A., Böhm K., **Weber D.**, Lang K., Schlaich F., Adeberg S., Paul A., Haefner M. F., Katayama S., Hörner-Rieber J., Hoegen P., Löw S., Debus J., Rieken S., Bernhardt D. (2019): Palliative Radiotherapy for Leptomeningeal Carcinomatosis-Analysis of Outcome, Prognostic Factors, and Symptom Response. *Frontiers in Oncology*, 8:641. doi: 10.3389/fonc.2018.00641.

El Shafie RA., Böhm K., **Weber D.**, Lang K., Schlaich F., Adeberg S., Paul A., Haefner MF., Katayama S., Sterzing F., Hörner-Rieber J., Löw S, Herfarth K., Debus J., Rieken S., Bernhardt D. (2019): Outcome and prognostic factors following palliative craniospinal irradiation for leptomeningeal carcinomatosis. *Cancer Management and Research*, 11: 789-801. doi: 10.2147/CMAR.S182154.

El Shafie R. A., Tonndorf-Martini E., Schmitt D., **Weber D.**, Celik A., Dresel T., Bernhardt D., Lang K., Hoegen P., Adeberg S., Paul A., Debus J., Rieken S. (2019): Pre-Operative Versus Post-Operative Radiosurgery of Brain Metastases - Volumetric and Dosimetric Impact of Treatment Sequence and Margin Concept. *Cancers*, 11(3): 294. doi: 10.3390/cancers11030294.

Schönenberger S., **Weber D.**, Ungerer M., Pfaff J., Schieber S., Uhlmann L., Heidenreich P., Bendszus M., Kieser M., Wick W., Möhlenbruch M., Ringleb P., Bösel J. (2019): The KEEP SIMPLEST Study: Improving In-House Delays and Periinterventional Management in Stroke Thrombectomy - A Matched Pair Analysis. *Neurocritical Care*. 31(1):46-55. doi: 10.1007/s12028-018-00667-3.

Kamradt M., Kaufmann-Kolle P., Andres E., Brand T., Klingenberg A., Glassen K., Poss-Doering R., Uhlmann L., Hees K., **Weber D.**, Gutscher A., Wambach V., Szecsenyi J., Wensing M. (2018): Sustainable reduction of antibiotics-induced antimicrobial resistance (ARena) in German ambulatory care: study protocol of a cluster randomized trial. *Implementation Science*, 13(1):23. doi: 10.1186/s13012-018-0722-0.

El Shafie R. A., Bougatf N., Sprave T., **Weber D.**, Oetzel D., Machmer T., Huber P. E., Debus J., Nicolay N. H. (2018): Oncologic therapy support via means of a dedicated mobile application (OPTIMISE-1) - a prospective pilot trial. *JMIR Research Protocols*. 7(3):e70. doi: 10.2196/resprot.8915.

Bischoff M. S., Meisenbacher K., Peters A. S., **Weber D.**, Bisdas T., Torsello G., Böckler D. (2018): Clinical significance of perioperative changes in ankle-brachial index with regard to extremity-related outcome in non-diabetic patients with critical limb ischemia. *Langenbeck's Archives of Surgery*, 403(6):741-748. doi: 10.1007/s00423-018-1689-7.

El Shafie R. A., Czech M., Kessel K. A., Habermehl D., **Weber D.**, Rieken S., Bougatf N., Jäkel O., Debus J., Combs S. E. (2018): Evaluation of particle radiotherapy for the re-irradiation of recurrent intracranial meningioma. *Radiation Oncology*, 13(1):86. doi: 10.1186/s13014-018-1026-x.

El Shafie R. A., Czech M., Kessel K. A., Habermehl D., **Weber D.**, Rieken S., Bougatf N., Jäkel O., Debus J., Combs S. E. (2018): Clinical outcome after particle therapy for meningiomas of the skull base: toxicity and local control in patients treated with active rasterscanning. *Radiation Oncology*, 13(1):54. doi: 10.1186/s13014-018-1002-5.

El Shafie R. A., **Weber D.**, Bougatf N., Sprave T., Oetzel D., Huber P. E., Debus J., Nicolay N. H. (2018): Supportive Care in Radiotherapy Based on a Mobile App: Prospective Multicenter Survey. *JMIR Mhealth Uhealth*, 6(8):e10916. doi: 10.2196/10916.

Pretzl B., El Sayed S., **Weber D.**, Eickholz P., Baeumer A. (2018): Tooth loss in periodontally compromised patients: Results 20 years after active periodontal therapy. *Journal of Clinical Periodontology*, doi: 45/11:1356-1364.

Conference contributions

The presenting author is underlined.

Weber D., Nakken S., Hovig E., Zucknick M. Predicting treatment response in personalized cancer therapy - method comparison and a neural network approach using prior informa-

tion. *Annual Conference of the International Society of Clinical Biostatistics (ISCB).* Poster presentation. July 2017, Vigo, Spain.

Weber D., Nakken S., Hovig E., Zucknick M. Predicting treatment response in personalized cancer therapy - method comparison and a neural network approach using prior information. *CEN-ISBS Vienna 2017 - Joint Conference on Biometrics & Biopharmaceutical Statistics.* Presentation. August 2017, Vienna, Austria.

Krisam J., **Weber D.**, Schlenk R., Kieser M. Enhancing Phase II trials by inclusion of matched control patients - the Matched-Threshold-Crossing (MTC) design. *Biometrisches Kolloquium 2018.* Presentation. March 2018, Frankfurt am Main, Germany.

Weber D., Jensen K., Kieser M. Comparison of Methods for Estimating Therapy Effects by Indirect Comparisons - A Simulation Study. *International Biometric Conference Barcelona 2018.* Poster presentation. July 2018, Barcelona, Spain.

Weber D., Jensen K., Kieser M. Comparison of Methods for Estimating Therapy Effects by Indirect Comparisons - A Simulation Study. *63. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS) e.V.* Presentation. September 2018, Osnabrück, Germany.

Weber D., Uhlmann L., Schönenberger S., Kieser M. Adaptive propensity score procedure improves matching in prospective observational trials. *DAGStat Conference 2019 (5. conference of the Deutsche Arbeitsgemeinschaft Statistik).* Presentation. March 2019, Munich, Germany. Presentation:

Weber D., Uhlmann L., Schönenberger S., Kieser M. Improve Matching in Prospective Observational Trials by an Adaptive Propensity Score Procedure. *Annual Conference of the International Society of Clinical Biostatistics (ISCB40).* Presentation. July 2019, Leuven, Belgium.

Weber D., Jensen K., Kieser M. Simulation in Evidence Synthesis - Incorporation of Multiple Studies. *10th International Workshop on Simulation and Statistics.* Presentation. September 2019, Salzburg, Austria.

Appendix A: Additional Tables and Figures

A.1 Generation of Evidence - Resampling CI Method

The results (power, type I error rate, mean matching rate, and mean sample size) for the time point of interim analysis for a small control group ($n_{control} = 50$) and a large control group ($n_{control} = 500$) are given in Figure 11 and 12.

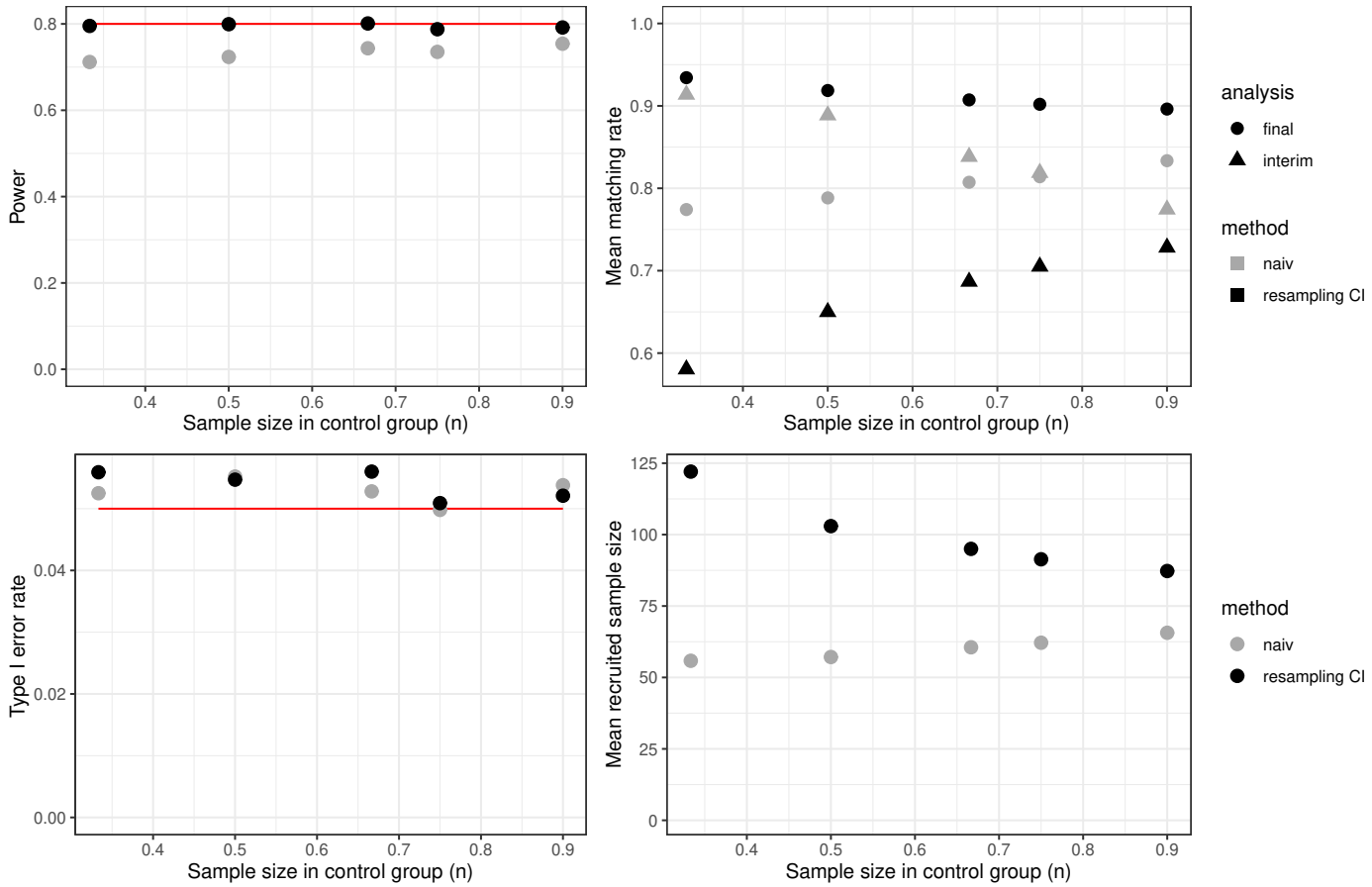


Figure 11: *Power, type I error rate, mean matching rate, and mean sample size in the treated group for different time points of the interim analysis ($n_{control} = 50$) (adapted from Weber et al. (2019)).*

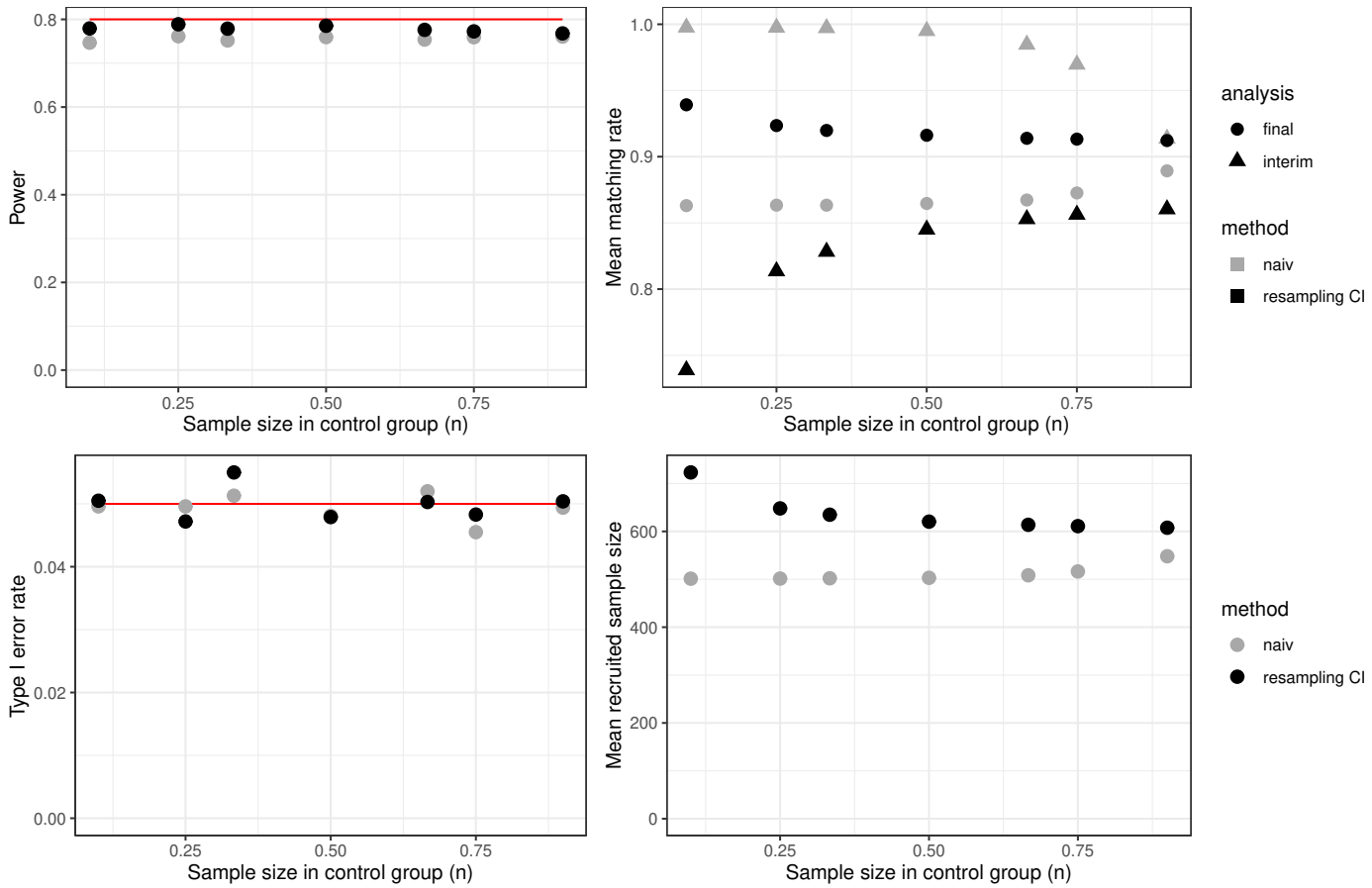


Figure 12: *Power, type I error rate, mean matching rate, and mean sample size in the treated group for different time points of the interim analysis ($n_{control} = 500$) (adapted from Weber et al. (2019)).*

Comment: Parts of the following Section A.2 are already included in the accepted manuscript Weber et al. (2020a). The manuscript has been written by myself, but may contain comments and corrections from the co-authors.

A.2 Synthesis of Evidence - Method Comparison

In addition to the setting where all regression models are adjusted for relevant confounders and effect modifiers, the method comparison was done for the situation where

- the regression models for estimating the treatment effect in trial AB and CB do not include an interaction for the effect modification (Tables 27 to 30),
- the regression models for estimating the treatment effect in trial CB are not adjusted for confounders (Tables 31 to 35).

The corresponding detailed results are given in the following Tables 27 to 35.

The ESS for the remaining scenarios are given in Tables 36 to 39.

Detailed results on the power of the indirect comparison depending on the planned power of head-to-head trials are shown in Table 40 (AB and CB trials are planned for the same level) and Table 41 (trial CB has a planned power of 80%, the power of the AB trial is varied).

Table 27: *Simulation results for scenario II. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial AB and CB do not include an interaction for the effect modification (setting 2) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.114	0.901	0.213	0.414	0.116	0.902	0.213	0.416	0.114	0.901	0.213	0.414
TTE	similar	moderate	0.085	0.885	0.134	0.219	0.086	0.886	0.134	0.219	0.085	0.885	0.134	0.219
TTE	similar	low	0.061	0.938	0.049	0.137	0.061	0.936	0.049	0.137	0.061	0.938	0.049	0.137
TTE	similar	no	0.053	0.947	-0.014	0.117	0.053	0.947	-0.014	0.117	0.053	0.947	-0.014	0.117
TTE	diff	high	0.247	0.942	0.057	0.278	0.151	0.880	0.258	0.630	0.130	0.896	0.205	0.424
TTE	diff	moderate	0.192	0.939	0.054	0.163	0.078	0.900	0.155	0.292	0.084	0.882	0.138	0.227
TTE	diff	low	0.124	0.946	-0.006	0.119	0.062	0.932	0.056	0.179	0.064	0.936	0.048	0.141
TTE	diff	no	0.093	0.907	-0.058	0.123	0.058	0.942	-0.009	0.148	0.055	0.945	-0.013	0.120
binary	similar	high	0.200	0.943	0.050	0.388	0.200	0.942	0.050	0.390	0.200	0.943	0.050	0.388
binary	similar	moderate	0.107	0.931	0.078	0.229	0.107	0.931	0.078	0.229	0.107	0.931	0.078	0.229
binary	similar	low	0.072	0.944	0.003	0.158	0.071	0.944	0.003	0.158	0.072	0.944	0.003	0.158
binary	similar	no	0.049	0.951	-0.019	0.144	0.049	0.951	-0.019	0.144	0.049	0.951	-0.019	0.144
binary	diff	high	0.315	0.945	-0.085	0.399	0.175	0.922	0.012	0.576	0.156	0.935	0.084	0.432
binary	diff	moderate	0.180	0.947	0.006	0.216	0.095	0.933	0.071	0.314	0.096	0.925	0.092	0.249
binary	diff	low	0.098	0.941	-0.036	0.160	0.065	0.946	0.005	0.207	0.059	0.949	0.017	0.165
binary	diff	no	0.064	0.936	-0.053	0.151	0.055	0.945	-0.017	0.186	0.050	0.950	-0.010	0.150

Table 28: *Simulation results for scenario III. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial AB and CB do not include an interaction for the effect modification (setting 2) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.121	0.900	0.202	0.381	0.121	0.899	0.202	0.383	0.121	0.900	0.202	0.381
TTE	similar	moderate	0.090	0.874	0.133	0.216	0.091	0.875	0.133	0.216	0.090	0.874	0.133	0.216
TTE	similar	low	0.060	0.940	0.046	0.133	0.060	0.939	0.046	0.133	0.060	0.940	0.046	0.133
TTE	similar	no	0.056	0.944	-0.014	0.114	0.056	0.944	-0.014	0.114	0.056	0.944	-0.014	0.114
TTE	diff	high	0.270	0.946	0.047	0.257	0.148	0.886	0.246	0.576	0.132	0.901	0.195	0.393
TTE	diff	moderate	0.208	0.942	0.050	0.157	0.089	0.896	0.145	0.276	0.091	0.884	0.132	0.22
TTE	diff	low	0.124	0.948	-0.006	0.116	0.060	0.935	0.056	0.172	0.063	0.933	0.049	0.139
TTE	diff	no	0.092	0.908	-0.057	0.120	0.056	0.944	-0.007	0.142	0.056	0.944	-0.012	0.117
binary	similar	high	0.209	0.945	0.038	0.389	0.209	0.945	0.038	0.390	0.209	0.945	0.038	0.389
binary	similar	moderate	0.122	0.938	0.067	0.227	0.121	0.938	0.067	0.227	0.122	0.938	0.067	0.227
binary	similar	low	0.066	0.952	0.004	0.156	0.067	0.951	0.004	0.156	0.066	0.952	0.004	0.156
binary	similar	no	0.052	0.948	-0.022	0.147	0.052	0.948	-0.022	0.147	0.052	0.948	-0.022	0.147
binary	diff	high	0.321	0.946	-0.090	0.393	0.173	0.926	0.009	0.566	0.163	0.937	0.081	0.424
binary	diff	moderate	0.184	0.950	0.008	0.216	0.094	0.938	0.069	0.309	0.095	0.927	0.094	0.250
binary	diff	low	0.098	0.941	-0.037	0.160	0.060	0.948	0.004	0.203	0.060	0.947	0.015	0.164
binary	diff	no	0.071	0.929	-0.053	0.153	0.055	0.945	-0.021	0.184	0.052	0.948	-0.012	0.150

Table 29: *Simulation results for scenario IV. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial AB and CB do not include an interaction for the effect modification (setting 2) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.550	0.852	-0.171	0.271	0.550	0.852	-0.171	0.273	0.550	0.852	-0.171	0.271
TTE	similar	moderate	0.526	0.881	-0.096	0.168	0.524	0.880	-0.096	0.168	0.526	0.881	-0.096	0.168
TTE	similar	low	0.377	0.822	-0.111	0.155	0.378	0.824	-0.111	0.155	0.377	0.822	-0.111	0.155
TTE	similar	no	0.301	0.699	-0.151	0.181	0.301	0.699	-0.151	0.181	0.301	0.699	-0.151	0.181
TTE	diff	high	0.615	0.837	-0.176	0.260	0.451	0.808	-0.151	0.362	0.565	0.832	-0.174	0.275
TTE	diff	moderate	0.567	0.871	-0.097	0.163	0.370	0.891	-0.092	0.210	0.523	0.875	-0.096	0.171
TTE	diff	low	0.412	0.808	-0.112	0.154	0.266	0.863	-0.107	0.178	0.376	0.822	-0.112	0.158
TTE	diff	no	0.323	0.677	-0.152	0.181	0.224	0.776	-0.149	0.194	0.298	0.702	-0.152	0.182
binary	similar	high	0.637	0.845	-0.363	0.519	0.634	0.845	-0.363	0.520	0.637	0.845	-0.363	0.519
binary	similar	moderate	0.398	0.903	-0.132	0.250	0.398	0.903	-0.132	0.250	0.398	0.903	-0.132	0.250
binary	similar	low	0.225	0.867	-0.129	0.202	0.225	0.867	-0.129	0.202	0.225	0.867	-0.129	0.202
binary	similar	no	0.148	0.852	-0.126	0.192	0.148	0.852	-0.126	0.192	0.148	0.852	-0.126	0.192
binary	diff	high	0.629	0.846	-0.376	0.535	0.435	0.845	-0.432	0.720	0.566	0.849	-0.387	0.567
binary	diff	moderate	0.400	0.902	-0.134	0.252	0.251	0.911	-0.140	0.335	0.362	0.905	-0.135	0.266
binary	diff	low	0.229	0.871	-0.129	0.201	0.153	0.901	-0.129	0.243	0.208	0.878	-0.128	0.207
binary	diff	no	0.148	0.852	-0.128	0.192	0.115	0.885	-0.129	0.225	0.140	0.860	-0.128	0.197

Table 30: *Simulation results for scenario V. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial AB and CB do not include an interaction for the effect modification (setting 2) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.580	0.846	-0.168	0.260	0.579	0.848	-0.168	0.260	0.580	0.846	-0.168	0.260
TTE	similar	moderate	0.552	0.879	-0.095	0.164	0.552	0.880	-0.095	0.164	0.552	0.879	-0.095	0.164
TTE	similar	low	0.391	0.822	-0.110	0.153	0.391	0.822	-0.110	0.153	0.391	0.822	-0.110	0.153
TTE	similar	no	0.311	0.689	-0.150	0.180	0.310	0.690	-0.150	0.180	0.311	0.689	-0.150	0.180
TTE	diff	high	0.631	0.836	-0.169	0.251	0.470	0.818	-0.150	0.333	0.576	0.833	-0.167	0.264
TTE	diff	moderate	0.586	0.878	-0.095	0.157	0.385	0.896	-0.089	0.199	0.533	0.885	-0.094	0.164
TTE	diff	low	0.420	0.803	-0.112	0.152	0.281	0.859	-0.108	0.173	0.388	0.816	-0.112	0.156
TTE	diff	no	0.331	0.669	-0.151	0.180	0.229	0.771	-0.149	0.192	0.305	0.695	-0.150	0.182
binary	similar	high	0.644	0.848	-0.367	0.520	0.641	0.849	-0.367	0.521	0.644	0.848	-0.367	0.520
binary	similar	moderate	0.405	0.908	-0.135	0.250	0.405	0.908	-0.135	0.251	0.405	0.908	-0.135	0.250
binary	similar	low	0.227	0.869	-0.131	0.204	0.226	0.869	-0.131	0.204	0.227	0.869	-0.131	0.204
binary	similar	no	0.143	0.857	-0.129	0.193	0.143	0.857	-0.129	0.193	0.143	0.857	-0.129	0.193
binary	diff	high	0.624	0.847	-0.368	0.527	0.422	0.848	-0.413	0.697	0.560	0.856	-0.372	0.554
binary	diff	moderate	0.408	0.906	-0.134	0.252	0.258	0.910	-0.139	0.333	0.364	0.906	-0.135	0.265
binary	diff	low	0.224	0.867	-0.129	0.200	0.159	0.899	-0.128	0.240	0.207	0.875	-0.129	0.207
binary	diff	no	0.147	0.853	-0.131	0.192	0.113	0.887	-0.131	0.224	0.140	0.860	-0.130	0.197

Table 31: *Simulation results for scenario I. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The regression models for estimating the treatment effect in trial CB are not adjusted for confounders (setting 3) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.385	0.919	-0.090	0.263	0.383	0.918	-0.090	0.265
TTE	similar	moderate	0.206	0.945	0.032	0.166	0.203	0.946	0.032	0.166
TTE	similar	low	0.066	0.943	0.039	0.136	0.066	0.944	0.039	0.136
TTE	similar	no	0.050	0.950	0.005	0.122	0.050	0.950	0.005	0.122
TTE	diff	high	0.430	0.913	-0.091	0.248	0.338	0.870	-0.060	0.397
TTE	diff	moderate	0.227	0.941	0.031	0.160	0.168	0.935	0.039	0.230
TTE	diff	low	0.066	0.945	0.036	0.129	0.067	0.942	0.041	0.172
TTE	diff	no	0.052	0.948	0.007	0.118	0.053	0.947	0.012	0.152
binary	similar	high	0.487	0.912	-0.228	0.438	0.485	0.913	-0.228	0.439
binary	similar	moderate	0.202	0.948	-0.004	0.212	0.201	0.948	-0.004	0.212
binary	similar	low	0.065	0.951	-0.001	0.154	0.066	0.951	-0.001	0.154
binary	similar	no	0.052	0.948	0.000	0.142	0.052	0.948	0.000	0.142
binary	diff	high	0.475	0.907	-0.241	0.457	0.346	0.881	-0.303	0.655
binary	diff	moderate	0.193	0.947	-0.001	0.214	0.140	0.940	-0.004	0.306
binary	diff	low	0.066	0.954	0.000	0.153	0.067	0.946	-0.002	0.207
binary	diff	no	0.050	0.950	0.003	0.142	0.050	0.950	0.002	0.183

Table 32: *Simulation results for scenario II. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial CB are not adjusted for confounders and effect modifiers (setting 3) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.064	0.758	0.628	0.943	0.067	0.760	0.628	0.946	0.064	0.758	0.628	0.943
TTE	similar	moderate	0.061	0.719	0.326	0.419	0.061	0.720	0.326	0.419	0.061	0.719	0.326	0.419
TTE	similar	low	0.077	0.823	0.172	0.245	0.078	0.825	0.172	0.245	0.077	0.823	0.172	0.245
TTE	similar	no	0.086	0.914	0.087	0.173	0.086	0.914	0.087	0.173	0.086	0.914	0.087	0.173
TTE	diff	high	0.073	0.801	0.697	1.157	0.125	0.790	0.873	1.624	0.083	0.790	0.707	1.181
TTE	diff	moderate	0.062	0.770	0.346	0.467	0.074	0.820	0.385	0.593	0.064	0.769	0.347	0.469
TTE	diff	low	0.074	0.854	0.173	0.264	0.072	0.886	0.185	0.327	0.074	0.853	0.173	0.264
TTE	diff	no	0.079	0.921	0.088	0.191	0.072	0.928	0.095	0.240	0.080	0.920	0.088	0.191
binary	similar	high	0.049	0.816	0.593	0.833	0.200	0.943	0.050	0.388	0.200	0.943	0.050	0.388
binary	similar	moderate	0.065	0.799	0.345	0.468	0.107	0.931	0.078	0.229	0.107	0.931	0.078	0.229
binary	similar	low	0.082	0.872	0.168	0.273	0.072	0.944	0.003	0.158	0.072	0.944	0.003	0.158
binary	similar	no	0.092	0.908	0.112	0.219	0.049	0.951	-0.019	0.144	0.049	0.951	-0.019	0.144
binary	diff	high	0.052	0.861	0.579	1.060	0.156	0.935	0.084	0.432	0.156	0.935	0.084	0.432
binary	diff	moderate	0.063	0.852	0.340	0.514	0.096	0.925	0.092	0.249	0.096	0.925	0.092	0.249
binary	diff	low	0.069	0.897	0.170	0.303	0.059	0.949	0.017	0.165	0.059	0.949	0.017	0.165
binary	diff	no	0.081	0.919	0.112	0.249	0.050	0.950	-0.010	0.150	0.050	0.950	-0.010	0.150

Table 33: *Simulation results for scenario III. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial CB are not adjusted for confounders and effect modifiers (setting 3) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.063	0.746	0.601	0.868	0.063	0.746	0.601	0.868	0.064	0.758	0.628	0.943
TTE	similar	moderate	0.064	0.709	0.325	0.416	0.064	0.709	0.325	0.416	0.061	0.719	0.326	0.419
TTE	similar	low	0.076	0.826	0.170	0.240	0.076	0.826	0.170	0.240	0.077	0.823	0.172	0.245
TTE	similar	no	0.088	0.912	0.088	0.171	0.088	0.912	0.088	0.171	0.086	0.914	0.087	0.173
TTE	diff	high	0.071	0.793	0.667	1.076	0.076	0.788	0.672	1.087	0.083	0.790	0.707	1.181
TTE	diff	moderate	0.066	0.763	0.336	0.456	0.067	0.761	0.336	0.457	0.064	0.769	0.347	0.469
TTE	diff	low	0.077	0.847	0.176	0.263	0.076	0.845	0.176	0.263	0.074	0.853	0.173	0.264
TTE	diff	no	0.083	0.917	0.091	0.189	0.083	0.917	0.091	0.189	0.080	0.920	0.088	0.191
binary	similar	high	0.055	0.817	0.600	0.905	0.209	0.945	0.038	0.390	0.209	0.945	0.038	0.389
binary	similar	moderate	0.065	0.809	0.340	0.467	0.121	0.938	0.067	0.227	0.122	0.938	0.067	0.227
binary	similar	low	0.083	0.876	0.174	0.277	0.067	0.951	0.004	0.156	0.066	0.952	0.004	0.156
binary	similar	no	0.091	0.909	0.111	0.222	0.052	0.948	-0.022	0.147	0.052	0.948	-0.022	0.147
binary	diff	high	0.052	0.857	0.614	1.115	0.173	0.926	0.009	0.566	0.163	0.937	0.081	0.424
binary	diff	moderate	0.063	0.843	0.348	0.523	0.094	0.938	0.069	0.309	0.095	0.927	0.094	0.250
binary	diff	low	0.070	0.899	0.171	0.306	0.060	0.948	0.004	0.203	0.060	0.947	0.015	0.164
binary	diff	no	0.076	0.924	0.111	0.248	0.055	0.945	-0.021	0.184	0.052	0.948	-0.012	0.150

Table 34: *Simulation results for scenario IV. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial CB are not adjusted for confounders and effect modifiers (setting 3) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.550	0.852	-0.171	0.271	0.550	0.852	-0.171	0.273	0.550	0.852	-0.171	0.271
TTE	similar	moderate	0.526	0.881	-0.096	0.168	0.524	0.880	-0.096	0.168	0.526	0.881	-0.096	0.168
TTE	similar	low	0.377	0.822	-0.111	0.155	0.378	0.824	-0.111	0.155	0.377	0.822	-0.111	0.155
TTE	similar	no	0.301	0.699	-0.151	0.181	0.301	0.699	-0.151	0.181	0.301	0.699	-0.151	0.181
TTE	diff	high	0.615	0.837	-0.176	0.260	0.451	0.808	-0.151	0.362	0.565	0.832	-0.174	0.275
TTE	diff	moderate	0.567	0.871	-0.097	0.163	0.370	0.891	-0.092	0.210	0.523	0.875	-0.096	0.171
TTE	diff	low	0.412	0.808	-0.112	0.154	0.266	0.863	-0.107	0.178	0.376	0.822	-0.112	0.158
TTE	diff	no	0.323	0.677	-0.152	0.181	0.224	0.776	-0.149	0.194	0.298	0.702	-0.152	0.182
binary	similar	high	0.637	0.845	-0.363	0.519	0.634	0.845	-0.363	0.520	0.637	0.845	-0.363	0.519
binary	similar	moderate	0.398	0.903	-0.132	0.250	0.398	0.903	-0.132	0.250	0.398	0.903	-0.132	0.250
binary	similar	low	0.225	0.867	-0.129	0.202	0.225	0.867	-0.129	0.202	0.225	0.867	-0.129	0.202
binary	similar	no	0.148	0.852	-0.126	0.192	0.148	0.852	-0.126	0.192	0.148	0.852	-0.126	0.192
binary	diff	high	0.629	0.846	-0.376	0.535	0.435	0.845	-0.432	0.720	0.566	0.849	-0.387	0.567
binary	diff	moderate	0.400	0.902	-0.134	0.252	0.251	0.911	-0.140	0.335	0.362	0.905	-0.135	0.266
binary	diff	low	0.229	0.871	-0.129	0.201	0.153	0.901	-0.129	0.243	0.208	0.878	-0.128	0.207
binary	diff	no	0.148	0.852	-0.128	0.192	0.115	0.885	-0.129	0.225	0.140	0.860	-0.128	0.197

Table 35: *Simulation results for scenario V. The rows include the two endpoints, similar and different population distributions, and the considered effect sizes, the columns the evaluation measures for these scenarios power/type I error, coverage (Cov), bias, and root mean squared error (RMSE). Accordingly, for no effect, the column Power shows the type I error rate (α). The upper part of the following tables contains the results for considering all confounders within MAIC and the lower part those for only effect modifiers are included in MAIC. The regression models for estimating the treatment effect in trial CB are not adjusted for confounders and effect modifiers (setting 3) (adapted from Weber et al. (2020a)).*

EP	Pat Char	Effect	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE	Power (α)	Cov	Bias	RMSE
TTE	similar	high	0.655	0.794	-0.207	0.277	0.650	0.791	-0.208	0.277	0.655	0.794	-0.207	0.277
TTE	similar	moderate	0.677	0.778	-0.156	0.202	0.674	0.777	-0.156	0.203	0.677	0.778	-0.156	0.202
TTE	similar	low	0.555	0.655	-0.181	0.212	0.555	0.656	-0.181	0.212	0.555	0.655	-0.181	0.212
TTE	similar	no	0.484	0.516	-0.224	0.248	0.483	0.517	-0.224	0.248	0.484	0.516	-0.224	0.248
TTE	diff	high	0.706	0.772	-0.208	0.270	0.531	0.782	-0.190	0.334	0.647	0.778	-0.206	0.280
TTE	diff	moderate	0.712	0.766	-0.156	0.198	0.500	0.826	-0.150	0.225	0.662	0.783	-0.155	0.202
TTE	diff	low	0.589	0.634	-0.183	0.212	0.433	0.734	-0.179	0.224	0.557	0.656	-0.183	0.214
TTE	diff	no	0.514	0.486	-0.225	0.248	0.388	0.612	-0.223	0.255	0.483	0.517	-0.225	0.249
binary	similar	high	0.644	0.848	-0.367	0.520	0.644	0.848	-0.367	0.520	0.637	0.845	-0.363	0.519
binary	similar	moderate	0.405	0.908	-0.135	0.250	0.405	0.908	-0.135	0.250	0.398	0.903	-0.132	0.250
binary	similar	low	0.227	0.869	-0.131	0.204	0.227	0.869	-0.131	0.204	0.225	0.867	-0.129	0.202
binary	similar	no	0.143	0.857	-0.129	0.193	0.143	0.857	-0.129	0.193	0.148	0.852	-0.126	0.192
binary	diff	high	0.624	0.847	-0.368	0.527	0.560	0.856	-0.372	0.554	0.566	0.849	-0.387	0.567
binary	diff	moderate	0.408	0.906	-0.134	0.252	0.364	0.906	-0.135	0.265	0.362	0.905	-0.135	0.266
binary	diff	low	0.224	0.867	-0.129	0.200	0.207	0.875	-0.129	0.207	0.208	0.878	-0.128	0.207
binary	diff	no	0.147	0.853	-0.131	0.192	0.140	0.860	-0.130	0.197	0.140	0.860	-0.128	0.197

Table 36: *Mean Effective Sample Size (ESS) for scenario II considering a time-to-event (TTE) as well as binary endpoints. The ESS is given for MAIC is adjusted for all confounders (Conf.) and for the case where it is only adjusted for effect modifiers (EM) (adapted from Weber et al. (2020a)).*

Endpoint	Population	Effect	Sample Size	ESS all Conf.	ESS only EM
TTE	similar	high	94	92.903	94.000
TTE	similar	moderate	396	394.759	396.000
TTE	similar	low	1044	1042.348	1044.000
TTE	similar	no	1578	1575.976	1578.000
TTE	different	high	94	38.927	78.638
TTE	different	moderate	396	161.868	332.960
TTE	different	low	1044	423.169	876.646
TTE	different	no	1578	639.260	1326.495
binary	similar	high	192	190.860	192.000
binary	similar	moderate	648	646.708	648.000
binary	similar	low	1600	1598.201	1600.000
binary	similar	no	2176	2173.999	2176.000
binary	different	high	192	80.000	161.920
binary	different	moderate	648	264.195	545.278
binary	different	low	1600	646.039	1344.015
binary	different	no	2176	877.268	1828.161

Table 37: *Mean Effective Sample Size (ESS) for scenario III considering a time-to-event (TTE) as well as binary endpoint. The ESS is given for MAIC is adjusted for all confounders and for the case where it is only adjusted for effect modifiers (EM) (adapted from Weber et al. (2020a)).*

Endpoint	Population	Effect	Sample Size	ESS all Conf.	ESS only EM
TTE	similar	high	94	92.873	94.000
TTE	similar	moderate	396	394.721	396.000
TTE	similar	low	1044	1042.343	1044.000
TTE	similar	no	1578	1576.018	1578.000
TTE	different	high	94	40.584	78.638
TTE	different	moderate	396	169.355	332.960
TTE	different	low	1044	443.025	876.646
TTE	different	no	1578	669.056	1326.495
binary	similar	high	192	190.923	192.000
binary	similar	moderate	648	646.681	648.000
binary	similar	low	1600	1598.250	1600.000
binary	similar	no	2176	2174.023	2176.000
binary	different	high	192	83.314	161.920
binary	different	moderate	648	276.291	545.278
binary	different	low	1600	676.922	1344.015
binary	different	no	2176	919.657	1828.161

Table 38: *Mean Effective Sample Size (ESS) for scenario IV considering a time-to-event (TTE) as well as binary endpoint. The ESS is given for MAIC is adjusted for all confounders and for the case where it is only adjusted for effect modifiers (EM) (adapted from Weber et al. (2020a)).*

Endpoint	Population	Effect	Sample Size	ESS all Conf.	ESS only EM
TTE	similar	high	94	92.898	94.000
TTE	similar	moderate	396	394.726	396.000
TTE	similar	low	1044	1042.335	1044.000
TTE	similar	no	1578	1575.942	1578.000
TTE	different	high	94	38.847	78.638
TTE	different	moderate	396	161.790	332.960
TTE	different	low	1044	423.595	876.646
TTE	different	no	1578	638.925	1326.495
binary	similar	high	192	190.867	192.000
binary	similar	moderate	648	646.692	648.000
binary	similar	low	1600	1598.270	1600.000
binary	similar	no	2176	2174.004	2176.000
binary	different	high	192	79.816	161.920
binary	different	moderate	648	264.143	545.278
binary	different	low	1600	646.770	1344.015
binary	different	no	2176	877.447	1828.161

Table 39: *Mean Effective Sample Size (ESS) for scenario V considering a time-to-event (TTE) as well as binary endpoint. The ESS is given for MAIC is adjusted for all confounders and for the case where it is only adjusted for effect modifiers (EM) (adapted from Weber et al. (2020a)).*

Endpoint	Population	Effect	Sample Size	ESS all Conf.	ESS only EM
TTE	similar	high	94	92.884	94.000
TTE	similar	moderate	396	394.747	396.000
TTE	similar	low	1044	1042.346	1044.000
TTE	similar	no	1578	1576.017	1578.000
TTE	different	high	94	40.588	78.638
TTE	different	moderate	396	169.266	332.960
TTE	different	low	1044	443.368	876.646
TTE	different	no	1578	668.927	1326.495
binary	similar	high	192	190.911	192.000
binary	similar	moderate	648	646.720	648.000
binary	similar	low	1600	1598.301	1600.000
binary	similar	no	2176	2173.999	2176.000
binary	different	high	192	83.332	161.920
binary	different	moderate	648	276.422	545.278
binary	different	low	1600	676.120	1344.015
binary	different	no	2176	919.719	1828.161

Table 40: *Power for indirect comparison for individual planned power is the same in aggregated data trial and individual patient data trial set to 90%, 95%, or 99%. The results are given for scenario I having similar patient populations without interaction. They demonstrate the influence of the power of the individual trials on the power of the indirect comparison. The power is given for both endpoints, in case of no effect the column shows the type I error rate (α) (adapted from Weber et al. (2020a)).*

		Binary		Time-to-event	
Effect	Planned Power	Bucher	MAIC	Bucher	MAIC
high	90%	0.317	0.316	0.338	0.337
moderate	90%	0.200	0.201	0.170	0.167
low	90%	0.081	0.080	0.075	0.075
no	90%	0.054	0.054	0.054	0.055
high	95%	0.371	0.371	0.436	0.432
moderate	95%	0.229	0.229	0.242	0.241
low	95%	0.087	0.088	0.073	0.072
no	95%	0.049	0.050	0.051	0.052
high	99%	0.497	0.498	0.549	0.549
moderate	99%	0.308	0.308	0.310	0.309
low	99%	0.101	0.100	0.077	0.077
no	99%	0.055	0.055	0.055	0.055

Table 41: *Power for indirect comparison under consideration of trials with individual planned power of 80% in the aggregated data trial. The results are given for scenario I having similar patient populations without interaction. They demonstrate the influence of the power of the individual trials on the power of the indirect comparison. The power is given for both endpoints, in case of no effect the column shows the type I error rate (α) (adapted from Weber et al. (2020a)).*

	Planned Power	Binary		Time-to-event	
Effect	IPD trial (AB)	Bucher	MAIC	Bucher	MAIC
high	80%	0.253	0.253	0.243	0.249
moderate	80%	0.160	0.162	0.173	0.173
low	80%	0.071	0.071	0.065	0.064
no	80%	0.046	0.046	0.060	0.058
high	90%	0.306	0.305	0.346	0.348
moderate	90%	0.188	0.187	0.179	0.183
low	90%	0.076	0.077	0.069	0.070
no	90%	0.048	0.048	0.047	0.048
high	95%	0.363	0.363	0.414	0.411
moderate	95%	0.215	0.216	0.225	0.226
low	95%	0.081	0.081	0.074	0.074
no	95%	0.053	0.053	0.041	0.041
high	99%	0.453	0.453	0.531	0.526
moderate	99%	0.267	0.266	0.234	0.233
low	99%	0.089	0.089	0.060	0.060
no	99%	0.049	0.050	0.067	0.066

Comment: Parts of the following Section A.3 are already included in the submitted manuscript Weber et al. (2020b). The manuscript has been written by myself, but may contain comments and corrections from the co-authors.

A.3 Synthesis of Evidence - Inclusion of multiple studies in indirect comparisons

This section contains the detailed results of the simulation study investigating the inclusion of multiple studies in indirect comparisons.

Three different settings are evaluated:

- all regression models are adjusted for confounders, meta-analysis is performed by a random effects model, and a variance value of $\sigma = 0.2$ is assumed in the generation of the true treatment effect (Tables 42 to 53),
- all regression models are adjusted for confounders, meta-analysis is performed by a fixed effects model, and a variance value of $\sigma = 0.2$ is assumed in the generation of the true treatment effect (Tables 54 to 59),
- all regression models are adjusted for confounders, meta-analysis is performed by a random effects model, and a variance value of $\sigma = 0.4$ is assumed in the generation of the true treatment effect (Tables 60 to 65).

Table 42: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario I and II. Power and coverage are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model.*

Popu-	Effect	Trials	True	Mean Effect				Power				Coverage			
lation	Size		Effect	Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.713	-0.713	-0.742	-0.742	0.808	0.804	0.444	0.444	0.881	0.885	0.951	0.951
similar	high	4	-0.69	-0.743	-0.743	-0.777	-0.777	0.846	0.843	0.459	0.459	0.878	0.878	0.953	0.953
similar	high	10	-0.69	-0.726	-0.726	-0.756	-0.756	0.842	0.845	0.445	0.445	0.869	0.873	0.949	0.949
similar	moderate	2	-0.22	-0.234	-0.234	-0.231	-0.231	0.547	0.551	0.250	0.250	0.902	0.902	0.952	0.952
similar	moderate	4	-0.22	-0.232	-0.232	-0.234	-0.234	0.571	0.567	0.267	0.267	0.883	0.879	0.948	0.948
similar	moderate	10	-0.22	-0.232	-0.232	-0.234	-0.234	0.512	0.610	0.278	0.278	0.879	0.879	0.958	0.958
similar	low	2	-0.05	-0.053	-0.053	-0.052	-0.052	0.150	0.152	0.091	0.089	0.905	0.903	0.940	0.942
similar	low	4	-0.05	-0.051	-0.051	-0.052	-0.052	0.184	0.186	0.090	0.090	0.879	0.878	0.943	0.943
similar	low	10	-0.05	-0.052	-0.052	-0.051	-0.051	0.209	0.207	0.087	0.087	0.875	0.878	0.946	0.946
similar	no	2	0	-0.003	-0.003	-0.002	-0.002	0.094	0.094	0.069	0.065	0.907	0.906	0.932	0.936
similar	no	4	0	0.002	0.002	0.002	0.002	0.109	0.110	0.060	0.060	0.892	0.890	0.940	0.940
similar	no	10	0	-0.001	-0.001	-0.001	-0.001	0.130	0.130	0.065	0.065	0.871	0.871	0.936	0.936
different	high	2	-0.69	-0.734	-0.756	-0.766	-0.766	0.822	0.635	0.456	0.456	0.866	0.821	0.947	0.947
different	high	4	-0.69	-0.725	-0.759	-0.746	-0.746	0.833	0.639	0.446	0.446	0.870	0.844	0.956	0.956
different	high	10	-0.69	-0.729	-0.755	-0.766	-0.766	0.850	0.6265	0.458	0.458	0.866	0.850	0.940	0.940
different	moderate	2	-0.22	-0.234	-0.239	-0.237	-0.237	0.538	0.3915	0.257	0.256	0.881	0.868	0.945	0.945
different	moderate	4	-0.22	-0.231	-0.232	-0.233	-0.233	0.562	0.388	0.273	0.273	0.883	0.876	0.950	0.950
different	moderate	10	-0.22	-0.231	-0.230	-0.236	-0.236	0.593	0.4015	0.282	0.282	0.864	0.851	0.933	0.933
different	low	2	-0.05	-0.050	-0.054	-0.051	-0.051	0.146	0.146	0.081	0.077	0.905	0.895	0.944	0.948
different	low	4	-0.05	-0.051	-0.056	-0.050	-0.050	0.159	0.174	0.079	0.079	0.885	0.881	0.947	0.947
different	low	10	-0.05	-0.045	-0.047	-0.046	-0.046	0.173	0.159	0.076	0.076	0.881	0.885	0.945	0.945
different	no	2	0	0.000	-0.002	-0.001	-0.001	0.099	0.106	0.059	0.054	0.901	0.895	0.941	0.947
different	no	4	0	0.002	0.003	0.002	0.002	0.099	0.1205	0.049	0.049	0.902	0.879	0.951	0.951
different	no	10	0	-0.005	-0.008	-0.006	-0.006	0.129	0.129	0.061	0.061	0.872	0.872	0.939	0.939

Table 43: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario I and II. The mean effect, bias, and the root mean squared error (RMSE) are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model.*

Population	Effect Size	Trials	Bias				MC Error for Bias				RMSE			
			Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.023	-0.023	-0.052	-0.052	0.007	0.007	0.010	0.010	0.300	0.302	0.433	0.433
similar	high	4	-0.053	-0.053	-0.087	-0.087	0.007	0.007	0.010	0.010	0.303	0.305	0.433	0.433
similar	high	10	-0.036	-0.036	-0.066	-0.066	0.007	0.007	0.010	0.010	0.301	0.302	0.436	0.436
similar	moderate	2	-0.014	-0.014	-0.011	-0.011	0.003	0.003	0.004	0.004	0.136	0.136	0.178	0.178
similar	moderate	4	-0.012	-0.012	-0.014	-0.014	0.003	0.003	0.004	0.004	0.135	0.135	0.178	0.178
similar	moderate	10	-0.012	-0.012	-0.014	-0.014	0.003	0.003	0.004	0.004	0.129	0.129	0.171	0.171
similar	low	2	-0.003	-0.003	-0.002	-0.002	0.002	0.002	0.003	0.003	0.094	0.094	0.117	0.117
similar	low	4	-0.001	-0.001	-0.002	-0.002	0.002	0.002	0.002	0.002	0.085	0.086	0.109	0.109
similar	low	10	-0.002	-0.002	-0.001	-0.001	0.002	0.002	0.002	0.002	0.078	0.078	0.103	0.103
similar	no	2	-0.003	-0.003	-0.002	-0.002	0.002	0.002	0.002	0.002	0.085	0.085	0.101	0.101
similar	no	4	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.073	0.073	0.091	0.091
similar	no	10	-0.001	-0.001	-0.001	-0.001	0.002	0.002	0.002	0.002	0.066	0.066	0.086	0.086
different	high	2	-0.044	-0.066	-0.076	-0.076	0.007	0.010	0.010	0.010	0.316	0.46	0.452	0.452
different	high	4	-0.035	-0.069	-0.056	-0.056	0.007	0.010	0.010	0.010	0.303	0.437	0.428	0.428
different	high	10	-0.039	-0.065	-0.076	-0.076	0.007	0.0098	0.010	0.010	0.302	0.442	0.438	0.438
different	moderate	2	-0.014	-0.019	-0.017	-0.017	0.003	0.005	0.004	0.004	0.143	0.200	0.184	0.184
different	moderate	4	-0.011	-0.012	-0.013	-0.013	0.003	0.004	0.004	0.004	0.135	0.191	0.178	0.178
different	moderate	10	-0.011	-0.010	-0.016	-0.016	0.003	0.004	0.004	0.004	0.133	0.195	0.182	0.182
different	low	2	0.000	-0.004	-0.001	-0.001	0.002	0.003	0.003	0.003	0.097	0.125	0.118	0.118
different	low	4	-0.001	-0.006	0.000	0.000	0.002	0.003	0.002	0.002	0.085	0.117	0.108	0.108
different	low	10	0.005	0.003	0.004	0.004	0.002	0.003	0.002	0.002	0.077	0.111	0.102	0.102
different	no	2	0.000	-0.002	-0.001	-0.001	0.002	0.002	0.002	0.002	0.087	0.108	0.103	0.103
different	no	4	0.002	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.073	0.098	0.090	0.090
different	no	10	-0.005	-0.008	-0.006	-0.006	0.002	0.002	0.002	0.002	0.067	0.096	0.086	0.086

Table 44: *One AgD trial (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. Power and coverage are presented for approach B.1 (pooled IPD) and B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. Results for scenario I and II are presented. All meta-analyses are based on a random effects model.*

Popu-	Effect	Trials	True	Mean Effect				Power				Coverage			
lation	Size		Effect	Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.711	-0.712	-0.719	-0.732	0.702	0.701	0.691	0.587	0.952	0.951	0.952	0.968
similar	high	4	-0.69	-0.720	-0.720	-0.734	-0.743	0.932	0.930	0.927	0.887	0.948	0.944	0.946	0.957
similar	high	10	-0.69	-0.698	-0.698	-0.720	-0.722	0.998	0.998	0.998	0.997	0.949	0.946	0.941	0.944
similar	moderate	2	-0.22	-0.237	-0.238	-0.238	-0.239	0.382	0.384	0.388	0.338	0.951	0.952	0.948	0.958
similar	moderate	4	-0.22	-0.232	-0.231	-0.234	-0.235	0.520	0.5175	0.526	0.510	0.949	0.948	0.948	0.953
similar	moderate	10	-0.22	-0.230	-0.230	-0.234	-0.234	0.678	0.678	0.692	0.691	0.941	0.941	0.939	0.940
similar	low	2	-0.05	-0.056	-0.056	-0.056	-0.056	0.092	0.093	0.091	0.082	0.946	0.946	0.947	0.952
similar	low	4	-0.05	-0.051	-0.051	-0.052	-0.052	0.099	0.098	0.100	0.100	0.935	0.934	0.936	0.937
similar	low	10	-0.05	-0.047	-0.047	-0.048	-0.048	0.098	0.096	0.100	0.100	0.940	0.941	0.942	0.942
similar	no	2	0	-0.002	-0.002	-0.002	-0.002	0.059	0.060	0.062	0.057	0.941	0.940	0.939	0.943
similar	no	4	0	0.003	0.003	0.002	0.002	0.055	0.055	0.055	0.055	0.945	0.945	0.945	0.945
similar	no	10	0	0.002	0.002	0.001	0.001	0.052	0.053	0.053	0.053	0.948	0.948	0.947	0.947
different	high	2	-0.69	-0.729	-0.749	-0.739	-0.750	0.716	0.495	0.700	0.608	0.948	0.918	0.947	0.958
different	high	4	-0.69	-0.714	-0.727	-0.726	-0.734	0.924	0.722	0.919	0.873	0.939	0.937	0.943	0.952
different	high	10	-0.69	-0.704	-0.707	-0.729	-0.731	1.000	0.946	0.999	0.996	0.952	0.943	0.940	0.945
different	moderate	2	-0.22	-0.232	-0.237	-0.234	-0.234	0.367	0.243	0.366	0.319	0.944	0.936	0.943	0.950
different	moderate	4	-0.22	-0.225	-0.226	-0.228	-0.228	0.476	0.327	0.488	0.479	0.951	0.946	0.947	0.950
different	moderate	10	-0.22	-0.231	-0.231	-0.236	-0.236	0.645	0.521	0.658	0.658	0.938	0.939	0.933	0.933
different	low	2	-0.05	-0.049	-0.051	-0.049	-0.049	0.081	0.073	0.081	0.073	0.941	0.941	0.938	0.945
different	low	4	-0.05	-0.051	-0.055	-0.052	-0.052	0.090	0.078	0.091	0.091	0.944	0.951	0.946	0.946
different	low	10	-0.05	-0.048	-0.049	-0.049	-0.049	0.097	0.094	0.098	0.098	0.942	0.943	0.941	0.941
different	no	2	0	-0.002	-0.004	-0.003	-0.003	0.064	0.062	0.063	0.058	0.936	0.939	0.938	0.942
different	no	4	0	-0.001	-0.002	-0.001	-0.001	0.052	0.055	0.052	0.052	0.949	0.946	0.949	0.949
different	no	10	0	0.000	0.000	-0.001	-0.001	0.056	0.058	0.058	0.058	0.944	0.943	0.943	0.943

Table 45: *One AgD trial (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. The mean effect, bias, and the root mean squared error (RMSE) are presented for approach B.1 (pooled IPD) and B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model.*

Population	Effect Size	Trials	Bias				MC Error for Bias				RMSE			
			Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.021	-0.022	-0.029	-0.042	0.007	0.007	0.007	0.007	0.291	0.293	0.299	0.305
similar	high	4	-0.030	-0.030	-0.044	-0.053	0.005	0.005	0.005	0.005	0.215	0.216	0.224	0.228
similar	high	10	-0.008	-0.008	-0.030	-0.032	0.003	0.003	0.004	0.004	0.145	0.146	0.163	0.160
similar	moderate	2	-0.017	-0.018	-0.018	-0.019	0.003	0.003	0.003	0.003	0.142	0.143	0.144	0.144
similar	moderate	4	-0.012	-0.011	-0.014	-0.015	0.003	0.003	0.003	0.003	0.117	0.117	0.118	0.118
similar	moderate	10	-0.010	-0.010	-0.014	-0.014	0.002	0.002	0.002	0.002	0.097	0.098	0.098	0.098
similar	low	2	-0.006	-0.006	-0.006	-0.006	0.002	0.002	0.002	0.002	0.107	0.107	0.107	0.107
similar	low	4	-0.001	-0.001	-0.002	-0.002	0.002	0.002	0.002	0.002	0.096	0.096	0.097	0.097
similar	low	10	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.087	0.087	0.087	0.087
similar	no	2	-0.002	-0.002	-0.002	-0.002	0.002	0.002	0.002	0.002	0.102	0.101	0.102	0.102
similar	no	4	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.090	0.090	0.090	0.090
similar	no	10	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.002	0.083	0.083	0.083	0.083
different	high	2	-0.039	-0.059	-0.049	-0.060	0.007	0.0098	0.007	0.007	0.305	0.444	0.316	0.321
different	high	4	-0.024	-0.037	-0.036	-0.044	0.005	0.007	0.005	0.005	0.217	0.296	0.228	0.230
different	high	10	-0.014	-0.017	-0.039	-0.041	0.003	0.005	0.004	0.004	0.148	0.200	0.178	0.174
different	moderate	2	-0.012	-0.017	-0.014	-0.014	0.003	0.005	0.003	0.003	0.151	0.202	0.153	0.153
different	moderate	4	-0.005	-0.006	-0.008	-0.008	0.003	0.003	0.003	0.003	0.120	0.152	0.121	0.121
different	moderate	10	-0.011	-0.011	-0.016	-0.016	0.002	0.003	0.002	0.002	0.103	0.119	0.104	0.104
different	low	2	0.001	-0.001	0.001	0.001	0.003	0.003	0.003	0.003	0.114	0.135	0.114	0.114
different	low	4	-0.001	-0.005	-0.002	-0.002	0.002	0.003	0.002	0.002	0.100	0.112	0.100	0.100
different	low	10	0.002	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.093	0.099	0.093	0.093
different	no	2	-0.002	-0.004	-0.003	-0.003	0.002	0.003	0.002	0.002	0.107	0.123	0.107	0.107
different	no	4	-0.001	-0.002	-0.001	-0.001	0.002	0.002	0.002	0.002	0.093	0.103	0.093	0.093
different	no	10	0.000	0.000	-0.001	-0.001	0.002	0.002	0.002	0.002	0.090	0.094	0.090	0.090

Table 46: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. Power and coverage are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model.*

Population	Effect Size	Trials	True Effect	Mean Effect						Power						Coverage					
				All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.711	-0.712	-0.710	-0.711	-0.718	-0.722	0.720	0.717	0.720	0.715	0.700	0.687	0.952	0.956	0.952	0.955	0.951	0.955
similar	high	4	-0.69	-0.717	-0.717	-0.717	-0.717	-0.728	-0.728	0.961	0.961	0.961	0.960	0.952	0.952	0.954	0.953	0.954	0.954	0.948	0.948
similar	high	10	-0.69	-0.701	-0.701	-0.700	-0.700	-0.712	-0.711	1.000	1.000	1.000	1.000	1.000	1.000	0.951	0.952	0.952	0.952	0.944	0.944
similar	moderate	2	-0.22	-0.236	-0.236	-0.236	-0.236	-0.237	-0.237	0.442	0.442	0.442	0.439	0.441	0.432	0.953	0.953	0.953	0.954	0.954	0.957
similar	moderate	4	-0.22	-0.230	-0.230	-0.230	-0.230	-0.231	-0.231	0.699	0.700	0.698	0.695	0.695	0.695	0.946	0.943	0.946	0.944	0.945	0.945
similar	moderate	10	-0.22	-0.229	-0.229	-0.229	-0.229	-0.231	-0.231	0.976	0.976	0.976	0.976	0.976	0.976	0.955	0.954	0.955	0.954	0.952	0.952
similar	low	2	-0.05	-0.054	-0.054	-0.054	-0.053	-0.054	-0.054	0.099	0.099	0.099	0.093	0.099	0.099	0.943	0.943	0.942	0.946	0.943	0.944
similar	low	4	-0.05	-0.050	-0.050	-0.049	-0.049	-0.050	-0.050	0.137	0.133	0.139	0.129	0.144	0.144	0.948	0.946	0.948	0.949	0.947	0.947
similar	low	10	-0.05	-0.050	-0.050	-0.050	-0.050	-0.050	-0.050	0.249	0.251	0.249	0.247	0.256	0.256	0.949	0.948	0.949	0.949	0.950	0.950
similar	no	2	0	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	0.057	0.058	0.057	0.055	0.057	0.057	0.944	0.942	0.944	0.946	0.943	0.944
similar	no	4	0	0.002	0.002	0.002	0.002	0.002	0.002	0.063	0.062	0.063	0.058	0.062	0.062	0.938	0.939	0.937	0.943	0.939	0.939
similar	no	10	0	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.053	0.054	0.054	0.053	0.054	0.054	0.948	0.947	0.947	0.948	0.946	0.946
different	high	2	-0.69	-0.730	-0.750	-0.730	-0.750	-0.739	-0.742	0.738	0.503	0.737	0.503	0.711	0.698	0.946	0.919	0.946	0.919	0.945	0.947
different	high	4	-0.69	-0.713	-0.726	-0.712	-0.726	-0.720	-0.720	0.956	0.749	0.956	0.750	0.946	0.946	0.958	0.942	0.957	0.942	0.949	0.949
different	high	10	-0.69	-0.705	-0.708	-0.704	-0.707	-0.718	-0.718	1.000	0.973	1.000	0.972	1.000	1.000	0.952	0.946	0.952	0.946	0.938	0.938
different	moderate	2	-0.22	-0.234	-0.239	-0.233	-0.238	-0.234	-0.235	0.419	0.271	0.418	0.268	0.423	0.415	0.945	0.940	0.945	0.942	0.942	0.945
different	moderate	4	-0.22	-0.229	-0.230	-0.228	-0.230	-0.230	-0.230	0.672	0.428	0.671	0.425	0.676	0.676	0.938	0.946	0.938	0.945	0.939	0.939
different	moderate	10	-0.22	-0.231	-0.230	-0.230	-0.230	-0.233	-0.233	0.971	0.791	0.970	0.789	0.968	0.968	0.943	0.936	0.943	0.936	0.936	0.936
different	low	2	-0.05	-0.050	-0.052	-0.049	-0.051	-0.050	-0.050	0.097	0.077	0.098	0.074	0.099	0.096	0.947	0.944	0.947	0.947	0.947	0.950
different	low	4	-0.05	-0.051	-0.054	-0.050	-0.054	-0.051	-0.051	0.131	0.102	0.132	0.100	0.133	0.133	0.950	0.950	0.950	0.952	0.949	0.949
different	low	10	-0.05	-0.048	-0.049	-0.048	-0.048	-0.048	-0.048	0.218	0.152	0.216	0.147	0.221	0.221	0.948	0.950	0.950	0.952	0.949	0.949
different	no	2	0	0.000	-0.002	0.000	-0.002	0.000	0.000	0.066	0.055	0.067	0.054	0.068	0.068	0.934	0.946	0.934	0.947	0.932	0.933
different	no	4	0	0.002	0.000	0.002	0.000	0.002	0.002	0.057	0.047	0.058	0.044	0.056	0.056	0.943	0.954	0.943	0.956	0.945	0.945
different	no	10	0	0.000	-0.001	0.000	0.000	-0.001	-0.001	0.060	0.057	0.060	0.057	0.060	0.060	0.941	0.943	0.941	0.943	0.941	0.941

Table 47: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. The mean effect, bias, and the root mean squared error (RMSE) are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model.*

Popu-	Effect	Trials	Bias						MC Error for Bias						RMSE					
lation	Size		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.021	-0.022	-0.020	-0.021	-0.028	-0.032	0.006	0.006	0.006	0.006	0.007	0.007	0.285	0.287	0.285	0.287	0.293	0.294
similar	high	4	-0.027	-0.027	-0.027	-0.027	-0.038	-0.038	0.005	0.005	0.005	0.005	0.005	0.005	0.202	0.203	0.202	0.203	0.210	0.210
similar	high	10	-0.011	-0.011	-0.010	-0.010	-0.022	-0.021	0.003	0.003	0.003	0.003	0.003	0.003	0.123	0.123	0.123	0.123	0.144	0.137
similar	moderate	2	-0.016	-0.016	-0.016	-0.016	-0.017	-0.017	0.003	0.003	0.003	0.003	0.003	0.003	0.130	0.130	0.130	0.130	0.132	0.132
similar	moderate	4	-0.010	-0.010	-0.010	-0.010	-0.011	-0.011	0.002	0.002	0.002	0.002	0.002	0.002	0.095	0.095	0.095	0.095	0.096	0.096
similar	moderate	10	-0.009	-0.009	-0.009	-0.009	-0.011	-0.011	0.001	0.001	0.001	0.001	0.001	0.001	0.058	0.058	0.058	0.058	0.059	0.059
similar	low	2	-0.004	-0.004	-0.004	-0.003	-0.004	-0.004	0.002	0.002	0.002	0.002	0.002	0.002	0.090	0.091	0.090	0.091	0.091	0.091
similar	low	4	0.000	0.000	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.064	0.064	0.064	0.064	0.064	0.064
similar	low	10	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.040	0.040	0.040	0.040	0.040	0.040
similar	no	2	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.082	0.082	0.082	0.082	0.082	0.082
similar	no	4	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.058	0.058	0.058	0.058	0.058	0.058
similar	no	10	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.035	0.035	0.035	0.035	0.036	0.036
different	high	2	-0.040	-0.060	-0.040	-0.060	-0.049	-0.052	0.007	0.010	0.007	0.010	0.007	0.007	0.298	0.439	0.298	0.439	0.309	0.311
different	high	4	-0.023	-0.036	-0.022	-0.036	-0.030	-0.030	0.005	0.006	0.005	0.006	0.005	0.005	0.201	0.285	0.201	0.284	0.211	0.211
different	high	10	-0.015	-0.018	-0.014	-0.017	-0.028	-0.028	0.003	0.004	0.003	0.004	0.004	0.003	0.123	0.181	0.123	0.181	0.158	0.150
different	moderate	2	-0.014	-0.019	-0.013	-0.018	-0.014	-0.015	0.003	0.004	0.003	0.004	0.003	0.003	0.138	0.193	0.138	0.193	0.140	0.140
different	moderate	4	-0.009	-0.010	-0.008	-0.010	-0.010	-0.010	0.002	0.003	0.002	0.003	0.002	0.002	0.096	0.134	0.096	0.134	0.097	0.097
different	moderate	10	-0.011	-0.010	-0.010	-0.010	-0.013	-0.013	0.001	0.002	0.001	0.002	0.001	0.001	0.061	0.085	0.061	0.085	0.063	0.063
different	low	2	0.000	-0.002	0.001	-0.001	0.000	0.000	0.002	0.003	0.002	0.003	0.002	0.002	0.095	0.120	0.095	0.120	0.095	0.095
different	low	4	-0.001	-0.004	0.000	-0.004	-0.001	-0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.065	0.085	0.065	0.085	0.065	0.065
different	low	10	0.002	0.001	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.040	0.053	0.040	0.053	0.040	0.040
different	no	2	0.000	-0.002	0.000	-0.002	0.000	0.000	0.002	0.002	0.002	0.002	0.002	0.002	0.085	0.105	0.085	0.105	0.085	0.085
different	no	4	0.002	0.000	0.002	0.000	0.002	0.002	0.001	0.002	0.001	0.002	0.001	0.001	0.058	0.073	0.058	0.072	0.058	0.058
different	no	10	0.000	-0.001	0.000	0.000	-0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.038	0.047	0.038	0.047	0.038	0.038

Table 48: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario III and IV. Power and coverage are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. An interaction is considered in the outcome generation process. All meta-analyses are based on a random effects model.*

Popu-	Effect	Trials	True	Mean Effect				Power				Coverage			
lation	Size		Effect	Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.468	-0.471	-0.528	-0.528	0.518	0.521	0.262	0.262	0.750	0.756	0.916	0.916
similar	high	4	-0.69	-0.495	-0.498	-0.551	-0.551	0.576	0.580	0.274	0.274	0.794	0.797	0.938	0.938
similar	high	10	-0.69	-0.514	-0.516	-0.550	-0.550	0.622	0.623	0.295	0.295	0.796	0.798	0.921	0.921
similar	moderate	2	-0.22	-0.152	-0.153	-0.179	-0.179	0.306	0.304	0.159	0.151	0.845	0.845	0.939	0.940
similar	moderate	4	-0.22	-0.184	-0.185	-0.214	-0.214	0.430	0.427	0.238	0.238	0.862	0.866	0.943	0.943
similar	moderate	10	-0.22	-0.222	-0.223	-0.239	-0.239	0.576	0.580	0.300	0.300	0.884	0.884	0.947	0.947
similar	low	2	-0.05	-0.067	-0.067	-0.087	-0.089	0.155	0.154	0.106	0.087	0.913	0.911	0.934	0.950
similar	low	4	-0.05	-0.102	-0.102	-0.127	-0.127	0.312	0.307	0.206	0.204	0.856	0.856	0.898	0.900
similar	low	10	-0.05	-0.136	-0.136	-0.153	-0.153	0.520	0.519	0.319	0.319	0.711	0.7085	0.827	0.827
similar	no	2	0	-0.038	-0.038	-0.055	-0.060	0.096	0.097	0.078	0.055	0.904	0.904	0.922	0.946
similar	no	4	0	-0.075	-0.075	-0.097	-0.098	0.215	0.218	0.177	0.168	0.786	0.782	0.824	0.833
similar	no	10	0	-0.105	-0.105	-0.119	-0.119	0.453	0.444	0.288	0.288	0.548	0.556	0.713	0.713
different	high	2	-0.69	-0.504	-0.324	-0.550	-0.550	0.587	0.276	0.283	0.283	0.800	0.683	0.916	0.916
different	high	4	-0.69	-0.531	-0.352	-0.570	-0.570	0.646	0.314	0.297	0.297	0.816	0.700	0.926	0.926
different	high	10	-0.69	-0.540	-0.352	-0.571	-0.571	0.652	0.300	0.306	0.306	0.819	0.686	0.929	0.929
different	moderate	2	-0.22	-0.218	-0.123	-0.234	-0.234	0.474	0.214	0.267	0.265	0.888	0.821	0.942	0.942
different	moderate	4	-0.22	-0.244	-0.149	-0.261	-0.261	0.601	0.235	0.325	0.325	0.884	0.847	0.938	0.938
different	moderate	10	-0.22	-0.260	-0.168	-0.266	-0.266	0.690	0.295	0.334	0.334	0.867	0.838	0.937	0.937
different	low	2	-0.05	-0.108	-0.047	-0.116	-0.117	0.275	0.118	0.1735	0.152	0.861	0.911	0.908	0.920
different	low	4	-0.05	-0.133	-0.075	-0.143	-0.143	0.453	0.191	0.266	0.266	0.751	0.883	0.860	0.861
different	low	10	-0.05	-0.152	-0.094	-0.156	-0.156	0.644	0.240	0.333	0.333	0.624	0.864	0.818	0.818
different	no	2	0	-0.079	-0.031	-0.086	-0.087	0.199	0.119	0.139	0.118	0.802	0.882	0.861	0.882
different	no	4	0	-0.103	-0.058	-0.112	-0.112	0.358	0.178	0.237	0.233	0.643	0.823	0.764	0.767
different	no	10	0	-0.127	-0.079	-0.133	-0.133	0.600	0.238	0.343	0.343	0.401	0.762	0.658	0.658

Table 49: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario III and IV. The mean effect, bias, and the root mean squared error (RMSE) are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. An interaction is considered in the outcome generation process. All meta-analyses are based on a random effects model.*

Population	Effect Size	Trials	Bias				MC Error for Bias				RMSE			
			Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	0.222	0.219	0.162	0.162	0.007	0.007	0.009	0.009	0.372	0.371	0.446	0.446
similar	high	4	0.195	0.192	0.139	0.139	0.006	0.006	0.009	0.009	0.339	0.339	0.419	0.419
similar	high	10	0.176	0.174	0.140	0.140	0.006	0.006	0.009	0.009	0.336	0.336	0.432	0.432
similar	moderate	2	0.068	0.067	0.041	0.041	0.003	0.003	0.004	0.004	0.155	0.155	0.184	0.184
similar	moderate	4	0.036	0.035	0.006	0.006	0.003	0.003	0.004	0.004	0.141	0.141	0.178	0.178
similar	moderate	10	-0.002	-0.003	-0.019	-0.019	0.003	0.003	0.004	0.004	0.127	0.127	0.171	0.171
similar	low	2	-0.017	-0.017	-0.037	-0.039	0.002	0.002	0.003	0.003	0.099	0.099	0.126	0.128
similar	low	4	-0.052	-0.052	-0.077	-0.077	0.002	0.002	0.003	0.003	0.102	0.103	0.134	0.134
similar	low	10	-0.086	-0.086	-0.103	-0.103	0.002	0.002	0.002	0.002	0.119	0.119	0.148	0.148
similar	no	2	-0.038	-0.038	-0.055	-0.060	0.002	0.002	0.002	0.002	0.096	0.096	0.117	0.121
similar	no	4	-0.075	-0.075	-0.097	-0.098	0.002	0.002	0.002	0.002	0.109	0.109	0.138	0.138
similar	no	10	-0.105	-0.105	-0.119	-0.119	0.002	0.002	0.002	0.002	0.127	0.127	0.150	0.150
different	high	2	0.186	0.366	0.140	0.140	0.007	0.009	0.009	0.009	0.343	0.551	0.439	0.439
different	high	4	0.159	0.338	0.120	0.120	0.006	0.009	0.009	0.009	0.329	0.540	0.428	0.428
different	high	10	0.150	0.338	0.119	0.119	0.006	0.010	0.009	0.009	0.322	0.541	0.421	0.421
different	moderate	2	0.002	0.097	-0.014	-0.014	0.003	0.004	0.004	0.004	0.142	0.220	0.184	0.184
different	moderate	4	-0.024	0.071	-0.041	-0.041	0.003	0.004	0.004	0.004	0.134	0.199	0.178	0.178
different	moderate	10	-0.040	0.052	-0.046	-0.046	0.003	0.004	0.004	0.004	0.133	0.196	0.178	0.178
different	low	2	-0.058	0.003	-0.066	-0.067	0.002	0.003	0.003	0.003	0.113	0.123	0.135	0.135
different	low	4	-0.083	-0.025	-0.093	-0.093	0.002	0.003	0.003	0.003	0.121	0.122	0.144	0.144
different	low	10	-0.102	-0.044	-0.106	-0.106	0.002	0.002	0.002	0.002	0.128	0.117	0.147	0.147
different	no	2	-0.079	-0.031	-0.086	-0.087	0.002	0.003	0.002	0.002	0.117	0.115	0.133	0.134
different	no	4	-0.103	-0.058	-0.112	-0.112	0.002	0.002	0.002	0.002	0.129	0.117	0.147	0.147
different	no	10	-0.127	-0.079	-0.133	-0.133	0.002	0.002	0.002	0.002	0.143	0.120	0.157	0.157

Table 50: *One AgD trial (aggregated data) and multiple IPD trials (individual patient data) - scenario III and IV. Power and coverage are presented for approach B.1 (pooled IPD) and B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. An interaction is considered in the outcome generation process. All meta-analyses are based on a random effects model.*

Population	Effect Size	Trials	True Effect	Mean Effect				Power				Coverage			
				Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.565	-0.567	-0.579	-0.584	0.482	0.479	0.482	0.399	0.914	0.921	0.923	0.939
similar	high	4	-0.69	-0.536	-0.539	-0.561	-0.564	0.670	0.669	0.690	0.648	0.894	0.891	0.914	0.925
similar	high	10	-0.69	-0.535	-0.537	-0.569	-0.570	0.8735	0.874	0.885	0.885	0.852	0.853	0.879	0.881
similar	moderate	2	-0.22	-0.254	-0.255	-0.256	-0.256	0.311	0.317	0.314	0.296	0.944	0.944	0.941	0.949
similar	moderate	4	-0.22	-0.253	-0.253	-0.255	-0.255	0.378	0.380	0.381	0.379	0.939	0.938	0.939	0.939
similar	moderate	10	-0.22	-0.253	-0.253	-0.257	-0.257	0.454	0.451	0.461	0.461	0.943	0.942	0.941	0.941
similar	low	2	-0.05	-0.167	-0.167	-0.168	-0.168	0.213	0.210	0.214	0.212	0.859	0.865	0.857	0.861
similar	low	4	-0.05	-0.166	-0.166	-0.167	-0.167	0.241	0.237	0.243	0.243	0.862	0.862	0.863	0.863
similar	low	10	-0.05	-0.167	-0.167	-0.168	-0.168	0.246	0.248	0.247	0.247	0.860	0.856	0.858	0.858
similar	no	2	0	-0.136	-0.136	-0.136	-0.136	0.165	0.162	0.163	0.162	0.836	0.838	0.838	0.838
similar	no	4	0	-0.140	-0.139	-0.140	-0.140	0.188	0.189	0.187	0.187	0.813	0.811	0.813	0.813
similar	no	10	0	-0.140	-0.140	-0.140	-0.140	0.188	0.188	0.189	0.189	0.812	0.812	0.812	0.812
different	high	2	-0.69	-0.557	-0.375	-0.575	-0.581	0.489	0.201	0.501	0.426	0.927	0.831	0.926	0.946
different	high	4	-0.69	-0.545	-0.356	-0.568	-0.571	0.720	0.270	0.734	0.678	0.903	0.762	0.921	0.932
different	high	10	-0.69	-0.534	-0.334	-0.567	-0.567	0.926	0.393	0.936	0.933	0.831	0.557	0.873	0.879
different	moderate	2	-0.22	-0.275	-0.178	-0.277	-0.277	0.423	0.160	0.423	0.388	0.934	0.933	0.934	0.941
different	moderate	4	-0.22	-0.273	-0.176	-0.276	-0.276	0.524	0.194	0.538	0.533	0.927	0.933	0.926	0.927
different	moderate	10	-0.22	-0.274	-0.178	-0.278	-0.278	0.624	0.290	0.632	0.632	0.917	0.927	0.913	0.913
different	low	2	-0.05	-0.172	-0.111	-0.173	-0.173	0.278	0.111	0.278	0.274	0.844	0.932	0.840	0.845
different	low	4	-0.05	-0.168	-0.110	-0.169	-0.169	0.296	0.142	0.300	0.300	0.828	0.922	0.827	0.827
different	low	10	-0.05	-0.162	-0.104	-0.163	-0.163	0.298	0.136	0.302	0.302	0.835	0.937	0.834	0.834
different	no	2	0	-0.137	-0.089	-0.138	-0.138	0.199	0.100	0.199	0.196	0.801	0.900	0.802	0.804
different	no	4	0	-0.134	-0.086	-0.134	-0.134	0.220	0.112	0.222	0.222	0.78	0.889	0.778	0.778
different	no	10	0	-0.141	-0.093	-0.141	-0.141	0.256	0.123	0.255	0.255	0.745	0.877	0.746	0.746

Table 51: *One AgD trial (aggregated data) and multiple IPD trials (individual patient data) - scenario III and IV. The mean effect, bias, and the root mean squared error (RMSE) are presented for approach B.1 (pooled IPD) and B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. An interaction is considered in the outcome generation process. All meta-analyses are based on a random effects model.*

Population	Effect Size	Trials	Bias				MC Error for Bias				RMSE			
			Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	0.125	0.123	0.111	0.106	0.007	0.007	0.007	0.007	0.328	0.328	0.333	0.334
similar	high	4	0.154	0.151	0.129	0.126	0.005	0.005	0.005	0.005	0.270	0.270	0.266	0.266
similar	high	10	0.155	0.153	0.121	0.120	0.004	0.004	0.004	0.004	0.232	0.231	0.216	0.216
similar	moderate	2	-0.034	-0.035	-0.036	-0.036	0.004	0.004	0.004	0.004	0.175	0.175	0.176	0.177
similar	moderate	4	-0.033	-0.033	-0.035	-0.035	0.003	0.003	0.003	0.003	0.157	0.158	0.158	0.158
similar	moderate	10	-0.033	-0.033	-0.037	-0.037	0.003	0.003	0.003	0.003	0.143	0.144	0.145	0.145
similar	low	2	-0.117	-0.117	-0.118	-0.118	0.003	0.003	0.003	0.003	0.186	0.186	0.186	0.186
similar	low	4	-0.116	-0.116	-0.117	-0.117	0.003	0.003	0.003	0.003	0.179	0.179	0.180	0.180
similar	low	10	-0.117	-0.117	-0.118	-0.118	0.003	0.003	0.003	0.003	0.174	0.174	0.174	0.174
similar	no	2	-0.136	-0.136	-0.136	-0.136	0.003	0.003	0.003	0.003	0.193	0.193	0.193	0.193
similar	no	4	-0.140	-0.139	-0.140	-0.140	0.003	0.003	0.003	0.003	0.195	0.195	0.195	0.195
similar	no	10	-0.140	-0.140	-0.140	-0.140	0.003	0.003	0.003	0.003	0.190	0.190	0.191	0.191
different	high	2	0.133	0.315	0.115	0.109	0.006	0.009	0.007	0.007	0.315	0.517	0.319	0.319
different	high	4	0.145	0.334	0.122	0.119	0.005	0.007	0.005	0.005	0.258	0.445	0.254	0.253
different	high	10	0.156	0.356	0.123	0.123	0.004	0.005	0.004	0.004	0.220	0.411	0.205	0.205
different	moderate	2	-0.055	0.042	-0.057	-0.057	0.004	0.005	0.004	0.004	0.169	0.213	0.171	0.171
different	moderate	4	-0.053	0.044	-0.056	-0.056	0.003	0.004	0.003	0.003	0.147	0.171	0.149	0.149
different	moderate	10	-0.054	0.042	-0.058	-0.058	0.003	0.003	0.003	0.003	0.135	0.143	0.138	0.138
different	low	2	-0.122	-0.061	-0.123	-0.123	0.003	0.003	0.003	0.003	0.176	0.157	0.177	0.177
different	low	4	-0.118	-0.060	-0.119	-0.119	0.003	0.003	0.003	0.003	0.167	0.144	0.168	0.168
different	low	10	-0.112	-0.054	-0.113	-0.113	0.003	0.003	0.003	0.003	0.157	0.125	0.157	0.157
different	no	2	-0.137	-0.089	-0.138	-0.138	0.003	0.003	0.003	0.003	0.182	0.161	0.182	0.182
different	no	4	-0.134	-0.086	-0.134	-0.134	0.003	0.003	0.003	0.003	0.177	0.152	0.177	0.177
different	no	10	-0.141	-0.093	-0.141	-0.141	0.003	0.003	0.003	0.003	0.178	0.147	0.179	0.179

Table 52: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario III and IV. Power and coverage are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. An interaction is considered in the outcome generation process. All meta-analyses are based on a random effects model.*

Popu-	Effect	Trials	True	Mean Effect						Power						Coverage					
lation	Size		Effect	All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.463	-0.466	-0.492	-0.495	-0.505	-0.507	0.396	0.400	0.432	0.430	0.426	0.412	0.845	0.848	0.864	0.869	0.874	0.879
similar	high	4	-0.69	-0.471	-0.475	-0.500	-0.504	-0.522	-0.522	0.697	0.700	0.749	0.752	0.743	0.743	0.795	0.799	0.842	0.839	0.860	0.860
similar	high	10	-0.69	-0.501	-0.504	-0.518	-0.521	-0.544	-0.544	0.985	0.988	0.992	0.992	0.992	0.992	0.655	0.665	0.712	0.713	0.780	0.780
similar	moderate	2	-0.22	-0.153	-0.153	-0.175	-0.177	-0.176	-0.177	0.220	0.215	0.250	0.224	0.254	0.245	0.916	0.914	0.939	0.943	0.941	0.944
similar	moderate	4	-0.22	-0.186	-0.186	-0.210	-0.210	-0.212	-0.212	0.484	0.488	0.566	0.558	0.570	0.570	0.933	0.937	0.943	0.946	0.946	0.946
similar	moderate	10	-0.22	-0.221	-0.221	-0.235	-0.235	-0.238	-0.238	0.933	0.935	0.954	0.954	0.952	0.952	0.945	0.946	0.943	0.943	0.942	0.942
similar	low	2	-0.05	-0.067	-0.067	-0.081	-0.087	-0.081	-0.083	0.111	0.111	0.130	0.094	0.133	0.123	0.945	0.944	0.944	0.961	0.941	0.946
similar	low	4	-0.05	-0.102	-0.102	-0.118	-0.121	-0.119	-0.119	0.303	0.304	0.363	0.314	0.367	0.367	0.893	0.893	0.854	0.876	0.850	0.850
similar	low	10	-0.05	-0.137	-0.137	-0.147	-0.147	-0.147	-0.147	0.807	0.808	0.845	0.838	0.848	0.848	0.559	0.552	0.481	0.493	0.476	0.476
similar	no	2	0	-0.038	-0.038	-0.048	-0.057	-0.049	-0.051	0.065	0.066	0.074	0.048	0.075	0.073	0.936	0.935	0.927	0.953	0.925	0.928
similar	no	4	0	-0.073	-0.073	-0.086	-0.091	-0.086	-0.086	0.199	0.200	0.237	0.187	0.240	0.240	0.802	0.801	0.763	0.814	0.760	0.760
similar	no	10	0	-0.106	-0.106	-0.114	-0.115	-0.114	-0.114	0.656	0.658	0.709	0.691	0.710	0.710	0.345	0.342	0.291	0.309	0.291	0.291
different	high	2	-0.69	-0.499	-0.318	-0.510	-0.330	-0.528	-0.529	0.445	0.167	0.464	0.176	0.472	0.463	0.885	0.789	0.894	0.798	0.904	0.907
different	high	4	-0.69	-0.512	-0.323	-0.522	-0.334	-0.542	-0.542	0.766	0.258	0.788	0.271	0.787	0.787	0.847	0.683	0.859	0.703	0.879	0.879
different	high	10	-0.69	-0.519	-0.320	-0.525	-0.326	-0.550	-0.550	0.993	0.474	0.994	0.487	0.993	0.993	0.708	0.409	0.727	0.423	0.795	0.795
different	moderate	2	-0.22	-0.216	-0.119	-0.225	-0.129	-0.226	-0.227	0.367	0.121	0.398	0.128	0.396	0.385	0.947	0.905	0.948	0.911	0.948	0.950
different	moderate	4	-0.22	-0.239	-0.143	-0.247	-0.152	-0.250	-0.250	0.714	0.199	0.740	0.219	0.738	0.738	0.953	0.909	0.944	0.918	0.943	0.943
different	moderate	10	-0.22	-0.259	-0.164	-0.263	-0.168	-0.266	-0.266	0.990	0.505	0.993	0.525	0.993	0.993	0.904	0.900	0.892	0.906	0.887	0.887
different	low	2	-0.05	-0.108	-0.047	-0.115	-0.057	-0.116	-0.116	0.207	0.069	0.226	0.062	0.226	0.220	0.906	0.955	0.900	0.961	0.898	0.902
different	low	4	-0.05	-0.134	-0.076	-0.140	-0.084	-0.141	-0.141	0.475	0.146	0.503	0.165	0.505	0.505	0.762	0.928	0.739	0.930	0.733	0.733
different	low	10	-0.05	-0.152	-0.094	-0.155	-0.098	-0.156	-0.156	0.928	0.383	0.935	0.405	0.936	0.936	0.373	0.882	0.352	0.866	0.344	0.344
different	no	2	0	-0.079	-0.032	-0.084	-0.040	-0.084	-0.085	0.155	0.067	0.168	0.058	0.167	0.161	0.846	0.933	0.833	0.943	0.833	0.840
different	no	4	0	-0.102	-0.055	-0.106	-0.062	-0.107	-0.107	0.361	0.118	0.386	0.126	0.387	0.387	0.640	0.882	0.615	0.875	0.614	0.614
different	no	10	0	-0.124	-0.077	-0.127	-0.080	-0.127	-0.127	0.862	0.344	0.871	0.372	0.874	0.874	0.139	0.656	0.129	0.629	0.126	0.126

Table 53: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario III and IV. The mean effect, bias, and the root mean squared error (RMSE) are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. An interaction is considered in the outcome generation process. All meta-analyses are based on a random effects model.*

Popu-	Effect	Trials	Bias						MC Error for Bias						RMSE					
lation	Size		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	0.227	0.224	0.198	0.195	0.185	0.183	0.006	0.006	0.006	0.006	0.007	0.007	0.365	0.364	0.348	0.347	0.351	0.351
similar	high	4	0.219	0.215	0.190	0.186	0.168	0.168	0.004	0.004	0.004	0.004	0.005	0.005	0.290	0.287	0.269	0.267	0.263	0.263
similar	high	10	0.189	0.186	0.172	0.169	0.146	0.146	0.003	0.003	0.003	0.003	0.003	0.003	0.224	0.222	0.210	0.208	0.195	0.195
similar	moderate	2	0.067	0.067	0.045	0.043	0.044	0.043	0.003	0.003	0.003	0.003	0.003	0.003	0.150	0.151	0.143	0.144	0.144	0.144
similar	moderate	4	0.034	0.034	0.010	0.010	0.008	0.008	0.002	0.002	0.002	0.002	0.002	0.002	0.104	0.104	0.101	0.101	0.101	0.101
similar	moderate	10	-0.001	-0.001	-0.015	-0.015	-0.018	-0.018	0.001	0.001	0.001	0.001	0.002	0.002	0.064	0.064	0.066	0.066	0.068	0.068
similar	low	2	-0.017	-0.017	-0.031	-0.037	-0.031	-0.033	0.002	0.002	0.002	0.002	0.002	0.002	0.095	0.095	0.100	0.105	0.101	0.102
similar	low	4	-0.052	-0.052	-0.068	-0.071	-0.069	-0.069	0.002	0.002	0.002	0.002	0.002	0.002	0.088	0.088	0.099	0.102	0.100	0.100
similar	low	10	-0.087	-0.087	-0.097	-0.097	-0.097	-0.097	0.001	0.001	0.001	0.001	0.001	0.001	0.099	0.099	0.108	0.109	0.109	0.109
similar	no	2	-0.038	-0.038	-0.048	-0.057	-0.049	-0.051	0.002	0.002	0.002	0.002	0.002	0.002	0.095	0.095	0.100	0.107	0.100	0.102
similar	no	4	-0.073	-0.073	-0.086	-0.091	-0.086	-0.086	0.002	0.002	0.002	0.002	0.002	0.002	0.099	0.099	0.109	0.113	0.109	0.109
similar	no	10	-0.106	-0.106	-0.114	-0.115	-0.114	-0.114	0.001	0.001	0.001	0.001	0.001	0.001	0.115	0.115	0.123	0.124	0.123	0.123
different	high	2	0.191	0.372	0.180	0.36	0.162	0.161	0.0061	0.0089	0.0061	0.0089	0.006	0.006	0.332	0.546	0.326	0.538	0.328	0.328
different	high	4	0.178	0.367	0.168	0.356	0.148	0.148	0.004	0.006	0.004	0.006	0.005	0.005	0.261	0.460	0.255	0.452	0.249	0.249
different	high	10	0.171	0.370	0.165	0.364	0.140	0.140	0.003	0.004	0.003	0.004	0.003	0.003	0.207	0.410	0.203	0.405	0.189	0.189
different	moderate	2	0.004	0.101	-0.005	0.091	-0.006	-0.007	0.003	0.004	0.003	0.004	0.003	0.003	0.136	0.216	0.137	0.211	0.138	0.138
different	moderate	4	-0.019	0.077	-0.027	0.068	-0.030	-0.030	0.002	0.003	0.002	0.003	0.002	0.002	0.097	0.152	0.099	0.148	0.101	0.101
different	moderate	10	-0.039	0.056	-0.043	0.052	-0.046	-0.046	0.001	0.002	0.001	0.002	0.001	0.001	0.072	0.100	0.075	0.098	0.077	0.077
different	low	2	-0.058	0.003	-0.065	-0.007	-0.066	-0.066	0.002	0.003	0.002	0.003	0.002	0.002	0.110	0.118	0.114	0.118	0.115	0.115
different	low	4	-0.084	-0.026	-0.090	-0.034	-0.091	-0.091	0.002	0.002	0.002	0.002	0.002	0.002	0.109	0.093	0.114	0.096	0.115	0.115
different	low	10	-0.102	-0.044	-0.105	-0.048	-0.106	-0.106	0.001	0.001	0.001	0.001	0.001	0.001	0.111	0.070	0.114	0.072	0.115	0.115
different	no	2	-0.079	-0.032	-0.084	-0.040	-0.084	-0.085	0.002	0.002	0.002	0.002	0.002	0.002	0.116	0.110	0.119	0.113	0.119	0.120
different	no	4	-0.102	-0.055	-0.106	-0.062	-0.107	-0.107	0.001	0.002	0.001	0.002	0.001	0.001	0.120	0.095	0.124	0.099	0.125	0.125
different	no	10	-0.124	-0.077	-0.127	-0.080	-0.127	-0.127	0.001	0.001	0.001	0.001	0.001	0.001	0.131	0.091	0.133	0.094	0.134	0.134

Table 54: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario I and II. Mean Effect, power, and coverage are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a fixed effects model.*

Popu-	Effect	Trials	True	Mean Effect				Power				Coverage			
lation	Size		Effect	Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.713	-0.713	-0.742	-0.742	0.808	0.804	0.444	0.444	0.881	0.885	0.951	0.951
similar	high	4	-0.69	-0.743	-0.743	-0.777	-0.777	0.846	0.843	0.459	0.459	0.878	0.878	0.953	0.953
similar	high	10	-0.69	-0.726	-0.726	-0.756	-0.756	0.842	0.845	0.445	0.445	0.869	0.873	0.949	0.949
similar	moderate	2	-0.22	-0.234	-0.234	-0.231	-0.231	0.547	0.551	0.250	0.250	0.902	0.902	0.952	0.952
similar	moderate	4	-0.22	-0.232	-0.232	-0.234	-0.234	0.571	0.567	0.267	0.267	0.883	0.879	0.948	0.948
similar	moderate	10	-0.22	-0.232	-0.232	-0.234	-0.234	0.512	0.610	0.278	0.278	0.879	0.879	0.958	0.958
similar	low	2	-0.05	-0.053	-0.053	-0.052	-0.052	0.150	0.152	0.091	0.091	0.905	0.903	0.940	0.940
similar	low	4	-0.05	-0.051	-0.051	-0.052	-0.052	0.184	0.186	0.090	0.090	0.879	0.878	0.943	0.943
similar	low	10	-0.05	-0.052	-0.052	-0.051	-0.051	0.209	0.207	0.087	0.087	0.875	0.878	0.946	0.946
similar	no	2	0	-0.003	-0.003	-0.002	-0.002	0.094	0.094	0.069	0.069	0.907	0.906	0.932	0.932
similar	no	4	0	0.002	0.002	0.002	0.002	0.109	0.110	0.060	0.060	0.892	0.890	0.940	0.940
similar	no	10	0	-0.001	-0.001	-0.001	-0.001	0.130	0.130	0.065	0.065	0.871	0.871	0.936	0.936
different	high	2	-0.69	-0.734	-0.756	-0.766	-0.766	0.822	0.6345	0.456	0.456	0.866	0.821	0.947	0.947
different	high	4	-0.69	-0.725	-0.759	-0.746	-0.746	0.833	0.639	0.446	0.446	0.870	0.844	0.956	0.956
different	high	10	-0.69	-0.729	-0.755	-0.766	-0.766	0.850	0.6265	0.458	0.458	0.866	0.850	0.940	0.940
different	moderate	2	-0.22	-0.234	-0.239	-0.237	-0.237	0.538	0.3915	0.257	0.257	0.881	0.868	0.945	0.945
different	moderate	4	-0.22	-0.231	-0.232	-0.233	-0.233	0.562	0.388	0.273	0.273	0.883	0.876	0.950	0.950
different	moderate	10	-0.22	-0.231	-0.230	-0.236	-0.236	0.593	0.4015	0.282	0.282	0.864	0.851	0.933	0.933
different	low	2	-0.05	-0.050	-0.054	-0.051	-0.051	0.146	0.146	0.081	0.081	0.905	0.895	0.944	0.944
different	low	4	-0.05	-0.051	-0.056	-0.050	-0.050	0.159	0.174	0.079	0.079	0.885	0.881	0.947	0.947
different	low	10	-0.05	-0.045	-0.047	-0.046	-0.046	0.173	0.159	0.076	0.076	0.881	0.885	0.945	0.945
different	no	2	0	0.000	-0.002	-0.001	-0.001	0.099	0.106	0.059	0.059	0.901	0.895	0.941	0.941
different	no	4	0	0.002	0.003	0.002	0.002	0.099	0.121	0.049	0.049	0.902	0.879	0.951	0.951
different	no	10	0	-0.005	-0.008	-0.006	-0.006	0.129	0.129	0.061	0.061	0.872	0.872	0.939	0.939

Table 55: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario I and II. The bias, its Monte Carlo (MC) standard error, and the root mean squared error (RMSE) are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a fixed effects model.*

Population	Effect Size	Trials	Bias				MC Error for Bias				RMSE			
			Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.023	-0.023	-0.052	-0.052	0.007	0.007	0.010	0.010	0.300	0.302	0.433	0.433
similar	high	4	-0.053	-0.053	-0.087	-0.087	0.007	0.007	0.010	0.010	0.303	0.305	0.433	0.433
similar	high	10	-0.036	-0.036	-0.066	-0.066	0.007	0.007	0.010	0.010	0.301	0.302	0.436	0.436
similar	moderate	2	-0.014	-0.014	-0.011	-0.011	0.003	0.003	0.004	0.004	0.136	0.136	0.178	0.178
similar	moderate	4	-0.012	-0.012	-0.014	-0.014	0.003	0.003	0.004	0.004	0.135	0.135	0.178	0.178
similar	moderate	10	-0.012	-0.012	-0.014	-0.014	0.003	0.003	0.004	0.004	0.129	0.129	0.171	0.171
similar	low	2	-0.003	-0.003	-0.002	-0.002	0.002	0.002	0.003	0.003	0.094	0.094	0.117	0.117
similar	low	4	-0.001	-0.001	-0.002	-0.002	0.002	0.002	0.002	0.002	0.085	0.086	0.109	0.109
similar	low	10	-0.002	-0.002	-0.001	-0.001	0.002	0.002	0.002	0.002	0.078	0.078	0.103	0.103
similar	no	2	-0.003	-0.003	-0.002	-0.002	0.002	0.002	0.002	0.002	0.085	0.085	0.101	0.101
similar	no	4	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.073	0.073	0.091	0.091
similar	no	10	-0.001	-0.001	-0.001	-0.001	0.002	0.002	0.002	0.002	0.066	0.066	0.086	0.086
different	high	2	-0.044	-0.066	-0.076	-0.076	0.007	0.010	0.010	0.010	0.316	0.46	0.452	0.452
different	high	4	-0.035	-0.069	-0.056	-0.056	0.007	0.010	0.010	0.010	0.303	0.437	0.428	0.428
different	high	10	-0.039	-0.065	-0.076	-0.076	0.007	0.0098	0.010	0.010	0.302	0.442	0.438	0.438
different	moderate	2	-0.014	-0.019	-0.017	-0.017	0.003	0.005	0.004	0.004	0.143	0.200	0.184	0.184
different	moderate	4	-0.011	-0.012	-0.013	-0.013	0.003	0.004	0.004	0.004	0.135	0.191	0.178	0.178
different	moderate	10	-0.011	-0.010	-0.016	-0.016	0.003	0.004	0.004	0.004	0.133	0.195	0.182	0.182
different	low	2	0.000	-0.004	-0.001	-0.001	0.002	0.003	0.003	0.003	0.097	0.125	0.118	0.118
different	low	4	-0.001	-0.006	0.000	0.000	0.002	0.003	0.002	0.002	0.085	0.117	0.108	0.108
different	low	10	0.005	0.003	0.004	0.004	0.002	0.003	0.002	0.002	0.077	0.111	0.102	0.102
different	no	2	0.000	-0.002	-0.001	-0.001	0.002	0.002	0.002	0.002	0.087	0.108	0.103	0.103
different	no	4	0.002	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.073	0.098	0.090	0.090
different	no	10	-0.005	-0.008	-0.006	-0.006	0.002	0.002	0.002	0.002	0.067	0.096	0.086	0.086

Table 56: *One AgD trial (individual patient data) and multiple IPD trials (aggregated data) - scenario I and II. Mean effect, power, and coverage are presented for approach B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a fixed effects model. Note, approach B.1 does not include a meta-analysis.*

Population	Effect Size	Trials	True Effect	Mean Effect		Power		Coverage	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.719	-0.719	0.691	0.691	0.952	0.952
similar	high	4	-0.69	-0.734	-0.734	0.927	0.927	0.946	0.946
similar	high	10	-0.69	-0.720	-0.720	0.998	0.998	0.941	0.941
similar	moderate	2	-0.22	-0.238	-0.238	0.388	0.388	0.948	0.948
similar	moderate	4	-0.22	-0.234	-0.234	0.526	0.526	0.948	0.948
similar	moderate	10	-0.22	-0.234	-0.234	0.692	0.692	0.939	0.939
similar	low	2	-0.05	-0.056	-0.056	0.091	0.091	0.947	0.947
similar	low	4	-0.05	-0.052	-0.052	0.100	0.100	0.936	0.936
similar	low	10	-0.05	-0.048	-0.048	0.100	0.100	0.942	0.942
similar	no	2	0	-0.002	-0.002	0.062	0.062	0.939	0.939
similar	no	4	0	0.002	0.002	0.055	0.055	0.945	0.945
similar	no	10	0	0.001	0.001	0.053	0.053	0.947	0.947
different	high	2	-0.69	-0.739	-0.739	0.700	0.700	0.947	0.947
different	high	4	-0.69	-0.726	-0.726	0.919	0.919	0.943	0.943
different	high	10	-0.69	-0.729	-0.729	0.999	0.999	0.940	0.940
different	moderate	2	-0.22	-0.234	-0.234	0.366	0.366	0.943	0.943
different	moderate	4	-0.22	-0.228	-0.228	0.488	0.488	0.947	0.947
different	moderate	10	-0.22	-0.236	-0.236	0.658	0.658	0.933	0.933
different	low	2	-0.05	-0.049	-0.049	0.081	0.081	0.938	0.938
different	low	4	-0.05	-0.052	-0.052	0.091	0.091	0.946	0.946
different	low	10	-0.05	-0.049	-0.049	0.098	0.098	0.941	0.941
different	no	2	0	-0.003	-0.003	0.063	0.063	0.938	0.938
different	no	4	0	-0.001	-0.001	0.052	0.052	0.949	0.949
different	no	10	0	-0.001	-0.001	0.058	0.058	0.943	0.943

Table 57: *One AgD trial (individual patient data) and multiple IPD trials (aggregated data) - scenario I and II. The bias, its Monte Carlo (MC) standard error, and the root mean squared error (RMSE) are presented for approach B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a fixed effects model. Note, approach B.1 does not include a meta-analysis.*

Population	Effect Size	Trials	Bias		MC Error for Bias		RMSE	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.029	-0.029	0.007	0.007	0.299	0.299
similar	high	4	-0.044	-0.044	0.005	0.005	0.224	0.224
similar	high	10	-0.030	-0.030	0.004	0.004	0.163	0.163
similar	moderate	2	-0.018	-0.018	0.003	0.003	0.144	0.144
similar	moderate	4	-0.014	-0.014	0.003	0.003	0.118	0.118
similar	moderate	10	-0.014	-0.014	0.002	0.002	0.098	0.098
similar	low	2	-0.006	-0.006	0.002	0.002	0.107	0.107
similar	low	4	-0.002	-0.002	0.002	0.002	0.097	0.097
similar	low	10	0.002	0.002	0.002	0.002	0.087	0.087
similar	no	2	-0.002	-0.002	0.002	0.002	0.102	0.102
similar	no	4	0.002	0.002	0.002	0.002	0.090	0.090
similar	no	10	0.001	0.001	0.002	0.002	0.083	0.083
different	high	2	-0.049	-0.049	0.007	0.007	0.316	0.316
different	high	4	-0.036	-0.036	0.005	0.005	0.228	0.228
different	high	10	-0.039	-0.039	0.004	0.004	0.178	0.178
different	moderate	2	-0.014	-0.014	0.003	0.003	0.153	0.153
different	moderate	4	-0.008	-0.008	0.003	0.003	0.121	0.121
different	moderate	10	-0.016	-0.016	0.002	0.002	0.104	0.104
different	low	2	0.001	0.001	0.003	0.003	0.114	0.114
different	low	4	-0.002	-0.002	0.002	0.002	0.100	0.100
different	low	10	0.001	0.001	0.002	0.002	0.093	0.093
different	no	2	-0.003	-0.003	0.002	0.002	0.107	0.107
different	no	4	-0.001	-0.001	0.002	0.002	0.093	0.093
different	no	10	-0.001	-0.001	0.002	0.002	0.090	0.090

Table 58: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II . Mean Effect, power, and coverage are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a fixed effects model.*

Popu-	Effect	Trials	True	Mean Effect						Power						Coverage					
lation	Size		Effect	All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.711	-0.712	-0.710	-0.711	-0.718	-0.718	0.720	0.717	0.720	0.715	0.700	0.700	0.952	0.956	0.952	0.955	0.951	0.951
similar	high	4	-0.69	-0.717	-0.717	-0.717	-0.717	-0.728	-0.728	0.961	0.961	0.961	0.960	0.952	0.952	0.954	0.953	0.954	0.954	0.948	0.948
similar	high	10	-0.69	-0.701	-0.701	-0.700	-0.700	-0.712	-0.712	1.000	1.000	1.000	1.000	1.000	1.000	0.951	0.952	0.952	0.952	0.944	0.944
similar	moderate	2	-0.22	-0.236	-0.236	-0.236	-0.236	-0.237	-0.237	0.442	0.442	0.442	0.441	0.441	0.441	0.953	0.953	0.953	0.954	0.954	0.954
similar	moderate	4	-0.22	-0.230	-0.230	-0.230	-0.230	-0.231	-0.231	0.699	0.700	0.698	0.695	0.695	0.695	0.946	0.943	0.946	0.944	0.945	0.945
similar	moderate	10	-0.22	-0.229	-0.229	-0.229	-0.229	-0.231	-0.231	0.976	0.976	0.976	0.976	0.976	0.976	0.955	0.954	0.955	0.954	0.952	0.952
similar	low	2	-0.05	-0.054	-0.054	-0.054	-0.053	-0.054	-0.054	0.099	0.099	0.099	0.100	0.099	0.099	0.943	0.943	0.942	0.943	0.943	0.943
similar	low	4	-0.05	-0.050	-0.050	-0.049	-0.049	-0.050	-0.050	0.137	0.133	0.139	0.133	0.144	0.144	0.948	0.946	0.948	0.948	0.947	0.947
similar	low	10	-0.05	-0.050	-0.050	-0.050	-0.050	-0.050	-0.050	0.249	0.251	0.249	0.247	0.256	0.256	0.949	0.948	0.949	0.949	0.950	0.950
similar	no	2	0	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	0.057	0.058	0.057	0.058	0.057	0.057	0.944	0.942	0.944	0.943	0.943	0.943
similar	no	4	0	0.002	0.002	0.002	0.002	0.002	0.002	0.063	0.062	0.063	0.062	0.062	0.062	0.938	0.939	0.937	0.938	0.939	0.939
similar	no	10	0	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.053	0.054	0.054	0.054	0.054	0.054	0.948	0.947	0.947	0.947	0.946	0.946
different	high	2	-0.69	-0.730	-0.750	-0.730	-0.750	-0.739	-0.739	0.738	0.503	0.737	0.503	0.711	0.711	0.946	0.919	0.946	0.919	0.945	0.945
different	high	4	-0.69	-0.713	-0.726	-0.712	-0.726	-0.720	-0.720	0.956	0.749	0.956	0.750	0.946	0.946	0.958	0.942	0.957	0.942	0.949	0.949
different	high	10	-0.69	-0.705	-0.708	-0.704	-0.707	-0.718	-0.718	1.000	0.973	1.000	0.972	1.000	1.000	0.952	0.946	0.952	0.946	0.938	0.938
different	moderate	2	-0.22	-0.234	-0.239	-0.233	-0.238	-0.234	-0.234	0.419	0.271	0.418	0.268	0.423	0.423	0.945	0.940	0.945	0.942	0.942	0.942
different	moderate	4	-0.22	-0.229	-0.230	-0.228	-0.230	-0.230	-0.230	0.672	0.428	0.671	0.425	0.676	0.676	0.938	0.946	0.938	0.945	0.939	0.939
different	moderate	10	-0.22	-0.231	-0.230	-0.230	-0.230	-0.233	-0.233	0.971	0.791	0.970	0.789	0.968	0.968	0.943	0.936	0.943	0.936	0.936	0.936
different	low	2	-0.05	-0.050	-0.052	-0.049	-0.051	-0.050	-0.050	0.097	0.077	0.098	0.075	0.099	0.099	0.947	0.944	0.947	0.944	0.947	0.947
different	low	4	-0.05	-0.051	-0.054	-0.050	-0.054	-0.051	-0.051	0.131	0.102	0.132	0.100	0.133	0.133	0.950	0.950	0.950	0.951	0.949	0.949
different	low	10	-0.05	-0.048	-0.049	-0.048	-0.048	-0.048	-0.048	0.218	0.152	0.216	0.147	0.221	0.221	0.948	0.950	0.950	0.952	0.949	0.949
different	no	2	0	0.000	-0.002	0.000	-0.002	0.000	0.000	0.066	0.055	0.067	0.055	0.068	0.068	0.934	0.946	0.934	0.946	0.932	0.932
different	no	4	0	0.002	0.000	0.002	0.000	0.002	0.002	0.057	0.047	0.058	0.046	0.056	0.056	0.943	0.954	0.943	0.955	0.945	0.945
different	no	10	0	0.000	-0.001	0.000	0.000	-0.001	-0.001	0.060	0.057	0.060	0.057	0.060	0.060	0.941	0.943	0.941	0.943	0.941	0.941

Table 59: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. The bias, its Monte Carlo (MC) standard error, and the root mean squared error (RMSE) are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a fixed effects model.*

Popu-	Effect	Trials	Bias						MC Error for Bias						RMSE					
lation	Size		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.021	-0.022	-0.020	-0.021	-0.028	-0.028	0.006	0.006	0.006	0.006	0.007	0.007	0.285	0.287	0.285	0.287	0.293	0.293
similar	high	4	-0.027	-0.027	-0.027	-0.027	-0.038	-0.038	0.005	0.005	0.005	0.005	0.005	0.005	0.202	0.203	0.202	0.203	0.210	0.210
similar	high	10	-0.011	-0.011	-0.010	-0.010	-0.022	-0.022	0.003	0.003	0.003	0.003	0.003	0.003	0.123	0.123	0.123	0.123	0.144	0.144
similar	moderate	2	-0.016	-0.016	-0.016	-0.016	-0.017	-0.017	0.003	0.003	0.003	0.003	0.003	0.003	0.130	0.130	0.130	0.130	0.132	0.132
similar	moderate	4	-0.010	-0.010	-0.010	-0.010	-0.011	-0.011	0.002	0.002	0.002	0.002	0.002	0.002	0.095	0.095	0.095	0.095	0.096	0.096
similar	moderate	10	-0.009	-0.009	-0.009	-0.009	-0.011	-0.011	0.001	0.001	0.001	0.001	0.001	0.001	0.058	0.058	0.058	0.058	0.059	0.059
similar	low	2	-0.004	-0.004	-0.004	-0.003	-0.004	-0.004	0.002	0.002	0.002	0.002	0.002	0.002	0.090	0.091	0.090	0.091	0.091	0.091
similar	low	4	0.000	0.000	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.064	0.064	0.064	0.064	0.064	0.064
similar	low	10	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.040	0.040	0.040	0.040	0.040	0.040
similar	no	2	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.082	0.082	0.082	0.082	0.082	0.082
similar	no	4	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.058	0.058	0.058	0.058	0.058	0.058
similar	no	10	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.035	0.035	0.035	0.035	0.036	0.036
different	high	2	-0.040	-0.060	-0.040	-0.060	-0.049	-0.049	0.007	0.010	0.007	0.010	0.007	0.007	0.298	0.439	0.298	0.439	0.309	0.309
different	high	4	-0.023	-0.036	-0.022	-0.036	-0.030	-0.030	0.005	0.006	0.005	0.006	0.005	0.005	0.201	0.285	0.201	0.284	0.211	0.211
different	high	10	-0.015	-0.018	-0.014	-0.017	-0.028	-0.028	0.003	0.004	0.003	0.004	0.004	0.004	0.123	0.181	0.123	0.181	0.158	0.158
different	moderate	2	-0.014	-0.019	-0.013	-0.018	-0.014	-0.014	0.003	0.004	0.003	0.004	0.003	0.003	0.138	0.193	0.138	0.193	0.140	0.140
different	moderate	4	-0.009	-0.010	-0.008	-0.010	-0.010	-0.010	0.002	0.003	0.002	0.003	0.002	0.002	0.096	0.134	0.096	0.134	0.097	0.097
different	moderate	10	-0.011	-0.010	-0.010	-0.010	-0.013	-0.013	0.001	0.002	0.001	0.002	0.001	0.001	0.061	0.085	0.061	0.085	0.063	0.063
different	low	2	0.000	-0.002	0.001	-0.001	0.000	0.000	0.002	0.003	0.002	0.003	0.002	0.002	0.095	0.120	0.095	0.120	0.095	0.095
different	low	4	-0.001	-0.004	0.000	-0.004	-0.001	-0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.065	0.085	0.065	0.085	0.065	0.065
different	low	10	0.002	0.001	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.040	0.053	0.040	0.053	0.040	0.040
different	no	2	0.000	-0.002	0.000	-0.002	0.000	0.000	0.002	0.002	0.002	0.002	0.002	0.002	0.085	0.105	0.085	0.104	0.085	0.085
different	no	4	0.002	0.000	0.002	0.000	0.002	0.002	0.001	0.002	0.001	0.002	0.001	0.001	0.058	0.073	0.058	0.072	0.058	0.058
different	no	10	0.000	-0.001	0.000	0.000	-0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.038	0.047	0.038	0.047	0.038	0.038

Table 60: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario I and II. Mean Effect, power, and coverage are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model. A variance value of $\sigma = 0.4$ is assumed in the generation of the true treatment effect.*

Population	Effect Size	Trials	True Effect	Mean Effect				Power				Coverage			
				Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.747	-0.748	-0.781	-0.781	0.807	0.805	0.431	0.431	0.866	0.862	0.946	0.946
similar	high	4	-0.69	-0.764	-0.764	-0.805	-0.805	0.833	0.834	0.453	0.453	0.856	0.856	0.939	0.939
similar	high	10	-0.69	-0.766	-0.766	-0.797	-0.797	0.853	0.853	0.448	0.448	0.859	0.861	0.943	0.943
similar	moderate	2	-0.22	-0.227	-0.227	-0.230	-0.230	0.514	0.513	0.249	0.249	0.863	0.862	0.933	0.933
similar	moderate	4	-0.22	-0.237	-0.237	-0.241	-0.241	0.591	0.593	0.295	0.295	0.860	0.863	0.941	0.941
similar	moderate	10	-0.22	-0.231	-0.231	-0.233	-0.233	0.603	0.602	0.285	0.285	0.854	0.851	0.939	0.939
similar	low	2	-0.05	-0.051	-0.051	-0.051	-0.051	0.181	0.181	0.111	0.108	0.859	0.861	0.913	0.915
similar	low	4	-0.05	-0.051	-0.051	-0.051	-0.051	0.220	0.222	0.111	0.111	0.845	0.847	0.922	0.922
similar	low	10	-0.05	-0.050	-0.050	-0.049	-0.049	0.241	0.241	0.118	0.118	0.824	0.822	0.908	0.908
similar	no	2	0	-0.001	-0.001	0.001	0.001	0.140	0.140	0.103	0.096	0.860	0.861	0.8975	0.904
similar	no	4	0	0.003	0.003	0.003	0.003	0.165	0.162	0.097	0.097	0.836	0.839	0.903	0.904
similar	no	10	0	0.000	0.000	0.001	0.001	0.175	0.176	0.086	0.086	0.825	0.824	0.914	0.914
different	high	2	-0.69	-0.763	-0.798	-0.797	-0.797	0.817	0.626	0.448	0.448	0.863	0.825	0.938	0.938
different	high	4	-0.69	-0.769	-0.808	-0.806	-0.806	0.840	0.652	0.460	0.460	0.865	0.816	0.943	0.943
different	high	10	-0.69	-0.769	-0.81	-0.815	-0.815	0.848	0.657	0.479	0.479	0.871	0.826	0.941	0.941
different	moderate	2	-0.22	-0.232	-0.232	-0.235	-0.235	0.5215	0.371	0.270	0.269	0.859	0.856	0.940	0.940
different	moderate	4	-0.22	-0.237	-0.236	-0.240	-0.240	0.582	0.4025	0.289	0.289	0.846	0.851	0.928	0.928
different	moderate	10	-0.22	-0.233	-0.238	-0.233	-0.233	0.595	0.423	0.274	0.274	0.855	0.8645	0.935	0.935
different	low	2	-0.05	-0.048	-0.049	-0.050	-0.050	0.167	0.166	0.105	0.101	0.872	0.8645	0.914	0.920
different	low	4	-0.05	-0.048	-0.050	-0.048	-0.048	0.209	0.191	0.119	0.119	0.839	0.844	0.913	0.913
different	low	10	-0.05	-0.057	-0.059	-0.057	-0.057	0.2435	0.190	0.125	0.125	0.847	0.857	0.916	0.916
different	no	2	0	0.003	0.003	0.002	0.002	0.133	0.137	0.090	0.078	0.868	0.864	0.910	0.922
different	no	4	0	-0.001	0.001	-0.001	-0.001	0.170	0.150	0.105	0.105	0.831	0.850	0.896	0.896
different	no	10	0	-0.001	-0.001	0.001	0.001	0.182	0.155	0.093	0.093	0.818	0.846	0.907	0.907

Table 61: *One IPD trial (individual patient data) and multiple AgD trials (aggregated data) - scenario I and II. The bias, its Monte Carlo (MC) standard error, and the root mean squared error (RMSE) are presented for approach A.1 (pooled AgD) and A.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model. A variance value of* $\sigma = 0.4$ *is assumed in the generation of the true treatment effect.*

Population	Effect Size	Trials	Bias				MC Error for Bias				RMSE			
			Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.057	-0.058	-0.091	-0.091	0.007	0.007	0.010	0.010	0.324	0.327	0.468	0.468
similar	high	4	-0.074	-0.074	-0.115	-0.115	0.007	0.007	0.011	0.011	0.334	0.335	0.487	0.487
similar	high	10	-0.076	-0.076	-0.107	-0.107	0.007	0.007	0.010	0.010	0.325	0.327	0.472	0.472
similar	moderate	2	-0.007	-0.007	-0.010	-0.010	0.003	0.003	0.004	0.004	0.145	0.146	0.190	0.190
similar	moderate	4	-0.017	-0.017	-0.021	-0.021	0.003	0.003	0.004	0.004	0.141	0.141	0.184	0.184
similar	moderate	10	-0.011	-0.011	-0.013	-0.013	0.003	0.003	0.004	0.004	0.137	0.137	0.179	0.179
similar	low	2	-0.001	-0.001	-0.001	-0.001	0.002	0.002	0.003	0.003	0.105	0.105	0.129	0.129
similar	low	4	-0.001	-0.001	-0.001	-0.001	0.002	0.002	0.003	0.003	0.095	0.095	0.120	0.120
similar	low	10	0.000	0.000	0.001	0.001	0.002	0.002	0.003	0.003	0.088	0.088	0.116	0.116
similar	no	2	-0.001	-0.001	0.001	0.001	0.002	0.002	0.003	0.003	0.095	0.095	0.113	0.113
similar	no	4	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.083	0.083	0.104	0.104
similar	no	10	0.000	0.000	0.001	0.001	0.002	0.002	0.002	0.002	0.074	0.074	0.095	0.095
different	high	2	-0.073	-0.108	-0.107	-0.107	0.007	0.011	0.011	0.011	0.334	0.481	0.480	0.480
different	high	4	-0.079	-0.118	-0.116	-0.116	0.007	0.011	0.011	0.011	0.331	0.489	0.493	0.493
different	high	10	-0.079	-0.120	-0.125	-0.125	0.007	0.010	0.010	0.010	0.320	0.473	0.481	0.481
different	moderate	2	-0.012	-0.012	-0.015	-0.015	0.003	0.005	0.004	0.004	0.153	0.206	0.191	0.191
different	moderate	4	-0.017	-0.016	-0.020	-0.020	0.003	0.005	0.004	0.004	0.143	0.201	0.189	0.189
different	moderate	10	-0.013	-0.018	-0.013	-0.013	0.003	0.004	0.004	0.004	0.135	0.190	0.181	0.181
different	low	2	0.002	0.001	0.000	0.000	0.002	0.003	0.003	0.003	0.106	0.134	0.130	0.130
different	low	4	0.002	0.000	0.002	0.002	0.002	0.003	0.003	0.003	0.099	0.128	0.123	0.123
different	low	10	-0.007	-0.009	-0.007	-0.007	0.002	0.003	0.003	0.003	0.087	0.116	0.114	0.114
different	no	2	0.003	0.003	0.002	0.002	0.002	0.003	0.003	0.003	0.096	0.118	0.114	0.114
different	no	4	-0.001	0.001	-0.001	-0.001	0.002	0.002	0.002	0.002	0.087	0.108	0.108	0.108
different	no	10	-0.001	-0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.077	0.102	0.099	0.099

Table 62: *One AgD trial (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. Mean effect, power, and coverage are presented for approach B.1 (pooled IPD) and B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. Results for scenario I and II are presented. All meta-analyses are based on a random effects model. A variance value of* $\sigma = 0.4$ *is assumed in the generation of the true treatment effect.*

Popu-	Effect	Trials	True	Mean Effect				Power				Coverage			
lation	Size		Effect	Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.742	-0.743	-0.751	-0.765	0.701	0.697	0.686	0.589	0.949	0.946	0.946	0.959
similar	high	4	-0.69	-0.748	-0.748	-0.761	-0.772	0.934	0.927	0.919	0.870	0.937	0.937	0.930	0.942
similar	high	10	-0.69	-0.734	-0.734	-0.758	-0.761	0.999	0.999	0.998	0.997	0.924	0.923	0.912	0.918
similar	moderate	2	-0.22	-0.226	-0.226	-0.227	-0.228	0.363	0.357	0.365	0.317	0.935	0.932	0.933	0.943
similar	moderate	4	-0.22	-0.235	-0.235	-0.237	-0.238	0.536	0.537	0.536	0.515	0.910	0.911	0.907	0.913
similar	moderate	10	-0.22	-0.227	-0.227	-0.231	-0.231	0.655	0.656	0.668	0.668	0.912	0.916	0.913	0.913
similar	low	2	-0.05	-0.051	-0.051	-0.052	-0.052	0.123	0.123	0.123	0.115	0.912	0.911	0.912	0.919
similar	low	4	-0.05	-0.051	-0.051	-0.052	-0.052	0.136	0.136	0.137	0.136	0.903	0.900	0.902	0.905
similar	low	10	-0.05	-0.050	-0.050	-0.051	-0.051	0.152	0.152	0.155	0.155	0.910	0.908	0.909	0.909
similar	no	2	0	-0.002	-0.002	-0.003	-0.003	0.097	0.096	0.097	0.090	0.904	0.905	0.904	0.910
similar	no	4	0	0.002	0.002	0.001	0.001	0.103	0.100	0.102	0.101	0.898	0.900	0.899	0.900
similar	no	10	0	-0.001	-0.001	-0.002	-0.002	0.087	0.086	0.085	0.085	0.913	0.914	0.915	0.915
different	high	2	-0.69	-0.759	-0.795	-0.766	-0.782	0.715	0.507	0.699	0.588	0.945	0.914	0.944	0.956
different	high	4	-0.69	-0.744	-0.763	-0.759	-0.772	0.931	0.721	0.931	0.881	0.937	0.932	0.932	0.941
different	high	10	-0.69	-0.741	-0.749	-0.765	-0.768	0.998	0.954	0.998	0.997	0.936	0.931	0.926	0.932
different	moderate	2	-0.22	-0.233	-0.234	-0.235	-0.235	0.384	0.246	0.387	0.340	0.919	0.924	0.918	0.930
different	moderate	4	-0.22	-0.232	-0.231	-0.235	-0.235	0.497	0.357	0.504	0.492	0.914	0.924	0.915	0.918
different	moderate	10	-0.22	-0.231	-0.233	-0.235	-0.235	0.632	0.530	0.640	0.640	0.931	0.933	0.931	0.931
different	low	2	-0.05	-0.050	-0.050	-0.050	-0.050	0.105	0.089	0.104	0.096	0.916	0.921	0.919	0.927
different	low	4	-0.05	-0.051	-0.052	-0.052	-0.052	0.130	0.104	0.130	0.128	0.906	0.922	0.906	0.907
different	low	10	-0.05	-0.052	-0.051	-0.053	-0.053	0.139	0.133	0.143	0.143	0.902	0.907	0.903	0.903
different	no	2	0	0.002	0.003	0.002	0.002	0.089	0.078	0.089	0.081	0.911	0.922	0.911	0.919
different	no	4	0	0.002	0.003	0.001	0.001	0.090	0.084	0.091	0.091	0.911	0.916	0.909	0.909
different	no	10	0	-0.001	0.000	-0.001	-0.001	0.083	0.078	0.082	0.082	0.918	0.922	0.919	0.919

Table 63: *One AgD trial (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. The bias, its Monte Carlo (MC) standard error, and the root mean squared error (RMSE) are presented for approach B.1 (pooled IPD) and B.2 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model. A variance value of* $\sigma = 0.4$ *is assumed in the generation of the true treatment effect. Results for are presented.*

Population	Effect Size	Trials	Bias				MC Error for Bias				RMSE			
			Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp		Pool AgD		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.052	-0.053	-0.061	-0.075	0.007	0.007	0.007	0.007	0.314	0.317	0.330	0.336
similar	high	4	-0.058	-0.058	-0.071	-0.082	0.005	0.005	0.005	0.005	0.236	0.237	0.250	0.256
similar	high	10	-0.044	-0.044	-0.068	-0.071	0.004	0.004	0.004	0.004	0.164	0.165	0.189	0.187
similar	moderate	2	-0.006	-0.006	-0.007	-0.008	0.003	0.004	0.004	0.004	0.154	0.155	0.155	0.156
similar	moderate	4	-0.015	-0.015	-0.017	-0.018	0.003	0.003	0.003	0.003	0.133	0.133	0.134	0.134
similar	moderate	10	-0.007	-0.007	-0.011	-0.011	0.002	0.002	0.002	0.002	0.107	0.107	0.108	0.108
similar	low	2	-0.001	-0.001	-0.002	-0.002	0.003	0.003	0.003	0.003	0.120	0.120	0.120	0.120
similar	low	4	-0.001	-0.001	-0.002	-0.002	0.002	0.002	0.002	0.002	0.108	0.108	0.108	0.108
similar	low	10	0.000	0.000	-0.001	-0.001	0.002	0.002	0.002	0.002	0.100	0.100	0.100	0.100
similar	no	2	-0.002	-0.002	-0.003	-0.003	0.003	0.003	0.003	0.003	0.111	0.111	0.111	0.111
similar	no	4	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.002	0.106	0.106	0.106	0.106
similar	no	10	-0.001	-0.001	-0.002	-0.002	0.002	0.002	0.002	0.002	0.096	0.096	0.096	0.096
different	high	2	-0.069	-0.105	-0.076	-0.092	0.007	0.010	0.007	0.007	0.323	0.465	0.334	0.342
different	high	4	-0.054	-0.073	-0.069	-0.082	0.005	0.007	0.005	0.006	0.230	0.323	0.252	0.279
different	high	10	-0.051	-0.059	-0.075	-0.078	0.004	0.005	0.004	0.004	0.163	0.216	0.179	0.181
different	moderate	2	-0.013	-0.014	-0.015	-0.015	0.004	0.005	0.004	0.004	0.164	0.213	0.164	0.165
different	moderate	4	-0.012	-0.011	-0.015	-0.015	0.003	0.004	0.003	0.003	0.134	0.165	0.135	0.136
different	moderate	10	-0.011	-0.013	-0.015	-0.015	0.002	0.003	0.002	0.002	0.108	0.122	0.109	0.109
different	low	2	0.000	0.000	0.000	0.000	0.003	0.003	0.003	0.003	0.122	0.144	0.122	0.122
different	low	4	-0.001	-0.002	-0.002	-0.002	0.003	0.003	0.003	0.003	0.112	0.122	0.112	0.112
different	low	10	-0.002	-0.001	-0.003	-0.003	0.002	0.003	0.002	0.002	0.104	0.111	0.105	0.105
different	no	2	0.002	0.003	0.002	0.002	0.003	0.003	0.003	0.003	0.115	0.133	0.115	0.115
different	no	4	0.002	0.003	0.001	0.001	0.002	0.003	0.002	0.002	0.106	0.115	0.106	0.106
different	no	10	-0.001	0.000	-0.001	-0.001	0.002	0.002	0.002	0.002	0.099	0.102	0.099	0.099

Table 64: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. Mean effect, power, and coverage are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model. A variance value of* $\sigma = 0.4$ *is assumed in the generation of the true treatment effect.*

Popu-	Effect	Trials	True	Mean Effect						Power						Coverage					
lation	Size		Effect	All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All	
				Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.69	-0.743	-0.744	-0.743	-0.743	-0.751	-0.755	0.715	0.710	0.714	0.708	0.694	0.684	0.953	0.949	0.953	0.950	0.942	0.944
similar	high	4	-0.69	-0.748	-0.748	-0.747	-0.747	-0.757	-0.757	0.958	0.956	0.958	0.955	0.947	0.946	0.935	0.936	0.935	0.936	0.934	0.934
similar	high	10	-0.69	-0.734	-0.734	-0.733	-0.733	-0.746	-0.746	0.100	0.100	0.100	0.100	0.100	0.100	0.940	0.941	0.941	0.942	0.929	0.929
similar	moderate	2	-0.22	-0.226	-0.227	-0.226	-0.226	-0.227	-0.227	0.410	0.411	0.407	0.404	0.406	0.399	0.936	0.938	0.937	0.939	0.931	0.932
similar	moderate	4	-0.22	-0.235	-0.235	-0.234	-0.234	-0.236	-0.236	0.698	0.698	0.697	0.694	0.692	0.692	0.919	0.918	0.919	0.921	0.913	0.913
similar	moderate	10	-0.22	-0.228	-0.228	-0.227	-0.227	-0.229	-0.229	0.966	0.967	0.966	0.966	0.966	0.966	0.934	0.934	0.934	0.935	0.933	0.933
similar	low	2	-0.05	-0.051	-0.051	-0.050	-0.050	-0.051	-0.051	0.132	0.132	0.131	0.118	0.130	0.128	0.913	0.910	0.913	0.919	0.913	0.916
similar	low	4	-0.05	-0.049	-0.049	-0.049	-0.049	-0.049	-0.049	0.179	0.177	0.179	0.166	0.179	0.179	0.908	0.909	0.907	0.915	0.909	0.909
similar	low	10	-0.05	-0.050	-0.050	-0.049	-0.049	-0.050	-0.050	0.274	0.272	0.268	0.265	0.271	0.271	0.917	0.916	0.914	0.917	0.917	0.917
similar	no	2	0	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.096	0.096	0.095	0.079	0.097	0.095	0.905	0.905	0.905	0.921	0.904	0.906
similar	no	4	0	0.003	0.003	0.003	0.003	0.003	0.003	0.094	0.095	0.093	0.082	0.092	0.092	0.906	0.906	0.907	0.918	0.909	0.909
similar	no	10	0	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.109	0.109	0.107	0.099	0.110	0.110	0.892	0.891	0.893	0.901	0.890	0.890
different	high	2	-0.69	-0.761	-0.797	-0.760	-0.796	-0.766	-0.771	0.737	0.514	0.736	0.514	0.715	0.705	0.944	0.915	0.945	0.914	0.943	0.944
different	high	4	-0.69	-0.747	-0.766	-0.746	-0.765	-0.757	-0.759	0.962	0.752	0.962	0.751	0.949	0.949	0.940	0.926	0.941	0.926	0.932	0.932
different	high	10	-0.69	-0.741	-0.749	-0.740	-0.748	-0.752	-0.752	1.000	0.978	1.000	0.978	1.000	1.000	0.932	0.927	0.932	0.928	0.924	0.924
different	moderate	2	-0.22	-0.231	-0.232	-0.230	-0.231	-0.232	-0.232	0.420	0.262	0.419	0.261	0.422	0.411	0.917	0.927	0.918	0.928	0.918	0.919
different	moderate	4	-0.22	-0.232	-0.231	-0.231	-0.230	-0.233	-0.233	0.676	0.442	0.671	0.439	0.671	0.671	0.931	0.939	0.932	0.938	0.931	0.931
different	moderate	10	-0.22	-0.229	-0.231	-0.228	-0.230	-0.230	-0.230	0.972	0.789	0.971	0.785	0.968	0.968	0.941	0.953	0.940	0.953	0.939	0.939
different	low	2	-0.05	-0.048	-0.048	-0.048	-0.047	-0.048	-0.048	0.121	0.097	0.120	0.089	0.122	0.118	0.917	0.928	0.917	0.934	0.918	0.920
different	low	4	-0.05	-0.049	-0.050	-0.049	-0.049	-0.049	-0.049	0.169	0.123	0.167	0.120	0.168	0.168	0.912	0.929	0.910	0.931	0.911	0.911
different	low	10	-0.05	-0.053	-0.053	-0.053	-0.052	-0.053	-0.053	0.293	0.196	0.289	0.194	0.294	0.294	0.911	0.926	0.909	0.926	0.909	0.909
different	no	2	0	0.003	0.003	0.003	0.004	0.003	0.003	0.098	0.080	0.098	0.075	0.098	0.093	0.903	0.920	0.903	0.926	0.902	0.907
different	no	4	0	-0.001	0.000	-0.001	0.001	-0.001	-0.001	0.104	0.078	0.103	0.077	0.101	0.101	0.896	0.923	0.897	0.923	0.899	0.899
different	no	10	0	-0.001	-0.001	-0.001	0.000	-0.001	-0.001	0.090	0.074	0.090	0.073	0.090	0.090	0.911	0.926	0.911	0.928	0.911	0.911

Table 65: *Multiple AgD trials (aggregated data) and multiple IPD trials (individual patient data) - scenario I and II. The bias, its Monte Carlo (MC) standard error, and the root mean squared error (RMSE) are presented for approach C.1 (pooled IPD, pooled AgD), C.2 (pooled IPD, all indirect comparisons), and C.3 (all indirect comparisons) and both methods for indirect comparisons, Bucher and MAIC. All meta-analyses are based on a random effects model. A variance value of $\sigma = 0.4$ is assumed in the generation of the true treatment effect.*

Popu-	Effect	Trials	Bias						MC Error for Bias						RMSE					
lation	Size		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp		Pooled IPD, AgD		Pooled IPD, All		All Ind. Comp	
			Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC	Bucher	MAIC
similar	high	2	-0.053	-0.054	-0.053	-0.053	-0.061	-0.065	0.007	0.007	0.007	0.007	0.007	0.007	0.306	0.309	0.306	0.309	0.321	0.322
similar	high	4	-0.058	-0.058	-0.057	-0.057	-0.067	-0.067	0.005	0.005	0.005	0.005	0.005	0.005	0.221	0.221	0.221	0.221	0.235	0.235
similar	high	10	-0.044	-0.044	-0.043	-0.043	-0.056	-0.056	0.003	0.003	0.003	0.003	0.003	0.003	0.136	0.136	0.135	0.136	0.163	0.157
similar	moderate	2	-0.006	-0.007	-0.006	-0.006	-0.007	-0.007	0.003	0.003	0.003	0.003	0.003	0.003	0.139	0.140	0.139	0.140	0.141	0.141
similar	moderate	4	-0.015	-0.015	-0.014	-0.014	-0.016	-0.016	0.002	0.002	0.002	0.002	0.002	0.002	0.103	0.103	0.103	0.103	0.104	0.104
similar	moderate	10	-0.008	-0.008	-0.007	-0.007	-0.009	-0.009	0.001	0.001	0.001	0.001	0.001	0.001	0.063	0.063	0.063	0.063	0.064	0.064
similar	low	2	-0.001	-0.001	0.000	0.000	-0.001	-0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.101	0.102	0.101	0.102	0.102	0.102
similar	low	4	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.073	0.073	0.073	0.073	0.073	0.074
similar	low	10	0.000	0.000	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.044	0.044	0.044	0.044	0.045	0.045
similar	no	2	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.092	0.092	0.092	0.092	0.092	0.092
similar	no	4	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.065	0.065	0.065	0.065	0.065	0.065
similar	no	10	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.042	0.042	0.042	0.042	0.042	0.042
different	high	2	-0.071	-0.107	-0.070	-0.106	-0.076	-0.081	0.007	0.010	0.007	0.010	0.007	0.007	0.317	0.461	0.317	0.461	0.328	0.330
different	high	4	-0.057	-0.076	-0.056	-0.075	-0.067	-0.069	0.005	0.007	0.005	0.007	0.005	0.0057	0.217	0.313	0.217	0.313	0.239	0.263
different	high	10	-0.051	-0.059	-0.050	-0.058	-0.062	-0.062	0.003	0.004	0.003	0.004	0.003	0.003	0.141	0.201	0.141	0.200	0.153	0.153
different	moderate	2	-0.011	-0.012	-0.010	-0.011	-0.012	-0.012	0.003	0.005	0.003	0.005	0.003	0.003	0.149	0.200	0.149	0.200	0.150	0.150
different	moderate	4	-0.012	-0.011	-0.011	-0.010	-0.013	-0.013	0.002	0.003	0.002	0.003	0.002	0.002	0.103	0.141	0.103	0.141	0.104	0.104
different	moderate	10	-0.009	-0.011	-0.008	-0.010	-0.010	-0.010	0.001	0.002	0.001	0.002	0.001	0.001	0.061	0.083	0.061	0.083	0.062	0.062
different	low	2	0.002	0.002	0.002	0.003	0.002	0.002	0.002	0.003	0.002	0.003	0.002	0.002	0.102	0.129	0.102	0.129	0.102	0.102
different	low	4	0.001	0.000	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.074	0.091	0.074	0.091	0.075	0.075
different	low	10	-0.003	-0.003	-0.003	-0.002	-0.003	-0.003	0.001	0.001	0.001	0.001	0.001	0.001	0.047	0.059	0.047	0.059	0.048	0.048
different	no	2	0.003	0.003	0.003	0.004	0.003	0.003	0.002	0.003	0.002	0.003	0.002	0.002	0.094	0.115	0.094	0.115	0.094	0.094
different	no	4	-0.001	0.000	-0.001	0.001	-0.001	-0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.068	0.081	0.068	0.081	0.068	0.068
different	no	10	-0.001	-0.001	-0.001	0.000	-0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.042	0.050	0.042	0.050	0.042	0.042

Appendix B: Additional Methodological Background

MAIC - Method of Moments

The method of moments is applied to estimate $\boldsymbol{\beta}$ in the following equation

$$\omega_i = \exp(\beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta})$$

such that the mean baseline characteristics of the IPD matches the AgD data. Therefore, the estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ solves the following equation

$$0 = \frac{\sum_{i:t_i \in AB}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{CB}) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{i:t_i \in AB} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}, \tag{B.1}$$

where t_i is the i^{th} patient in trial AB for $i = 1, \ldots, n_{AB}$. By definition the correct weights for balancing the populations are provided by this formula because a logistic model ($\boldsymbol{x}_i$ contains all confounders and a correctly specified model) provides a consistent estimate of the treatment effect.

The finite solutions to formula B.1 are unique and will converge to the true $\boldsymbol{\beta}$ if the logistic regression model is correctly specified. To see this first note that

$$0 = \frac{\sum_{i:t_i \in AB}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{CB}) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\sum_{i:t_i \in AB} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \qquad \Longleftrightarrow \qquad 0 = \sum_{i:t_i \in AB} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{CB}) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

It can be assumed, that $\overline{\boldsymbol{x}}_{CB} = 0$ (normalize all baseline characteristics by $\overline{\boldsymbol{x}}_{CB}$). This results in the following equation

$$0 = \sum_{i:t_i \in AB} \boldsymbol{x}_i \exp(\boldsymbol{x}_i' \boldsymbol{\beta}). \tag{B.2}$$

The right hand side of equation B.2 is a function of $\boldsymbol{\beta}$ and a possible first derivative of the following function

$$Q(\boldsymbol{\beta}) = \sum_{i:t_i \in AB} \exp(\boldsymbol{x}_i' \boldsymbol{\beta}).$$

The second derivative of $Q(\boldsymbol{\beta})$ is given by

$$Q''(\boldsymbol{\beta}) = \sum_{i:t_i \in AB} \boldsymbol{x}_i \boldsymbol{x}_i' \exp(\boldsymbol{x}_i' \boldsymbol{\beta}).$$

This function $Q''(\boldsymbol{\beta})$ is positive definite for all $\boldsymbol{\beta}$, hence $Q(\boldsymbol{\beta})$ is convex and any finite solution of Equation B.1 is unique and coincide with the global minimum of $Q(\boldsymbol{\beta})$ (Signorovitch et al., 2010).

Appendix C: Implementations in R

In this chapter, selected R codes used for the simulations are given.

C.1 Generation of Evidence - Resampling CI Method

The following R function includes the data generating process, the naïve approach, and the resampling CI method.

```
##############################################################################
# Function for simulation of resampling CI method and naive approach
# The data generation process is included in the function
##############################################################################

matching_resamplingCI_fct = function(
  # number of simulation runs
  n_sim=10000,
  ## data generation process
  # regression coefficients for group model (logistic model)
  alpha0 = -.15,
  a_age = .002,
  a_sex = 0.02,
  a_diab = 0.026,
  # regression coefficients for outcome model (logistic model)
  beta0 = -0.5,
  b_age = -0.07,
```

```
  b_ASPECTS = 0.2,
  b_NIHSS = 0.1,
  b_group = 1,
  # number of patients in group 0
  n = 100,

  # distribution parameters equal in both groups
  age_mean = 70,
  age_sd = 15,
  sex_p = 0.5,
  diab_p = 0.2,
  # distribution parameters for ("new") treatment group
  group1_ASPECTS_p = .75,
  group1_NIHSS_mean = 16,
  group1_NIHSS_sd = 4,
  # distribution parameters for (already existing) control group
  group0_ASPECTS_p = .8,
  group0_NIHSS_mean = 17,
  group0_NIHSS_sd = 5,

  # fraction of patients for interims analysis (a vector if multiple time
      points for the interim analysis)
  interim = 0.5,
  # number of matching partners: 1:k matching
  k = 1,

  # caliper for propensity score matching
  ca = 0.2,
  # number of resampling steps
  boot = 200,
  # quantile for resampling CI method
  q_CI = .99
){

  require(Matching)
  require(boot)

  # calculating the number of patients recruited for the interim analysis
  n_interim = interim * n * k
```

```
n_interim_group0 = interim * n

# initializing relevant outcome values
# interim analyis
p_value_interim_H0_naiv = rep(NA, n_sim)
p_value_interim_H1_naiv = rep(NA, n_sim)
p_value_interim_H0 = rep(NA, n_sim)
p_value_interim_H1 = rep(NA, n_sim)
matching_rate = rep(NA, n_sim)
matching_rate_sd = rep(NA, n_sim)
matching_rate_interim_naiv = rep(NA, n_sim)

# number of patients added to the number used at interim analysis
n_add_naiv = rep(NA, n_sim)

# naive approach
n_naiv = rep(NA, n_sim)
p_value_naiv_H0 = rep(NA, n_sim)
p_value_naiv_H1 = rep(NA, n_sim)
matching_rate_naiv = rep(NA, n_sim)

# resampling CI method
n_CI = rep(NA, n_sim)
p_value_CI_H0 = rep(NA, n_sim)
p_value_CI_H1 = rep(NA, n_sim)
matching_rate_CI = rep(NA, n_sim)

results = list()

for(i in 1:length(n_interim)){

  for(s in 1:n_sim){
    # set seed
    set.seed(123456+s)

    # initializing data set
    # generating a big data set (for having enough patients for the final
        analysis)
    data = matrix(NA, ncol=8, nrow=n*40*k*2)
```

```
colnames(data) = c("sex", "diab", "age", "group", "ASPECTS",
                   "NIHSS_Aufnahme", "NIHSS_dif_H0", "NIHSS_dif_H1")

# baseline variables used in group model: e.g. sex, diabetes, and age
data[,1] = sample(c(0,1), n*40*k*2, prob=c(sex_p, 1-sex_p), replace=T)
data[,2] = sample(c(0,1), n*40*k*2, prob=c(1-diab_p, diab_p), replace=T)
data[,3] = rnorm(n*40, mean=age_mean, sd=age_sd)

# group model
data[,4] = rbinom(n*40*k*2,1, inv.logit(alpha0 + a_sex* data[,1] +
                                        a_diab * data[,2] +
                                        a_age * data[,3]))
# order data by group
data = data[order(data[,4]),]

n_group0 = length(which(data[,4]==0))
n_group1 = length(which(data[,4]==1))

# baseline variables used in outcome model: e.g. ASPECTS, NIHSS
data[,5] = c(rbinom(n_group0, 10, group0_ASPECTS_p),
             rbinom(n_group1, 10, group1_ASPECTS_p))
data[,6] = c(round(rnorm(n_group0, mean=group0_NIHSS_mean,
                         sd=group0_NIHSS_sd)),
             round(rnorm(n_group1, mean=group1_NIHSS_mean,
                         sd=group1_NIHSS_sd)))

# outcome model under H0
data[,7] = c(sample(c(0,1), n_group0, prob=c(0.5, 0.5), replace=T),
             sample(c(0,1), n_group1, prob=c(0.5, 0.5), replace=T))

# outcome model under H1
data[,8] = rbinom(n*40*k*2,1, inv.logit(beta0 +
                                        b_group * data[,4] +
                                        b_age * data[,3] +
                                        b_ASPECTS * data[,5]
))

#####################################################################
# naive approach
```

```
#########################################################################

# interim analysis
data_matching = as.data.frame(data[c(1:n, (n_group0+1):(n_group0+ceiling(
    n_interim[i]))),])

# model for propensity score
fit1 <- glm(group ~ diab + NIHSS_Aufnahme + age, data = data_matching,
    family = "binomial")

glm.fitted <- log(fit1$fitted / (1 - fit1$fitted))

# propensity score matching
rr <- Match(Tr = data_matching$group, X = glm.fitted,
            replace = F, M = k, ties = F, caliper = ca)

# McNemar Test for H1 (difference exists)
test_interim_H1_naiv = mcnemar.test(matrix(c(
  sum(ifelse(data_matching$NIHSS_dif_H1[rr$index.control]<1 &
               data_matching$NIHSS_dif_H1[rr$index.treated]<1,1,0)),
  sum(ifelse(data_matching$NIHSS_dif_H1[rr$index.control]>0 &
               data_matching$NIHSS_dif_H1[rr$index.treated]<1,1,0)),
  sum(ifelse(data_matching$NIHSS_dif_H1[rr$index.control]<1 &
               data_matching$NIHSS_dif_H1[rr$index.treated]>0,1,0)),
  sum(ifelse(data_matching$NIHSS_dif_H1[rr$index.control]>0 &
               data_matching$NIHSS_dif_H1[rr$index.treated]>0,1,0))),
  nrow=2),
  correct=F)

# McNemar Test for H0 (no difference exists)
test_interim_H0_naiv = mcnemar.test(matrix(c(
  sum(ifelse(data_matching$NIHSS_dif_H0[rr$index.control]<1 &
               data_matching$NIHSS_dif_H0[rr$index.treated]<1,1,0)),
  sum(ifelse(data_matching$NIHSS_dif_H0[rr$index.control]>0 &
               data_matching$NIHSS_dif_H0[rr$index.treated]<1,1,0)),
  sum(ifelse(data_matching$NIHSS_dif_H0[rr$index.control]<1 &
               data_matching$NIHSS_dif_H0[rr$index.treated]>0,1,0)),
  sum(ifelse(data_matching$NIHSS_dif_H0[rr$index.control]>0 &
               data_matching$NIHSS_dif_H0[rr$index.treated]>0,1,0))),
```

```
    nrow=2),
    correct=F)

  # save p-values of interim analysis
  p_value_interim_H0_naiv[s] = test_interim_H0_naiv$p.value
  p_value_interim_H1_naiv[s] = test_interim_H1_naiv$p.value

  # calculate matching rate at interim analysis
  matching_rate_interim_naiv[s] = (n_interim[i]-rr$ndrops)/n_interim[i]
  # number of patients additionally needed for final analysis
  n_add_naiv[s] = ceiling(n/matching_rate_interim_naiv[s] - n_interim[i])

  rm(rr, glm.fitted, fit)

  # recruit additional data for naiv method
  data_all_naiv = as.data.frame(
    data[c(1:n, (n_group0+1):(n_group0+n_interim[i] +ceiling(n_add_naiv[s])
        )),]
    )

  ##############################################
  # final analysis

  # propensity score matching with complete data
  fit2 <- glm(group ~ diab +NIHSS_Aufnahme +age, data = data_all_naiv,
      family = "binomial")

  glm.fitted2 <- log(fit2$fitted / (1 - fit2$fitted))

  rr2 <- Match(Tr = data_all_naiv$group, X = glm.fitted2,
               replace = F, M = k, ties = F, caliper = ca)

  # McNemar Test for H1 (difference exists)
  test_naiv_H1 = mcnemar.test(matrix(c(
    sum(ifelse(data_all_naiv$NIHSS_dif_H1[rr2$index.control]<1 &
                 data_all_naiv$NIHSS_dif_H1[rr2$index.treated]<1,1,0)),
    sum(ifelse(data_all_naiv$NIHSS_dif_H1[rr2$index.control]>0 &
                 data_all_naiv$NIHSS_dif_H1[rr2$index.treated]<1,1,0)),
    sum(ifelse(data_all_naiv$NIHSS_dif_H1[rr2$index.control]<1 &
```

```
                data_all_naiv$NIHSS_dif_H1[rr2$index.treated]>0,1,0)),
  sum(ifelse(data_all_naiv$NIHSS_dif_H1[rr2$index.control]>0 &
                data_all_naiv$NIHSS_dif_H1[rr2$index.treated]>0,1,0))),
  nrow=2),
  correct=F)

# McNemar Test for H0 (no difference exists)
test_naiv_H0 = mcnemar.test(matrix(c(
  sum(ifelse(data_all_naiv$NIHSS_dif_H0[rr2$index.control]<1 &
                data_all_naiv$NIHSS_dif_H0[rr2$index.treated]<1,1,0)),
  sum(ifelse(data_all_naiv$NIHSS_dif_H0[rr2$index.control]>0 &
                data_all_naiv$NIHSS_dif_H0[rr2$index.treated]<1,1,0)),
  sum(ifelse(data_all_naiv$NIHSS_dif_H0[rr2$index.control]<1 &
                data_all_naiv$NIHSS_dif_H0[rr2$index.treated]>0,1,0)),
  sum(ifelse(data_all_naiv$NIHSS_dif_H0[rr2$index.control]>0 &
                data_all_naiv$NIHSS_dif_H0[rr2$index.treated]>0,1,0))),
  nrow=2),
  correct=F)

# save p-values of final analysis
p_value_naiv_H1[s] = test_naiv_H1$p.value
p_value_naiv_H0[s] = test_naiv_H0$p.value

# matching rate at final analysis
matching_rate_naiv[s] = length(rr2$index.treated)/n

rm(rr2, glm.fitted2, fit2)

##########################################################################
# resamping approach of matching rate at interim analysis
##########################################################################

# interim analysis
matching_rate_boot = rep(NA, boot)

for(b in 1:boot){
  # sample data for matching (groups of equal size)
  data_matching = as.data.frame(
    data[c(sample(1:n, ceiling(n_interim_group0[i]), replace=F),
```

```
          (n_group0+1):(n_group0+ceiling(n_interim[i]))),])

    # model for propensity score
    fit3 <- glm(group ~ diab + NIHSS_Aufnahme + age,
                data = data_matching,
                family = "binomial")

    glm.fitted3 <- log(fit3$fitted / (1 - fit3$fitted))

    # propensity score matching
    rr3 <- Match(Tr = data_matching$group, X = glm.fitted3,
                 replace = F, M = k, ties = F, caliper = ca)

    # McNemar Test for H1 (difference exists)
    test_interim_H1 = mcnemar.test(matrix(c(
      sum(ifelse(data_matching$NIHSS_dif_H1[rr3$index.control]<1 &
                   data_matching$NIHSS_dif_H1[rr3$index.treated]<1,1,0)),
      sum(ifelse(data_matching$NIHSS_dif_H1[rr3$index.control]>0 &
                   data_matching$NIHSS_dif_H1[rr3$index.treated]<1,1,0)),
      sum(ifelse(data_matching$NIHSS_dif_H1[rr3$index.control]<1 &
                   data_matching$NIHSS_dif_H1[rr3$index.treated]>0,1,0)),
      sum(ifelse(data_matching$NIHSS_dif_H1[rr3$index.control]>0 &
                   data_matching$NIHSS_dif_H1[rr3$index.treated]>0,1,0))),
      nrow=2),
      correct=F)

    # McNemar Test for H0 (no difference exists)
    test_interim_H0 = mcnemar.test(matrix(c(
      sum(ifelse(data_matching$NIHSS_dif_H0[rr3$index.control]<1 &
                   data_matching$NIHSS_dif_H0[rr3$index.treated]<1,1,0)),
      sum(ifelse(data_matching$NIHSS_dif_H0[rr3$index.control]>0 &
                   data_matching$NIHSS_dif_H0[rr3$index.treated]<1,1,0)),
      sum(ifelse(data_matching$NIHSS_dif_H0[rr3$index.control]<1 &
                   data_matching$NIHSS_dif_H0[rr3$index.treated]>0,1,0)),
      sum(ifelse(data_matching$NIHSS_dif_H0[rr3$index.control]>0 &
                   data_matching$NIHSS_dif_H0[rr3$index.treated]>0,1,0))),
      nrow=2),
      correct=F)
```

```
    p_value_interim_H0[s] = test_interim_H0$p.value
    p_value_interim_H1[s] = test_interim_H1$p.value

    # matching rate
    matching_rate_boot[b] = (n_interim[i]-rr3$ndrops)/n_interim[i]

  }

  # mean/sd of resampled matching rates at interim analysis
  matching_rate[s] = mean(matching_rate_boot)
  matching_rate_sd[s] = sd(matching_rate_boot)

  # Lower CI limit to calculate the number of patients additionally needed
      for final analysis
  CI_lowerLimit = matching_rate[s] -
    qnorm(q_CI) * sqrt((matching_rate[s]*(1-matching_rate[s])))/n)
  n_CI[s] = ceiling(n/CI_lowerLimit - n_interim[i])

  rm(rr3, glm.fitted3, fit3)

  # recruit missing data for CI method
  data_all_CI = as.data.frame(
    data[c(1:n, (n_group0+1):(n_group0+n_interim[i] +n_CI[s])),]
    )

  ################################################
  # final analysis

  # Propensity Score matching with complete data
  fit4 <- glm(group ~ NIHSS_Aufnahme + age + diab,
              data = data_all_CI,
              family = "binomial")

  glm.fitted4 <- log(fit4$fitted / (1 - fit4$fitted))

  rr4 <- Match(Tr = data_all_CI$group, X = glm.fitted4,
               replace = F, M = k, ties = F, caliper = ca)

  # McNemar Test for H1 (difference exists)
```

```
test_CI_H1 = mcnemar.test(matrix(c(
  sum(ifelse(data_all_CI$NIHSS_dif_H1[rr4$index.control]<1 &
               data_all_CI$NIHSS_dif_H1[rr4$index.treated]<1,1,0)),
  sum(ifelse(data_all_CI$NIHSS_dif_H1[rr4$index.control]>0 &
               data_all_CI$NIHSS_dif_H1[rr4$index.treated]<1,1,0)),
  sum(ifelse(data_all_CI$NIHSS_dif_H1[rr4$index.control]<1 &
               data_all_CI$NIHSS_dif_H1[rr4$index.treated]>0,1,0)),
  sum(ifelse(data_all_CI$NIHSS_dif_H1[rr4$index.control]>0 &
               data_all_CI$NIHSS_dif_H1[rr4$index.treated]>0,1,0))),
  nrow=2),
  correct=F)

# McNemar Test for H0 (no difference exists)
test_CI_H0 = mcnemar.test(matrix(c(
  sum(ifelse(data_all_CI$NIHSS_dif_H0[rr4$index.control]<1 &
               data_all_CI$NIHSS_dif_H0[rr4$index.treated]<1,1,0)),
  sum(ifelse(data_all_CI$NIHSS_dif_H0[rr4$index.control]>0 &
               data_all_CI$NIHSS_dif_H0[rr4$index.treated]<1,1,0)),
  sum(ifelse(data_all_CI$NIHSS_dif_H0[rr4$index.control]<1 &
               data_all_CI$NIHSS_dif_H0[rr4$index.treated]>0,1,0)),
  sum(ifelse(data_all_CI$NIHSS_dif_H0[rr4$index.control]>0 &
               data_all_CI$NIHSS_dif_H0[rr4$index.treated]>0,1,0))),
  nrow=2),
  correct=F)

# save p-values of final analysis
p_value_CI_H0[s] = test_CI_H0$p.value
p_value_CI_H1[s] = test_CI_H1$p.value

# matching rate of final analysis
matching_rate_CI[s] = length(rr4$index.treated)/n

rm(rr4, glm.fitted4, fit4)

##############################################
# save results

if(s==n_sim){
  results[[i]] = list(
```

```
          p_value_interim_H0 = p_value_interim_H0,
          p_value_interim_H1 = p_value_interim_H1,
          p_value_interim_H0_naiv = p_value_interim_H0_naiv,
          p_value_interim_H1_naiv = p_value_interim_H1_naiv,

          matching_rate = matching_rate,
          matching_rate_sd = matching_rate_sd,
          matching_rate_interim_naiv = matching_rate_interim_naiv,
          matching_rate_interim_CI = CI_lowerLimit,

          n_naiv = n_add_naiv+n_interim[i],
          p_value_naiv_H0 = p_value_naiv_H0,
          p_value_naiv_H1 = p_value_naiv_H1,
          matching_rate_naiv = matching_rate_naiv,

          n_CI = n_CI+n_interim[i],
          p_value_CI_H0 = p_value_CI_H0,
          p_value_CI_H1 = p_value_CI_H1,
          matching_rate_CI = matching_rate_CI
        )
      }
    }
  }
  return(results)
}
```

C.2 Generation of Evidence - Iterative Matching Procedure

The following R functions include the data generating process (data_fct), the function performing the iterative matching (iterative_matching_fct), and one shell function (Simulation) applying the data function and the iterative matching procedure function.

```
#############################################################################
# Functions for simulation of iterative matching procedure
# First function: data generation
# Second function: iterative matching procedure
```

```
# Third function: simulation function (using 1st and 2nd function)
################################################################################

# function for data generation

data_fct <- function(
  # control data
  # binary confounder 1, 2 (here: FLT3, Zyto high) joint distribution
  # vector: no/no, no/yes, yes/no, yes/yes
  binary_proportions_control = c(.48, .32, .18, .02),
  binary_proportions_intervention = c(.48, .32, .18, .02),

  # continuous confounder normally distributed
  cont_mean_control = 55,
  cont_sd_control = 15,
  cont_mean_intervention = 55,
  cont_sd_intervention = 15,

  # outcome model coefficients
  outcome_intercept = 2,
  outcome_binary1 = -.2,
  outcome_binary2 = -.5,
  outcome_cont = -.05,
  outcome_interaction = 0,
  outcome_intervention=log(0.7/0.3),

  # sample size
  n_control = 1000,
  n_intervention = 100,

  # random number seed
  seed=1234
){
  require(boot)
  set.seed(seed)

  # total number of observations
  n = sum(n_control, n_intervention)
```

```
# initializing data tables
data_control = matrix(NA, ncol=5, nrow = n_control)
data_intervention = matrix(NA, ncol=5, nrow = n_intervention)

# group variable
data_control[,1] = rep(0, n_control)
data_intervention[,1] = rep(1, n_intervention)

# binary variables in control group
data_control[, 2] <- sample(c(0,1), n_control, replace=T,
                            prob=c(sum(binary_proportions_control[1:2]),
                                   sum(binary_proportions_control[3:4])))
tmp_rows_control <- which(data_control[, 2]==0)
data_control[tmp_rows_control, 3] <- sample(
  c(0,1), length(tmp_rows_control), replace=T,
  prob=c(binary_proportions_control[1], binary_proportions_control[2]))
data_control[-tmp_rows_control, 3] <- sample(
  c(0,1), n_control-length(tmp_rows_control), replace=T,
  prob=c(binary_proportions_control[3], binary_proportions_control[4]))
rm("tmp_rows_control")

# binary variables in intervention group
data_intervention[, 2] <- sample(
  c(0,1), n_intervention, replace=T,
  prob=c(sum(binary_proportions_intervention[1:2]),
         sum(binary_proportions_intervention[3:4])))
tmp_rows_intervention <- which(data_intervention[, 2]==0)
data_intervention[tmp_rows_intervention, 3] <- sample(
  c(0,1), length(tmp_rows_intervention), replace=T,
  prob=c(binary_proportions_intervention[1],
         binary_proportions_intervention[2]))
data_intervention[-tmp_rows_intervention, 3] <- sample(
  c(0,1), n_intervention-length(tmp_rows_intervention), replace=T,
  prob=c(binary_proportions_intervention[3],
         binary_proportions_intervention[4]))
rm("tmp_rows_intervention")

# continuous variable
data_control[,4] <- rnorm(n_control, cont_mean_control, cont_sd_control)
```

```
  data_intervention[,4] <- rnorm(n_intervention, cont_mean_intervention,
                                 cont_sd_intervention)

  data <- rbind(data_control, data_intervention)

  # outcome generation
  data[,5] <- rbinom(n,1, inv.logit(outcome_intercept +
                                    outcome_intervention*data[,1]+
                                    outcome_interaction*data[,1]*data[,4]+
                                    outcome_binary1 * data[,2] +
                                    outcome_binary2 * data[,3] +
                                    outcome_cont * data[,4]))

  colnames(data) <- c("group", "binary1", "binary2", "cont", "outcome")
  data <- as.data.frame(data)
  return(data)
}

################################################################################
# function for iterative matching procedure

iterative_matching_fct <- function(
  # sample size
  n_control = 1000,
  n_interim_intervention = 50,

  # data (control group = 0, intervention group = 1) ordered by group
  data=NULL,

  # column of binary variable 1
  binary_col_1 = 2,
  # column of binary variable 2
  binary_col_2 = 3,
  # column of continuous variable
  cont_col = 4,
  # column for group variable
  group_col = 1,

  # caliper for propensity score matching
```

```
  ca = 0.2,

  # maximum number of matching partners
  k = 10,

  # tolerance for matching rate
  tolerance = 0.1,

  # random number seed
  seed
){
  require(Matching)
  require(boot)

  set.seed(seed)

  matching_rate <- rep(NA, k)
  matching_pairs <- list()

  # matching data

  data_control <- data[which(data$group==0),]
  data_intervention <- data[which(data$group==1),]

  data_matching <- as.data.frame(rbind(data_control,
          data_intervention[1:n_interim_intervention,]))

  # Propensity Score matching
  # 1:1
  fit <- glm(group ~ .,
             data = data_matching,
             family = "binomial")

  glm.fitted <- log(fit$fitted / (1 - fit$fitted))

  rr <- Match(Tr = data_matching$group, X = glm.fitted,
              replace = F, M = 1, ties = F, caliper = ca)

  matching_rate[1] <- (n_interim_intervention-rr$ndrops)/n_interim_intervention
```

```
matching_pairs[[1]] <- cbind(rr$index.treated,
                             rr$index.control)

mr_actual <- matching_rate[1]

k_final <- 0

# 1:k matching
for(l in 2:k){
  if((n_control/2) >= ((k+1)*n_interim_intervention)){
    if((matching_rate[1]-tolerance) <= mr_actual){

      k_final <- k_final + 1

      fit <- glm(group ~ .,
                 data = data_matching,
                 family = "binomial")

      glm.fitted <- log(fit$fitted / (1 - fit$fitted))

      rr <- Match(Tr = data_matching$group, X = glm.fitted,
                  replace = F, M = 1, ties = F, caliper = ca)

      matching_rate[l] <- (n_interim_intervention-rr$ndrops)/
        n_interim_intervention

      matching_pairs[[l]] <- cbind(rr$index.treated,
                                   rr$index.control)

      mr_actual <- matching_rate[l]
    }
  }
}
matching_pairs_final <- matching_pairs[[k_final]]

# save results
results <- list(matching_rate = matching_rate,
                matching_pairs = matching_pairs,
```

```
                matching_pairs_final = matching_pairs_final,
                k_final = k_final)
  return(results)
}

#########################################################################
# function for simulating iterative matching procedure

Simulation <- function(
  # control data
  # binary covar 1, 2 (here: FLT3, Zyto high)
  # no/no, no/yes, yes/no, yes/yes
  binary_proportions_control = c(.48, .32, .18, .02),

  # cont. covar normally distributed
  cont_mean_control = 55,
  cont_sd_control = 15,

  # intervention data
  # binary covar 1, 2 (here: FLT3, Zyto high)
  binary_proportions_intervention = c(.48, .32, .18, .02),

  # cont. covar normally distributed
  cont_mean_intervention = 55,
  cont_sd_intervention = 15,

  # outcome model
  outcome_intercept = 2,
  outcome_binary1 = -.2,
  outcome_binary2 = -.5,
  outcome_cont = -.05,
  outcome_interaction = 0,
  outcome_intervention=log(0.7/0.3),

  # sample size in control group
  n_control = 1000,

  # maximal number of intervention patients
  n_intervention_max = 100,
```

```
  # sample size at interim analysis (intervention patients)
  n_interim_intervention = 25,

  # planned sample size for final analysis
  n_intervention_plan = 50,

  # column for group variable
  group_col = 1,

  # column of binary variable 1
  binary_col_1 = 2,

  # column of binary variable 2
  binary_col_2 = 3,

  # column of continuous variable
  cont_col = 4,

  # column for outcome variable
  out_col = 5,

  # caliper for propensity score matching
  ca = 0.2,

  # maximum number of matching partners
  k = 10,

  # tolerance for matching rate
  tolerance = 0.1,

  # Number of simulation runs
  s=1,

  # random number seed
  seed=1234
){
  source("data_fct.R")
  source("iterative_matching_fct.R")
```

```
matching_result_interim <- list()

for (i in 1:s){
  # data generation
  data <- data_fct(
    # binary baseline variables (joint distribution)
    binary_proportions_control = binary_proportions_control,
    binary_proportions_intervention = binary_proportions_intervention,

    # continuous confounder normally distributed
    cont_mean_control = cont_mean_control,
    cont_sd_control = cont_sd_control,
    cont_mean_intervention = cont_mean_intervention,
    cont_sd_intervention = cont_sd_intervention,

    # coefficients for outcome model
    outcome_intercept = outcome_intercept,
    outcome_binary1 = outcome_binary1,
    outcome_binary2 = outcome_binary2,
    outcome_cont = outcome_cont,
    outcome_interaction = outcome_interaction,
    outcome_intervention = outcome_intervention,

    # sample sizes
    n_intervention = n_intervention_max,
    n_control = n_control,

    # random number seed
    seed = seed+i)

  # Iterative matching procedure at interim analysis using the before
      generated data set
  matching_result_interim[[i]] <- iterative_matching_fct(
    # sample sizes
    n_control = n_control,
    n_interim_intervention = n_interim_intervention,

    data = data,
```

```
      # columns for baseline variables in data set
      binary_col_1 = binary_col_1,
      binary_col_2 = binary_col_2,
      cont_col = cont_col,
      group_col = group_col,

      # caliper width for propensity score matching
      ca = ca,

      # maximum number of matching partners
      k = k,

      # tolerance for matching rate
      tolerance=tolerance,

      # seed
      seed = seed+i)
  }
  return(matching_result_interim)
}
```

C.3 Synthesis of Evidence

This section contains selected function of the method comparison. first the shell function of the method comparison is given (indComp), followed by functions conducting the indirect comparison by Bucher (Bucher_fct) and MAIC (MAIC_fct).

```
##############################################################################
# Function to conduct the method comparison for indirect comparisons
# considering the method of Bucher and
# the matching adjusted indirect comparison (MAIC)
# The shell function is for a time-to-event endpoints because it includes
# the data generation proces (indComp)
# Functions for the indirect comparison are independent prom the endpoint
##############################################################################
```

```
# The following functions are needed: sampleSizeTtE - sample size calculation
#                                     patChar_fct - baseline characteristics
#                                     surv_times - survival times
#                                     directComp - direct comparison
#                                     aggData - calculates aggregated data
#                                     Bucher_fct - indirect comparison
#                                     MAIC_fct - indirect comparison
################################################################################

indComp <- function(
  # Input for sampleSizeTtE function
  # Direct Comparison AB
  # signifiCBnce level
  alpha_n_AB = 0.05,
  # power
  beta_AB = 0.2,
  # propotion of samples in control group
  v_AB = 0.5,
  # hazard ratio under alternative hypothesis (not log HR)
  hr_AB,
  # probability for event
  phi_AB,

  ## Direct Comparison CB
  # significance level
  alpha_n_CB = 0.05,
  # power
  beta_CB = 0.2,
  # propotion of samples in control group
  v_CB = 0.5,
  # hazard ratio under alternative hypothesis (not log HR)
  hr_CB,
  # probability for event
  phi_CB,

  ## Direct Comparison AB
  # Input for patChar_fct function:
  # names of patient characteristic variables (character vector)
  varnames_AB,
```

```
# type of patient characteristics
# 0 - cont
# 1 - catergorical (k=2)
vartypes_AB,
# vector of means for cont variables or reference cat for categorical
    variables
mu_AB,
# sd for cont Variables, else NA
sd_AB,
# Covariance matrix for patient variables
C_AB,
# list containing probabilities for categorical variables, if 2 categories
    only the probability for categorie "1"
prob_AB,

## Direct Comparison CB
# Input for patChar_fct function:
# names of patient characteristic variables (character vector)
varnames_CB,
# type of patient characteristics
# 0 - cont
# 1 - catergorical (k=2)
# 2 - catergorical (k>2)
vartypes_CB,
# vector of means for cont variables or reference cat for categorical
    variables
mu_CB,
# sd for cont Variables, else NA
sd_CB,
# Covariance matrix for patient variables
C_CB,
# list containing probabilities for categorical variables, if 2 categories
    only the probability for categorie "1"
prob_CB,

# Direct Comparison AB
# Input for surv_times function
lambda_mort_AB,
nue_mort_AB,
```

```
  beta_mort_AB,
  lambda_cens_AB,
  nue_cens_AB,
  max_time_AB,

  # Direct Comparison CB
  # Input for surv_times function
  lambda_mort_CB,
  nue_mort_CB,
  beta_mort_CB,
  lambda_cens_CB,
  nue_cens_CB,
  max_time_CB,

  # Input for directComp function
  # method for handling ties (as in coxph)
  method = "breslow",
  # robust=T --> using robust variance estimator
  robust = TRUE,

  # Input for MAIC_fct function
  # significance level
  alpha = .05,
  # print effective samplesize additionally to results of indirect comparison
  print_ES = TRUE
){
  # load functions
  source('SampleSizeTtE.R')
  source('PatCharFct.R')
  source('Surv_data.R')
  source('directComp.R')
  source('aggregatedData.R')
  source('Bucher.R')
  source('MAIC.R')

  # list of results
  results = list()

  # Sample size for trial AB
```

```
sampleSize_AB <- sampleSizeTtE(alpha = alpha_n_AB,
                               # power
                               beta = beta_AB,
                               # propotion of samples in control group
                               v = v_AB,
                               # hazard ratio under alternative hypothesis (
                                   not log HR)
                               hr = hr_AB,
                               # probability for event
                               phi = phi_AB)

# Sample size for trial CB
sampleSize_CB <- sampleSizeTtE(alpha = alpha_n_CB,
                               # power
                               beta = beta_CB,
                               # propotion of samples in control group
                               v = v_CB,
                               # hazard ratio under alternative hypothesis (
                                   not log HR)
                               hr = hr_CB,
                               # probability for event
                               phi = phi_CB)

# Patient characteristics for trial AB
Pat_AB = patChar_fct( n_exp=sampleSize_AB[2],
                      # number of patients in control arm
                      n_cont=sampleSize_AB[1],
                      # names of patient characteristic variables (character
                          vector)
                      varnames=varnames_AB,
                      # type of patient characteristics
                      vartypes=vartypes_AB,
                      # vector of means for cont variables or reference cat
                          for categorical variables
                      mu=mu_AB,
                      # sd for cont Variables, else NA
                      sd=sd_AB,
                      # Covariance matrix for patient variables
                      C=C_AB,
```

```
                        # list containing probabilities for categorical
                            variables, if 2 categories only the probability for
                             categorie "1"
                        prob = prob_AB)

# Patient characteristics for trial AB
Pat_CB = patChar_fct( n_exp=sampleSize_CB[2],
                        # number of patients in control arm
                        n_cont=sampleSize_CB[1],
                        # names of patient characteristic variables (character
                            vector)
                        varnames=varnames_CB,
                        # type of patient characteristics
                        vartypes=vartypes_CB,
                        # vector of means for cont variables or reference cat
                            for categorical variables
                        mu=mu_CB,
                        # sd for cont Variables, else NA
                        sd=sd_CB,
                        # Covariance matrix for patient variables
                        C=C_CB,
                        # list containing probabilities for categorical
                            variables, if 2 categories only the probability for
                             categorie "1"
                        prob = prob_CB)

# Survival times for trial AB
Data_AB = surv_times(data=Pat_AB,
                        lambda_mort=lambda_mort_AB,
                        nue_mort=nue_mort_AB,
                        beta_mort=beta_mort_AB,
                        lambda_cens=lambda_cens_AB,
                        nue_cens=nue_cens_AB,
                        max_time=max_time_AB)

# Survival times for trial CB
Data_CB = surv_times(data=Pat_CB,
                        lambda_mort=lambda_mort_CB,
                        nue_mort=nue_mort_CB,
```

```
                    beta_mort=beta_mort_CB,
                    lambda_cens=lambda_cens_CB,
                    nue_cens=nue_cens_CB,
                    max_time=max_time_CB)

# Direct comparison for trial AB
results$directAB = directComp(
  # data with covariates and survival information
  data = Data_AB,
  # method for handling ties (as in coxph)
  method = method,
  # robust=T --> using robust variance estimator
  robust = robust
)

# Direct comparison for trial CB

results$directCB = directComp(
  # data with covariates and survival information
  data = Data_CB,
  # method for handling ties (as in coxph)
  method = method,
  # robust=T --> using robust variance estimator
  robust = robust
)

# aggregated data of trial AB
results$AGGR_AB = aggData(varnames = varnames_AB,
                          vartypes = vartypes_AB,
                          data = Data_AB)

# aggregated data of trial CB
results$AGGR_CB = aggData(varnames = varnames_CB,
                          vartypes = vartypes_CB,
                          data = Data_CB)

# indirect comparison according to Bucher
results$Bucher <- Bucher_fct(results$directAB$logTE_AB,
                             results$directCB$logTE_CB,
```

```
                              results$directAB$SElogTE_AB,
                              results$directCB$SElogTE_CB)

  # indirect comparison using MAIC
  results$MAIC <- MAIC_fct(IPD_data = Data_AB,
                           # aggregated data (mean baseline values)
                           aggr_data = results$AGGR_CB,
                           # results of aggregated trial
                           results_aggr = results$directCB,
                           # significance level
                           alpha = alpha,
                           # robust variance method
                           robust = robust,
                           # print effective samplesize additionally to results
                               of indirect comparison
                           print_ES = print_ES)

  return(results)
}

##########################################################################
# Function to conduct an indirect comparison by the method of Bucher
##########################################################################

Bucher_fct <- function(
  # results of direct comparisons
  # comparison between treatments A and B (AB), C and B (CB)
  # the logarithm of the treatment effect (odds ratio (OR)/hazard ratio (HR))
  logTE_AB,
  logTE_CB,

  # and the standard error of the log treatment effect
  SElogTE_AB,
  SElogTE_CB,

  # significance level (for calculation of confidence interval)
  alpha = .05
){
  # log estimate (OR/HR) of the indirect comparison
```

```
  logTE <- logTE_AB - logTE_CB

  # variance of direct comparisons
  Var_AB <- SElogTE_AB^2
  Var_CB <- SElogTE_AB^2

  # variance of indirect comparison
  Var_logTE <- Var_AB + Var_CB

  # confidence intervall for indirect log Odds Ratio or Hazard Ratio (Estimate)
  logCI <- logTE + c(-1,1) * qnorm(1-alpha/2) * sqrt(Var_logTE)

  # save indirect treatment effect (OR/HR) and the corresponding 95% CI
  result_indComp <- exp(c(logTE, logCI))
  names(result_indComp) <- c("exp(coef)", "lowerCI", "upperCI")

  return(result_indComp)
}

##############################################################################
# Function to conduct an indirect comparison by the
# matching adjusted indirect comparison (MAIC)
##############################################################################

MAIC_fct <- function(
  # individual patient data (IPD)
  IPD_data,

  # aggregated data (mean baseline values, including proportions for
      categorical variables)
  aggr_data,

  # results of aggregated trial (treatment effect and the corresponding 95%
      confidence interval)
  results_aggr,

  # significance level
  alpha = .05,
```

```
  # robust variance method (using a sandwich estimator to calculate the
      variance of the estimate obtained by MAIC)
  robust = TRUE,

  # print effective sample size additionally to results of indirect comparison
  print_ES = TRUE,

  # vector with names of variables, which are not used for matching
  prog_var
){
  # identify variable names used for matching
  names_match <- names(aggr_data)
  names_match <- names_match[which(!(names_match %in% prog_var))]

  # exclude variables which are not needed for matching
  IPD_data_matching <- IPD_data[, names_match]

  # center the IPD by the mean baseline characteristics of the aggregated data
  IPD_centered <- t(t(IPD_data_matching)-aggr_data[names_match])

  k <- ncol(IPD_centered)

  # function to optimize
  fct <- function(beta, X){
    Xmatrix <- as.matrix(X)
    sum(exp(Xmatrix %*% beta))
  }

  # optimise betas in function "fct"
  # Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, non-linear optimization
  # initial vector of weights: all 0
  beta_hat <- optim(rep(0,k), fct, method="BFGS", X=IPD_centered)$par

  # calculate weights
  weights <- exp(as.matrix(IPD_centered) %*% beta_hat)

  # fit weighted regression model (cox/logistic)
  results_IPD <- directComp_adj(IPD_data,
                                weights=weights,
```

```
                                robust=robust)

  # final comparison using Bucher Method
  result_indComp <- Bucher_fct(logTE_AB = results_IPD$logTE,
                               logTE_CB = results_aggr$logTE,
                               SElogTE_AB = results_IPD$selogTE,
                               SElogTE_CB = results_aggr$selogTE)

  if(print_ES){
    # calculate effective sample size (ESS)
    n <- nrow(IPD_data)
    n_effective <- (sum(weights))^2 / (sum(weights^2))
    samplesize <- c(n, n_effective)
    names(samplesize) <- c("n IPD", "n effective")

    # return results of indirect comparison and ESS
    return(list(result_indComp = result_indComp,
                sampleSize = samplesize))

  }else{
    # return results of indirect comparison
    return(result_indComp)
  }
}
```

Curriculum Vitae

Dorothea Weber

born 12 July 1991 in Malsch, Germany

Education

University of Heidelberg Since 12/2016

Doctoral student (Dr. sc. hum.)

Technical University Munich 10/2014 - 11/2016

Master of Science Mathematics in Bioscience (M.Sc.)

Karlsruhe Institute of Technology 10/2010 - 12/2013

Bachelor of Science Mathematics (B.Sc.)

Bertha-von-Suttner Schule Ettlingen 09/2007 - 07/2010

A-Level (Abitur)

Wilhelm-Lorenz Realschule Ettlingen 09/2001 - 07/2007

Secondary Education (Mittlere Reife)

Pestalozzischule Ettlingen 09/1997 - 07/2001

Professional experience

University of Heidelberg Since 12/2016

Research fellow at the Institute of Medical Biometry and Informatics

HelmholtzZentrum münchen 10/2015 - 03/2016

Research assistant at Institute of Genetic Epidemiology

University Hospital Basel, Switzerland 08/2015 - 10/2015

Intern at the Swiss Transplant Cohort Study

German Cancer Research Centre (DKFZ), Heidelberg 03/2013 - 06/2013

Intern at the Division of Biostatistics

Karlsruhe Institute of Technology 03/2012 - 09/2014

Teaching assistant at the Faculty for Mathematics

Acknowledgments

First, I would like to thank my supervisor Prof. Dr. sc. hum. Meinhard Kieser for providing me the opportunity to write this thesis at the Institute of Medical Biometry and Informatics, proposing the subject of this thesis, and his support and guidance throughout my dissertation. I have greatly benefited from his expertise and am deeply grateful for his supervision.

I also want to thank my co-supervisor Dr. rer. nat. Katrin Jensen for her support and the useful discussions and suggestions. Moreover, I grateful acknowledge Dr. med. Silvia Schönenberger for her collaboration and for providing the medical data set.

Furthermore, I would like to thank my dear colleagues at the Institute of Medical Biometry and Informatics for fruitful discussions and for creating such a nice, cooperative, and thriving work environment.

Finally, I would like to thank my family and friends for their unconditional support, encouragement, and the good moments we spent together, all of which helped me to keep my motivation and cheerfulness high during the last years.

Eidesstattliche Erklärung

1. Bei der eingereichten Dissertation zu dem Thema *Applying Matching Procedures in the Generation and Synthesis of Evidence* handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In-oder Auslands als Bestandteil einer Prüfungs-oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift