Medical Decision Making
2020, Vol. 40(5) 644–654

DOI: 10.1177/0272989X20929309
journals.sagepub.com/home/mdm

Comparison of Methods for Estimating Therapy Effects by Indirect Comparisons: A Simulation Study

Dorothea Weber, Katrin Jensen, and Meinhard Kieser

Objective. In evidence synthesis, therapeutic options have to be compared despite the lack of head-to-head trials. Indirect comparisons are then widely used, although little is known about their performance in situations where cross-trial differences or effect modification are present. **Methods**. We contrast the matching adjusted indirect comparison (MAIC), simulated treatment comparison (STC), and the method of Bucher using a simulation study. The different methods are evaluated according to their power and type I error rate as well as with respect to the coverage, bias, and the root mean squared error (RMSE) of the effect estimate for practically relevant scenarios using binary and time-to-event endpoints. In addition, we investigate how the power planned for the head-to-head trials influences the actual power of the indirect comparison. **Results**. Indirect comparisons are considerably underpowered. None of the methods had substantially superior performance. In situations without cross-trial differences and effect modification, MAIC and Bucher led to similar results, while Bucher has the advantage of preserving the within-study randomization. MAIC and STC could enhance power in some scenarios but at the cost of a potential type I error inflation. Adjusting MAIC and STC for confounders that did not modify the effect led to higher bias and RMSE. **Conclusion**. The choice of effect modifiers in MAIC and STC influences the precision of the indirect comparison. Therefore, a careful selection of effect modifiers is warranted. In addition, missed differences between trials may lead to low power and partly high bias for all considered methods, and thus, results of indirect comparisons should be interpreted with caution.

Keywords
anchored indirect comparison, evidence synthesis, Bucher, MAIC, population adjustment

Date received: July 25, 2019; accepted: May 3, 2020

In medical practice, physicians frequently face situations where various therapy options exist. Ideally, all these therapies were previously compared in several clinical trials. However, often only 2-arm trials were conducted comparing just a subset of all possible therapies. In situations where a head-to-head comparison is missing, the question arises whether and how reliable and valid conclusions on the choice of the best treatment option can be drawn. In the last years, so-called indirect comparisons have attracted considerable attention.[1,2] Especially in the field of health technology assessments (HTAs), indirect comparisons are of increasing interest, because valid comparator treatments are defined for early benefit

Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Baden-Württemberg, Germany (DW, KJ, MK). The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author(s) received no financial support for the research, authorship, and/or publication of this article.

Corresponding Author
Dorothea Weber, Institute of Medical Biometry and Informatics, University of Heidelberg, Marsilius Arkaden, Im Neuenheimer Feld 130.3, Heidelberg, Baden-Württemberg, 69120, Germany (weber@imbi.uni-heidelberg.de).

assessment, and pharmaceutical industry frequently lacks for direct comparisons with this valid comparator.[3] Simply merging the results from different trials to get an estimate for the missing comparison of interest can lead to severe bias due to cross-trial differences. For example, worse baseline disease status in one of the trials may suggest that a treatment is more or less effective.[4] Therefore, powerful methods for indirect comparisons are needed.

In case individual patient data (IPD) are available, using these data may increase the reliability of the results and may reduce the uncertainty in treatment effects compared to situations where only aggregated data are available. We consider the situation where 2 treatments, A and C, are compared to a common comparator B in head-to-head trials. IPD are available for the trial A v. B (AB), whereas for the trial C v. B (CB), only aggregated data (AgD) are accessible from published results (see Figure 1). We are interested to find a treatment effect between A and C (AC) under the assumption that the population of interest is given by the population considered in the CB trial. The method of Bucher,[5] the matching-adjusted indirect comparison (MAIC),[4] and the simulated treatment comparison (STC)[6] address this setting of an anchored indirect comparison. In early benefit assessment, the acceptance of population-adjusted indirect comparisons with a common comparator differs between health regulatory authorities. For example, the Institute for Quality and Efficiency in Healthcare (IQWiG) in Germany approves indirect comparisons by the method of Bucher and the MAIC, but they are frequently rejected due to the lack of suitable data.[7,8] The National Institute for Health and Care Excellence (NICE) accepts indirect comparisons conducted by the method of Bucher as well as the MAIC and STC in case there is evidence that the population adjustment produces less biased effect estimates.[9]

Little is known about how indirect comparisons perform in situations where interactions are present, assumptions of the methods for indirect comparisons are violated such as differences in the patient population, or when cross-trial differences exist, like different confounder adjustment of regression models for evaluating the treatment effect. To examine those situations, simulation studies covering a variety of practically relevant scenarios are needed.[10–13]

Our simulation study has 2 aims. First, we investigate the method of Bucher, the MAIC, and the STC in a wide range of practically relevant scenarios where assumptions are violated and cross-trial differences exist. Those scenarios are of particular interest for benefit assessment because they are likely to be rejected by health regulatory authorities. The method of Bucher is applicable even if only aggregated data are available for both trials AB and CB. A problem arising for the method of Bucher may be the insufficient comparability of studies according to important effect modifiers. MAIC and STC address this problem of differing patient populations by a matching procedure. However, individual patient data need to be available for one trial and aggregated data for the other trial to conduct an indirect comparison by MAIC or STC.

Published results of simulation studies in the context of indirect comparisons show unsatisfactory results, especially according to power, which means to detect a treatment difference by the indirect comparison.[3,14] Snapinn and Jiang[15] showed that the sample size needed for an indirect comparison is always higher than for the underlying direct comparison. Therefore, as the second aim of this simulation study, we investigate the influence of the power provided by the sample size calculation for the head-to-head trials on the power of the indirect comparison.

Methods

The most commonly applied methods to conduct indirect comparisons are the method of Bucher, the MAIC, and the STC. The situation that can be addressed includes 1 study for each used comparison, which means in Figure 1, 1 study for the comparison A v. B and 1 study for C v. B. Even if only the treatment effects and their corresponding variances for the 2 studies are available, the method of Bucher can be applied. These treatment effects can either be adjusted (if available) or unadjusted. IPD are not required to conduct an indirect comparison. The treatment effects are calculated for each trial separately taking into consideration the randomization within the trial. Hence, the method of Bucher preserves the within-study randomization. A common comparator is needed for calculating an indirect comparison (in Figure 1 treatment B is the common comparator of the trials). The assumptions for applying the method of Bucher are shared effect modifiers and comparable study populations for important effect modifiers. Then, the effect estimate δ (log odds ratio for binary data, log hazard ratio for time-to-event data) for the indirect comparison AC is given by

$$\delta_{AC} = \delta_{AB} - \delta_{CB}, \quad (1)$$

with δ_{AB} the effect estimate of the trial AB and δ_{CB} accordingly. The variance of δ_{AC} is given by

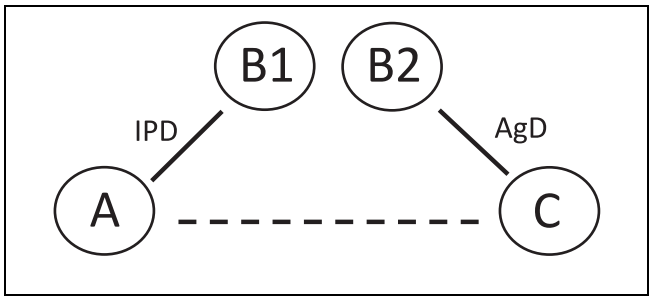


Figure 1 Indirect comparison. The plot shows the situation of the indirect comparison A v. C considered in this simulation study. To illustrate that cross-trial differences may exist, treatment B is described as B1 for the individual patient data (IPD) trial and B2 for the aggregated data (AgD) trial.

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

Without loss of generality, we assume IPD are available for the comparison AB, and only AgD are given for the trial comparing CB.

The MAIC approach needs IPD for at least 1 trial, because the aim is to match the IPD to the AgD of the other trial. The matching procedure selects a weight for each patient to reach similarity in the summary measures of the baseline characteristics of the IPD and AgD trial and follows the idea of propensity score matching. The odds between being a patient in trial AB v. trial CB provides the weights for balancing the populations. Since IPD of baseline characteristics are available for only one of the trials, the maximum likelihood method cannot be applied. Instead, the method of moments addresses this setting. The IPD is centered according to the aggregated data $\bar{x}_{CB}$ including the means and proportions, respectively, of all matching baseline characteristics.

The weights are optimized using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.[16] The matching procedure can be based on means and proportions as well as on higher-order moments. For the sake of simplicity, only means and proportions were used within this simulation study. The assigned weight estimate $\hat{\omega}_i$ is used in the direct comparison AB to obtain the weighted effect estimate $\delta_{AB}^{\text{weighted}}$. At last, the method of Bucher uses the weighted effect estimate $\delta_{AB}^{\text{weighted}}$ of the MAIC and the effect estimate of the trial CB δ_{CB} for calculation of the effect estimate δ_{AC} of the indirect comparison following equation (1). For more detailed information about these methods, see Bucher et al.[5] and Signorovitch et al.[4]

For the MAIC, the effective sample size (ESS)[10] is calculated to measure the differences in baseline characteristics between the trials. The set of estimated weights ω contains information about these differences. The ESS is calculated as follows:

$$n_{\text{effective}} = \frac{\left(\sum_{t=A,B}\sum_{i=1}^{N_{t(AB)}}\hat{\omega}_{it}\right)^2}{\sum_{t=A,B}\sum_{i=1}^{N_{t(AB)}}\hat{\omega}_{it}^2},$$

where $t = A, B,$ denotes the treatment arms of the trial AB and $N_{t(AB)}$ the sample size of treatment arm t.

The STC introduced by Ishak et al.[6] is based on a regression model for the IPD, which is substituted in mean covariate values. This substitution is done by including covariate adjustment that centers all effect modifiers by the mean values of the aggregated data. The centered treatment effect $\delta_{AB}^{\text{centered}}$ denotes the regression coefficient of the treatment covariate. Then an indirect comparison is conducted using the effect estimates $\delta_{AB}^{\text{centered}}$ and δ_{CB} to calculate the indirect treatment effect δ_{AC}. To supply unbiased estimates, the regression model needs to be specified correctly. Ishak et al.[6] propose to simulate the missing arm, but we do not follow this suggestion as it introduces additional variation. Instead, we proceed by substituting in mean covariate values directly.[9] More details are provided by Ishak et al.[6] and Phillippo et al.[9]

Simulation Study

We perform a simulation study for a wide range of practically relevant scenarios to investigate the method of Bucher, MAIC, and STC for indirect comparisons for time-to-event and binary endpoints. We assess and compare the statistical properties of the methods, including bias in the estimated therapy effects, root mean squared error (RMSE), coverage, type I error rates, and power. We performed 10,000 simulation runs for each scenario. For the method of Bucher, we assume that there are no differences between trials AB and CB with respect to effect modifiers. For the underlying indirect comparisons, nevertheless, this assumption needs to be evaluated for each situation in practice. All simulations were done with the statistical software R version 3.3.3.[17]

Data Simulation Procedure

The simulation setting included 2 studies, one comparing treatment A v. B and another study comparing treatment C v. B. A study comparing A v. C was assumed to be unavailable (Figure 1). The true treatment effects of treatments AB, BC, and AC are expressed on the log hazard ratio (HR) scale for a time-to-event setting or the log odds ratio (OR) scale for a binary endpoint, respectively.

Table 1 Patient Characteristics and Covariance Matrices Used for the Data-Generating Process for the 2 Trials

	Population		
	Similar	**Different**	
Variable	**AB/CB**	**AB_1**	**CB_2**
Continuous, mean (SD)	55 (15)	55 (15)	65 (10)
Binary1 (= 1), %	0.7	0.7	0.5
Binary2 (= 1), %	0.8	0.8	0.6
Binary3 (= 1), %	0.4	0.4	0.45
Covariance	$\begin{pmatrix} 225 & 0.25 & 0.05 & 0.01 \\ 0.25 & 0.2 & 0.01 & 0 \\ 0.05 & 0.01 & 0.15 & 0.05 \\ 0.01 & 0 & 0.05 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 225 & 0.25 & 0.05 & 0.01 \\ 0.25 & 0.2 & 0.01 & 0 \\ 0.05 & 0.01 & 0.15 & 0.05 \\ 0.01 & 0 & 0.05 & 0.1 \end{pmatrix}$	$\begin{pmatrix} 100 & 0 & 0.05 & 0.01 \\ 0 & 0.25 & -0.01 & 0 \\ 0.05 & -0.01 & 0.1 & 0.05 \\ 0.01 & 0 & 0.05 & 0.15 \end{pmatrix}$

The data-generating process includes 1 continuous and 3 binary variables, and covariances between the variables were considered to include correlations between the variables (Table 1). In the following, *similar populations* means that data for the trials AB and CB follow the same distribution, and the term *different* corresponds to divergence in the distribution parameters. For generating the time-to-event endpoint, the event time and the censoring time are sampled from a Weibull distribution ($\lambda_{event} = 0.0002$, $v_{event} = 1.8$, $\lambda_{censoring} = 0.00012$, $v_{censoring} = 2$, max. time = 100). The endpoint is generated by a Cox proportional hazard model from those times and variables. In addition, the binary endpoint is generated by a logistic regression model incorporating the variables as covariates ("confounders"). The outcome generation model with the link function $g(\cdot)$ looks as follows:

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

Table 2 contains the values for log HR and log OR of the confounders in the models with respect to the assumed treatment effect (see Table 3). Furthermore, some simulation scenarios cover an interaction term between a binary variable and the treatment. This variable is then called an effect modifier. The following equation shows the inclusion of the interaction between binary variable 2 (b 2) and treatment in the outcome generation model:

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

In addition, Table 2 contains the corresponding log HR and log OR. If the interaction term is included in only one of the trials, the shared effect modifier assumption is violated.

Table 2 Regression Coefficients in Terms of Log Hazard Ratios (HRs) for Cox Proportional Hazards Models and Log Odds Ratios (ORs) for Logistic Regression Models Considered for Simulation of Outcomes

Variable	**Time-to-Event Log HR**	**Binary Log OR**
Continuous (c)	−0.0051	0.06
Binary1 ($b1$ = 1)	−0.2	−1.76
Binary2 ($b2$ = 1)	0.18	1.26
Binary3 ($b3$ = 1)	−0.14	−0.2
Interaction		
Treatment and binary1 (= 1)	0.02	0.04

We limit the simulation study to these described clinically inspired data because we do not aim to examine the influence of the number or distributions of patient characteristics itself but rather the violation of assumptions and occurrence of cross-trial differences.

The true effect size of the trial AC is simulated as high, medium, low, or no effect, with the exact values given in Table 3. This classification of treatment effects is traced back to the benefit assessment of new drugs, which aims to test whether a new drug results in an added benefit compared to the current standard of practice. Effect sizes in terms of log HRs for time-to-event endpoints are classified according to Skipka et al.[18] For the ease of comparability, the log ORs for binary endpoints are set to similar values. Sample size calculations are based on established formulas[19,20] (the effects given in Table 3), a 5% type I error rate, and 80% power. The target population for the indirect comparison is given by the CB population.

Table 3 Values for Log Odds Ratios for Binary Endpoints Including the Binary Event Rates (p_1, p_2) and Log Hazard Ratios for Time-to-Event Endpoints for Different Effect Classes

Time to Event	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
Binary	AC	AB (p_1, p_2)	CB (p_1, p_2)
High	−0.48	−0.70 (0.45, 0.62)	−0.22 (0.45, 0.51)
Moderate	−0.23	−0.45 (0.45, 0.56)	−0.22 (0.45, 0.51)
Low	−0.06	−0.28 (0.45, 0.52)	−0.22 (0.45, 0.51)
No	0	−0.22 (0.45, 0.51)	−0.22 (0.45, 0.51)

Table 4 Considered Simulation Scenarios

	Population		Confounders		Interactions	
	Similar	Different	Similar	Different	AB	CB
I	x		x			
		x	x			
II	x		x		x	x
		x	x		x	x
III	x			x	x	x
		x		x	x	x
IV	x		x			x
		x	x			x
V	x			x		x
		x		x		x

Evaluation Measures

The performance of the 3 approaches is evaluated by the bias of the effect estimate (i.e., the difference to the true treatment effect), the RMSE, the power, the actual type I error rate, and the 2-sided 95% confidence interval (CI) coverage, where the CI for the effect estimate in the regression model relies on a normal approximation. The power is assessed by the proportion of simulation runs where 0 (no effect) is not included in the 2-sided 95% CI of the effect estimate when in fact an effect exists. The power is calculated for the categories high, medium, and low effect. If there is no effect, we are 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 2-sided 95% CI of the effect estimate for the indirect comparison. The aim is to minimize bias, RMSE, and inflation of the type I error rate, whereas power ought to reach high values and the coverage should be around 95%. All evaluation measures are calculated for the indirect comparison AC and correspond to the primary treatment effect.

The Simulation Scenarios

We analyze indirect comparisons for a binary and a time-to-event endpoint. In Table 4, the simulation scenarios are depicted. They are characterized by the following 4 aspects:

1. similar or different distributions of patient characteristics (proportions of categorical variables, mean, and variance of continuous variables substantially differ, and the cutoff between similar and different depends on the variable and the objective of the comparison),
2. inclusion of an effect modification (interaction term between a binary variable and treatment),
3. similar or different confounders, and
4. differences in the presence of the interaction.

In addition, for the method of Bucher and MAIC, we investigate the influence of the power planned for the sample size calculation of the head-to-head trials (AB and CB) on the power of the indirect comparison (AC). The sample size calculations are based on established formulas[18,19] (the effects given in Table 3), a type I error rate of 5%, and 80%, 90%, 95%, or 99% power. The characteristics such as actual type I error rate, 95% CI coverage, and bias of the effect estimate in the indirect comparison are evaluated as well.

Results

This section is split according to the evaluation measures. For scenarios I to V, we focus on the situation including individual studies that are planned for 80% power and a 5% type I error rate. The effect estimates are calculated by a logistic or a Cox regression model. They can either be adjusted for effect modifiers and confounders or include the treatment group as a single covariate. Independent of the method for the indirect comparison, we consider 3 settings for the calculation of direct effect estimates: regression models in trials AB and CB include all relevant effect modifiers and confounders, regression models in trials AB and CB only include confounders (effect modifiers are treated as confounders; note that in this case, STC is not considered, since its concept is based on including effect modifiers), and the

regression model for trial CB is not adjusted for effect modifiers or confounders. The aim is to investigate the influence of this adjustment of the direct comparison to the indirect comparison. In particular, for the method of Bucher where usually AgD are used only, unadjusted effect estimates might be available. The fifth paragraph describes the influence of the planned power of the individual trials on that of the indirect comparison. Within each subsection, we describe the results for both endpoints. The detailed results of the different scenarios and endpoints based on the described evaluation measures can be found in Supplemental Tables S7 to S20. The ESS is influenced by the differences in distributions of variables considered in the MAIC procedure. Therefore, it is independent of interactions or adjustment of regression models, and hence, the results are similar for all considered scenarios (see Supplemental Tables S21–S25).

Power

In scenario I, no effect modification is present, but all confounds are assumed to modify the effect, and MAIC and Bucher show identical results for equal patient groups. For those 2 methods, the power decreases for lower treatment effects, whereas STC reaches higher values for smaller treatment effects. When characteristics differ between trials, STC loses power. Scenario II includes an interaction that makes MAIC achieve higher power values when there are differences in the confounder and effect modifier distributions for time-to-event (TTE) endpoints. When adjusting MAIC only for the effect modifiers, power is decreasing to a small amount. STC produces comparable power values for equal and different patient characteristic distributions. These power values do not vary by a high amount between scenarios. Only small differences are observed for STC when centering is considered for effect modifiers only. For binary endpoints and present effect modification, MAIC results in higher power values compared to Bucher and STC, even when characteristics are equal. Furthermore, a small increase is observed when MAIC adjusts for effect modifiers only. In case confounder overlap differs (scenario III), we observe similar results as for scenario II for all 3 methods. In scenario IV, the power values are relatively high and are comparable for both methods when populations are similar. Adjusting MAIC for all confounders leads to power loss compared to only adjusting for effect modifiers when population distributions differ. In case the effect modification is not considered within the regression models for MAIC and Bucher, we observe higher power values in scenarios where the effect modification is present in both trials. When only unadjusted effect estimates are available for CB (regression models are not adjusted for confounders and effect modifiers) for TTE endpoints, power values decrease for scenarios where effect modification is only present in CB.

Type I Error Rate

In scenario I, Bucher and MAIC show type I error rates around 5% for both endpoints. For STC, however, type I error rate is around 5% only for binary endpoints, and highly inflated type I error rates are observed for TTE endpoints. Scenario II leads to type I error rates around 5% for all methods. If confounder overlap differs (scenario III), the type I error rate remains as in scenario II. However, if effect modification is only present in trial CB, we observe a type I error rate inflation for Bucher and MAIC as well. When populations are similar, the 2 methods perform equally, but in case of differences, MAIC leads to lower type I error rates, which are still highly inflated. In scenarios III and IV, STC leads to the smallest type I error rates. For binary endpoints, MAIC yields 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, the type I error rate is controlled for scenarios with effect modification in both trials.

Coverage

Given scenario I and binary endpoints, all methods show comparable results. When an effect modifier is present (scenario II), MAIC reaches coverage over 90%, whereas Bucher leads to values lower than 90%. When performing STC, the coverage is below 50% for high and medium treatment effects and increases for lower treatment effects for TTE endpoints. When additionally confounders differ between trials in the binary case, MAIC reaches higher coverage, whereas for TTE endpoints, Bucher and MAIC show similar results. In scenarios IV and V, we observe higher values for MAIC when population distributions differ, but all values are below 85%.

Bias and RMSE

The bias and the RMSE are slightly higher for MAIC when all confounders are considered in the matching step of MAIC in scenario I for both endpoints. STC shows

even higher bias and RMSE. When effect modifiers are present and regression models are adjusted, Bucher and MAIC give similar results for TTE endpoints, and STC shows lower bias and RMSE. In the binary setting, MAIC results in lower bias and RMSE compared to Bucher and STC. In case confounder overlap differs (scenario III), we observe only small differences to scenario II. Scenarios IV and V show the lowest bias and RMSE values for TTE endpoints, differences between Bucher and MAIC are negligible for both endpoints, and for STC, higher values are observed. When effect modification is not considered in the regression models for effect estimates of AB and CB, we even observe a slightly smaller bias and RMSE. When CB is not adjusted for any confounder, bias and RMSE increase for Bucher and MAIC, whereas STC leads only to marginally higher bias and to a lower RMSE compared to the other methods.

Effective Sample Size

The effective sample size for the MAIC procedure equals the actual sample size for similar population distributions. When characteristics differ between trials and all confounders are considered within MAIC, ESS results in less than half of the actual sample size. Adjusting MAIC solely for effect modifiers reduces the ESS by 15% to 20% only.

Influence of Power of Direct Comparisons

We varied the power used for the sample size calculation for the trials AB and CB to investigate its influence on the power of the indirect comparison. We analyze trials powered at 80%, 90%, 95%, and 99%, including all possible power combinations. The power of the indirect comparison increases with increasing power of the head-to-head comparisons. However, even for trials powered at 99%, the indirect comparison reaches a power of less than 60% when all method assumptions are met (Supplemental Table S26). Higher treatment effects in the indirect comparison gain more power by increasing the power in head-to-head trials. The results for a fixed power of 80% in the AgD trial CB and increasing power of the IPD trial AB are plotted in Figure 2. The type I error rate remains at around 5% for all power scenarios. The results for the bias of the effect estimate, RMSE, and the coverage of the 95% CI are unchanged for all power scenarios. These measures are already discussed above.

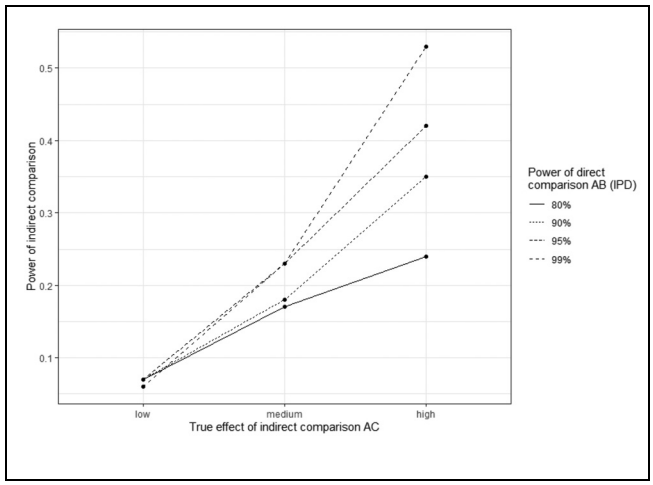


Figure 2 Power of indirect comparison. The plot shows the power of the indirect comparison (A v. C) depending on its true effect. The power for the sample size calculation in direct trial C v. B (aggregated data are available) is set to 80%, and the power for A v. B (individual patient data is available) is varied. Equal population distributions are considered, which makes the results hold true for the method of Bucher and matching adjusted indirect comparison (MAIC) because they perform similarly in this scenario. IPD, individual patient data.

Discussion

This simulation study observes that indirect comparisons are highly underpowered in scenarios commonly met in practice. Scenarios, where the assumptions of the methods (see Methods) hold true, perform better, but the performance is far from being good in terms of power. This observation is in line with other publications that include simulation studies on indirect comparisons.[3,14] Our results also provide a rationale for why, as often in practice, the application of indirect comparisons in early benefit assessment leads to the conclusion that there is no additional benefit.[21] Even if there would be an actual benefit, it will rarely be demonstrated by the indirect comparison due to low power and wide confidence intervals.

The results show that there are situations where the method of Bucher performs better than the matching approaches and vice versa. Noticeable differences between the matching methods are observed as well. Deviations from the method assumptions result in biased effect estimates. While MAIC results in less biased estimates in some scenarios, STC shows higher bias and RMSE in most of the scenarios. The superiority of the MAIC over the method of Bucher is linked to the

presence of effect modification. However, there are also situations when the MAIC leads to higher bias and less power compared to the method of Bucher. This may be explained by the 1-arm weighting when models are already adjusted for all influencing confounders. Then, the weighting may result in more biased effects for the indirect comparison, which seems to be connected to the MAIC adjusting for confounders that are not effect modifiers. This situation, where models are adjusted for all relevant confounders and effect modifiers, may not always be given in practice, and moreover, this assumption cannot be checked and increases bias and RMSE. For STC, similar patterns are observed, and the performance decreases when confounders are treated as effect modifiers. When the set of confounders differs between trials, the results remain comparable. However, when the overlap of effect modifiers differs, type I error rate is highly inflated, and in case of binary endpoints, bias and RMSE are also higher as in other scenarios. Only STC leads to results comparable to the situation where effect modification is present in both trials. We observe situations where ignoring the effect modification leads to better results. This may be since we evaluate the marginal effect of the treatment and do not evaluate the interaction term itself because it cannot be assessed by the marginal effect of the interaction.[22] A limitation of the MAIC and STC approach is the shift of the IPD towards the AgD trial, which means that the AgD trial defines the target population. Higher planned power in the sample size calculation of the direct trials increases the power of the indirect comparison. Hence, investigators can influence the power of a later indirect comparison by planning the head-to-head trials on a higher power level.

One strength of our simulation study is the variety of clinically relevant scenarios considered, including confounders, correlations, interactions (effect modification), and adjustment of regression models, which are evaluated and compared within this work. The sample size of the 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, which makes the scenarios practically relevant. Nevertheless, the following limitations apply to the simulation study. We only considered 1 clinically inspired data set, and it is assumed that the interaction terms have the same sign and that the treatment effect modifiers are known. In addition, the overlap of populations is good enough to expect matching to work well. Note that this is mainly relevant for variables considered in the matching procedure of MAIC.

The MAIC in its original form can only be used considering 1 study per treatment comparison. However, there are commonly at least 2 or even more studies per treatment comparison available. Further research is needed to develop, evaluate, and compare the method of Bucher, MAIC, and STC using several studies per treatment comparison and to expand these methods to network settings. By using several studies, the decision is based on a higher number of patients, which may lead to more precise estimates and therefore to a higher power for the indirect comparison.

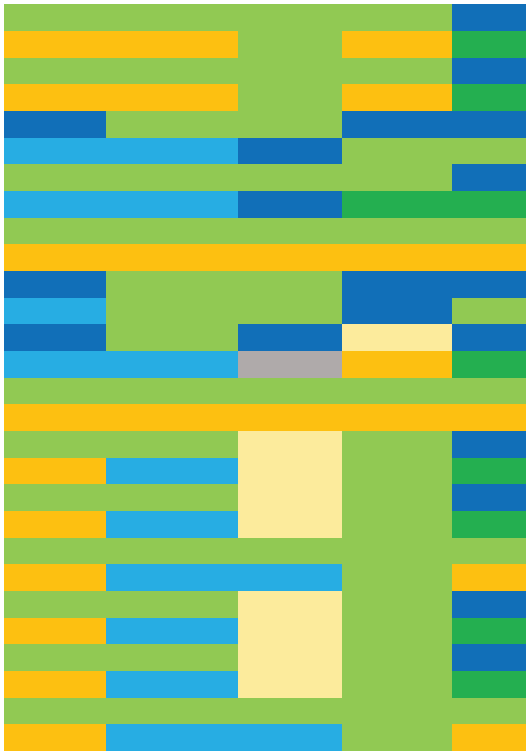
A further limitation of this simulation study is that we just compared the method of Bucher, MAIC, and STC since they are widely used and accepted in the field of HTA. Nevertheless, there are other methods available for indirect comparisons (e.g., Droitcour et al.,[23] Nie et al.,[24] Jansen[25]) whose properties are currently not sufficiently examined.[3]

Conclusion

Indirect comparisons allow for estimation of treatment effects when studies comparing these therapies directly are not available. An important step prior to conducting an indirect comparison is the identification of possible underlying differences between the trials. Based on this knowledge, the method for the indirect comparison should be chosen carefully to avoid bias. 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, in case of effect modification in 1 or both trials and differences with respect to effect modifiers as well as adjustment of regression models, the MAIC provides less biased effect estimates and higher coverage. Matching variables in MAIC, as well as effect modifiers in STC, need to be chosen carefully because including confounders, which do not modify the effect, influences the precision of the indirect comparison. A summary of the performance in the considered scenarios is given in Table 5 for TTE and in Table 6 for binary endpoints. Nevertheless, results of indirect comparisons should be interpreted with caution, and one should be aware of the potentially low power if no treatment effect can be demonstrated in an indirect comparison.

Table 5 Overview of Situations and the Best-Performing Method with Regard to the Considered Simulation Scenarios for TTE Endpoints[a]

Theoretical Situation						Results				
Scenario	**End point**	**Population**	**Confounder**	**Interaction**	**Analysis Deviations from Simulated Situation**	**Power**	**Coverage**	**Type I Error**	**Bias/ RMSE**	**Sample Size**
I	TTE	Equal	Similar	No						
	TTE	Different	Similar	No						
	TTE	Equal	Similar	No	CB not adjusted for confounders					
	TTE	Different	Similar	No						
II	TTE	Equal	Similar	Yes						
	TTE	Different	Similar	Yes						
	TTE	Equal	Similar	Yes	CB not adjusted for confounders					
	TTE	Different	Similar	Yes						
STC not applicable	TTE	Equal	Similar	Yes	Effect modification not adjusted					
STC not applicable	TTE	Different	Similar	Yes						
III	TTE	Equal	Different	Yes						
	TTE	Different	Different	Yes						
	TTE	Equal	Different	Yes	CB not adjusted for confounders					
	TTE	Different	Different	Yes						
STC not applicable	TTE	Equal	Different	Yes	Effect modification not adjusted					
STC not applicable	TTE	Different	Different	Yes						
IV	TTE	Equal	Similar	Yes (CB)						
	TTE	Different	Similar	Yes (CB)						
	TTE	Equal	Similar	Yes (CB)	CB not adjusted for confounders					
	TTE	Different	Similar	Yes (CB)						
STC not applicable	TTE	Equal	Similar	Yes (CB)	Effect modification not adjusted					
STC not applicable	TTE	Different	Similar	Yes (CB)						
V	TTE	Equal	Different	Yes (CB)						
	TTE	Different	Different	Yes (CB)						
	TTE	Equal	Different	Yes (CB)	CB not adjusted for confounders					
	TTE	Different	Different	Yes (CB)						
STC not applicable	TTE	Equal	Different	Yes (CB)	Effect modification not adjusted					
STC not applicable	TTE	Different	Different	Yes (CB)						

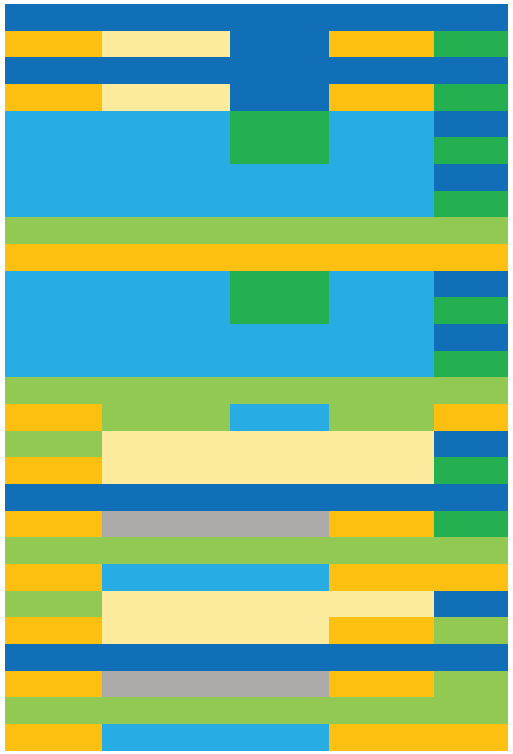


MAIC, matching adjusted indirect comparison; RMSE, root mean squared error; STC, simulated treatment comparison; TTE, time to event.

[a]MAIC, light blue; Bucher, orange; STC, pink; MAIC and Bucher perform similar, light green; STC and Bucher perform similar, dark green; MAIC and STC perform similar, brown; all methods perform similar, dark blue. This table is available in color online.

Table 6 Overview of Situations and the Best-Performing Method with Regard to the Considered Simulation Scenarios for Binary Endpoints[a]

Theoretical Situation						Results				
Scenario	**Endpoint**	**Population**	**Confounder**	**Interaction**	**Analysis Deviations from Simulated Situation**	**Power**	**Coverage**	**Type I Error**	**Bias/ RMSE**	**Sample Size**
I	Binary	Equal	Similar	No						
	Binary	Different	Similar	No						
	Binary	Equal	Similar	No	CB not adjusted for confounders					
	Binary	Different	Similar	No						
II	Binary	Equal	Similar	Yes						
	Binary	Different	Similar	Yes						
	Binary	Equal	Similar	Yes	CB not adjusted for confounders					
	Binary	Different	Similar	Yes						
STC not applicable	Binary	Equal	Similar	Yes	Effect modification not adjusted					
STC not applicable	Binary	Different	Similar	Yes						
III	Binary	Equal	Different	Yes						
	Binary	Different	Different	Yes						
	Binary	Equal	Different	Yes	CB not adjusted for confounders					
	Binary	Different	Different	Yes						
STC not applicable	Binary	Equal	Different	Yes	Effect modification not adjusted					
STC not applicable	Binary	Different	Different	Yes						
IV	Binary	Equal	Similar	Yes (CB)						
	Binary	Different	Similar	Yes (CB)						
	Binary	Equal	Similar	Yes (CB)	CB not adjusted for confounders					
	Binary	Different	Similar	Yes (CB)						
STC not applicable	Binary	Equal	Similar	Yes (CB)	Effect modification not adjusted					
STC not applicable	Binary	different	Similar	Yes (CB)						
V	Binary	Equal	Different	Yes (CB)						
	Binary	Different	Different	Yes (CB)						
	Binary	Equal	Different	Yes (CB)	CB not adjusted for confounders					
	Binary	Different	Different	Yes (CB)						
STC not applicable	Binary	Equal	Different	Yes (CB)	Effect modification not adjusted					
STC not applicable	Binary	Different	Different	Yes (CB)						



MAIC, matching adjusted indirect comparison; RMSE, root mean squared error; STC, simulated treatment comparison; TTE, time to event.

[a]MAIC, light blue; Bucher, orange; STC, pink; MAIC and Bucher perform similar, light green; STC and Bucher perform similar, dark green; MAIC and STC perform similar, brown; all methods perform similar, dark blue. This table is available in color online.

Acknowledgments

We thank the 3 reviewers who committed their time and efforts to improve this article by their very helpful comments.

ORCID iD

Dorothea Weber https://orcid.org/0000-0003-4850-9116

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://journals.sagepub.com/home/mdm.

References

1. Signorovitch J, Swallow E, Kantor E, et al. Everolimus and sunitinib for advanced pancreatic neuroendocrine tumors: a matching-adjusted indirect comparison. *Exp Hematol Oncol*. 2013;2(1):32.
2. Nash P, McInnes IB, Mease PJ, et al. Secukinumab versus adalimumab for psoriatic arthritis: comparative effectiveness up to 48 weeks using a matching-adjusted indirect comparison. *Rheumatol Ther*. 2018;5(1):99–122.
3. Kühnast S, Schiffner-Rohe J, Rahnenführer J, Leverkus F. Evaluation of adjusted and unadjusted indirect comparison methods in benefit assessment. *Methods Inf Med*. 2017;58(1):43–58.
4. Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials. *Pharmacoeconomics*. 2010;28(10):935–45.
5. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50(6):683–91.
6. Ishak KJ, Rael M, Phatak H, Masseria C, Lanitis T. Simulated treatment comparison of time-to-event (and other non-linear) outcomes. *Value Health*. 2015;18(7):719.
7. Institute for Quality and Efficiency in Healthcare (IQWiG). Allgemeine Methoden: Version 5.0. 2017. Available from: https://www.iqwig.de/en/methods/methods-paper.3020.html.
8. Institute for Quality and Efficiency in Healthcare (IQWiG). IQWiG's results on AMNOG at a glance. July 22, 2019. Available from: https://www.iqwig.de/en/press/amnog-at-a-glance.7723.html.
9. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Nice DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to nice. 2016. Available from: http://nicedsu.org.uk/technical-support-documents/population-adjusted-indirect-comparisons-maic-and-stc/.
10. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*. 2018;38(2):200–211.
11. Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ*. 2009;338:b1147.
12. Glenny A, Altman D, Song F, Sakarovitch C, Deeks J. Indirect comparisons of competing interventions. *Health Technol Assess*. 2005;9(26):1–134.
13. Petto H, Kadziola Z, Brnabic A, Saure D, Belger M. Alternative weighting approaches for anchored matching-adjusted indirect comparisons via a common comparator. *Value Health*. 2019;22(1):85–91.
14. Mills EJ, Ghement I, O'Regan C, Thorlund K. Estimating the power of indirect comparisons: a simulation study. *PLoS ONE*. 2011;6:1–8.
15. Snapinn S, Jiang Q. Indirect comparisons in the comparative efficacy and non-inferiority settings. *Pharm Stat*. 2011;10(5):420–6.
16. Broyden CG. The convergence of a class of double-rank minimization algorithms: 1. General considerations. *IMA J Appl Math*. 1970;6(1):76–90.
17. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017.
18. Skipka G, Wieseler B, Kaiser T, et al. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biometrical J*. 2015;56(3):261–7.
19. Schoenfeld D. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983;39(2):499–503.
20. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.
21. Ruof J, Schwartz FW, Schulenburg JM, Dintsios CM. Early benefit assessment (EBA) in Germany: analysing decisions 18 months after introducing the new AMNOG legislation. *Eur J Health Econ*. 2014;15(6):577–89.
22. Norton EC, Wang H, Ai C. Computing interaction effects and standard errors in logit and probit models. *Stata J*. 2004;4(2):154–67.
23. Droitcour J, Silberman G, Chelimsky E. Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *Int J Technol Assess Health Care*. 1993;9(3):440–9.
24. Nie L, Zhang Z, Rubin D, Chu J. Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference. *Ann Appl Stat*. 2013;7(3):1796–813.
25. Jansen JP. Network meta-analysis of individual and aggregate level data. *Res Synthesis Methods*. 2012;3(2):177–90.