Song et al. BMC Medical Research Methodology 2012, 12:138
http://www.biomedcentral.com/1471-2288/12/138

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons

Fujian Song*, Allan Clark, Max O Bachmann and Jim Maas

Abstract

Background: Indirect treatment comparison (ITC) and mixed treatment comparisons (MTC) have been increasingly used in network meta-analyses. This simulation study comprehensively investigated statistical properties and performances of commonly used ITC and MTC methods, including simple ITC (the Bucher method), frequentist and Bayesian MTC methods.

Methods: A simple network of three sets of two-arm trials with a closed loop was simulated. Different simulation scenarios were based on different number of trials, assumed treatment effects, extent of heterogeneity, bias and inconsistency. The performance of the ITC and MTC methods was measured by the type I error, statistical power, observed bias and mean squared error (MSE).

Results: When there are no biases in primary studies, all ITC and MTC methods investigated are on average unbiased. Depending on the extent and direction of biases in different sets of studies, ITC and MTC methods may be more or less biased than direct treatment comparisons (DTC). Of the methods investigated, the simple ITC method has the largest mean squared error (MSE). The DTC is superior to the ITC in terms of statistical power and MSE. Under the simulated circumstances in which there are no systematic biases and inconsistencies, the performances of MTC methods are generally better than the performance of the corresponding DTC methods. For inconsistency detection in network meta-analysis, the methods evaluated are on average unbiased. The statistical power of commonly used methods for detecting inconsistency is very low.

Conclusions: The available methods for indirect and mixed treatment comparisons have different advantages and limitations, depending on whether data analysed satisfies underlying assumptions. To choose the most valid statistical methods for research synthesis, an appropriate assessment of primary studies included in evidence network is required.

Keywords: Indirect comparison, Mixed treatment comparison, Network meta-analysis, Inconsistency, Bias, Type I error, Statistical power, Simulation evaluation

Background

Indirect and mixed treatment comparisons have been increasingly used in health technology assessment reviews [1-4]. Indirect treatment comparison (ITC) refers to a comparison of different treatments using data from separate studies, in contrast to a direct treatment comparison (DTC) within randomised controlled trials. Statistical methods have been developed to indirectly compare multiple treatments and to combine evidence from direct and indirect comparisons in mixed treatment comparison (MTC) or network meta-analysis [5-9].

The existing simple [5] or complex [6-8] statistical methods for ITC and MTC are theoretically valid if certain assumptions can be fulfilled [2,10]. The relevant assumptions could be specifically classified according to a conceptual framework that delineates the homogeneity assumption for conventional meta-analysis, the similarity assumption for adjusted ITC, and the consistency

* Correspondence: Fujian.Song@uea.ac.uk
Norwich Medical School, Faculty of Medicine and Health Science, University of East Anglia, Norwich, Norfolk NR4 7TJ, UK

assumption for pooling direct and indirect estimates by MTC [2,11]. Among the basic assumptions, heterogeneity in meta-analysis and inconsistency between direct and indirect estimates can be quantitatively investigated. The presence of inconsistency between direct and indirect estimates has been empirically investigated in meta-epidemiological studies and numerous cases reports [12-16]. A range of statistical methods have been suggested to investigate the inconsistency in network meta-analysis [5,7,9,17-19].

The statistical properties of simple adjusted ITC [5] have been previously evaluated in simulation studies [1,20,21]. However, there are no simulation studies that formally evaluate methods for Bayesian network meta-analysis. In this simulation study, we comprehensively evaluated properties and the performance of commonly used ITC and MTC methods. Specifically, the objectives of the study are (1) to investigate bias, Type I error and statistical power of different comparison models for estimating relative treatment effects, and (2) to investigate bias, Type I error and statistical power of different comparison models for quantifying inconsistency between direct and indirect estimates.

Methods

Comparison models investigated

We investigated the performance of the following ITC and MTC statistical models.

Adjusted indirect treatment comparison (AITC)

This frequentist based method is also called as Bucher's method [5], based on the assumption that indirect evidence is consistent with the direct comparison. Suppose that treatment A and B are compared in RCT-1 (with d_{AB} as its result, logOR for example), and treatment A and C compared in RCT-2 (with d_{AC}as its result). Then treatment A can be used as a common comparator to adjust the indirect comparison of treatment B and C:

$$d_{BC}^{Ind} = d_{AB} - d_{AC}$$

Its variance is:

$$Var\left(d_{BC}^{Ind}\right) = Var(d_{AB}) + Var(d_{AC})$$

When there are multiple trials that compared treatment A and B or treatment A and C, results from individual trials can be combined using fixed-effect or random-effects model. Then the pooled estimates of d_{AB} and d_{AC} are used in the AITC.

Consistency frequentist MTC (CFMTC)

The results of frequentist ITC (using the Bucher's method) can be combined with the result of frequentist DTC in a MTC. The frequentist combination of the DTC and ITC estimate is weighted by the corresponding inverse of variance, as for pooling results from two individual studies in meta-analysis [22].

This MTC is termed 'consistency MTC', as it assumes that the result of direct comparison of treatment B and C statistically equals to the result of indirect comparison of B and C based on the common comparator A [9]. Suppose a network of three sets of trials that compared A vs. B, A vs. C, and B vs. C, we only need to estimate two basic parameters d_{AB} and d_{AC}, and the third contrast (functional parameter) can be derived by $d_{BC} = d_{AB} - d_{AC}$.

Consistency Bayesian MTC (CBMTC)

As the CFMTC, this model is also based on the assumption that ITC is consistent with DTC [8]. Suppose that several treatments (A, B, C, and so on) are compared in a network of trials. We need to select a treatment (treatment A, for example, placebo or control) as the *reference* treatment. In each study, we also consider a treatment as the *base* treatment (b). Below is the general model for the consistency MTC:

$$\theta_{kt} = \begin{cases} \mu_{kb} & b = A, B, C, \quad if\ t = b \\ \mu_{kb} + \delta_{kbt} & t = B, C, D, \quad if\ t\ is\ after\ b \end{cases}$$

$$\delta_{kbt} \sim N\left(d_{bt}, \tau^2\right)$$

$$d_{bt} = d_{At} - d_{Ab}$$

$$d_{AA} = 0$$

Here θ_{kt} is the underlying outcome for treatment t in study k, μ_{kb} is the outcome of treatment b, and δ_{kbt} is the relative effect of treatment t as compared with treatment b in study k. The trial specific relative effect δ_{kbt} is assumed to have a normal distribution with a mean d_{bt} and variance τ^2 (i.e., between study variance). When $\tau^2 = 0$, this model provides results as a fixed-effect analysis.

Random Inconsistency Bayesian MTC (RIBMTC)

Some authors assumed that inconsistencies (that is, the differences between d_{BC} from direct comparisons and d_{BC}^{Ind} based on indirect comparison) have a common normal distribution with mean 0 and variance σ_{ω}^2 [7,9]. These methods have been termed the "random inconsistency model" [23]. In this study, we evaluated the random inconsistency model by Lu and Ades [9]. This model can be expressed by the following:

$$d_{BC} = d_{AB} - d_{AC} + \omega_{BC},$$

and

$$\omega_{BC} \sim N(0, \sigma_{\omega}^2).$$

Here ω_{BC} is termed inconsistency factor (ICF).

Inconsistency Bayesian Meta-Analysis (IBMA)

In the inconsistency Bayesian meta-analysis (IBMA), each of the mean relative effects (d_{xy}) is separately estimated without using indirect treatment comparison information. The IBMA analysis is equivalent to a series of pair-wise DTC meta-analyses, although a common between-study variance (τ^2) across different contrasts is assumed [24].

We originally intended to include the Lumley's frequentist method for network meta-analysis [7]. However, it was excluded because of convergence problems during computer simulations.

Inconsistency test

Let d_{BC} denote the natural log OR estimated by the DTC, and d_{BC}^{Ind} denote the log OR estimated by the ITC. The inconsistency (ω_{BC}) in the results between the direct and indirect comparison of treatment B and C can be calculated by the following:

$$\omega_{BC} = d_{BC} - d_{BC}^{Ind}$$

When the estimated ω_{BC} is greater than 0, it indicates that the treatment effect is over-estimated by the ITC as compared with the DTC. For Bucher's method [5,12], the calculation of inconsistency was based on the pooled estimates of d_{BC} and d_{BC}^{Ind} by meta-analyses. The variance of the estimated inconsistency was calculated by:

$$Var(\omega_{BC}) = Var(d_{BC}) + Var(d_{BC}^{Ind})$$

where Var(d_{BC}) and Var(d_{BC}^{Ind}) are the variance of d_{BC} and d_{BC}^{Ind} respectively. The null hypothesis that the DTC estimate equals to the ITC estimate was tested by Z statistic

$$Z_{BC} = \frac{\omega_{BC}}{\sqrt{Var(\omega_{BC})}}$$

If the absolute value of Z_{BC} is greater than 1.96, the observed inconsistency is considered to be statistically significantly different from zero.

The estimate of inconsistency is not applicable when the consistency Bayesian MTC model [8] is used. With the inconsistency Bayesian meta-analysis (IBMA), the estimate of d_{BC} is naturally available, and d_{BC}^{Ind} can be easily estimated based on d_{AB} and d_{AC}, as by the "node-splitting" method [17,24]. The point estimate of inconsistency in Bayesian MTC was the average (mean value) of the simulated results. The significance of the inconsistency was based on the estimated 95% intervals. If the 95% intervals did not contain the zero, the observed inconsistency was considered to be statistically significant.

The random inconsistency Bayesian MTC (RIBMTC) model assumes that the inconsistency within a network of trials is normally distributed with mean $\omega = 0$ and variance σ_{ω}^2 [9]. We also recorded the estimated ω and σ_{ω}^2 by the RIBMTC model.

Simulation scenarios

In this study, a simple network of two-arm trials with a closed loop was simulated to separately compare three treatments: treatment 1 (T_1, placebo), treatment 2 (T_2, an old drug), and treatment 3 (T_3, a new drug) (Figure 1). The comparison of T_2 and T_3 was considered as the main interest. Trials that compared T_1 vs. T_2 and trials that compared T_1 vs. T_3 were used for the indirect comparison of T_2 and T_3. Given the available resource, a limited number of simulation scenarios were adopted in this study. The following simulation parameters were decided after considering characteristics of published meta-analyses (also see Table 1).

- The number of patients in each arm of a pair-wise trial is 100. The number of trials for each of the three contrasts is 1, 5, 10, 20, 30 and 40. A scenario of imbalanced number of trials (including a single trial for one of the three sets) is also included.

- We use odds ratio (OR) to measure the outcome [25]. The assumed true $OR_{12} = 0.8$, and the true $OR_{13} = 0.8$ or 0.6. When OR is less than 1 (or *log OR < 0*), it indicates that the risk of events is reduced by the second of the two treatments compared.

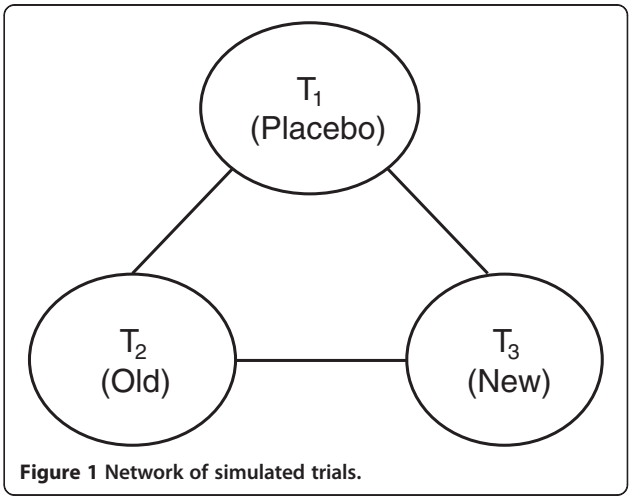

Figure 1 Network of simulated trials.

Table 1 Simulation input parameters

Parameters	Values
Number of studies	3×40; 3×20; 3×10; 3×5; 3×1; 5/1/5
Number of patients per study	2×100
Between trial heterogeneity: τ^2	0.00; 0.05; 0.10; 0.15
Treatment effect: log OR, θ_{12}	log(0.8)
Treatment effect: log OR, θ_{13}	log(0.8); log(0.6)
Bias: ROR_{12}	0.00; 0.80
Bias: ROR_{13}	0.00; 0.80
Bias: ROR_{23}	0.00; 0.80
Baseline risk: P_1	10%; 20%

(Note: these input values could be combined differently for a large number of possible simulation scenarios).

- The true $\log OR_{23}$ is calculated by: $\log OR_{23} = \log OR_{13} - \log OR_{12}$.
- The baseline risk in the control arm is assumed to be 20% or 10%.
- It is assumed that heterogeneity is constant across different comparisons, and there are four levels of between study variance: $\tau^2 = 0.00$, 0.05, 0.10, and 0.15 respectively [26].
- The trial-specific natural log OR (d_{kij}) in study k used to generate simulated trials is based on the assumed true log OR and the between-trial variance: $d_{kij} \sim N(d_{ij}, \tau^2)$.
- Given the baseline risk (P_{k1}) and the trial-specific OR, the risk in the treatment arm in study k is calculated by:
- $$P_{kt} = \frac{P_{k1} \times Exp(d_{k1t})}{1 - P_{k1} + P_{k1} \times Exp(d_{k1t})}.$$
- Bias in a clinical trial can be defined as a systematic difference between the estimated effect size and the true effect size [27]. It is assumed here that all bias, where it exists, will result in an over-estimated treatment effect of active drugs (T_2 and T_3) as compared with placebo (T_1), and an over-estimated treatment effect of the new drug (T_3) relative to the old drug (T_2). The extent of bias and inconsistency is measured by ratio of odds ratios (ROR). When ROR = 1, it indicates that there is no bias. When ROR = 0.8, it means that the effect (OR) of a treatment is over-estimated by 20%.

A network of trials was randomly generated, using assumed input parameters (Table 1). For each arm of the simulated trial, the number of events was randomly generated according to the binomial distribution:

$$r_{ki} \sim Binomial(N_{ki}, P_{ki})$$

Here, N_{ki} is the number of patients in the arm of treatment i, and P_{ki} is the risk of events given treatment i in study k. If the simulated number of events is zero, we added 0.5 to the corresponding cells of the 2x2 table for conducting inverse variance weighted meta-analysis.

Data analysis

AITC and MTC were conducted using data from the simulated trials by fixed-effect and random-effects meta-analyses. For frequentist ITC, we used inverse variance weights to pool results of multiple trials in meta-analysis, and used the DerSimonian-Laird method for random-effects meta-analyses [22].

The performance of the ITC and MTC methods was measured by the type I error rate or statistical power, observed bias and mean squared error (MSE). We estimated the rate of type I error (when the null hypothesis is true) and the statistical power (when the null hypothesis is false) by the proportion of significant estimates (two sided $\alpha < 0.05$) for the frequentist methods, or the proportion of estimates with a 95% interval that did not contain the zero treatment effect for the Bayesian methods.

We generated 5000 simulated results for each of the simulation scenarios in Table 1, and calculated the bias and mean squared error (MSE) as:

$$Bias\ (\widehat{\theta}) = \frac{1}{5000}\sum_{c=1}^{5000} (\widehat{\theta}_c - \theta)$$

$$MSE\ (\widehat{\theta}) = \frac{1}{5000}\sum_{c=1}^{5000} (\widehat{\theta}_c - \theta)^2$$

where ϑ is the true parameter value, $\widehat{\theta}_c$ is the estimated value from the c^{th} simulated data set. Monte Carlo 95% intervals for estimated mean bias and inconsistency were based on the 2.5% and 97.5% percentiles of the corresponding estimates.

Computing implementation

Bayesian network meta-analyses were implemented by Markov chain Monte Carlo (MCMC) methodology [8]. Vague or non-informative priors were used for MCMC simulations. Each simulation comprised 20,000 'burn-in' iterations followed by 40,000 posterior mean sample iterations. Posterior mean samples collected were thinned by a ratio of 5:1 to resulting in 8,000 final

posterior mean samples from each MCMC simulation. We used R 2.13.0 [28] and related packages (RJAGS) to generate data and to sample Bayesian posterior distributions. All simulations were carried out on the High Performance Computing Cluster supported by the Research Computing Service at the University of East Anglia.

Results

For the purpose of simplification, we only presented the results of selected representative scenarios below.

Estimating relative treatment effects

MSE and bias

As expected, mean squared error (MSE) is positively associated with the small number of studies, and large heterogeneity in meta-analysis (Figure 2). Of the comparison methods investigated, the AITC method has the largest MSE. With the existence of heterogeneity, there are no noticeable differences in MSE between the fixed-effect and random-effects models.

When there is no bias in simulated trials, the results of the all comparison methods are on average unbiased

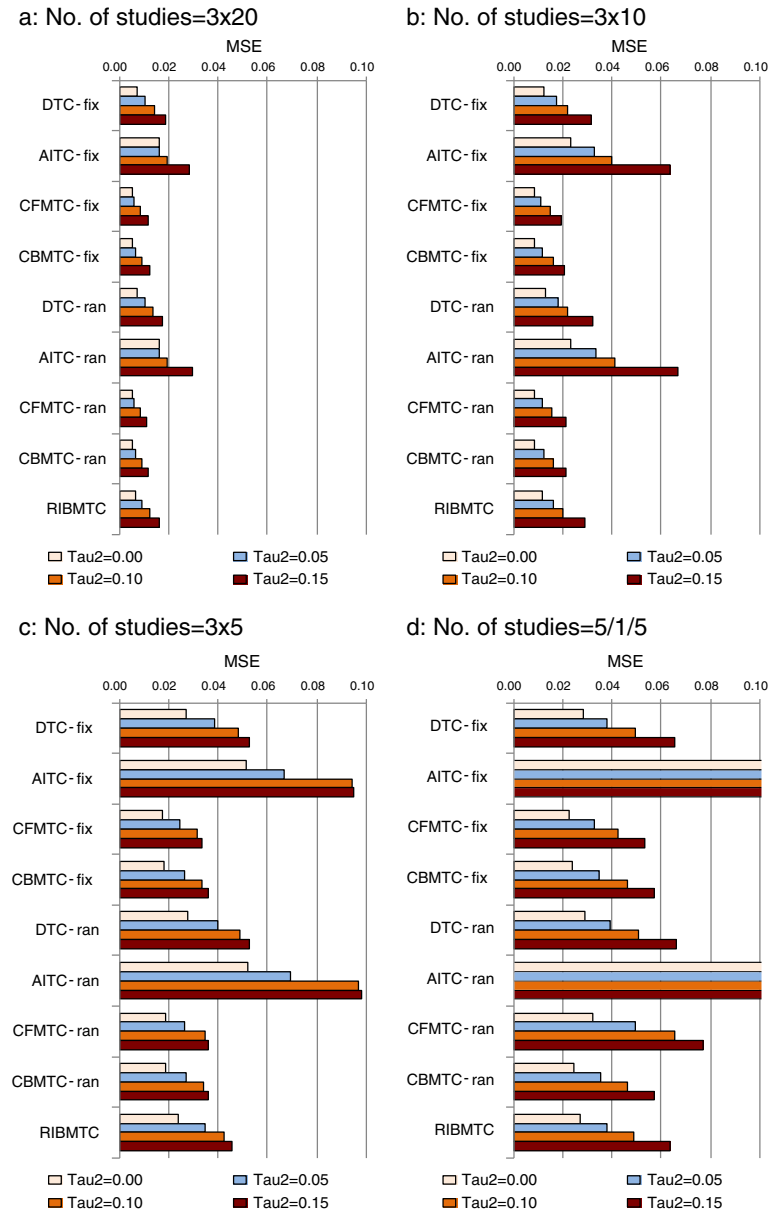

Figure 2 Mean squared error (MSE) by different comparison models (Note: baseline risk 20%; zero treatment effect; without systematic bias in trials; fix, fixed effect; ran, random-effects; Tau2 refers τ^2).

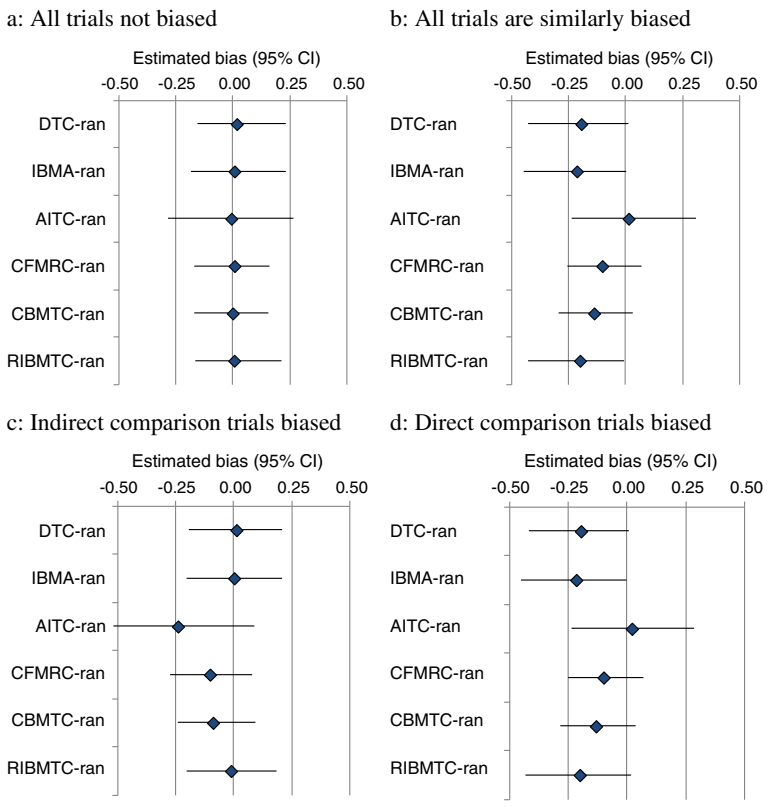

Figure 3 Bias by different comparison methods (Note: selected simulation scenarios, baseline risk = 20%; $\tau^2 = 0.05$; number of studies =3x20; random-effects analyses).

(Figure 3a). When all trials are similarly biased, the DTC and the inconsistency Bayesian MTC (RIBMTC) are fully biased, while the AITC is not biased (Figure 3b). When only the trials involved in AITC are biased, the DTC and inconsistency MTC models are unbiased (Figure 3c). The extent of bias in the consistency MTC models (both CFMTC and CBMTC) lies between the DTC and ITC. The impacts of biases in primary studies on the validity of different comparison methods are summarised in Table 2.

Type I error

Assuming zero heterogeneity across studies, there are no clear differences in the rate of type I error between different MTC methods (Figure 4). The extent of heterogeneity was clearly associated with inflated rates of type I error. In the presence of great heterogeneity, the rate of type I error is particularly large when fixed-effect models are applied. The random-effects models tend to have values closer to 0.05. However, random-effects models no longer have advantages when there is only a

Table 2 Impact of simulated biases on the results of different comparison methods

Comparison methods	Actual true biases			
	Trials not biased	All trials similarly biased	One set of AIC trials biased	DC trials biased
Direct comparison (DTC)	Not biased	Fully biased	Not biased	Fully biased
Indirect comparison (AITC)	Not biased	Not biased	Fully biased	Not biased
Consistency frequentist MTC	Not biased	Moderately biased	Moderately biased	Moderately biased
Consistency Bayesian MTC	Not biased	Moderately biased	Moderately biased	Moderately biased
Inconsistency Bayesian meta-analysis	Not biased	Fully biased	Not biased	Fully biased
Random inconsistency Bayesian MTC (RIBMTC)	Not biased	Fully biased	Not biased	Fully biased

(Note: "Fully biased" – the bias equals the bias in trials; "Moderately biased" – as a result of combining biased direct estimate and unbiased indirect estimate, or a result of combining unbiased direct estimate and biased indirect estimate).

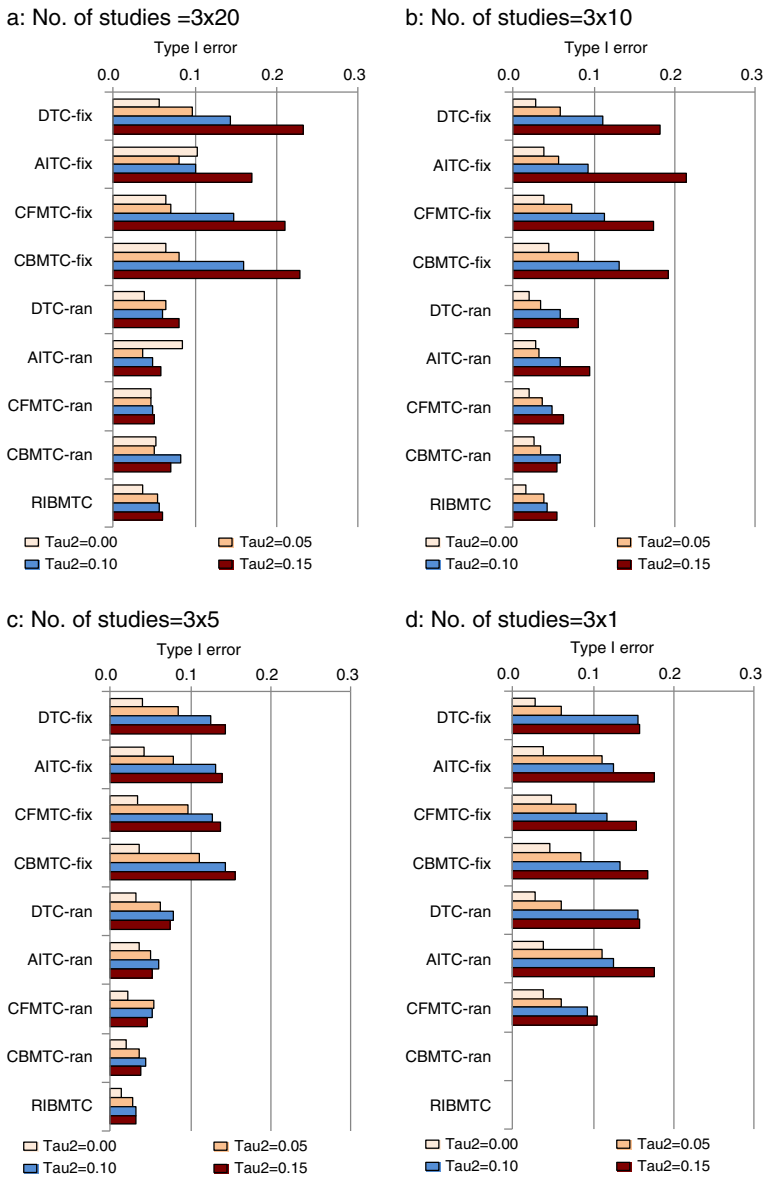

Figure 4 Type I error – proportion of significant results when true treatment effect is zero, impact of number of studies and assumed heterogeneity (Note: baseline risk =20%; fix, fixed effect; ran, random-effects; Tau2 refers τ^2).

single study available for each of the three comparisons (Figure 4d). When there is only a single study for each of the three contrasts, the rate of type I error is zero by Bayesian random-effects models (CBMTC and RIBMTC), which seems due to the unchanged vague or non-informative priors [26]. Within the fixed-effect models the different methods have similar type I error rates, as well as within the random-effects models (Figure 4).

As expected, the higher baseline risk (20%) is associated with the higher rate of type I error as compared with the lower baseline risk (10%) (data not shown).

Statistical power

As expected, the statistical power (*1-β*) is positively associated with the number of studies (Figure 5). As compared with the DTC, the statistical power of AITC is low. The pooling of DTC and AITC evidence in MTC increases the statistical power (Figure 5).

With a larger number of studies, the statistical power of all methods is reduced by the presence of heterogeneity (Figure 5a-b). The association between heterogeneity and statistical power becomes unclear when the number of studies is small (Figure 5c-d). When there is only a single study, the statistical power of all the methods is extremely low, and it is zero by the Bayesian

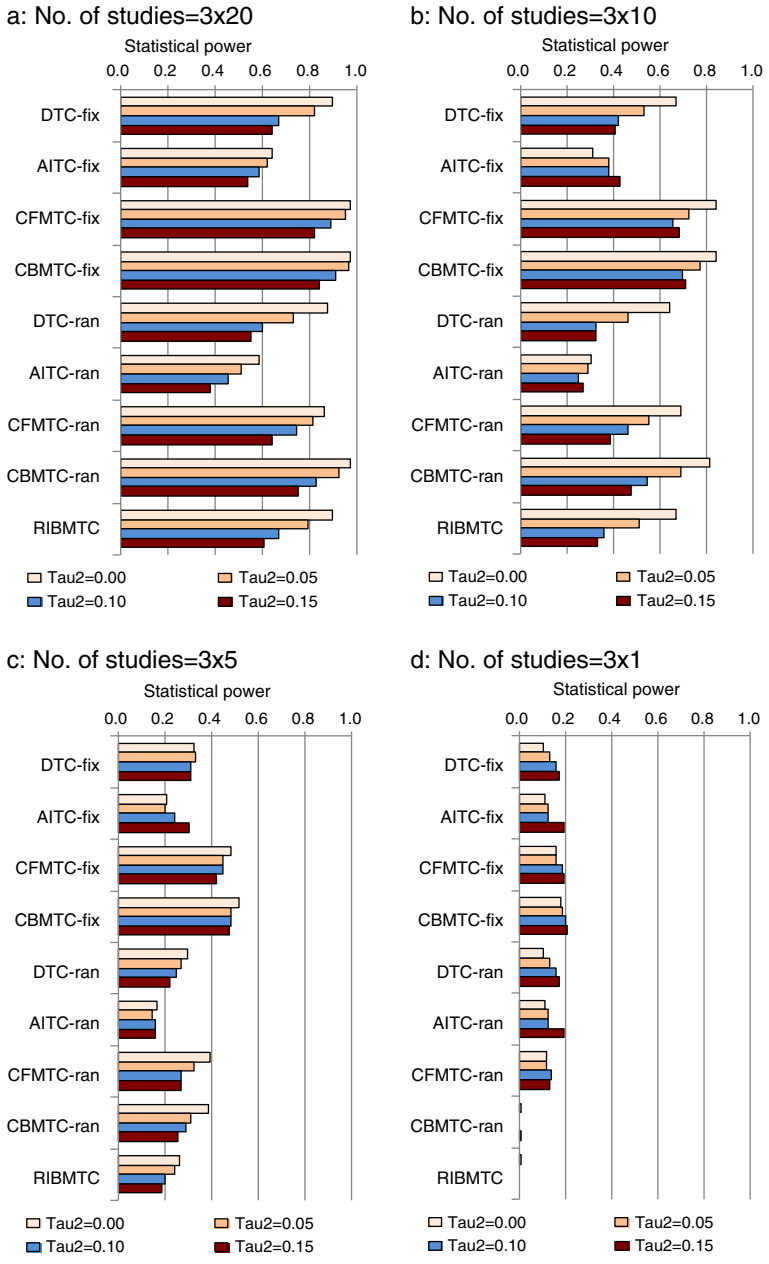

Figure 5 Statistical power to detect treatment effect (OR23 = 0.75), impact of number of studies and assumed heterogeneity (Note: Baseline risk =20%; fix, fixed effect; ran, random-effects; Tau2 refers τ^2).

random-effects models (again, due to vague or non-informative priors) (Figure 5d).

A expected, the statistical power is reduced when the baseline risk is lowered from 20% to 10% (data no shown).

Inconsistency detecting

The estimated inconsistencies by the different comparison methods are on average unbiased, but the 95% intervals are wide (Figure 6). The 95% interval of the estimated inconsistency by the RIBMTC method is much wider than by other methods.

Heterogeneity is positively associated with the rate of type I error for detecting inconsistency by the fixed-effect models, while the number of studies does not noticeably affect the rate of type I error (Figure 7). However, when there is only a single study for each of the three contrasts, the Bayesian random-effects method has zero type I error (due to the vague or non-informative priors for τ), and the rate of type I error by frequentist random-

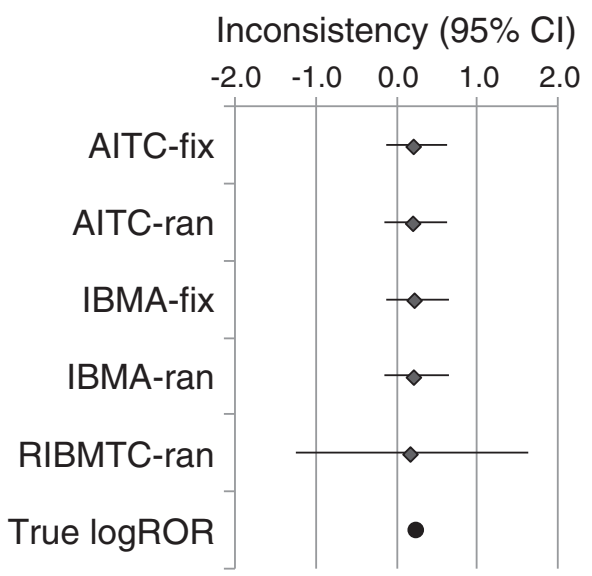

Figure 6 Estimated inconsistency (log ROR) – a selected simulation scenario (Note: true logROR = 0.223; baseline risk = 20%; number of studies =3x20; $\tau^2 = 0.10$; Tau2 refers τ^2).

effects model was similar to the fixed-effect models (Figure 7e). When there is imbalanced and singleton number of trials, the frequentist random-effects model has larger type I errors than the Bayesian random-effects method (Figure 7f).

The statistical power to detect the specified inconsistency ($P < 0.05$) increases with the increasing number of studies (Figure 8). However, the statistical power is still lower than 70% even when there are 120 studies (200 patients in each study) in the trial network (Figure 8a). By fixed-effect model, the existence of heterogeneity generally increases the power to detect inconsistency. However, the impact of heterogeneity on the power of random-effects models is unclear. When there is only one study for each of the three contrasts, the power by Bayesian random-effects model is about zero (given vague or non-informative priors for τ^2) (Figure 8e).

Discussion

Summary of findings

Mean squared error (MSE) reflects a combination of both bias and random error, which is clearly associated with the number of studies, heterogeneity, and the baseline risk. When simulated studies are not biased, the AITC method had the largest MSE, as compared with DTC and MTC methods. Given the same comparison approach, there are no noticeable differences in estimated MSE between the fixed-effect and random-effects models.

When simulated trials are unbiased, the results of all comparison methods investigated are good at predicting the true magnitude and direction of the effect. However, there are simulation scenarios under which AITC could be biased. When all trials are similarly biased, the results of AITC will be less biased than the results of DTC. This finding is consistent with the result of a previous study that evaluated the impacts of biases in trials involved in AITC [29]. Bias by MTC will lie between the bias by DTC and AITC (Table 2).

It should be noted that, in addition to the scenarios simulated in this study, bias in original trials may also be magnified if the two sets of trials for the AITC are biased in opposite directions. For example, it is possible that the relative effect of a treatment versus the common comparator is over-estimated in one set of trials, and under-estimated in another set of trials. Under this circumstance, the AITC estimate will be biased and the extent of such bias will be greater than the extent of bias in the original studies.

Estimating comparative treatment effect

The type I error of ITC and MTC methods are associated with the extent of heterogeneity, whether a fixed-effect or random-effects meta-analysis is used, and the level of baseline risk. There are no noticeable differences in type I error between different comparison methods.

As expected, the number of studies is clearly associated with the statistical power to detect specified true treatment effect. The AITC method has the lowest statistical power. When there is no assumed inconsistency or bias, the MTC increases the statistical power as compared with the power of DTC alone. There are no noticeable differences in the statistical power between different MTC methods.

Inconsistency testing

We found that the all comparison methods are on average unbiased for estimating the inconsistency between the direct and indirect estimates. The 95% intervals by the RIBMTC method are much wider than that by other methods. Heterogeneity inflates the type I error in the detection of inconsistencies by fixed-effect models. When there are singleton studies in the trial network, the frequentist based random-effects model has relatively larger type I error than the Bayesian random-effects model.

As expected, the power to detect inconsistency is positively associated with the number of studies and the use of fixed-effect models. For the inconsistency detection, heterogeneity increases the power of fixed-effect models, but reduces the power of random-effects models when the number of studies is large.

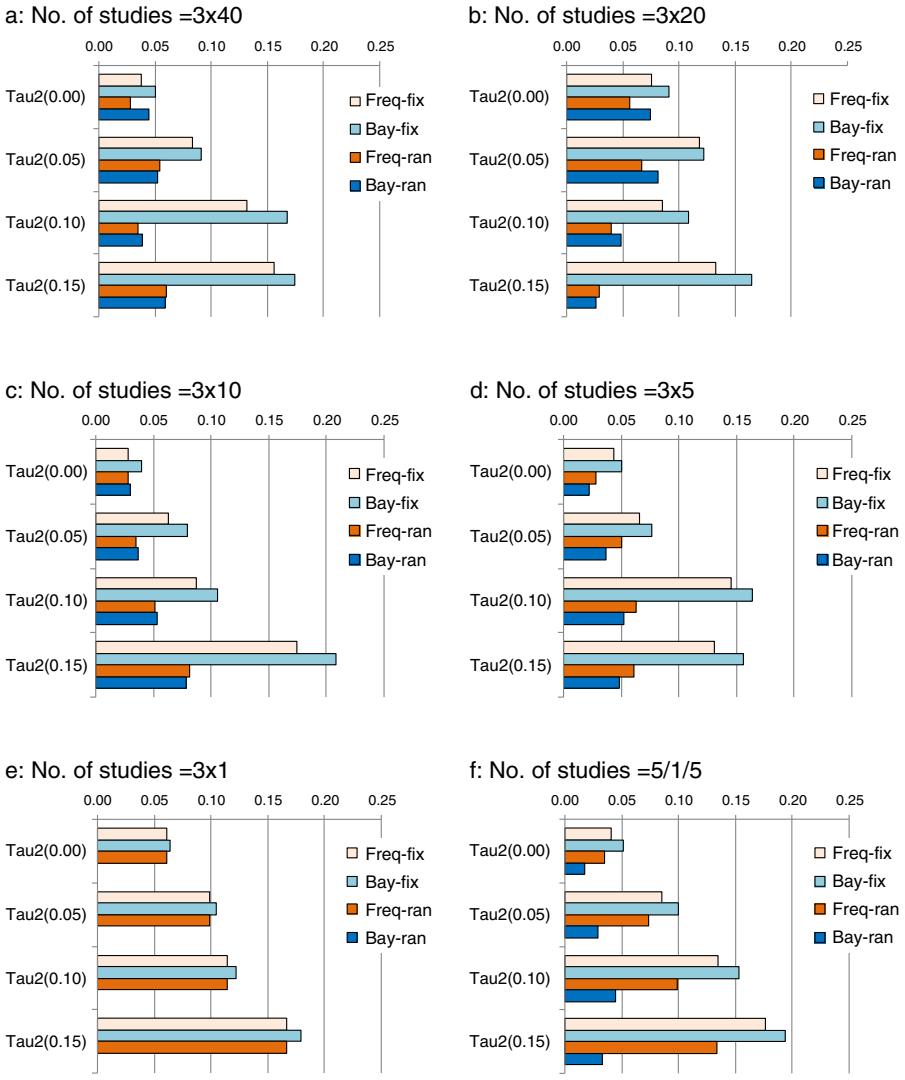

Figure 7 Type I error for inconsistency detection: impact of heterogeneity and number of studies (Note: baseline risk =20%, true lnROR = 0; tau2 refers τ^2; Freq-fix, frequentist fixed-effect; Freq-ran, frequentist random-effects; Bay-fix, Bayesian fixed-effect; Bay-ran, Bayesian random-effects).

Comparing with previous studies

Methods of frequentist based indirect comparison have been investigated in several previous simulation studies [1,20,21]. A study found that the Bucher's method and logistic regression generally provided unbiased estimates [1]. The simulation scenarios evaluated in that study was limited by using data from a single trial. In another study, Wells and colleagues simulated variance, bias and MSE by the DTC and AITC method [21]. It was reported that the observed variance, bias and MSE for the AITC were larger than that for the DTC, particularly when the baseline risk was low [21]. A more recent simulation study by Mills and colleagues reported findings from an investigation of the Bucher's ITC method [20]. They found that the AITC method lacks statistical power, particularly in the presence of heterogeneity, and has high risk of over-estimation when only a single trial is available in one of the two trial sets. However, they did not compare the performance of the AITC and the corresponding DTC or MTC [20].

Bayesian MTC methods have not been investigated in previous simulation studies. In the current study, we investigated the performance of statistical methods for DTC, AITC, frequentist and Bayesian MTC. The simulation results reveal the complex impacts of biases in primary studies on the results of direct, indirect and mixed treatment comparisons. When the simulated primary studies are not systematically biased, the AITC and MTC methods are not systematically biased, although the AITC method has the largest MSE. Depending on

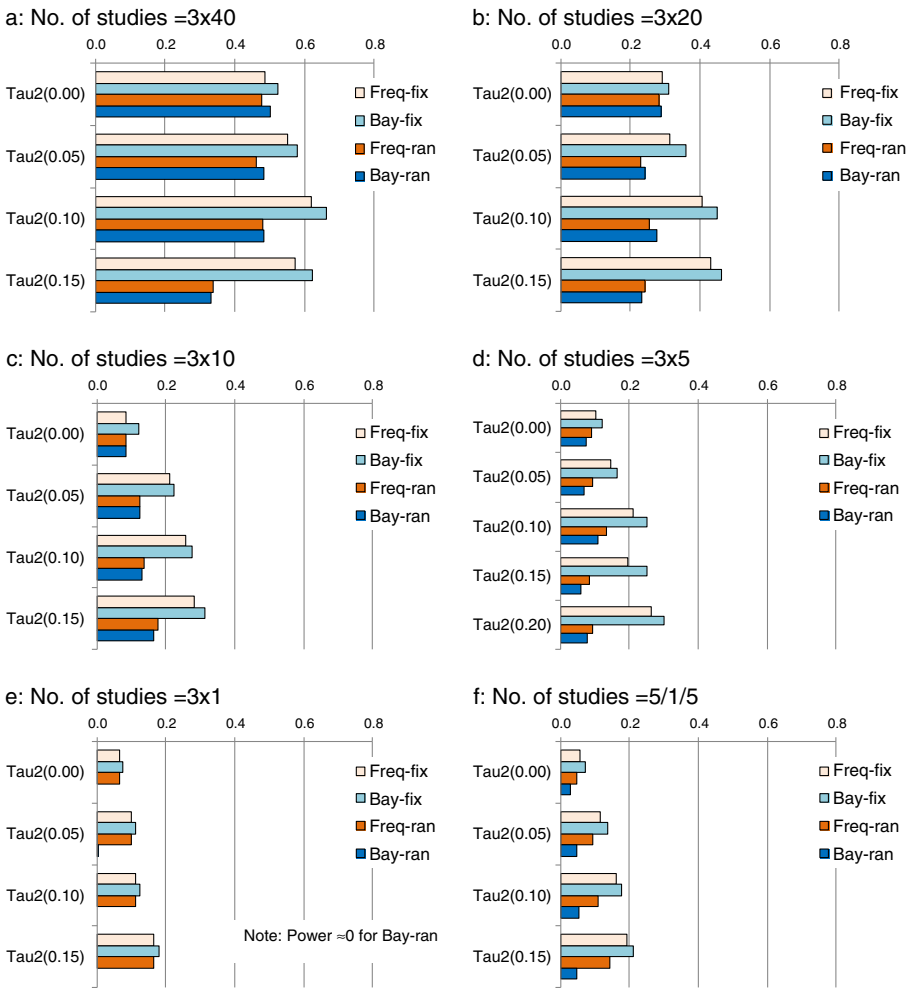

Figure 8 Statistical power to detect inconsistency: impact of heterogeneity and number of studies (Note: baseline risk =20%, true lnROR = 0.223; tau2 refers τ^2. Freq-fix, frequentist fixed-effect; Freq-ran, frequentist random-effects; Bay-fix, Bayesian fixed-effect; Bay-ran, Bayesian random-effects).

the extent and direction of bias in primary studies, the AITC and MTC estimates could be more or less biased than the DTC estimates.

In the existence of heterogeneity and a small number of studies, AITC and MTC methods have indeed the inflated rate of type I error and a low statistical power. It is important to note that the performance of the corresponding DTC is similarly affected. The performance of the DTC method is superior to the performance of the AITC method. However, the statistical power of MTC is generally higher than the corresponding DTC.

It is the first time that the power to detect inconsistency in network meta-analysis has been investigated by simulations. The low power to detect inconsistency in network meta-analysis seems similar to the low power to detect heterogeneity in pair-wise meta-analysis [30].

Limitations of the study

Due to the restriction of available resource, a limited number of simulation scenarios were considered. Clearly, the performance of a model will depend on whether the simulation scenario matches the model's assumptions. For example, the fixed-effect model should not be used when there is heterogeneity across multiple studies, in order to avoid the inflated type I error.

In this paper, the simple network containing three sets of two-arm trials with a single completed loop is considered. We evaluated the methods for detecting inconsistency, and did not consider models for investigating causes of inconsistency. Therefore, further simulation studies are required to evaluate complicated networks involving more than three different treatments and containing trials with multiple arms. In addition, further simulation studies are required to evaluate the performance of regression models that incorporate study-level

covariates for investigating the causes of heterogeneity and inconsistency in network meta-analysis [18,19,31].

For MCMC simulations, we used vague or non-informative priors [32]. When the number of studies involved is large, finding of the study were unlikely to be different if more informative priors had been used. However, further research is required to investigate whether an informed prior for between-study variance would be more appropriate when the number of studies involved in a Bayesian meta-analysis is very small [26].

Implications to practice and research

The results of any comparison methods (including direct comparison trials) may be biased as a consequence of bias in primary trials involved. To decide which comparison method may provide more valid or less biased results, it is helpful if we can estimate the extent and direction of possible biases in primary studies. Empirical evidence indicated the existence of bias in randomised controlled trials [33-35], particularly in trials that had outcomes subjectively measured without appropriate blinding [36,37]. Although it is usually difficult to estimate the magnitude of bias, the likely direction of bias may be estimated. For example, it may be assumed that possible bias was likely to result in an over-estimation of treatment effect of active or new drugs when they are compared with placebo or old drugs [38]. More complicated models could also be explored for estimating bias in evidence synthesis [39-41].

For detecting inconsistency, the fixed-effect methods have a higher rate of type I errors as well as a higher statistical power as compared with the random-effects methods. The performances of the Bayesian and frequentist methods are generally similar. When there are singleton trials in evidence network, the rate of type I error by frequentist random-effects method is larger than by the Bayesian random-effects method. This is due to the under-estimation of between-study variance by the frequentist method, while the Bayesian method provides an estimate of between-study variance using all data available in the whole network of trials [32]. However, when there is a single study for each of the all comparisons, Bayesian random-effects models should be avoided.

Imbalanced distribution of effect-modifiers across studies may be a common cause of both heterogeneity in pair-wise meta-analysis and evidence inconsistency in network meta-analysis [17]. However, it is helpful to distinguish the heterogeneity in pair-wise meta-analysis and inconsistency in network meta-analysis. Under the assumption of exchangeability, the results of direct and indirect comparisons could be consistent in the presence of large heterogeneity in meta-analyses. For example, the inflated type I error rate in detecting inconsistency by the fixed-effect models can be corrected by the use of random-effects models. It is also possible to observe significant inconsistencies between direct and indirect estimates when there is no significant heterogeneity in the corresponding pair-wise meta-analyses. The association between heterogeneity and the statistical power to detect inconsistency is complex, depending on whether the fixed-effect or random-effects model is used and the number of studies involved.

A major concern is the very low power of commonly used methods to detect inconsistency in network meta-analysis when it does exist. Therefore, inconsistency in network meta-analysis should not be ruled out based only on the statistically non-significant result of a statistical test. For all network meta-analysis, trial similarity and evidence consistency should be carefully examined [2,42].

Conclusions

Of the comparison methods investigated, the indirect comparison has the largest mean squared error and thus the lowest certainty. The direct comparison is superior to the indirect comparison in terms of statistical power and mean squared error. Under the simulated circumstances in which there are no systematic biases and inconsistencies, the performances of mixed treatment comparisons are generally better than the performance of the corresponding direct comparisons.

When there are no systematic biases in primary studies, all methods investigated are on average unbiased. Depending on the extent and direction of biases in different sets of studies, indirect and mixed treatment comparisons may be more or less biased than the direct comparisons. For inconsistency detection in network meta-analysis, the methods evaluated are on average unbiased. The statistical power of commonly used methods for detecting inconsistency in network meta-analysis is low.

In summary, the statistical methods investigated in this study have different advantages and limitations, depending on whether data analysed satisfies the different assumptions underlying these methods. To choose the most valid statistical methods for network meta-analysis, an appropriate assessment of primary studies included in the evidence network is essential.

Abbreviations

AITC: Adjusted indirect treatment comparison; CBMTC: Consistency Bayesian mixed treatment comparison; CFMTC: Consistency frequentist mixed treatment comparison; DTC: Direct treatment comparison; IBMA: Inconsistency Bayesian meta-analysis; ITC: Indirect treatment comparison; MCMC: Markov chain Monte Carlo; MSE: Mean squared error; MTC: Mixed treatment comparison; OR: Odds ratio; RCT: Randomised controlled trial; ROR: Ratio of odds ratios; RIBMTC: Random inconsistency Bayesian mixed treatment comparison.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
FS, AC and MOB conceived the idea and designed research protocol. JM, AC and FS developed simulation programmes and conducted computer simulations. FS analysed data and prepared the draft manuscript. All authors commented on the manuscript. FS had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Acknowledgment
This study was funded by UK Medical Research Council (Methodological Research Strategic Grant: G0901479). The research presented was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service (RSCSS) at the University of East Anglia.

Received: 15 June 2012 Accepted: 4 September 2012
Published: 12 September 2012

References

1. Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, Bradburn M, Eastwood AJ: **Indirect comparisons of competing interventions.** *Health Technol Assess* 2005, **9**(26):1–134. iii-iv.
2. Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG: **Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: a survey of published systematic reviews.** *BMJ* 2009, **338:**b1147. doi:10.113/bmj.b1147.
3. Donegan S, Williamson P, Gamble C, Tudur-Smith C: **Indirect comparisons: a review of reporting and methodological quality.** *PLoS One* 2010, **5**(11):e11054.
4. Edwards SJ, Clarke MJ, Wordsworth S, Borrill J: **Indirect comparisons of treatments based on systematic reviews of randomised controlled trials.** *Int J Clin Pract* 2009, **63**(6):841–854.
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):683–691.
6. Higgins JP, Whitehead A: **Borrowing strength from external trials in a meta-analysis.** *Stat Med* 1996, **15**(24):2733–2749.
7. Lumley T: **Network meta-analysis for indirect treatment comparisons.** *Stat Med* 2002, **21**(16):2313–2324.
8. Lu G, Ades AE: **Combination of direct and indirect evidence in mixed treatment comparisons.** *Stat Med* 2004, **23**(20):3105–3124.
9. Lu G, Ades AE: **Assessing evidence inconsistency in mixed treatment comparisons.** *J Am Stat Assoc* 2006, **101**(474):447–459.
10. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L, Cappelleri JC: **Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1.** *Value Health* 2011, **14**(4):417–428.
11. Jansen JP, Schmid CH, Salanti G: **Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons.** *J Clin Epidemiol* 2012, **65**(7):798–807.
12. Song F, Altman DG, Glenny AM, Deeks JJ: **Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses.** *BMJ* 2003, **326**(7387):472–475.
13. Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, Holland R, Chen YF, Glenny AM, Deeks JJ, *et al*: **Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study.** *BMJ* 2011, **343:**d4909.
14. Chou R, Fu R, Huffman LH, Korthuis PT: **Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses.** *Lancet* 2006, **368**(9546):1503–1515.
15. Madan J, Stevenson MD, Cooper KL, Ades AE, Whyte S, Akehurst R: **Consistency between direct and indirect trial evidence: is direct evidence always more reliable?** *Value Health* 2011, **14**(6):953–960.
16. Gartlehner G, Moore CG: **Direct versus indirect comparisons: a summary of the evidence.** *Int J Technol Assess Health Care* 2008, **24**(2):170–177.
17. Dias S, Welton NJ, Caldwell DM, Ades AE: **Checking consistency in mixed treatment comparison meta-analysis.** *Stat Med* 2010, **29**(7–8):932–944.
18. Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ: **Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation.** *Stat Med* 2009, **28**(14):1861–1881.
19. Salanti G, Marinho V, Higgins JP: **A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered.** *J Clin Epidemiol* 2009, **62**(8):857–864.
20. Mills EJ, Ghement I, O'Regan C, Thorlund K: **Estimating the power of indirect comparisons: a simulation study.** *PLoS One* 2011, **6**(1):e16237.
21. Wells GA, Sultan SA, Chen L, Khan M, Coyle D: *Indirect evidence: indirect treatment comparisons in meta-analysis*. Ottawa, Canada: Canadian Agency for Drugs and Technologies in Health; 2009.
22. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Controlled Clin Trials* 1986, **7:**177–188.
23. Salanti G, Higgins JP, Ades AE, Ioannidis JP: **Evaluation of networks of randomized trials.** *Stat Methods Med Res* 2008, **17**(3):279–301.
24. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE: *NICE DSU Technical Support Document 4: Inconsistency in Network of Evidence Based on Randomised Controlled Trials*: 2011. Available from http://www.nicedsu.org.uk.
25. Eckermann S, Coory M, Willan AR: **Indirect comparison: relative risk fallacies and odds solution.** *J Clin Epidemiol* 2009, **62**(10):1031–1036.
26. Pullenayegum EM: **An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes.** *Stat Med* 2011, **30**(26):3082–3094.
27. Higgins JP, Altman DG: **Chapter 8: Assessing risk of bias in included studies**. In *Cochrane Handbook for Systematic Reviews of Interventions*. Edited by Higgins J, Green S. Chichester: Wiley; 2008.
28. R_Development_Core_Team: R: *A language and environment for statistical computing. In. Vienna, Austria*. Vienna, Austria: R Foundation for Statistical Computing; 2008.
29. Song F, Harvey I, Lilford R: **Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions.** *J Clin Epidemiol* 2008, **61**(5):455–463.
30. Hardy RJ, Thompson SG: **Detecting and describing heterogeneity in meta-analysis.** *Stat Med* 1998, **17**(8):841–856.
31. Nixon RM, Bansback N, Brennan A: **Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis.** *Stat Med* 2007, **26**(6):1237–1254.
32. Dias S, Welton NJ, Sutton AJ, Ades AE: *NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials*: 2011. Available from: http://www.nicedsu.org.uk.
33. Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Krebs Seida J, Klassen TP: **Risk of bias versus quality assessment of randomised controlled trials: cross sectional study.** *BMJ* 2009, **339:**b4012.
34. Juni P, Altman DG, Egger M: **Systematic reviews in health care: assessing the quality of controlled clinical trials.** *BMJ* 2001, **323**(7303):42–46.
35. Schulz KF, Chalmers I, Hayes RJ, Altman DG: **Empirical evidence of bias.** *Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA* 1995, **273**(5):408–412.
36. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA: **Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study.** *BMJ* 2008, **336**(7644):601–605.
37. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, Ravaud P, Brorson S: **Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors.** *BMJ* 2012, **344:**e1119.
38. Chalmers I, Matthews R: **What are the implications of optimism bias in clinical research?** *Lancet* 2006, **367**(9509):449–450.
39. Thompson S, Ekelund U, Jebb S, Lindroos AK, Mander A, Sharp S, Turner R, Wilks D: **A proposed method of bias adjustment for meta-analyses of published observational studies.** *Int J Epidemiol* 2011, **40**(3):765–777.
40. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG: **Bias modelling in evidence synthesis.** *J R Stat Soc A* 2009, **172**(1):21–47.

41. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JA: **Models for potentially biased evidence in meta-analysis using empirically based priors.** *J R Stat Soc A* 2009, **172**(Part 1):119–136.
42. Xiong T, Parekh-Burke S, Loke YK, Abdelhamid A, Sutton AJ, Eastwood AJ, Holland R, Chen YF, Walsh T, Glenny AM, *et al*: **Assessment of trial similarity and evidence consistency for indirect treatment comparison: an empirical investigation.** *J Clin Epidemiol* 2012, In press.

doi:10.1186/1471-2288-12-138
Cite this article as: Song *et al.*: Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. *BMC Medical Research Methodology* 2012 **12**:138.