A comparison of Bayesian and frequentist methods in random-effects network meta-analysis of binary data

Svenja E. Seide | Katrin Jensen | Meinhard Kieser

Institute of Medical Biometry and Informatics, Ruprecht-Karls University Heidelberg, Heidelberg, Germany

The performance of statistical methods is often evaluated by means of simulation studies. In case of network meta-analysis of binary data, however, simulations are not currently available for many practically relevant settings. We perform a simulation study for sparse networks of trials under between-trial heterogeneity and including multi-arm trials. Results of the evaluation of two popular frequentist methods and a Bayesian approach using two different prior specifications are presented. Methods are evaluated using coverage, width of intervals, bias, and root mean squared error (RMSE). In addition, deviations from the theoretical surface under the cumulative rankings (SUCRAs) or P-scores of the treatments are evaluated. Under low heterogeneity and when a large number of trials informs the contrasts, all methods perform well with respect to the evaluated performance measures. Coverage is observed to be generally higher for the Bayesian than the frequentist methods. The width of credible intervals is larger than those of confidence intervals and is increasing when using a flatter prior for between-trial heterogeneity. Bias was generally small, but increased with heterogeneity, especially in netmeta. In some scenarios, the direction of bias differed between frequentist and Bayesian methods. The RMSE was comparable between methods but larger in indirectly than in directly estimated treatment effects. The deviation of the SUCRAs or P-scores from their theoretical values was mostly comparable over the methods but differed depending on the heterogeneity and the geometry of the investigated network. Multivariate meta-regression or Bayesian estimation using a half-normal prior scaled to 0.5 seems to be promising with respect to the evaluated performance measures in network meta-analysis of sparse networks.

Keywords
Bayesian and frequentist methods, binary data, multi-arm trials, network meta-analysis, random-effects model, simulation study

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. Research Synthesis Methods published by John Wiley & Sons Ltd.
Network meta-analysis, also known as mixed-treatment or multiple treatment comparison, is the quantitative synthesis of evidence for multiple treatments of the same indication. As a generalization of pairwise meta-analysis, one of its advantages is the fact that direct and indirect evidence can be combined into the same analysis with the potential to estimate treatment effects more precisely. Furthermore, network meta-analysis allows to rank the available treatments according to their safety or efficacy. This is a highly relevant question in practice as patients and medical professionals are interested in receiving/giving the best available treatment with respect to a certain outcome. Although a considerable amount of methodological research in network meta-analysis has been published lately, only a limited number of simulation studies compares these existing methodological advances. Unlike in other areas of biomedical research, methods are commonly evaluated on an empirical example, resulting in the notion of differences between candidate models rather than the evaluation against known theoretical values. One challenge inherent in network meta-analysis are multi-arm trials, which potentially contribute multiple treatment effect estimates to a network, thus violating the independence assumption between observed trial results. This dependence needs to be reflected in the statistical model used. However, published simulation studies are restricted either to a common-effect model, where the strong assumption that one true effect size is shared by all included studies is necessary, or to simulation studies for network meta-analyses that relax this assumption of between-study homogeneity and allow (constant) between-study heterogeneity in the different treatment contrasts that are currently restricted to a network of two-arm trials. However, this neglects the empirical situation where multi-arm trials are often part of a network of treatments. Both of these restrictions are the exception in practice rather than the rule.

Empirically, networks of interventions are often sparse with respect to both the number of available direct comparisons relative to the number of potential comparisons (the network density) and the number of trials that contribute to such direct evidence, which may complicate the estimation. Sparse networks in the sense that they have a low network density are likely in situations where a large number of treatments (l) is compared, as the number of potential comparisons (0.5l(l − 1)) depends directly on l. Network meta-analyses in situations where networks are sparse are therefore increasingly observed in practice. For example, Petropoulou et al. performed a systematic review including network meta-analyses with at least four treatments. They observed the median number of treatments in a network meta-analysis to be 7 and a majority of networks being analyzed under the random-effects model. Furthermore, Tonin et al. conducted a systematic review on network meta-analyses performed during the last few years and identified 167 network meta-analyses where the mean number of treatments was 9. Reliable estimation of between-study heterogeneity is known to be problematic when only few studies are available in pairwise meta-analysis and potentially pose also a challenge in network meta-analysis. Depending on the network geometry, an additional challenge may be the absence of a common comparator over all trials. Then, different network clusters may occur where subnetworks may be less sparse, but only few connections between these subnetworks are available, either through a common comparator or a bridge in the network. It is still unclear how the geometry of the network and the precision with which a bridge between such clusters is estimated may influence the overall network meta-analysis' results.

While network meta-analysis in practice encounters these challenges, little is known about the performance of different network meta-analytical methods in such situations. To add to existing evidence, we conduct a simulation study investigating the characteristics of random-effects network meta-analyses with multi-arm trials and a binary endpoint. Like all statistical models, network meta-analysis is based upon a number of modeling assumptions that should be satisfied in order to avoid misleading interpretations. Three important assumptions in network meta-analysis concern transitivity, consistency, and heterogeneity of the network. Transitivity assumes missing treatments in a trial to be missing at random, therefore justifying the use of indirect estimates as valid replacements of missing direct estimates stemming from head-to-head trials. This assumption cannot be statistically tested but needs to be evaluated conceptually. Consistency assumes exchangeability in the sense that available direct and indirect evidence for each contrast in the network are similar and can be consistently combined. This assumption may be tested in some statistical models, eg, when a design-by-treatment interaction term is explicitly modeled, or a node-splitting model is adapted. Although consistency is one of the key assumptions in network meta-analysis, we will not evaluate the case of inconsistent networks. This is due to the fact that violations of the exchangeability assumption may not only result in inconsistency but can also manifest as heterogeneity, and we want to concentrate on the latter. In addition, if multi-arm trials are part of the network, the parametrization of such a node-split model may influence the estimation results, resulting in...
different modeling options for multi-arm trials. The assumption of heterogeneity concerns the exchangeability between trials that contribute data to the same contrast.\textsuperscript{32} A common but simplifying assumption allows variation between trials by employing a random-effect but assumes constant and common variance over different contrasts in the network. We simulated the simplified case, a consistent, transitive network with constant and common heterogeneity variances and multi-arm trials within in the network. We restrict ourselves to a binary endpoint and use the odds ratio (OR) as effect measure. To reflect empirically common situations, we will concentrate on sparse networks where only few trials are available per contrast and investigate situations with and without a common comparator.

This paper is organized as follows. In Section 2, we present the evaluated data models and estimation methods. An empirical example data set to motivate our simulation study and the considered scenarios are discussed in Section 3. Section 4 is dedicated to the simulation study and the considered scenarios are discussed in Section 3. Section 4 is dedicated to the simulation results, which we will discuss in Section 5.

2 | ESTIMATION METHODS AND THEIR UNDERLYING DATA MODELS

For binary outcomes, the input (aggregated) data in a network meta-analysis can be either provided on arm level (e.g., observed number of events and number of patients) or contrast level (e.g., estimate of the relative treatment effect such as log-OR and its standard error for any two treatments in a trial). As contrast-level data can be easily retrieved from arm-level data while the opposite is not possible without additional assumption, we will simulate arm-level data. Careful evaluation of clinical and methodological heterogeneity across trials is important. If there is unexplained heterogeneity, either identified through clinical or statistical investigations, a random-effect rather than a common-effect model is preferred. Detailed reviews about assessing and dealing with between-trial heterogeneity (\(\tau^2\)) in the context of network meta-analysis can, for example, be found in Gagnier\textsuperscript{33} or Dias et al.\textsuperscript{10} In this simulation study, we generally refer to the random-effects model for estimation, even though we will include homogeneity in our investigation in some scenarios (by setting \(\tau^2 = 0\)).

In the following, we review three commonly used statistical methods for network meta-analysis along with their underlying data models, stemming from two different statistical inference frameworks that are typically used in network meta-analysis: frequentist and Bayesian approaches. For convenience and identifiability with the supplementary R code,\textsuperscript{34} we will abbreviate the presented methods by the name of the respective R extension used to estimate the results of the simulation study. The investigated methods are the graph-theoretical approach “netmeta” by Rücker,\textsuperscript{35} the approach based on multivariate meta-regression “mvmeta” as presented by White et al.,\textsuperscript{28} and a generalized linear mixed model as described by Dias et al.\textsuperscript{10} “gemtc.” While the first two apply frequentist methods for estimation, the latter uses a Bayesian approach for which we will evaluate two different prior specifications for the between-study heterogeneity \(\tau^2\). As data models, we review the normal-normal hierarchical model and the binomial-normal hierarchical model, which are both commonly used in practice. The methods included in this simulation study will consist of a mixture of two-stage methods using an approximate likelihood and one-stage methods using an exact likelihood. Both frequentist methods are two-stage methods, while the Bayesian method is a one-stage method.

2.1 | Netmeta

Exploiting the similarity between networks of treatments and electrical networks, Rücker\textsuperscript{35} proposed a graph-theoretical model for network meta-analysis. The vector of relative treatment effects \(\theta\) for the different contrasts is estimated employing a linear model assuming normality for the observed treatment effects \(Y = X\theta + e\) and a normally distributed error term \(e \sim N(0, \Sigma)\). Each trial \(i\) in the network consists of \(k \geq 2\) arms that contribute evidence on the different treatment effects or contrasts. In a random-effects network meta-analysis, it is additionally assumed that the study-specific treatment effect for a treatment \(k\) contrasted against the study-specific reference treatment \(1\) (\(\theta_{i,1k}\)) varies randomly around the underlying theoretical treatment effect \(\theta_{1k}\) with variance \(\tau^2\), such that \(\theta_{i,1k} \sim N(\theta_{1k}, \tau^2)\). This between-study heterogeneity is assumed to be common between and constant for all different treatment effects in the network meta-analysis and is estimated by the generalized DerSimonian-Laird methods of moments as described by Jackson et al.\textsuperscript{36} The heterogeneity variance \(\tau^2\) is incorporated into the individual studies’ weights that are used to weight each study by the inverse of its variance. Under the presence of multi-arm trials, these study-specific variances are adjusted to account for the correlation introduced by multiple use of one study arm. The treatment effects in this normal-normal hierarchical model are then estimated with the frequentist approach of weighted least-squares using the inverse variance weights, which include the estimated study-specific and between-study variances. As the question which treatment is “best” with respect to some
outcome is highly relevant in practice, the P-score will be used to evaluate the ranking of treatments in netmeta. The P-score evaluates the certainty that a treatment is better than another one, averaged over all competing treatments except the one under consideration.37

2.2 Mvmeta

White et al28 suggest to estimate a network meta-analysis by using a multivariate random-effects network meta-regression. There, the available treatment contrasts of the full network of treatments in each study are modeled as \( y_i \sim N(X_i \theta, S_i + \Sigma) \). The different endpoints are modeled simultaneously under the additional assumption of a correlation coefficients between those endpoints. In the common case that a trial does not include all potential treatment arms, the respective elements are treated as missing in this trial. The design matrix \( X_i \) codes the information which treatment of that study is contrasted against the respective study-specific reference treatment. The covariance matrix \( S_i \) contains the study-specific variance terms and the covariances between the different treatment arms within a trial, and \( \Sigma \) contains the between-study heterogeneity covariance matrix. The latter is assumed to be similar over all trials, while the first is needed to be estimated for each trial separately. However, this model requires explicit formulation of covariance terms or at least a covariance structure as an additional assumption. The unknown parameters in this model are the regression coefficients \( \theta \) that are used to estimate the treatment effects (per endpoint) and the between-study covariance matrix \( \Sigma \). This covariance matrix is difficult to estimate, and different estimation techniques may be employed. Following the suggestion of White et al,28 we use restricted maximum likelihood (REML) to avoid the negative bias associated with maximum likelihood estimation of variance components. In multivariate meta-regression, one main advantage is the simultaneous modeling of different endpoints of interest within one regression. As we are only considering a single endpoint during the simulation study, we are not using this feature of mvmeta within our comparisons. To evaluate the certainty that a treatment is best, the parametric resampling approach by White et al28,38 to the Bayesian surface under the cumulative ranking (SUCRA) is used by Dias et al.10 The events per trial arm are modeled as realizations from a binomial distribution with probability equal to the (observed) event rate in the trial \( r_{ik} \sim Bin(p_{ik}, n_{ik}) \). Evidence on the relative treatment effects per contrast is then synthesized by fitting a generalized linear model under a Bayesian framework. Considering the binomial data structure, a logit link is used to fit a logistic regression model \( \logit(p_{ik}) = \mu_i + \theta_{i1k} \). There, the relative treatment effect \( \theta_{i1k} \) is added to a trial-specific baseline term, which is regarded as nuisance term adjusting for the baseline treatment of the considered trial. When a random-effects model is used, the trial-specific treatment effects are assumed to be normally distributed around the true treatment effect with the between-trial heterogeneity variance \( \tau^2 \), ie, \( \theta_{i1k} \sim N(\theta_{b1k}, \tau^2_{b1k}) \). The heterogeneity structure in this generalized linear mixed model corresponds to the one used in netmeta. In this approach, the SUCRA,39 which is the inversely scaled average rank of treatment37 in a network, is used to evaluate the posterior probability that a treatment is best.

This model is estimated in a Bayesian framework using two different prior specifications for the between-trial heterogeneity. As suggested by Dias et al,10 we use a uniform prior \((U(0,4))\) in one specification. This prior assumes values of the between-trial heterogeneity lying between 0 and 4 and being equally likely. Considering that the prior is defined on the log OR scale, it assigns considerable mass to high heterogeneity. We therefore followed Friede et al40 who recommend the use of a half-normal prior \((HN(0.5)\) or \(HN(1))\) in pairwise binary meta-analysis of few studies. We chose the half-normal prior scaled around \((0.5)\) for our simulation study. For the treatment effects, we used the same uniform priors \((U(0,100))\) for all treatments in all simulated scenarios. All Bayesian models were estimated by Markov-Chain-Monte-Carlo simulation.

3 APPLICATION IN RHEUMATOID ARTHRITIS AND SIMULATION SETTINGS

To inspire the theoretical values of the simulation study, we use an empirical network meta-analysis published by Warren et al.14 The reported network consists of 13 primary studies each contributing evidence in two to four treatment arms to a network on treatments of rheumatoid arthritis. Several anti-tumor-necrosis factor (anti-TNF) treatments are available for patients suffering from rheumatoid arthritis to suppress the related inflammatory response of the immune system. All of these treatments are compared with placebo in the study published by Warren et al,14 resulting in a star-shaped geometry.
Anti-TNF treatments are associated with (serious) adverse events, such as, beside others, lymphomas, infections in the respiratory tract, congestive heart failure, and injection side reactions. In the empirical investigation, the occurrence of such adverse events was investigated. In total, eight different treatments form the evidence network and 76 adverse events in 7233 participants were observed. To evaluate a potential association between these treatments and the risk of treatment-related malignancy, Warren et al performed different (network) meta-analyses, depending on whether different doses of the same treatment arms are lumped or split in the respective meta-analysis model. The empirical data set is shown in Table 1. In the notation used by Warren et al, we choose model “D” to define theoretical values for the simulation. We used the empirical example to define the number of treatments, number of studies per arm, network geometry, and the OR as effect measure.

Several typical challenges in network meta-analyses are prominent in this data set. Firstly, the network is sparse with respect to the number of contrasts, which are informed by direct head-to-head comparisons, forming a network where consistency is difficult to verify for some treatments. Secondly, it is sparse with respect to the number of trials, which are available to estimate most of the available direct comparisons, leading to problems in the estimation of between-study heterogeneity and treatment effects. Thirdly, the empirically observed network is also

| Study               | Treatment 1 | Patients Arm 1 | Events Arm 1 | Treatment 2 | Patients Arm 2 | Events Arm 2 |
|---------------------|-------------|----------------|--------------|-------------|----------------|--------------|
| Ericson 1999        | Placebo     | 105            | 0            | Etanercept medium | 111           | 0            |
| Ericson 1999        | Placebo     | 105            | 0            | Etanercept low   | 343           | 2            |
| Moreland 1999       | Placebo     | 80             | 0            | Etanercept medium | 78            | 1            |
| Moreland 1999       | Placebo     | 80             | 0            | Etanercept low   | 76            | 0            |
| Genovese 2002       | Placebo     | 217            | 4            | Etanercept medium | 207           | 5            |
| Genovese 2002       | Placebo     | 217            | 4            | Etanercept low   | 208           | 5            |
| Combe 2006          | Placebo     | 50             | 0            | Etanercept medium | 204           | 1            |
| Van der Heijde 2006 | Placebo     | 228            | 1            | Etanercept medium | 454           | 10           |
| Weismann or Baumgartner 2004 | Placebo | 229            | 2            | Etanercept medium | 266           | 2            |
| Furst 2003          | Placebo     | 318            | 0            | Adalimumab medium | 318           | 4            |
| Weinblatt 2003      | Placebo     | 62             | 0            | Adalimumab medium | 67            | 0            |
| Weinblatt 2003      | Placebo     | 62             | 0            | Adalimumab low    | 69            | 0            |
| Weinblatt 2003      | Placebo     | 62             | 0            | Adalimumab high   | 73            | 1            |
| Keystone 2004       | Placebo     | 200            | 1            | Adalimumab medium | 419           | 8            |
| Van de Putte 2004   | Placebo     | 110            | 1            | Adalimumab medium | 225           | 2            |
| Van de Putte 2004   | Placebo     | 110            | 1            | Adalimumab low    | 106           | 1            |
| Van de Putte 2004   | Placebo     | 110            | 1            | Adalimumab high   | 103           | 1            |
| Bredfeld 2006       | Placebo     | 257            | 4            | Adalimumab medium | 542           | 6            |
| Maini 2004          | Placebo     | 88             | 1            | Infliximab medium | 86            | 1            |
| Maini 2004          | Placebo     | 88             | 1            | Infliximab high   | 254           | 8            |
| St Clair 2004       | Placebo     | 291            | 0            | Infliximab medium | 372           | 0            |
| St Clair 2004       | Placebo     | 291            | 0            | Infliximab high   | 377           | 4            |
sparse with respect to the number of events observed. As mentioned above, we will not investigate this aspect of sparsity, even though it is widely discussed for various situations and methods, but refer to relevant literature, both for pairwise\textsuperscript{41–44} and network meta-analysis.\textsuperscript{45,46} The complete network would consist of 28 direct treatment effect estimates out of which 12 are available (42%). The mean path length in this network is 1.57, resulting from a star-shaped network where contrasts that are not directly compared can always be reached by two steps in the path using the common comparator. With respect to important network characteristics, this empirical network of trials is therefore in line with the overview on typical network meta-analyses reported by Petropoulou et al.\textsuperscript{23} and Tonin et al.\textsuperscript{24}

In order to compare the performance of different meta-analytical approaches, we conduct a simulation study based on our empirical example data. The simulation scenarios are similar to the empirical data with respect to the number of treatments and the treatment effects. The network geometry, which was initially star-shaped, was varied such that a non–star-shaped network with one main bridge is evaluated in addition to the original star-shaped one. We additionally varied the sparsity in the network by removing all contrasts not strictly necessary to keep a connected network. This means that for a network of eight treatments, only seven contrasts are informed by direct evidence. We extended the original empirical data set to scenarios where few studies are included per contrast and to more pronounced between-trial heterogeneity. We vary the number of trials per contrast from 1 over 2 to 5 to reflect values frequently observed in practice\textsuperscript{24} and also to reflect values that are observed in the original data set.\textsuperscript{14}

To allow comparisons between the different numbers of trials within one network meta-analysis, we set the number of trials to 10 for one contrast, the contrast $\theta_{12}$ in all simulations. Contrary to the empirical data observed in Warren et al.\textsuperscript{14} we explicitly simulate data sets that are not rare with respect to the number of events.

We generated data for two- and multi-arm trials by drawing study-specific treatment effects from a multivariate normal distribution centered around the vector of theoretical treatment effects and varying with the between-study heterogeneity of the respective scenario. The covariance matrix is simply the chosen heterogeneity between-study heterogeneity of the respective scenario. $\tau^2$ at the diagonal entries and 0.5 $\tau^2$ at the nondiagonal elements. In more complicated settings, eg, when considering different missingness patterns in the network, this assumption could be relaxed. We then back-calculated the event rates by assuming an overall event risk of 0.5 and using an inverse logit transformation to simultaneously determine all event rates for one trial. We then draw from independent binomial distributions for each trial arm to incorporate within-trial sampling error. By doing so, we ensure that all treatment effects within one trial, which are calculated from these simulated events, are loop-consistent. At the same time, we avoid pre-defining a single event rate and ensure that all trial arms incorporate the variation inherent in the simulation. This data-generating mechanism is described in more detail elsewhere.\textsuperscript{47} For sake of simplicity, we set the number of patients per arm to $n_a = 100$ for all trial-arms, which is not varied for the different scenarios. Results for all scenarios setting $n_a = 10\,000$ can be found in Data S1. The different simulation scenarios are summarized in Table 2, where we additionally included network graphs for the four different combinations of geometry and sparsity we investigated in the simulation study.

We then investigate the candidate methods by using several performance measures described in Morris et al.\textsuperscript{48} In particular, we evaluate the coverage of the treatment effect estimates, the width of the respective confidence or credible intervals for the treatment effect estimate, the bias, the root mean squared error (RMSE), and the relative frequency of the treatments’ rankings. The number of repetitions is set to 2000 in all simulation scenarios. For Bayesian estimation, we used four chains, a thinning factor of 10, and a burn-in of 5000 observations.

3.1 Software

All simulations and network meta-analyses are performed in R\textsuperscript{44} and its extensions tidyverse,\textsuperscript{49} mvtnorm,\textsuperscript{50} netmeta,\textsuperscript{51} mvmeta,\textsuperscript{52} and gemtc.\textsuperscript{53} We used mvtnorm for generating the necessary data; netmeta, mvmeta, and gemtc to estimate the network meta-analyses with the respective models; and tidyverse for data handling and visualization. For the Bayesian estimation, iterative estimation using Markov-Chain-Monte-Carlo simulations is used, which is performed in jags\textsuperscript{54} by using the interface provided in gemtc.

4 RESULTS

Results are split in five different paragraphs along the evaluated measures. They are illustrated in Figures 1–5. Each of Figures 1–4 is divided into four panels (A, B, C, and D) and the legend including the respective network as in Table 2. Results for the coverage of the 95% confidence or credible interval, respectively, are presented in panel A as radar graph where the different contrasts in the network are presented at the spokes of the graph. In the best case, the coverage for all contrasts in one simulated scenario should fall around 95% with the
uncertainty determined by the Monte-Carlo error of the simulation. Using 2000 repetitions, this area is roughly ±1% (highlighted in gray). The width of the 95% confidence or credible intervals are shown in panel B, where the contrasts are presented on the x-axis, and the transparency indicates whether the estimation of the confidence or credible interval stems from direct or indirect evidence. The bias of the estimated treatment effects is shown in panel C using the same scheme for x-axis and transparency as in panel B. Negative (positive) values indicate underestimation (overestimation) when compared with the theoretical treatment effects. Results illustrating the RMSE are illustrated in panel D, again using the same scheme for x-axis and transparency as in panel B. Results are shown for all possible contrasts including both direct and indirect evidence in the network. Figure 5 illustrates the mean deviation of the SUCRA or P-score, respectively, from its theoretical value. This figure is again split into four panels, each of which illustrates one combination of geometry and sparsity of this performance measure. A bar chart with the different treatments on the x-axis is used for illustration, where again negative (positive) values indicate underestimation (overestimation) for a treatment in the network. The same colors as in Figures 1–4 are used to indicate the methods. In addition to the legend, the network graphs as in Table 2 and the theoretical ranking for each of the two geometries are given between the four panels of Figure 5. In each of the panels of all figures, 3 × 3 facets compare the different scenarios reaching from no heterogeneity to high heterogeneity (columns) and from 1 to 5 trials per contrast (rows). We additionally simulated networks where the number of patients per arm was 10 000 rather than 100 (results can be found in Data S1).

### 4.1 Coverage

All methods performed well with respect to coverage of treatment effects when between-study heterogeneity is absent or small, especially as the number of trials per contrast increases. For all approaches, coverage is higher than the nominal value in scenarios with no or low heterogeneity, and only one or two trials per contrast are available, irrespective of the geometry or sparsity in the network. However, especially for no heterogeneity and 1 trial per contrast, this overconservativeness is more pronounced in gemtc and more so when using the wider U(0,4) prior for the between-study heterogeneity. The difference between the methods decreases when including five trials per contrast.

For all methods, coverage decreases with increasing heterogeneity. For τ = 1 and only 1 trial per contrast, coverage drops below the nominal value of 95% for all methods except gemtc with the U(0,4) prior on between-study heterogeneity. When additionally increasing the number of trials per contrast, coverage of gemtc and mvmeta gets closer to the nominal level in all scenarios. However, when using netmeta, the coverage is observed to decrease even further when increasing the number of

### Table 2 Parameters for simulation scenarios

| Parameter                              | Label (If any) | Values                                      |
|----------------------------------------|----------------|---------------------------------------------|
| Number of trials per contrast          | n_t           | 1, 2, 5 (10 for \( \theta_{12} \) in all scenarios) |
| Between-study heterogeneity \( \tau \) |                | 0, 0.22, 1                                  |
| Treatment effect \( \theta_{1k} \)     |                | \( \theta_{12} = \log(0.881) = -0.1267 \)   |
|                                        |                | \( \theta_{13} = \log(10.000) = 2.3026 \)   |
|                                        |                | \( \theta_{14} = \log(2.423) = 0.8850 \)    |
|                                        |                | \( \theta_{15} = \log(2.279) = 0.8237 \)    |
|                                        |                | \( \theta_{16} = \log(2.53) = 0.8123 \)     |
|                                        |                | \( \theta_{17} = \log(4.110) = 1.4134 \)    |
|                                        |                | \( \theta_{18} = \log(1.336) = 0.2897 \)    |
| Geometry                               | Star-shaped    | Non–star-shaped                             |
| Sparsity                               | Sparse         | Nonsparse                                   |
|                                        | Sparse         | Nonsparse                                   |
trials per contrast under high heterogeneity. The decrease in coverage is observed irrespective of the geometry and sparsity of the evaluated network. In particular, it is also present in the sparse scenarios, which do not include multi-arm trials. When evaluating the estimated between-trial heterogeneity $\tau^2$ derived using netmeta, we observed that between-trial heterogeneity is underestimated. In the different scenarios simulated with a theoretical between-trial heterogeneity of $\tau^2 = 1$, we observed average $\tau^2$ values around 0.9 in all scenarios. This might be due to using the generalized method of moments estimator, whose pairwise version (the DerSimonian-Laird estimator) is known to underestimate $\tau^2$ in the pairwise setting. Coverage in scenarios with high heterogeneity and only one or two trials per contrast is slightly higher using gemtc with the HN(0.5) prior than using mvmeta, except for two trials per contrast in both sparse scenarios, where this difference is reversed.

Estimates for indirect comparisons where more than two steps in the network are necessary (e.g., $\theta_{26}$ in the networks that are not star-shaped) provide a comparable or slightly higher coverage than those from direct evidence or with only one intermediate treatment. When the sparsity in the network is reduced through additional connections, coverage increases for all methods, especially in scenarios under heterogeneity.

### 4.2 Width of intervals

Width of 95% confidence or credible intervals, respectively, is generally wider for indirect estimates than for direct estimates, irrespective of the method used. In all methods, as expected, width of intervals increases for increasing between-study heterogeneity but decreases when including more trials per contrast. This can also be observed for contrast $\theta_{12}$, for which we included 10 studies in all simulations as internal control and for which the intervals are observed to be clearly narrower in all scenarios, however, as expected the difference in width of

---

**FIGURE 1** Results of the simulation study for a star-shaped, sparse network, showing between-trial standard deviation ($\tau$) in columns and number of trials per contrast ($n_t$) in rows. All arms with $n_a = 100$ patients. Transparency indicates indirect comparisons. Treatment contrasts on the x-axis in all panels. A, Coverage of the 95% confidence or credible interval with expected coverage in gray on the y-axis. B, Width of the 95% confidence or credible interval. C, Bias. D, Root mean squared error [Colour figure can be viewed at wileyonlinelibrary.com]
intervals is most pronounced in scenarios with only one trial per contrast.

With respect to the different methods, Bayesian credible intervals are, in general, wider than the frequentist confidence intervals. Credible intervals stemming from estimations where the flatter U(0,4) prior was used for between-study heterogeneity are, as expected, always wider than those from estimation where the HN(0.5) prior was used. Confidence intervals calculated using mvmeta are narrower than those using gemtc with the HN(0.5), except for the scenarios with high heterogeneity and two trials per contrast in both sparse scenarios. There, as in the case of coverage, the relation is reversed. In the other scenarios, confidence intervals from mvmeta are comparable with or wider than those estimated from netmeta. However, all differences between methods become smaller with increasing number of trials or decreasing heterogeneity.

The geometry of the scenario did influence the number and position of indirectly estimated contrasts and whether or not contrasts, which need more than one intermediate step for estimation are present (eg, contrast \(\theta_{26}\), where both treatments 1 and 5 are needed as intermediates). Confidence and credible intervals are for all methods wider for indirect than for direct treatment effects, and even more so when an additional intermediate step was necessary due to the network geometry. The increase in width is observed to be most pronounced when using gemtc with the U(0,4) prior and is generally slightly larger under high heterogeneity and in nonsparse scenarios. While this is expected from the fact that the variances of the directly observed contrasts are summed up when estimating those of the indirect treatment effects, this may be kept in mind in empirical estimation of network meta-analysis.

4.3 | Bias

Bias is generally low in all methods over all scenarios with a range from -0.1 to 0.1. Using gemtc, bias is observed to be higher than using either frequentist
estimation method, as long as heterogeneity was absent or low. In these scenarios, bias is comparable in either of the frequentist methods and did not differ between the two prior specifications for gemtc. In scenarios with high heterogeneity and when more than one trial per contrast is available, the bias increases considerably for some contrasts using the frequentist estimation. This leads to a higher magnitude of bias for these contrasts using frequentist than using Bayesian methods. In these scenarios, gemtc using the U(0,4) prior shows a slightly higher bias than gemtc using HN(0.5) and mvmeta a slightly higher bias than netmeta. In some scenarios and contrasts, the direction of bias differed between frequentist and Bayesian methods but never within frequentist or Bayesian estimation. Following Morris et al., we evaluated the Monte Carlo standard error of the bias estimate. This Monte Carlo standard error was generally low over all considered scenarios and estimation methods with a minimum of 0.0022 and a maximum of 0.0636. Although generally low, it slightly increases with increasing heterogeneity (0.035 to 0.08 on average) and is slightly lower for mvmeta than for the other methods in star-shaped scenarios (0.001 on average).

### 4.4 Root mean squared error

The different methods all behave very similarly with respect to the RMSE, and only slight differences are observed. When increasing heterogeneity or decreasing the number of trials per contrast, the RMSE increases for all methods. When between-trial heterogeneity is high and especially when the number of trials per contrast is simultaneously low, the differences between the evaluated methods, as well as the differences between direct and indirect evidence are more pronounced, irrespective of the geometry or whether or not the network is sparse. With respect to the evaluated methods, no systematic differences were observed throughout the investigated scenarios.
The main interest for many clinical practitioners lies in the comparison of treatments with respect to one or multiple outcomes. One advantage of network meta-analysis is the fact that treatments, even those which have not been compared in a head-to-head trial, may be ranked. Various approaches to expressing the relative merit of a treatment over other ones exist with different advantages and disadvantages. A straightforward approach to ranking would be to derive the ranking per method in each iteration of the simulation study and to compare how the treatments are ranked with the known theoretical values. Another option would be to present the SUCRA for gemtc or its frequentist counterpart, the P-scores when reporting netmeta. For mvmeta, the parametric resampling approach reported in White et al28,38 is used to approximate the SUCRA value. These values are numerical presentations of the overall ranking with the advantage that solely a single number per treatment is required (instead of one number per treatment and possible ranking positions 1-8 in our case).

Depending on the geometry of the simulated network, the theoretical ranking of treatments differs. For star-shaped networks (results illustrated in Figures 1 and 2) the theoretical ranking is 3 > 7 > 4 > 5 > 6 > 8 > 1 > 2. For the non-star-shaped networks (results illustrated in Figures 3 and 4), the two best-ranked and two worst-ranked treatments stay the same, while the order in the middle is changed. This results in the following order for non-star-shaped networks 3 > 7 > 6 > 8 > 4 > 5 > 1 > 2. The theoretical SUCRA or P-score values of a treatment depend on its position in this ranking. Using perfect data, one would expect the best treatment to have a value of 1, the second best, a value of 6/7, the third best, a value of 5/7, etc down to 0 for the worst treatment.

The mean deviations of the SUCRAs or P-scores from theoretical values are illustrated in FIGURE 4 Results of the simulation study for a non-star-shaped network including additional contrasts, showing between-trial standard deviation (τ) in columns and number of trials per contrast (n) in rows. All arms with n = 100 patients. Transparency indicates indirect comparisons. Treatment contrasts on the x-axis in all panels. A, Coverage of the 95% confidence or credible interval with expected coverage in grey on the y-axis. B, Width of the 95% confidence or credible interval. C, Bias. D, Root mean squared error [Colour figure can be viewed at wileyonlinelibrary.com]
Figure 5, again using bar charts. In case of a perfect performance, we would expect this deviation to be zero for all treatments in scenarios and methods. Regarding the simulation results, the deviation increases with increasing heterogeneity and with decreasing number of trials per contrast. In general, all compared methods perform very similarly. Particularly, the direction of the deviation is similar for all treatments in all investigated scenarios. The SUCRA or P-score for the best treatment (3) is underestimated in all scenarios. This underestimation is observed to be smaller in star-shaped networks, irrespective of the estimation method, the amount of heterogeneity, and the number of trials per contrast. Similarly, the SUCRA or P-score for treatment 2, the worst treatment, is overestimated in all scenarios. However, no difference with respect network’s geometry is observed. In scenarios with high heterogeneity, especially for non–star-shaped networks, deviations for SUCRAs from gemtc using the U(0,4) prior and mvmeta tend to be slightly higher than those for P-scores and SUCRAs for gemtc using the HN(0.5) prior.

When the numbers of estimated ranks for each position are determined, two main results can be observed. The number of correct ranks is in all scenarios, and over all methods not very high (30%-50%). Furthermore, this number is decreasing when heterogeneity increases or when only one trial per contrast is available. The estimated rank, however, is in most cases not far away from the theoretical value. Regarding treatment 3, which is in fact the best, it is estimated to be the second best with a high relative frequency and is rarely estimated to be third best or even worse. Again, if heterogeneity is high or the number of trials per contrast is low, a treatment’s ranking is spread over more positions, and the frequency of estimating the right ranking is lower, which corresponds to the higher variation in the deviation of SUCRAs or P-scores in these scenarios. The geometry of the network does not seem to influence the estimation results in any method. However, contrary to the deviation of SUCRAs or P-score values from their theoretical values, additional connections and multi-arm trials in the network increase the relative frequency with which a treatment is ranked correctly.
5 | DISCUSSION AND CONCLUSIONS

Even though network meta-analysis is widely used, simulation studies comparing estimation models and methods are rare. For the practically highly relevant scenario of a network of trials comparing a binary endpoint under heterogeneity and including multi-arm trials, no simulation study is, to our best knowledge, currently available. With this work, we aim at contributing evidence here. The simulation study performed has a clear focus on sparse networks with respect to the available direct estimable contrasts and the number of trials per contrast.

Regarding the network geometry and combing the results of all performance measures used, we observe that all evaluated analysis methods perform better for direct than for indirect evidence. With respect to coverage, indirect estimates from more than one intermediate step slightly increase in coverage as compared with indirect estimates where only one step is necessary. Contrary to what we expected and contrary to the behavior of the other investigated methods, coverage decreased using netmeta when increasing the number of trials per contrasts in scenarios with high heterogeneity. When evaluating the estimated between-trial heterogeneity obtained using netmeta, it was observed that between-trial heterogeneity was underestimated in all scenarios with theoretical between-trial heterogeneity present. This might be partly due to the generalized methods of moments estimator in netmeta. Its pairwise equivalent, the DerSimonian-Laird estimator, is commonly known to underestimate between-trial heterogeneity, and different adjustments are used to correct for this underestimation in the pairwise case. Another potential reason might be the inverse variance weighting used on binary data. In the current version of the netmeta package (since version 1.0-0), the Mantel-Haenzel method is implemented for network meta-analysis under the common-effect model. For future investigations, it might be interesting to investigate whether using the Mantel-Haenzel method or an adjustment similar to those in pairwise meta-analysis might lead to an improved performance.

The width of confidence and credible intervals increases when indirect evidence is used as compared with direct evidence and increases even further when more than intermediate step is necessary. The geometry of network seems therefore relevant when considering the potential trade-off between coverage and precision, especially when heterogeneity is high and/or only few studies per contrast are present. Additionally, the presence of multi-arm trials, which increases the connectivity within the network, seems to have a high influence of the estimation results. In both geometries, coverage is closer to the nominal value in these scenarios, and width of confidence and credible intervals is on average slightly narrower. Under high heterogeneity, confidence intervals of netmeta tend to decrease more strongly in width than confidence or credible intervals of the other investigated methods when increasing the number of trials per contrast. This corresponds to the scenarios where a decrease in coverage was unexpectedly observed. For these scenarios, bias, although generally low, is also larger in some contrasts than those of the other methods.

There are several additional aspects of network meta-analytical methods that are omitted from the simulation but that are potentially of interest. First of all, our investigation is limited to three commonly used meta-analytical methods. Further, modeling approaches, such as frequentist one-stage methods, Bayesian two-stage estimation, arm-based models, or models for informative missing data structures, may also potentially be investigated in the future. We did not investigate inconsistent networks nor settings where between-trial heterogeneity is present but not constant over contrasts. Furthermore, we limited our evaluation to the case of binary data and to the OR as an effect measure. Although concentrating on binary endpoints, we did not investigate the third aspect of sparsity, rare events, that is relevant for this data type. Different endpoints that require using a different likelihood in the Bayesian one-step method and a different transformation to obtain a treatment effect in the frequentist two-step methods are potentially of interest. Even though we would not expect fundamentally different results with respect to the performance of the evaluated methods in other endpoints, we would expect that the network's sparsity and geometry, as well as the presence of heterogeneity, would have a high influence in all types of endpoints. Furthermore, we also did not investigate scenarios where effect measure, link function, or likelihood are miss-specified.

The Bayesian estimation using the flatter U(0,4) prior on between-trial heterogeneity results in coverage above the nominal value in most simulation settings. Accordingly, credible intervals estimated using this approach are the widest investigated, irrespective of the considered scenario. Coverage estimated using gemtc or mvmeta was overconservative in scenarios with no heterogeneity and only one trial per contrasts but performed good in the other investigated scenarios. Except for the sparse scenarios with high heterogeneity and only one or two treatments per contrast, coverage for gemtc using the HN(0.5) prior was always slightly higher than those for mvmeta. We observe that coverage was below the nominal value in netmeta and considerably lower than those of the other approaches in scenarios with high heterogeneity in both network geometries and irrespective of whether networks were sparse or not. While netmeta
has smaller intervals than both gemtc approaches and mvmeta, this partly seems to come at the cost of coverage, especially in scenarios with high heterogeneity. Furthermore, when heterogeneity is present, coverage decreases when the number of trials per contrast increases. Potentially, confidence intervals that are based on the normal distribution and that do not take into account the additional uncertainty introduced by estimating between-trial heterogeneity could lead to confidence intervals, which are too narrow in scenarios of high heterogeneity using a moderate number of trials per contrast. We did investigate neither into the behavior when using quantiles from another distribution nor into scenarios with a larger number of trials per contrast. In most scenarios, credible intervals of gemtc using the HN(0.5) prior for between-trial heterogeneity are slightly wider than the credible intervals estimated using mvmeta. As for the coverage, this relation is reversed for sparse scenarios with high heterogeneity and two trials per contrast. While bias was slightly higher using either gemtc approach in scenarios with absent or low heterogeneity, netmeta and mvmeta showed a considerably higher bias in some contrasts under scenarios with high heterogeneity, irrespective of the geometry or sparsity of the network. Deviations of the SUCRA or P-score values are comparable over the methods in most scenarios. However, considering scenarios with high heterogeneity, the average deviation is in some contrasts slightly higher when using mvmeta. This might be due to the approximation of the SUCRA value by the parametric resampling used and might improve when simulating a larger amount of samples. Gemtc using the half-normal prior for the between-study heterogeneity delivered, with respect to all evaluated measures, results in-between the Bayesian approach using the U(0,4) prior and the frequentist methods. In the investigated setting and with respect to the considered performance measures, the HN(0.5) seems to be preferable when compared with the wider U(0,4) prior. While mvmeta performed slightly worse with respect to coverage and slightly better with respect to width of intervals than gemtc with HN(0.5), none of these two candidate methods is clearly preferable with respect to bias and RMSE. Regarding the deviation from the theoretical SUCRA, both methods are comparable in most scenarios with absent or low heterogeneity. However, some treatments in the scenarios of high heterogeneity show a higher deviation for mvmeta than for gemtc using HN(0.5). This prior specification, which has been recommended for use in pairwise binary random-effects meta-analysis with few studies, also offers a promising compromise in our simulation when using a Bayesian framework, while mvmeta is a good choice when considering frequentist estimation. Both methods perform well with respect to coverage, precision, bias, RMSE, and ranking, each with slight nuances with respect to the trade-off between the different evaluated performance measures.

ACKNOWLEDGEMENTS
The authors would like to thank two reviewers who committed their time and efforts to significantly improving and clarifying this manuscript by providing very helpful comments and questions.

CONFLICT OF INTEREST
The authors declare to have no conflict of interest.

DATA AVAILABILITY STATEMENT
All data generated or analyzed during this study are included in this article (and its supplementary information files) by providing necessary, previously unpublished R Code to generate and analyse the synthetic data.

ORCID
Svenja E. Seide https://orcid.org/0000-0002-9113-7373
Katrin Jensen https://orcid.org/0000-0002-5088-5279
Meinhard Kieser https://orcid.org/0000-0003-2402-4333

REFERENCES
1. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002;21(16):2313-2324.
2. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc. 2006;101(474):447-459.
3. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods. 2012;3(2):80-97.
4. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004;23(20):3105-3124.
5. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. Syst Rev. 2017 Apr;6(1):79.
6. Veroniki AA, Straus SE, Rücker G, Tricco AC. Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? J Clin Epidemiol. 2018 August;100:122-129.
7. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ. 2011;343:d4909.
8. Lee AW. Review of mixed treatment comparisons in published systematic reviews shows marked increase since 2009. J Clin Epidemiol. 2014;67(2):138-143.
9. Greco T, Biondi-Zoccai G, Saleh O, et al. The attractiveness of network meta-analysis: a comprehensive systematic and narrative review. Heart Lung Vessel. 2015;2(7):133-142.
10. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network Meta-Analysis for Decision-Making. Statistics in Practice, Chichester, UK: John Wiley & Sons; 2018.
11. Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. Clin Epidemiol. 2014;6:451-460.
12. Zhang J, Carlin BP, Neaton JD, et al. Network meta-analysis of randomized clinical trials: reporting the proper summaries. Clin Trials. 2013;11(2):246-262.
13. Franchini AJ, Dias S, Ades AE, Jansen JP, Welton NJ. Accounting for correlation in network meta-analysis with multi-arm trials. Res Synth Methods. 2012;3(2):142-160.
14. Warren FC, Abrams KR, Sutton AJ. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. Stat Med. 2014;33(14):2449-2466.
15. Owen RK, Tincello DG, Keith RA. Network meta-analysis: development of a three-level hierarchical modeling approach incorporating dose-related constraints. Value Health. 2015;18 (1):116-126.
16. Günhan BK, Freide T, Held L. A design-by-treatment interaction model for network meta-analysis and meta-regression with integrated nested Laplace approximations. Res Synth Methods. 2018;9(2):179-194.
17. Rücker G, Cates CJ, Schwarzer G. Methods for including information from multi-arm trials in pairwise meta-analysis. Res Synth Methods. 2017;8(4):392-403.
18. Song F, Clark A, Bachmann MO, Maas J. Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. BMC Med Res Methodol. 2012;12:138.
19. Mills EI, Ghenemt I, O’Regan C, Thorlund K. Estimating the power of indirect comparisons: a simulation study. PLoS One. 2011;6(1):e16237.
20. Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G. Evaluation of inconsistency in networks of interventions. Int J Epidemiol 2013 03;42(1):332–345.
21. Nikolakopoulou A, Chaimani A, Veroniki AA, Vasiliadis HS, Schmid CH, Salanti G. Characteristics of networks of interventions: a description of a database of 186 published networks. PLoS One 2014 01;9(1):1–10.
22. Pinart M, Kranz J, Jensen K, et al. Optimal dosage and duration of pivmecillinam treatment for uncomplicated lower urinary tract infections: a systematic review and meta-analysis. Int J Infect Dis. 2017;58:96-109.
23. Petropoulou M, Nikolakopoulou A, Veroniki AA, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. J Clin Epidemiol. 2017;82:20-28.
24. Tonin FS, Borba HH, Mendes AM, Wiens A, Fernandez-Llimos F, Pontarolo R. Description of network meta-analysis geometry: a metrics design study. PLoS One 2019 02;14(2):1–14.
25. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res Synth Methods. 2019;10(1):83-98.
26. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010;29(7–8):932-944.
27. Carroll K, Hemings R. On the need for increased rigour and care in the conduct and interpretation of network meta-analyses in drug development. Pharm Stat. 2016;15(2):135-142.
28. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Methods. 2012;3 (2):111-125.
29. Jackson D, Barrett JK, Rice S, White IR, Higgins JPT. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. Stat Med. 2014;33 (21):3639-3654.
30. van Valkenhoef G, Dias S, Ades AE, Welton NJ. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Res Synth Methods. 2016;7(1): 80-93.
31. Yu-Kang T. Node-splitting generalized linear mixed models for evaluation of inconsistency in network meta-analysis. Value Health. 2016;19(8):957-963.
32. 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.
33. Gagnier J. Appraising Between-Study Heterogeneity. In: Biondi-Zoccai G, ed. Network Meta-Analysis: Evidence Synthesis with Mixed Treatment Comparison Hauppauge. NY: Nova Science Publishers; 2014:171-190.
34. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2018, https://www.R-project.org/.
35. Rücker G. Network meta-analysis, electrical networks and graph theory. Res Synth Methods. 2012;3(4):312-324.
36. Jackson D, White IR, Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. Stat Med. 2012;31(29):3805-3820.
37. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Med Res Methodol. 2015;15(1):58.
38. White IR. Multivariate random-effects meta-regression: updates to mvmeta. Statata Journal. 2011;11(2):255-270(16).
39. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol. 2011;64(2):163-171.
40. Freide T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. Res Synth Methods. 2017;8(1):79-91.
41. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med. 2004;23(9):1351-1375.
42. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med. 2010;29(29):3046-3067.
43. Kuss O. Statistical methods for meta-analyses including information from studies without any events—add nothing to nothing and succeed nevertheless. Stat Med. 2015;34(7):1097-1116.
44. Efthimiou O. Practical guide to the meta-analysis of rare events. Evid Based Ment Health. 2018;21(2):72-76.
45. Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. Biometrical Journal Biometrische Zeitschrift. 2015;57(4):633-648.
46. Efthimiou O, Rücker G, Schwarzer G, Higgins JPT, Egger M, Salanti G. Network meta-analysis of rare events using the mantel-Haenszel method. Stat Med. 2019;38(16):2992-3012.
47. Seide SE, Jensen K, Kieser M. Simulation and data-generation for random-effects network meta-analysis of binary outcome. Stat Med. 2019;38(17):3288-3303.
48. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-2102.
49. Wickham H. tidyverse: Easily Install and Load the 'Tidyverse'; 2017, https://CRAN.R-project.org/package=tidyverse, r package version 1.2.1.
50. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, et al. mvtnorm: Multivariate Normal and t Distributions; 2018, https://CRAN.R-project.org/package=mvtnorm, r package version 1.0–8.
51. Rücker G, Schwarzer G, Krahn U, König J. netmeta: Network Meta-Analysis using Frequentist Methods; 2018, https://CRAN.R-project.org/package=netmeta, r package version 0.9–8.
52. Gasparini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. Stat Med. 2012;31(29):3821-3839.
53. van Valkenhoef G, Kuiper J. gemtc: Network Meta-Analysis Using Bayesian Methods; 2016, https://CRAN.R-project.org/package=gemtc, r package version 0.8–2.
54. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing; 2003. p. 1–8.
55. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-188.
56. Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud C. Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses—an empirical assessment of 920 Cochrane primary outcome meta-analyses. Res Synth Methods. 2011;2(4):238-253.
57. Jackson Daniel VAAVW. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods. 2016;7(1):55-79.
58. Rücker G, Krahn U, König J, Efthimiou O, Schwarzer G. netmeta: Network Meta-Analysis using Frequentist Methods; 2019, https://CRAN.R-project.org/package=netmeta, r package version 1.1–0.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Seide SE, Jensen K, Kieser M. A comparison of Bayesian and frequentist methods in random-effects network meta-analysis of binary data. Res Syn Meth. 2020;11:363–378. https://doi.org/10.1002/jrsm.1397