Bayesian model-based outlier detection in network meta-analysis

Silvia Metelli1* | Dimitris Mavridis2† | Perrine Créquit1,3‡ | Anna Chaimani1

1Inserm Research Center of Epidemiology and Statistics, Université Paris Cité, France
2Department of Primary Education, University of Ioannina, Greece
3Direction de la recherche Clinique, Hôpital Foch, Suresnes, France

Correspondence
Silvia Metelli PhD, Inserm Research Center of Epidemiology and Statistics, Université Paris Cité, France
Email: silvia.metelli@u-paris.fr

Present address
†Inserm Research Center of Epidemiology and Statistics, Université Paris Cité, Paris, 75004, France

Funding information
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101031840.

In network meta-analysis, some of the collected studies may deviate markedly from the others, for example having very unusual effect sizes. These deviating studies can be regarded as outlying with respect to the rest of the network and can be influential on the pooled results. Thus, it could be inappropriate to synthesise those studies without further investigation. In this paper, we propose two Bayesian methods to detect outliers in a network meta-analysis via: (a) a mean-shifted outlier model and (b), posterior predictive $p$-values constructed from ad-hoc discrepancy measures. The former method uses Bayes factors to formally test each study against outliers while the latter provides a score of outlyingness for each study in the network, allowing to numerically quantify the uncertainty associated with being outlier. Furthermore, we present a simple method based on informative priors as part of the network meta-analysis model to down-weight the detected outliers. We conduct extensive simulations to evaluate the effectiveness of the proposed methodology while comparing it to some alternative outlier detection tools. Two case studies are then used to demonstrate our methods in practice.

KEYWORDS
outlying studies, indirect treatment effects, Bayes factors, posterior predictive checking, down-weighting
1 | INTRODUCTION

In medical statistics, meta-analyses and network meta-analyses (NMAs) (Lumley, 2002; Lu and Ades, 2004) have become crucial tools to quantitatively pool results from independent studies and assess treatment efficacy and cost-effectiveness. In pairwise meta-analysis, only two treatments at the time can be compared, while network meta-analysis allows for the simultaneous comparison of multiple (≥ 3) treatments, forming a so-called network of treatments. By integrating into a single model direct and indirect evidence across trials, network meta-analysis has the potential to provide a more precise, global estimate of the relative effect of any pair of treatments included in the network. To avoid misleading conclusions and provide valuable information for clinical decisions, the network needs to be carefully screened looking for studies with markedly different or extreme effect sizes, namely outlying studies. Outliers may occur for many different reasons, including very small sample sizes or study-specific effect sizes whose distribution depart from the conventional normal curve (e.g. heavy-tailed or skewed distribution of the effect sizes). Such studies can substantially influence and alter the conclusions of the analysis and need proper investigation.

Whilst many different issues of network meta-analysis methodology, such as inconsistency and heterogeneity, have received large attention in the literature, outlying studies - although intrinsically related to the presence of inconsistency and heterogeneity in the network - have not been widely studied. To date, no specific guidelines exist for how these studies should be treated in the general context of evidence synthesis. Several outlier detection methods have been recently developed for pairwise meta-analysis (Viechtbauer and Cheun, 2010; Gumedze and Jackson, 2011; Zhao et al., 2017; Mavridis et al., 2017) but little work has been done to extend the methods to network meta-analysis. Moreover, most of the available techniques are based on useful yet heuristic diagnostics measures such as studentised residuals or the Cook’s distance while only a few rely on probabilistic model-based approaches. Among these, a frequentist ‘variance shift’ outlier model has been proposed for univariate meta-analysis (Gumedze and Jackson, 2011) while two ‘mean-shift’ models have been later developed for a bivariate model for diagnostic test accuracy (DTA) meta-analyses and subsequently for a full multivariate model for network meta-analysis (Negeri and Beyene, 2020; Noma et al., 2020). In both cases, the methodology made use of a frequentist likelihood ratio test (LRT) as a test statistic for assessing whether each included study was outlying and the parametric bootstrap approach to approximate the sampling distribution of the observed LRT statistic. Bayesian approaches are attractive in network meta-analysis (Dias et al., 2018) and have the advantage of using the exact likelihood for the data (i.e. binomial for binary data) rather than relying on normal approximations. However, outlier detection in the Bayesian framework has not been sufficiently explored, with exception of one method for pairwise meta-analysis of DTAs (Matsushima et al., 2020) and one introducing a Bayesian p-value for network meta-analysis but mainly focusing on arm-based models for continuous outcomes (Zhang et al., 2015).

A comprehensive assessment against outlyingness should not merely focus on the statistical detection of extreme effect measures (or variances) of the studies included; rather, it should try to understand the causes behind it through a careful appraisal of the characteristics of each included study. A related question then arises about how these studies should be treated while ensuring that the validity and robustness of the synthesis process is maintained. The debate was initially centered around whether or not outliers should be removed from the analysis (Hedges and Olkin, 1985). Conducting sensitivity analyses with and without outliers to monitor the changes in the summary effects is surely useful, but clinicians might still not reach consensus about which scenario should be used for their final clinical decisions. Therefore, more tailored strategies for treating outliers are necessary. For example, methods have been proposed in pairwise meta-analysis for building heterogeneity measures which are minimally
affected by the presence of outliers (Lin and Hodges, 2017) or down-weighting the apparent outlying studies without removing them (Gumedze and Jackson, 2011). In network meta-analysis, this also ensures that the connectivity of the network is maintained.

In this paper, we suggest to employ a two-step procedure: first, a probabilistic outlier detection model is used to quantify outlying behaviour and then, the studies associated with high probability of being outliers can undergo down-weighting, if appropriate. As a first step towards this, we propose an intuitive Bayesian mean-shift model that detects deviating studies within the network using Bayes factors; then we seek to complement Bayes factor detection with Bayesian model checking, which allows to better quantify the associated uncertainty for each study to be outlier. Specifically, we propose the use of posterior predictive $p$-values under ad-hoc discrepancy measures, which are well-suited to capture local deviations in the model. As a second step, informative beta priors are conveniently incorporated into the network meta-analysis model to down-weight the outliers identified. The performance of our methods is assessed and compared using both simulated and real data.

The rest of the paper is structured as follows. Section 2 describes two examples of real networks of treatments while in Section 3 we briefly introduce the most commonly used random effects network meta-analysis model. In Section 4, we describe our proposed approaches: first, a mean-shifted model with Bayes factors and then, posterior predictive checks with ad-hoc discrepancy measures; while the down-weighting scheme is described in Section 5. In Section 6, we perform an extensive simulation study and in Section 7 we present an application to the two real networks previously introduced. Finally, we conclude with a discussion in Section 8.

2 | EXEMPLAR DATA

We introduce two real data sets, each forming a network of treatments, which we later use to demonstrate our methods. The first example is a network of treatments for non-small cell lung cancer (NSCLC) and the second is a smaller network of non-pharmacological interventions for smoking cessation. Non-small cell lung cancer represents approximately 85% of all lung cancer cases, and most patients have wild-type or unknown status for epidermal growth factor receptor (EGFR) which often leads to a diagnosis of advanced-stage disease. According to specific eligibility criteria, patients with advanced-stage diagnosis might receive second-line treatments instead of palliative care. Despite the American Society of Clinical Oncology recommends two cytotoxic drugs and two EGFR-tyrosine kinase inhibitors (Master et al., 2015), many new treatments have been recently approved by the US Food and Drug Administration (FDA) and over the years, more than forty treatments have been assessed in randomised trials for second-line treatment of advanced NSCLC (Créquit et al., 2016). Clearly, simultaneously comparing the relative efficacy and safety of all available treatments in a network can better assist clinical decision-making. Créquit et al. 2017 conducted an extensive systematic review and NMA for second-line treatments of advanced NSCLC (involving a total of 39,388 patients), forming a network of $N$ = 112 randomised controlled trials (RCTs) comparing 62 different treatments, many of which informed by one or very few studies only. This makes it a good candidate to suspect the presence of trials with outlying results. As a second example, we use a well-known network of $N$ = 24 RCTs investigating four different counselling programs to aid smoking cessation (involving a total of 16,737 participants, Hasselblad 1998, where the four counselling interventions are defined as self-help, individual counselling, group counselling, and no contact. This network is mainly used in this article for comparison purposes, as it has been recently tested for the presence of outliers in a network meta-analysis application (Petropoulou et al., 2021).
In Figure 1, we show the network geometries for both data sets. For NSCLC data, each treatment was further grouped to one out of five treatment classes (Dual Targeted Therapy, Chemotherapy plus Targeted Therapy, Immunotherapy, Monochemotherapy and Targeted Therapy) to facilitate the visualisation of the available evidence. The full NSCLC network at treatment level can be found in the Supplementary material.

3 \ NETWORK META-ANALYSIS RANDOM EFFECTS MODEL

3.1 \ Basic elements and notation

Network meta-analyses expand the scope of more conventional pairwise meta-analyses to simultaneously compare multiple treatments in a connected network of evidence, where the information of the relative treatment effects (e.g. log odds ratios) is pooled across multiple studies (Lu and Ades, 2006). More specifically, consider a collection of studies $i = 1, \ldots, N$, where each study $i$ only compares a subset $K_i$ of the full set of $\{1, \ldots, K\}$ treatments. Let $k_i = |K_i|$ be the cardinality of $K_i$, then in most NMAs, $k_i$ is 2 (“two-arm study”) or 3 and rarely we have studies with four or more arms (“multi-arm studies”). In the following, we focus on NMAs with a binary outcome (e.g. death, no death), so for each study $i$ we have data $D = \{(r_{ik}, n_{ik}) : i = 1, \ldots, N; k \in K_i\}$, where $r_{ik}$ is the number of observed events and $n_{ik}$ the total number of participants for the $k^{th}$ treatment in the $i^{th}$ study. The corresponding probability of the event will be denoted by $\pi_{ik}$.

3.2 \ Standard model
In each study $i$, let one treatment be seen as the baseline treatment, $b_i$ (simply denoted as $b$ in the following for convenience). Without loss of generality, the baseline treatment can be considered a reference (e.g. placebo) against which each other treatment $k \in \{1, \ldots, K\}/\{b\}$ is compared. Then, the commonly used random-effects network meta-analysis for the binomial data can be written as

$$r_{ik} \sim \text{Binomial} (n_{ik}, \pi_{ik}), \quad i = 1, \ldots, N, k \in K_i$$

$$\logit (\pi_{ik}) = \mu_i + \theta_{bk} + \delta_{i,bk}, \quad k \neq b,$$

where $\mu_i$ represents the log odds of the baseline treatment $b$ in each study $i$, since for $k = b$ the logit expression in (1) simply reduces to $\logit (\pi_{ik}) = \mu_i$. This parameter is generally considered a nuisance while the main interest lies in the mean relative effect $\theta_{bk}$. Likewise, for continuous outcome data or log odds and risk ratios we can formulate the same network meta-analysis model using a normal likelihood with the identity link function instead of the logit one. To be identifiable, the model requires an arbitrary reference treatment whose effect is set to zero. Here, we choose reference $b = 1$ so that $\theta = (\theta_{12}, \theta_{13}, \ldots, \theta_{1K})^T$ is a vector of treatment effects relative to the reference treatment, which are called the basic parameters. Then, assuming statistical consistency, i.e. agreement between direct and indirect evidence, we have $\theta_{hk} = \theta_{1k} - \theta_{1h}$ for every treatment pairs $(h, k) \in \{1, \ldots, K\}$. In words, when we have both direct and indirect evidence for a particular comparison, then consistency holds in the data if no discrepancy exists in the treatment effects obtained under both types of evidence. All other relative effects can be obtained as linear combinations of the basic parameters in $\theta$.

Study-specific heterogeneity is captured by the random effects $\delta_{i,bk}$, which represent the relative effects between treatment $k$ and $b$ for the $i^{th}$ study. We assume exchangeability of the $\delta_{i,bk}$ so that the NMA model provides estimates for the $\theta_{bk}$'s, and the between-study heterogeneity variance of the random effects $\tau_{bk}^2$. The specific distributional assumptions made on $\delta_{i,bk}$ are discussed separately below. Suppose $\delta_i = (\delta_{i,12}, \delta_{i,13}, \ldots, \delta_{i,1K})^T$ is the vector of study-specific relative effects of treatment $k$ versus $b = 1$. Then, $\delta_i \in \mathbb{R}^{k_i-1}$ is assumed to follow a multivariate normal distribution,

$$\delta_i \sim N \left(0, \Psi_i^2 \right).$$

Following Higgins and Whitehead (1996) and Lumley (2002), we assume throughout the paper a common heterogeneity, i.e. $\tau_{bk} = \tau$, for all comparisons and the $(k_i - 1) \times (k_i - 1)$ matrix $\Psi_i^2$ to be homogeneous and symmetric with $\tau^2$ elements on the diagonal representing treatment-specific variances and $\tau^2/2$ elements off-diagonal, representing between-study covariance. This, along with the consistency equation, ensures that in each study $i$, with treatment pair $(h, k)$, $\text{Var} (\delta_{i,hk}) = \text{Var} (\delta_{i,1h}) + \text{Var} (\delta_{i,1k}) - 2 \text{Cov} (\delta_{i,1h}, \delta_{i,1k}) \Leftrightarrow \tau^2 = 2\tau^2 - 2\text{Cov} (\delta_{i,1h}, \delta_{i,1k}) \Leftrightarrow \text{Cov} (\delta_{i,1h}, \delta_{i,1k}) = \tau^2/2$.

Finally, the observed data $D$ are described by the following likelihood function:

$$P \left(D \mid \mu, \theta, \tau^2 \right) = \prod_{i=1}^N \prod_{k \in K_i} \left( \frac{n_{ik}}{r_{ik}} \right) \left[ \logit^{-1} (\pi_{ik}) \right]^{r_{ik}} \left[ 1 - \logit^{-1} (\pi_{ik}) \right]^{n_{ik} - r_{ik}}$$

where $\mu = (\mu_1, \ldots, \mu_N)$ is the vector of study-specific baseline parameters (intercepts), while $\theta$ and $\tau^2$ are the parameters of primary interest. To estimate parameters, NMA models are often rely on maximum likelihood, but the hierarchical structure of random-effect components typically requires
numerical optimisation methods or restricted maximum likelihood (REML) techniques. Here, we take a Bayesian approach, and so parameters of interest are assigned independent prior distributions, $P(\mu), P(\theta)$ and $P(\tau^2)$. Posterior inference is then conducted on the joint posterior distribution of parameters,

$$P(\mu, \theta, \tau^2 | D) \propto P(D | \mu, \theta, \tau^2) P(\mu) P(\theta) P(\tau^2).$$  \hspace{1cm} (4)

As exact inference on this posterior distribution is not analytically tractable, Markov chain Monte Carlo (MCMC) simulation is used to perform posterior inference.

4 \hspace{1cm} OUTLIER DETECTION

4.1 \hspace{1cm} Mean-shift model with Bayes factor tests

We define outliers in a network meta-analysis as studies with ‘shifted’ effect sizes and we propose a mean-shifted model to identify such studies. This model assumes ‘shifted’ location parameters for the effect sizes of study $i$, meaning that the underlying relative effect in the $i$-th study may diverge from those of the other studies. In practice, this means that model 1 is replaced by

$$r_{ik} \sim \text{Binomial} \left( n_{ik}, \pi_{ik} \right), \hspace{1cm} i = 1, \ldots, N, k \in K_i$$

$$\logit(\tilde{\pi}_{ik}) = \mu_i + \theta_{bk} + \eta_{bk} + \delta_{i,bk}, \hspace{1cm} k \neq b.$$  \hspace{1cm} (5)

In this case, the likelihood function contains an additional parameter vector:

$$P(D | \mu, \theta, \eta, \tau^2) = \prod_{i=1}^{N} \prod_{k \in K_i} \left( \binom{n_{ik}}{r_{ik}} \right) \left[ \logit^{-1}(\tilde{\pi}_{ik}) \right]^{r_{ik}} \left[ 1 - \logit^{-1}(\tilde{\pi}_{ik}) \right]^{n_{ik}-r_{ik}}$$  \hspace{1cm} (6)

where $\eta = (\eta_{12}, \eta_{13}, \ldots, \eta_{1K})^T$ is a vector of mean-shift location parameters, i.e. the grand mean parameters of study $i$ may deviate from the grand mean of the other studies, and this implicitly means that we are assuming a mean-shifted model for the random-effect of each study $i$. Then, the outlier detection problem can be cast as follows: if the mean-shifted model with non-zero shift factors is more plausible than the ordinary network meta-analysis model, then the $i$-th study can be seen a potential outlier. This corresponds to testing the following hypothesis for each study $i$:

$$H_0 : \eta_{bk} = 0 \hspace{1cm} \text{vs.} \hspace{1cm} H_1 : \eta_{bk} \neq 0, \hspace{1cm} b \neq k, \forall \hspace{0.2cm} k \in \{1, \ldots, K\} / \{1\}$$  \hspace{1cm} (7)

In a Bayesian hypothesis testing context, the test above can be formally assessed through Bayes factors (BFs). Suppose model 0 ($H_0$) is the standard model and model 1 ($H_1$) is the mean-shift outlier model, then the Bayes factor would take the following form:

$$BF_{1:0} = \frac{P(D | H_1)}{P(D | H_0)} = \frac{\int_{\theta_1} P(D | \theta_1) P_1(\theta_1) d\theta_1}{\int_{\theta_0} P(D | \theta_0) P_0(\theta_0) d\theta_0},$$  \hspace{1cm} (8)

with $\theta_1 = (\mu, \theta, \tau^2) \in \Theta_1$, $\theta_0 = (\mu, \theta, \tau^2) \in \Theta_0$, $P_0(\theta_0)$ and $P_1(\theta_1)$ being the prior distributions
for the parameters of interest. The Bayes factor can be interpreted as an updating factor of prior beliefs, and represents how likely the data were predicted by $H_1$ compared to $H_0$. This also provides a fair comparison between two models of different parameter dimension, since the Bayesian paradigm embodies a natural penalty against overfitting, i.e. the Occam’s razor principle.

This model can be seen as the Bayesian counterpart of the location-shift model introduced by Noma et al. [2020] where potential outliers were searched via bootstrap-adjusted Likelihood Ratio tests. Our model has the additional flexibility provided by prior information, which is crucial in NMA outlier-detection as we often encounter small or sparse networks informed by a few studies only. Furthermore, being ratios of probabilities, Bayes factors also give an indication about the size of the evidence. Indeed, they represent the relative probability assigned to the observed data under each of the two hypotheses, and so they not only provide evidence in favour of outlyingness, $H_1$, as the classical hypothesis testing, but also in favour of $H_0$.

### 4.2 Posterior predictive model checking

An alternative possibility for Bayesian model-based outlier detection is posterior predictive checking (Meng, 1994; Gelman et al., 1996), which is a commonly used tool for the identification of divergent observations of Bayesian models. The idea is to construct a discrepancy measure which captures deviation between the observed data and the posterior predictive distribution of the assumed model, which is for us the standard random effects network meta-analysis model. First, we take the posterior predictive distribution by simulating replicated data from the fitted model. Then we compare the replicated to the observed data to look for systematic discrepancies that will show us whether the observed data could have been plausible under the hypothesised model. The discrepancy measure $f$ is often taken to be the omnibus $\chi^2$ measure proposed by Gelman et al. 1996. This approach was followed in Zhang et al. 2015 to construct a ‘Bayesian $p$-value’.

However the method is primarily built for arm-based NMA models and continuous data where absolute treatment effects are assumed exchangeable while here we focus on contrast-based modelling where exchangeability is assumed on relative treatment effects, as introduced in Section 3. In simulations (see Supplementary material), we found that Gelman’s discrepancy performs poorly in the present context. Omnibus discrepancy measures are useful but provide less power with respect to measures designed to test specific features of the data (e.g. extremeness), suggesting the need for a discrepancy measure more capable to detect local deviations in the model. Thus, we propose two different choices for $f$: first, we make use of the single log-likelihood contribution of each study $i$ and then we leverage the Stahel-Donoho outlyingness (SDO) measure (Stahel, 1981; Donoho, 1982) to construct an ‘outlyingness score’. For each study $i$ with arm data $D_{i,k} = (r_{ik}, n_{i,k})$, the two discrepancy measures are respectively given by

$$ f^L_i = \sum_{k \in K_i} \log P(D_{i,k} | \theta_0), $$

$$ f^{SDO}_i = \sum_{k \in K_i} \frac{|x_{i,k} - \text{med}(x)|}{\text{MAD}(x)}, $$

where $x_{ik} = r_{ik}/n_{ik}$, $\text{med}(x)$ is the median and $\text{MAD}(x) = \text{med}_i(|x_i - \text{med}_j(x_j)|)$ is the median absolute deviation of the observed proportions $x_{ik} : i = 1, \ldots, N; k \in K_i$. The first proposal is somewhat related to the omnibus $\chi^2$ measure but captures different aspects of the relationship between data structure and the parameters and avoids producing extremely small values in presence of studies
with small variances, while the second is specifically aimed at detecting asymmetry in the data. Note that the first measure depends both on data and model parameters while the second depends on the data only.

The values of the discrepancy measure for the observed data are compared to values of the posterior predictive distribution: large differences indicate lack of fit. Specifically, we use posterior predictive $p$-values, which calculate a tail-area probability given that the assumed model is true, and so quantify the extremeness of the observed value rather than offering a strict accept-reject decision rule as in standard hypothesis testing. An extreme $p$-value implies that the observed data would be unlikely to occur in replications of the data if the model was true and so, may represent an outlier. Here, posterior predictive $p$-values quantify the uncertainty associated with each study in the network by measuring departure of each study from the assumed model. For each study $i$ let $D_i = \{(r_{i,k}, n_{i,k}) : k \in K_i\}$, then the posterior predictive $p$-value is as follows:

$$p_{f_i} \equiv P\{f_i(D^*_i|\theta_0) \geq f_i(D_i|\theta_0)|D\} = \int P\{f_i(D^*_i|\theta_0) \geq f_i(D_i|\theta_0)|\theta_0\} P(\theta_0|D) \, d\theta_0,$$

where $D$ is the observed data, $D^*$ a hypothetical replicated data set generated from the model predictive distribution, and $P\{\cdot|D\}$ the joint posterior distribution of $(\theta_0, D^*)$ given $D$. This can be easily estimated from the MCMC samples as

$$p_{f_i} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\{f_i(D^*_i|\theta_0(s)) \geq f_i(D_i|\theta_0(s))\}$$

where $S$ is the number of MCMC simulations and $\theta_0(s)$ the simulated parameter values at step $s$. Plugging-in the discrepancies of (9) and (10) into the $p$-value in (11) we obtain our proposed posterior predictive $p$-values under the two different discrepancies, which we respectively denote $p_{\text{L}}$ and $p_{\text{SDO}}$.

5 | DOWN-WEIGHTING OUTLIERS

Statistical detection of outlying effects in a network meta-analysis should always be complemented by an accurate investigation of the causes underlying the observed outlyingness. Before taking any decision, investigators should carefully check the characteristics of all included studies looking for possible explanations, such as systematic differences that modify the observed effect and produce extreme results. In particular, characteristics of the trial design, conduct, participants, interventions and outcomes should be explicitly assessed. Placing more stringent inclusion criteria in the systematic review may not always capture differences when they are subtle, and thus, a thorough assessment of the nature and reliability of the data is always necessary.

When no clear causes are identified, it is possible to construct systems to down-weight the effect of outlying studies towards the overall network estimates, which seems a more reasonable choice compared to removing outlying studies tout-court from the analysis. Indeed, the latter approach comes at the risk of disconnecting the network graph and this would prevent the whole NMA analysis. The risk is particularly high for sparse networks, where some comparisons might be informed by only one study.
We propose a computationally simple scheme that consists of two-stages: first, we screen the studies looking for outliers using the two methods described in the previous sections, i.e. for each study $i = 1, \ldots, N$ we calculate Bayes Factors and posterior predictive $p$-values. If a study is associated with either a Bayes Factor above the chosen outlying threshold and/or a posterior predictive $p$-value below the significance threshold, further investigation is conducted. Then, if down-weighting is deemed appropriate, a second stage of analysis is performed where informative power priors (Ibrahim and Chen, 2003) are used to automatically raise the likelihood of each outlying study $j$ to a power strictly between 0 and 1, to reduce its impact on the overall results. Here, that power represents a down-weighting factor $w_j \in (0, 1)$. At the second stage, the joint posterior in (4) is modified to

$$P(\mu, \theta, \tau^2 | \mathcal{D}) \propto P(D^o | \mu, \theta, \tau^2) P(D^o | \mu, \theta, \tau^2) P(\mu)P(\theta)P(\tau^2),$$

where $\mathcal{D}^o = \{(r_{jk}, n_{jk}) : j = 1, \ldots, N_o ; k \in K_j \}$ is the sub-set containing data for outlying studies, with size $N_o$. Analogously, we can define $\mathcal{D}^o = \mathcal{D} \setminus \mathcal{D}^o$ as the set of data for the remaining $N - N_o$ non-outlying studies. In expression (13), $P(D^o | \mu, \theta, \tau^2)$ is defined as in (3) with the only difference of using the restricted set of data $\mathcal{D}^o$ while

$$P(D^o | \mu, \theta, \tau^2) = \prod_{j=1}^{N_o} \prod_{k \in K_j} \left[ \left( \frac{n_{jk}}{r_{jk}} \right) \left[ \logit^{-1}(\pi_{jk}) \right]^{r_{jk}} \left[ 1 - \logit^{-1}(\pi_{jk}) \right]^{n_{jk} - r_{jk}} \right]^{w_j},$$

where $\logit(\pi_{jk}) = \mu_j + \theta_{bk} + \delta_{j, bk}$ for each outlying study $j$. As per Bayesian approach, the down-weighting factors $w_j$ are treated themselves as random variables and hence assigned their own prior distributions. We choose informative beta priors $w_j \sim \text{Beta}(a_j, b_j)$, so that the hyperparameters $a_j$ and $b_j$ can be specified to reflect how unusual the outlying study $j$ appears to be: they can be centered at values $\leq 0.5$ if we seek to apply a severe down-weighting - for example if there is additional external evidence supporting our hypothesis - or conversely, centered at values $\geq 0.5$ if we seek to apply a moderate down-weight - for example when being more uncertain about whether the study is an actual outlier or not. Examples of beta distributions reflecting different prior belief scenarios can be found in the Supplementary material. Ideally, external opinion should be used to elicit the beta distribution incorporating information from experts about their level of trust of suspicious effect sizes or studies. This approach would be particularly beneficial in the presence of so-called “mega-trials” with large discrepancies between fixed and random effects pooled estimates. In presence of heterogeneity, a random effect model would indeed give a large weight to small studies: if appropriate, our scheme could down-weight such studies according to expert information.

6 | SIMULATION STUDY

We conducted a simulation study to assess the performance of our outlier detection tools on binary outcome data. We constructed four different network geometries and we analyzed a number of different scenarios, varying the amount of heterogeneity and number of outliers included in the network. For each scenario, we simulated $\tau = 1000$ data sets for two- and multi-arm trials by drawing study-specific treatment effects $\theta$ and covariance matrix $\Psi_\delta$, as defined in Section 5. In all scenarios we sampled two MCMC chains, with 50000 iterations and a burn-in period of size 10000. Vague normal priors, $N(0, 1000)$, were used for the fixed effect and for each basic parameter and location-shift parameter. A vague uniform distribution, $U(0, 5)$, was used for the heterogeneity $\tau^2$ and a beta prior centred
around 0.5, i.e. \( w_i \sim \text{Beta}(3, 3) \), was used for the down-weighting factors, to reflect a moderate down-weighting.

The proposed methods have been then compared to similar approaches available in the literature, namely the Likelihood Ratio approach proposed in [Noma et al. 2020] where bootstrapped outlier \( p \)-values are approximated, and the Bayesian \( p \)-value proposed in [Zhang et al. 2015]. In addition, we have compared the methods to two cross-validatory leave-one-out alternatives, namely a recently developed Forward Search (FS) algorithm (Petropoulou et al. 2021) and the Conditional Predictive Ordinate (CPO) diagnostics (Gelfand, 1995). The former monitors several diagnostic measures in a forward fashion, i.e. starting from a basic ‘outlier-free’ set of studies and sequentially adding the remaining studies, while the latter is a Bayesian diagnostic to detect surprising observations. CPO values where estimated via integrated nested Laplace approximations (INLA) (Rue et al. 2009; Held et al. 2010; Sauter and Held, 2015), which is an available alternative to MCMC to estimate Bayesian models. Specifically, INLA offers a convenient and fast way to perform cross-validatory posterior predictive checking in a Bayesian framework without re-running the model during the backward search. Extreme values of the diagnostics measures monitored during the FS search (here Cook’s distance larger than 1) as well as large CPO values (typically extreme values above 70) may indicate outlying observations (Ntzoufras, 2009).

6.1 Simulation settings and data generation

The number of studies per comparison was set to 10 for all comparisons in an ideally balanced design and ranged from 1 to 9 to reflect values more often encountered in practice in three different unbalanced designs. The number of patients per trial arm was simulated from a uniform distribution \( U (50, 200) \) rounding to the closest integer. The number of studies per comparison was set to 10 for all comparisons in an ideally balanced design and ranged from 1 to 9 to reflect values more often encountered in practice in three different unbalanced designs. The number of patients per trial arm was simulated from a uniform distribution \( U (50, 200) \) rounding to the closest integer. To calculate the probability of an event in each study treatment arm we first draw baseline risks, i.e. event rates for the reference (treatment 1), from a uniform distribution \( \pi_{treatment \, 1} \sim U (0.4, 0.6) \) and then back-calculate the probability of an event in each study treatment arm using both baseline risks, assuming an overall event risk of 0.5. Then, the study arm-specific number of events was generated from a binomial distribution, using the probability of an event and the number of patients per trial-arm. We set the underlying true log odds ratios \( \theta^{true} \) of each treatment versus reference to be fixed at equal intervals between 0 and 1. Variances of the simulated log odds ratios are sampled from \( U \left( s_{\min}^2, s_{\max}^2 \right) \) with \( \left( s_{\min}^2, s_{\max}^2 \right) \) either \((2^2, 3.5^2)\) or \((0.5^2, 2^2)\) to represent different between-study variation. To study our detection power at varying heterogeneity, we analyse several between-study heterogeneity values chosen accordingly to the predictive distributions for heterogeneity estimated empirically by [Turner et al. 2012] who elicit predictive distributions for heterogeneity expected in future meta-analyses. Different distributions are obtained for different settings defined by the type of outcome and intervention comparison. Specifically, we take \( \tau^2 \in \{0, 0.032, 0.096, 0.287\} \), which respectively correspond to no heterogeneity, first, second and third quartiles of the estimated distribution of heterogeneity (we choose the setting with subjective outcome and pharmacological vs. pharmacological intervention comparisons). Again, this choice of outcome and comparison was made to reflect common settings of published meta-analyses (Nikolakopoulou et al. 2014). Finally, we contaminate the data with 1 or 3 outlying log odds ratios, which are sampled from \( N \left( \theta_{bk} \pm C, \tau^2 \right) \), where \( C = 2.5 \sqrt{s_{\max}^2 + \tau^2} \) or \( C = 3 \sqrt{s_{\max}^2 + \tau^2} \), corresponding respectively to less extreme and more extreme outliers. Number of events are then
sampled accordingly from binomial distributions. Overall, we explored 32 different scenarios. A detailed summary of the different scenarios is reported in Table 1.

**Table 1** Summary of the different scenarios analysed in the simulation study.

| Network geometry | Design | Scenario | Heterogeneity | Number of outliers |
|------------------|--------|----------|---------------|--------------------|
| **Balanced**     | fairly-connected network | 1 | 0 | 1 |
|                  |        | 2 | 0.032 | 1 |
|                  |        | 3 | 0.096 | 1 |
|                  |        | 4 | 0.287 | 1 |
|                  |        | 5 | 0    | 3 |
|                  |        | 6 | 0.032 | 3 |
|                  |        | 7 | 0.096 | 3 |
|                  |        | 8 | 0.287 | 3 |
| **Unbalanced**   | well-connected network | 9 | 0 | 1 |
|                  |        | 10 | 0.032 | 1 |
|                  |        | 11 | 0.096 | 1 |
|                  |        | 12 | 0.287 | 1 |
|                  |        | 13 | 0    | 3 |
|                  |        | 14 | 0.032 | 3 |
|                  |        | 15 | 0.096 | 3 |
|                  |        | 16 | 0.287 | 3 |
| **Unbalanced**   | fairly-connected network | 17 | 0 | 1 |
|                  |        | 18 | 0.032 | 1 |
|                  |        | 19 | 0.096 | 1 |
|                  |        | 20 | 0.287 | 1 |
|                  |        | 21 | 0    | 3 |
|                  |        | 22 | 0.032 | 3 |
|                  |        | 23 | 0.096 | 3 |
|                  |        | 24 | 0.287 | 3 |
| **Unbalanced**   | poorly-connected network | 25 | 0 | 1 |
|                  |        | 26 | 0.032 | 1 |
|                  |        | 27 | 0.096 | 1 |
|                  |        | 28 | 0.287 | 1 |
|                  |        | 29 | 0    | 3 |
|                  |        | 30 | 0.032 | 3 |
|                  |        | 31 | 0.096 | 3 |
|                  |        | 32 | 0.287 | 3 |

### 6.2 Simulations results

The main results of the simulation study are reported in Table 2 and Table 3. The case of an unbalanced design with a fairly-connected network contaminated with either one or three outliers is chosen as a representative case often encountered in practice. To explore false-positive detections, we also included
comparisons with the scenario where no outliers were induced. Additional results for all the included studies and remaining scenarios are reported in the Supplementary material.

To assess our methods, we report the following performance measures. First, we calculate mean Bayes Factors and mean posterior predictive p-values (under both discrepancy definitions). Then, we calculate and report the proportion of false-positive detections when no synthetic outliers are induced in the network and we compare with the false-positive rates obtained under competing methods. In addition, to assess the benefit of the down-weighting scheme we report the estimate relative bias for each treatment contrast, defined as \( (\hat{\theta}_{MC} - \theta_{true})/\theta_{true} \), with \( \hat{\theta}_{MC} \) Monte Carlo average of estimated effects. The evidence from Bayes factors is typically quantified as weak, moderate, strong or decisive through heuristic classification schemes (see Kass and Raftery 1995 table in Supplementary material). Following Kass and Raftery, we consider a study to show weak evidence of outlyingness if the Bayes Factor is above 3.2, decisive evidence if above 100, while as per standard convention we consider a study to show some evidence of outlyingness if the p-value is below 0.05. Similarly, details about the thresholds for detection used in the CPOs and FS algorithm are discussed in the Supplementary material.

**TABLE 2** Mean Bayes factors and mean posterior predictive p-values for the induced outliers of 1000 simulated data sets for the unbalanced design with a fairly connected network of 27 studies (scenarios 17-24 in Table 1). At varying scenarios, either a single outlier or three outliers (outlier 1, outlier 2, outlier 3) are induced in the network. Here, BF: Bayes Factor test; LR: Likelihood Ratio test as in Noma et al. 2020 (bootstrapped p-values reported); \( p_L \), \( p_{SDO} \) and \( p_G \) posterior predictive p-values under under likelihood-based discrepancy in (9), Stahel-Donoho outlyingness discrepancy in (10), and Gelman’s Omnibus \( \chi^2 \) as in Zhang et al. 2015, CPO: conditional predictive ordinate values; FS: Forward search algorithm as in Mavridis et al 2017 (Cook’s distance reported); \( \tau^2 \): heterogeneity (thresholds for detection: \( BF > 3.2 \) for Bayes factors; \( p < 0.05 \) for p-values, \( CPO > 70 \) for conditional predictive ordinates and Cook’s distance > 1 for FS algorithm).

| \( \tau^2 \) | Induced outliers | BF | \( p_L \) | \( p_{SDO} \) | LR | \( p_G \) | CPO | FS |
|-------|-----------------|----|--------|---------|-----|--------|------|----|
| 0     | a single outlier | 2063.3 | <0.001 | <0.001 | 0.01 | <0.001 | 118 | 3.7 |
|       | outlier 1       | 511.1 | 0.01   | 0.05   | 0.01 | 0.01   | 78  | 3.2 |
|       | outlier 2       | 118.2 | 0.01   | 0.001  | 0.02 | 0.01   | 115 | 2.1 |
|       | outlier 3       | 284.1 | 0.001  | 0.01   | 0.01 | 0.05   | 101 | 2.7 |
| 0.032 | a single outlier | 1540.1 | 0.002  | <0.0001 | 0.01 | 0.01   | 132 | 3.0 |
|       | outlier 1       | 287.1 | 0.05   | 0.01   | 0.01 | 0.07   | 68  | 2.8 |
|       | outlier 2       | 452.1 | 0.001  | 0.001  | 0.01 | 0.02   | 80  | 2.2 |
|       | outlier 3       | 32.1  | 0.01   | 0.01   | 0.06 | 0.06   | 99  | 2.9 |
| 0.096 | a single outlier | 11.1  | 0.05   | 0.05   | 0.02 | 0.07   | 35  | 1.0 |
|       | outlier 1       | 9.1   | 0.06   | 0.06   | 0.04 | 0.13   | 32  | 0.8 |
|       | outlier 2       | 2.7   | 0.03   | 0.05   | 0.05 | 0.05   | 46  | 0.6 |
|       | outlier 3       | 2.8   | 0.04   | 0.05   | 0.04 | 0.10   | 39  | 0.6 |
| 0.287 | a single outlier | 3.5   | 0.04   | 0.08   | 0.06 | 0.07   | 30  | 0.3 |
|       | outlier 1       | 2.5   | 0.22   | 0.12   | 0.10 | 0.40   | 10  | 0.6 |
|       | outlier 2       | 1.3   | 0.10   | 0.12   | 0.10 | 0.25   | 22  | 0.6 |
|       | outlier 3       | 0.98  | 0.17   | 0.15   | 0.12 | 0.31   | 29  | 0.5 |

Based on Table 2, both our Bayes Factor tests and posterior predictive p-values are able to detect the majority of the artificial outliers induced in the network (demonstrated by either large Bayes Factors and/or small p-value). In particular, our posterior predictive p-value based on the likelihood is able
TABLE 3 Proportion of false-positive detections, when no outlier is induced in the network, at varying simulation scenario. Here, BF: Bayes Factor tests; LR: Likelihood Ratio test as in Noma et al. 2020; \( p_L \), \( p_{SDO} \) and \( p_G \) posterior predictive \( p \)-values under under likelihood-based discrepancy in (9), Stahel-Donoho outlyingness discrepancy in (10), and Gelman’s Omnibus \( \chi^2 \) as in Zhang et al. 2015, CPO: conditional predictive ordinate values; FS: Forward search algorithm as in Mavridis et al. 2017; \( \tau^2 \): heterogeneity.

| Design                        | \( \tau^2 \) | BF  | LR test | \( p_L \) | \( p_{SDO} \) | \( p_G \) | CPO | FS |
|-------------------------------|--------------|-----|---------|-----------|-------------|-----------|-----|----|
| Balanced fairly-connected     | 0            | 0   | 0       | 0         | 0           | 0         | 0   | 0  |
| (100 studies in total)        | 0.032        | 0   | 0       | 0         | 0           | 0         | 0   | 0  |
|                               | 0.096        | 0.1 | 0.1     | 0.1       | 0.01        | 0.01      | 0.02| 0.02|
|                               | 0.287        | 0.6 | 0.06    | 0.03      | 0.03        | 0.06      | 0.11| 0.06|
| Unbalanced well-connected     | 0            | 0   | 0       | 0         | 0           | 0         | 0.03| 0  |
| (35 studies in total)         | 0.032        | 0.03| 0.03    | 0.03      | 0           | 0         | 0.03| 0.03|
|                               | 0.096        | 0.06| 0.07    | 0.04      | 0           | 0         | 0.09| 0.06|
|                               | 0.287        | 0.11| 0.11    | 0.08      | 0.08        | 0.09      | 0.11| 0.10|
| Unbalanced fairly-connected   | 0            | 0   | 0       | 0         | 0           | 0         | 0.06| 0.06|
| (27 studies in total)         | 0.032        | 0.06| 0.08    | 0.06      | 0           | 0.13      | 0.13| 0.06|
|                               | 0.096        | 0.12| 0.12    | 0.06      | 0.06        | 0.13      | 0.02| 0.07|
|                               | 0.287        | 0.12| 0.13    | 0.12      | 0.12        | 0.13      | 0.02| 0.14|

to identify some outliers in two highly heterogeneous scenarios where the Bayes factor tests fail (outlier 1 and outlier 2 for \( \tau^2 = 0.096 \), single outlier for \( \tau^2 = 0.287 \). As we can see from Table 2 the detection performance is slightly higher when only one outlier is present in the network, and this might be due to the fact that multiple outliers can shift the overall network meta-analysis model estimates to an extent to which they are not anymore recognised as deviating. As might be expected, the detection becomes difficult at increasing heterogeneity. All the induced outliers are detected only when heterogeneity is absent or low, \( \tau^2 = 0 \) and \( \tau^2 = 0.032 \); while only some outliers are detected for \( \tau^2 = 0.096 \) and very few when \( \tau^2 = 0.287 \) (see Supplementary material for all remaining scenarios). As expected, Bayes Factors and Likelihood Ratio tests have similar performance, while \( p \)-values based on Gelman’s discrepancy (Zhang et al. 2015) perform quite poorly in this context. This seems to be in line with results reported in Zhang et al. (2015), which comment that their measure “fails to uncover any outlyingness under the contrast-based framework, with all Bayesian \( p \)-values simply around 0.50”. Cross-validatory CPO diagnostic and Forward Search based on Cook’s distance perform quite well when low or moderate heterogeneity is present, but largely fail in highly heterogeneous scenarios. CPO diagnostics do not always discriminate well outliers from influential data, as points with high leverage may have small CPOs, independently of whether or not they are outliers.

Table 3 shows that methods based on posterior predictive \( p \)-values led to smallest false-positive rates, on average and in unbalanced cases, the rate was slightly higher under the likelihood-based discrepancy compared to the other discrepancy measures. This might be due to the fact that the likelihood contribution of each study is itself affected by the heterogeneity parameter, which in these
cases lead to very small $f_i(D_i^* | \theta_0)$. Conversely, $f_i(D_i | \theta_0)$ can be large, as the values observed values can be quite dispersed. Clearly, this leads to very small $p$-values, likely to be falling below the threshold of outlyingness (see Supplementary material). This seems also in agreement with the findings in Zhang et al. [2015].

Finally, to assess the performance of our down-weighting scheme on each contrast estimate, we computed the estimate relative bias. Figure 2 reports the effect of the down-weighting method on the estimate biases at varying heterogeneity, for the unbalanced scenario with poorly connected network and three artificial outliers, which is associated with the highest down-weighting benefit. The contrast estimates which show highest bias refer to the treatment comparisons of the outlying studies and in some cases, of treatment comparisons informed by very few studies. Full results for the other scenarios can be found in the Supplementary material. In all scenarios, down-weighting the suspicious studies is almost always associated with less biased estimates, with magnitude of benefit increasing at larger heterogeneity, in particular for those contrast for which direct evidence is available.

**FIGURE 2** Relative bias plot for the network meta-analysis estimates with and without down-weight out of 1000 simulated data sets, at varying heterogeneity, for the case of an unbalanced design with poorly connected network and three induced outliers.

7 | APPLICATIONS

In this section we apply our proposed outlier-detection tools to the two motivating networks of interventions previously described in Section 2 and illustrated in Figure 1. For lung cancer data, we used objective response (ObR) - defined as a complete response or a partial response according to the Response Evaluation Criteria in Solid Tumors (Therasse et al., 2017) - while smoking cessation data report the number of individuals who successfully quit smoking after 6 to 12 months. In both cases, the odds ratio (OR) was used as a summary measure.
In Figure 3 we report the estimated Bayes factors for each study and the posterior predictive distributions for the detected outliers under the likelihood-based discrepancy. Similar results were achieved under the Stahel-Donoho outlyingness (SDO) discrepancy and can be found in the Supplementary material. In both data sets, we used 50000 iterations for two MCMC chains and a burn-in period of 10000 samples. Vague normal priors, $N(0, 1000)$, were used for the fixed effect and for each basic parameter and location-shift parameter, and a vague uniform distribution, $U(0, 5)$, was used for the heterogeneity $\tau^2$. Our diagnostic tools detected three potential outliers in the NSCLC network. Here, study 44 and 42 were associated with large Bayes factors and relatively small predictive $p$-values, in support of a strong or decisive evidence in favour of outlyingness, while study 7 is associated with a relatively low Bayes factor and high $p$-value, suggesting a low evidence of outlyingness. In the smoking cessation network, one potential outlier was identified (study 3), associated with moderate Bayes factor and predictive $p$-value. This study was also identified as outlying in [Petropoulou et al., 2021].

In the lung cancer network, most included studies have unknown status for epidermal growth factor receptor (EGFR), while study 44 and study 42 included Asian patients with respectively wild-type mutation and KRAS (Kirsten Rat Sarcoma Virus) mutation. Compared to the few other included studies with these types of mutations, study 44 and study 42 (both comparing Monochemotherapy vs Immunotherapy) have considerably larger proportions of nonsmokers, and these patients are known to vastly differ from smokers in terms of driver mutations and therapy responsiveness (Immunotherapy in particular).

Further, the impact of so-called ‘small-study effects’ was assessed graphically though comparison-adjusted funnel plots, which can in some cases raise additional flags of outlyingness. Here, study 3 in the smoking cessation data creates an asymmetry in the plot (see Figure 4) but interestingly, neither study 42 nor study 44 are identified as suspicious, supporting the need of sophisticated methods to be used rather than relying on simple visual inspection of funnel plots or standardised residuals. A second stage of analysis was then performed to down-weight these potential outliers, as described in Section 5. The choices of the beta hyperparameters were made according to the degree of outlyingness of each study. For lung cancer data, a Beta(3, 3) - which is centred around 0.5 - was used for study 7 in the lung cancer data and study 3 in the smoking cessation data to reflect the large uncertainty about outlyingness. For study 42 and 44, we employed a beta distribution more concentrated in the range $(0, 0.5)$, i.e. Beta(2, 5), as we have stronger evidence in favour of outlyingness and so wish to apply a more severe downgrading. We refer the reader to Figure 1 in Supplementary material for a visual inspection of the chosen beta distributions.

Finally, we assessed the robustness of our results comparing the network estimates with and without down-weighting, and when outliers are removed from the network. For lung cancer data, we observe a reduction in the heterogeneity estimates both when the three studies are down-weighed and excluded. Study 44 was associated with the highest contribution matrix percentages (full contribution matrix reported in the Supplementary material). The contribution matrix measures how much each direct treatment effect contributes to the effect estimate from network meta-analysis and can support detection of influential studies. However, as shown in Figure 5 moderate changes in the comparative ORs were observed in the overall estimates, where the most significant change is in the effect of Immunotherapy vs. Targeted therapy, which changed from 0.72 (95%CI: 0.64-0.80) to 0.67 (95%CI: 0.60-0.75). For the smoking cessation data, the down-weighting of Study 3 (No contact vs. Individual Counselling) markedly reduced the estimated heterogeneity (from $\tau^2 = 0.541$ to $\tau^2 = 0.162$) and thus, the standard error estimates of the ORs became smaller as a whole. In particular, the comparative OR of Individual Counselling vs. No Contact was changed from 2.09 (95%CI: 1.35-3.19) to 1.67 (95%CI: 1.26-2.75) with down-weighting and 1.58 (95%CI: 1.21-2.09) with
FIGURE 3  Bayes factors with thresholds of evidence (annotated values on the linear scale) and histograms of draws from the posterior predictive distribution for the replicated vs. realised likelihood (vertical line) for the potential outliers identified under the likelihood-based discrepancy measure $f$, alongside a randomly chosen non-outlying study used as comparison (with annotated posterior predictive $p$-values). The two upper plots correspond to lung cancer data, and the lower two plots to smoking cessation data.
study exclusion. Here, down-weighting study 3 appears a more conservative choice, as with relatively small networks the exclusion of even a single study can affect significantly the overall estimates.

**FIGURE 4** Comparison-adjusted funnel plots centered at comparison-specific effect with pseudo confidence intervals at 90%, 95% and 99% approximate confidence levels for the lung cancer data (left panel) and smoking cessation data (right panel). Each comparison-specific effect is plotted against their reversed standard error to further investigate the distribution of the effect sizes. Treatments ordered from oldest to newest in both networks. Studies in red correspond to the potential outliers detected.

8 | DISCUSSION

In this paper, we have proposed two model-based methods to detect outlying studies in network meta-analysis, leveraging Bayes factors and posterior predictive assessments, and we have further presented a simple scheme to down-weight the studies detected. We have focused on binary data, but the methods can be applied to any type of data. All proposed methodology was tested both on simulated and empirical data.

In simulations, we have identified most of the artificially induced outliers, although both methods fail to some degree to detect outliers with poorly connected networks, with few studies per comparison and mostly, at increasing heterogeneity. This is relatively expected as outliers may in fact cause heterogeneity to be overestimated and in turn affect procedures to detect them, especially when there is not enough information available in the network. Posterior predictive p-values achieved the best detection power in comparison with Bayes factors, under both the likelihood-based and SDO-based discrepancy. The forward search (FS) algorithm and, in several scenarios, the cross-validation conditional predictive ordinates (CPO) computed via INLA were also outperformed. Likewise posterior predictive p-values, CPO is a Bayesian diagnostic tool based on predictive densities but does not always discriminate well outliers from influential data, as points with high leverage may have small CPOs, independently of whether or not they are outliers. This suggests that the use of posterior predictive p-values with discrepancy measures able to capture extreme deviations are essential to improve the detection performance within network meta-analyses, as also pointed out by Zhang et al. 2015. When we used our approaches in an network meta-analysis of 112 randomised controlled
trials comparing second-line treatments for advanced NSCLC, we identified one clear and two potential outliers corresponding to very large and moderate Bayes factors and posterior predictive $p$-values. In the well-known smoking cessation data, we identified one potential outlier, with a moderate Bayes factor and $p$-value. The down-weighting scheme yielded a significant reduction in the bias of the relative effect sizes estimates in simulations, suggesting the scheme to be effective; which was also confirmed on real data by an overall reduction in heterogeneity and more precise confidence intervals of the network meta-analysis estimates. In the smoking cessation data, it also led to a clear reduction in the contrast estimate related to the outlying study, suggesting it to be also influential.

With both simulated and real data, the different detection methods were not always in full agreement, confirming that it is good practice to jointly assess more than one measure when searching for outliers. Indeed, our proposed tools should not be seen as competing alternatives, but rather as complementing each other and should ideally be used in combination. This is because they capture different aspects of the modelling mechanism: while Bayes factors can be used to compare models (in our case a standard model versus an outlier mean-shift model), posterior predictive $p$-values can only assess discrepancy between the observed data and some assumed model. A reason in support of the Bayes factor is that it is based on weighing the alternative models by the posterior evidence in favour of each of them and thus can also measures evidence in favour of the null hypothesis. Similarly, posterior predictive $p$-values can represent powerful tools for assessing outliers in a Bayesian fashion, but require careful choice of the discrepancy measure, that should always be chosen according to the scientific context and question of interest.

Our proposed tools present also limitations. For example, Bayes factors are known to be dependent on the choice of the prior distributions and thus caution is needed, especially when informative priors
are used into the network meta-analysis model. Moreover, our Bayes factor test depends on how the alternative model is defined. In this paper, the outlier model was constructed as a mean-shift model, but more sophisticated approaches, for example incorporating both a shift in mean and in variance, could be considered. Under certain circumstances, this would aid to account into the model for sample size or related phenomena such as small-study effects. Overall, the method searches for one outlier at the time, making it subject to well-known masking problems (e.g. when a cluster of outliers shift the model parameters to a degree that makes these observations not being identifiable as outliers). Accounting for multiple outliers simultaneously is a topic of further research which would require external knowledge about the groups of studies to be tested to achieve computational feasibility. The posterior predictive \( p \)-value assessment could alternatively be carried out in a cross-validatory leave-one-out setting but it would become computationally intensive, which can be problematic when the network is large (Marshall and Spiegelhalter 2003). Regarding the discrepancy measures chosen, one limitation of the Stahel-Donoho measure is that it implicitly assumes the non-outlier data to be symmetrically distributed and thus it may fail to detect asymmetry in very skewed data. Other choices can include the skewness-adjusted outlyingness (AO) measure (Brys et al. 2003). Assessment of inconsistency was out of scope in this paper, but we should acknowledge that outlying studies can also be the primary source of inconsistency; in which case differentiating between outlyingness and inconsistency would be difficult: as with heterogeneity, outliers may contribute significantly to an increased inconsistency in the network whilst at the same time affecting the inconsistency checking procedures.

In conclusion, our methods have shown encouraging outlier detection results, but we advise that they should always be used in conjunction with clinical expertise and judgement. Looking at future work, we are interested in extending the methodology in a multiple outcome framework (Efthimiou et al. 2015), to see whether a study has an outlying behaviour in all the reported outcomes. Clearly, this would allow to draw more precise conclusions about the outlyingness of each study in the network. Finally, our simple down-weighting scheme could be refined to allow automatic down-weight of the outliers, rather than specifying the down-weighting factors for outlying studies only at a second stage of analysis. Again, expert information could be used for constructing more appropriate down-weighting factors and further sensitivity analyses may be added to compare the choice of different prior weights. The source code for the proposed methods, which we further plan to incorporate into an R package to facilitate broader usage, is freely available at https://github.com/silviametelli/Bayes-NMA-outlier-detection

**REFERENCES**

Lumley T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* 21, 2313-2324.

Lu, A. G. and Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* 23, 3105–3124.

Lu, A. G. and Ades, A. E. (2006) Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 474, 447–459.

Higgins, J. P. and Whitehead, A. (1996) Borrowing strength from external trials in a meta-analysis, *Statistics in Medicine* 15: 2733–49.

Viechtbauer W. and Cheung M. W. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* 1, 112–125.

Gumedze, F. N. and Jackson, D. (2011). A random effects variance shift model for detecting and
accommodating outliers in meta-analysis. *BMC Medical Research Methodology* **11**, 1–19.

Zhao H., Hodges, J. S. and Carlin, B.P. (2017). Diagnostics for generalized linear hierarchical models in network meta-analysis. *Research Synthesis Methods* **8**, 333–342.

Mavridis, D., Moustaki, I., Wall, M. and Salanti, G. (2017). Detecting outlying studies in meta-regression models using a forward search algorithm. *Research Synthesis Methods* **8**, 199–211.

Negeri, Z. F. and Beyene, J. (2020). Statistical methods for detecting outlying and influential studies in meta-analysis of diagnostic test accuracy studies. *Statistical Methods in Medical Research* **9**, 1227–1242.

Noma H., Gosho M., Ishii R., Oba, K. and Furukawa, T. A. (2020). Outlier detection and influence diagnostics in network meta-analysis. *Research Synthesis Methods* **11**, 891–902.

Dias S., Ades, A. E., Welton, N. J., Jansen, J. p. and Sutton, A. J. (2018) Network Meta-Analysis for Decision Making. John Wiley & Sons.

Matsushima, Y., Noma, H., Yamada, T. and Furukawa, T. A. (2020). Influence diagnostics and outlier detection for meta-analysis of diagnostic test accuracy. *Research Synthesis Methods* **11**, 237–247.

Zhang, J., Fu, H. and Carlin, B. P. (2015). Detecting outlying trials in network meta-analysis. *Statistics in Medicine* **34**, 2695–2707.

Hedges, L. V. and Olkin, I. (1985). Statistical Method for Meta-Analysis, Orlando, FL: Academic Press.

Lin, L., Chu, H. and Hodges, J. S. (2017). Alternative measures of between-study heterogeneity in meta-analysis: reducing the impact of outlying studies. *Biometrics* **13**, 156–166.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics* **22**, 1142–1160.

Gelman, A., Meng, X. L. and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733–807.

Stahel, W. (1981). Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen, PhD thesis, ETH Zürich.

Donoho, D. (1982). Breakdown properties of multivariate location estimators, Ph.D. Qualifying paper, Dept. Statistics, Harvard University, Boston.

Petropoulou, M., Salanti, G., Rücker, G., Schwarzer, G., Moustaki, I. and Mavridis, D. (2021). A forward search algorithm for detection of extreme study effects in network meta-analysis. *Statistics in Medicine* **1–15**.

Gelfand, A. E. (1995). Model Determination Using Sampling-Based Methods, In: Gilks W, Richardson S and Spiegelhalter D (eds) Markov Chain Monte Carlo In Practice, London, Chapman Hall.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B* **71**, 319–392.

Held, L., Schrödle, B. and Rue, H. (2010). Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA. In: Kneib T and Tutz G. (eds) Statistical Modelling and Regression Structures. Physica-Verlag HD.

Sauter, R. and Held, L. (2015). Network meta-analysis with integrated nested Laplace approximations. *Biometrical Journal* **57**, 1038–1050.

Ntzoufras I. (2009). Bayesian Modeling Using WinBUGS. John Wiley & Sons, West Sussex, England.
Turner, R.M., Davey, J., Clarke, M.J., Thompson, S. G. and Higgins, J. P. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International Journal of Epidemiology* **41**, 818–82.

Nikolakopoulou, A., Chaimani, A., Veroniki, A. A., Vasiliadis, H. S., Schmid, C. H. and Salanti, G. (2014). Characteristics of Networks of Interventions: A Description of a Database of 186 Published Networks. *Plos One* **9**, e86754.

Créquit P, Trinquart L, Yavchitz A., Ravaud P. (2016). Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. *BMC Medicine*, **14** (8).

Créquit, P., Chaimani, A., Yavchitz, A., Attiche, N., Cadranel, J., Trinquart, L. and Ravaud, P. (2017). Comparative efficacy and safety of second-line treatments for advanced non-small cell lung cancer with wild-type or unknown status for epidermal growth factor receptor: a systematic review and network meta-analysis. *BMC Medicine* **15**, 193.

Masters, G.A., Temin, S., Azzoli, C.G., Giaccone, G., Baker, S., Brahmer, J.R., et al (2015). Systemic therapy for stage IV non-small-cell lung cancer with wild-type or unknown status for epidermal growth factor receptor: American Society of Clinical Oncology Clinical Practice Guideline Update. *Journal of Clinical Oncology* **33** (30), 3488–515.

Hasselblad, V (1998). Meta-analysis of multitreatment studies. *Medical Decision Making*, **18** (1), 37–43.

Bayarri, M. and Berger, J. O. (2000). P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142.

Dahl, F. A. (2006). On the conservativeness of posterior predictive p-values. *Statistics and Probability Letters* **76**, 1170–1174.

Gelman A. (2013) Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics* **7**, 2595–2602.

Ibrahim, J. G. and Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46–60.

Verweij, J., Van Glabbeke, M., van Oosterom, A. T., Christian, M. C. and Gwyther, S. G. (2000). New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *Journal of the National Cancer Institute* **92**, 205–216.

Papakonstantinou, T., Nikolakopoulou, A., Rucker, G., Chaimani, A., Schwarzer, G., Egger, M. and Salanti, G. (2018). Estimating the contribution of studies in network meta-analysis: paths, flows and streams [version 3; peer review: 2 approved, 1 approved with reservations]. *F1000Research* **7**, 610.

Madan, J., Stevenson, M. D., Cooper, K. L., Ades, A.E., Whyte, S. and Akehurst, R. (2011) Consistency between direct and indirect trial evidence: is direct evidence always more reliable? *Value in Health* **14**, 953–960.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22**, 1649–1660.

Brys, G. Hubert, M. and Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics* **13**, 996–1017.

Efthimiou, O., Mavridis, D., Riley, R. D., Cipriani, A. and Salanti, G. (2015). Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics* **16**, 84–97.