Improved methods to construct prediction intervals for network meta-analysis

Hisashi Noma, PhD*
Department of Data Science, The Institute of Statistical Mathematics, Tokyo, Japan
ORCID: http://orcid.org/0000-0002-2520-9949

Yasuyuki Hamura, PhD
Graduate School of Economics, Kyoto University, Kyoto, Japan

Shonosuke Sugasawa, PhD
Faculty of Economics, Keio University, Tokyo, Japan

Toshi A. Furukawa, MD, PhD
Departments of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

*Corresponding author: Hisashi Noma
Department of Data Science, The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
TEL: +81-50-5533-8440
e-mail: noma@ism.ac.jp
Abstract

Network meta-analysis has played an important role in evidence-based medicine for assessing the comparative effectiveness of multiple available treatments. The prediction interval has been one of the standard outputs in recent network meta-analysis as an effective measure that enables simultaneous assessment of uncertainties in treatment effects and heterogeneity among studies. To construct the prediction interval, a large-sample approximating method based on the $t$-distribution has generally been applied in practice; however, recent studies have shown that similar $t$-approximation methods for conventional pairwise meta-analyses can substantially underestimate the uncertainty under realistic situations. In this article, we performed simulation studies to assess the validity of the current standard method for network meta-analysis, and we show that its validity can also be violated under realistic situations. To address the invalidity issue, we developed two new methods to construct more accurate prediction intervals through bootstrap and Kenward–Roger-type adjustment. In simulation experiments, the two proposed methods exhibited better coverage performance and generally provided wider prediction intervals than the ordinary $t$-approximation. We also developed an R package, PINMA (https://cran.r-project.org/web/packages/PINMA/), to perform the proposed methods using simple commands. We illustrate the effectiveness of the proposed methods through applications to two real network meta-analyses.

Key words: network meta-analysis; prediction interval; bootstrap; Kenward-Roger–type adjustment; higher-order approximation.
1. Introduction

Network meta-analysis has played an important role in evidence-based medicine, enabling a comprehensive synthesis of existing evidence and assessment of comparative effectiveness for multiple available treatments \(^1,2\). In network meta-analyses, heterogeneity among different studies is an important issue and random-effects models are therefore generally adopted. Conventionally, the “grand mean” parameters have been addressed as a primary estimand; however, the ability of the grand mean to express a summarized measure of treatment effects is substantially limited because it is solely a point measure that corresponds to the mean of the random-effects distribution. Quantifying the true effect and effectiveness in real-world applications of the treatment requires more appropriate measures that suitably reflect the degree of heterogeneity and magnitude of the treatment effect. The prediction interval, defined as an interval that covers the true treatment effect in a future study with a certain probability (e.g., 95%), has been proposed to address this problem \(^3,4\).

In recent systematic reviews, the prediction interval has been one of the standard outputs used as an effective measure that enables the simultaneous assessment of uncertainties in treatment effects and heterogeneity among studies \(^5,6\). To construct the prediction interval, large-sample approximating methods that use the \(t\)-distribution have generally been applied in practice, as originally proposed by Higgins et al. \(^3\) for conventional meta-analysis and extended to network meta-analysis in Chaimani and Salanti \(^7\). The \(t\)-distribution approximation is computationally efficient and has been effectively used as an approximation to bootstrap and permutation approaches in inference problems \(^8\); however, in the case of applications in conventional pairwise meta-analyses, recent studies have revealed that the \(t\)-approximation methods possibly underestimate the uncertainty and heterogeneity under certain conditions, especially
when the number of studies $N$ is not large (e.g., $\leq 10$) \textsuperscript{9-11}. The prediction interval has been gaining prominence in practice, and the accuracy of the resultant prediction interval is an important issue in circumventing misleading evidence. However, for network meta-analysis, the accuracy of prediction intervals obtained by the current standard method has not been investigated and alternative effective methods have not been well discussed. In addition, recent studies have reported that the validity of inference methods for network meta-analysis can be violated under moderate numbers of synthesized studies \textsuperscript{12-14}.

In the present work, we assess the properties of the $t$-approximation method to construct prediction intervals for network meta-analysis under realistic situations via simulation experiments and we show that an invalidity problem similar to that encountered with conventional pairwise meta-analysis can occur. In addition, to address the invalidity problem, we develop two new improved methods to construct prediction intervals for network meta-analysis. The first method is a bootstrap method that uses an estimated predictive distribution of the effect of a future study via bootstrap resampling as an alternative to the $t$-distribution. Parametric bootstrap methods have been conventionally applied to prediction problems of random effects parameters in general mixed-effects models, and the asymptotic approximations have been theoretically shown to be improved in general \textsuperscript{15-17}. However, parametric bootstrap methods have not been applied to the prediction problem of effect size of a future study generated from the random-effects distribution. In the present study, we therefore provide a new method to apply a parametric bootstrap method to construct prediction intervals in network meta-analysis. The second method is to use a higher-order asymptotic method for improving the $t$-approximation of the prediction distribution by the Kenward–Roger-type approximation \textsuperscript{18}. The Kenward–Roger method \textsuperscript{18} has been widely used for linear mixed-effects models that improve the conventional restricted maximum likelihood (REML)
method involving the Satterthwaite-type approximation. The Kenward–Roger methods were originally developed as inference methods, not as prediction tools; however, Partlett and Riley 9 showed that they can be effectively applied to the prediction issues in conventional pairwise meta-analysis. Recently, Noma et al. 19 developed Kenward–Roger-type approximation methods for network meta-analysis. In the present work, we propose a new method that applies the improved approximation method 19 to the prediction problem. We also developed an R package, PINMA (https://cran.r-project.org/web/packages/PINMA/), for implementing these new methods. We illustrate the effectiveness of the proposed methods by applying them to two real data examples of network meta-analyses 20,21.

2. Multivariate random-effects model for network meta-analysis

We here consider the contrast-based multivariate random-effects model for network meta-analysis 22,23. We suppose the synthesis of $N$ trials and compare $p + 1$ treatments. Let $Y_{ij}$ denote an estimator of a treatment effect in contrast to a reference treatment (e.g., placebo) for the $j$th treatment in the $i$th trial ($i = 1, 2, \ldots, N; j = 1, 2, \ldots, p$). Commonly used effect measures are the mean difference, standardized mean difference, risk difference, risk ratio, odds ratio, and hazard ratio, where the ratio measures are usually transformed on a logarithmic scale. We then consider the following random-effects model for the outcome random vector \( Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})^T \):

\[
Y_i = \theta_i + \epsilon_i \tag{(*)}
\]

\[
\theta_i = \mu + \epsilon_i
\]

where \( \theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})^T \) and \( \mu = (\mu_1, \ldots, \mu_p)^T \); \( \theta_1, \ldots, \theta_N \) are the random-effects parameters, and \( \mu \) expresses an average treatment effect among the $N$ trials; \( \epsilon_i \) and \( \epsilon_i \) are independent random variation terms within and across studies ($p \times 1$ random
vectors) and are assumed to be distributed as \( \mathbf{e}_i \sim \text{MVN}(\mathbf{0}, \mathbf{S}_i) \) and \( \mathbf{e}_i \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma}) \), respectively. \( \mathbf{S}_i \) (a \( p \times p \) matrix) is the within-study covariance matrix,

\[
\mathbf{S}_i = \begin{pmatrix}
    s_{i1}^2 & \rho_{i12}s_{i1}s_{i2} & \cdots & \rho_{i1p}s_{i1}s_{ip} \\
    \rho_{i12}s_{i2}s_{i1} & s_{i2}^2 & \cdots & \rho_{i2p}s_{i2}s_{ip} \\
    \vdots & \vdots & \ddots & \vdots \\
    \rho_{ip1}s_{ip}s_{i1} & \rho_{ip2}s_{ip}s_{i2} & \cdots & s_{ip}^2
\end{pmatrix}
\]

which is usually assumed to be known and fixed to its valid estimate; several suitable estimators that can be obtained from summary statistics have been proposed \(^{24,25}\). In addition, \( \mathbf{\Sigma} \) is the between-studies variance–covariance matrix:

\[
\mathbf{\Sigma} = \tau^2 \mathbf{P} = \tau^2 \begin{pmatrix}
    1 & 1/2 & \cdots & 1/2 \\
    1/2 & 1 & \cdots & 1/2 \\
    \vdots & \vdots & \ddots & \vdots \\
    1/2 & 1/2 & \cdots & 1
\end{pmatrix}
\]

for \( \tau^2 > 0 \). Note that the correlation structure of \( \mathbf{\Sigma} \) can be assumed to be unstructured; however, in general practice of network meta-analysis, there are rarely a sufficient number of studies to estimate all of the variance–covariance parameters. Thus, most network meta-analyses adopt the equal-variance assumption for the \( p \) components of \( \mathbf{e}_i \); then, all the pairwise correlation coefficients should be equal to 0.50 because of the consistency restriction \(^{26,27}\). In the present work, we also consistently adopt this standard assumption. For trials that do not include a reference treatment, the data augmentation approach of White et al. \(^{23}\) can be adopted; a quasi-small dataset is added into the reference arm (e.g., 0.001 events for 0.01 patients for a binary outcome). Because most clinical trials typically involve only two or three arms, many of the components of \( \mathbf{Y}_i \) are undefined and we formally use the sub-vector of them, corresponding to available treatment pairs.
For standard inference of the model parameters, REML estimation is the most widely adopted method in practice \(^{22,28}\). The log-likelihood function of the multivariate meta-analysis model is given as \(^{28}\)

\[
\ell(\mu, \tau^2) = \text{const.} - \frac{1}{2} \sum_{i=1}^{N} \{ \log(\det(W_i^{-1})) + (y_i - \mu)^T W_i (y_i - \mu) \}
\]

where \(W_i = (\Sigma + S_i)^{-1}\). \(Y_i\) and \(S_i\) involve missing components, as previously mentioned, and these are shrunk to the sub-vector and sub-matrix for the observed components in the log-likelihood function. In addition, the log-restricted likelihood function is written as \(^{28}\)

\[
\ell_{RL}(\mu, \tau^2) = \text{const.} + \ell(\mu, \tau^2) - \frac{1}{2} \log(\det(\sum_{i=1}^{N} W_i))
\]

The REML estimators of \{\(\hat{\mu}, \hat{\tau}^2\)\} are obtained by maximizing \(\ell_{RL}(\mu, \tau^2)\). We denote the covariance matrix of \(\hat{\mu}\) as \(\Phi\), and the ordinary estimator is given as \(\hat{\Phi} = (\sum_{i=1}^{N} \bar{W}_i)^{-1}\), where \(\bar{W}_i = (\bar{\Sigma} + \bar{S}_i)^{-1}\). The ordinary Wald-type tests and confidence intervals are constructed using the standard error estimator. In the following discussions, we used the REML estimator as a standard method; however, it can be replaced by another adequate estimator such as the ordinary maximum likelihood (ML) estimator or Jackson’s method-of-moments estimator \(^{29,30}\).

3. Prediction interval using the approximation by \(t\)-distribution

For the multivariate random-effects model (*), the prediction interval is derived by the predictive distribution of the effect of a future study \(\theta_{new}^{XY}\) for a certain comparison \(X\) vs. \(Y\) among \(p\) treatments. Using the frequentist framework of Higgins et al. \(^3\), we can consider an approximate sampling distribution of \(\hat{\mu}^{XY}\) that is the REML estimator of \(\mu^{XY}\), which is a component of \(\mu\) or an adequate contrast of several components of \(\mu\):
\[ \hat{\mu}^{XY} \sim N\{\mu^{XY}, SE(\hat{\mu}^{XY})^2\} \]

where \( SE(\hat{\mu}^{XY}) \) is the standard error of \( \hat{\mu}^{XY} \). The effect of an independent future study \( \theta_{\text{new}}^{XY} \) follows the common random-effects distribution,

\[ \theta_{\text{new}}^{XY} \sim N(\mu^{XY}, \tau^2) \]

implying that \( \theta_{\text{new}}^{XY} - \hat{\mu}^{XY} \sim N\{0, \tau^2 + SE(\hat{\mu}^{XY})^2\} \) and, therefore, that

\[ \theta_{\text{new}}^{XY} \sim N\{\hat{\mu}^{XY}, \tau^2 + SE(\hat{\mu}^{XY})^2\} \]

Note that the heterogeneity variance \( \tau^2 \) is unknown in practice and must be estimated. If it is substituted by the REML estimator \( \hat{\tau}^2 \), the normal approximation might be inadequate because the uncertainty of \( \hat{\tau}^2 \) is ignored. Higgins et al. 3 proposed using the \( t \)-distribution with \( N - 2 \) degrees of freedom to improve the approximation of the predictive distribution. Chaimani and Salanti 7 provided a modified degree of freedom for network meta-analysis with the number of direct comparison trials adjusted; that is,

\[ W = \frac{\theta_{\text{new}}^{XY} - \hat{\mu}^{XY}}{\sqrt{\hat{\tau}^2 + SE(\hat{\mu}^{XY})^2}} \sim t_{N-N_{XY}-1} \]

where \( N_{XY} \) is the number of direct comparisons for \( X \) vs. \( Y \) on the network. The approximate 100(1 - \( \alpha \))% prediction interval formula is then provided as

\[ \hat{\mu}^{XY} \pm t_{N-N_{XY}-1}^{\alpha} \sqrt{\hat{\tau}^2 + SE(\hat{\mu}^{XY})^2} \]

where \( t_{N-N_{XY}-1}^{\alpha} \) is the 100(1 - \( \alpha / 2 \))% percentile of the \( t \)-distribution with \( N - N_{XY} - 1 \) degrees of freedom. This procedure is currently a standard method to calculate prediction intervals in network meta-analysis and can be calculated using a standard statistical package (e.g., network graphs 7 for Stata).

4. Bootstrap prediction interval

The \( t \)-approximation is a computationally efficient and effective method in practice.
However, the parametric approximation is not rigorously justified and can be violated under certain conditions; actual numerical evidence is shown in Section 6. To improve the accuracy of the prediction interval, one effective approach is to use a bootstrap method that enables direct estimation of the sampling distribution of the predictive statistic $W$. Bootstrap methods have generally been applied to prediction problems of random-effects models, especially for prediction of random effects (e.g., constructing prediction intervals of the random effects parameters $\theta_1, \ldots, \theta_N$ in the case of the random-effects model (*); however, there are no methods that provide a bootstrap-based prediction interval for a future observation $\theta_{new}$.

We here develop a bootstrap procedure to construct the prediction interval for $\theta_{new}$. The basic concept is to substitute the $t$-distribution into the bootstrap sampling distribution of the predictive statistic $W$ to improve the approximation. The concrete algorithm is constructed as follows.

**Algorithm (Bootstrap prediction interval)**

1. For the multivariate random-effects model (*), calculate the REML estimate $\{\mu, \tau^2\}$ for $\{\mu, \tau \}$.  
2. Resample $Y^{(b)}_1, Y^{(b)}_2, \ldots, Y^{(b)}_N$ from the estimated distribution of (*) with the parameters substituted with $\{\mu, \tau\}$ via parametric bootstrap, $B$ times ($b = 1, 2, \ldots, B$).
3. Compute the REML estimates of model parameters $\{\mu^{(1)}, \tau^{2(1)}\}, \{\mu^{(2)}, \tau^{2(2)}\}, \ldots, \{\mu^{(B)}, \tau^{2(B)}\}$ for the $B$ bootstrap samples.
4. Simulate the effects of a future study $\theta_{new}^{XY(1)}, \theta_{new}^{XY(2)}, \ldots, \theta_{new}^{XY(B)}$ for treatment pair $X$ vs. $Y$ from the estimated random-effects distribution $N(\mu^{XY}, \tau^{2})$.
5. Calculate the prediction statistic,
\[ W^{(b)} = \frac{\theta_{\text{new}}^{XY(b)} - \hat{\mu}^{XY(b)}}{\sqrt{\hat{\tau}^2(b) + SE(\hat{\mu}^{XY(b)})^2}} \]

where \( \hat{\mu}^{XY(b)} \) is the REML estimator of \( \mu^{XY} \) for the \( b \)th bootstrap sample (\( b = 1, 2, \ldots, B \)).

6. Denote \( Q_1^{BS} \) and \( Q_2^{BS} \) as the 100\( \alpha/2 \)% and 100(1\( - \alpha/2 \))% quantiles of \( W^{(1)}, W^{(2)}, \ldots, W^{(B)} \), then construct the bootstrap 100(1\( - \alpha \))% prediction interval by

\[ \left( \hat{\mu}^{XY} + Q_1^{BS} \sqrt{\hat{\tau}^2 + SE(\hat{\mu}^{XY})^2}, \hat{\mu}^{XY} + Q_2^{BS} \sqrt{\hat{\tau}^2 + SE(\hat{\mu}^{XY})^2} \right) \]

The approximation error of coverage is generally improved by the bootstrap method 15-17. Note that the bootstrap quantiles \( Q_1^{BS} \) and \( Q_2^{BS} \) can be unrealistically small values when the true heterogeneity variance \( \tau^2 \) is small (e.g., \( \leq 0.20 \); see Section 6). The resultant prediction interval is then unnaturally narrow and can underestimate the heterogeneity and statistical uncertainty, which causes the \( \tau^2 \) estimate to be nearly 0. The bootstrap estimates \( \hat{\tau}^2(1), \hat{\tau}^2(2), \ldots, \hat{\tau}^2(B) \) can then also be values near 0, and \( W^{(1)}, W^{(2)}, \ldots, W^{(B)} \) vary in a narrow range. To circumvent these unnatural operating characteristics, we propose substituting \( Q_1^{BS} \) or \( Q_2^{BS} \) for 1.96 (the 97.5th percentile of the standard normal distribution) when the absolute value of \( Q_1^{BS} \) or \( Q_2^{BS} \) is less than 1.96; the results are then considered unnatural. The variation of \( W \) is considered to be at least larger than the standard normal distribution obtained as described in Section 3 because the uncertainty of \( \hat{\tau}^2 \) is added to the standardized quantity. The operating characteristics can be checked in the numerical studies presented in Section 6. In addition, the bootstrap algorithm is similarly applicable if the estimating method is changed to other methods (e.g., the ML estimation).
5. Prediction interval using the Kenward–Roger-type approximation

5.1 Improved variance estimator

For inferences of $\mu$ in the random-effects model (*), the ordinary Wald-type approximation of the REML estimator is not good, especially under small $N$ settings. Noma et al. 19 developed improved inference methods using the Kenward-Roger–type approximations 18. The higher-order approximation methods were not originally considered to apply to the construction of prediction intervals, but Partlett and Riley 9 proposed applying the same approximation to construct a prediction interval of pairwise meta-analysis. We here propose using the Kenward–Roger-type adjustment for the network meta-analysis.

First, Noma et al. 19 derived improved covariance estimators of the REML estimator $\hat{\mu}$. They were motivated by the tendency of the ordinary asymptotic covariance estimator $\Phi$ to generally underestimate the true variability of $\hat{\mu}$, mainly because it does not consider the variability of the heterogeneity variance estimator $\hat{\tau}^2$ 19. The variance estimator of $\hat{\tau}^2$ based on the expected information is given as

$$ \hat{\omega} = 2 \left\{ \sum_{i=1}^{N} \text{tr}(\hat{W}_iP\hat{W}_iP) - \text{tr}(2K^{-1}H - K^{-1}JK^{-1}J) \right\}^{-1} $$

where

$$ K = \sum_{i=1}^{N} \hat{W}_iJ = \sum_{i=1}^{N} \hat{W}_iP\hat{W}_iP, \quad J = \sum_{i=1}^{N} \hat{W}_iP\hat{W}_iP\hat{W}_i $$

Then, using the Kackar–Harville-type approximation 31, we can derive the improved asymptotic covariance estimator of $\hat{\mu}$ as

$$ \hat{\Phi}_{KR} = \Phi + 2\Phi\{\hat{\omega}(H - JK^{-1}J)\} \Phi $$

We denote the standard error estimator for $\mu^{XY}$ obtained by the components of $\hat{\Phi}_{KR}$ as $\hat{SE}_{KR}(\hat{\mu}^{XY})$. Substituting the standard error estimator into the predictive statistic is
expected to improve the approximation

\[
W_{KR} = \frac{\theta_{new}^{XY} - \hat{\mu}^{XY}}{\sqrt{\hat{\tau}^2 + \bar{SE}_{KR}(\hat{\mu}^{XY})^2}}
\]

### 5.2 Improved approximation for sampling distribution of \( W_{KR} \)

Second, to improve the approximation of the sampling distribution for a test statistic of the Wald-type test of \( H_0: \mu^{XY} = 0 \) vs. \( H_1: \mu^{XY} \neq 0 \), \( T^{XY} = \frac{\hat{\mu}^{XY}}{SE(\hat{\mu}^{XY})} \), Noma et al. 19 proposed using the \( t \)-distribution with an adjusted degree of freedom derived by the higher-order asymptotics considered by Kenward and Roger 18. The adjusted degree of freedom is

\[
\nu_{KR}^{XY} = \frac{2\kappa_{XY}^2}{\sigma \eta_{XY}}
\]

where \( \kappa_{XY} \) and \( \eta_{XY} \) are the corresponding diagonal elements of \( K^{-1} \) and \( K^{-1}JK^{-1} \) or their combinations (e.g., if \( X \) is the reference and \( Y \) is the \( k \)th treatment, they correspond to the \( k \)th diagonal components of them).

Then, according to the approximation method of Partlett and Riley 9, the approximate distribution of \( W_{KR} \) is the \( t \)-distribution with \( \nu_{KR}^{XY} - 1 \) degrees of freedom. Therefore, an improved approximate 100\((1 - \alpha)\)% prediction interval formula is obtained as

\[
\hat{\mu}^{XY} \pm t_{\nu_{KR}^{XY}-1}^{\alpha/2} \sqrt{\hat{\tau}^2 + \bar{SE}_{KR}(\hat{\mu}^{XY})^2}
\]

Note that the degree of freedom \( \nu_{KR}^{XY} \) can sometimes be a small value. The resultant prediction interval then becomes quite wide and cannot be defined when \( \nu_{KR}^{XY} \leq 1 \). To address this problem, we propose truncating \( \nu_{KR}^{XY} \) at 3 (if \( \nu_{KR}^{XY} < 3 \), it is rounded to 3), which is an adequate criterion that the operating characteristics of the inference methods

10
were well by numerical evidence in simulation studies; see e-Appendix in Supporting Information.

6. Simulations

To assess the operating characteristics of the proposed methods, we conducted simulation studies under practical situations, especially under moderate numbers of synthesized studies ($N$). The simulation settings mimicked the real data of network meta-analysis of Siontis et al. 21 (Section 7.2). Figure 1 shows the network geometric graph 32 of the simulation settings that compare six treatments (A, B, C, D, E, and F, where A is formally regarded as a reference). The number of synthesized trials $N$ was set to 10, 14, or 18; we adopted $N = 18$ and considered smaller settings ($N = 10, 14$) because the coverage properties would worsen under small $N$ settings. Larger $N$ settings are plausible, but the overall tendency is considered to be preserved. Outcome data of individual trials were simulated following binomial distributions, $X_{ij} \sim \text{Binomial} (n_{ij}, p_{ij})$ ($i = 1, \ldots, N; j = 0, 1, \ldots, 5$, corresponding to treatments A–F). For the outcome measure, we adopted the odds ratio (OR), i.e., $\theta_{ij} = \log \frac{p_{ij}(1 - p_{0j})}{p_{0j}(1 - p_{ij})}$ ($j = 1, 2, \ldots, 5$). The response rates of the reference treatment groups $p_{0j}$ were simulated from a uniform distribution, U(0.020, 0.436). The sample sizes of individual trials were set to be equal among all the arms within a trial (i.e., $n_{0j} = n_{1j} = n_{2j}$), and random integer values generated by a discrete uniform distribution on (30, 1, 126) were assigned, which is the range of actual sample sizes of individual trials of Siontis et al. 21 In addition, the random effects parameters $\theta_i$ were generated from a multivariate normal distribution MVN($\mu, \Sigma$). The grand mean parameter $\mu$ was set to be $\mu_1 = 1.057, \mu_2 = -0.362, \mu_3 = -0.508, \mu_4 = 1.415, \mu_5 = -1.317$, consistent with the REML estimate of the dataset of Siontis et al. 21 (see Section 7.2). For the heterogeneity covariance matrix $\Sigma$, we applied the equal
variance assumption $\tau^2 = \tau_1^2 = \cdots = \tau_5^2$, which implies equal correlation coefficients ($\kappa_{ij}$ s) of 0.50. The degree of heterogeneity was varied among $\tau = 0.10, 0.20, 0.40, 0.60, 0.80, 1.00, \text{ and } 1.20$. The parameter settings of 21 scenarios are presented in Figure 1. The response-rate parameters $p_{ij}$s of individual studies were identified by $\theta_{ij}$ and $p_0$.

From the generated individual binomial data, we calculated log-OR estimate $Y_i$ and its covariance matrix estimate $S_i$. We then calculated 95% prediction intervals for 5 contrasts on the network (A vs. B, C, D, E, and F) using the two proposed methods and the current standard ordinary $t$-approximation method. For the bootstrap method, we conducted 1,000 resamplings to calculate the bootstrap distributions. We also replicated the simulations 2,000 times for each scenario. Among the 2,000 simulations, we calculated the empirical estimates of coverage probabilities and expected widths of the 95% prediction intervals; the coverage probabilities were calculated by empirical rates of coverage events of randomly simulated $\theta_{XY}^{XY}$ from the $N(\mu^{XY}, \tau^2)$ according to the definition of the prediction interval.

The results of the simulations are presented in Figures 2 and 3. In general, the ordinary $t$-approximation method showed undercoverage performance, especially for small $N$. The undercoverage was moderated when $N = 18$. However, even under the $N = 18$ settings, the coverage probabilities were substantially less than the nominal level (95%) under certain scenarios (e.g., under $\tau = 0.20$ or 0.40).

For the bootstrap method, the coverage probabilities were generally larger than those obtained by the $t$-approximation. In particular, the bootstrap method led to conservative results, where the coverage probabilities were greater than 95% and the expected widths were relatively large. These results might indicate that the bootstrap method provides an excessively wide prediction interval; however, the relative expected widths did not
substantially differ from those obtained using the other two methods, especially when \( N = 14 \) and 18. Under a few scenarios, the bootstrap prediction interval was considerably undercoveraged, i.e., when \( \tau \) was small (0.10 or 0.20). In these cases, the expected widths could be wider than those obtained by the \( t \)-approximation. We confirmed that, under these scenarios, the widths of the bootstrap prediction intervals were generally wider than those obtained by the \( t \)-approximation; however, their widths were especially small in certain cases and these cases caused the undercoveraged results. The REML estimates of \( \tau \) were especially small (approximately 0) in these cases. Bootstrap resampling was then performed from an especially narrow range of the estimated random effects distributions with small \( \hat{\tau} \). This resampling led to the narrow prediction interval. To maintain the conservative property, an ad hoc strategy to circumvent the undercoverage property is to truncate \( \tau \) at a certain value (e.g., 0.10) for implementing the bootstrap resampling; however, this approach has a substantial limitation that does not involve the fixed-effect model.

For the Kenward–Roger-type adjustment, the coverage probabilities were also larger than those obtained using the \( t \)-approximation. In addition, the expected widths were larger than those of the \( t \)-approximation and adequately adjusted the undercoverage property by the higher-order asymptotic approximation. Furthermore, the coverage probabilities were closer to the nominal level (95%) and the expected widths were smaller than those of the bootstrap methods for many scenarios. Under small \( \tau \) scenarios, the coverage probabilities were less than 95% under several scenarios; however, the degrees were moderate compared with the bootstrap method for many scenarios. The coverage probabilities were equivalent to the ordinary \( t \)-approximation under some scenarios (e.g., scenarios 16–21 for the comparison A vs. F). However, under these settings, both methods exhibited adequate coverage probabilities (95%), and the two methods provide similar
prediction intervals in general.

7. Applications

7.1 Renal outcome of antihypertensive drugs

As an illustrative example, we applied the proposed methods to a network meta-analysis dataset to assess the effects of antihypertensive drugs on renal outcome in patients with diabetes. We considered doubling the serum creatinine level as the outcome. The authors of this study performed a network meta-analysis of 13 clinical trials (total participants: 25,055) comparing an angiotensin-converting enzyme (ACE) inhibitor, angiotensin-receptor blocker (ARB), calcium-channel blocker (CCB), β-blocker, combination of ACE inhibitor and diuretic, and a placebo. The network plot of this analysis is presented in Figure 4(A). We performed the network meta-analysis using the contrast-based random-effects model (*) for the log odds ratio and calculated the 95% prediction intervals via the ordinary t-approximation, bootstrap, and Kenward–Roger-type adjustments. The reference treatment was set to placebo.

Table 1 presents the results of the network meta-analysis. The number of bootstrap resampling was 3,600. The comparative OR estimates and 95% confidence intervals indicate that an ACE inhibitor was the best treatment to achieve the renal outcome. The estimate of the between-studies standard deviation by the REML method was 0.065 (log-OR scale). The 95% prediction intervals of the Kenward–Roger-type adjustment were narrower than those of the ordinary t-approximation. In addition, the 95% prediction intervals of the bootstrap method were wider than those of the ordinary t-approximation. In the simulation studies, the t-approximation method showed undercoverage properties under similar settings; thus, the prediction intervals might underestimate the uncertainty. In addition, the coverage rates of the Kenward–Roger-type adjustment were better than those of the t-approximation and the expected widths of these methods were comparable. The Kenward–Roger-type adjustment probably provides more accurate prediction
intervals. In addition, the 95% prediction intervals of the bootstrap method were wider than those of the $t$-approximation for some treatment pairs and were narrower for the others. In the simulation studies, the operating characteristics of the bootstrap method were somewhat unstable under similar settings. The Kenward–Roger-type adjustment would be recommended for such cases.

7.2 Noninvasive diagnostic modalities for detecting coronary artery disease
As a second example, we reanalyzed the network meta-analysis of Siontis et al. 21, who evaluated the comparative effectiveness of noninvasive diagnostic modalities for detecting coronary artery disease. Six treatments were compared using a dataset of 18 clinical trials ($n = 11,329$) for patients with low-risk acute coronary syndrome: coronary computed tomographic angiography (CCTA), cardiovascular magnetic resonance (CMR), exercise echocardiography (ECG), single-photon emission computed tomography–myocardial perfusion imaging (SPECT–MPI), standard care, and stress echocardiography. Here, we considered the subsequent downstream testing as the outcome, which requires additional diagnostic investigations because of test failure or diagnostic uncertainty. We adopted OR as the effect measure. The network plot is presented in Figure 4(B).

Table 2 presents the results of this network meta-analysis. The reference treatment was set to “CCTA.” The value of $\tau$ was estimated as 1.099 by the REML estimation (log-OR scale). The number of bootstrap resamples was set to 3,600. Overall, the widths of the 95% prediction intervals obtained by the bootstrap method were larger than those obtained by the other two methods. In particular, the upper limits of the prediction intervals were substantially larger for some comparisons, possibly because these limits were transformed by the exponential functions from the original scale (log-OR scale). Given the simulation results, the bootstrap prediction interval showed conservative properties under similar settings; thus, the interval might be too wide. However, the
validity was retained. In addition, the Kenward–Roger-type adjustment provided wider prediction intervals than the ordinary $t$-approximation for CCTA vs. exercise ECG, SPECT–MPI, and stress echo and also provided narrower prediction intervals for the remaining two comparisons. The simulation results indicated that both of these two cases possibly occur; however, the Kenward–Roger-type adjustment provided relatively adequate results in general; the coverage probabilities were closer to the nominal level. The proposed methods would provide more accurate predictive intervals for this real data example.

8. Discussion

Network meta-analysis has been widely used in systematic reviews of medical studies, and the resultant evidence has been adopted in public health, clinical practice, health technology assessment, and policy making. If misleading evidence has been reported as a result of the adoption of inadequate statistical methods, the influences in various applications might be enormous. The $t$-distribution approximation method has been the standard method for constructing the prediction interval and has been adopted in standard statistical packages for network meta-analysis (e.g., network graph of Stata). However, our simulation results indicate that the $t$-approximation method is possibly invalid under realistic situations in practice. For conventional pairwise meta-analysis, this problem has been well discussed and several alternative effective methods have been developed. However, no such effective methods have been developed for network meta-analysis. In addition, as shown in Section 6, this issue becomes more serious for network meta-analysis, where the statistical information is split between several treatment comparisons.
In the present work, we developed two new methods to construct prediction intervals for network meta-analysis. Both of these methods showed improved performance compared with the $t$-approximation and can be easily implemented using the R package \texttt{PINMA}. The bootstrap method requires a few minutes to calculate the prediction interval, but this requirement would not be problematic in practice. By contrast, the Kenward–Roger-type adjustment can be rapidly implemented via matrix computations. These two methods would be alternative standard methods for network meta-analysis.

One of the substantial limitations of the proposed methods is the undercoverage properties when $\tau$ is approximately 0. In the simulation studies, all three methods showed undercoverage properties in general. The instability of inferences for random-effects models under these settings is generally well known \cite{12,13}, and a similar issue would occur for the prediction problem. Notably, for the confidence intervals of the overall mean, overcoverage properties have been generally known \cite{33,34}; however, for the prediction intervals, undercoverage properties have also been reported in previous simulation studies \cite{9,10,11}. Nonetheless, the two proposed methods at least improve the coverage performance compared with that of the ordinary $t$-approximation method. If exact prediction methods are developed in the future, this issue will be completely resolved; however, a feasible solution to this problem might never be developed. The development of exact methods would be a difficult but interesting topic for future studies. Comparing these two methods, we recommend the Kenward–Roger approach over the bootstrap approach because the coverage probabilities were more stable and more closely spread around the nominal level; the bootstrap-based prediction intervals also showed low coverage probabilities when $\tau^2$ was close to 0.

Notably, for the bootstrap approach, we adopted the parametric bootstrap because the nonparametric bootstrap cannot be adequately applied straightforwardly to the
heterogeneous multivariate data with unequal covariance matrices. We also suggested truncating to 1.96 when the absolute value of the bootstrap quantiles is smaller than this threshold. This approach is similar to that proposed by Knapp and Hartung\textsuperscript{35}, which was later found to be overly conservative in mixed-effects meta-regression models\textsuperscript{36}. Because mixed-effects meta-regression models can be considered a special case of the network meta-analysis models, it would be reasonable for the bootstrap method to yield coverage probabilities clearly greater than the nominal level for many of the conditions depicted in the simulation studies. Morey et al.\textsuperscript{37} and Wiksten et al.\textsuperscript{38} have discussed related issues. We also adopted truncation of degree of freedom in the Kenward-Roger approach. Further theoretical researches would be needed for the properties of these truncation in future studies.

In conclusion, the two proposed methods would provide new effective tools for constructing better prediction intervals in the practice of network meta-analysis. Accurate prediction methods would improve the statistical evaluations of systematic reviews that use network meta-analysis.

**Acknowledgements**

This study was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (grant numbers: JP22H03554, JP22K19688).
References


**Table 1.** Results of network meta-analysis for the effects of antihypertensive drugs on renal outcome in patients with diabetes (comparative odds ratios) †.

<table>
<thead>
<tr>
<th>Placebo vs.</th>
<th>ACE inhibitor</th>
<th>ACE inhibitor + diuretic</th>
<th>ARB</th>
<th>β-blocker</th>
<th>CCB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate and 95% CI for overall mean</td>
<td>0.644</td>
<td>1.225</td>
<td>0.751</td>
<td>4.482</td>
<td>1.166</td>
</tr>
<tr>
<td></td>
<td>(0.476, 0.871)</td>
<td>(0.808, 1.856)</td>
<td>(0.619, 0.912)</td>
<td>(0.898, 22.37)</td>
<td>(0.886, 1.535)</td>
</tr>
<tr>
<td>95% prediction interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinary t-approximation</td>
<td>(0.419, 0.989)</td>
<td>(0.752, 1.996)</td>
<td>(0.571, 0.987)</td>
<td>(0.746, 26.92)</td>
<td>(0.830, 1.639)</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>(0.429, 0.966)</td>
<td>(0.663, 2.263)</td>
<td>(0.524, 1.076)</td>
<td>(0.890, 22.56)</td>
<td>(0.751, 1.810)</td>
</tr>
<tr>
<td>Kenward–Roger-type adjustment</td>
<td>(0.455, 0.896)</td>
<td>(0.803, 1.869)</td>
<td>(0.572, 0.987)</td>
<td>(0.880, 21.30)</td>
<td>(0.842, 1.607)</td>
</tr>
</tbody>
</table>

† ACE: angiotensin-converting enzyme; ARB: angiotensin-receptor blocker; CCB: calcium-channel blocker.
Table 2. Results of network meta-analysis for noninvasive diagnostic modalities for detecting coronary artery disease (comparative odds ratios) †.

<table>
<thead>
<tr>
<th>CCTA vs.</th>
<th>CMR</th>
<th>Exercise ECG</th>
<th>SPECT–MPI</th>
<th>Standard care</th>
<th>Stress echo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate and 95% CI for overall mean</td>
<td>2.878 (0.319, 25.94)</td>
<td>0.697 (0.141, 3.442)</td>
<td>0.602 (0.121, 2.982)</td>
<td>4.117 (1.740, 9.740)</td>
<td>0.268 (0.043, 1.677)</td>
</tr>
<tr>
<td>95% prediction interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordinary t-approximation</td>
<td>(0.097, 85.56)</td>
<td>(0.035, 13.72)</td>
<td>(0.030, 11.88)</td>
<td>(0.251, 67.57)</td>
<td>(0.012, 6.053)</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>(0.085, 94.66)</td>
<td>(0.031, 17.90)</td>
<td>(0.023, 17.46)</td>
<td>(0.271, 63.21)</td>
<td>(0.010, 9.423)</td>
</tr>
<tr>
<td>Kenward–Roger-type adjustment</td>
<td>(0.106, 78.01)</td>
<td>(0.030, 16.04)</td>
<td>(0.024, 15.30)</td>
<td>(0.273, 62.06)</td>
<td>(0.010, 7.251)</td>
</tr>
</tbody>
</table>

† CCTA: coronary computed tomographic angiography; CMR: cardiovascular magnetic resonance; ECG: electrocardiography; SPECT-MPI: single-photon emission computed tomography–myocardial perfusion imaging.
Figure 1. Network geometric graph and settings of the 21 scenarios of simulation studies.

<table>
<thead>
<tr>
<th>Design</th>
<th>$N = 10$</th>
<th>$N = 14$</th>
<th>$N = 18$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A vs. C</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A vs. D</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A vs. E</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>B vs. E</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C vs. E vs. F</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C vs. F</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>D vs. E</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 2. Simulation results: Coverage probabilities and expected widths of 95% prediction intervals (black: ordinary t-approximation; red: bootstrap; blue: Kenward–Roger-type adjustment).
Figure 3. Simulation results: Coverage probabilities and expected widths of 95% prediction intervals (black: ordinary t-approximation; red: bootstrap; blue: Kenward–Roger-type adjustment).
Figure 4. Network geometric graphs for the network meta-analyses of (A) antihypertensive drugs for renal outcome and (B) noninvasive diagnostic modalities for detecting coronary artery disease.
e-Appendix: Truncation of $\nu_{KR}^{XY}$ for the Kenward-Roger–type approximation

In Section 5.1, we proposed truncating the degree of freedom $\nu_{KR}^{XY}$ at 3 (if $\nu_{KR}^{XY} < 3$, it is rounded to 3) for the Kenward-Roger–type approximation. When $\nu_{KR}^{XY} \leq 1$, since the prediction interval cannot be defined, some truncations must be needed. Also, we confirmed the resultant interval can be quite wide and only provides vague information if $\nu_{KR}^{XY}$ is truncated at around 1. In e-Figures 1 and 2, we present the coverage probabilities and expected widths of the Kenward-Roger method truncating at 1.01, 3, 5 and 10 for the simulation studies in Section 6 (comparisons A vs. B and A vs. E).

At first, the expected widths of the Kenward-Roger method truncating at 1.01 were infinite for many scenarios. Second, the coverage probabilities of the Kenward-Roger method truncating at 10 was markedly smaller than the nominal level (0.95) and it would have undercoverage properties under wide settings. Third, the coverage probabilities of the Kenward-Roger method truncating at 3 and 5 would be acceptable levels, but the former showed larger coverage probabilities for many settings. Comparing the expected widths, the two methods were comparable except for a few scenarios (e.g., scenarios 1-3). Therefore, considering the validity of coverage probability and precision, we propose to adopt truncating at 3 for the Kenward-Roger–type approximation. It can be optionally changed by the analysts.
e-Figure 1. Simulation results for the Kenward-Roger method: comparison A vs. B. The degree of freedom $v_{KR}$ was truncated at 1.01, 3, 5, and 10.
e-Figure 2. Simulation results for the Kenward-Roger method: comparison A vs. E. The degree of freedom $\nu_{KR}$ was truncated at 1.01, 3, 5, and 10.
R example code

```r
# "Improved methods to construct prediction intervals for network meta-analysis"
# by Hisashi Noma, Yasuyuki Hamura, Shonosuke Sugawara and Toshi A. Furukawa

# R example code for constructing improved prediction intervals using the bootstrap
# and Kenward-Roger–type adjustment methods of network meta-analysis

# The "PINMA" package
# CRAN page: https://cran.r-project.org/web/packages/PINMA/

###

install.packages("PINMA") # install the "PINMA" package from CRAN
library("PINMA") # load the "PINMA" package

# Analyzing the example dataset "dstr" using the ordinary random-effects model

data(dstr) # load the example data
help(dstr)
attach(dstr)
edat <- data.edit(study, trt, d, n) # transforming the arm-level data to the
contrast-based summaries

y <- edat$y
S <- edat$S
tPI(y, S) # the ordinary t-approximation method for constructing prediction
# interval
help(tPI)
PBS(y, S) # the bootstrap method for constructing prediction interval
help(PBS)
KR(y, S) # the Kenward-Roger–type adjustment method for constructing
# prediction interval
help(KR)
```