Published in final edited form as:

Stat Methods Med Res. 2019 May ; 28(5): 1293–1310. doi:10.1177/0962280218754928.

Bayesian hierarchical methods for meta-analysis combining randomized-controlled and single-arm studies*

Jing Zhang[†], Chia-Wen Ko[‡], Lei Nie[‡], Yong Chen[§], Ram Tiwari[‡]

[†]Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, MD 20742, USA. Address correspondence to Jing Zhang, PhD, Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, MD 20742, USA.

[‡]Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA.

[§]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

Abstract

Meta-analysis of interventions usually relies on randomized controlled trials (RCTs). However, when the dominant source of information comes from single-arm studies, or when the results from RCTs lack generalization due to strict inclusion and exclusion criteria, it is vital to synthesize both sources of evidence. One challenge of synthesizing both sources is that single-arm studies are usually less reliable than RCTs due to selection bias and confounding factors. In this paper, we propose a Bayesian hierarchical framework for the purpose of bias reduction and efficiency gain. Under this framework, three methods are proposed: bivariate generalized linear mixed effects models (BGLMMs), hierarchical power prior (HPP) model and hierarchical commensurate prior (HCP) model. Design difference and potential biases are considered in all models, within which the HPP and HCP models further offer to downweight single-arm studies flexibly. The HCP model is recommended as the primary method for evidence synthesis because of its accuracy and robustness. We illustrate our methods by applying all models to two motivating datasets and evaluate their performance through simulation studies. We finish with a discussion of the advantages and limitations of our methods, as well as directions for future research in this area.

Keywords

Evidence synthesis; Different types of studies; Bias; Efficiency; Downweighting; Markov chain Monte Carlo

---

*This article reflects the views of the authors and should not be construed to represent the views or policies of the U.S. Food and Drug Administration.

jzhang86@umd.edu. Telephone: 301-405-3085.

1 Introduction

Evidence synthesis is becoming increasingly important because of the explosive growth of evidence in the scientific literature and the need for timely and informed decisions in public health and medicine¹. Meta-analysis, a systematic identification, appraisal, and statistical aggregation of all relevant evidence on a specified topic according to a predetermined and explicit method², commonly constitutes the main statistical component of evidence synthesis³.

However, the existing methods for meta-analysis usually restrict the synthesis to evidence from *randomized controlled trials* (RCTs) with at least two treatment arms, even when a wealth of evidence from single-arm studies exists. A common limitation of this strategy is that it may ignore important information and deviate the ultimate goal to make informed decision on the basis of *all* available evidence. In addition, good and conclusive evidence from RCTs may not even exist because some interventions are not suitable for randomization due to economic, ethical or operational consideration. For example, in the field of interventional cardiology or oncology, randomization is difficult but a large body of single-arm clinical databases and registries are often available⁴⁵. Furthermore, RCTs are usually conducted with very strict inclusion and exclusion criteria of patient population, where children, elderly, and patients with comorbidity are usually excluded, thus can lack generalization⁶⁷. In short, exploiting data from single-arm trials is of great interest in these cases and methods synthesizing both evidence are essential⁸.

Synthesis of RCTs and single-arm studies poses new challenges, such as the need to account for the difference between these two designs and the inherent biases from single-arm studies⁹. The main appeal of RCTs in health care comes from its potential to balance unknown prognostic factors at baseline, while no other study design offers such power, which is the whole premise for conducting RCTs in the first place³. As single-arm study outcomes are usually interpreted relative to historical controls, they are subject to selection bias in terms of patient selection and treatment allocation, and other potential confounding factors. To the best of our knowledge, few papers have sensibly synthesized RCTs and single-arm studies. Begg and Pilote¹⁰ and Li and Begg¹¹ proposed models to combine results from controlled and uncontrolled studies using maximum likelihood estimators in a frequentist framework, where inherent biases from uncontrolled studies were ignored. Begg and Pilote¹⁰ further suggested a preliminary analysis to test for bias before using their model. Schmitz et al.⁷ and Soares et al.¹² proposed Bayesian alternatives to combine controlled and uncontrolled studies, but single-arm studies were either excluded or matched with a control arm from other trials. On the other hand, since there are good RCTs and poorly conducted RCTs and the situation is the same for other types of studies, some works³¹³¹⁴ suggest considering the risk of bias of each study in the synthesis, randomized or not, based on large databases of empirical data or expert opinion, which we believe is impractical.

In this article, we propose a Bayesian framework to flexibly synthesize evidence coming from RCTs and single-arm studies and to adaptively adjust for the difference between designs and potential biases. Under this framework, we provide three methods, all of which

adjust for design difference and potential biases. In addition, two of these methods further provide strategies to downweight single-arm studies. The proposed models are developed under the *arm-based* (AB) framework¹⁵¹⁶ where absolute treatment effects are assumed to be exchangeable across studies. It has been criticized that this assumption is stronger than the assumption of exchangeability of relative treatment effects in the *contrast-based* (CB) framework. However, the AB framework does offer advantages in ease of model fitting and interpretation when single-arm studies are present¹⁵. The pros and cons of the arm-based and contrast-based frameworks have been compared in published works¹⁷¹⁸¹⁹²⁰¹⁶²¹²² and will also be discussed in this paper. All the three models proposed are expected to gain precision by including single-arm studies compared with a meta-analysis of solely RCTs. In addition, the Bayesian framework provides a straightforward way to identify and address the problems brought by distinct studies²³ as well as by small sample sizes that prevail in meta-analysis.

This article is organized as follows. We begin by introducing our two motivating datasets in Section 2. We then present our proposed methods in Section 3, and apply them to the two motivating datasets in Section 4. The performance of our methods is evaluated through extensive simulations in Section 5. We close with some concluding remarks and suggested avenues for future research in Section 6.

2 Motivating Data

2.1 Deep vein thrombosis (DVT) data

Low molecular weight heparins (LMWH) is usually indication for prophylaxis of deep vein thrombosis (DVT) in patients with severely restricted mobility after major orthopedic surgery. However, LMWH has been associated with the risk of serious or fatal hemorrhagic adverse events. In addition, the need for parenteral or subcutaneous administration is another disadvantage. In the past decade, research has focused on finding new active drugs that can be given orally and can improve the benefit and risk profile. In trials to develop new active drugs, the main objective is to demonstrate that these new drugs are non-inferior to LMWH in prophylaxis of DVT, which relies on a reasonable non-inferiority margin. To determine the margin, the treatment effect of LMWH compared with placebo has to be estimated as the first step. Thus we collected RCTs comparing LMWH versus placebo and single-arm studies of LMWH to facilitate the determination.

As is shown in Table 1, this dataset includes 12 studies, 7 of which are RCTs comparing LMWH versus placebo and the remaining 5 are single-arm studies of LMWH. Let N represent the sample size (i.e., the number of patients assigned in each treatment arm), y represent the number of DVT events, and $\%$ represent the corresponding event probability. The author and year of publication for each study are also presented. Ignoring the 5 single-arm studies can lead to a substantial loss of information.

2.2 Advanced follicular lymphoma (FL) data

Many pharmaceutical companies are currently developing biosimilar products against the US-licensed product Rituxan (i.e., rituximab), which has oncology indications in patients

with advanced follicular lymphoma (FL). Rituximab is usually used as an add-on therapy to cyclophosphamide, vincristine, and prednisone (CVP) in the advanced FL patient population. However, it is hard to achieve a reliable estimation on the treatment effect of rituximab as an add-on therapy due to the fact that there is only one RCT conducted since 1997 comparing rituximab+CVP (R-CVP) versus CVP, and that all the other available evidence is from single-arm studies of either R-CVP or CVP alone. Accordingly designing an equivalence trial to support the similarity or equivalence between the biosimilar product and rituximab is complicated. In this sense, rigorous methods that can combine both sources of evidence are essential and may help the regulatory agencies address their commitment to best practice for the conduct of meta-analysis.

We collected all available data through literature search from 1997, the year that rituximab was approved, up to 2014. Table 2 shows our assembled dataset of 12 studies including 1 RCT and 11 single-arm studies, to compare R-CVP versus CVP alone. The author and year for each study, the sample size N , the number of responses y , and the response probability % are reported. In this example, as all studies but one are single-arm, it is evident that excluding these single-arm studies will hinder well-informed and comprehensive decisions.

3 Statistical Methods

Without loss of generality, we consider a setting of meta-analysis of RCTs with two arms. The extension to multiple-arm RCTs will be discussed in Subsection 3.4. Suppose there are totally M studies, n of which are RCTs comparing Treatments 1 and 2, m of which are single-arm studies containing only Treatment 1, and l of which are single-arm studies containing only Treatment 2. Note that $M = n + m + l$. Let N_{1i} and y_{1i} denote the sample size and number of events for Treatment 1 in the i^{th} study of the n RCTs, and let N_{2i} and y_{2i} denote these quantities for Treatment 2, where $i = 1, \dots, n$. Let N_{3i} and y_{3i} represent the sample size and number of events for Treatment 1 in the i^{th} study among the m single-arm studies, where $i = n + 1, \dots, n + m$, and let N_{4i} and y_{4i} be the corresponding quantities for Treatment 2 in the i^{th} study among the l single-arm studies, where $i = n + m + 1, \dots, n + m + l$. Further, let p_{1i} and p_{2i} denote the event probabilities for Treatments 1 and 2 in the n RCTs, let p_{3i} denote the event probability for Treatment 1 in the m single-arm studies, and let p_{4i} denote that for Treatment 2 in the l single-arm studies.

We assume that the numbers of events in these $(m + n + l)$ studies follow independent binomial distributions as follows

$$\begin{aligned} y_{1i} &\sim \text{Bin}(N_{1i}, p_{1i}), & i = 1, \dots, n \\ y_{2i} &\sim \text{Bin}(N_{2i}, p_{2i}), & i = 1, \dots, n \\ y_{3i} &\sim \text{Bin}(N_{3i}, p_{3i}), & i = n + 1, \dots, n + m \\ y_{4i} &\sim \text{Bin}(N_{4i}, p_{4i}), & i = n + m + 1, \dots, n + m + l. \end{aligned} \quad (1)$$

Under (1), we propose three Bayesian hierarchical methods to model the event probabilities p_{1i} , p_{2i} , p_{3i} , and p_{4i} and discuss their respective strengths and weaknesses in the following subsections.

3.1 Bivariate generalized linear mixed effects models

Bivariate generalized linear mixed effects models (BGLMMs) are proposed by adaptation of Zhang et al.¹⁵. We start with **BGLMM (I)**, formulated as,

$$\begin{aligned} g(p_{1i}) &= \mu_1 + \nu_{1i}, \\ g(p_{2i}) &= \mu_2 + \nu_{2i}, \quad (\nu_{1i}, \nu_{2i}) \sim N(\mathbf{0}, \Sigma), \quad i = 1, \dots, n \\ g(p_{3i}) &= \mu_1 + \nu_{3i}, \quad \nu_{3i} \sim N(0, \sigma_3^2), \quad i = n + 1, \dots, n + m \\ g(p_{4i}) &= \mu_2 + \nu_{4i}, \quad \nu_{4i} \sim N(0, \sigma_4^2), \quad i = n + m + 1, \dots, n + m + l, \end{aligned} \quad (2)$$

where μ_1 and μ_2 are the overall effects for Treatments 1 and 2, respectively; (ν_{1i}, ν_{2i})

follows a bivariate normal distribution with a variance-covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

accounting for the correlation between two treatment arms and heterogeneity across n RCTs; and ν_{3i} and ν_{4i} follow normal distributions accounting for the heterogeneities across the m and l single-arm studies, respectively. Here $g(\cdot)$ is some proper link function. The probit link $\Phi^{-1}(\cdot)$ is recommended because it enables a closed-form expression for the population-averaged treatment-specific event probabilities, i.e.,

$$\pi_k = E(p_{ik} | \mu_k, \sigma_k) = \int_{-\infty}^{+\infty} \Phi(\mu_k + \nu_{ki}) \phi(z) dz = \Phi\left(\frac{\mu_k}{\sqrt{1 + \sigma_k^2}}\right) \text{ for } k = 1, 2^{24}. \Phi \text{ and } \phi \text{ are the}$$

cumulative distribution function and probability density function of the standard normal distribution. Then the marginal *odds ratio* ($OR_{21} = \pi_2(1 - \pi_1)/\{(1 - \pi_2)\pi_1\}$), *risk difference* ($RD_{21} = \pi_2 - \pi_1$) and *risk ratio* ($RR_{21} = \pi_2/\pi_1$) can be calculated correspondingly¹⁵. Since noninformative improper priors may lead to improper posteriors, we use their proper approximations, i.e., vague proper priors. Specifically, a $N(0, 1000)$ prior is assigned to μ_1 and μ_2 , and an inverse-gamma prior $IG(0.001, 0.001)$ is assigned to σ_3^2 and σ_4^2 . Here the $IG(0.001, 0.001)$ prior (with density plot showed in the appendix) is an approximation to the Jeffreys prior²⁵²⁶²⁷. An inverse-Wishart ($R, 2$) prior can be used for the covariance matrix Σ , where R is a 2×2 matrix with diagonal elements 1 and off-diagonal elements 0.005. This corresponds to the 95% Bayesian credible intervals (CIs) for the standard deviation parameters ranging from 0.45 to 33.56, and the 95% CIs for the correlation parameters ranging from -1.00 to 1.00, which together imply that the population-averaged treatment-specific event probability ranges from 0 to 1. BGLMM (I) allows different levels of heterogeneity for single-arm studies and RCTs, but assumes the same overall treatment effects for the two types of studies to enable borrowing information for estimation.

The BGLMM (I) can be extended by allowing different overall treatment effects parameters μ_3 and μ_4 for single-arm studies; namely

$$\begin{aligned} g(p_{3i}) &= \mu_3 + \nu_{3i}, \quad \nu_{3i} \sim N(0, \sigma_3^2), \quad i = n + 1, \dots, n + m \\ g(p_{4i}) &= \mu_4 + \nu_{4i}, \quad \nu_{4i} \sim N(0, \sigma_4^2), \quad i = n + m + 1, \dots, n + m + l. \end{aligned} \quad (3)$$

Here similar vague proper priors can be assigned to μ_3, μ_4, σ_3^2 and σ_4^2 . To allow borrowing information between single-arm studies and RCTs, we can use the estimated posterior medians $\hat{\mu}_3$ and $\hat{\mu}_4$ from (3) to form a prior distribution for the overall treatment effect parameters μ_1 and μ_2 in model $g(p_{1i}) = \mu_1 + \nu_{1i}$, $g(p_{2i}) = \mu_2 + \nu_{2i}$ and $(\nu_{1i}, \nu_{2i}) \sim N(\mathbf{0}, \Sigma)$, for $i = 1, \dots, n$. More specifically, $\mu_1 \sim N(\hat{\mu}_3, \tau_1^2)$ and $\mu_2 \sim N(\hat{\mu}_4, \tau_2^2)$, where τ_1^2 and τ_2^2 represent a priori belief regarding the possible extent of the biases from the single-arm studies. A sensible choice for τ_1^2 and τ_2^2 are $\hat{var}(\mu_3)$ and $\hat{var}(\mu_4)$, implying that the single-arm studies are used at the face value. In order to account for additional variability, τ_1^2 and τ_2^2 can be expressed as $\tau_1^2 = \hat{var}(\mu_3) + \zeta_1^2$ and $\tau_2^2 = \hat{var}(\mu_4) + \zeta_2^2$, where prior distributions for ζ_1^2 and ζ_2^2 are usually based on empirical data including existing evidence and expert opinion on the difference between the two types of studies. Alternatively, we can define $\tau_1^2 = \hat{var}(\mu_3) \cdot c_1$ and $\tau_2^2 = \hat{var}(\mu_4) \cdot c_2$, where c_1 and c_2 are independent and follow an inverse gamma distribution. In this paper, we use the face values for illustration. This model is referred to as the **BGLMM (II)**.

On the other hand, the BGLMM (I) can be simplified as $g(p_{1i}) = \mu_1 + \nu_{1i}$, $\nu_{1i} \sim N(0, \sigma_1^2)$, $i = 1, \dots, n + m$ and $g(p_{2i}) = \mu_2 + \nu_{2i}$, $\nu_{2i} \sim N(0, \sigma_2^2)$, $i = 1, \dots, n, n + m + 1, \dots, n + m + l$. We call this model *naive pooling* (**NP**) (**I**). This model assumes that all available information is from single-arm studies, synthesizes information for two treatment arms separately, and as such ‘breaks randomisation’. Another simplification is $g(p_{1i}) = \mu_1 + \nu_{1i}$, $g(p_{2i}) = \mu_2 + \nu_{2i}$ with $(\nu_{1i}, \nu_{2i}) \sim N(\mathbf{0}, \Sigma)$, $i = 1, \dots, n + m + l$. We name it **NP (II)**. In this model, missing entries are imputed automatically with Markov chain Monte Carlo (MCMC) algorithm and the correlation between two arms is incorporated as well. Note that neither the design difference between the two study types nor the potential biases from single-arm studies are considered in naive pooling models.

3.2 Hierarchical power prior (HPP) model

Power priors, introduced by Ibrahim and Chen²⁸, offer a simple way to incorporate and downweight historical data by raising the historical likelihood to a power $\alpha_0 \in [0, 1]$ and re-standardizing the result to a proper distribution. Along the lines of this work, we can raise the likelihoods of the m and l single-arm studies to powers $\alpha_{01} \in [0, 1]$ and $\alpha_{02} \in [0, 1]$, respectively, and specify priors on α_{01} and α_{02} , to incorporate and downweight information.

Based upon the BGLMM (I), we build up the model as follows. For simplicity of notation, we let θ denote the parameters of interest in the BGLMM (I), i.e., $\theta = (\mu_1, \mu_2, \Sigma, \sigma_3, \sigma_4)$, and let α_0 denote the power parameters, i.e., $\alpha_0 = (\alpha_{01}, \alpha_{02})$. Let $\mathbf{y}_1 = (y_{11}, \dots, y_{1n})$, $\mathbf{y}_2 = (y_{21}, \dots, y_{2n})$, $\mathbf{y}_3 = (y_{3,n+1}, \dots, y_{3,n+m})$, and $\mathbf{y}_4 = (y_{4,n+m+1}, \dots, y_{4,n+m+l})$. Then the joint prior for (θ, α_0) is as follows

$$\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}_0) \propto \left\{ \prod_{i=n+1}^{n+m} f(y_{3i} | \mu_1, \sigma_3^2) \right\}^{\alpha_{01}} \left\{ \prod_{i=n+m+1}^{n+m+l} f(y_{4i} | \mu_2, \sigma_4^2) \right\}^{\alpha_{02}} \times \pi_0(\boldsymbol{\theta}) \pi_0(\boldsymbol{\alpha}_0), \quad (4)$$

where $\pi_0(\boldsymbol{\theta})$ denotes the initial joint priors for the parameters of interest and $\pi_0(\boldsymbol{\alpha}_0)$ denotes the joint priors for the power parameters $\boldsymbol{\alpha}_0 = (\alpha_{01}, \alpha_{02})$. A natural prior choice for $\pi_0(\alpha_{01})$ and $\pi_0(\alpha_{02})$ is *Beta*(a, b) with positive hyperparameters a and b . These hyperparameters represent the priori belief of the degree of borrowing from the single-arm studies: if $a \gg b$, the model strongly encourages borrowing; if $a \ll b$, it strongly discourages borrowing. Note that by specifying a prior for $\boldsymbol{\alpha}_0$ rather than directly assigning specific numerical values to $\boldsymbol{\alpha}_0$, we allow more flexibility in weighting single-arm studies²⁸. However, it has a tendency to downweight the single-arm studies overmuch, therefore we use a *Beta*(10, 1) prior. This prior distribution has a mean value of 0.91 and a 95% CI from 0.69 to 0.99 indicating a priori that there is moderate to strong similarity between single-arm studies and RCTs, which is the underlying reason for synthesizing the two types of studies. Here a truncated gamma prior or a truncated normal prior can also work well as they have similar theoretical, computational and practical properties as a *beta* prior. More prior specifications for the power parameters have been discussed in several papers^{29,30,31}. Gamalo et al.³² further proposed an ‘order restricted power prior’, which incorporated dependence of the powers but was more suitable when combining information from multiple historical studies.

Finally the joint posterior distribution for $(\boldsymbol{\theta}, \boldsymbol{\alpha}_0)$ is proportional to

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{\alpha}_0 | y_1, y_2, y_3, y_4) &\propto \prod_{i=1}^n f(y_{1i}, y_{2i} | \mu_1, \mu_2, \boldsymbol{\Sigma}) \\ &\times \left\{ \prod_{i=n+1}^{n+m} f(y_{3i} | \mu_1, \sigma_3^2) \right\}^{\alpha_{01}} \left\{ \prod_{i=n+m+1}^{n+m+l} f(y_{4i} | \mu_2, \sigma_4^2) \right\}^{\alpha_{02}} \\ &\times \pi_0(\boldsymbol{\theta}) \pi_0(\boldsymbol{\alpha}_0), \end{aligned} \quad (5)$$

where $\prod_{i=1}^n f(y_{1i}, y_{2i} | \mu_1, \mu_2, \boldsymbol{\Sigma})$, $\prod_{i=n+1}^{n+m} f(y_{3i} | \mu_1, \sigma_3^2)$ and $\prod_{i=n+m+1}^{n+m+l} f(y_{4i} | \mu_2, \sigma_4^2)$ are the likelihood functions specified in the BGLMM (I) for the n RCTs, and m and l single-arm studies, respectively. We refer to model (5) as the *hierarchical power prior* (HPP) model.

The HPP model allows us to downweight single-arm studies through power parameters defined on the interval [0, 1], but it does not directly parameterize the agreement of information from the two types of studies. In addition, the HPP tends to over-attenuate the impact of the single-arm studies and thus needs fairly large power parameters to ensure sufficient borrowing.

3.3 Hierarchical commensurate prior (HCP) model

As a solution to the problems in the HPP model, we propose a *hierarchical commensurate prior* (HCP) model to offer further flexibility in downweighting single-arm studies and to enable the measurement of commensurability, i.e., the degree to which the two sources of information are commensurate. The core idea of this method is based on a work by Hobbs et al.³³.

We assume different mean treatment effect parameters for the two types of studies (i.e., μ_1 and μ_2 for RCTs and μ_3 and μ_4 for single-arm studies). This dichotomous parameterization allows us to measure the degrees of commensurability between μ_1 and μ_3 and between μ_2 and μ_4 . More specifically, we let μ_1 and μ_2 follow normal distributions with means μ_3 and μ_4 and precisions τ_1 and τ_2 , where τ_1 and τ_2 parameterize commensurability. Now the joint commensurate prior is formulated as follows

$$\pi(\theta, \tau_1, \tau_2) \propto \prod_{i=n+1}^{n+m} f(y_{3i} | \mu_3, \sigma_3^2) \prod_{i=n+m+1}^{n+m+l} f(y_{4i} | \mu_4, \sigma_4^2) \times N(\mu_1 | \mu_3, 1/\tau_1) N(\mu_2 | \mu_4, 1/\tau_2) \pi_0(\theta), \quad (6)$$

where $\theta = (\mu_1, \mu_2, \Sigma, \sigma_3, \sigma_4)$ and $\pi_0(\theta)$ represents the initial priors. Vague priors are assigned again, namely a $N(0, 1000)$ prior to μ_3 and μ_4 and a $\text{Gamma}(0.01, 0.01)$ prior to τ_1 and τ_2 . Note that other prior specifications of τ_1 and τ_2 can also be adopted here; for example, a ‘spike and slab’ distribution^{33,34}. Larger values of τ_1 and τ_2 indicate increased commensurability and induce increased borrowing of strength from the single-arm data. Specifically, if τ_1 and τ_2 approach zero, then $\pi(\theta, \tau_1, \tau_2)$ approaches $\pi_0(\theta)$, effectively ignoring the single-arm studies; if τ_1 and τ_2 approach infinity, then $\pi(\theta, \tau_1, \tau_2)$ approaches $\prod_{i=n+1}^{n+m} f(y_{3i} | \mu_1, \sigma_3^2) \prod_{i=n+m+1}^{n+m+l} f(y_{4i} | \mu_2, \sigma_4^2) \times \pi_0(\theta)$, equivalent to pooling the RCTs and single-arm studies together without downweighting.

The joint posterior distribution is formulated as

$$\begin{aligned} \pi(\theta, \tau_1, \tau_2 | y_1, y_2, y_3, y_4) &\propto \prod_{i=1}^n f(y_{1i}, y_{2i} | \mu_1, \mu_2, \Sigma) \quad (7) \\ &\times \prod_{i=n+1}^{n+m} f(y_{3i} | \mu_3, \sigma_3^2) \prod_{i=n+m+1}^{n+m+l} f(y_{4i} | \mu_4, \sigma_4^2) \\ &\times N(\mu_1 | \mu_3, 1/\tau_1) N(\mu_2 | \mu_4, 1/\tau_2) \pi(\theta), \end{aligned}$$

where $\prod_{i=1}^n f(y_{1i}, y_{2i} | \mu_1, \mu_2, \Sigma)$ is based on $g(p_{1i}) = \mu_1 + v_{1i}$, $g(p_{2i}) = \mu_2 + v_{2i}$, and $(v_{1i}, v_{2i}) \sim N(\mathbf{0}, \Sigma)$, for $i = 1, \dots, n$.

The HCP model offers more flexibility in downweighting single-arm studies according to the degree of commensurability between single-arm studies and RCTs. The more commensurate the two sources of information are, the more information the HCP model borrows from the single-arm studies.

The Deviance Information Criterion (DIC)³⁵ and residual deviance $\bar{D}_{res}$ are reported. The DIC is defined by $\bar{D}(\theta) + p_D$, where $\bar{D}(\theta) = E_{\theta | \text{Data}} [D(\theta)]$ is the posterior expectation of the deviance, and $p_D = \bar{D}(\theta) - D(\bar{\theta})$ is the effective number of model parameters. Here $\bar{\theta}$ refers to the posterior expectation of θ , i.e., $E_{\theta | \text{Data}}(\theta)$. DIC rewards better fitting models through $\bar{D}(\theta)$ and penalizes more complex models through p_D . $\bar{D}_{res}$ is defined as the posterior mean of the deviance under the current model, minus the deviance for the saturated model. All models can be implemented via MCMC methods using the WinBUGS software³⁷. A generous burn-in period of 500,000 iterations was used, with 500,000 subsequent iterations retained for posterior treatment effect estimates. Two parallel chains were generated using different initial values. The convergence of Markov chain Monte Carlo (MCMC) chains can be assessed by visual inspection of the chains and by the Gelman-Rubin convergence statistic³⁸.

3.4 Extension to settings with more than two treatments

In previous subsections, we have focused on comparisons of two treatments. The proposed methods can be adapted to cases where simultaneous comparisons of multiple treatments are of interest. There are two types of scenarios: (1) when we have multiple-arm RCTs; and (2) when we want to compare multiple treatments but no trial has simultaneously compared them. Multiple-arm RCTs are usually conducted, for instance, when two different doses of Treatment A are compared with control B or when Treatment A is compared with previous gold-standard Treatment B and control C. This seems to be the natural outgrowth of our methods and does not require additional modeling skills. For the second scenario, ‘network meta-analysis’¹⁵¹⁸¹⁶³⁹⁴⁰ can be applied which combines evidence from both direct comparisons of treatments within RCTs and indirect comparisons across trials through a common comparator. This technique is increasingly popular recently in evidence-based medicine and comparative effectiveness research. To apply this technique, inconsistency between direct and indirect comparisons needs special attention in addition to heterogeneity between designs. Our proposed models will also need to be tailored and refined. Finally, our Bayesian framework eases the inferential burden for both scenarios where large covariance matrices are usually involved. In contrast, frequentist inference may encounter convergence problems because the (restricted) maximum likelihood estimate of the between-study correlation is often at or very close to boundary of its parameter space and thus the estimated covariance matrix is singular⁴¹⁴².

4 Data Analysis

In this section, we apply the statistical methods proposed in previous section to the motivating datasets introduced in Section 2 and compare the results. In the first example, the dataset contains multiple RCTs and multiple single-arm studies; whereas in the second example, all studies but one are single-arm.

4.1 DVT data

In this dataset, the primary outcome is the occurrence of DVT event, thus a lower event probability implies a better treatment. Table 3 presents the synthesis results from the proposed models of the 12 studies comparing LMWH versus placebo. Here π_1 and π_2 denote the population-averaged DVT event probabilities for LMWH and placebo, respectively. OR_{21} , RD_{21} and RR_{21} show the relative treatment effect of placebo versus LMWH. In Table 3, posterior medians and 95% posterior CIs of these parameters are reported. The posterior medians of π_1 are 0.16, 0.17, 0.17, 0.18, 0.19, and 0.19 from the NPs (I) - (II), BGLMMs (I) - (II), HPP and HCP, respectively; and those for π_2 are 0.50, 0.50, 0.50, 0.49, 0.50 and 0.50. It seems that the estimated population-averaged event probabilities of LMWH are slightly larger from more complex models. The relative summaries of the comparison of placebo versus LMWH (i.e., OR_{21} , RD_{21} , and RR_{21}) are smaller in more complex models accordingly. Let us take OR_{21} as an example. It is 5.29 in the NP (I), 4.88 in the NP (II), 4.65 in the BGLMM (I), 4.54 in the BGLMM (II), 4.28 in the HPP and 4.33 in the HCP. It seems that the HPP and HCP models are more conservative than the other models, and the BGLMMs (I) and (II) are more conservative than the NPs (I) and (II). Note that though there are marginal differences in magnitude of the estimated values, all models concur that LMWH is statistically significantly more efficacious than placebo in the prevention of DVT.

Though the BGLMM (I) shows a better fit with $\bar{D}_{res} = 18.27$, the DIC values from the BGLMM (I) and HCP models are comparable (123.76 and 123.83, respectively), implying similar goodness of fit after penalizing model complexity. In addition, the two DIC values are smaller than those from the other models. We choose the HCP as our final model because it is the most flexible model in terms of downweighting information from single-arm studies. We now present the results from the HCP model. Patients taking placebo have a higher population-averaged DTV event probability (0.50; 95% CI = 0.34 to 0.67) than patients taking LMWH (0.19; 95% CI = 0.11 to 0.31). The OR_{21} , RD_{21} and RR_{21} are 4.33 (95% CI = 1.92 to 9.96), 0.31 (95% CI = 0.14 to 0.48) and 2.65 (95% CI = 1.54 to 4.54), respectively. Note that the DIC value from the BGLMM (II) is calculated based solely on the data from RCTs, thus it is not presented in Table 3.

To visually compare different models, Figure 1 depicts the posterior medians and 95% CIs of all parameters. We observe that none of the 95% CIs of RD_{21} from these six models intersect the dashed line $x = 0$, and none of the 95% CIs of OR_{21} and RR_{21} intersect the dashed line $x = 1$, indicating statistically significant results from all the models. We also observe that the lengths of the intervals from the HPP and HCP modes are comparable to

those from the other models, thus these two models do not lose much efficiency when available information are sufficient though they have more parameters to estimate.

Prior sensitivity analysis is conducted for the HPP and HCP models with the results presented in Table 4. We start with the HCP model. The first four investigate the sensitivity of μ_1 and μ_2 to more informative normal distributions, namely $N(0, 100)$ and $N(0, 10)$, and heavy-tailed t -distributions with degrees of freedom 2 and 5. The estimated π_1 and π_2 are all close to those from the HCP in Table 3, thus the HCP model is robust to these other prior distributions. The following four explore the sensitivity of commensurate parameter τ to more and less informative Gamma priors including $\text{Gamma}(0.1, 0.1)$, $\text{Gamma}(0.001, 0.001)$, $\text{Gamma}(1, 0.5)$ and $\text{Gamma}(3, 1)$. It seems that the HCP is not particularly sensitive to these prior distributions either. Next, Wishart priors $W(I_2, 2)$ and $W(R^{-1}, 2)$ with R being a 2×2 matrix with diagonal elements equal to 0.1 and off-diagonal elements equal to 0.005 are used to assess the sensitivity of the precision matrix Σ^{-1} . Comparison of Tables 3 and 4 suggests that more informative Wishart priors for Σ^{-1} do not have substantial effect on the estimates. Finally, prior sensitivity analysis for α_0 in the HPP is conducted using $\text{Beta}(5, 1)$, $\text{Beta}(3, 2)$, $\text{Beta}(1, 1)$, $\text{Beta}(2, 2)$ and $\text{Beta}(2, 5)$ as priors and they all yield similar results.

4.2 Advanced FL data

We now present the synthesis results for the advanced FL dataset in Table 5. Here π_1 and π_2 denote the estimates of the population-averaged response probabilities for CVP and R-CVP, respectively. They are 0.64 and 0.85 in the NP (I) and 0.63 and 0.84 in the NP (II). The BGLMMs (I) and (II) obtain slightly smaller estimates; namely, $\pi_1 = 0.61$ and $\pi_2 = 0.79$ in both models. In BGLMM (II)*, $\mu_1 \sim N(\hat{\mu}_3, \hat{v}ar(\mu_3) \cdot c_1)$ and $\mu_2 \sim N(\hat{\mu}_4, \hat{v}ar(\mu_4) \cdot c_2)$, where $c_1, c_2 \sim IG(2, 3)$. Similar results to those from the BGLMM (II) are obtained. The HPP and HCP models perform comparably; more specifically, $\pi_1 = 0.60$ and $\pi_2 = 0.77$ in the HPP model, and $\pi_1 = 0.59$ and $\pi_2 = 0.76$ in the HCP model.

In terms of relative summaries, the NPs (I) and (II) produce statistically significant results and suggest that R-CVP is better than CVP alone. For example, for NP (I), OR_{21} has a posterior median 3.25 with a 95% CI (1.81, 6.05); RD_{21} has a posterior median 0.21 with a 95% CI (0.10, 0.35); and RR_{21} has a posterior median 1.33 with a 95% CI (1.13, 1.70). However, the BGLMM (I), BGLMM (II), HPP and HCP do not show statistically significant results in the comparison of R-CVP versus CVP; namely, the 95% CIs for OR_{21} and RR_{21} from these models cover 1 and those for RD_{21} cover 0. $\bar{D}_{res}$ values are similar from different models. The DIC disfavors the HPP model, but does not differentiate the NP (I), NP (II), BGLMM (I), and HCP model. Again the DIC value from BGLMM (II) is not presented here.

Figure 2 provides visual illustration to augment Table 5. Under the BGLMM (I), BGLMM (II), HPP and HCP models, the 95% CIs for RD_{21} pass through the dashed line $x = 0$, and those for OR_{21} and RR_{21} pass through the dashed line $x = 1$. However, under the NPs (I) and (II), the 95% CIs for RD_{21} and those for OR_{21} and RR_{21} do not intersect the lines $x = 0$ and $x = 1$, respectively. This conflict may be due to possible over-parameterization of the BGLMM (I), BGLMM (II), HPP and HCP models under this extremely sparse dataset which

contains only 1 RCT. Accordingly, there may have been insufficient power making inference from these models. If however we had multiple RCTs as in the DVT dataset, these models would make perfect sense and were expected to outperform the NPs (I) and (II). We also observe that the HPP and HCP models obtain wider 95% CIs, which we believe is again due to the fact that the data are very sparse whereas the two models have additional parameters to estimate (i.e., α_{01} and α_{02} in the HPP model, and τ_1 and τ_2 in the HCP model).

5 Simulations

In this section we evaluate the performance of the proposed models via simulations under various scenarios. Relative biases (i.e., biases / true values) and coverage probabilities of the 95% CI are used to compare point and interval estimates from these models. 1000 replicates of simulations are conducted for all scenarios.

5.1 Simulation Setting

5.1.1 Simulation 1—We generate a pool of evidence comprising 30 studies to compare 2 treatments under 5 scenarios. Figure 3 shows the data structures under these 5 scenarios labeled as S1–S5. In S1, $n = 15$, $m = 10$, and $l = 5$; in S2, $n = m = l = 10$; in S3, $n = 5$, $m = 10$, and $l = 15$; in S4, $n = 3$, $m = 12$, and $l = 15$; and in S5, $n = 1$, $m = 15$, and $l = 14$. Here n represents the number of RCTs, m represents the number of single-arm studies for Treatment 1, and l represents the number of single-arm studies for Treatment 2. The 5 scenarios enable us to evaluate our methods under both balanced (i.e., S2) and unbalanced cases (i.e., S1, 3, 4, and 5). For the unbalanced case, we investigate both the case where majority of studies are RCTs (i.e., S1) and the case where majority of studies are single-arm studies (i.e., S3–5), which simulate the motivating data sets.

For RCTs, event probabilities p_{1i} and p_{2i} in Trial i are generated according to $\Phi(p_{1i}) = \mu_1 + \nu_{1i}$, $\Phi(p_{2i}) = \mu_2 + \nu_{2i}$, $(\nu_{1i}, \nu_{2i}) \sim N\left(0, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$ where $i = 1, \dots, n$. For single-arm studies of Treatment 1, event probabilities are generated according to $\Phi(p_{1i}) = \mu_1 + 0.2 + \nu_{1i}$, $\nu_{1i} \sim N(0, \sigma_3^2)$ where $i = n + 1, \dots, n + m$; for single-arm studies of Treatment 2, event probabilities are generated through $\Phi(p_{2i}) = \mu_2 + 0.4 + \nu_{2i}$, $\nu_{2i} \sim N(0, \sigma_4^2)$ where $i = n + m + 1, \dots, n + m + l$. Note that 0.2 and 0.4 represent the systematic differences between the single-arm studies and RCTs for Treatments 1 and 2, respectively. For simplicity we assign 100 participants to each treatment arm within each trial. Then the numbers of events are generated according to $y_{1i} \sim \text{Bin}(100, p_{1i})$ and $y_{2i} \sim \text{Bin}(100, p_{2i})$ with $i = 1, \dots, n + m + l$. We let $\mu_1 = 0.40$, $\mu_2 = 1.10$, $\sigma_1 = 0.60$, $\sigma_2 = 0.70$, $\rho = 0.60$, $\sigma_3 = 0.80$, and $\sigma_4 = 1.00$. These correspond to the true values of the population-averaged event probabilities $\pi_1 = 0.63$ and $\pi_2 = 0.82$.

5.1.2 Simulation 2—In order to explore whether the proposed models are robust when there are larger heterogeneity between the two types of studies and when the study sample sizes vary across the two types of studies, we conduct additional simulations under 6 scenarios. For all scenarios, we let $n = m = l = 10$ as in S2 and thus label them S2.1–S2.6.

For S2.1 and S2.2, the systematic differences between RCTs and single-arm studies remain 0.2 and 0.4; for S2.3 and S2.4, they are set to be larger values of 0.4 and 0.6; for S2.5 and S2.6, they have even larger values of 0.6 and 0.8. In S2.1, S2.3, and S2.5, 200 participants are assigned to RCTs and 80 are assigned to single-arm studies; in S2.2, S2.4, and S2.6, 250 participants are assigned to RCTs and 60 are assigned to single-arm studies. Basic parameters including μ_1 , μ_2 , σ_1 , σ_2 , ρ , σ_3 and σ_4 are set to be the same values as in Simulation 1. Then y_{1i} and y_{2i} are generated in the same way.

5.2 Simulation Results

5.2.1 Simulation 1—The NPs (I) and (II), BGLMM (I), HPP, and HCP are fitted to the simulated datasets under the 5 scenarios described in Section 5.1 (note that BGLMM (II) is not explored in simulations because its prior specification for μ_1 and μ_2 relies on a separate analysis of single-arm studies). We also synthesize the information from only RCTs through $g(p_{1i}) = \mu_1 + v_{1i}$, $g(p_{2i}) = \mu_2 + v_{2i}$, with $(v_{1i}, v_{2i}) \sim N(\mathbf{0}, \Sigma)$, under Scenarios 1–4 (note that Scenario 5 contains only 1 RCT and thus is not suitable here). The synthesis results of solely RCTs are denoted by RCTs* in Table 6 and compared with those from the proposed methods.

Let us first compare the point estimates and interval lengths based on the data of solely RCTs with those based on both RCTs and single-arm studies. As is shown in Table 6, the estimates of π_1 and π_2 from the RCTs* are closer to those from the HPP and HCP models than those from the NP (I), NP (II) and BGLMM (I), under all scenarios. In other words, the HPP and HCP models downweight the single-arm studies to a larger degree than the other models. Posterior credible interval lengths of π_1 and π_2 from all proposed models (except some π_2 under the NP (II)) are shorter than those from the RCTs*. It suggests that the proposed methods do gain precision by including single-arm studies in synthesis. Note that the exception of NP (II) is understandable because it hypothetically assumes the single-arm studies to be two-arm with one arm missing and imputes these missing data.

Now let us compare the performance of the proposed models. Table 6 presents the relative biases of π_1 and π_2 and coverage probabilities of their 95% CIs from the proposed models under the 5 scenarios. Note that the MSEs of π_1 and π_2 are not presented here because they are consistently very small (at 3 decimal points) under all scenarios for all models. The relative biases are larger from the NP (I), NP (II) and BGLMM (I) than those from the HPP and HCP under Scenarios 1–4. Under Scenario 5, the relative biases from the BGLMM (I) and HCP are smaller than those from the other models. In short, the HCP seems to be the best model in terms of relative bias throughout all simulation scenarios.

Considering the observed coverage probabilities at the nominal level of 0.95, the NP (I), HPP and HCP give close to the nominal coverage under Scenarios 1 and 2, while the NP (II) and BGLMM (I) produce coverage probabilities a little further from the nominal level. With the number of RCTs becomes smaller and smaller from Scenarios 3 to 5 (i.e., we have less and less information for estimation), the HPP and HCP lose precision dramatically yielding coverage probabilities that are too high. This phenomenon is particularly evident under Scenario 5 with the coverage probabilities under these two models nearly 100%, which is consistent with what we have observed in the real data analysis of the advanced FL dataset.

Although the NP (I), NP (II) and BGLMM (I) also produce unsatisfactory coverage probabilities under Scenarios 3–5 (more specifically, lower than the nominal level), they are not as extreme as those from the HPP and HCP. Therefore sparse data should be handled more carefully.

Figure 4 further compares the performance of different models across scenarios to augment Table 6. The relative biases for π_1 and π_2 are shown in panel A of Figure 4, and coverage probabilities are shown in panel B. The x-axis lists the 5 scenarios. In panel A, the relative biases from each model under different scenarios are labeled by points and connected through lines. Looking at the trend of each line, relative biases tend to become larger going from Scenarios 1 to 5; in other words, more sparse information may yield more biased estimation. Comparison across different models shows that the relative biases from the HCP for both π_1 and π_2 are closer to 0 than the other models, under all scenarios. Thus we draw a conclusion that the HCP offers the most accurate estimation among all models. In panel B, the coverage probabilities from each model under different scenarios are labeled by points and connected through lines. The trend of each line also reveals that sparse information can lead to unsatisfactory coverage probability, especially for the HPP and HCP models under Scenario 5, where the coverage probabilities are nearly 100%.

5.2.2 Simulation 2—Figure 5 summarizes the results of Simulation 2. Panel A shows the relative biases for π_1 and π_2 and panel B shows the coverage probabilities. Let us first look at the relative bias for π_1 displayed on the left side of panel A. The comparison across models shows that the NP (I), NP (II) and BGLMM (I) produce more biased estimates than the HPP and HCP models under all scenarios. The HPP performs similarly to the HCP under S2.1–S2.4, but is inferior to the HCP under S2.5–S2.6. In short, the HCP produces the most accurate estimates. All lines except the one for the HCP show an upward trend along the increasing heterogeneity between RCTs and single-arm studies. The NP (I), NP (II) and BGLMM (I) produce dramatically larger biases when the heterogeneity becomes larger. Take NP (II) as an example, the relative bias is around 0.05 under S2.1 and S2.2, around 0.10 under S2.3 and S2.4, and 0.15 under S2.5 and S2.6. The HPP also shows an ascending trend but not as sharply as the aforementioned three models. The performance of the HCP seems stable across different scenarios, suggesting its robustness to larger heterogeneity between RCTs and single-arm studies. Another phenomenon we observed is that the relative biases are about the same for S2.1 and S2.2, as well as for S2.3 and S2.4, and for S2.5 and S2.6. Thus the performance of the models is not affected by heterogeneous study sample sizes between the two types of studies. We observe similar pattern for π_2 on the right side of panel A.

In terms of coverage probability for π_1 , all models except NP (II) broadly achieve the 95% nominal level under S2.1 and S2.2. The coverage probabilities from the NP (I), NP (II) and BGLMM (I) drop sharply along with the increase of heterogeneity between RCTs and single-arm studies. The HPP model performs generally well across all scenarios except under S2.5 and S2.6 where the coverage probabilities are only around 0.90. The HCP performs well under all scenarios. For π_2 , all models work well under S2.1 and S2.2. From S2.3 to S2.6, the coverage probabilities from the HPP and HCP continue achieving the nominal level, while those from the NP (I), NP (II) and BGLMM (I) drop sharply.

6 Discussion

For many biomedical research areas, such as interventional cardiology, orphan diseases, and rare cancer trials, a vast of information may be available from single-arms studies, while there may be only limited number of RCTs. It is important to synthesize single-arm studies and RCTs in order to improve statistical efficiency on the relative effectiveness of treatment options and to better inform decision-making process. However, few methods have synthesized these two sources of information. In this paper, we proposed a Bayesian framework with three methods to flexibly combine these two sources of evidence with the advantage of improved precision compared to traditional synthesis of RCTs only.

The proposed methods were evaluated empirically through two real data analyses and extensive simulations. In summary, the NPs (I) and (II) make a strong assumption of no difference between study designs thus are not recommended to synthesize RCTs and single-arm studies. However, when available information from RCTs is extremely sparse, as is shown in the advanced FL data set and in Scenario 5 of simulations, these two models may provide inferior yet reasonable alternatives. On the other hand, the BGLMM (I), BGLMM (II), HPP and HCP models account for design difference between single-arm and two-arm trials and adjust for potential bias from single-arm studies. Both the HPP and HCP models provide ways to downweight the single-arm studies and automate the prior elicitation procedures. Furthermore, the HCP model offers a way to measure commensurability between information from RCTs and single-arm studies and downweight single-arm studies accordingly. Typically, an adequate number of studies in meta-analysis is needed to ensure identifiability, sufficient precision and a reasonable power, especially for complex models. In general, the HCP offers more flexibility to integrate information of different qualities or from different designs, and thus is recommended as the primary method.

Criticisms can be made. Many researchers dislike the AB framework, arguing that the assumption of exchangeable absolute treatment effects “breaks randomization”²⁰. We admit that it may be stronger than the assumption of exchangeable treatment contrasts in the CB framework. However, as Senn⁴³ discussed, little harm is likely to be done in practice with this stronger assumption. In addition, the assumption of exchangeable event rates or studies is reasonable in most conditions, presented in highly-cited papers including Van Houwelingen et al.⁴⁴ and Shuster et al.⁴⁵. In meta-analysis, randomized studies are usually collected and screened with predetermined inclusion and exclusion criteria that are primarily concerned with characteristics of the target population. A (usually contrast-based) meta-analysis is then conducted under the fundamental assumption that relative treatment effects are exchangeable across studies. However, this assumption would be flawed if the collected studies were heterogeneous, and thus did not represent a common target population. For example, if a study drug performs better on younger or healthier people, the exchangeability of relative effects may not hold. In contrast, exchangeability of event rates may be a fair assumption to make, given that the meta-analysis data often arise from broadly similar study populations.

Hong et al.²¹ instead talked about advantages of the AB framework over the CB framework. First, the AB framework estimates absolute effects and based on which any relative effects

on any scale can be inferred, whereas the CB framework requires a proper selection of scale, which is difficult and empirically unverifiable. Second, the AB framework can provide more useful and straightforward interpretations when individual-patient data are available. Third, in hierarchical Bayesian theory, methods and computing, correlations in all data sources should be modeled, which is what the AB framework does by incorporating correlations among all treatment arms. In contrast, the CB framework requires additional assumptions or external information to capture and estimate the unknown within-study correlations.

One limitation of our methods is that we only considered single-arm studies in addition to RCTs in the synthesis process though we may have access to pseudo-RCTs and observational studies which may contain additional information and are worthy of consideration as well. Moreover, our methods only consider the comparison of two interventions, but usually multiple treatments are available for a particular disease and thus their simultaneous comparisons are of interest to patients, clinicians, and decision makers. Tapping into and utilizing the data abundance of current era would enhance the power of evidence synthesis and facilitate better informed decisions. Our future endeavor will focus on more generalized evidence synthesis and more complex data structures involving various designs and multiple treatments.

Future work also includes reducing the lag between knowledge generation and implementation. First, concerns on additional heterogeneity from different designs, differential publication selection processes for different types of studies (i.e., differential publication biases issue), and inconsistency between direct and indirect evidence may hinder the application and popularization of more generalized evidence synthesis. Therefore reliable measures to detect these issues and robust approaches to address them are essential and await exploration. Second, most of the current available R packages on meta-analysis are focused on single study type⁴⁶, thus work remains to be done in developing user-friendly software for meta-analysis synthesizing various types of studies.

Acknowledgment

Dr. Jing Zhang was supported in part by a start-up fund from the University of Maryland. The authors would like to express their appreciation to Dr. Rajeshwari Sridhara for her comments and suggestions, which substantially improved the quality of the article.

Appendices

A Inverse-gamma Prior Densities

Figure A.1:
Density plots for inverse-gamma (IG) priors.

References

1. Donna F, Berlin J, Morton S, Olkin I, Williamson G, Rennie D, et al. Meta-analysis of observational studies in epidemiology. JAMA 2000; 283(15):2009–2012.
2. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the quorum statement. The Lancet 1999; 354(9193): 1896–1900.
3. Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, Ades AE. Evidence Synthesis for Decision Making in Healthcare. John Wiley & Sons, West Sussex, 2012.
4. Navarese EP, Lupi A, Bu on A, Sansa M, Franchi E, Aurelio A, Bongo AS. Meta-analysis of non-randomized studies in interventional cardiology: a critical appraisal. Italian Journal of Public Health 2012; 6(4):303–306.

5. Jung SH. Randomized Phase II Cancer Clinical Trials. Chapman & Hall, Boca Raton, 2013.
6. Zhang J, Yuan Y. Industry-funded clinical trials: Beneficial or harmful? *Clinical Research and Regulatory Affairs* 2015; 32(4):111–114.
7. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in Medicine* 2013; 32(17):2935–2949. [PubMed: 23440610]
8. Sutton AJ, Higgins J. Recent developments in meta-analysis. *Statistics in Medicine* 2008; 27(5): 625–650. [PubMed: 17590884]
9. Huston P, Naylor CD. Health services research: reporting on studies using secondary data sources. *Canadian Medical Association Journal* 1996; 155(12):1697–1702. [PubMed: 8976336]
10. Begg CB, Pilote L. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991; 47(3):899–906. [PubMed: 1742445]
11. Li Z, Begg CB. Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association* 1994; 89(428):1523–1527.
12. Soares MO, Dumville JC, Ades A, Welton NJ. Treatment comparisons for decision making: facing the problems of sparse and few data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2014; 177(1):259–279.
13. Welton N, Ades A, Carlin J, Altman D, Sterne J. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172(1):119–136.
14. Turner RM, Spiegelhalter DJ, Smith G, Thompson SG. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172(1):21–47.
15. Zhang J, Carlin BP, Neaton JD, Soon GG, Nie L, Kane R, Virnig BA, Chu H. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. *Clinical Trials* 2014; 11(2): 246–262. [PubMed: 24096635]
16. Hong H, Chu H, Zhang J, Carlin BP. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods* 2016; 7(1):6–22. [PubMed: 26536149]
17. Zhang J, Fu H, Carlin BP. Detecting outlying trials in network meta-analysis. *Statistics in Medicine* 2015; 34(19):2695–2707. [PubMed: 25851533]
18. Zhang J, Chu H, Hong H, Virnig BA, Carlin BP. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Statistical Methods in Medical Research* 2017; 26(5):2227–2243. [PubMed: 26220535]
19. Zhang J, Yuan Y, Chu H. The impact of excluding trials from network meta-analyses—an empirical study. *PloS One* 2016; 11(12):e0165889. [PubMed: 27926924]
20. Dias S, Ades A. Absolute or relative effects? Arm-based synthesis of trial data. *Research Synthesis Methods* 2016; 7(1):23–28. [PubMed: 26461457]
21. Hong H, Chu H, Zhang J, Carlin BP. Rejoinder to the discussion of ‘A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons’ by S. Dias and A.E. Ades. *Research Synthesis Methods* 2016; 7(1):29–33. [PubMed: 26461816]
22. Lin L, Zhang J, Hodges JS, Chu H. Performing arm-based network meta-analysis in R with the pcnetmeta package. *Journal of Statistical Software* 2017; 80(5):doi:10.18637/jss.v080.i05.
23. Larose DT, Dey DK. Grouped random effects models for bayesian meta-analysis. *Statistics in Medicine* 1997; 16(16):1817–1829. [PubMed: 9280035]
24. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; 44:1049–1060. [PubMed: 3233245]
25. Berger J The case for objective Bayesian analysis. *Bayesian Analysis* 2006; 1(3):385–402.
26. Gelman A Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis* 2006; 1(3):515–534.
27. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. CRC press, Boca Raton, 2014.
28. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Statistical Science* 2000; 15(1):46–60.

29. Ibrahim JG, Chen MH, Sinha D. On optimality properties of the power prior. *Journal of the American Statistical Association* 2003; 98(461):204–213.
30. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. *Environmetrics* 2006; 17(1):95–106.
31. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Statistics in Medicine* 2015; 34(28):3724–3749. [PubMed: 26346180]
32. Gamalo MA, Tiwari RC, LaVange LM. Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharmaceutical Statistics* 2014; 13(1):25–40. [PubMed: 23913880]
33. Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 2011; 67(3):1047–1056. [PubMed: 21361892]
34. Murray TA, Hobbs BP, Lystig TC, Carlin BP. Semiparametric bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. *Biometrics* 2014; 70(1):185–191. [PubMed: 24308779]
35. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; 64(4):583–639.
36. Dias S, Sutton AJ, Ades A, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making* 2013; 33(5):607–617. [PubMed: 23104435]
37. Lunn DJ, Thomas A, Best N, Spiegelhalter D. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; 10(4):325–337.
38. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; 7(4):457–472.
39. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006; 101(474):447–459.
40. Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* 2009; 10(4):792–805. [PubMed: 19687150]
41. Chen Y, Luo S, Chu H, Su X, Nie L. An empirical bayes method for multivariate meta-analysis with an application in clinical trials. *Communications in Statistics-Theory and Methods* 2014; 43(16):3536–3551. [PubMed: 25089070]
42. Chen Y, Hong C, Riley RD. An alternative pseudolikelihood method for multivariate random effects meta-analysis. *Statistics in Medicine* 2015; 34(3):361–380. [PubMed: 25363629]
43. Senn S, Hans van houwelingen and the art of summing up. *Biometrical Journal* 2010; 52(1):85–94. [PubMed: 20140900]
44. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. *Statistics in Medicine* 1993; 12(24):2273–2284. [PubMed: 7907813]
45. Shuster JJ, Jones LS, Salmon DA. Fixed vs random effects meta-analysis in rare event studies: The rosiglitazone link with myocardial infarction and cardiac death. *Statistics in Medicine* 2007; 26(24):4375–4385. [PubMed: 17768699]
46. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods* 2015; 6(1):45–62. [PubMed: 26035469]
47. Lovenox (enoxaparin sodium injection) for subcutaneous and intravenous use. http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/020164s102lbl.pdf. Accessed: 2016-01-04.
48. Lassen MR, Borris LC, Christiansen HM, Boll KL, Eiskjær SP, Nielsen BW, Schøstt P, Olsen AD, Rodenberg JC, Lucht U. Prevention of thromboembolism in 190 hip arthroplasties: Comparison of LMW heparin and placebo. *Acta Orthopaedica* 1991; 62(1):33–38.
49. Hoek JA, Nurmohamed MT, Hamelynck KJ, Marti RK, Knipscheer HC, Ten Cate H, Büller HR, Magnani HN, ten Cate JW. Prevention of deep vein thrombosis following total hip replacement by low molecular weight heparinoid. *Thrombosis and Haemostasis* 1992; 67(1):28–32. [PubMed: 1615479]

50. Kalodiki EP, Hoppensteadt DA, Nicolaides AN, Fareed J, Gill K, Regan F, Al-Kutoubi A, Cunningham DA, Birch R, Harris N, et al. Deep venous thrombosis prophylaxis with low molecular weight heparin and elastic compression in patients having total hip replacement. a randomised controlled trial. *International Angiology* 1996; 15(2):162–168. [PubMed: 8803642]
51. Samama CM, Clergue F, Barre J, Montefiore A, Ill P, Samii K. Low molecular weight heparin associated with spinal anaesthesia and gradual compression stockings in total hip replacement surgery. arar study group. *British Journal of Anaesthesia* 1997; 78(6):660–665. [PubMed: 9215015]
52. ARIXTRA (fondaparinux sodium) solution for subcutaneous injection. http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/021345s032lbl.pdf. Accessed: 2016-01-04.
53. Iprivask 15 mg (desirudin for injection), for subcutaneous injection. http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/021271s006lbl.pdf. Accessed: 2016-01-04.
54. Marcus R, Imrie K, Belch A, Cunningham D, Flores E, Catalano J, Solal-Celigny P, Offner F, Walewski J, Raposo J, et al. CVP chemotherapy plus rituximab compared with CVP as first-line treatment for advanced follicular lymphoma. *Blood* 2005; 105(4):1417–1423. [PubMed: 15494430]
55. Kalinka E, Juszczy ski P, Woszczek G, Borowiec M, Robak T, Kowalski M, Bili ski P, Coiffier B, Salles G, Warzocha K. The role of HLA DRB1 genetic polymorphisms in non-Hodgkin's lymphomas. *Przegląd Lekarski* 2004; 62(12):1422–1426.
56. Hagenbeek A, Eghbali H, Monfardini S, Vitolo U, Hoskin PJ, de Wolf-Peters C, MacLennan K, Staab-Renner E, Kalmus J, Schott A, et al. Phase III intergroup study of fludarabine phosphate compared with cyclophosphamide, vincristine, and prednisone chemotherapy in newly diagnosed patients with stage III and IV low-grade malignant Non-Hodgkin's lymphoma. *Journal of Clinical Oncology* 2006; 24(10):1590–1596. [PubMed: 16575010]
57. Levitt M, Gharibo M, Strair R, Schaar D, Rubin A, Bertino J. Accelerated r-cop: A pilot study for the treatment of advanced low grade lymphomas that has a high complete response rate. *Journal of Chemotherapy* 2009; 21(4):434–438. [PubMed: 19622463]
58. Moccia AA, Hoskins P, Klasa R, Savage KJ, Shenkier T, Slack GW, Gascoyne RD, Connors JM, Sehn LH. Front-line therapy with Rituximab, Cyclophosphamide, Vincristine, and Prednisone (R-CVP) followed by 2 years of Rituximab maintenance for follicular lymphoma (FL) is associated with excellent outcomes and improved progression-free survival (PFS) in comparison to no maintenance. *Blood* 2010; 116(21):753–753.
59. Salles G, Seymour JF, Offner F, López-Guillermo A, Belada D, Xerri L, Feugier P, Bouabdallah R, Catalano JV, Brice P, et al. Rituximab maintenance for 2 years in patients with high tumour burden follicular lymphoma responding to rituximab plus chemotherapy (prima): a phase 3, randomised controlled trial. *The Lancet* 2011; 377(9759):42–51.
60. Nastoupil L, Sinha R, Byrtek M, Taylor M, Cerhan JR, Friedberg J, Link BK, Ziemiecki RM, Hirata JH, Flowers CR. A comparison of the effectiveness of first-line chemoimmunotherapy regimens for follicular lymphoma (FL) used in the united states. *ASH Annual Meeting Abstracts* 2011; 118(21):97.
61. Federico M, Luminari S, Dondi A, Tucci A, Vitolo U, Rigacci L, Di Raimondo F, Carella AM, Pulsoni A, Merli F, et al. R-CVP versus R-CHOP versus R-FM for the initial treatment of patients with advanced-stage follicular lymphoma: results of the FOLL05 trial conducted by the fondazione italiana linfomi. *Journal of Clinical Oncology* 2013; 31(12):1506–1513. [PubMed: 23530110]
62. Bastion Y, Blay JY, Divine M, Brice P, Bordessoule D, Sebban C, Blanc M, Tilly H, Lederlin P, Deconinck E, et al. Elderly patients with aggressive non-Hodgkin's lymphoma: disease presentation, response to treatment, and survival—a Groupe d'Etude des Lymphomes de l'Adulte study on 453 patients older than 69 years. *Journal of Clinical Oncology* 1997; 15(8):2945–2953. [PubMed: 9256139]
63. Arranz R, Garcia-Alfonso P, Sobrino P, Zamora P, Carrion R, Garcia-Larana J, Pérez G, López J, Lavilla E, Lozano M, et al. Role of interferon alfa-2b in the induction and maintenance treatment of low-grade non-Hodgkin's lymphoma: results from a prospective, multicenter trial with double randomization. *Journal of Clinical Oncology* 1998; 16(4):1538–1546. [PubMed: 9580385]

64. Hagenbeek A, Carde P, Meerwaldt JH, Somers R, Thomas J, De Bock R, Raemaekers J, van Hoof A, De Wolf-Peeters C, Van Glabbeke M. Maintenance of remission with human recombinant interferon alfa-2a in patients with stages III and IV low-grade malignant non-Hodgkin's lymphoma. European Organization for Research and Treatment of Cancer Lymphoma Cooperative Group. *Journal of Clinical Oncology* 1998; 16(1):41–47. [PubMed: 9440721]
65. Hochster H, Weller E, Ryan T, Habermann T, Gascoyne R, Frankel S, Horning S. Results of e1496: A phase III trial of CVP with or without maintenance rituximab in advanced indolent lymphoma (NHL). *ASCO Annual Meeting Proceedings* 2004; 22(14):6502.

Figure 1: Graphical comparisons of the estimates from various models for the deep vein thrombosis (DVT) data set. Both posterior medians and 95% credible intervals are plotted. The dotted vertical lines are $x = 0$ and $x = 1$ respectively.

Figure 2:
Graphical comparisons of estimates from various models for the advanced follicular lymphoma (FL) data set. Both posterior medians and 95% credible intervals are plotted. The dotted vertical lines are $x = 0$ and $x = 1$ respectively.

Figure 3:

Data structures under various simulation scenarios. S1 to S5 refer to Scenario 1 to Scenario 5. RCT represents the randomized controlled trials, Single1 represents the single-arm studies of Treatment 1, and Single2 represents the single-arm studies of Treatment 2.

Figure 4:
Relative biases and coverage probabilities for π_1 and π_2 obtained from various models under Scenarios 1 to 5.

Figure 5:
Relative biases and coverage probabilities for π_1 and π_2 obtained from various models under Scenarios 2.1 to 2.6.

Table 1:

Deep vein thrombosis (DVT) data set. LMWH=Lower molecular weight heparin. N denotes sample size, y denotes number of DVT events, and % denotes crude event probability for each arm.

Trial	Author (year)	Treatment	N	y	%
1	Lovenox (2013) ⁴⁷	LMWH	50	5	10
1	Lovenox (2013) ⁴⁷	Placebo	50	23	46
2	Lovenox (2013) ⁴⁷	LMWH	47	5	11
2	Lovenox (2013) ⁴⁷	Placebo	52	32	62
3	Lovenox (2013) ⁴⁷	LMWH	90	6	7
3	Lovenox (2013) ⁴⁷	Placebo	89	18	20
4	Lassen (1991) ⁴⁸	LMWH	93	29	31.18
4	Lassen (1991) ⁴⁸	Placebo	97	44	45.36
5	Hoek (1992) ⁴⁹	LMWH	97	15	15.46
5	Hoek (1992) ⁴⁹	Placebo	99	56	56.57
6	Kalodiki (1996) ⁵⁰	LMWH	32	12	37.50
6	Kalodiki (1996) ⁵⁰	Placebo	14	13	92.86
7	Samama (1997) ⁵¹	LMWH	78	11	14.10
7	Samama (1997) ⁵¹	Placebo	75	28	37.33
8	Arixtra (2014) ⁵²	LMWH	797	66	8.3
9	Arixtra (2014) ⁵²	LMWH	919	85	9.2
10	Iprivask (2014) ⁵³	LMWH	768	197	25.7
11	Lovenox (2013) ⁴⁷	LMWH	208	22	11
12	Lovenox (2013) ⁴⁷	LMWH	199	27	14

Table 2:

Advanced follicular lymphoma (FL) data set. CVP = cyclophosphamide, vincristine, and prednisone; R-CVP=Rituximab+CVP. N denotes sample size, y denotes number of responses, and % denotes crude response probability for each arm.

Study	Author (year)	Treatment	N	y	%
1	Marcus (2005) ⁵⁴	R-CVP	162	131	80.86
1	Marcus (2005) ⁵⁴	CVP	159	91	57.23
2	Kalinka (2005) ⁵⁵	CVP	10	6	60.00
3	Hagenbeek (2006) ⁵⁶	CVP	187	97	51.87
4	Levitt (2009) ⁵⁷	R-CVP	7	7	100.00
5	Moccia (2010) ⁵⁸	R-CVP	251	216	86.06
6	Salles (2011) ⁵⁹	R-CVP	272	223	81.99
7	Nastoupil (2011) ⁶⁰	R-CVP	247	217	87.85
8	Federico (2013) ⁶¹	R-CVP	168	148	88.10
9	Bastion (1997) ⁶²	CVP	219	99	45.21
10	Arranz (1997) ⁶³	CVP	71	53	74.65
11	Hagenbeek (1998) ⁶⁴	CVP	315	249	79.05
12	Hochster (2004) ⁶⁵	CVP	282	223	79.08

Table 3:

Posterior medians and 95% credible intervals under various models for the deep vein thrombosis (DVT) data set. π_1 denotes the population-averaged DVT event probability for LMWH, and π_2 denotes the population-averaged DVT event probability for placebo. OR_{21} , RD_{21} and RR_{21} represent the odds ratio, risk difference, and risk ratio of placebo versus LMWH. The DIC and $\bar{D}_{res}$ from the BGLMM (II) are calculated based on a subset of the data and the $\bar{D}_{res}$ from the HPP is calculated based on a downweighted likelihood, thus they are inappropriate for comparison and labeled by “—”.

	NP (I)	NP (II)	BGLMM (I)	BGLMM (II)	HPP	HCP
π_1	0.16 (0.12,0.22)	0.17 (0.11,0.25)	0.17 (0.12,0.25)	0.18 (0.12,0.26)	0.19 (0.11,0.30)	0.19 (0.11,0.31)
π_2	0.50 (0.34,0.68)	0.50 (0.34,0.66)	0.50 (0.34,0.67)	0.49 (0.33,0.66)	0.50 (0.33,0.67)	0.50 (0.34,0.67)
OR_{21}	5.29 (2.46,11.91)	4.88 (2.38,10.20)	4.65 (2.23,10.33)	4.54 (2.18,10.01)	4.28 (1.97,9.84)	4.33 (1.92,9.96)
RD_{21}	0.34 (0.17,0.52)	0.32 (0.17,0.49)	0.32 (0.16,0.50)	0.31 (0.15,0.49)	0.31 (0.14, 0.49)	0.31 (0.14,0.48)
RR_{21}	3.14 (1.91,4.88)	2.94 (1.84,4.62)	2.83 (1.76,4.45)	2.78 (1.72,4.44)	2.62 (1.57,4.47)	2.65 (1.54,4.54)
DIC	128.92	124.10	123.76	—	125.23	123.83
$\bar{D}_{res}$	20.10	23.34	18.27	—	—	26.19

Table 4:

Prior sensitivity analysis. $se_1: \mu_1, \mu_2 \sim N(0, 100)$; $se_2: \mu_1, \mu_2 \sim N(0, 10)$; $se_3: \mu_1, \mu_2 \sim t\text{-dist}(0, df=2)$; $se_4: \mu_1, \mu_2 \sim t\text{-dist}(0, df=5)$; $se_5: \tau \sim \text{Gamma}(0.1, 0.1)$; $se_6: \tau \sim \text{Gamma}(0.001, 0.001)$; $se_7: \tau \sim \text{Gamma}(1, 0.5)$; $se_8: \tau \sim \text{Gamma}(3, 1)$; $se_9: \Sigma^{-1} \sim W(I_2, 2)$; $se_{10}: \Sigma^{-1} \sim W(R^{-1}, 2)$, where R is a known 2×2 matrix with diagonal elements equal to 0.1 and off-diagonal elements equal to 0.005; $se_{11}: \alpha_0 \sim \text{Beta}(5, 1)$; $se_{12}: \alpha_0 \sim \text{Beta}(3, 2)$; $se_{13}: \alpha_0 \sim \text{Beta}(1, 1)$; $se_{14}: \alpha_0 \sim \text{Beta}(2, 2)$; $se_{15}: \alpha_0 \sim \text{Beta}(2, 5)$.

	se₁	se₂	se₃	se₄	se₅
π_1	0.19 (0.11,0.31)	0.19 (0.11,0.31)	0.19 (0.12,0.32)	0.19 (0.11,0.31)	0.19 (0.11,0.33)
π_2	0.50 (0.34,0.67)	0.50 (0.33,0.67)	0.51 (0.35,0.67)	0.50 (0.35,0.66)	0.50 (0.34,0.67)
	se₆	se₇	se₈	se₉	se₁₀
π_1	0.18 (0.12,0.30)	0.19 (0.11,0.33)	0.19 (0.11,0.33)	0.19 (0.11,0.31)	0.18 (0.12,0.27)
π_2	0.50 (0.34,0.66)	0.50 (0.33,0.68)	0.50 (0.33,0.68)	0.50 (0.34,0.68)	0.49 (0.35,0.64)
	se₁₁	se₁₂	se₁₃	se₁₄	se₁₅
π_1	0.19 (0.11,0.32)	0.19 (0.11,0.33)	0.20 (0.11,0.35)	0.19 (0.11,0.34)	0.19 (0.11,0.34)
π_2	0.50 (0.33,0.68)	0.50 (0.33,0.68)	0.51 (0.34,0.69)	0.50 (0.33,0.69)	0.50 (0.33,0.69)

Table 5:

Posterior medians and 95% credible intervals under various models for the advanced follicular lymphoma (FL) data set. π_1 denotes the population-averaged response probability for CVP, and π_2 denotes the population-averaged response probability for R-CVP. OR_{21} , RD_{21} and RR_{21} represent the odds ratio, risk difference, and risk ratio of R-CVP versus CVP. The DIC and $\bar{D}_{res}$ from the BGLMM (II) are calculated based on a subset of the data and the $\bar{D}_{res}$ from the HPP is calculated based on a downweighted likelihood, thus they are inappropriate for comparison and labeled by “—”.

	NP (I)	NP (II)	BGLMM (I)	BGLMM (II)	BGLMM (II)*	HPP	HCP
π_1	0.64 (0.50,0.75)	0.63 (0.47,0.76)	0.61 (0.49,0.73)	0.61 (0.50,0.73)	0.60 (0.44,0.76)	0.60 (0.43,0.78)	0.59 (0.28,0.83)
π_2	0.85 (0.81,0.88)	0.84 (0.70,0.91)	0.79 (0.59,0.86)	0.79 (0.59,0.87)	0.78 (0.58,0.88)	0.77 (0.56,0.91)	0.76 (0.12,0.93)
OR_{21}	3.25 (1.81,6.05)	3.02 (1.11,7.96)	2.31 (0.97,4.78)	2.31 (0.96,4.99)	2.29 (0.86,6.23)	2.12 (0.68,7.88)	2.14 (0.33,14.73)
RD_{21}	0.21 (0.10,0.35)	0.20 (0.02,0.38)	0.17 (−0.01,0.32)	0.17 (−0.01,0.32)	0.17 (−0.03,0.37)	0.16 (−0.08,0.39)	0.16 (−0.23,0.52)
RR_{21}	1.33 (1.13,1.70)	1.32 (1.03,1.79)	1.27 (0.99,1.62)	1.28 (0.99,1.61)	1.28 (0.95,1.80)	1.26 (0.89,1.88)	1.26 (0.67,2.76)
DIC	85.22	85.95	85.16	—	—	94.59	85.29
$\bar{D}_{res}$	13.66	12.24	13.03	—	—	—	12.55

Table 6:

Point estimates, interval lengths, relative biases and coverage probabilities for the population-averaged event probabilities π_1 and π_2 under various scenarios. Estimate represents point estimate; Length represents interval length; ReBias represents relative bias; and CP represents coverage probability. S_1 to S_5 refer to Scenario 1 to Scenario 5. RCTs* shows the synthesis results of solely RCTs. Note that relative biases and coverage probabilities from RCTs* are not of interest and thus are not presented.

			NP (I)	NP (II)	BGLMM (I)	HPP	HCP	RCTs*
S_1	Estimate	π_1	0.65	0.66	0.65	0.64	0.64	0.63
		π_2	0.82	0.83	0.82	0.82	0.82	0.81
	Length	π_1	0.17	0.17	0.17	0.18	0.20	0.21
		π_2	0.16	0.32	0.15	0.16	0.17	0.18
	ReBias	π_1	0.03	0.04	0.03	0.01	0.01	—
		π_2	0.01	0.01	0.01	0	0	—
	CP (100%)	π_1	92.8	91.3	92.5	95.6	95.7	—
		π_2	96.0	94.7	94.4	96.3	96.2	—
	Estimate	π_1	0.66	0.66	0.66	0.64	0.64	0.63
		π_2	0.83	0.84	0.83	0.81	0.82	0.81
S_2	Length	π_1	0.19	0.19	0.19	0.22	0.24	0.26
		π_2	0.17	0.33	0.16	0.18	0.20	0.22
	ReBias	π_1	0.03	0.05	0.04	0.01	0.01	—
		π_2	0.02	0.02	0.02	-0.01	0	—
	CP (100%)	π_1	92.8	91.0	91.5	97.4	97.0	—
		π_2	93.9	92.5	95.3	96.6	96.8	—
	Estimate	π_1	0.66	0.67	0.66	0.64	0.64	0.63
		π_2	0.84	0.85	0.84	0.80	0.82	0.81
	Length	π_1	0.23	0.22	0.23	0.27	0.33	0.38
		π_2	0.17	0.36	0.20	0.23	0.28	0.32
S_3	ReBias	π_1	0.04	0.06	0.05	0	0.02	—
		π_2	0.03	0.04	0.03	-0.02	0	—
	CP (100%)	π_1	93.2	92.3	95.5	99.5	99.6	—
		π_2	92.5	89.9	97.9	98.6	99.4	—
	Estimate	π_1	0.67	0.67	0.66	0.63	0.64	0.62
		π_2	0.84	0.85	0.84	0.79	0.81	0.80
	Length	π_1	0.23	0.22	0.24	0.31	0.40	0.49
		π_2	0.18	0.36	0.24	0.28	0.36	0.43
	ReBias	π_1	0.05	0.06	0.04	0	0.01	—
		π_2	0.03	0.04	0.03	-0.03	0	—
S_4	CP (100%)	π_1	92.5	91.5	96.8	99.8	100	—
		π_2	93.3	92.0	99.8	99.1	99.5	—

		NP (I)	NP (II)	BGLMM (I)	HPP	HCP	RCT _S *	
S_5	ReBias	π_1	0.06	0.06	0.02	−0.03	−0.01	—
		π_2	0.04	0.04	0.01	−0.06	−0.02	—
	CP (100%)	π_1	90.7	89.5	99.0	100	100	—
		π_2	92.8	91.3	99.7	99.3	99.9	—

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript