ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)
67, *Part* 4, *pp.* 1047–1069

Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis

Hwanhee Hong,

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

Haoda Fu

Eli Lilly and Company, Indianapolis, USA

and Bradley P. Carlin

University of Minnesota, Minneapolis, USA

[Received September 2016. Revised February 2018]

Summary. In network meta-analysis, it is often desirable to synthesize different types of studies, featuring aggregated data and individual patient level data. However, existing methods do not sufficiently consider the quality of studies across different types of data and assume that the treatment effects are exchangeable across all studies regardless of these types. We propose Bayesian hierarchical network meta-analysis models that allow us to borrow information adaptively across aggregated data and individual patient level data studies by using power and commensurate priors. The power parameter in the power priors and spike-and-slab hyperprior in the commensurate priors govern the level of borrowing information among study types. We incorporate covariate-by-treatment interactions to deliver personalized decision making and model any ecological fallacy. The methods are validated and compared via extensive simulation studies and then applied to an example in diabetes treatment comparing 28 oral antidiabetic drugs. We compare results across model and hyperprior specifications. Finally, we close with a discussion of our findings, limitations and future research.

Keywords: Adaptive borrowing; Commensurate prior; Data integration; Individual patient data; Network meta-analysis; Power prior

1. Introduction

Network meta-analysis (NMA) is a popular statistical methodology for combining multiple independent studies to compare multiple treatments directly and indirectly, and perhaps to find the best among the candidates. NMA usually combines aggregated (or study level) data (AD), which include only summary statistics for each study. A large number of statistical methods for analysing AD have been established and discussed (Lumley, 2002; Lu and Ades, 2004, 2006; Hong *et al.*, 2016). Recent advances in data sharing systems and computing power enable us to integrate individual patient data (IPD) in meta-analysis, for which various methods have also been proposed (Goldstein *et al.*, 2000; Jackson *et al.*, 2006; Riley *et al.*, 2008; Debray *et al.*, 2012, 2013; Hong *et al.*, 2015). These models have been extended to simultaneous combination

Address for correspondence: Hwanhee Hong, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 North Broadway, Baltimore, MD 21205, USA.
E-mail: hhong@jhu.edu

 0035–9254/18/671047

of both AD and IPD, sometimes adjusted by covariates that are available in both sources of data (Saramago *et al.*, 2012; Donegan *et al.*, 2013).

Existing NMA models combining AD and IPD typically use the same weight on each source when estimating treatment effects. However, if one source is considered better than the other (for example, IPD often provide richer covariate information), we would prefer to weight the two sources differently and to borrow information differentially across them. In addition, in practice, the number of available studies reporting IPD (the 'gold standard' of data) is typically much smaller than the number reporting only AD, because of confidentiality and competitiveness issues that often limit access to IPD. This heightens the need for flexible, adaptive statistical methods that can intelligently combine all sources of information, weighting each according to their reliability.

In Bayesian analysis, adaptively informative priors can be used when synthesizing results across studies, especially when combining results from primary and auxiliary data. These approaches include power priors (Ibrahim and Chen, 2000; Neelon and O'Malley, 2010) and commensurate priors (Hobbs *et al.*, 2011, 2012, 2013). Suppose that D and D_0 denote primary and auxiliary data respectively. Both modelling approaches require parameters in the two sources of data, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, to be analogous. Power priors assume that $\boldsymbol{\theta} \equiv \boldsymbol{\theta}_0$ and downweight the auxiliary data by raising its likelihood to a power $\alpha_0 \in [0, 1]$, the posterior arising as

$$p(\boldsymbol{\theta}|\mathbf{D}, \mathbf{D}_0) \propto L(\boldsymbol{\theta}|\mathbf{D}) L(\boldsymbol{\theta}|\mathbf{D}_0)^{\alpha_0} \pi(\boldsymbol{\theta}), \tag{1}$$

where $L(\boldsymbol{\theta}|\mathbf{D})$ and $L(\boldsymbol{\theta}|\mathbf{D}_0)$ denote the likelihoods from the primary and auxiliary data respectively. In this paper, we presume that IPD are the primary source of data and AD are the auxiliary source of data.

Although power priors are intuitively appealing, they can be difficult to implement in standard Markov chain Monte Carlo software, and developing expert sampling algorithms can be challenging. Commensurate priors, by contrast, assume that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and instead specify a hierarchical model with posterior

$$p(\boldsymbol{\theta}|\mathbf{D}, \mathbf{D}_0) \propto L(\boldsymbol{\theta}|\mathbf{D}) L(\boldsymbol{\theta}_0|\mathbf{D}_0) \pi(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{\eta}) \pi(\boldsymbol{\theta}_0) \pi(\boldsymbol{\eta}). \tag{2}$$

The $\pi(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{\eta})$ component of posterior (2) is called the *commensurate prior*, where the *commensurability parameter* $\boldsymbol{\eta}$ controls the influence of the supplemental information (in our case, the AD). The standard commensurate prior assumes that $E[\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{\eta}] = \boldsymbol{\theta}_0$, and that $\text{var}(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \boldsymbol{\eta})$ is a diagonal matrix with diagonal given by the reciprocals of the components of $\boldsymbol{\eta}$. Hence, a commensurate prior 'centres' $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$ and $\boldsymbol{\eta}$ controls the variance of $\boldsymbol{\theta}$ about $\boldsymbol{\theta}_0$. When there are only two sources of data (say, one IPD and one AD), estimating the components of $\boldsymbol{\eta}$ is difficult, especially when the posteriors for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ are not highly separated.

In this paper, we adapt power and commensurate prior methods for combining IPD and AD in Bayesian network meta-analysis, where we borrow information adaptively from each source. The remainder of our paper is structured as follows. First, Section 2 introduces our motivating diabetes data set. Section 3 then develops our adaptive NMA data integration methods. We use extensive simulation studies in Section 4 to assess the performance of our models and then apply them to the diabetes data in Section 5. Finally, Section 6 discusses our work and its limitations, and suggests avenues for further research.

2. Motivating data

Our motivating data include a total of 41 randomized controlled trials (RCTs) investigating the efficacy of 28 treatments for patients diagnosed with type 2 diabetes in lowering their lev-

Table 1. Oral antidiabetic drugs' full names, their abbreviations, drug code and the number of studies investigating each treatment†

Drug	*Abbreviation*	*Drug code*	*Number of AD*	*Number of IPD*
Placebo	Placebo	1	17	7
Metformin	MET	2	9	1
Pioglitazone	PIO	3	2	12
Rosiglitazone	ROS	4	1	1
Acarbose	ACA	5	1	
Dapagliflozin	DAP	6	1	
Glipizide	GLI	7	1	
Glyburide	GLY	8	2	
Rimonabant	RIM	9	1	
Saxagliptin	SAX	10	1	
Sitagliptin	SIT	11	7	
Troglitazone	TRO	12	1	
Vildagliptin 50 mg	VIL50	13	5	
Vildagliptin 100 mg	VIL100	14	4	
Metformin + alogliptin	MET+ALO	15	1	
Metformin + glipizide	MET+GLI	16	1	
Metformin + glyburide	MET+GLY	17	1	
Metformin + pioglitazone	MET+PIO	18	1	
Metformin + sitagliptin	MET+SIT	19	4	
Metformin + vildagliptin 50 mg	MET+VIL50	20	2	
Metformin + vildagliptin 100 mg	MET+VIL100	21	2	
Pioglitazone + sitagliptin	PIO+SIT	22	1	
Pioglitazone + vildagliptin 50 mg	PIO+VIL50	23	1	
Pioglitazone + vildagliptin 100 mg	PIO+VIL100	24	1	
Sulphonylurea + vildagliptin 50 mg	SUL+VIL50	25	1	
Sulphonylurea + vildagliptin 100 mg	SUL+VIL100	26	1	
Glyburide + troglitazone	GLY+TRO	27	1	
Sulphonylurea	SUL	28		4

†Missing cells in the last two columns indicate no study available.

els of glycated haemoglobin A1c, HbA1c. Of the 41 studies, 29 have only AD and 12 have IPD. These 28 treatments are all oral antidiabetic treatments. These 12 studies containing IPD data are all from phase III RCTs evaluating oral antidiabetic treatments that were conducted by Eli Lilly and Company between the years 2000 and 2010. The 29 AD studies are from the recent (1998–2009) published literature on major phase III registration trials to evaluate oral antidiabetic treatments. These studies were selected before conducting our analysis, and the selection method was objective and followed standard meta-analytic guidelines (see, for example, the white paper published by the Food and Drug Administration, which is available from `http://www.fda.gov/downloads/Drugs/NewsEvents/UCM372069.pdf`). Detailed information on the 12 IPD studies is available in Hong *et al.* (2015). The 41 studies are from independent RCTs, i.e. no study is reported in both the AD and the IPD collections. Our outcome of interest is the change in HbA1c from baseline to 24 weeks. The IPD outcomes were measured between 24 and 28 weeks; the AD outcomes were measured between 12 and 52 weeks. It is clinically reasonable to believe that changes in HbA1c from baseline are stable between the 12- and 52-week end points (American Diabetes Association, 2015). We consider two covariates, the patient's baseline age and HbA1c measurement, as this information is available across all AD and IPD studies.

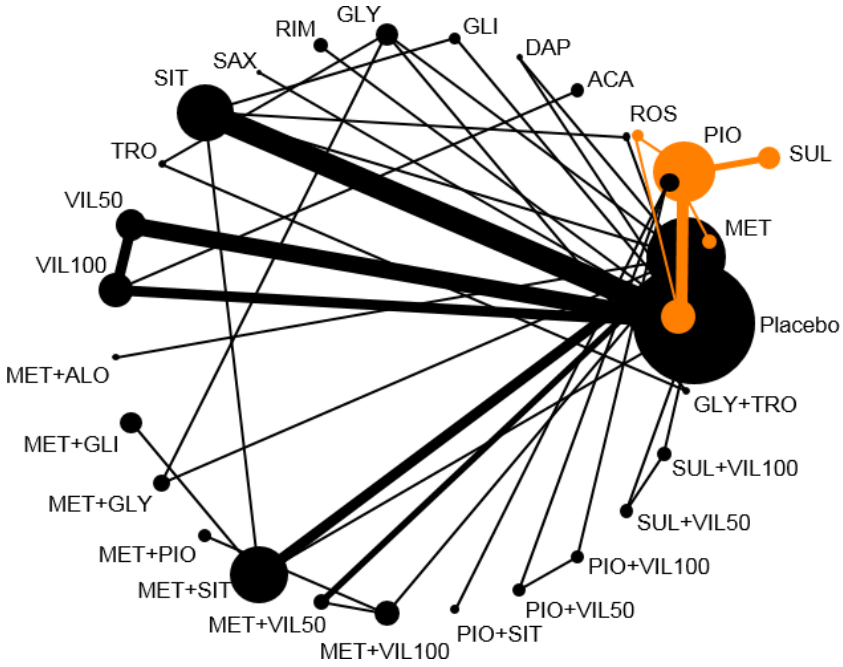

Fig. 1. Network graph for both AD and IPD; ●—, AD; ●—, IPD

Table 1 shows the drugs' full names, their abbreviations, the drug code that we assigned for the analyses and the number of AD and IPD studies investigating each drug. Out of the 28 drugs, only four (placebo, metformin, pioglitazone and rosiglitazone) have results from both AD and IPD studies. For placebo and metformin, the numbers of IPD studies (seven and one respectively) are relatively smaller than the numbers of AD studies (17 and nine respectively), whereas the opposite is true for pioglitazone. All the other drugs have results only from AD, except for sulphonylurea. Fig. 1 overlays two network plots across our studies; the network for AD is black and that for IPD is orange. The size of each node corresponds to the number of studies investigating the drug, and the thickness of each edge is proportional to the total number of samples for the relationship. As we observe in Table 1, the AD network is much bigger than the IPD network because we could obtain AD from publicly available peer-reviewed journals, whereas IPD in our setting were available only from studies performed by one firm (Eli Lilly and Company).

11 of the 29 AD studies do not report standard errors for the outcomes. As such, we simply impute them all to be 0.1, since 14 of the 18 studies (77%) reporting standard errors have roughly this standard error, and those of the remaining studies are between 0.06 and 0.2.

3. Methods

In NMA hierarchical models, two model parameterizations are often used: contrast based and arm based. The former essentially assumes only that treatment contrasts (with respect to some arbitrary baseline treatment) are exchangeable across studies, whereas the latter assumes that the arms testing a particular treatment are themselves exchangeable across studies. Some investigators are still uncomfortable with the arm-based parameterization, believing it to 'break the randomization'; see Hong *et al.* (2016), its discussion and rejoinder for a full (and some-

times contentious) discussion of this issue. In practice, results when combining AD under the two NMA parameterizations are often very similar, and arm-based approaches provide more straightforward parameter interpretations when including treatment-by-individual covariate interactions in the NMA with IPD (Hong *et al.*, 2015, 2016; Zhang *et al.*, 2015; Zhao *et al.*, 2016). Since in this paper we wish to focus on data integration methods and not to revisit NMA orthodoxy, we consider only the more conceptually straightforward arm-based models. Similar integration methods can be used with the contrast-based parameterization, and we provide the contrast-based counterparts to our approaches in the on-line supplementary materials.

We observe continuous outcome measurements y_{ik}^{AD} from AD and y_{ijk}^{IPD} from IPD, where i, j and k index study, patient and treatment respectively. Here, $i=1,\ldots,I$, $j=1,\ldots,n_i$ and $k=1,\ldots,K$, where $I^{\mathrm{AD}}=29$ and $I^{\mathrm{IPD}}=12$ are the total numbers of AD and IPD studies, with $I=I^{\mathrm{AD}}+I^{\mathrm{IPD}}=41$. Also, n_i is the number of patients in study i, and $K=28$ is the total number of treatments. The likelihoods that are assumed for AD and IPD are

$$y_{ik}^{\mathrm{AD}} \sim N(\mu_{ik}^{\mathrm{AD}}, \mathrm{se}_{ik}^2), \qquad i=1,\ldots,I^{\mathrm{AD}}, \tag{3}$$

and

$$y_{ijk}^{\mathrm{IPD}} \sim N(\mu_{ijk}^{\mathrm{IPD}}, \sigma_{ik}^2), \qquad i=I^{\mathrm{AD}}+1,\ldots,I, \tag{4}$$

where μ_{ik}^{AD} and μ_{ijk}^{IPD} are the underlying true means from AD and IPD respectively, se_{ik} is the known (but often estimated) standard error in AD and σ_{ik} is the unknown within-study standard deviation for IPD.

In the remainder of this section, we consider covariate-adjusted models that include treatment-by-covariate interactions. For simplicity, we consider a single covariate x when describing models, but multiple covariates can be applied with a vector $\mathbf{x}$. When proper covariates are not available, the interactions can simply be excluded from the models below and fitted as covariate-unadjusted models.

3.1. Network meta-analysis with a single data type

We first overview NMA models when a single data type (AD or IPD) is available. First, when we have only AD, we use the arm-based NMA model incorporating a missing data framework and accounting for correlation between treatment effects that was proposed by Hong *et al.* (2016). In addition, we allow the model to have treatment-by-covariate interactions. Using likelihood (3), suppose that $\bar{x}_{ik}$ is the aggregate covariate measured for patients who are assigned to the kth treatment in study i. We model the unknown parameter μ_{ik}^{AD} as

$$\mu_{ik}^{\mathrm{AD}} = \theta_k^{\mathrm{AD}} + \nu_{ik}^{\mathrm{AD}} + \beta_{ik}^{\mathrm{AD}} \bar{x}_{ik}, \tag{5}$$

where θ_k^{AD} is the kth treatment effect, the ν_{ik}^{AD} are study level treatment random effects and the β_{ik}^{AD} are study level treatment-by-covariate interaction random effects in the AD studies. We assume the random-effects distribution $\boldsymbol{\nu}_i^{\mathrm{AD}} = (\nu_{i1}^{\mathrm{AD}},\ldots,\nu_{iK}^{\mathrm{AD}})^{\mathrm{T}} \sim^{\mathrm{ind}} N(\mathbf{0}, \boldsymbol{\Sigma})$. We consider two hyperprior specifications for $\boldsymbol{\Sigma}$. First, $\boldsymbol{\Sigma}$ might have diagonal elements equal to τ^2 and off-diagonal elements equal to $\rho\tau^2$ (the so-called equicorrelated model). We denote this equicorrelated covariance matrix as $\boldsymbol{\Sigma} = \mathrm{diag}(\tau^2) + \rho\tau^2 I(i \neq j)$, where $\mathrm{diag}(e)$ is a diagonal matrix with diagonal elements e and $I(i \neq j)$ is a matrix having diagonal elements 0 and off-diagonal elements 1. Here, we assign hyperprior distributions $\rho \sim \mathrm{uniform}(0,1)$, which assumes positive correlation between treatment effects, and $\tau \sim \mathrm{uniform}(0,10)$. Our second $\boldsymbol{\Sigma}$ specification assumes that $\boldsymbol{\Sigma}^{-1}$ follows a vague $\mathrm{Wishart}(\boldsymbol{\Omega}, \gamma)$ distribution having a mean of $\gamma\boldsymbol{\Omega}^{-1}$ and degrees of freedom parameter γ set equal to the dimension of $\boldsymbol{\Sigma}$. In our data example, this second option may be less

suitable because many treatments were investigated from a single AD study, so the data struggle to identify all the elements of an inverse-Wishart-distributed (but otherwise unstructured) $\boldsymbol{\Sigma}$ matrix.

For the treatment-by-covariate interaction random effects, we can assume a similar equicorrelated model such that $\boldsymbol{\beta}_i^{\mathrm{AD}} = (\beta_{i1}^{\mathrm{AD}}, \ldots, \beta_{iK}^{\mathrm{AD}}) \sim^{\mathrm{ind}} N(\mathbf{b}^{\mathrm{AD}}, \boldsymbol{\Sigma}_x)$, where $\mathbf{b}^{\mathrm{AD}} = (b_1^{\mathrm{AD}}, \ldots, b_K^{\mathrm{AD}})^{\mathrm{T}}$, $\boldsymbol{\Sigma}_x = \mathrm{diag}(\tau_x^2) + \rho_x \tau_x^2 I(i \neq j)$ with $\rho_x \sim \mathrm{uniform}(0, 1)$ and $\tau_x \sim \mathrm{uniform}(0, 10)$. In addition, we assume that the mean interaction effects b_k^{AD} are exchangeable, e.g. $b_k^{\mathrm{AD}} \sim^{\mathrm{ind}} N(b^{\mathrm{AD}}, \tau_b^2)$, where b^{AD} and τ_b follow vague normal and uniform prior distributions respectively. We do not consider an unstructured Wishart prior for $\boldsymbol{\Sigma}_x^{-1}$, again because of identifiability concerns.

When we have only IPD, we can use the arm-based covariate-adjusted model that was proposed by Hong *et al.* (2015). The likelihood can be assumed to be likelihood (4), and the individual level covariates are now x_{ijk}. The unknown mean response μ_{ijk}^{IPD} can now be modelled as

$$\mu_{ijk}^{\mathrm{IPD}} = \theta_k^{\mathrm{IPD}} + \nu_{ik}^{\mathrm{IPD}} + \beta_{ik}^{\mathrm{IPD}} x_{ijk}, \tag{6}$$

where ν_{ik}^{IPD} and $\beta_{ik}^{\mathrm{IPD}}$ have similar interpretations to those of ν_{ik}^{AD} and β_{ik}^{AD} above with similar hyperprior specifications. We assume that $\boldsymbol{\nu}_i^{\mathrm{IPD}} = (\nu_{i1}^{\mathrm{IPD}}, \ldots, \nu_{iK}^{\mathrm{IPD}})^{\mathrm{T}} \sim^{\mathrm{ind}} N(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\beta}_i^{\mathrm{IPD}} = (\beta_{i1}^{\mathrm{IPD}}, \ldots, \beta_{iK}^{\mathrm{IPD}}) \sim^{\mathrm{ind}} N(\mathbf{b}^{\mathrm{IPD}}, \boldsymbol{\Psi}_x)$, where $\mathbf{b}^{\mathrm{IPD}} = (b_1^{\mathrm{IPD}}, \ldots, b_K^{\mathrm{IPD}})^{\mathrm{T}}$, $\boldsymbol{\Psi} = \mathrm{diag}(\psi^2) + \varphi\psi^2 I(i \neq j)$, $\boldsymbol{\Psi}_x = \mathrm{diag}(\psi_x^2) + \varphi_x \psi_x^2 I(i \neq j)$ and $b_k^{\mathrm{IPD}} \sim^{\mathrm{ind}} N(b^{\mathrm{IPD}}, \psi_b^2)$. We then assign vague hyperprior distributions to ψ, φ, ψ_x, φ_x, b^{IPD} and ψ_b.

Model (6) can be generalized to investigate the existence of an ecological fallacy by dividing the treatment-by-covariate interactions into two factors: *within-trial* and *across-trial* interactions (Riley *et al.*, 2008; Saramago *et al.*, 2012; Donegan *et al.*, 2013; Cooper *et al.*, 2009; Hong *et al.*, 2015). The covariate-adjusted model differentiating the two different interaction sources can be written as

$$\mu_{ijk}^{\mathrm{IPD}} = \theta_k^{\mathrm{IPD}} + \nu_{ik}^{\mathrm{IPD}} + \beta_{ik}^{\mathrm{w}}(x_{ijk} - \bar{x}_{ik}) + \beta_{ik}^{\mathrm{a}} \bar{x}_{ik}, \tag{7}$$

where $\bar{x}_{ik}$ is the mean covariate for patients who are assigned to treatment k in study i, β_{ik}^{w} is the within-trial interaction and β_{ik}^{a} is the across-trial interaction. Note that model (6) is a special case of model (7) when assuming that $\beta_{ik}^{\mathrm{w}} \equiv \beta_{ik}^{\mathrm{a}}$. We can assign prior distributions to β_{ik}^{w} and β_{ik}^{a} similarly to perhaps for $\beta_{ik}^{\mathrm{IPD}}$, with equicorrelated covariance matrices. Donegan *et al.* (2013) and Hong *et al.* (2015) pointed out that computational implementations of this model appear not to converge well. To enhance convergence, we can reduce model complexity by assuming fixed within-trial and across-trial interactions, such as $\beta_{ik}^{\mathrm{w}} \equiv b^{\mathrm{w}}$ and $\beta_{ik}^{\mathrm{a}} \equiv b^{\mathrm{a}}$, when such assumptions are plausible. A large difference between b^{w} and b^{a} suggests a potential ecological fallacy in the data.

3.2. Adaptive integration of aggregated level data and individual level patient data

3.2.1. Existing models

When combining AD and IPD, as a starting part for integrative models, we can simply assume that the treatment effects in AD and IPD are identical and exchangeable across studies. In equations (5) and (6), we can replace θ_k^{AD} and θ_k^{IPD} with θ_k, and β_{ik}^{AD} and $\beta_{ik}^{\mathrm{IPD}}$ with β_{ik}. When considering model (7) instead of model (6), we can replace θ_k^{AD} and θ_k^{IPD} with θ_k and assume that β_{ik}^{AD} and β_{ik}^{a} are identical. We estimate β_{ik}^{w} by using only IPD, because only IPD can inform within-study interactions. We then assign a vague normal prior to θ_k; exchangeable normal shrinkage priors assigned to β_{ik}, β_{ik}^{a} and β_{ik}^{w} complete the specification.

3.2.2. Power priors

As mentioned near equation (1), we consider the IPD as the primary data and the AD as the auxiliary data in our setting, because AD are typically less informative and may be biased relative to IPD. The power prior parameter decides how much the AD likelihood contributes to the posterior distribution of the parameters of interest. We assume identicality of the treatment effects between AD and IPD for each treatment k, as in Section 3.2.1. Suppose that $\boldsymbol{\Theta}$ is a vector having all model parameters including θ_k and β_{ik}; then the power prior of $\boldsymbol{\Theta}$ given the likelihood of $\{y_{ik}^{\mathrm{AD}}\}$ can be written as

$$\pi(\boldsymbol{\Theta} \mid \{y_{ik}^{\mathrm{AD}}\}, \alpha_{ik}) \propto \left\{ \prod_{i=1}^{I^{\mathrm{AD}}} L(\boldsymbol{\Theta} \mid y_{ik}^{\mathrm{AD}})^{\alpha_{ik}} \right\} \pi_0(\boldsymbol{\Theta}), \tag{8}$$

where $\pi_0(\boldsymbol{\Theta})$ is often assumed to be uniform. Here, α_{ik} governs the effect of AD on the analysis, ranging from 0 (no influence) to 1 (fully combining), for $k = 1, \ldots, K$ and study i. In addition, we use study- and treatment-specific power parameters α_{ik} instead of a universal α_0, and this allows more flexibility of NMA integrative models. Neelon and O'Malley (2010) showed that estimating the posterior distribution of the power parameter in a regression setting does not correctly capture the degree of downweighting of the auxiliary data, and they recommended sensitivity analyses over a range of fixed power parameters. We heed this warning and take the α_{ik} to be fixed. The choice of power parameters can be a subjective decision based on expert opinion. When integrating AD and IPD, we can simply downweight the evidence from AD by setting $\alpha_{ik} < 1$ for $i = 1, \ldots, I^{\mathrm{AD}}$ on the basis of the age or perceived quality of each trial. Finally, the posterior distribution of $\boldsymbol{\Theta}$ can be written as

$$q(\boldsymbol{\Theta} \mid \{y_{ijk}^{\mathrm{IPD}}\}, \{y_{ik}^{\mathrm{AD}}\}, \alpha_{ik}) \propto \left\{ \prod_{i=1}^{I^{\mathrm{AD}}} L(\boldsymbol{\Theta} \mid y_{ik}^{\mathrm{AD}})^{\alpha_{ik}} \right\} \left\{ \prod_{i=I^{\mathrm{AD}}+1}^{I} \prod_{j=1}^{n_i} L(\boldsymbol{\Theta} \mid y_{ijk}^{\mathrm{IPD}}) \right\} \pi_0(\boldsymbol{\Theta}). \tag{9}$$

Hyperprior specifications for all parameters are the same as described in Section 3.2.1.

Fig. 2(a) shows the directed acyclic graph (DAG) for the integrative NMA model using power priors specifically for our diabetes data. In this DAG, we depict only an unadjusted model, with circle nodes for unknown parameters and square nodes for observed data or fixed quantities. A DAG for an adjusted model can have additional nodes for treatment-by-covariate interactions. In our diabetes data, we could completely rely on AD trials for treatments for which evidence is available from AD only (treatments 5–27) by setting $\alpha_{i5}, \ldots, \alpha_{i,27} \equiv 1$ and then downweight AD trials for treatments 1–4, for which evidence comes from both AD and IPD, i.e. set $\alpha_{i1}, \ldots, \alpha_{i4} < 1$. Because only IPD provide information about the last drug in Table 1, sulphonylurea, there is no $L(\boldsymbol{\Theta} \mid y_{ik}^{\mathrm{AD}})$ term when $k = 28$.

3.2.3. Commensurate priors

To incorporate the commensurate prior (2), we start with models (5) and (6). We add the IPD and AD superscripts back onto our parameter vectors and assign commensurate priors for θ_k^{IPD} and b_k^{IPD} to borrow information from the AD as

$$\theta_k^{\mathrm{IPD}} \overset{\mathrm{ind}}{\sim} N(\theta_k^{\mathrm{AD}}, \eta_k^2) \quad \text{and} \quad b_k^{\mathrm{IPD}} \overset{\mathrm{ind}}{\sim} N(b_k^{\mathrm{AD}}, \eta_{bk}^2) \qquad \text{for } k = 1, \ldots, 4, \tag{10}$$

thus centring the IPD treatment effect of interest near the corresponding AD effect for treatments having both AD and IPD sources. Previous work in this area has often used *spike-and-slab* hyperpriors on the components of the $\boldsymbol{\eta}$-vectors (Hobbs *et al.*, 2012; Murray *et al.*, 2014, 2015, 2016). This distribution places probability p_k on a 'slab' of values that are close to the origin,

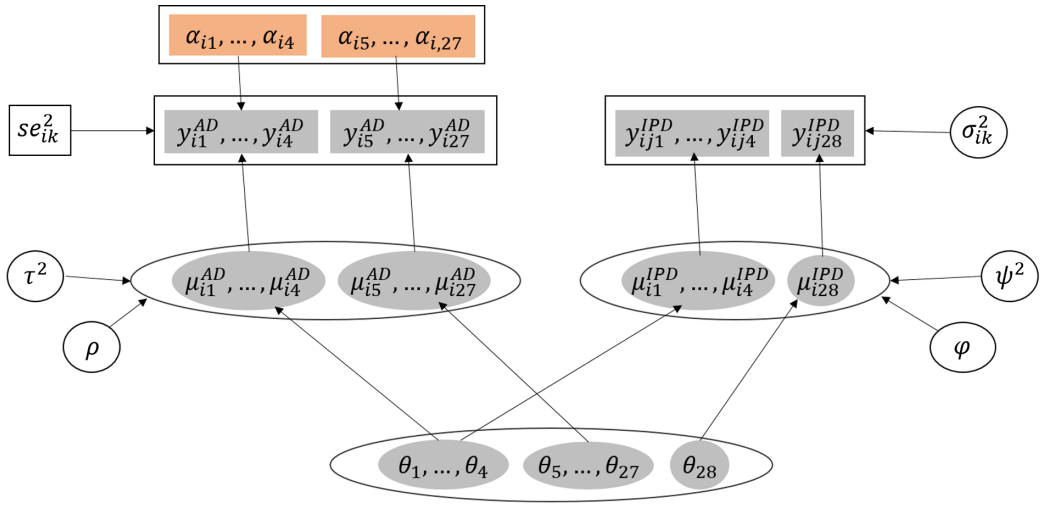

Fig. 2. DAGs for Bayesian NMA hierarchical models integrating AD and IPD adaptively (a) with power priors and (b) with commensurate priors: ○, unknown parameters; □, observed data or fixed quantities

and the remaining $1 - p_k$ at a point mass (the 'spike') or a small collection or larger values designed to encourage borrowing from the auxiliary data. In our case, the treatment-specific spike-and-slab hyperprior that we select for our precisions $1/\eta_k^2$ is

$$\frac{1}{\eta_k^2} \sim \begin{cases} N(20, 1) & \text{if } c_k = 0, \\ \text{gamma}(0.1, 0.1)I(0.1, 5) & \text{if } c_k = 1, \end{cases} \tag{11}$$

where $c_k \sim^{\text{IID}} \text{Bernoulli}(p_k)$ and we can either set $p_k = 0.5$ or assign it an informative hyperprior distribution. A similar hyperprior is assigned to $1/\eta_{bk}^2$ given slab probabilities p_{bk}. Here, the 'quasi-spike' at 20 corresponds to high borrowing of AD information, whereas the 'slab'

over (0.1, 5) corresponds to essentially disregarding the AD on this parameter. The hyperprior specification for other parameters is the same as described in Section 3.1.

Fig. 2(b) shows the DAG for the integrative NMA model using commensurate priors for our diabetes example. Commensurate priors are assigned to parameters for treatment effects and treatment-by-covariate interactions for treatments 1–4. The same parameters for treatments 5–27 are estimated by using only AD; the parameters for treatment 28 are estimated by using only IPD. After estimating all parameters, the posterior Bayesian inference is made on the parameter vector $(\theta_1^{\mathrm{IPD}}, \ldots, \theta_4^{\mathrm{IPD}}, \theta_5^{\mathrm{AD}}, \ldots, \theta_{27}^{\mathrm{AD}}, \theta_{28}^{\mathrm{IPD}})$.

3.2.4. Summary of models

Table 2 shows model names and detailed hyperprior specifications of all unadjusted and adjusted models fitted to our diabetes data example. We use similar models in our simulation studies that are discussed in Section 4.1.

For covariate-unadjusted models and covariate-adjusted models assuming $b^{\mathrm{w}} = b^{\mathrm{a}}$ (i.e. no ecological fallacy), we first fit AD-only models using 29 AD studies (denoted by 'AD1') and using the 29 AD studies and summary statistics calculated from 12 IPD studies (denoted by 'AD2'). Then, we fit an IPD-only model using the 12 IPD studies, and we call this model 'IPD'. For these three models, we adopt equicorrelated models for the random treatment effects $\boldsymbol{\nu}_i^{\mathrm{AD}}$ and $\boldsymbol{\nu}_i^{\mathrm{IPD}}$.

Moving to integrative models, the first basic integrative model fits both AD1 and IPD models by assuming that $\theta_k^{\mathrm{AD}} \equiv \theta_k^{\mathrm{IPD}} \equiv \theta_k$ and $\beta_{ik}^{\mathrm{AD}} \equiv \beta_{ik}^{\mathrm{IPD}} \equiv \beta_{ik}$, where the second assumption applies only to covariate-adjusted models, as described in Section 3.2.1. We call this model 'I-vague' because it does not use a special class of priors but assigns vague priors to all model parameters. For integrative models adapting power priors, we consider four settings of α_{ik} (denoted by 'I-pp1, I-pp2, I-pp3' and 'I-pp4'). The model I-pp1 employs $\alpha_{ik} = 0.8$ for the AD studies with drug naive participants, and $\alpha_{ik} = 0.5$ otherwise. Here, we implicitly assume that drug naive AD studies are slightly more reliable than AD studies with participants who have previously taken other related medications. Other factors such as the year of publication or study type (e.g. double blind *versus* single blind) can also be used to determine study reliability. However, in our diabetes data, the AD studies are fairly similar in terms of publication year and study design, because the publication years are between 2005 and 2009 except for one study that was published in 1998, and all studies were double blind. Models I-pp2, I-pp3 and I-pp4 use α_{ik} equal to 0.2, 0.5 or 0.8 for AD studies investigating the four drugs ($k = 1, \ldots, 4$) having evidence available from both AD and IPD, and use $\alpha_{ik} = 1$ for the other AD studies. Finally, we consider models with two different commensurate prior specifications (denoted by 'I-cp1' and 'I-cp2'): I-cp1 assumes a common η across treatments by replacing η_k, p_k and c_k in expression (11) with η, p and c, whereas I-cp2 assumes a treatment-specific η_k as in expression (11). We assign non-informative priors to p and p_k, as we do not have strong prior information about them. For all covariate-adjusted models except those using power priors, we employ exchangeable normal shrinkage priors to covariate-by-treatment interactions, such as $\boldsymbol{\beta}_i \sim N(\mathbf{b}, \boldsymbol{\Sigma}_x)$, where $b_k \sim N(b, \tau_b^2)$. For covariate-adjusted models with power priors, we assume a common fixed covariate-by-treatment interaction $\beta_{ik} \equiv b$ to ensure model convergence.

For the adjusted models assuming $b^{\mathrm{w}} \neq b^{\mathrm{a}}$, we first fit an IPD-only model, denoted by 'IPDef', where 'ef' stands for ecological fallacy. We denote integrative models by 'Ief' and consider only a subset of our selected priors (Ief-vague, Ief-pp1, Ief-pp3 and Ief-cp2), because the general findings of the Ief models are broadly similar to those of the corresponding adjusted models assuming no ecological fallacy.

Table 2. Model specification and fit statistics

Data	*Model*	*Hyperprior specification*	*Number of studies*	$\bar{D}$	p_D	*DIC*
Covariate-unadjusted models						
AD	AD1	$\boldsymbol{\nu}_i^{\mathrm{AD}} \sim^{\mathrm{ind}} N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\tau^2) + \rho\tau^2 I(i \neq j)$	29	−118.0	71.6	−46.4
	AD2	Same as the covariate-unadjusted AD1 model; summary statistics of IPD are combined with AD	41	−157.8	96.5	−61.3
IPD	IPD	$\boldsymbol{\nu}_i^{\mathrm{IPD}} \sim^{\mathrm{ind}} N(0, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \mathrm{diag}(\psi^2) + \varphi\psi^2 I(i \neq j)$	12	11209.0	54.5	11263.5
IPD+AD	I-vague	$\theta_k^{\mathrm{AD}} \equiv \theta_k^{\mathrm{IPD}} \equiv \theta_k$	41	11090.5	123.9	11214.3
	Power prior approach					
	I-pp1	$\alpha_{ik} = 0.8$ for drug naive AD trial i; $\alpha_{ik} = 0.5$ otherwise	41	1431136.0	121.3	1431257.3
	I-pp2	$\alpha_{ik} = 0.2$ for $k = 1, \ldots, 4$	41	1431050.1	115.3	1431165.4
	I-pp3	$\alpha_{ik} = 0.5$ for $k = 1, \ldots, 4$	41	1431047.2	109.3	1431156.5
	I-pp4	$\alpha_{ik} = 0.8$ for $k = 1, \ldots, 4$	41	1431019.0	241.4	1431260.4
	Commensurate prior approach					
	I-cp1	$\eta^{-2} \sim$ spike-and-slab prior, where $c \sim$ Bernoulli(0.5)	41	11090.7	123.7	11214.4
	I-cp2	$\eta_k^{-2} \sim$ spike-and-slab prior, where $c_k \sim$ Bernoulli(0.5)	41	11090.7	123.7	11214.4
Covariate-adjusted models						
AD	AD1	$\boldsymbol{\beta}_i^{\mathrm{AD}} \sim^{\mathrm{ind}} N(b_k^{\mathrm{AD}}, \boldsymbol{\Sigma}_x)$, where $b_k^{\mathrm{AD}} \sim N(b^{\mathrm{AD}}, \tau_b^2)$	29	−118.0	71.9	−46.1
	AD2	Same as the covariate-adjusted AD1 model; summary statistics of IPD are combined with AD	41	−157.4	97.7	−59.7
IPD ($b^{\mathrm{w}} = b^{\mathrm{a}}$)	IPD	$\boldsymbol{\beta}_i^{\mathrm{IPD}} \sim^{\mathrm{ind}} N(b_k^{\mathrm{IPD}}, \boldsymbol{\Psi}_x)$, where $b_k^{\mathrm{IPD}} \sim N(b^{\mathrm{IPD}}, \psi_b^2)$	12	9937.2	97.1	10034.2
IPD+AD ($b^{\mathrm{w}} = b^{\mathrm{a}}$)	I-vague	$b_k^{\mathrm{AD}} \equiv b_k^{\mathrm{IPD}} \equiv b_k$	41	9819.3	169.2	9988.6
	Power prior approach					
	I-pp1	$\alpha_{ik} = 0.8$ for drug naive AD trial i; $\alpha_{ik} = 0.5$ otherwise	41	1430205.9	121.1	1430327.1
	I-pp2	$\alpha_{ik} = 0.2$ for $k = 1, \ldots, 4$	41	1430124.9	116.6	1430241.5
	I-pp3	$\alpha_{ik} = 0.5$ for $k = 1, \ldots, 4$	41	1430108.8	150.7	1430259.5
	I-pp4	$\alpha_{ik} = 0.8$ for $k = 1, \ldots, 4$	41	1430086.3	171.4	1430257.7
	Commensurate prior approach					
	I-cp1	$\eta_b^{-2} \sim$ spike-and-slab prior, where $c_b \sim$ Bernoulli(0.5)	41	9819.5	170.1	9989.6
	I-cp2	$\eta_{bk}^{-2} \sim$ spike-and-slab prior, where $c_{bk} \sim$ Bernoulli(0.5)	41	9819.4	168.6	9988.0
IPD ($b^{\mathrm{w}} \neq b^{\mathrm{a}}$)	IPDef	$\beta_{ik}^{\mathrm{w}} \equiv b^{\mathrm{w}}$ and $\beta_{ik}^{\mathrm{a}} \equiv b^{\mathrm{a}}$	12	10282.1	56.4	10338.5
IPD+AD ($b^{\mathrm{w}} \neq b^{\mathrm{a}}$)	Ief-vague	$\beta_{ik}^{\mathrm{w}} \equiv b^{\mathrm{w}}$ and $\beta_{ik}^{\mathrm{a}} \equiv b^{\mathrm{a}} \equiv b^{\mathrm{AD}}$	41	10163.7	125.8	10289.5
	Power prior approach					
	Ief-pp1	$\beta_{ik}^{\mathrm{w}} \equiv b^{\mathrm{w}}$ and $\beta_{ik}^{\mathrm{a}} \equiv b^{\mathrm{a}} \equiv b^{\mathrm{AD}}$	41	1430207.4	115.1	1430322.6
	Ief-pp3	$\beta_{ik}^{\mathrm{w}} \equiv b^{\mathrm{w}}$ and $\beta_{ik}^{\mathrm{a}} \equiv b^{\mathrm{a}} \equiv b^{\mathrm{AD}}$	41	1430117.1	126.5	1430243.6
	Commensurate prior approach					
	Ief-cp2	$b^{\mathrm{a}} \sim N(b^{\mathrm{AD}}, \eta_b^2)$ and $\eta_b^{-2} \sim$ spike-and-slab prior, where $c_b \sim$ Bernoulli(0.5)	41	10163.5	126.9	10290.4

3.3. Surface under the cumulative ranking curve

One of the advantages of Bayesian NMA is that it can make inference about treatment effects by ranking all treatments in terms of the outcome (Salanti *et al.*, 2011; Hong *et al.*, 2013). We can find the best treatment by using rank probabilities. Salanti *et al.* (2011) proposed a simple numerical summary of the cumulative rank probabilities for each treatment called the *surface under the cumulative ranking* (SUCRA). SUCRA estimates the area under the step function of the cumulative rank probabilities. Suppose that $P_k(r)$ is the posterior probability of treatment k being the rth in rank, where $r = 1$ is best. The cumulative probabilities are calculated as $\text{Cum}_k(r) = \text{Prob}(k \text{ is ranked } r \text{ or less}) = \Sigma_{h=1}^{r} P_k(h)$, whence SUCRA is defined as

$$\text{SUCRA}_k = \sum_{r=1}^{K-1} \frac{\text{Cum}_k(r)}{K-1}.$$

A large SUCRA value (i.e. close to 1) indicates better performance in terms of outcomes of interest. SUCRA enables us to compare all treatments by using a single number instead of comparing a vector of $P_k(r)$ for each treatment, and as such we make our inferences about treatment effects by using SUCRA.

3.4. Computation

For both simulation studies and diabetes data analyses, we use the R package `R2jags` (Su and Yajima, 2015) implementing Bayesian analysis using Markov chain Monte Carlo sampling in JAGS (Plummer, 2003). For data analyses, a total of 50 000 samples were employed for inference from two parallel chains by saving every other sample (i.e. the thinning interval for our Markov chain Monte Carlo chains is 2) after a 50 000-iteration burn-in for each chain. For our simulation studies, we used a total of 20 000 samples after a 20 000-iteration burn-in from a single chain. To investigate convergence of all chains, we monitored trace plots and $\hat{R}$-statistics (Gelman and Rubin, 1992) for all parameters. JAGS code for selected integrative models is available from

`http://wileyonlinelibrary.com/journal/rss-datasets`

4. Simulation studies

4.1. Settings

4.1.1. Covariate-unadjusted model

In this simulation study, we shall compare the performance of the power and commensurate prior approaches when a few low quality studies are included. We investigate this with unadjusted models and calculate bias, mean-squared error (MSE) and coverage probability of treatment effect estimates under five scenarios. We estimate the coverage probability as the empirical proportion of times that a parameter's 95% credible interval includes its true value out of 1000 replications of simulated data, where the target is the nominal level, 0.95. Here, the 95% credible intervals are calculated by using 0.025 and 0.975 posterior quantiles. Table 3 displays the various data structures that were considered. Panel (a) displays the base structure, which is assumed to include only high quality studies comparing three treatments (A, B and C): 15 AD studies and four IPD studies. Here, we want to mimic a feature of our diabetes data example, which features far fewer IPD studies than AD studies. Structure 2 in panel (b) assumes that three AD studies (20%) comparing treatments A and B have low quality, either because the outcome that is associated with treatment B is heterogeneous (noisy) or the effect of treatment B is biased. Similarly, structure 3 in panel (c) assumes that one A–B IPD study (25%) is of low quality.

Table 3. Data structure for simulation studies

Data	*A*	*B*	*C*	*Number of studies*
(a) Structure 1				
AD	✓	✓		5
	✓		✓	5
		✓	✓	5
IPD	✓	✓		2
	✓		✓	2
(b) Structure 2				
AD	✓	✓		2
	✓	✓		3 low quality
	✓		✓	5
		✓	✓	5
IPD	✓	✓		2
	✓		✓	2
(c) Structure 3				
AD	✓	✓		5
	✓		✓	5
		✓	✓	5
IPD	✓	✓		1
	✓	✓		1 low quality
	✓		✓	2

We consider five scenarios:

(a) structure 1 with high quality studies;
(b) structure 2 with noisy outcomes from the three low quality AD studies;
(c) structure 2 with a biased treatment effect for the three low quality AD studies;
(d) structure 3 with noisy outcomes from the one low quality IPD study;
(e) structure 3 with a biased treatment effect from the one low quality IPD study.

We denote these five scenarios as cases 1.1–1.5.

We first generate individual level data for the total number of 19 studies, and then we calculate the sample mean and sample standard deviation for the 15 AD studies. We assume that the number of subjects in each arm of each study is 100. The data are generated by $y_{ijk} \sim^{\text{ind}} N(\theta_k, \sigma^2)$ for $i = 1, \ldots, 19$, $j = 1, \ldots, 200$ and $k \equiv \text{A}, \text{B}, \text{C}$. Here, θ_k is the true effect for treatment k. To generate 'high quality' studies, we set $(\theta_\text{A}, \theta_\text{B}, \theta_\text{C}) = (0, 1, 2)$ and $\sigma = 1$. For our 'low quality' studies with noisy outcomes, we set $\sigma = 2$ for generating outcome associated with treatment B (i.e. $y_{ijB} \sim^{\text{ind}} N(1, 2^2)$). For low quality studies with biased treatment effect for B, we set $\boldsymbol{\theta} = (0, 1.5, 2)^\text{T}$ and $\sigma = 1$. Our simulations replicate the data $\{y_{ijk}\}$ 1000 times.

We fit seven unadjusted models: AD1, IPD, I-vague, I-cp2 and I-pp for three settings, $\alpha_{ik} \equiv \alpha = 0.2$, 0.5, 0.8, denoted by I-pp5, I-pp6 and I-pp7. For case 1.2 and case 1.3, we fit three additional power prior models with $\alpha_{ik} = 0.2, 0.5, 0.8$ for $k \equiv \text{B}$, and $\alpha_{ik} = 1$ otherwise, denoted by I-pp8, I-pp9 and I-pp10. Note that the model names for power priors in the simulation study are different from those used in the data analysis as power prior specifications are slightly different between the simulation study and the data analysis. In addition, because only these two scenarios include low quality AD studies, we want to investigate different behaviours of common power parameters across all studies and study-specific power parameters.

4.1.2. Covariate-adjusted model

In our second simulation study, we compare the performance of our adjusted models in terms of bias, MSE and coverage probability of treatment effect estimates under two scenarios:

(a) only the AD studies have evidence of an ecological fallacy (denoted by case 2.1) and
(b) both AD and IPD studies have evidence of ecological fallacy (denoted by case 2.2).

We no longer consider noisy outcomes or biased treatment effects in a subset of studies. Instead, here we focus on the situation where AD and IPD do not agree on the size of treatment-by-covariate interactions (i.e. an ecological fallacy exists), although they agree on the size of treatment effects.

We consider structure 1 in Table 3 with a single covariate $x_{ijk} \sim^{\text{ind}} N(1,1)$. We generate AD and IPD studies as follows:

$$\begin{aligned} y_{ik}^{\text{AD}} &= \bar{y}_{ik} + \beta_k^{\text{AD}} \bar{x}_{ik}, \\ y_{ijk}^{\text{IPD}} &\overset{\text{ind}}{\sim} N\{\theta_k + \beta_k^{\text{w}}(x_{ijk} - \bar{x}_{ik}) + \beta_k^{\text{a}} \bar{x}_{ik}, 1\}, \end{aligned} \tag{12}$$

where $\bar{y}_{ik}$ and $\bar{x}_{ik}$ are sample means of $y_{ijk} \sim^{\text{ind}} N(\theta_k, 1)$ and x_{ijk} for treatment k in each study i respectively. Here, θ_k is the true treatment effect in both AD and IPD studies, β^{AD} is the true treatment-by-covariate interaction effect in AD studies and β_k^{w} and β_k^{a} are coefficients for within- and across-trial interactions respectively, for the kth treatment. We again replicate the simulated data 1000 times.

For case 2.1, we set $(\beta_{\text{A}}^{\text{AD}}, \beta_{\text{B}}^{\text{AD}}, \beta_{\text{C}}^{\text{AD}}) = (0,0,0)$, $(\beta_{\text{A}}^{\text{w}}, \beta_{\text{B}}^{\text{w}}, \beta_{\text{C}}^{\text{w}}) = (1,1,1)$ and $(\beta_{\text{A}}^{\text{a}}, \beta_{\text{B}}^{\text{a}}, \beta_{\text{C}}^{\text{a}}) = (1,1,1)$. Here, $\beta_k^{\text{w}} = \beta_k^{\text{a}}$ but β_k^{AD} and β_k^{a} are different, indicating that only the AD studies have evidence of an ecological fallacy. For case 2.2, we set $(\beta_{\text{A}}^{\text{AD}}, \beta_{\text{B}}^{\text{AD}}, \beta_{\text{C}}^{\text{AD}}) = (0,0,0)$, $(\beta_{\text{A}}^{\text{w}}, \beta_{\text{B}}^{\text{w}}, \beta_{\text{C}}^{\text{w}}) = (1,1,1)$ and $(\beta_{\text{A}}^{\text{a}}, \beta_{\text{B}}^{\text{a}}, \beta_{\text{C}}^{\text{a}}) = (0,0,0)$. Here, $\beta_k^{\text{w}} \neq \beta_k^{\text{a}}$ but β_k^{AD} and β_k^{a} are the same, indicating that both AD and IPD studies have evidence of ecological fallacy.

We fit five adjusted models assuming that $\beta_k^{\text{w}} = \beta_k^{\text{a}}$, AD1, IPD, I-vague, I-pp6 (assigning $\alpha = 0.5$) and I-cp2, and four adjusted models assuming that $\beta_k^{\text{w}} \neq \beta_k^{\text{a}}$, IPDef, Ief-vague, Ief-pp with $\alpha = 0.5$ and Ief-cp2. We assume a common fixed covariate-by-treatment interaction in all models, as the simulated data have a common interaction effect size across treatments.

4.2. Results

4.2.1. Covariate-unadjusted model

Table 4 shows bias, MSE and coverage probability of θ_{B} from the simulation study with covariate-unadjusted models. In case 1.3, when partial AD studies have biased results, all models except IPD, I-pp8 and I-cp2 provide larger bias and MSE with poorer coverage probability than estimates under case 1.1. In case 1.3, the AD1 model gives the largest bias (0.150) and the bias decreases in integrative models. In addition, in models using power priors, the bias grows smaller as we use smaller α (0.112, 0.111 and 0.091 for I-pp7, I-pp6 and I-pp5, when $\alpha = 0.8, 0.5, 0.2$ respectively) and even smaller when adopting study- and treatment-specific α_{ik} in I-pp8, I-pp9 and I-pp10. Model I-cp2 provides the smallest bias and MSE with the best coverage among all integrative models.

In case 1.5, where a single IPD study has biased results, model AD1 provides unbiased estimates with small MSE and good coverage probability, as we expected. Model IPD gives the largest bias among all models, but its coverage probabilities are not as poor as those for model AD1 under case 1.3. Models I-vague and I-pp7 perform better in terms of bias, MSE and coverage probability than the other models. When using power priors, in contrast with the results under case 1.3, bias decreases as we borrow more information from the AD studies (i.e.

Table 4. Bias, MSE and coverage probability of θ_B from the simulation study with covariate-unadjusted models

Unadjusted model	*Parameter*	*Results for the following cases:*				
		Case 1.1	*Case 1.2*	*Case 1.3*	*Case 1.4*	*Case 1.5*
AD1	Bias	0.000	0.000	0.150	0.000	0.000
	MSE	0.001	0.001	0.024	0.001	0.001
	Coverage	0.974	0.973	0.109	0.974	0.974
IPD	Bias	−0.004	−0.004	−0.004	−0.005	0.245
	MSE	0.005	0.005	0.005	0.008	0.065
	Coverage	0.995	0.995	0.995	0.993	0.879
I-vague	Bias	0.000	0.000	0.111	0.000	0.015
	MSE	0.001	0.001	0.013	0.001	0.001
	Coverage	0.971	0.972	0.360	0.972	0.948
I-pp5, $\alpha_{ik} \equiv \alpha = 0.2$	Bias	−0.001	−0.002	0.091	−0.001	0.055
	MSE	0.001	0.001	0.009	0.001	0.004
	Coverage	1.000	0.999	0.750	1.000	0.990
I-pp6, $\alpha_{ik} \equiv \alpha = 0.5$	Bias	−0.001	−0.001	0.111	0.000	0.025
	MSE	0.001	0.001	0.013	0.001	0.002
	Coverage	0.992	0.992	0.345	0.991	0.984
I-pp7, $\alpha_{ik} \equiv \alpha = 0.8$	Bias	0.001	0.002	0.112	0.000	0.017
	MSE	0.001	0.001	0.013	0.001	0.001
	Coverage	0.979	0.985	0.349	0.980	0.962
I-pp8, $\alpha_{ik} = 0.2$ for $k \equiv B$	Bias	—	0.000	0.040	—	—
	MSE	—	0.001	0.003	—	—
	Coverage	—	0.976	0.870	—	—
I-pp9, $\alpha_{ik} = 0.5$ for $k \equiv B$	Bias	—	0.000	0.085	—	—
	MSE	—	0.001	0.008	—	—
	Coverage	—	0.976	0.477	—	—
I-pp10, $\alpha_{ik} = 0.8$ for $k \equiv B$	Bias	—	0.000	0.105	—	—
	MSE	—	0.001	0.012	—	—
	Coverage	—	0.973	0.373	—	—
I-cp2	Bias	−0.003	−0.004	0.013	−0.004	0.193
	MSE	0.004	0.004	0.004	0.006	0.041
	Coverage	0.996	0.996	0.995	0.995	0.902

α is close to 1) because the AD studies are helping to estimate the correct treatment effect in this scenario. This shows that fully borrowing information from high quality AD studies helps to obtain a better estimate when our database includes some poorly conducted IPD studies. The performance of the commensurate prior model is poorer in case 1.5 than in case 1.3. This is because commensurate prior models rely more on IPD than AD, as our implementation assumes that IPD are the gold standard.

When partial AD or IPD studies have noisy outcomes (case 1.2 and case 1.4), all models perform reasonably well. Across all scenarios, commensurate prior models always select the spike component of the prior regardless of noisy or biased results in the AD studies. The bias, MSE and coverage probabilities for θ_A and θ_C do not change much across different scenarios (the results are not shown).

4.2.2. Covariate-adjusted model

Table 5 shows bias, MSE and coverage probability of θ_B, and the mean of covariate-by-treatment interactions from the simulation study with covariate-adjusted models. In Table 5, we show

Table 5. Bias, MSE and coverage probability of θ_B and mean of covariate-by-treatment interactions from the simulation study with covariate-adjusted models

Adjusted model	*Parameter*	*Results for case 2.1*		*Results for case 2.2*	
		$\beta_k^a = \beta_k^w$	$\beta_k^a \neq \beta_k^w$	$\beta_k^a = \beta_k^w$	$\beta_k^a \neq \beta_k^w$
AD1	Bias(θ_B)	−0.010	—	−0.010	—
	MSE(θ_B)	0.045	—	0.045	—
	Coverage(θ_B)	0.960	—	0.959	—
	Mean(b or b^w)	0.012	—	0.012	—
	Mean(b^a)	—	—	—	—
IPD or IPDef	Bias(θ_B)	−0.002	0.010	−1.000	0.008
	MSE(θ_B)	0.006	0.376	1.011	0.376
	Coverage(θ_B)	0.994	0.997	0.000	0.996
	Mean(b or b^w)	1.002	1.002	0.999	1.002
	Mean(b^a)	—	0.989	—	−0.009
I-vague or Ief-vague	Bias(θ_B)	−0.977	−0.045	−0.976	−0.004
	MSE(θ_B)	0.958	0.045	0.956	0.034
	Coverage(θ_B)	0.000	0.961	0.000	0.964
	Mean(b or b^w)	0.984	1.002	0.980	1.002
	Mean(b^a)	—	0.050	—	0.005
I-pp or Ief-pp	Bias(θ_B)	−0.981	−0.072	−0.981	0.000
	MSE(θ_B)	0.965	0.050	0.964	0.035
	Coverage(θ_B)	0.000	0.988	0.000	0.990
	Mean(b or b^w)	0.988	1.002	0.984	1.002
	Mean(b^a)	—	0.058	—	0.006
I-cp2 or Ief-cp2	Bias(θ_B)	−0.005	0.026	−0.927	−0.005
	MSE(θ_B)	0.006	0.039	0.871	0.039
	Coverage(θ_B)	0.993	0.992	0.151	0.986
	Mean(b or b^w)	1.001	1.002	0.988	1.002
	Mean(b^a)	—	0.966	—	0.004

results from models assuming that $\beta_k^a = \beta_k^w$ and assuming that $\beta_k^a \neq \beta_k^w$ in separate columns, with the same hyperprior specifications used in each row. For interaction parameters, we report the mean of b in models with no ecological fallacy, whereas we report means of b^w and b^a separately in models assuming ecological fallacy.

In case 2.1, where only AD studies have evidence of an ecological fallacy, I-vague and I-pp give large bias (−0.977 and −0.981) and MSE (0.958 and 0.965) with a zero coverage probability. However, I-cp2 gives an unbiased estimate because the model always selects the slab prior component for η_b, i.e. the model does not borrow any information about the interaction from the AD studies. Moving to the models assuming ecological fallacy, model IPDef provides larger MSE of θ_B (0.376) than does model IPD (0.006), because the true data-generating mechanism here also assumes no ecological fallacy, though mean(b^w) and mean(b^a) are correctly estimated under IPDef. Models Ief-vague and Ief-pp provide smaller bias and MSE with better coverage probability compared with I-vague and I-pp. Ief-cp2 gives larger bias and MSE than does I-cp2.

In case 2.2, where both AD and IPD studies have evidence of an ecological fallacy, all models assuming no ecological fallacy except AD1 perform poorly. However, all models assuming ecological fallacy correctly yield unbiased $\hat{\theta}_B$ and estimate b^w and b^a. Model I-cp2 selects the spike component of the prior for η_b about half the time (460/1000), whereas model Ief-CP2 always selects the spike component. Overall, these simulation study results suggest fitting NMA inte-

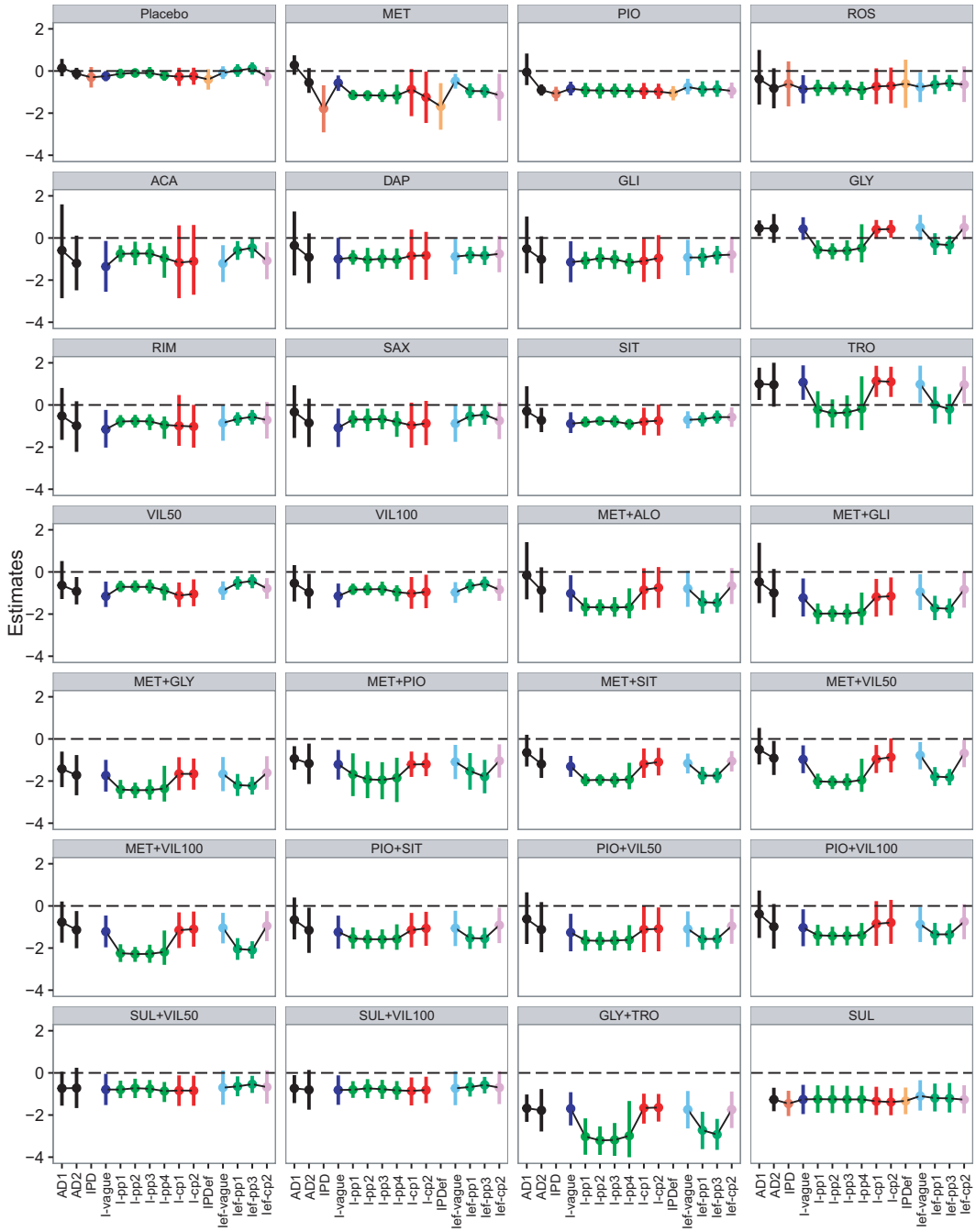

Fig. 3. Treatment effect estimates obtained from adjusted models

grative models that differentiate within- and across-trial interactions when there is any evidence of an ecological fallacy in the data.

5. Data analysis results

5.1. Model comparison

Table 2 shows the deviance information criterion (DIC), which is a measure that summarizes model fit ($\bar{D}$) and complexity (p_D), for each model listed, where smaller DIC values indicate preferred models (Spiegelhalter *et al.*, 2002). The DIC can only be compared among models using the same likelihood for the observed data (i.e. y_{ik}^{AD} and y_{ijk}^{IPD} in our models). As such, for example, comparing the DIC of model AD1 and IPD or comparing DIC between power prior models is not meaningful. Among unadjusted models, we can compare the DIC of the two models using commensurate priors, but they produce the same DIC numerically for our diabetes data.

However, we can compare DIC values across the covariate-unadjusted and adjusted models. For example, the DIC of the unadjusted IPD model is 11 263.5 and the DIC values for the adjusted IPD model and IPDef are 10 034.2 and 10 338.5 respectively, i.e. including covariate-by-treatment interactions in our IPD models is preferred. Model IPDef has a larger DIC than the adjusted IPD model. This could be because IPDef estimates both within- and across-trial interactions, or because IPDef assumes a common fixed interaction, rather than assigning a shrinkage prior to treatment-specific interactions as in the adjusted IPD model. All integrative adjusted models have smaller DIC than the corresponding unadjusted models.

As such, we show diabetes data analysis results from covariate-adjusted models in this section and results from the covariate-unadjusted models in the on-line supplementary materials.

5.2. Covariate-adjusted models

In our adjusted models, we centre the baseline age and HbA1c at 57 years and 9%, their average values respectively, to obtain meaningful treatment effect estimates. We interpret $\hat{\theta}_k$ as the expected HbA1c-level change for those whose age is 57 years and baseline HbA1c is 9%. Since the purpose of our treatments is to reduce HbA1c-level, we expect $\hat{\theta}_k$ to be negative. The interaction parameters ($\hat{b}_{\mathrm{age}}$ and $\hat{b}_{\mathrm{A1c}}$) are interpreted as the expected additional HbA1c-level change per unit increase in baseline age and HbA1c. Our diabetes data show modest evidence of an ecological fallacy when comparing AD and IPD studies (see Fig. A3 in the on-line supplementary materials)

Fig. 3 displays treatment effect estimates and associated 95% credible intervals that were obtained from the 15 covariate-adjusted models. We use colours to distinguish groups of models: black is for AD models, orange is for IPD models, blue is for the basic integrative model, green is for integrative models with power priors and red is for integrative models with commensurate priors. We use lighter colours for models assuming ecological fallacy: yellow is for IPDef, light blue is for Ief-vague, light green is for Ief-pp1 and Ief-pp3, and pink is for Ief-cp2. In what follows we refer to each panel of Fig. 3 as an element of a matrix (e.g. panel (1,1) is for placebo).

In panels (1,2) and (1,3), we see that model AD1 (the first black line) and the IPD model (the orange line) provide somewhat different effect estimates for MET and PIO. For MET and ROS, power (green) and commensurate (red) prior models give similar point estimates, but the commensurate prior models give wider 95% credible intervals. This is because commensurate prior models depend more on IPD than on AD studies, but there is only one IPD study investigating MET and ROS. The four power prior models provide similar effect estimates regardless of the

selection of power parameters across all drugs. However, these estimates are somewhat different from those under I-vague (blue) and I-cp1 and I-cp2 (red), specifically for treatments having evidence only from AD studies (see the second to seventh rows, except panel (7,4)). This happens even when we assign non-zero α_{ik} only to treatments having evidence from both AD and IPD (I-pp2, I-pp3 and I-pp4) because we assume that random effects are correlated across k. From Fig. 2(a), we can see that $\theta_5, \ldots, \theta_{27}$ are associated with only partial data from AD studies, whereas $\theta_1, \ldots, \theta_4$ are associated with both AD and IPD studies and power prior parameters $\alpha_{i1}, \ldots, \alpha_{i4}$. However, the μ_{ik}^{AD} are correlated with each other through ρ^{AD}. As a result, the power prior parameters affect estimation of $\theta_5, \ldots, \theta_{27}$ indirectly. This issue disappears when we set $\rho^{\mathrm{AD}} \equiv 0$ (i.e. the estimates for $\theta_5, \ldots, \theta_{27}$ under model I-vague are the same as those under I-pp2, I-pp3 and I-pp4). Models assuming an ecological fallacy do not yield noticeably different treatment effect estimates compared with models assuming no ecological fallacy, although we found quite significant measures of ecological fallacy in terms of baseline HbA1c (see Web Table A1).

Fig. 4 displays rankings of the 28 treatments across the 15 adjusted models based on SUCRA. Note that the ranks in models IPD and IPDef are not comparable with the other models because there are only five treatments to be compared in the IPD-only models. Across all models except the IPD-only models and I-vague, GLY+TRO is the best treatment and MET+GLY is the runner-up. Note that these two treatments tie in rankings under model I-cp2. MET is the best treatment in both IPD-only models, but it is one of the least-favoured treatments in model AD1, AD2, I-vague and Ief-vague and emerges as the fifth and sixth best treatment under models I-cp2 and Ief-cp2 respectively. Overall, the rankings vary when adapting different priors, except for a

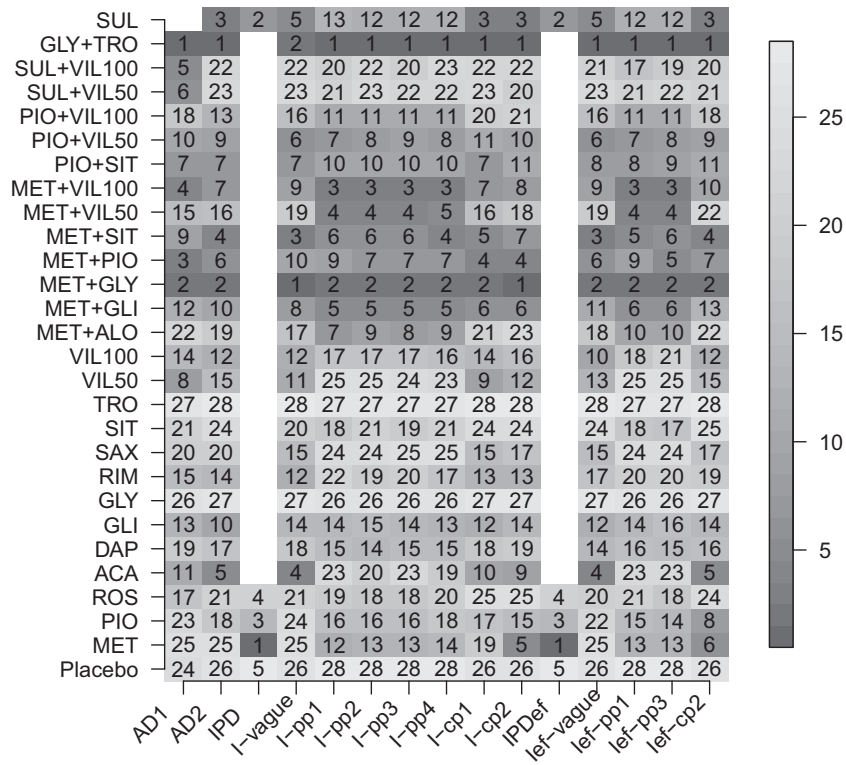

Fig. 4. Rankings of 28 treatments based on SUCRA under 15 covariate-adjusted models

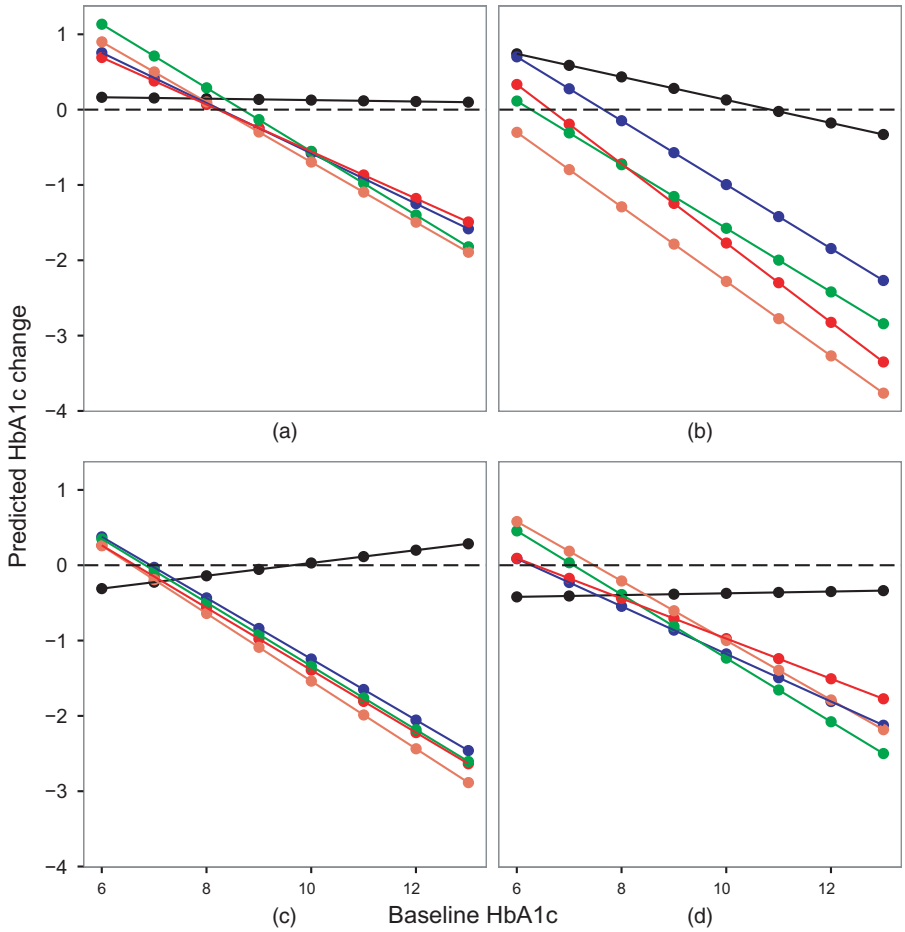

Fig. 5. Predicted HbA1c-changes under five covariate-adjusted models (●, AD1; ●, IPD; ●, I-vague; ●, I-pp1; ●, I-cp2) for those whose age is 57 years at various levels of baseline HbA1c for the four selected drugs (a) placebo, (b) metformin, (c) pioglitazone and (d) rosiglitazone having evidence from both AD and IPD

few treatments such as GLY+TRO and MET+GLY. The rankings tend to be more consistent in the covariate-unadjusted models than in the adjusted models (see Web Fig. A2).

Finally, Fig. 5 shows predicted HbA1c-change outcomes under five adjusted models (AD1, IPD, I-vague, I-pp1 and I-cp2) for a person aged 57 years at various levels of baseline HbA1c. Here, we show only four selected drugs (placebo, metformin, pioglitazone and rosiglitazone) having evidence from both AD and IPD. The predicted outcomes under AD1 are fairly different from those under models IPD and I-vague. Indeed, for PIO and ROS, model AD1 yields the *opposite* direction of predicted outcomes from the other models as baseline HbA1c increases. This shows that estimating the covariate-by-treatment interaction effect by using only AD can lead to incorrect inferences regarding effect modification.

6. Discussion

In this paper, we have adapted power and commensurate prior models to NMA integrating IPD and AD. Our simulation studies show that commensurate priors perform better when AD

studies have biased results, whereas power priors give better performance when IPD studies have biased results. In addition, covariate-by-treatment interactions should be carefully modelled when there is evidence of an ecological fallacy. We applied various models under different parameter specifications to our diabetes data and compared results in terms of effect estimates, rankings and predicted outcomes.

In NMA, effect heterogeneity across studies and inconsistency of evidence (often defined as a discrepancy between direct and indirect comparisons) should be considered carefully. First, all models that were fitted in this paper assume effect heterogeneity across studies by adding random-effect terms. In terms of the estimation of effect heterogeneity, using a power prior could affect the estimation of the effect heterogeneity across AD studies (τ^2 in Fig. 2), but not across IPD studies (ψ^2 in Fig. 2). This is because power parameters control for the AD likelihood function and this could directly impact the estimation of τ^2. In contrast, using a commensurate prior does not greatly influence the estimation of either τ^2 or ψ^2. Since the commensurate prior is assigned to the group mean level (θ_1^{IPD}, θ_2^{IPD}, θ_3^{IPD} and θ_4^{IPD} in Fig. 2(b)), it does not affect study level estimation. Our data analysis results (covariate-unadjusted models) show that the posterior means of ψ^2 across all models are quite consistent, whereas the posterior means of τ^2 when using power priors are slightly different from those under other models. Second, our arm-based models implicitly assumed consistency. Recently, Zhao *et al.* (2016) have proposed methods to detect inconsistency in arm-based NMA models with only AD studies by computing *discrepancy factors* and then identifying trials that are the sources of inconsistency by investigating extreme study-specific random-effect estimates. Later, Zhao *et al.* (2017) introduced diagnostic methods to detect influential observations that might cause inconsistency and studied the effect of these influential observations on conclusions drawn from an NMA. However, further methodological development is required to apply these approaches to our analysis because it combines IPD and AD by assigning different prior weights to different sources of evidence.

We presume that the IPD are the primary data and the AD are the auxiliary data in this paper. It might seem unfair to downweight a large and well-conducted controlled AD study just because the data have been aggregated. However, when we consider a covariate-adjusted model for investigating interaction effects between treatment and individual level covariates, the AD study may not deliver accurate interaction information to potential ecological bias. Although this setting seems a little restricted, our approaches can be adapted flexibly for useful extensions to current evidence synthesis methodology. First, when both AD and IPD include low quality trials, we can define all low quality trials, regardless of their sources, as auxiliary data, and the remaining trials as primary data. Then, we downweight the likelihood by adding a power parameter α_{ik} to both $L(\boldsymbol{\Theta} \mid y_{ijk}^{\mathrm{AD}})$ and $L(\boldsymbol{\Theta} \mid y_{ijk}^{\mathrm{IPD}})$ in expression (9) for the auxiliary trials (both AD and IPD). Similarly, we can modify models by using commensurate priors. Second, our approaches would give a nice practical extension to the situation where evidence from RCTs is pooled with non-randomized evidence. In this case, we might want to downweight the non-randomized evidence to reflect biases that are associated with such study designs in the overall pooled results. We can easily apply our methods by setting RCTs (regardless of IPD or AD) as the primary data and non-randomized studies as the auxiliary data. Third, our models can be generalized to binary or count outcome cases. For example, we can replace the normal likelihood functions equations (3) and (4) with binomial or Poisson distributions for binary or count outcomes respectively and replace the identity link function in equations (5)–(7) with logit or log-links for binary or count outcomes respectively.

Applying our approaches in practice requires many subjective decisions about how to define auxiliary data and what model should be used under which conditions. First, we could use study design information (e.g. random-sequence generation, allocation concealment and blinding of

participants) to measure risk of bias for each study (Higgins *et al.*, 2011). For example, open label RCTs are believed to have a higher risk of bias than double-blind RCTs. As such, we can have some sense of the risk of bias for each trial before we pool studies by using our proposed methods in practice. Determining how much we borrow from studies having high or low risk of bias (i.e. how to decide the value of power parameter α_{ik}) after assessing the risk of bias is not an easy decision, because risk of bias is a somewhat qualitative measure rather than quantitative. Given that there is no clear guideline, we recommend fitting several models under various assumptions (e.g. using different α_{ik}-values) and then comparing the results. Further research is required to inform such a guideline. Finally, choosing between the two types of prior might not be a straightforward decision, especially when both AD and IPD include a low risk of bias studies. There is no single, foolproof metric measuring the different degrees of risk of bias in AD and IPD studies, and the performance of models using a power or commensurate prior would highly depend on this difference. Therefore, we recommend fitting models by using both priors and then comparing the results. For general advice in this practical setting, further investigation is needed.

Our diabetes data analyses have several limitations. First, our AD study collection is likely to exclude many relevant studies because it was assembled for methods development purposes, not for a systematic review. Therefore, our Section 5 results should be interpreted cautiously. Another limitation is that we considered only treatment efficacy here, and not treatment safety. Although some treatments have better efficacy, they may cause additional safety profiles, such as hypoglycaemic events. In addition, many treatments' evidence arose entirely from a single study, so it would be wise to collect more studies to obtain reliable data on the comparative effectiveness of these treatments. Next, all IPD studies are only from a single trial sponsor, and this could possibly lead to selection bias. IPD studies from multiple agents can be combined, and we can extend our approaches to borrow information adaptively across IPD studies from different agents. Finally, we assumed simple linear covariate-by-treatment interactions in our adjusted models. However, non-linear interactions could be implemented when the data suggest a non-linear trend.

In future work, we shall conduct more simulation studies to answer many unanswered practical questions, such as how to measure different degrees of bias in AD and IPD studies, implications of the use of power and commensurate priors, how to decide the power parameters in practice and how many AD and IPD studies are required to obtain reliable estimates. We hope to develop methods to detect inconsistency when combining AD and IPD in NMA. In addition, we hope to extend our proposed methods to multiple outcomes (e.g. efficacy and safety) and provide decision-making tools that consider the outcomes simultaneously. Finally, we hope to apply our methods to an even larger network including more IPD studies from several agents, so that we can estimate more accurate, reliable and clinically meaningful treatment effects and make better decisions.

Acknowledgements

We appreciate Eli Lilly and Company for providing relevant data sets for this research. The work of the third author was supported in part by National Cancer Institute grant 1-R01-CA157458-01A1.

References

American Diabetes Association (2015) Approaches to glycemic treatment. *Diab. Care*, **38**, suppl. 1, 41–48.

Cooper, N. J., Sutton, A. J., Morris, D., Ades, A. E. and Welton, N. J. (2009) Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statist. Med.*, **28**, 1861–1881.

Debray, T. P. A., Koffijberg, H., Vergouwe, Y., Moons, K. G. M. and Steyerberg, E. W. (2012) Aggregating published prediction models with individual participant data: a comparison of different approaches. *Statist. Med.*, **31**, 2697–2712.

Debray, T. P. A, Moons, K. G. M., Ahmed, I., Koffijberg, H. and Riley, R. D. (2013) A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statist. Med.*, **32**, 3158–3180.

Donegan, S., Williamson, P., D'Alessandro, U., Garner, P. and Smith, C. T. (2013) Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: individual patient data may be beneficial if only for a subset of trials. *Statist. Med.*, **32**, 914–930.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Goldstein, H., Yang, M., Omar, R., Turner, R. and Thompson, S. (2000) Meta-analysis using multilevel models with an application to the study of class size effects. *Appl. Statist.*, **49**, 399–412.

Higgins, J. P. T., Altman, D. G. and Sterne, J. A. C. (2011) Assessing risk of bias in included studies. In *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011)* (eds J. P. T. Higgins and S. Green), ch. 8. London: Cochrane Collaboration. (Available from `www.handbook.cochrane.org`.)

Hobbs, B. P., Carlin, B. P., Mandrekar, S. and Sargent, D. J. (2011) Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, **67**, 1047–1056.

Hobbs, B. P., Carlin, B. P. and Sargent, D. J. (2013) Adaptive adjustment of the randomization ratio using historical control data. *Clin. Trials*, **10**, 430–440.

Hobbs, B. P., Sargent, D. J. and Carlin, B. P. (2012) Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Baysn Anal.*, **7**, 639–674.

Hong, H., Carlin, B. P., Shamliyan, T. A., Wyman, J. F., Ramakrishnan, R., Sainfort, F. and Kane, R. L. (2013) Comparing Bayesian and frequentist approaches for multiple outcome mixed treatment comparisons. *Med. Decsn Makng*, **33**, 702–714.

Hong, H., Chu, H., Zhang, J. and Carlin, B. P. (2016) A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons (with discussion). *Res. Synth. Meth.*, **7**, 6–22.

Hong, H., Fu, H., Price, K. L. and Carlin, B. P. (2015) Incorporation of individual patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. *Statist. Med.*, **34**, 2794–2819.

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

Jackson, C., Best, N. and Richardson, S. (2006) Improving ecological inference using individual-level data. *Statist. Med.*, **25**, 2136–2159.

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

Lu, G. and Ades, A. E. (2006) Assessing evidence inconsistency in mixed treatment comparisons. *J. Am. Statist. Ass.*, **101**, 447–459.

Lumley, T. (2002) Network meta-analysis for indirect treatment comparisons. *Statist. Med.*, **21**, 2313–2324.

Murray, T. A., Hobbs, B. P. and Carlin, B. P. (2015) Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. *Ann. Appl. Statist.*, **9**, 1549–1570.

Murray, T. A., Hobbs, B. P., Lystig, T. C. and Carlin, B. P. (2014) Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. *Biometrics*, **70**, 185–191.

Murray, T. A., Hobbs, B. P., Sargent, D. J. and Carlin, B. P. (2016) Flexible Bayesian survival modeling with nonparametric time-dependent and shape-restricted covariate effects. *Baysn Anal.*, **11**, 381–402.

Neelon, B. and O'Malley, A. J. (2010) Bayesian analysis using power priors with application to pediatric quality of care. *J. Biometr. Biostatist.*, **1**, article 103.

Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd Int. Wrkshp Distributed Statistical Computing, Vienna, Mar. 20th–22nd, 2002* (eds K. Hornik, F. Leisch and A. Zeibeis), vol. 2.

Riley, R. D., Lambert, P. C., Staessen, J. A., Wang, J., Gueyffier, F., Thijs, L. and Boutitie, F. (2008) Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statist. Med.*, **27**, 1870–1893.

Salanti, G., Ades, A. E. and Ioannidis, J. P. (2011) Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J. Clin. Epidem.*, **64**, 163–171.

Saramago, P., Sutton, A. J., Cooper, N. J. and Manca, A. (2012) Mixed treatment comparisons using aggregate and individual participant level data. *Statist. Med.*, **31**, 3516–3536.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.

Su, Y. and Yajima, M. (2015) R2jags: using R to run 'JAGS'. *R Package Version 0.5-7*. (Available from `https://CRAN.R-project.org/package=R2jags`.)

Zhang, J., Chu, H., Hong, H., Neaton, J. D., Virnig, B. A. and Carlin, B. P. (2015) Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Statist. Meth. Med. Res.*, to be published, doi 10.1177/0962280215596185.

Zhao, H., Hodges, J. S. and Carlin, B. P. (2017) Diagnostics for generalized linear hierarchical models in network metaanalysis. *Res. Synth. Meth.*, **8**, 333–342.
Zhao, H., Hodges, J. S., Ma, H., Jiang, Q. and Carlin, B. P. (2016) Hierarchical Bayesian approaches for detecting inconsistency in network meta-analysis. *Statist. Med.*, **35**, 3524–3536.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supplementary materials for "Power and commensurate priors for synthesizing aggregate and individual patient-level data in network meta-analysis"'.