Hierarchical Bayesian Approaches for Detecting Inconsistency in Network Meta-Analysis

Hong Zhao [*] James S. Hodges [*] Haijun Ma [†] Qi Jiang[†] Bradley P. Carlin [*]

February 19, 2015

Abstract

Network meta-analysis (NMA), also known as multiple treatment comparisons, is commonly used to incorporate and compare direct and indirect evidence comparing treatments. With recent advances in methods and software, Bayesian approaches to NMA have become quite popular and allow models of previously unanticipated complexity. However, when direct and indirect evidence conflict in a NMA, the model is said to suffer from *inconsistency*. Current inconsistency detection in NMA is usually based on contrast-based (CB) models; however, this approach has certain limitations. In this work, we propose an arm-based (AB) random effects model, where we detect discrepancy of direct and indirect evidence for comparing two treatments using the fixed effects in the model, while flagging extreme trials using the random effects. Our approaches permit users to address issues previously tackled via CB models. We compare sources of inconsistency identified by our approach and existing loop-based CB methods using real and simulated datasets, and demonstrate that our methods can offer more powerful inconsistency detection.

Keywords: Multiple treatment comparisons, Bayesian analysis, Inconsistency

---

[*] Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, MN 55455 USA. Address correspondence to Dr. Bradley P. Carlin, Division of Biostatistics, University of Minnesota School of Public Health, Mayo Mail Code 303, Minneapolis, MN 55455, USA. E-mail: brad@biostat.umn.edu

[†] Amgen Inc., Thousand Oaks, CA, 91320, USA.

1 Introduction

In comparative effectiveness research, *network meta-analysis* (NMA), also known as *mixed treatment comparisons* [1, 2], is an extension of the pairwise meta-analysis method [3] to compare the results from two or more studies that have at least one treatment in common. This enables both direct and indirect comparisons, and addresses the comparative effectiveness or safety of the interventions based on all sources of data. Because of its enormous potential, interest in this methodology has grown substantially, and the application of NMA is increasingly common [4]. Moreover, drug regulators and other national health agencies have increasingly adopted such methods [5, 6].

One of the key assumptions for NMA is *consistency* [1, 7–9]. As shown in the NMA network graph in Figure 1, each vertex represents a treatment and each edge represents a pairwise comparison. In Figure 1(a), the comparison of P vs. A and P vs. B is direct evidence, and there is no head-to-head comparison between A and B. To make inference about A vs. B, we can only use the indirect information from studies of P vs. A and P vs. B. In Figure 1(b), where we instead have direct evidence between A and B, we can now compare A and B using both direct and indirect information. However, when direct and indirect evidence conflict in a NMA, the model is said to suffer from *inconsistency.* Using Figure 1(b) as an example, consistency holds if

$$d_{AB} = d_{PB} - d_{PA}, \tag{1}$$

where d_{AB} is the treatment effect (e.g., log odds ratio) of B vs. A. Inconsistency can arise due to many causes, including non-comparability of trials, different control groups, or differences in patient characteristics. If inconsistency is present, information from different sources may disagree, and the treatment effects estimates obtained from the NMA may be biased or hard to interpret. Many remedies have been suggested, including adding fixed covariate effects to

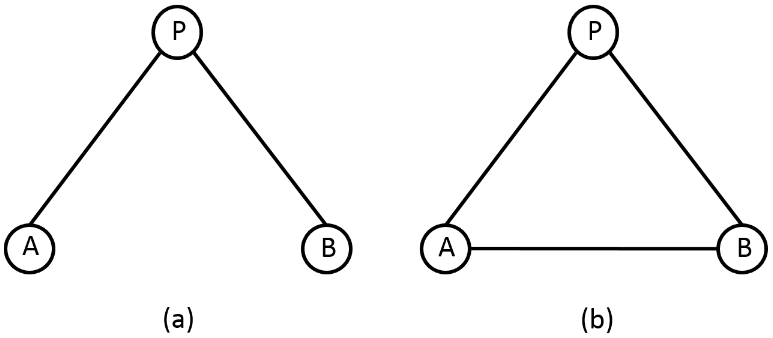


Figure 1: Networks between treatments. Each vertex represents a treatment and each edge represents a pairwise comparison. (a) Indirect comparison between A and B. (b) Both direct and indirect comparison between A and B.

account for different patient characteristics. As a first step, it is important to detect inconsistency and its likely sources. But as the number of studies increases, network complexity often precludes a quick diagnosis of inconsistency and identification of which studies or treatments are these sources.

In this paper, we will review inconsistency detection methods using various NMA models and propose another. Bayesian hierarchical models for NMA with binary outcomes have been well studied. The logit model initially proposed by Lu and Ades [2] is a *contrast-based* (CB) model, which uses the log odds ratio to estimate relative effects of two treatments. Furthermore, a fixed effects model is used when we assume treatment effects do not vary between studies, while a random effects model is implemented when heterogeneity between trials is allowed.

Inconsistency detection is thoroughly discussed by Lu and Ades [7] for CB models using two illustrative datasets [10, 11]. They proposed examining *loop-based inconsistency* by adding one parameter called an *inconsistency factor* (ICF), w, to the consistency relationship as shown in (1): $d_{AB} = d_{PB} - d_{PA} + w_{ABP}$, where three treatments are connected in a cycle, like the loop ABP in Figure 1(b). The posterior distribution of w reflects the extent of inconsistency in a particular evidence loop. Most recently, the back-calculation and node-splitting methods proposed by Dias et al. [8] have streamlined the process of inconsistency detection by looking in more detail at specific comparisons. The first method is useful when only pooled summaries of the pairwise contrasts are available, while the second method is more general and can be applied to networks with trial-level data. Both approaches extend inconsistency detection to any network, not merely triangular node structures. Higgins et al. [9] conducted inconsistency detection in NMA using multivariate meta-regression, extending the full design-by-treatment interaction model proposed by Lumley [1] to include multi-arm trials. Higgins et al. compared their method to Lu and Ades' method, and concluded that the inconsistency model proposed by Lu and Ades is a restricted version of their full design-by-treatment interaction model.

All the inconsistency detection methods mentioned above are based on CB models. However, these methods suffer from certain limitations due to their focus on relative effects [12, 13]. First, they require one arm in each study to be designated as the "baseline". Since many NMAs do not have a common "control" arm (such as a placebo or "standard of care"), any parameters for such "baseline" groups are generally hard to interpret. Moreover, for binary outcomes, CB models for NMA typically analyze only a summary statistic, often the odds ratio (OR). Because they treat the underlying "baseline" risks as nuisance parameters, CB approaches fail to estimate the treatment-specific response proportions [13]. Finally, the CB model restricts the variance of a baseline effect to always be smaller than that of the other treatments [12]. For these reasons, several authors have suggested an alternative method for NMA often called the *arm-based* (AB) model, which models the absolute (rather than relative) effect of each treatment. The AB model requires an assumption of exchangeability of treatment arms across studies, while the CB model assumes exchangeability of only the *relative* treatment effects compared to baseline ("treatment contrasts"), measured on an appropriate scale (e.g., OR) across studies. However, even this standard CB assumption is difficult to justify, especially for binary outcomes where the standard logit scale used is arbitrary and chosen primarily for statistical convenience.

In this paper, we aim to develop an inconsistency detection method using the arm-based random effects (ABRE) model, which is distinct from loop-based and other inconsistency detection methods for CB models. Lu and Ades presented the first key CB inconsistency analysis, and while later researchers have produced arguably better CB methods [8, 9], they all suffer from complications arising from the CB formulation. Therefore, like earlier authors we review and compare to Lu and Ades's method, since it is relatively simple and the comparison can be made for both loop- or non-loop-based methods. Under our AB model framework, we look for possible inconsistency in two ways: (1) by using estimates of the fixed effects in the ABRE models to test the discrepancy of the direct and indirect evidence for comparing two treatments, either loop-based or not; and (2) by using estimates of specific ABRE random effects to detect inconsistency at certain trial-by-arm combinations, once inconsistency has been detected through the ABRE model fixed effects .

The rest of our paper is organized as follows. In Section 2, two motivating datasets used in Lu and Ades [7] are described; they will be used throughout our analysis. Details of current models for NMA will be illustrated and methods to detect inconsistency investigated in Section 3, including our proposed methods using AB models. This section also applies our methods to one of the example datasets, and compares our results to those of Lu and Ades [7]. Section 4 then evaluates our methods via simulation. Finally, in Section 5 we summarize and offer directions for future research.

2 Motivating Examples

2.1 Thrombolytic Drugs Dataset

Our first illustrative dataset is described in a systematic review [10], and compares eight thrombolytic drugs for use after acute myocardial infarction with the primary outcome being 30-35 day mortality. Twenty-eight trials were conducted to study eight drugs: reteplase (Ret),

Table 1: Thrombolytic drugs dataset (total number of events/ total number of subjects).

Study number	SK(1)	AtPA(2)	t –PA(3)	SK +tPA(4)	Ten(5)	Ret(6)	UK(7)	ASPAC(8)
1	1472/20163	652/10344		723/10328				
2	1455/13780		1418/13746					1448/13773
3	9/130		6/123					
4	5/63		2/59					
5	3/65		3/64					
6	887/10396		929/10372					
7	7/85		4/86					
8	12/147		7/143					
9	10/135		5/135					
10	4/107			6/109				
11	285/2992					270/2994		
12	3/58							2/52
13	3/86							6/89
14	3/58							2/58
15	13/182							11/188
16	10/203						7/198	
17		522/8488			523/8461			
18		356/4921				757/10138		
19		13/155				7/169		
20		2/26					7/54	
21		12/268					16/350	
22		5/210						17/211
23		3/138						13/147
24			8/132				4/66	
25			10/164				6/166	
26			6/124				5/121	
27			13/164					10/161
28			7/93					5/90

streptokinase (SK), urokinase (UK), alteplase (tPA), anistreptilase (ASPAC), accelerated alteplase (AtPA), tenecteplase (Ten), and streptokinase plus alteplase (SK + tPA). The dataset is shown in Table 1, which displays total number of events over total number of subjects for the treatment groups in each trial. The evidence network is plotted in Figure 2(a), with each vertex representing a treatment and each edge representing a pair of treatments for which at least one direct comparison exists. The indices of the trials having each pairwise comparison are shown in square brackets to the left of or above the corresponding edge.

2.2 Smoking Cessation Dataset

Our second illustrative NMA dataset compares smoking cessation strategies reported by the Agency for Health Care Policy and Research (AHCPR) Smoking Cessation Guideline Panel [11], which has been analyzed by Lu and Ades [7] among others. It consists of 24 studies to compare the relative effect of three treatments (B: self-help; C: individual counseling; and D:

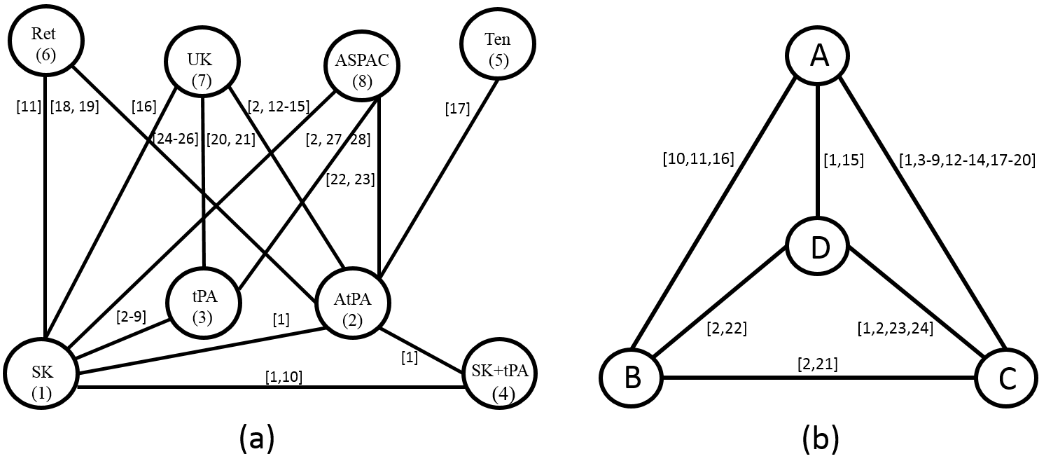


Figure 2: (a) Network for thrombolytic drugs dataset. Each vertex represents a treatment and each edge represents a pair of treatments for which at least one direct comparison exists. The indices of the trials having each comparison are shown in square brackets to the left of or above the corresponding edge. (b) Network for smoking cessation dataset. Each vertex represents a treatment and the indices of trials having each comparison are shown in square brackets on the corresponding edge.

group counseling) vs. the baseline treatment (A: no contact). While we do not show these data explicitly, Figure 2(b) gives the evidence network showing the connection between all treatment groups, where again the indices of trials having each comparison are shown in square brackets on the corresponding edge.

3 Current Models for Inconsistency Detection in NMA

In this section, we will review various models for NMA, and investigate the associated methods to detect inconsistency in the network. We assume the outcome Y_{ik} for each study follows a binomial distribution as in our sample datasets,

$$Y_{ik} \overset{ind}{\sim} Bin(n_{ik}, p_{ik}),\ i = 1, \dots, I,\ k = 1, \dots, K, \tag{2}$$

where Y_{ik} is the total number of subjects with events, n_{ik} is the total number of subjects, and p_{ik} is the probability of the outcome in the k^{th} treatment arm from the i^{th} study.

3.1 Contrast-Based Random Effects (CBRE) Models

The NMA logistic model proposed by Lu and Ades [7] is introduced for a single outcome. Using a contrast-based model, we wish to estimate relative effects of treatment pairs. One can use random effects to capture heterogeneity between studies, namely

$$logit(p_{ik}) = \alpha_{iB} + \delta_{iBk}, \tag{3}$$

where B refers to the baseline treatment. Unless there is a common treatment to all studies, the baseline treatments in the studies will be different. Here, α_{iB} is the log odds of the response for the baseline treatment in study i, and δ_{iBk} is the log odds ratio of treatment k versus baseline for the i^{th} study. An independent normal specification for the random effects δ_{iBk} is assumed:

$$\delta_{iBk} \overset{ind}{\sim} N(d_k - d_B, \sigma^2), \tag{4}$$

where δ_{iBk} follows a normal distribution with mean equal to the contrast $d_k - d_B$, and δ_{iBk} varies across i, capturing the variability in the log odds of contrast for different studies. This assumes exchangeability of these differences in d_k on one specific scale used for analyzing binary data, the log-odds scale. From the posterior distribution of the d_k, the relative effects of each treatment can be calculated. In this random effects model, the same variance σ^2 is assumed for all treatment groups, so it is called a *homogeneous random effects* model. If a trial has more than two arms, we need to assume a particular variance-covariance structure for the vector $\vec{\delta_i} = (\delta_{iB2}, ..., \delta_{iBK})'$. The vector $\vec{\delta_i}$ then follows a multivariate normal distribution, with the correlation between any two treatment effects equal to 0.5 under consistency by construction [7]. If we change the variance σ^2 in (4) to σ^2_{Bk}, we obtain a *heterogeneous random effects* model, associating different variances with specific pairwise comparisons. As mentioned by Lu and Ades [14], estimating the parameters in the heterogeneity model is a difficult problem due to the implicit constraints on the variances from the NMA structure, which becomes quite complicated when the NMA includes many multi-arm trials. Therefore, we focus on the standard CBRE model with homogeneous random effects.

3.2 Arm-Based Random Effects (ABRE) Model

The term “arm-based” is used here to refer to a model proposed by Hong et al. [12] and Zhang et al. [13], although the term has been used somewhat differently elsewhere [15, 16]. In this model, instead of estimating a mean contrast in each trial, we estimate the logit response probability for each treatment for binary data:

$$logit(p_{ik}) = \mu_k + \eta_{ik}, \tag{5}$$

where μ_k is the (fixed-effect) mean outcome for treatment k, and η_{ik} is the random effect for treatment k in study i. Then the random effects $\vec{\eta_i}$ for study i can be structured as:

$$\vec{\eta_i} = (\eta_{i1}, \dots, \eta_{iK})' \sim MVN(\mathbf{0}, \Sigma), \tag{6}$$

where Σ is an $K \times K$ unstructured covariance matrix to allow correlation between treatment arms in each trial. Compared to the CB model framework, AB models are more straightforward to interpret, especially when implemented in a missing-data framework that imputes values for any missing treatment arms, thus allowing use of a common baseline across all treatments [12]. However, ABRE models do have slightly more parameters to be estimated, since they model the absolute effect of each treatment, rather than relative effects as in CBRE models. As noted, these two approaches make different assumptions of exchangeability: we assume trials are exchangeable according to the levels of treatment outcomes, while CB methods assume that trials are exchangeable according to treatment contrasts.

3.3 Priors for CBRE and ABRE Models and Model Selection

An important issue in Bayesian modeling is the choice of prior distributions for each of the model parameters. This could be a traditional informative prior, which might come from a literature review or explicitly from an earlier data analysis. But in the situation where there is no previous information about the parameters, we often choose proper, weakly informative prior densities, and let the data drive the posterior distribution [17].

In this paper, to keep our modeling somewhat generic, we use weakly informative priors for both CBRE and ABRE models. Such priors for CBRE models are also described in Lu and Ades [7]. For all our models, the fixed effects (α_{iB}, d_k, and μ_k) are assumed to follow a $N(0, 1000)$ distribution, which is very vague but proper. We also assume all CBRE ICFs (the ws) independently follow a $N(0, \sigma_w^2)$ distribution. In the homogeneous random effects model, we adopt uniform priors $\sigma \sim U(0, 2)$ and $\sigma_w \sim U(0, 2)$ for the standard deviations of the random effects and the ICFs, respectively. For the precision matrix of the random effects, we choose a Wishart prior $\Sigma_k^{-1} \sim W(V, n)$, where the degrees of freedom $n = K$, the number of treatments, and V is a $K \times K$ matrix with diagonal elements equal to 0.1 and off-diagonal elements equal to 0.005. Using the R function rWishart(), we can calculate that the above prior corresponds to a 95% prior credible set of $0.14 - 11.52$ for the standard deviation parameters and a 95% credible set of -1.00 to 1.00 for the correlation parameters, confirming it is weakly informative.

For Bayesian model selection and comparison, we use the Deviance Information Criterion (DIC). DIC is a Bayesian generalization of the Akaike information criterion (AIC), and is calculated as the sum of p_D and $\bar{D}$, where $\bar{D}$ is a measure of goodness of fit and p_D is the effective number of parameters in the model [18]. A model with smaller DIC (say, by 5-10 units) is preferred.

3.4 Software

`WinBUGS` was used to obtain MCMC samples for all our Bayesian NMA models. Standard diagnostics, including trace plots and sample autocorrelations, were implemented to check

MCMC convergence. Lu and Ades [7] provide `WinBUGS` codes for CBRE models for the above two examples. `WinBUGS` code for our ABRE models are available at www.biostat.umn.edu/ brad/-software.html.

3.5 Inconsistency Detection in CB models

In Section 1, we introduced loop-based inconsistency as proposed by Lu and Ades [7]. In this approach, the number of *inconsistency degrees of freedom* (ICDF) must be determined. ICDF is informally defined as "the number of independent 'loops' of evidence" in the network [19]. When there is no multi-arm trial in the network or when each pair of arms compared in a multi-arm trial is also compared in another trial, ICDF is calculated as $T - K + 1$, where T is the total number of direct pairwise comparisons and K is the number of treatments. Using the smoking cessation dataset (Figure 2(b)) as an example, $T = 6$ and $K = 4$, so there are $ICDF = T - K + 1 = 3$ independent "loops" of evidence for estimating inconsistency in the dataset. To model this inconsistency, (1) is modified by adding an inconsistency factor w for each loop:

$$\begin{aligned} d_{BC} &= d_{AC} - d_{AB} + w_{ABC}, \\ d_{BD} &= d_{AD} - d_{AB} + w_{ABD}, \\ \text{and } d_{CD} &= d_{AD} - d_{AC} + w_{ACD}, \end{aligned} \tag{7}$$

where the posterior distributions of the 3 added parameters measure the effect of inconsistency in the 3 respective loops. However, it is not clear how big the inconsistency factor should be for the network to qualify as "inconsistent". For example, using the thrombolytic drugs dataset, although the point estimate for one of the w factors is fairly different from 0, the 95% posterior credible interval (CI) for this w factor covers 0, suggesting no "significant" inconsistency. Also, it is not clear which three independent loops from the four that are possible (ABC, ABD, ACD and BCD) should be selected in (7). With a different parameterization (i.e., selecting a different 3 loops), the estimates of the relative effects using CB models would be different.

A further complication is that when some treatment arms are involved only in multi-arm trials, the above ICDF formula needs to be reduced by S, which is the number of inconsistency loops where direct comparisons of pairs of arms are only present in multi-arm trials, not in another trial [7]. As shown in Table 1, there are $T = 13$ direct comparisons in the thrombolytic drugs dataset: 1-2, 1-3, 1-4, 1-6, 1-7, 1-8, 2-4, 2-5, 2-6, 2-7, 2-8, 3-7, and 3-8. Since the comparisons of 1-2 and 2-4 are only estimated in the multi-arm Trial 1, the inconsistency relation for loop 124 cannot be estimated; there is no indirect evidence regarding this loop. Therefore, in this dataset $S = 1$ and $ICDF = T - K + 1 - S = 13 - 8 + 1 - 1 = 5$. A valid set of inconsistency equations modifying model (1) is thus

$$\begin{aligned} d_{62} &= d_{61} - d_{21} + w_{126}, \\ d_{72} &= d_{71} - d_{21} + w_{127}, \\ d_{82} &= d_{81} - d_{21} + w_{128}, \\ d_{73} &= d_{71} - d_{31} + w_{137}, \\ \text{and } d_{83} &= d_{81} - d_{31} + w_{138}. \end{aligned} \tag{8}$$

Here, the parameterization happens to be unique since there are only five independent loops. However, this will not happen in general. Also, when many multi-arm trials are presented in a NMA, it may become difficult to calculate S, whence there will be no general formula to determine ICDF [7].

3.6 Inconsistency Detection in AB Models

Since AB models implicitly assume consistency, which we will investigate in detail later (Section 3.7.1), instead of using loop-based w's we propose using the ABRE fixed and random effects to detect inconsistency. For the thrombolytic drugs dataset, Lu and Ades [7] argue that the posterior distribution of w_{128} indicates evidence for a potential inconsistency problem in loop 128 associated with Trials 22 and 23. Therefore, we use this dataset to show how one might detect inconsistency in AB models. Analysis of inconsistency detection for the smoking cessation dataset using AB models has also been performed. However, like Lu and Ades [7], our results also suggest absence of serious inconsistency in this dataset, so we omit these results.

3.6.1 Inconsistency Detection in AB Models using Fixed Effects

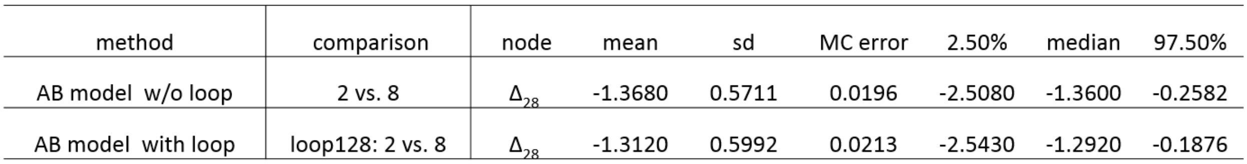
To directly measure the discrepancy of direct and indirect evidence for comparing two treatments (say, A vs. B) in an arm-based model, we can divide the trials into 4 groups: (*i*) trials that include both A and B, (*ii*) trials that include A but not B, (*iii*) trials that include B but not A, and (*iv*) trials that include neither A nor B. Following (5), the fixed effects estimating the log odds of an event under treatments A and B in the first group can be denoted as $\mu_A^{(i)}$ and $\mu_B^{(i)}$, respectively. Similarly, the fixed effects estimating the log odds of an event under treatment A in the second group and treatment B in the third group can be denoted as $\mu_A^{(ii)}$ and $\mu_B^{(iii)}$. The discrepancy between A and B using arm-based models can then be tested by computing the posterior distribution of the *discrepancy factor*

$$\Delta_{AB} = (\mu_A^{(i)} - \mu_B^{(i)}) - (\mu_A^{(ii)} - \mu_B^{(iii)}), \tag{9}$$

which is the difference in treatment effects in trials including both arms (the direct evidence) minus the difference in trials including just one arm (the indirect evidence). If zero is in the far tail of this posterior distribution, we conclude that the two sources of evidence for comparing A and B are discrepant, and thus inconsistency exists. This method allows us to check the inconsistency of the direct and indirect evidence from all sources associated with these two treatments. This is because group (*ii*) above includes trials with Arm A plus other arms that are never paired with B in any trial, and group (*iii*) includes trials with Arm B plus other arms that are never paired with A in any trial.

We have applied the above method to calculate the discrepancy factors for 10 different comparisons in the thrombolytic drugs dataset: 1-2, 1-3, 1-6, 1-7, 1-8, 2-6, 2-7, 2-8, 3-7, and 3-8. Comparison of 1 vs. 4 was not done since it is only present in the multi-arm Trial 1. The first row of Table 2 (“AB model w/o loop”) summarizes the result for the discrepancy factor which is significantly different from 0 in terms of its 95% posterior CI. In this approach,

Table 2: Discrepancy factors for thrombolytic drugs dataset using ABRE model. "AB model w/o loop" refers to the method described in Section 3.6.1 and "AB model with loop" refers to the method described in Section 3.7.2.



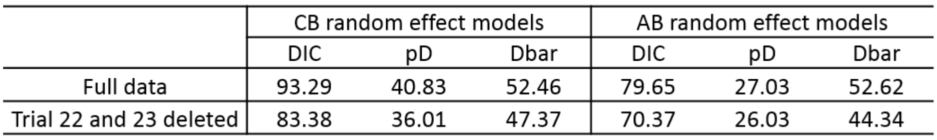
method	comparison	node	mean	sd	MC error	2.50%	median	97.50%
AB model w/o loop	2 vs. 8	Δ_{28}	-1.3680	0.5711	0.0196	-2.5080	-1.3600	-0.2582
AB model with loop	loop128: 2 vs. 8	Δ_{28}	-1.3120	0.5992	0.0213	-2.5430	-1.2920	-0.1876

without defining loops, we have successfully detected the discrepancy of sources of evidence for comparing 2 vs. 8, which agrees with the conclusion by Lu and Ades. These authors found a posterior estimate for w_{128} of 0.678 with 95% posterior CI $(-0.03, 1.72)$, indicating the presence of inconsistency in loop 128. A similar result was found by White et al. [20], also suggesting inconsistency around loop 128 from certain *designs* (to refer only to the set of treatments compared in a trial) using a frequentist method. The "AB model with loop" row will be discussed in Section 3.7.2.

3.6.2 Inconsistency Detection in AB Models by Random Effects

One disadvantage of detecting inconsistency in AB models through fixed effects (Section 3.6.1) is that although it detects inconsistent comparisons, it does not identify the source of the inconsistency as arising from certain trial and treatment combinations. Without defining specific w factors in our AB models, effects of inconsistency, if any, must manifest in either the fixed or the random effects in the ABRE model. Therefore, after detecting inconsistency with fixed effects, we can use the random effects η_{ik} in our ABRE model to investigate the most extreme η_{ik}, say, around the top 5% in absolute value. Using the thrombolytic drugs dataset, the posterior means of these random effects are as follows: the random effects for Treatments 3, 8 and 1 in Trial 2 have the estimated values 0.51, 0.45 and 0.40, respectively, followed by Treatment 3 in Trial 6 and Treatment 2 in Trial 22, where Trials 2 and 6 have large sample sizes. The remaining random effects deemed to be large are for Trial 11, Treatment 7 in Trial

Table 3: Model comparisons with DIC for thrombolytic drugs dataset



	CB random effect models			AB random effect models		
	DIC	pD	Dbar	DIC	pD	Dbar
Full data	93.29	40.83	52.46	79.65	27.03	52.62
Trial 22 and 23 deleted	83.38	36.01	47.37	70.37	26.03	44.34

20, Treatment 2 in Trials 23 and 18, and Treatment 6 in Trial 19. As in Lu and Ades [7], we also identify Trials 22 and 23 for Treatment 2 (AtPA) as extreme cases. These two observations consider the comparison of 2 vs. 8, which was first identified as discrepant using our method in Section 3.6.1. Therefore, Treatment 2 in Trials 22 and 23 are identified as the sources of inconsistency in the dataset.

We then fit the dataset without Trials 22 and 23 using the same ABRE model. The results for CBRE and ABRE models are compared using DIC, for the datasets with and without Trials 22 and 23. As shown in Table 3, the ABRE models have smaller DIC compared to CBRE models in general, indicating that the model-based imputation of the unobserved arms yields better DIC performance than ignoring such information for these data. Moreover, the ABRE and CBRE models have similar $\bar{D}$, indicating similar goodness of fit. Therefore, the reduction in DIC for the ABRE model is mainly due to the reduction in p_D, i.e., the ABRE has a smaller effective number of parameters.

3.7 Relation between Inconsistency Detection in the CB and AB Models

In this section, we will show how our methods for detecting inconsistency using AB models are related to the method using CB models. We cannot use inconsistency factors to study loops in our AB models since consistency is assumed implicitly. However, we can investigate inconsistency in a loop-based manner within the AB approach by defining discrepancy factors using a different subsetting method for groups.

3.7.1 No Inconsistency Factors in AB Models

In our ABRE model framework, the relative effect comparing two treatments (for example, d_{XY}) is calculated as the difference in the fixed mean effects of treatments X and Y, which has a posterior distribution. Thus, we cannot use (7) to check for inconsistency because the inconsistency factors w are not identified in the ABRE model. To illustrate this point using the thrombolytic drugs dataset described in Section 2.1, we applied the ABRE model in Section 3.2 with inconsistency factors for the five evidence loops (126, 127, 128, 137 and 138) shown in (8). Again, we choose this dataset because Lu and Ades [7] identified significant inconsistency in particular for loop 128. As in the CBRE model, we assigned independent $N(0, \sigma_w^2)$ priors to the w factors, where $\sigma_w \sim U(0, 2)$. The resulting posterior means for all w's are near 0 with wide posterior CIs. Also, the posterior distribution of σ_w has the same mean and variance as the prior distribution $\sigma_w \sim U(0, 2)$, indicating that the w factors are not identified by the data in the ABRE model. We further confirmed this using more diffuse uniform priors, such as $\sigma_w \sim U(0, 10)$ or $\sigma_w \sim U(0, 100)$ for the inconsistency factors. The ABRE model without the inconsistency factor was also fit, and estimates for all remaining parameters were very similar in the two fits.

3.7.2 Loop-based Inconsistency Detection using Fixed Effects in ABRE Models

To directly compare the AB and CB loop-based methods to detect inconsistency, we can redefine the 4 subgroups in the AB method's discrepancy factor defined in Section 3.6.1, by dividing trials into smaller groups corresponding to specific loops. For example, to detect the discrepancy of sources of evidence for comparing A vs. B in loop ABC, we can divide the trials into 4 groups: (*i*) trials that include both A and B, (*ii*) trials that include A and C but not B, (*iii*) trials that include B and C but not A, and (*iv*) other trials. Then the posterior distribution of $\Delta_{AB} = (\mu_A^{(i)} - \mu_B^{(i)}) - (\mu_A^{(ii)} - \mu_B^{(iii)})$ can be used to detect the discrepancy between the direct and indirect evidence for comparing A vs. B in loop ABC. This is a more direct analogue to Lu and Ades' method in the AB model framework. However, this throws away useful information because groups (*ii*) and (*iii*) only include the trials within a specific loop, and exclude other trials involving only one of arms A and B. As in Section 3.6.1, we applied this method to the thrombolytic drugs dataset, computing discrepancy factors for 15 different pairwise comparisons, for loop 126, 127, 128, 137, and 138, with 3 comparisons in each loop: 1-2, 1-6, 2-6 for loop 126; 1-2, 1-7, 2-7 for loop 127, etc. As shown in the bottom row of Table 2 ("AB model with loop"), zero is not contained in the 95% posterior CI for the discrepancy factor for comparison 2-8 in loop 128, indicating that inconsistency exists in loop 128 using loop-based AB model method.

4 Simulation Studies

In this section, we generate more extreme inconsistency structures in datasets from CBRE models to evaluate the performance of inconsistency detection using our ABRE models, and compare the results to those of the loop-based method of Lu and Ades [7].

4.1 Simulation Settings

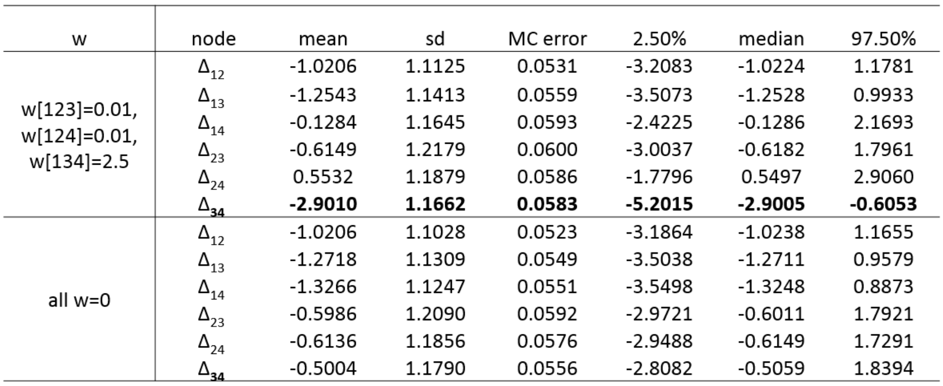
We simulate data for a network meta-analysis from the CBRE model for 30 trials (each with only 2 treatment arms) to compare 4 treatments using binary outcomes as in the smoking cessation dataset. In the CBRE models, the true values of $\alpha_{iB}, i = 1, ..., 30$ were assigned as a sequence of length 30 from -2 to -3. True values of d_{12}, d_{13} and d_{14} were chosen as 0.5, 0.9 and 1.2 respectively, with $w_{123} = w_{124} = 0.01$ and $w_{134} = 2.5$ (what we call the *alternative scenario*). Using (7), the true values of d_{23}, d_{24} and d_{34} were calculated as 0.41, 0.71, and 2.8, respectively. The standard deviation σ in (4) was set equal to 2, and we set $n_{ik} = 100$ for the k^{th} treatment arm in the i^{th} study. Then artificial data Y_{ik} can be generated according to (2) and (3). From this data generation scheme, the evidence in loop 134 is inconsistent, especially in the comparison of 3 vs. 4. This can be confirmed by comparing the randomly generated datasets using the above settings to consistent datasets generated assuming all $w = 0$ but otherwise using the above specification (the *null scenario*). For the experimental design, since there are 6 pairwise comparisons for these 4 treatments (1-2, 1-3, 1-4, 2-3, 2-4, and 3-4), we assigned 5 trials to each comparison, 30 trials in total, each with two arms. We generated 1000 simulated datasets for each scenario using the `Brugs` package in `R`, where we call `OpenBUGS` [21] from `R`, once for each simulated dataset.

4.2 Simulation Results

In this simulation, our hypothesis is that when analyzing data simulated from the alternative scenario, we will be able to detect inconsistency using the fixed effects in AB models for the comparison 3 vs. 4 as stated above. We also perform the analysis using datasets from the null scenario, which should show no evidence of inconsistency using the same method (Section 3.6.1). If the discrepancy factors indicate inconsistency in the dataset, we further investigate the sources of inconsistency by considering the estimated random effects in the AB models. To check the accuracy of the data generation, we first use CB models to fit the datasets from the alternative scenario, and evaluate the coverage probabilities of 95% posterior CIs for these true values. The coverage probabilities of the w factors (w_{123}, w_{124} and w_{134}) and the relative treatment effects (d_{12}, d_{13} and d_{14}) are (0.927, 0.956, 0.935) and (0.936, 0.933 and 0.904), respectively, indicating good data generation and not very influential prior distributions.

Our proposed methods using AB models (as in Section 3.6.1) were then applied. Table 4 displays 2.5% and 97.5% posterior quantiles for the discrepancy factors for these 6 pairwise comparisons averaged over 1000 simulated datasets under each scenario. The posterior mean of Δ_{34} in the simulation for the alternative scenario is -2.90, with 95% Bayesian CI $(-5.20, -0.60)$.

Table 4: Discrepancy factors for simulated datasets using ABRE model. The entry of each column is the summary of results from 1000 datasets under each scenario.

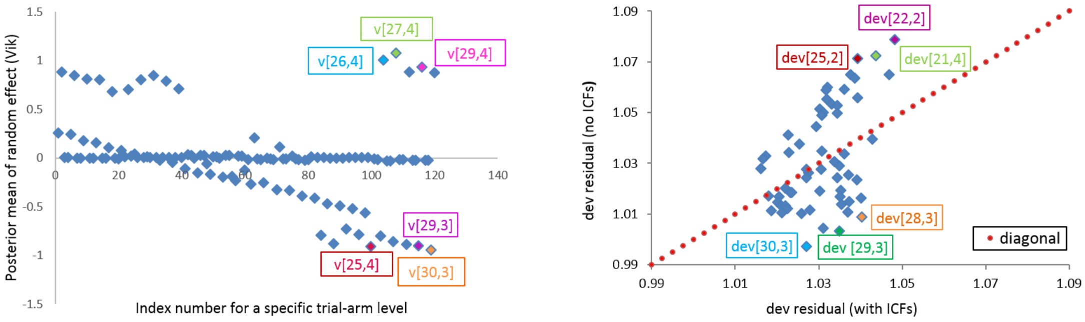


w	node	mean	sd	MC error	2.50%	median	97.50%
w[123]=0.01, w[124]=0.01, w[134]=2.5	Δ_{12}	-1.0206	1.1125	0.0531	-3.2083	-1.0224	1.1781
	Δ_{13}	-1.2543	1.1413	0.0559	-3.5073	-1.2528	0.9933
	Δ_{14}	-0.1284	1.1645	0.0593	-2.4225	-0.1286	2.1693
	Δ_{23}	-0.6149	1.2179	0.0600	-3.0037	-0.6182	1.7961
	Δ_{24}	0.5532	1.1879	0.0586	-1.7796	0.5497	2.9060
	$\boldsymbol{\Delta_{34}}$	**-2.9010**	**1.1662**	**0.0583**	**-5.2015**	**-2.9005**	**-0.6053**
all w=0	Δ_{12}	-1.0206	1.1028	0.0523	-3.1864	-1.0238	1.1655
	Δ_{13}	-1.2718	1.1309	0.0549	-3.5038	-1.2711	0.9579
	Δ_{14}	-1.3266	1.1247	0.0551	-3.5498	-1.3248	0.8873
	Δ_{23}	-0.5986	1.2090	0.0592	-2.9721	-0.6011	1.7921
	Δ_{24}	-0.6136	1.1856	0.0576	-2.9488	-0.6149	1.7291
	$\boldsymbol{\Delta_{34}}$	-0.5004	1.1790	0.0556	-2.8082	-0.5059	1.8394

Since zero is not in the 95% Bayesian CI, we conclude that the two sources of evidence for comparing Arm 3 and 4 are discrepant, and thus inconsistency exists. On the other hand, all the 95% Bayesian CIs for the discrepancy factors cover zero when we set all $w = 0$, indicating no evidence of inconsistency for null datasets using our AB inconsistency detection approach.

The sources of inconsistency for the alternative dataset were further investigated with the AB model random effects (as in Section 3.6.2). As shown in Figure 3a, we have examined the most extreme elements (the top 5% in absolute value) at specific trial-arm levels using the posterior mean of η_{ik} averaged over 1000 simulated datasets: Treatment 4 in Trials 25, 26, 27, 29, and Treatment 3 in Trials 29 and 30. Five out of these six extreme random effects are from comparison 3 vs. 4, suggesting that both inconsistency detection methods using AB models work well on the simulated datasets. We have also checked the other extreme observations in Figure 3a, and found most of them to be from the 3 vs. 4 comparison as well.

Furthermore, Treatment 4 in Trial 26, 27, 29 and Treatment 3 in Trials 29 and 30 are identified as extreme observations by both loop and non-loop methods using AB models (Sections 3.6.1 and 3.7.2). After deletion of these observations, none of the discrepancy factors is significantly different from 0 based on their 95% Bayesian CIs, indicating successful identification of the source of inconsistency in the alternative dataset. For example, the posterior mean of



(a) **(b)**

Figure 3: Simulated datasets under alternative scenario: the most inconsistent trial by treatment combinations are shown in rectangle using either AB or CB models. (a) Extreme random effects in fitting ABRE models. (b) Residuals in fitting CBRE models with and without inconsistency factors.

Δ_{34} after the deletion is -2.78, with 95% Bayesian CI $(-5.77, 0.21)$. However, since the data were generated using the CB model with $w_{134} = 2.5$, inconsistency is present to some degree in all the studies in which both arms 3 and 4 appear. Here, we only delete the top 5 outlying observations, therefore, the posterior mean of the discrepancy factor for comparing 3 vs. 4 is still large, although the CI no longer covers 0.

We have also detected inconsistency using CB models as described in Lu and Ades [7] for comparison. CBRE models with and without ICFs were applied to the simulated alternative datasets, and the large value for $w_{134} = 2.31$ suggests inconsistency in loop 134, though it is not significantly different from 0 according to its 95% Bayesian CI $(-0.30, 4.96)$. Furthermore, the sources of the inconsistency are investigated by comparing the mean residual deviance with and without ICFs using CBRE models. As shown in Figure 3b, Treatment 4 in Trial 21, Treatment 2 in Trials 22 and 25, and Treatment 3 in Trials 28, 29 and 30 are identified as outliers using the CBRE model. Three of these outliers are from loop 134, suggesting slightly weaker inconsistency detection than using our Section 3.6.2 method. Moreover, our discrepancy factor of Δ_{34} *is* significantly different from 0 using the method proposed in Section 3.6.1, indicating more powerful inconsistency detection.

5 Discussion and Future Work

In this paper, we have proposed methods for detection of inconsistency using an arm-based random effects model for network meta-analysis (NMA). Our methods can consider loops but do not need to, and they permit users to address issues previously tackled by CB models, which has limitations due to its focus on relative effects. Compared to Lu and Ades [7], our approach can examine specific comparisons in detail, showing more clearly how indirect evidence combines with or adds to direct evidence to form the NMA estimates (c.f. Dias et al. [8]). Moreover, after the Lu and Ades [7] approach detects inconsistency (say, for loop 128 for the thrombolytic drugs dataset), it still needs to examine the source of inconsistency by checking each comparison in the loop. Using the fixed effects in ABRE models to check the discrepancy of the different sources of evidence for comparing two treatments seems to us a simpler and more direct approach for inconsistency detection, and one that can be done using either all sources of information, or only the information in a specific loop.

We identified similar sources of inconsistency using ABRE models as using CB methods in both examples and in simulated datasets. The significance levels of the inconsistency factors are only summarized in Lu and Ades [7]; they do not describe how big the ICFs should be to declare network inconsistency. On the other hand, our methods flag discrepancy factors significantly different from 0 based on their 95% posterior CIs in both the illustrative and simulated datasets, providing more objective evidence of inconsistency.

We further identified the sources of inconsistency by the most extreme fitted random effects, which we recommend be confirmed using the discrepancy factor method described in Section 3.6.1. For example, in the thrombolytic drugs dataset, none of the trials detected by the random effects method contains comparison of 1 vs. 2, and only Trials 22 and 23 consider the comparison of 2 vs. 8; therefore, these two trials can be identified as the sources of the inconsistency problem. Here, using both of our proposed methods in concert can identify trial-level inconsistency that cannot be detected by the discrepancy factor method alone.

There are of course some limitations or potential concerns associated with our proposed NMA approaches. First, a common concern with AB models is that they "break the randomization" by using fixed effects for treatments across trials. But since we are using a random effects model with unstructured covariance matrix on the random effect terms, this concern is mitigated since it gives the desired correlations between treatments in a specific trial. Second, there exist other good methods for inconsistency detection, such as the node-splitting method of Dias et al. [8] and the multivariate meta-regression of Higgins et al. [9]. We found that our AB approach is a bit like Dias' method, and performs as well. Lastly, the node-splitting method compares "direct evidence" on X vs. Y with what would be predicted from all the remaining evidence (not just the trials including X vs. Y). Although we reported only the discrepancy factor results based on first 3 of the 4 groups described in Section 3.6.1, a re-analysis of our data using the discrepancy factors arising from all 4 groups ("direct" evidence vs. all remaining evidence) gave similar results.

To obtain even better inconsistency detection, more work needs to be done to improve our AB models. Future work looks to extending our methods to continuous, count, or time-to-event outcomes, though the latter will likely require individual-level patient data except for the simplest of models. Rather than noninformative priors, weakly or even informative priors can be used if we have suitable information from historical or observational studies. Finally, individual-level data can be incorporated with the aggregated data summaries used above, allowing borrowing of strength from patient characteristics to better investigate treatment effects and assess inconsistency.

References

1. Lumley T. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* 2002; **21**:2313–2324.

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

3. Dersimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.

4. Achana F, Cooper N, Dias S, Lu G, Rice S, Kendrick D, Sutton A. Extending methods for investigating the relationship between treatment effect and baseline risk from pairwise meta-analysis to network meta-analysis. *Statistics in Medicine* 2013; **32**:752–771.

5. Pharmaceutical Benefits Advisory Committee. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (version 4.3).* Canberra: Australian Government Department of Health and Ageing, 2008.

6. Wells G, Sultan S, Chen L, Khan M, Coyle D. *Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis.* Ottawa: Canadian Agency for Drugs and Technologies in Health, 2009.

7. Lu G, Ades A. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006; **101**:447–459.

8. Dias S, Welton N, Caldwell D, Ades A. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* 2010; **29**:932–944.

9. Higgins J, Jackson D, Barrett J, Lu G, Ades A, White I. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods* 2012; **3**:98–110.

10. Boland A, Dundar Y, Bagust A, Haycox A, Hill R, Mujica Mota R, Walley T, Dickson A. Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation. *Health Technology Assessment* 2003; **7**:1–136.

11. Fiore M, Bailey W, Cohen S, et al. *Smoking Cessation, Clinical Practice Guideline No. 18.* Agency for Health Care Policy and Research, U.S. Department of Health and Human Services, Rockville, MD, 1996.

12. Hong H, Chu H, Zhang J, Carlin B. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research Synthesis Methods* 2015; **0**:to appear.

13. Zhang J, Carlin B, Neaton J, Soon G, Nie L, Kane R, Virnig B, Chu H. Network meta-analysis of randomized clinical trials: Reporting the proper summaries. *Clinical Trials* 2014; **11**:246–262.

14. Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* 2009; **10**:792–805.

15. Salanti G, Higgins J, Ades A, Ioannidis J. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* 2008; **17**:279–301.

16. Spiegelhalter D, Abrams K, Myles J. *Bayesian approaches to clinical trials and health-care evaluation.* John Wiley & Sons, 2004.

17. Carlin B, Louis T. *Bayesian Methods for Data Analysis.* 3rd edn. Chapman and Hall/CRC Press, 2009.

18. Spiegelhalter D, Best N, Carlin B, van der Linde A. The deviance information criterion: 12 years on (with discussion). *J. Roy. Statist. Soc., Ser. B* 2014; **76**:485–493.

19. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades A. *NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials.* NICE Decision Support Unit, 2011.

20. White I, Barrett J, Jackson D, Higgins J. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods* 2012; **3**:111–125.

21. Lunn D, Spiegelhalter DJ, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* 2009; **28**:3049–3067.