Assessing Evidence Inconsistency in Mixed Treatment Comparisons

Guobing Lu and A. E. Ades

Randomized comparisons among several treatments give rise to an incomplete-blocks structure known as mixed treatment comparisons (MTCs). To analyze such data structures, it is crucial to assess whether the disparate evidence sources provide consistent information about the treatment contrasts. In this article we propose a general method for assessing evidence inconsistency in the framework of Bayesian hierarchical models. We begin with the distinction between basic parameters, which have prior distributions, and functional parameters, which are defined in terms of basic parameters. Based on a graphical analysis of MTC structures, evidence inconsistency is defined as a relation between a functional parameter and at least two basic parameters, supported by at least three evidence sources. The inconsistency degrees of freedom (ICDF) is the number of such inconsistencies. We represent evidence consistency as a set of linear relations between effect parameters on the log odds ratio scale, then relax these relations to allow for inconsistency by adding to the model random inconsistency factors (ICFs). The number of ICFs is determined by the ICDF. The overall consistency between evidence sources can be assessed by comparing models with and without ICFs, whereas their posterior distribution reflects the extent of inconsistency in particular evidence cycles. The methods are elucidated using two published datasets, implemented with standard Markov chain Monte Carlo software.

KEY WORDS: Bayesian hierarchical model; Direct and indirect evidence; Evidence cycle; Inconsistency degrees of freedom; Inconsistency factor; Mixed treatment comparison; Random-effects model; WinBUGS.

1. INTRODUCTION

Meta-analysis has focused on summary relative effect measures based on comparisons of two treatments. In the decision-making context, it is perhaps more common to encounter mixed treatment comparisons (MTCs), in which more than two treatments must be compared on an evidence base containing different randomized pairwise or multiarm comparisons. Such data structures, which are analogous to unbalanced incomplete-block designs (Scheffé 1959; Gleser and Olkin 1994), have been explored by several authors. There has been recent interest in whether estimated log odds ratios (LORs), say $\hat{\theta}_{AB}$, based on direct randomized comparisons between treatments A and B, are empirically consistent with indirect estimates $\tilde{\theta}_{AB} = \hat{\theta}_{AC} - \hat{\theta}_{BC}$ based on A versus C and B versus C comparisons (Bucher, Guyatt, Griffith, and Walter 1997; Song, Altman, Glenny, and Deeks 2003). There have also been a range of proposals about how direct and indirect evidence can be combined, including situations in which there may be four or more treatments (Eddy, Hasselblad, and Shachter 1992; Hasselblad and McCrory 1995; Higgins and Whitehead 1996; Domenici, Parmigiani, Wolpert, and Hasselblad 1999; Lumley 2002; Whitehead 2002; Pasty et al. 2003; Lu and Ades 2004).

Some skepticism has been expressed about the validity of mixed treatment comparisons. Some authors have considered these comparisons to be vulnerable to "biases" (Bucher et al. 1997; Song et al. 2003; Yazdanpanah et al. 2004), and they are not accorded a flattering position in the "evidence hierarchy" (McAlistair, Laupacis, Wells, and Sackett 1999), even though all of the aforementioned proposals respect the randomization structure in the evidence. The Cochrane Collaboration (Alderson, Green, and Higgins 2003) explicitly warns against combining direct and indirect evidence. The concern is that statistical combination of potentially very disparate sources of information may be inappropriate, because the information provided by various sources about the model parameters (LORs) may be "inconsistent." In particular, "direct" evidence might be inconsistent with "indirect" evidence, because, for example, they may involve different patient populations or procedures. Statistical methods to assess evidence inconsistency thus become a crucial requirement in MTC analysis.

Our primary objective in this article is to examine inconsistency in MTC evidence structures. We borrow from Lumley's (2002) idea of "incoherence" in networks of pairwise comparisons, but propose a somewhat different analysis. We first distinguish between evidence inconsistency and statistical heterogeneity, with the latter existing between trials making the same comparison and the former lying between different types of comparisons that form evidence cycles (Sec. 4.1). There may be several different inconsistencies in an MTC structure, each corresponding to a particular evidence cycle supported by three or more types of sources. We then characterize an MTC structure by an index called *inconsistency degrees of freedom* (ICDF) that coincides with the number of independent evidence cycles. Under a framework of Bayesian hierarchical modeling, we represent evidence consistency as a set of linear relations between effect parameters on the log odds scale, then relax those relations associated with independent evidence cycles by introducing some random terms, called *inconsistency factors* (ICFs), to allow for evidence inconsistency. We then assess consistency between evidence sources by comparing models with and without ICFs.

The article is organized as follows. We begin by analyzing MTC data structures and parameterization in Section 2, where the comparative relations of interest are elucidated by an undirected graph. In Section 3 we propose Bayesian hierarchical models with fixed or random effects for fitting MTCs under the assumption that the available evidence sources are consistent in estimating all treatment contrasts. In Section 4 we propose a general approach to assessing evidence inconsistency in MTCs based on the concepts of ICDF and ICF.

Guobing Lu is Senior Research Fellow (E-mail: *Guobing.Lu@bristol.ac.uk*) and Tony Ades is Professor (E-mail: *Tony.Ades@bristol.ac.uk*), MRC, Health Services Research Collaboration, University of Bristol, Bristol BS8 2PR, U.K. This research is supported by the Medical Research Council, U.K. The authors thank colleagues who commented on an earlier version of this article presented to the meta-analysis seminar at the Royal Statistical Society, London, June 2004.

Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Applications and Case Studies
DOI 10.1198/016214505000001302

Two datasets, one on smoking cessation and the other of thrombolytic drugs, are used for illustration. Computations are carried out by Markov chain Monte Carlo (MCMC) with WinBUGS software (Spiegelhalter, Thomas, Best, and Lunn 2001). (WinBUGS codes for all of the analyses presented are available at *http://www.hsrc.ac.uk/Current_research/research_programmes/mpes.htm*.) Finally, in Section 6 we discuss the relationship between our approach and previous work on MTCs.

2. MTC DATA STRUCTURES

Suppose that N randomized controlled trials (RCTs) make mixed comparisons between K treatments. In trial i, let r_{ik} denote the number of successful outcomes on treatment k from n_{ik} observations, and let p_{ik} be the probability of a successful outcome on treatment k; then r_{ik} has a binomial distribution $\text{bin}(p_{ik}, n_{ik})$. We use X, Y, and Z to represent variable treatments and use A, B, C, and so on to represent fixed treatments. For notational convenience, we introduce an index set $\mathcal{T}_i$ representing the treatments compared in trial i, and denote the set of all treatments by $\mathcal{T}$. Then the full data are represented as

$$\mathcal{D} = \{(r_{ik}, n_{ik}) : i = 1, \ldots, N; k \in \mathcal{T}_i\}, \tag{1}$$

which can be considered an incomplete $N \times K$ matrix with some empty entries. This type of data structure is analogous to that arising from incomplete-blocks designs (Scheffe 1959; Gleser and Olkin 1994).

2.1 Example 1: Smoking Cessation Data

The data were reported by AHCPR Smoking Cessation Guideline Panel (Fiore et al. 1996) and investigated by Hasselblad (1998). They consist of 24 studies comparing the relative effects of self-help (B), individual counseling (C), and group counseling (D) versus no contact (A) (Table 1). Clearly,

Table 1. Smoking Cessation Rates (r_{ik}/n_{ik}) (Hasselblad 1998)

Baseline	Study number	No contact (A)	Self-help (B)	Individual counseling (C)	Group counseling (D)
$G_{(A)}$	1	9/140		23/140	10/138
$G_{(B)}$	2		11/78	12/85	29/170
$G_{(A)}$	3	79/702	77/694		
	4	18/671	21/535		
	5	8/116	19/146		
	6	75/731		363/714	
	7	2/106		9/205	
	8	58/549		237/1,561	
	9	0/33		9/48	
	10	3/100		31/98	
	11	1/31		26/95	
	12	6/39		17/77	
	13	95/1,107		134/1,031	
	14	15/187		35/504	
	15	78/584		73/675	
	16	69/1,177		54/888	
	17	64/642		107/761	
	18	5/62		8/90	
	19	20/234		34/237	
	20	0/20			9/20
$G_{(B)}$	21		20/49	16/43	
	22		7/66		32/127
$G_{(C)}$	23			12/76	20/74
	24			9/55	3/26

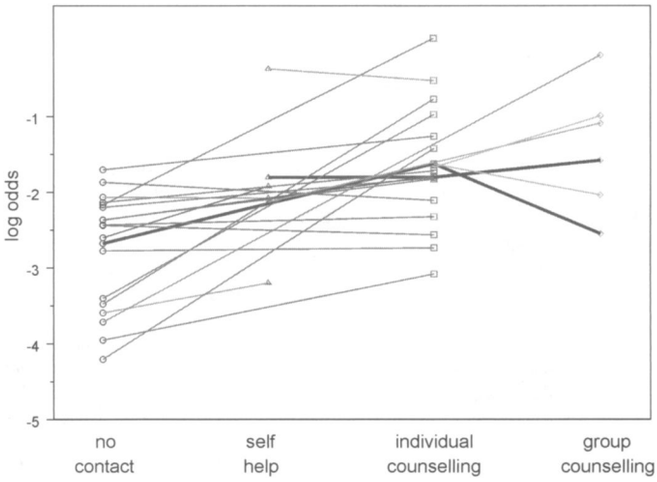

Figure 1. Smoking Cessation Data Represented on the Log Odds Scale. The thick lines correspond to two three-arm trials.

$\mathcal{T} = \{A, B, C, D\}$, $\mathcal{T}_1 = \{A, C, D\}$, $\mathcal{T}_2 = \{B, C, D\}$, $\mathcal{T}_3 = \{A, C\}$, and so on. Here the prime aim of an MTC analysis is to estimate the relative effects of treatments B, C, and D versus the baseline "treatment," A, by synthesizing all of the evidence provided from the data. Figure 1 plots the log odds of smoking cessation rates directly calculated from the 24 studies, giving a visual representation of the available evidence on the various comparisons between the four treatments.

2.2 Graph Representation

The comparative relations can be elucidated by an undirected graph $\mathcal{G} = (\mathcal{T}, \mathcal{E})$, where $\mathcal{T}$ is the set of vertices indicating treatments and $\mathcal{E}$ is the set of undirected edges, a subset of the set $\mathcal{T} \times \mathcal{T}$ of pairs of distinct vertices, denoting comparisons that have been made in one or more trials. For the smoking cessation data, we may write

$$\mathcal{E}_{\text{smoking}} = \{(X, Y) : X \prec Y, X, Y \in \mathcal{T}\},$$

where "$X \prec Y$" means "X is prior to Y" in alphabetic order (e.g., $A \prec B \prec C$). Note that we identify (X, Y) with (Y, X), because there are no directions attached to edges. Each edge represents a comparative relation corresponding to a treatment contrast parameter.

Obviously, a graph corresponding to a proper MTC problem should be connected; otherwise, it can be decomposed into two or more separated subgraphs corresponding to two or more independent subproblems, which can be investigated separately.

The graphic representation of the smoking cessation data is given in Figure 2. The graph is complete, each pair of treatments having been compared in at least one trial. The ordinals of trials that share a common comparison are shown in braces attached to corresponding edges.

2.3 Parameterization

The MTC problem can be recognized as an example of multiparameter evidence synthesis (Eddy et al. 1992; Hasselblad and McCrory 1995; Ades and Cliffe 2002), in which parameters are related to one another by the structure of the problem. A joint probability model for these parameters should reflect

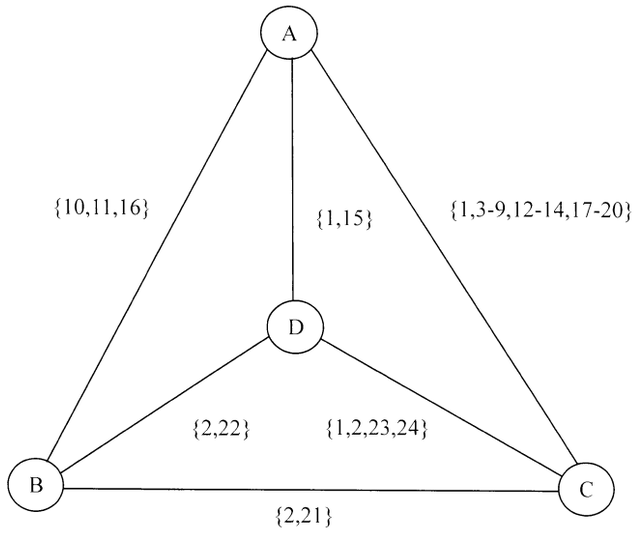

Figure 2. Evidence Network for the Smoking Cessation Data. The figures in braces attached to edges are the ordinals of trials making the corresponding comparisons.

the dependence among them (Gelman, Carlin, Stern, and Rubin 2004). The parameters of interest in smoking cessation data are the six LORs relating each treatment to each of the others, $d_{XY} \equiv \mathrm{LOR}(X, Y)$, $X \prec Y \in \mathcal{T}$. In a single two-arm randomized trial of X versus Y, the full parameterization includes a treatment effect, d_{XY}, relative to a baseline, $\mu = \mathrm{logit}(p_X)$, considered a nuisance parameter,

$$\mathrm{logit}(p_Y) = \mu + d_{XY}. \tag{2}$$

If we extend this to a set $\mathcal{S}_{XY}$ of trials comparing treatments X and Y, then we assume that each trial estimates a baseline parameter μ_i and a single treatment effect d_{XY} (fixed-effects model) or a parameter μ_i and a trial-specific treatment effect δ_{iXY} (random-effects model). A basic assumption in Bayesian hierarchical modeling is that the random effects $\{\delta_{iXY}, i \in \mathcal{S}_{XY}\}$ are exchangeable with a common mean d_{XY} (Bernardo and Smith 1994; Carlin and Louis 2001; Gelman et al. 2004).

If treatment A, the "no contact" treatment, is considered the baseline for relative effects, then, using the terminology of Eddy et al. (1992), the three effect parameters d_{AX} for $X = B, C$, or D may be treated as the *basic parameters*. Then the other three are treated as the *functional parameters* that can be represented as functions of the basic parameters through the following linear relations:

$$\begin{aligned} d_{BC} &= d_{AC} - d_{AB}, \\ d_{BD} &= d_{AD} - d_{AB}, \\ d_{CD} &= d_{AD} - d_{AC}. \end{aligned} \tag{3}$$

Functional parameters must be able to be written in terms of basic parameters; no recursive definition is allowed. Each relation corresponds to a cycle of edges in the graph. Any statistical models built on the relations between basic and functional parameters, such as (3), may be called *models under evidence consistency*, or simply *consistency models*.

In general, any subset of effect parameters can be chosen as basic parameters as long as their corresponding edges can form a *spanning tree* in the graph $\mathcal{G}$, that is, a connected subgraph consisting of all vertices but containing no cycles. Clearly, one new edge added to a spanning tree represents one functional parameter and forms one cycle in the tree. By the time that three new edges are added to the spanning tree formed by AB, AC, and AD in Figure 2, a total of seven cycles have been created: however, only three of these are independent in the sense that if we know that the relations in these three are consistent, then we know that all seven are consistent. Dividing the parameter set $\{d_{XY} : X \prec Y \in \mathcal{T}\}$ into "basic" and "functional" seems to lose symmetry among the roles of the six effect parameters in the model. This is not the case, however. Mathematically, it just provides a parametric way of representing a three-dimensional space that consists of six correlated parameters. The apparent "reduction" in dimensionality is caused by the consistency constraints (3). This is an important insight: If a saturated model has six parameters, and the imposition of consistency relations (3) produces a parameter space of three dimensions, then this suggests that the remaining three parameters represent inconsistency.

We remark that in an incomplete MTC graph, such as that for the thrombolytic drugs data (see Fig. 5), the edge set $\mathcal{E}_{\mathrm{data}}$ is only a proper subset of $\{(X, Y) : X \prec Y, X, Y \in \mathcal{T}\}$. There are some comparisons that have never been made in the available data. Clearly, the effect contrasts associated with them can be represented in terms of the basic parameters and thus can be estimated on purely indirect evidence. However, these contrasts have nothing to do with assessment of evidence consistency, because there is no direct evidence on them. Therefore, by functional parameters we mean only those for which direct evidence is available.

3. MODELS UNDER CONSISTENCY

Bayesian hierarchical models are particularly suitable for fitting data with group structure (Gelman et al. 2004). They have been applied to meta-analysis by many authors, including DuMouchel and Harris (1983), Smith, Spiegelhalter, and Thomas (1995), Higgins and Whitehead (1996), Larose and Dey (1997), Dominici et al. (1999), Prevost, Abrams, and Jones (2000), and Marshall and Spiegelhalter (2001). Our application of this approach to MTCs closely follows that of Higgins and Whitehead (1996), although our account extends to a wider range of MTC structures.

3.1 The General Formulation

For MTC data $\mathcal{D}$ in (1), let the index $b(i)$ denote the baseline treatment specified to trial i and let the set $\mathcal{T}_{(i)} = \mathcal{T}_i \setminus \{b(i)\}$ denote the treatments compared in trial i exclusive of the baseline. Let $\mathbf{d} = (\mathbf{d}_b^T, \mathbf{d}_f^T)^T$ be the vector of parameters for population treatment contrasts on the LOR scale, which are associated with all edges in the MTC graph. Here the superscript T denotes vector or matrix transposition, $\mathbf{d}_b$ is the $(K-1)$ vector of basic parameters, and $\mathbf{d}_f$ is the M vector of functional parameters. Further, let $\mathbf{V}_b$ and $\mathbf{V}_f$ be the corresponding covariance matrices for the treatment contrasts in studies if the effects are considered random. Under the assumption that the evidence about $\mathbf{d}$ provided by all available sources is consistent, the following

consistency relations hold

$$\mathbf{d}_f = \mathbf{F}\mathbf{d}_b, \tag{4}$$

where $\mathbf{F}$ is a known $M \times (K-1)$ matrix representing the linear connections between basic and functional parameters. Then the variance–covariance structure associated with the full-effect parameters can be represented as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_b & \mathbf{V}_b\mathbf{F}^T \\ \mathbf{F}\mathbf{V}_b^T & \mathbf{V}_f \end{pmatrix}, \tag{5}$$

where $\mathbf{V}_f = \mathbf{F}\mathbf{V}_b\mathbf{F}^T$. The fact that we can write $\mathbf{d}_f$ and $\mathbf{V}_f$ in terms of the elements of $\mathbf{d}_b$ and $\mathbf{V}_b$ means that evidence on the functional parameters and their between-trials variances provides information on the basic parameters and their between-trials variance, and vice versa.

Under consistency, a Bayesian random-effects model can be established on the following hierarchical structure:

Level 1: $\prod_{i=1,\ldots,N}\prod_{k\in\mathcal{T}_i}\binom{n_{ik}}{r_{ik}}p_{ik}^{r_{ik}}(1-p_{ik})^{n_{ik}-r_{ik}}$, the likelihood
Linkage: $p_{ik} = g^{-1}(\mu_i + X_{ik}\delta_{ib(i)k})$, $i = 1,\ldots,N$, $k \in \mathcal{T}_i$
Level 2: $f(\delta_{ib(i)k}, k \in \mathcal{T}(i)|\mathbf{d}_b, \mathbf{V}_b)$, $i = 1,\ldots,N$
Level 3: $\pi(\mu_i)$, $\pi(\mathbf{d}_b)$, and $\pi(\mathbf{V}_b)$, the prior distributions.

Here $g(\cdot)$ is the linking function representing the measure scale of the treatment effect, usually taken as the logit function $g(t) = \log(t/(1-t))$; μ_i is the baseline parameter in trial i, corresponding to the specified baseline treatment $b(i)$; and the X_{ik}'s are indicators for baseline treatments given by

$$X_{ik} = \begin{cases} 1, & k \neq b(i) \\ 0, & k = b(i). \end{cases} \tag{6}$$

Thus, for $k \neq b(i)$, δ_{ik} represents the relative random effect of treatment k versus $b(i)$ on log odds scale in the ith study. The conditional densities $f(\cdot)$ at level 2 is usually taken to be a $|\mathcal{T}_{(i)}|$-dimensional normal distribution (a univariate normal if the ith trial is a two-arm comparison or a multivariate normal distribution if the ith trial is a multiarm comparison) with mean vector $(d_{b(i)k}, k \in \mathcal{T}_{(i)})$ and covariance matrix $(v_{k,l}, k, l \in \mathcal{T}_{(i)})$, where the v_{kl}'s are (k,l) entries of matrix $\mathbf{V}$ in (5).

To establish a fixed-effects model, the random effects $\delta_{ib(i)k}$ are replaced by the fixed effects $d_{b(i)k}$ in the linkage, and the procedure at level 2 is ignored. We now apply these models to the smoking cessation data.

3.2 Fixed Effects and Random-Effects Models for Smoking Cessation Data

For the smoking cessation data, the LORs attaching to the functional parameters, and the between-trial variances can now be specified as follows: If

$$\mathbf{d}_b = (d_{AB}, d_{AC}, d_{AD})^T$$

is the vector of basic parameters that spans the space of effect parameters, then the vector of three functional parameters, denoted by $\mathbf{d}_f = (d_{BC}, d_{BD}, d_{CD})^T$, can be expressed as

$$\mathbf{d}_f = \mathbf{F}\mathbf{d}_b, \quad \text{where } \mathbf{F} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Furthermore, suppose that the covariance matrix associated with the basic parameters has the following structure:

$$\mathbf{V}_b = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix},$$
$$\rho_{12}, \rho_{13}, \rho_{23} \geq 0. \tag{7}$$

Nonnegative correlations are proposed here because the between-trials variances in LORs are related through the between-trial variances in their common baselines. Then the variance–covariance structure associated with the full-effect parameters can be represented through (5) with $\mathbf{V}_f = \mathbf{F}\mathbf{V}_b\mathbf{F}^T$ given by

$$\begin{pmatrix} \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2 & \sigma_1^2 - \rho_{12}\sigma_1\sigma_2 - \rho_{13}\sigma_1\sigma_3 + \rho_{23}\sigma_2\sigma_3 & \rho_{12}\sigma_1\sigma_2 - \rho_{13}\sigma_1\sigma_3 - \sigma_2^2 + \rho_{23}\sigma_2\sigma_3 \\ * & \sigma_1^2 + \sigma_3^2 - 2\rho_{13}\sigma_1\sigma_3 & \rho_{12}\sigma_1\sigma_2 - \rho_{13}\sigma_1\sigma_3 - \rho_{23}\sigma_2\sigma_3 + \sigma_3^2 \\ * & * & \sigma_2^2 + \sigma_3^2 - 2\rho_{23}\sigma_2\sigma_3 \end{pmatrix}. \tag{8}$$

Thus the diagonal elements of $\mathbf{V}$ provide a variance structure for describing between-trial heterogeneity. For example, the random effect δ_{iAB} can be considered a sample from normal distribution with expectation d_{AB} and variance σ_1^2, whereas δ_{iBC} has normal distribution with expectation $d_{AC} - d_{AB}$ and variance $\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2$. This provides the covariance structure needed for describing multiarm trials. For example, trial 1 consists of three arms (i.e., baseline A and treatments C and D), and thus the two random effects δ_{1AC} and δ_{1AD} may be assigned to the bivariate normal distribution

$$\begin{pmatrix} \delta_{1AC} \\ \delta_{1AD} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} d_{AC} \\ d_{AD} \end{pmatrix}, \begin{pmatrix} \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}\right). \tag{9}$$

For two-arm trials, because there is only one effect parameter is involved in each trial, only the diagonal elements of $\mathbf{V}$ are needed for modeling.

We partition the dataset into three groups (Table 1), $G_{(A)}$, $G_{(B)}$, and $G_{(C)}$, according to their baselines, namely,

$$G_{(X)} = \{\text{trial } i : b(i) = X\}, \qquad X = A, B, \text{ or } C.$$

Under the foregoing settings, the fixed-effects model can be written as

$$\begin{aligned} r_{ik} &\sim \mathrm{bin}(p_{ik}, n_{ik}), \qquad i = 1,\ldots,N, k \in \mathcal{T}_i; \\ \mathrm{logit}(p_{iX}) &= \mu_{(X)i} \quad \text{for } i \in G_{(X)}, X = b(i); \\ \mathrm{logit}(p_{iY}) &= \mu_{(X)i} + d_{XY}, \qquad Y \in \mathcal{T}_{(i)}. \end{aligned} \tag{10}$$

By replacing the fixed-effects parameters, d_{XY}, with the random-effects versions, δ_{iXY}, the random-effects model is given by

$$\begin{aligned} \mathrm{logit}(p_{iX}) &= \mu_{(X)i} \quad \text{for } i \in G_{(X)}, X = b(i); \\ \mathrm{logit}(p_{iY}) &= \mu_{(X)i} + \delta_{iXY}; \\ \delta_{iXY} &\sim \mathrm{N}(d_{XY}, \sigma_{XY}^2), \qquad Y \in T_{(i)}, i \neq 1, 2, \end{aligned} \tag{11}$$

where the variance σ_{XY}^2 can be specified according to the diagonal elements of $\mathbf{V}$ such as $\sigma_{AB}^2 = \sigma_1^2$ and $\sigma_{BC}^2 = \sigma_1^2 + \sigma_2^2 - \rho_{12}\sigma_1\sigma_2$, and so on. For the first two three-arm trials, the joint

distributions for $(\delta_{AC}, \delta_{AD})^T$ and $(\delta_{BC}, \delta_{BD})^T$ can be specified according to the structure of $\mathbf{V}$, the distribution of the former having been given in (9).

A commonly used case occurs when $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$ and $\rho_{12} = \rho_{13} = \rho_{23} = 1/2$. This in turn implies that $\text{diag}(\mathbf{V}_f) = \text{diag}(\mathbf{V}_b)$, and thus all random effects have the same variance, which is just the case of homogeneity of between-trial variation (Higgins and Whitehead 1996; Lu and Ades 2004). Thus $\delta_{iXY} \sim \text{N}(d_{XY}, \sigma^2)$ for $i \neq 1, 2$, and for the first two three-arm trials, the covariance matrix is given by

$$\begin{pmatrix} \sigma^2 & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 \end{pmatrix}.$$

3.3 Priors

When there is no extra information about the parameters besides the available data, the prior densities can be specified by the "just proper noninformative" distributions (Spiegelhalter, Abrams, and Myles 2004), say,

$$\mu_{(X)i} \sim \text{N}(0, 1{,}000), \qquad d_{XY} \sim \text{N}(0, 1{,}000), \qquad X \prec Y. \tag{12}$$

For σ^2 in the homogeneity model, we adopt the uniform $\sigma \sim \text{U}(0, 2)$ because it gives appropriate weight to the possibility of very low or even zero variance (Warn, Thompson, and Spiegelhalter 2002). In the case of variance heterogeneity, there are six parameters to specify in $\mathbf{V}_b$. We may assume $\sigma_{AB} = \sigma_1$, $\sigma_{AC} = \sigma_2$, $\sigma_{AD} = \sigma_3$, and $iid \sim \text{U}(0, 2)$, and then $\rho_{12}, \rho_{13}, \rho_{23} \sim \text{U}(0, 1)$. Alternatively, noting that the σ_{XY}^2, as the diagonal elements of $\mathbf{V}$ in (5), obey the inequalities

$$\begin{aligned} (\sigma_1 - \sigma_2)^2 \leq \sigma_{BC}^2 \leq \sigma_1^2 + \sigma_2^2, \\ (\sigma_1 - \sigma_3)^2 \leq \sigma_{BD}^2 \leq \sigma_1^2 + \sigma_3^2, \\ (\sigma_2 - \sigma_3)^2 \leq \sigma_{CD}^2 \leq \sigma_2^2 + \sigma_3^2, \end{aligned} \tag{13}$$

when ρ_{12}, ρ_{13}, and ρ_{23} are all positive, we may specify the between-trials variances attaching to the functional parameters as

$$\begin{aligned} \sigma_{BC}/\sigma_1, \sigma_2 &\sim \text{U}\big(|\sigma_1 - \sigma_2|, (\sigma_1^2 + \sigma_2^2)^{1/2}\big), \\ \sigma_{BD}/\sigma_1, \sigma_3 &\sim \text{U}\big(|\sigma_1 - \sigma_3|, (\sigma_1^2 + \sigma_3^2)^{1/2}\big), \\ \sigma_{CD}/\sigma_2, \sigma_3 &\sim \text{U}\big(|\sigma_2 - \sigma_3|, (\sigma_2^2 + \sigma_3^2)^{1/2}\big). \end{aligned} \tag{14}$$

In evidence structures where most of the pairwise comparisons were informed by a large number of trials, this would allow a detailed exploration of variations in between-trial heterogeneity. But the data in Table 1 provide evidence on only one of the between-trials variances, so we adopted the following, less-structured model with hyperparameters σ_0^2 and v_{XY}:

$$\begin{aligned} \log \sigma_{XY} &= \log \sigma_0 + v_{XY}, \\ \sigma_0 &\sim \text{U}(0, 2), \qquad v_{XY} \sim \text{N}(0, \psi^2), \end{aligned} \tag{15}$$

where ψ is a prespecified constant that gauges the degree of heterogeneity in between-trial variance. We set the value of $\log(2\psi)$ to 2, expressing a prior belief that the true values of the σ_{XY} were unlikely to differ from σ_0 by more than a factor of 2 either way.

3.4 Goodness of Fit

The goodness-of-fit of a model to data can be measured by the sum of residual deviance (McCullagh and Nelder 1989; Congdon 2003). For binomial data, it is given by

$$\begin{aligned} \bar{D} &= \sum_i \text{Dev}_i \\ &= \sum_i \sum_{k \in \mathcal{T}_i} 2\left\{ r_{ik} \log\left[\frac{r_{ik}}{n_{ik} p_{ik}}\right] \right. \\ &\qquad \left. + (n_{ik} - r_{ik}) \log\left[\frac{n_{ik} - r_{ik}}{n_{ik} - n_{ik} p_{ik}}\right] \right\}. \end{aligned}$$

For model selection and comparison, we use the deviance information criterion (DIC) $DIC = \bar{D} + p_D$ (Spiegelhalter, Best, Carlin, and van der Linde 2002), where p_D is an estimate of the effective number of parameters, particularly useful in Bayesian hierarchical modeling.

3.5 Results for the Smoking Cessation Data

Both fixed-effects and random-effects models are applied to the smoking cessation data. The results given in Table 2 (and all results thereafter) are based on a WinBUGS run of 20,000 updates after a 30,000-run burn-in, using the method of Gelman and Rubin (1992) to check convergence. In addition to the effect parameters d_{XY}, the posterior probabilities that "X is the 'best' treatment among A, B, C, and D" are also of practical interest in MTC meta-analysis. They can be defined through rank statistics and are given by

$$\Pr\{X \text{ is the best} | \mathcal{D}\} = \Pr\{\text{rank}(T_X) = 1 | \mathcal{D}\} \quad \text{for } X \in \mathcal{T}, \tag{16}$$

where T_X is some "loss function" associated with each treatment. Thus the event "X is the best" can be read as "P_X is the maximum among P_A, P_B, P_C, and P_D." These posterior probabilities provide an intuitive way of ranking treatments. Moreover, they are easy to evaluate in an MCMC framework. For the purpose of illustration, we choose $T_X = 1 - P_X$ as such a loss function for ranking treatments, where P_X is the average smoking cessation rate on treatment X. MTC analyses have been embedded in cost-effectiveness models with realistic loss functions (e.g., Bridle et al. 2004), and Pr{X is most cost-effective} forms the basis for cost-effectiveness acceptability curves (Fenwick, Claxton, and Sculpher 2001).

The fixed-effects model fits poorly with $\bar{D}$ as large as 267, compared with 54 in the random-effects model. In a well-fitting model, $\bar{D}$ is expected to approximate the number of binomial data points (50). The clear advantage of the random-effects model persists when we apply a penalty for extra parameters, with DIC of 99 compared with 294 under the fixed-effects model. The data do not distinguish between random-effects models under homogeneity and heterogeneity assumptions. However, given the a priori likelihood of extreme differences between protocols in each of the three active interventions, the heterogeneity model seems more reasonable on epidemiologic grounds and perhaps gives a fairer account of the posterior uncertainty in mean treatment differences.

Table 2. MTC Analysis Under Consistency Models for Smoking Cessation Data

Statistical quantity	Fixed effects	Random effects		
		Homogeneous	Heterogeneous	Homogeneous with w-factors
DIC	294.1	99.0	99.0	99.3
$\bar{D}$	267.1	54.2	54.0	54.0
pD	27.0	44.8	45.0	45.3
d_{AB}	$.223_{(.127)}$	$.494_{(.399)}$	$.505_{(.397)}$	$.449_{(.441)}$
d_{AC}	$.766_{(.059)}$	$.844_{(.236)}$	$.838_{(.239)}$	$.860_{(.248)}$
d_{AD}	$.839_{(.175)}$	$1.101_{(.437)}$	$1.13_{(.453)}$	$1.088_{(.516)}$
d_{BC}	$.543_{(.135)}$	$.350_{(.411)}$	$.333_{(.403)}$	$.258_{(.519)}$
d_{BD}	$.616_{(.192)}$	$.607_{(.486)}$	$.625_{(.489)}$	$.636_{(.546)}$
d_{CD}	$.074_{(.171)}$	$.257_{(.412)}$	$.293_{(.425)}$	$.226_{(.545)}$
σ^2		.731		.783
σ_0^2			.795	
σ_{AB}^2			1.008	
σ_{AC}^2			.768	
σ_{AD}^2			1.103	
σ_{BC}^2			.781	
σ_{BD}^2			.776	
σ_{CD}^2			.842	
σ_w^2				.607
$\Pr(\sigma_w^2 > \sigma^2)$				.266
w_{ABC}				$-.154_{(.484)}$
w_{ABD}				$-.004_{(.519)}$
w_{ACD}				$-.003_{(.510)}$
Pr(*A* 1st)	0	0	0	0
Pr(*B* 1st)	0	.057	.053	.065
Pr(*C* 1st)	.330	.235	.214	.283
Pr(*D* 1st)	.670	.708	.733	.652

NOTE: Reported values are the posterior means (standard deviation).

4. ASSESSING EVIDENCE INCONSISTENCY

4.1 Heterogeneity and Inconsistency

Following Lumley (2002), we first distinguish between evidence inconsistency and statistical heterogeneity in MTC problems. In pairwise meta-analysis, the term "heterogeneity" is used to describe between-trials variability in effect size estimates that exceeds the variation expected from sampling error alone (Thompson, Smith, and Sharp 1997; Sutton, Abrams, Jones, Sheldon, and Song 1998; Higgins, Thompson, Deeks, and Altman 2002). Heterogeneity is a form of between-study discrepancy, caused by artificial or real differences in a treatment effect across the different trials (Berry 1998; Glasziou and Sanders 2002). In MTCs, there are more than one pairwise comparison, each with its own measure of between-trial variation and is related to the others in complex ways, as shown in $\mathbf{V}_b$ and $\mathbf{V}_f$.

Evidence inconsistency is another sort of discrepancy that lies between the pairwise comparisons rather than between individual trials. Although the relationships (3) between basic and functional parameters are the logical results of their definitions, the evidence provided by different groups of trials may not conform to these relationships. For example, the parameter d_{BC} is supported directly by evidence from trial group $\{2, 21\}$ and is also supported in two ways by indirect evidence: first, by the evidence about d_{AC} from trial group $\{1, 3–9, 12–14, 17–20\}$ and evidence about d_{AB} from trial group $\{10, 11, 16\}$, and second, by indirect evidence on d_{BD} from $\{2, 22\}$ and on d_{CD} from $\{1, 2, 23, 24\}$. If the discrepancy in the results of the studies goes beyond that explained by sampling error and between-trial heterogeneity, then we may say the direct and indirect evidence are inconsistent as sources of effect size estimation.

4.2 Inconsistency Factors and Inconsistency Degrees of Freedom

If the equations in (3) express consistency relations, by definition, inconsistency must be

$$\begin{aligned} d_{BC} &= d_{AC} - d_{AB} + w_{ABC}, \\ d_{BD} &= d_{AD} - d_{AB} + w_{ABD}, \\ d_{CD} &= d_{AD} - d_{AC} + w_{ACD}, \end{aligned} \tag{17}$$

where the terms w, called ICFs, represent the discrepancy between the evidence supporting the functional parameter on the left side and the difference between the basic parameters on the right side. In the general formulation of the model, ICFs are added to the equations in (4),

$$\mathbf{d}_f = \mathbf{F}\mathbf{d}_b + \mathbf{w}. \tag{18}$$

Because each consistency relationship lies on a cycle, the ICFs are, in fact, attached to corresponding cycles rather than to individual edges. The size of an ICF, say w_{ABC} in (17), can reflect evidence inconsistency among items informing the cycle *ABCA,* not merely to an "inaccuracy" or "bias" in d_{BC}.

The introduction of these ICFs in (17) effectively restores the dimensionality of the model (which had been reduced from six

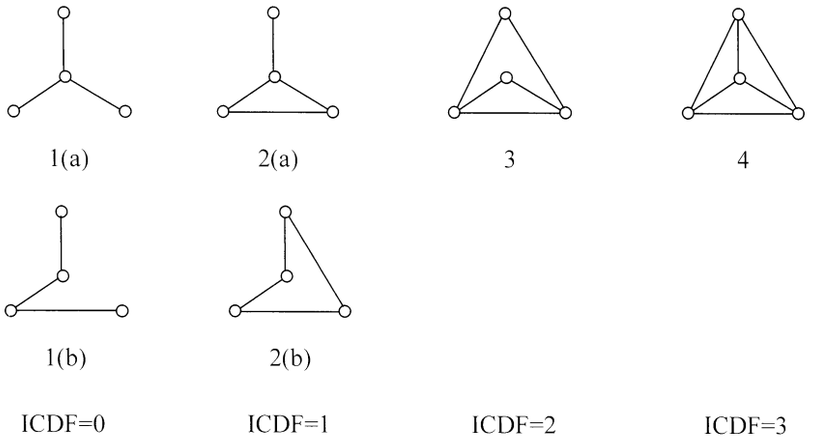

Figure 3. Four-Treatment MTCs With Different Inconsistency Degrees of Freedom. As ICDF varies from 0 to 3, the complexity of the MTC structure increases.

to three by the consistency relations) back to six. The increase in dimension corresponds to the potential number of inconsistencies in the structure, which we call the ICDF. Obviously, the greater the ICDF, the more complex the MTC structure should be (see Fig. 3). For most MTC evidence structures, including pairwise comparison networks, the natural intuition that ICDF is the number of functional parameters turns out to be correct, that is,

$$\text{ICDF} = \#(\mathbf{d}_f) = T - K + 1, \tag{19}$$

where K is the number of treatments, and T is the number of comparisons informed by data. However, a slight modification of (19) is needed when certain treatment contrasts are provided only from multi-arm trials.

Because of an essential prerequisite that *multi-arm trials must be internally consistent*, any treatment contrasts informed solely by a multi-arm trial can be inconsistent only with (two or more) contrasts informed by two or more other evidence sources. We thus define *inconsistency* in the context of MTC as a relation between a functional parameter and $R \geq 2$ basic parameters, in which all $R + 1$ parameters are supported by at least three independent sources of evidence. The ICDF in an MTC evidence structure is just the number of independent inconsistencies.

It follows immediately that $0 \leq \text{ICDF} \leq (K - 1)(K - 2)/2$, and in a network consisting only of pairwise comparisons, $\text{ICDF} = \#(\mathbf{d}_f)$, the number of independent functional parameters. However, in general, the foregoing definition gives $\text{ICDF} = \#(\mathbf{d}_f) - S$, where S is the number of independent inconsistency relations in which the corresponding parameters are supported by no more than two independent sources of evidence. So there is no inconsistency in the structure $\{A \text{ vs } B, A \text{ vs } B \text{ vs } C\}$ as $\#(\mathbf{d}_f) = S = 1$, even though it provides evidence on d_{AB}, d_{AC}, and d_{BC}. The calculation of S may become difficult when the MTC net includes many multi-arm trials, and in such cases we have not managed to find a general formula or a mechanical routine to count it. In practice, an inconsistency model must be programmed very carefully, and the ICDF may have to be counted "by hand."

Another difficulty created by multi-arm trials lies in the choice of basic parameters. For example, in the structure $\{A \text{ vs } B, A \text{ vs } C, A \text{ vs } B \text{ vs } C\}$, a consistency model could equivalently have the three-arm trial a bivariate normal on pairs of parameters (d_{AB}, d_{AC}), (d_{AB}, d_{BC}), or (d_{BC}, d_{AC}) given the relation $d_{BC} = d_{AC} - d_{AB}$. The three versions of the MTC model are essentially equivalent. In an inconsistency model, however, if the first parameterization is chosen, then the inconsistency relation $d_{BC} = d_{AC} - d_{AB} + w_{ABC}$ cannot be expressed, because the parameter d_{BC} is already implicitly defined by the parameterization of the three-arm trial. Therefore, to estimate an inconsistency model for a dataset where certain contrasts are examined *only* in multi-arm trials, care must be taken to parameterize the likelihood to express these contrasts in model parameters.

Finally, because an ICF that is positive on one parameterization may become negative on another, there is no possibility that "indirect" evidence could systematically overestimate or underestimate the "true" effects as some have feared (Bucher et al. 1997; Song et al. 2003).

4.3 Assessing Inconsistency

If we assume that the ICFs have a common normal distribution with mean 0 and variance, say, σ_w^2, then the posterior mean of each ICF is a measure of the extent of inconsistency within the corresponding evidence cycle. For example, suppose that the absolute value for the posterior mean of w_{ABC} is relatively large and the corresponding values for the other two ICFs are small; this casts doubt on the consistency of the evidence cycle $ABCA$. In addition, we can compare σ_w^2, taken as an overall measure of inconsistency, with the between-trials heterogeneity σ^2. The posterior probability

$$\Pr(\sigma_w^2 > \sigma^2 | \mathcal{D}) \tag{20}$$

may give an approximate summary, with a high value signaling potential evidence inconsistency. The suggestion is that σ^2 might serve as a benchmark by which to assess σ_w^2, not that σ^2 and σ_w^2 are necessarily related.

4.4 Results for Smoking Cessation Data (Continued)

For assessing evidence consistency in the smoking cessation data, we add three w-factors to the random-effects model using the relations (17) instead of (3). Results with and without

the w-factors are shown in Table 2. Posterior means and standard variations for all effect parameters are similar, and neither the global model fit statistics nor the inconsistency p value of .27 suggest the presence of serious inconsistency. The three w-factors are all small, both in relation to the between-trial variance and in relation to the mean effect estimates. Nevertheless, both the posterior mean inconsistency variance, σ_w^2, and heterogeneity, σ^2, are large, and investigators cannot confidently conclude that there is no inconsistency.

4.5 Theoretical Notes on Alternative Parameterizations

Note 1. Consider a consistency model. For a specified parameterization, as mentioned earlier, the basic parameters form a basis for the space of the effect parameter vector $\mathbf{d} = (\mathbf{d}_b^T, \mathbf{d}_f^T)^T$ under the consistency relations (4). Let $\mathbf{d}'_b$ and $\mathbf{d}'_f$ be the basic and functional parameters under another parameterization with consistency relation equations $\mathbf{d}'_f = \mathbf{F}'\mathbf{d}'_b$. Clearly, $\mathbf{d}_b$ and $\mathbf{d}'_b$ have the same dimension $(K-1)$, $\mathbf{d}_f$ and $\mathbf{d}'_f$ have the same dimension M, and the two sets of consistency relations are equivalent in the sense that

$$\{\mathbf{d} : \mathbf{d}_f = \mathbf{F}\mathbf{d}_b\} = \{\mathbf{d} : \mathbf{d}'_f = \mathbf{F}'\mathbf{d}'_b\} \equiv \mathcal{H}. \tag{21}$$

That is, under the assumption of consistency, alternative parameterizations are merely different representations of the same hyperplane $\mathcal{H}$. Thus any two parameterizations are equivalent when there is no prior information about particular basic parameters.

The matrix $\mathbf{F}'$ can be derived as follows. Let $\mathbf{P}$ be a permutation matrix such that

$$\begin{pmatrix} \mathbf{d}_b \\ \mathbf{d}_f \end{pmatrix} = \mathbf{P}\begin{pmatrix} \mathbf{d}'_b \\ \mathbf{d}'_f \end{pmatrix}, \quad \text{where } \mathbf{P} \equiv (\mathbf{P}_1 \quad \mathbf{P}_2) \equiv \begin{pmatrix} \mathbf{P}_{(1)} \\ \mathbf{P}_{(2)} \end{pmatrix},$$

where $\mathbf{P}_1$ is $T \times (K-1)$, $\mathbf{P}_2$ is $T \times M$, $\mathbf{P}_{(1)}$ is $(K-1) \times T$, $\mathbf{P}_{(2)}$ is $M \times T$, and $T = M + (K-1)$. Recall that a permutation matrix $\mathbf{P}$ is a square matrix with entries all 0 and 1 and precisely one 1 in each row and column of $\mathbf{P}$. Clearly, $\mathbf{P}$ is an orthogonal matrix, and left (right) multiplication of a matrix $\mathbf{A}$ by a permutation matrix $\mathbf{P}$ permutes the rows (columns) of $\mathbf{A}$ (Horn and Johnson 1985). Thus $\mathbf{P}^{-1} = \mathbf{P}^T$, and $\mathbf{d}' = \mathbf{P}^T\mathbf{d}$. The hyperplane (21) can be rewritten as

$$\{\mathbf{d} : (-\mathbf{F}, \mathbf{I}_M)\mathbf{d} = \mathbf{0}\} = \{\mathbf{d} : (-\mathbf{F}, \mathbf{I}_M)\mathbf{P}\mathbf{d}' = \mathbf{0}\}, \tag{22}$$

where $\mathbf{I}_M$ is an M-order identity matrix. It then follows that

$$\mathbf{F}' = [(-\mathbf{F}, \mathbf{I}_M)\mathbf{P}_2]^{-1}[(\mathbf{F}, -\mathbf{I}_M)\mathbf{P}_1] \equiv \mathbf{B}^{-1}\mathbf{A}. \tag{23}$$

Similarly,

$$\mathbf{F} = \left[(-\mathbf{F}', \mathbf{I}_M)\mathbf{P}_{(2)}^T\right]^{-1}\left[(\mathbf{F}', -\mathbf{I}_M)\mathbf{P}_{(1)}^T\right] \equiv \mathbf{B}'^{-1}\mathbf{A}'. \tag{24}$$

Note that both $\mathbf{F}$ and $\mathbf{F}'$ are matrices with entries 0, 1, or -1, because each row of them is associated with a cycle. Furthermore, we have a 1–1 linear transformation between $\mathbf{d}_b$ and $\mathbf{d}'_b$,

$$\mathbf{d}'_b = \mathbf{P}_1^T\begin{pmatrix} \mathbf{I}_{K-1} \\ \mathbf{F} \end{pmatrix}\mathbf{d}_b \qquad \text{or} \qquad \mathbf{d}_b = \mathbf{P}_{(1)}\begin{pmatrix} \mathbf{I}_{K-1} \\ \mathbf{F}' \end{pmatrix}\mathbf{d}'_b. \tag{25}$$

Note 2. For an inconsistency model, suppose that $\mathbf{w}$ and $\mathbf{w}'$ are the vectors of ICFs associated with consistency equations under the foregoing two parameterizations. Then

$$\mathbf{w} = \mathbf{d}_f - \mathbf{F}\mathbf{d}_b \qquad \text{and} \qquad \mathbf{w}' = \mathbf{d}'_f - \mathbf{B}^{-1}\mathbf{A}\mathbf{d}'_b$$

imply that

$$\mathbf{B}\mathbf{w}' = \mathbf{B}\mathbf{d}'_f - \mathbf{A}\mathbf{d}'_b = (-\mathbf{F}, \mathbf{I}_M)\mathbf{P}\mathbf{d}' = (-\mathbf{F}, \mathbf{I}_M)\mathbf{d} = \mathbf{w}. \tag{26}$$

Similarly,

$$\mathbf{w}' = (-\mathbf{F}', \mathbf{I}_M)\mathbf{P}_{(2)}^T\mathbf{w} = \mathbf{B}'\mathbf{w}, \qquad \text{and thus} \qquad \mathbf{B}' = \mathbf{B}^{-1}. \tag{27}$$

Formulas (26) and (27) reveal that ICFs under one parameterization can be represented as linear combinations of ICFs under another parameterization. Thus all of the possible inconsistencies on an MTC net can be detected by the inconsistencies on the cycles to which the specified $\mathbf{w}$-factors are attached, and in this sense an inconsistency model can be indexed by the "core" parameter set $\{\mathbf{d}_b, \mathbf{w}\}$ of any parameterization, provided that there is no prior information about particular effect parameters and there are no evidence cycles with inconsistencies of particular interest. By combining (25)–(27), we find that the reparameterization of MTC is equivalent to making a 1–1 linear transformation between the corresponding core parameters. This therefore implies that, setting aside the effects of prior information and nuisance parameters, the posterior estimation of parameters of interest, the model fit, and the deviance statistics should not be affected by the choices of basic parameters.

Clearly, one significantly large value of ICF, under whatever parameterization, is sufficient to reject the assumption of consistency. For accepting consistency or for locating the source of inconsistency, however, one needs to examine ICFs attached to all evidence cycles, including the vector $\mathbf{w}$ but also its corresponding linear combinations. This may be more difficult harder as the complexity of MTC increases. As we show in next example, some statistical diagnostics, such as deviance analysis and cross-validation, may help in such situations.

5. ANOTHER EXAMPLE: META–ANALYSIS FOR THROMBOLYTIC DRUGS

The dataset consists of 28 trials comparing $K = 8$ thrombolytic treatments after acute myocardial infarction: streptokinase (*SK*), alteplase (*tPA*), accelerated alteplase (*AtPA*), reteplase (*Ret*), tenecteplase (*Ten*), streptokinase plus alteplase ($SK + tPA$), urokinase (*UK*), and anistreptilase (*ASPAC*). This is a set of treatments and studies defined in a comprehensive systematic review (Boland et al. 2003), except that the trials involving either *ASPAC* or *UK* were excluded from the original analyses, because these treatments were no longer available in the United Kingdom. The log odds of mortality at 30–35 days in the 28 studies are plotted in Figure 4.

5.1 Parameterization

Among the total $K(K-1)/2 = 28$ potential pairs of comparisons, $T = 13$ pairwise comparisons are independently supported by direct evidence from the data (Fig. 5). For describing all possible treatment effects in a model, we need to specify $K - 1 = 7$ basic parameters that can form a spanning tree. It would be natural to choose the effects of the seven new treatments (i.e., *AtPA*, *tPA*, $SK + tPA$, *Ten*, *Ren*, *UK* and *ASPAC*) relative to the old treatment *SK*. However, no direct evidence is

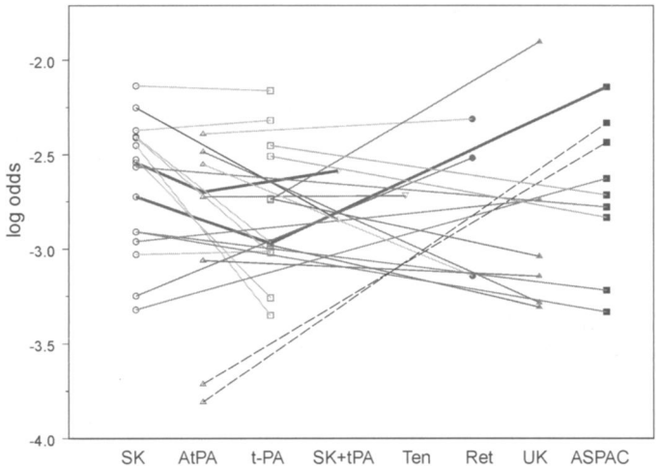

Figure 4. Thrombolytic Drugs Data Represented on the Log Odds Scale. The thick lines correspond to the three-arm trials, and the two dashed lines correspond to trials 22 and 23.

available for an *SK* versus *Ten* comparison. Treatment *Ten* is an isolated vertex (see Fig. 5) compared only with treatment *AtPA* in trial 17, and thus we must have $d_{AtPA,Ten}$, instead of $d_{SK,Ten}$, as a basic parameter. Therefore, we have the following seven basic parameters (represented by the solid lines in Fig. 5): $d_{SK,AtPA}$, $d_{SK,tPA}$, $d_{SK,SK+tPA}$, $d_{SK,Ret}$, $d_{SK,UK}$, $d_{SK,ASPAC}$, and $d_{AtPA,Ten}$.

One of the functional parameters $d_{AtPA,SK+tPA}$ is only estimated by the three-arm study 1. The relationship $d_{AtPA,SK+tPA} = d_{SK,SK+tPA} - d_{SK,AtPA} + w_{SK,AtPA,SK+tPA}$ relating this parameter to basic parameters does not qualify as an inconsistency relation, however, because one of the basic parameters on the right side, $d_{SK,AtPA}$ is only estimated by the same trial. (This is the $\{A$ vs B, A vs B vs $C\}$ structure discussed in Sec. 4.2.) Thus the three parameters in the relationship are not supported by three separate evidence sources, so there is no information with which to estimate the w-factor. (It also follows that, because trial 1 does not generate an inconsistency factor, it does not matter how the bivariate normal is set up.) This gives ICDF $= \#(\mathbf{d}_f) - S = 13 - 7 - 1 = 5$, with the inconsistency

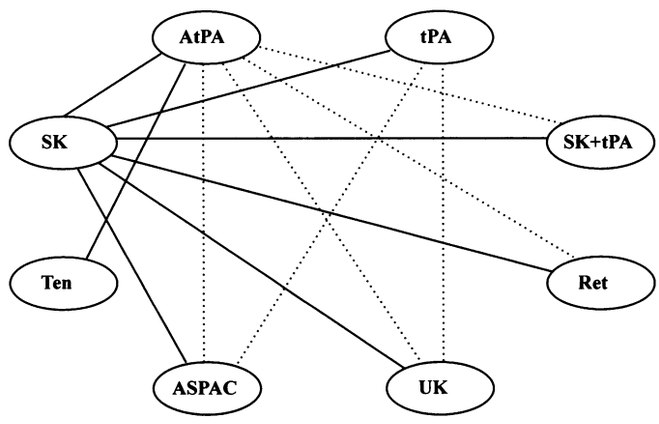

Figure 5. Thrombolytic Drugs Data: Network of 13 Types of Comparisons. The solid lines represent comparisons whose treatment contrasts are specified as basic parameters. All solid lines form a spanning tree. The dotted lines represent comparisons associated with functional parameters.

relations

$$\begin{aligned} d_{AtPA,Ret} &= d_{SK,Ret} - d_{SK,AtPA} + w_{SK,AtPA,Ret}, \\ d_{AtPA,UK} &= d_{SK,UK} - d_{SK,AtPA} + w_{SK,AtPA,UK}, \\ d_{AtPA,ASPAC} &= d_{SK,ASPAC} - d_{SK,AtPA} + w_{SK,AtPA,ASPAC}, \\ d_{tPA,UK} &= d_{SK,UK} - d_{SK,tPA} + w_{SK,tPA,UK}, \\ d_{tPA,ASPAC} &= d_{SK,ASPAC} - d_{SK,tPA} + w_{SK,tPA,ASPAC}. \end{aligned} \tag{28}$$

5.2 Results of Homogeneous Variance Random-Effects Models With and Without Inconsistency Factors

In the random-effects consistency model, the posterior mean of the between-trials variance is very small (.020), suggesting that similar results would be obtained with a fixed-effects model. Results with and without the inconsistency factors are given in Table 4. Both the goodness-of-fit statistics, *DIC* and $\bar{D}$, take reduced values when the inconsistency factors are included in the model. Taken together with the high inconsistency probability, $\Pr\{\sigma_w^2 > \sigma^2|\mathcal{D}\}$, .932, this strongly suggests the presence of inconsistency between sources of evidence on posterior treatment effects. The particularly high value of $w_{SK,AtPA,ASPAC}$, .678, suggests the presence of inconsistency in the $\{SK, AtPA, ASPAC\}$ evidence cycle.

We also compare the mean residual deviance for individual data points in models with and without inconsistency factors. Figure 6 reveals four obvious outliers, the arms of trials 22 and 23, in the model without w-factors. These are the two trials comparing *AtPA* and *ASPAC*, confirming suspicions about the $\{SK, AtPA, ASPAC\}$ cycle. Whether or not we can infer that trials 22 and 23 are the deviant data source is taken up in Section 5.4.

An inconsistency model with four ICFs, based on the data remaining after trials 22 and 23 are removed (see Table 4), has ICFs and inconsistency variance substantially reduced. Interestingly, the posterior uncertainty in the mean treatment effects is returned to the level seen in the original consistency model based on full data.

5.3 Cross-Validation

Cross-validation is another way of checking the compatibility of two independent sets of evidence, in this case trials 22 and 23 versus the remaining data. Let $\mathcal{D}_{AtPA,ASPAC}$ denote the subset of data comprising outcomes from trials 22 and 23 [i.e., $\{(r_{i,AtPA}, n_{i,AtPA}), (r_{i,ASPAC}, n_{i,ASPAC}), i = 22, 23\}$], and let $\mathcal{D}_{\setminus AtPA,ASPAC}$ denote the remaining data (i.e., $\mathcal{D}_{\setminus AtPA,ASPAC} = \mathcal{D} \setminus \mathcal{D}_{AtPA,ASPAC}$). Following Gelman et al. (2004), we can generate replicated samples for trials 22 and 23 (with sample sizes $n_{i,AtPA}$ and $n_{i,ASPAC}$ fixed) from the posterior predictive distribution based on the remaining data, $\mathcal{D}_{\setminus AtPA,ASPAC}$, under an inconsistency model with ICDF $= 4$, then compare these samples with the observed data by means of a prechosen statistic. Here we choose the pooled LOR for the two *AtPA* versus *ASPAC* trials as such a checking function. In detail, we used Woolf's (1955) statistic, a precision-weighted average of the trial log LORs, with .5 added to all cell frequencies (Gart and Zweiful 1967). Based on 20,000 updates in MCMC simulations, the predictive distribution of this statistic was approximately normal with mean .0046 and standard deviation .241, compared with

Table 3. Thrombolytic Drugs Data (death in 30 or 35 days/patients)

Baseline	Study number	SK (1)	AtPA (2)	t − PA (3)	SK + tPA (4)	Ten (5)	Ret (6)	UK (7)	ASPAC (8)
1	1	1,472/20,163	652/10,344		723/10,328				
1	2	1,455/13,780		1,418/13,746					1,448/13,773
1	3	9/130		6/123					
1	4	5/63		2/59					
1	5	3/65		3/64					
1	6	887/10,396		929/10,372					
1	7	7/85		4/86					
1	8	12/147		7/143					
1	9	10/135		5/135					
1	10	4/107			6/109				
1	11	285/2,992					270/2,994		
1	12	10/203						7/198	
1	13	3/58							2/52
1	14	3/86							6/89
1	15	3/58							2/58
1	16	13/182							11/188
2	17		522/8,488			523/8,461			
2	18		356/4,921				757/10,138		
2	19		13/155				7/169		
2	20		2/26					7/54	
2	21		12/268					16/350	
2	22		5/210						17/211
2	23		3/138						13/147
3	24			8/132				4/66	
3	25			10/164				6/166	
3	26			6/124				5/121	
3	27			13/164					10/161
3	28			7/93					5/90

an observed value of 1.270. The Bayesian p value is .0027, which constitutes strong evidence for a conflict between trials 22 and 23 comparing *AtPA* and *ASPAC*, and data sources for other comparisons.

5.4 Locating the Source of Inconsistency

Both cross-validation and the posterior distribution of $w_{SK,AtPA,ASPAC}$ point strongly to a potential inconsistency problem associated with trials 22 and 23. Furthermore, removing this set of trials effectively resolves the inconsistency, in the sense that the model based on the reduced dataset has satisfactory global deviance statistics and reasonably small inconsistency factors. However, we cannot immediately conclude that trials 22 and 23 are the "cause" of the inconsistency, because it remains entirely possible that removing several other types of data would have the same effect because inconsistency is a property of evidence cycles, not of individual data points.

The deviance statistics pointing specifically to trials 22 and 23 are also potentially misleading. Apparently, poor fit in one part of an evidence network compared with another only reflects unequal *amounts* of evidence coming from different

Table 4. Meta-Analysis for Thrombolytic Drugs Data Under Random Effects Models

Statistical quantity	No w-factors	With w-factors: Full data	With w-factors: 22 and 23 deleted
DIC	95.8	92.6	83.4
$\bar{D}$	58.0	52.5	47.4
pD	37.8	40.1	36.0
$d_{SK,AtPA}$	$-.219_{(.126)}$	$-.195_{(.147)}$	$-.149_{(.109)}$
$d_{SK,tPA}$	$-.010_{(.088)}$	$-.040_{(.103)}$	$-.029_{(.084)}$
$d_{SK,SK+tPA}$	$-.056_{(.134)}$	$-.044_{(.147)}$	$-.028_{(.117)}$
$d_{SK,Ten}$	$-.212_{(.199)}$	$-.187_{(.221)}$	$-.141_{(.170)}$
$d_{SK,Ret}$	$-.167_{(.142)}$	$-.087_{(.169)}$	$-.093_{(.126)}$
$d_{SK,UK}$	$-.236_{(.232)}$	$-.313_{(.372)}$	$-.181_{(.279)}$
$d_{SK,ASPAC}$	$.040_{(.102)}$	$.009_{(.119)}$	$-.022_{(.103)}$
σ^2	.020	.023	.014
σ_w^2		.489	.141
$\Pr(\sigma_w^2 > \sigma^2)$		.932	.766
$W_{SK,AtPA,Ret}$		$-.122_{(.226)}$	$-.030_{(.144)}$
$W_{SK,AtPA,UK}$		$.183_{(.417)}$	$.019_{(.239)}$
$W_{SK,AtPA,ASPAC}$		$.678_{(.474)}$	
$W_{SK,tPA,UK}$		$.014_{(.369)}$	$-.005_{(.226)}$
$W_{SK,tPA,ASPAC}$		$-.190_{(.291)}$	$-.031_{(.211)}$

NOTE: Reported values are posterior means (standard deviations).

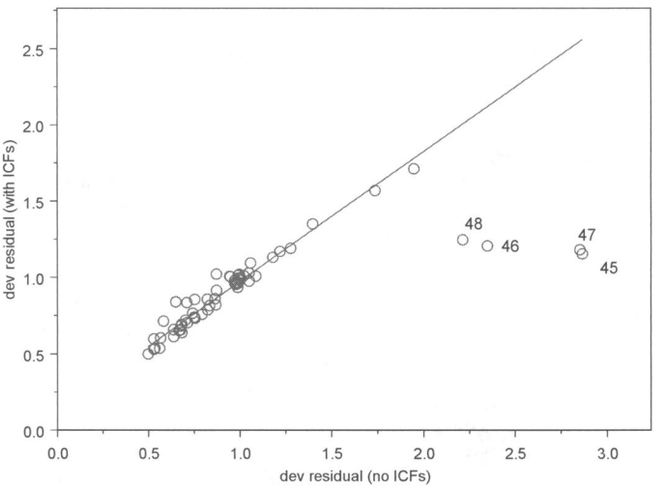

Figure 6. Thrombolytic Drugs Data: Residuals in Fitting Random-Effects Models With and Without Inconsistency Factors. Points 45 and 46 correspond to arms of trial 22; points 47 and 48 correspond to arms of trial 23.

sources. For example, if the evidence informing d_{AB} and d_{AC} were based on 10 very large trials each, but these were inconsistent with two medium-sized trials informing d_{BC}, then we would expect the deviance diagnostics to light up most strongly for d_{BC}, although clearly either of the other two elements could just as easily be the culprit. Only if each of the three data sources were of the same size would they all be equally "deviant."

Along with the number and size of the trials informing a particular edge, we should also consider how many evidence cycles are "corroborating" each edge. If it is not the *AtPA* versus *ASPAC* evidence that is deviant, it must instead be the *SK* versus *AtPA* or the *SK* versus *ASPAC* evidence. Table 3 shows that *SK* versus *AtPA* is supported by the very large trial 1 and corroborated by three further evidence cycles involving $SK + tPA$, *UK*, and *Ret*. Similarly, *SK* versus *ASPAC* is supported by the very large trial 2, as well as a corroborating cycle involving *tPA*. The sizes and number of trials supporting each edge contribute to our statistical confidence in the evidence on that edge, that is, to internal validity. But it is the number of corroborating cycles that gives us confidence in its external validity. Because both the other two edges in the inconsistency cycle are corroborated, we may tentatively conclude that the *AtPA* versus *ASPAC* evidence is indeed "deviant." The remarks in Section 4.5 show that model fit will not be substantially affected by parameterization, given minimally informative priors on nuisance parameters, $\{\mathbf{d}_b, \mathbf{w}\}$ and σ_w. Therefore, we can be reasonably confident that the consideration of deviance statistics and cross-validation would lead us to the same conclusions about the existence of inconsistency and its likely origin, given any parameterization. Reparameterization will define different evidence cycles, but the indirect evidence that corroborates each edge remains the same.

6. DISCUSSION

6.1 Relation to Previous Work on MTCs

Analysis of mixed treatment comparisons data remains relatively underresearched when one considers that these evidence structures are far more common than the standard pairwise meta-analysis. There is growing interest in this area, however, with empirical work comparing direct and indirect inferences through a third comparator (Bucher et al. 1997; Song et al. 2003), and applications in major medical journals (Pasty et al. 2003). Domenici et al. (1999) (see also Parmigiani 2002) provided an outstanding example of the potential for synthesis of trial evidence in a model that recognizes both within- and between-class treatment effects and draws together information on binary, ordinal, and continuous outcomes. The role of MTC synthesis is particularly clear in medical decision making where a choice must be made on which of several treatments to adopt in national treatment recommendations. In the United Kingdom, using MTC methods for these purposes is likely to become routine.

This article has concentrated on datasets with a single outcome, and a relatively homogeneous internal structure, in an attempt to precisely define the concept of inconsistency in MTC and then to come as close as possible to identifying its source—a task made difficult by the fact that consistency is a property of evidence cycles and that several cycles share common edges. Our consistency models consolidate a range of similar proposals for fixed-effects models (Hasselblad 1998; Ades 2003) and random-effects models (Higgins and Whitehead 1996; Hasselblad 1998; Lumley 2002; Lu and Ades 2004) for MTC within a single Bayesian framework. They all take the relative treatment effects as the only parameters of interest, with the trial-specific baseline effects relegated to the status of nuisance parameters. This distinguishes them from similar proposals in which the study-specific baselines are given hierarchical distributions (Lu and Ades 2004) or incorporated in multivariate distributions with treatment effect parameters (Berkey, Hoaglin, Antczak-Nouckoms, Mosteler, and Colditz 1998; Arends, Hoes, Lubsen, Grobbee, and Stijnen 2000; Van Houwelingen, Arends, and Stijnen 2002). The present proposals for random-effects models are, in principle, the same as those proposed by Higgins and Whitehead (1996) and Whitehead (2002), although our presentation is explicit about how these models extend to a wider range of MTC structures, including those where more than one "baseline" treatment must be specified.

The ability to statistically combine a large number of independent evidence sources to inform a smaller number of parameters raises the possibility of inconsistency between them. This article follows a series of recent attempts to examine evidence consistency in a variety of different evidence structures (Ades and Cliffe 2002; Ades 2003; Welton and Ades 2005), based on studies with completely unrelated designs, subjects, and outcomes. Because each inconsistency reflected a type of potential bias that was unique to individual evidence sources, an exchangeability assumption for ICFs would be epidemiologically implausible. In MTC structures, where the evidence is relatively homogeneous, exchangeability seems reasonable, and our proposals are in the same spirit as the work of Lumley (2002), whose model included treatment-by-comparator interaction terms $\xi_{jk} \sim \mathrm{N}(0, \omega^2)$, where ω^2 is termed "incoherence" and is equivalent to our σ_w^2 inconsistency variance. However, the two approaches would yield different measures of both inconsistency variance and between-trial heterogeneity, because Lumley uses an approximation to handle multi-arm trials, and the number of terms ξ_{jk} being estimated appears to be equal to the number of treatment comparisons, T, rather than the ICDF.

6.2 The Response to Inconsistency, and Potential Inconsistency

It is useful to separate detection and response to specific "extreme" inconsistencies in the evidence, as exemplified by $w_{SK.AtPA.ASPAC}$, from the interpretation of the overall measure of inconsistency represented by σ_w^2. Analyses of inconsistency factors, cross-validation, posterior contributions to deviance, or extreme residuals (Lu and Ades 2004) all offer means of identifying evidence cycles or edges that exhibit high levels of inconsistency. It may be also worth trying to successively remove edges that contribute most to deviance, a process analogous to that proposed by DuMouchel and Harris (1983) in their cross-species and cross-carcinogen syntheses of dose–response curves. They named their analog to σ_w^2 the "error of uncertain relevance," aptly expressing doubt as to whether the "indirect" evidence is relevant to the "direct" comparison of interest. However, the final decision as to which edge is the most likely "culprit" must rely on clinical and epidemiologic knowledge of the

trials involved in both the inconsistent cycle and the trials supplying "corroborating" evidence.

An often-expressed concern about a combination of "direct" and "indirect" evidence is that it may be "biased" because it breaks randomization (Bucher et al. 1997; Song et al. 2003; Yazdapanah et al. 2004). If each trial validly estimates a causal effect, the combined estimate—a weighted average—should also be valid. The issue is more one of generalizability. If the BC trials involve a different population to the AC and AB trials, then the combined estimates of d_{AB} and d_{AC} may not be valid for either population, even if all three sets of trials had, in fact, compared all three treatments. As with pairwise meta-analysis, the question for the investigator is whether inference based on the patient groups and protocols studied previously generalizes to the target population of current interest.

Assuming that there are no clearly "deviant" data items in the evidence base, one proposal (Lumley 2002) would be to report effect estimates from an MTC incorporating inconsistency. This could be seen as a natural way of down-weighting "indirect" evidence and incorporating our uncertainty about the size of the inconsistency variance. This seems particularly appealing when we wish to borrow strength from indirect data without assuming trials exchangeability of relative treatment effects across the entire evidence ensemble, or where we need a prediction for a treatment comparison that has never been made before. One difficulty with this approach is that for most datasets, the degrees of freedom for assessing inconsistency is small and measures of σ_w^2 will be highly uncertain, possibly resulting in unrealistically high degrees of uncertainty in predicted treatment effects. Further case studies will hopefully provide a basis for constructing reasonable informative priors.

[Received September 2004. Revised August 2005.]

REFERENCES

Ades, A. E. (2003), "A Chain of Evidence With Mixed Comparisons: Model for Multi-Parameter Synthesis and Consistency of Evidence," *Statistics in Medicine*, 22, 2995–3016.

Ades, A. E., and Cliffe, S. (2002), "Markov Chain Monte Carlo Estimation of a Multi-Parameter Decision Model: Consistency of Evidence and the Accurate Assessment of Uncertainty," *Medical Decision Making*, 22, 359–371.

Alderson, P., Green, S., and Higgins, J. P. T. (2003), *Cochrane Reviewers' Handbook 4.2.2.*, Chichester, U.K.: Wiley.

Arends, L. R., Hoes, A. W., Lubsen, J., Grobbee, D. E., and Stijnen, T. (2000), "Baseline Risk as Predictor of Treatment Benefit: Three Clinical Meta-Reanalysis," *Statistics in Medicine*, 19, 3479–3518.

Berkey, C. S., Hoaglin, D. C., Antczak-Nouckoms, A., Mosteler, F., and Colditz, G. A. (1998), "Meta-Analysis of Multiple Outcomes by Repressions With Random Effects," *Statistics in Medicine*, 17, 2537–2550.

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Berry, S. M. (1998), "Understanding and Testing for Heterogeneity Across 2×2 Tables: Application to Meta-Analysis," *Statistics in Medicine*, 17, 2353–2369.

Boland, A., Dundar, Y., Bagust, A., Haycox, A., Hill, R., Mujica Mota, R., Walley, T., and Dickson, R. (2003), "Early Thrombolysis for the Treatment of Acute Myocardial Infarction: A Systematic Review and Economic Evaluation," *Health Technology Assessment*, 7, 1–136.

Bridle, C., Bagnall, A., Duffy, S., Riemsma, R., Palmer, S., Darba, J., and Sculpher, M. J. (2004), *A Rapid and Systematic Review of the Clinical and Cost-Effectiveness of Newer Drugs for Treatment of Mania Associated With Bipolar Affective Disorder*, Health Technology Assessment, Vol. 8, No. 19.

Bucher, H., Guyatt, G., Griffith, L., and Walter, S. (1997), "The Results of Direct and Indirect Treatment Comparison in Meta-Analysis of Randomised Controlled Trials," *Journal of Clinical Epidemiology*, 50, 683–691.

Carlin, B. P., and Louis, T. A. (2001), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), New York: Chapman & Hall.

Congdon, P. (2003), *Applied Bayesian Modelling*, Chichester, U.K.: Wiley.

Domenici, F., Parmigiani, G., Wolpert, R. L., and Hasselblad, V. (1999), "Meta-Analysis of Migraine Headache Treatments: Combining Information From Heterogenous Designs," *Journal of the American Statistical Association*, 94, 16–28.

DuMouchel, W. H., and Harris, J. H. (1983), "Bayesian Methods for Combining the Results of Cancer Studies in Humans and Other Species," *Journal of the American Statistical Association*, 78, 293–307.

Eddy, D. M., Hasselblad, V., and Shachter, R. (1992), *Meta-Analysis by the Confidence Profile Method*, London: Academic Press.

Fenwick, E., Claxton, K., and Sculpher, M. J. (2001), "Representing Uncertainty: The Role of Cost-Effectiveness Acceptability Curves," *Health Economics*, 10, 779–787.

Fiore, M. C., Bailey, W. C., Choen, S. J., et al. (1996), *Smoking Cessation*, Clinical Practice Guideline No. 18 (AHCPR Publication no. 96–0692), Rockville, MD: Agency for Health Care Policy and Research, U.S. Department of Health and Human Services.

Gart, J. J., and Zweiful, J. R. (1967), "On the Bias of Various Estimators of the Logit and Its Variance, With Application to Quantal Bioassay," *Biometrika*, 54, 471–475.

Gelman, A., Carlin, J. G., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), London: Chapman & Hall/CRC.

Gelman, A., and Rubin, R. D. (1992), "Inferences From Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

Glasziou, P. P., and Sanders, S. L. (2002), "Investigating Causes of Heterogeneity in Systematic Reviews," *Statistics in Medicine*, 21, 1503–1511.

Gleser, L. J., and Olkin, I. (1994), "Stochastically Dependent Effect Sizes," in *Handbook of Research Synthesis*, eds. H. M. Cooper and L. V. Hedges, New York: Russell Sage Foundation, pp. 340–355.

Hasselblad, V. (1998), "Meta-Analysis of Multi-Treatment Studies," *Medical Decision Making*, 18, 37–43.

Hasselblad, V., and McCrory, D. C. (1995), "Meta-Analytic Tools for Medical Decision Making: A Practical Guide," *Medical Decision Making*, 15, 81–96.

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

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2002), "Statistical Heterogeneity in Systematic Reviews of Clinical Trials: A Critical Appraisal of Guidelines and Practice," *Journal of Health Services Research and Policy*, 7, 51–61.

Higgins, J. P. T., and Whitehead, A. (1996), "Borrowing Strength From External Trials in a Meta-Analysis," *Statistics in Medicine*, 15, 2733–2749.

Horn, R. A., and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge, U.K.: Cambridge University Press.

Larose, D. T., and Dey, D. K. (1997), "Grouped Random-Effects Models for Bayesian Meta-Analysis," *Statistics in Medicine*, 16, 1817–1829.

Lu, G., and Ades, A. E. (2004), "Combination of Direct and Indirect Evidence in Mixed Treatment Comparisons," *Statistics in Medicine*, 23, 3105–3124.

Lumley, T. (2002), "Network Meta-Analysis for Indirect Treatment Comparisons," *Statistics in Medicine*, 21, 2313–2324.

Marshall, E. C., and Spiegelhalter, D. J. (2001), "Institutional Performance," in *Multilevel Modelling of Health Statistics*, eds. A. H. Leyland and H. Goldstein, Chichester, U.K.: Wiley, pp. 127–142.

McAlistair, F. A., Laupacis, A., Wells, G. A., and Sackett, D. L. (1999), "User's Guides to Medical Literature, XIX: Applying Clinical Results, B: Guidelines for Determining Whether a Drug Is Exerting (More Than) a Class Effect," *Journal of the American Medicine Association*, 28, 1371–1377.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman & Hall.

Parmigiani, G. (2002), *Modeling in Medical Decision Making: A Bayesian Approach*, Chichester, U.K.: Wiley.

Pasty, B. M., Lumley, T., Furberg, C. D., Schellenbaum, G., Pahor, M., Alderman, M. H., and Weiss, N. S. (2003), "Health Outcomes Associated With Various Antihypertensive Therapies Used as First-Line Agents," *Journal of the American Medical Association*, 289, 2534–2544.

Prevost, T. C., Abrams, K. R., and Jones, D. R. (2000), "Hierarchical Models in Generalized Synthesis of Evidence: An Example Based on Studies of Breast Cancer Screening," *Statistics in Medicine*, 19, 3359–3376.

Scheffe, H. (1959), *The Analysis of Variance*, New York: Wiley.

Shachter, R. D., Eddy, D. M., and Hasselblad, V. (1990), "An Influence Diagram Approach to Medical Technology Assessment," in *Influence Diagrams, Belief Nets and Decision Analysis*, eds. R. M. Oliver and J. Q. Smith, New York: Wiley, pp. 321–350.

Smith, T. C., Spiegelhalter, D. J., and Thomas, S. L. (1995), "Bayesian Approaches to Random-Effects Meta-Analysis: A Comparative Study," *Statistics in Medicine*, 14, 2685–2699.

Song, F., Altman, D. G., Glenny, A. M., and Deeks, J. J. (2003), "Validity of Indirect Comparison for Estimating Efficacy of Competing Interventions: Empirical Evidence From Published Meta-Analyses," *British Medical Journal*, 326, 472–476.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Chichester, U.K.: Wiley.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society*, Ser. B, 64, 1–34.

Spiegelhalter, D. J., Thomas, A., Best, N., and Lunn, D. (2001), *WinBUGS User Manual: Version 1.4*, Cambridge, U.K.: MRC Biostatistics Unit.

Sutton, A., Abrams, K. R., Jones, D. K., Sheldon, T. A., and Song, F. (1998), "Systematic Reviews of Trials and Other Studies," *Health Technology Assessment*, Vol. 2, No. 19, NHS R&D HTA Programme, U.K.

Thompson, S. G., Smith, T. C., and Sharp, S. J. (1997), "Investigating Underlying Risk as a Source of Heterogeneity in Meta-Analysis," *Statistics in Medicine*, 16, 2741–2758.

Van Houwelingen, H. C., Arends, L. R., and Stijnen, T. (2002), "Advanced Methods in Meta-Analysis: Multivariate Approach and Meta-Regression," *Statistics in Medicine*, 21, 589–624.

Warn, D. E., Thompson, S. G., and Spiegelhalter, D. J. (2002), "Bayesian Random Effects Meta-Analysis of Trials With Binary Outcomes: Method for Absolute Risk Difference and Relative Risk Scales," *Statistics in Medicine*, 21, 1601–1623.

Welton, N. J., and Ades, A. E. (2005), "A Toxoplasmosis Incidence in the U.K.: Evidence Synthesis and Consistency of Evidence," *Applied Statistics*, 54, 385–404.

Whitehead, A. (2002), *Meta-Analysis of Controlled Clinical Trials*, Chichester, U.K.: Wiley.

Woolf, B. (1955), "On Estimating the Relation Between Blood Group and Disease," *Annals of Human Genetics*, 19, 251–253.

Yazdanpanah, Y., Sissoko, D., Egger, M., Mouton, Y., Zwahlen, M., and Chene, G. (2004), "Clinical Efficacy of Antiretroviral Combination Therapy Based on Protease Inhibitors or Non-Nucleoside Analogue Reverse Transcriptase Inhibitors: Indirect Comparison of Controlled Trials," *British Medical Journal*, 328, 249–253.