Statistical Science
2013, Vol. 28, No. 3, 376–397
DOI: 10.1214/13-STS426

Conflict Diagnostics in Directed Acyclic Graphs, with Applications in Bayesian Evidence Synthesis

Anne M. Presanis, David Ohlssen, David J. Spiegelhalter and Daniela De Angelis

Abstract. Complex stochastic models represented by directed acyclic graphs (DAGs) are increasingly employed to synthesise multiple, imperfect and disparate sources of evidence, to estimate quantities that are difficult to measure directly. The various data sources are dependent on shared parameters and hence have the potential to conflict with each other, as well as with the model. In a Bayesian framework, the model consists of three components: the prior distribution, the assumed form of the likelihood and structural assumptions. Any of these components may be incompatible with the observed data. The detection and quantification of such conflict and of data sources that are inconsistent with each other is therefore a crucial component of the model criticism process. We first review Bayesian model criticism, with a focus on conflict detection, before describing a general diagnostic for detecting and quantifying conflict between the evidence in different partitions of a DAG. The diagnostic is a p-value based on splitting the information contributing to inference about a "separator" node or group of nodes into two independent groups and testing whether the two groups result in the same inference about the separator node(s). We illustrate the method with three comprehensive examples: an evidence synthesis to estimate HIV prevalence; an evidence synthesis to estimate influenza case-severity; and a hierarchical growth model for rat weights.

Key words and phrases: Conflict, directed acyclic graph, evidence synthesis, graphical model, model criticism.

1. INTRODUCTION

Bayesian evidence synthesis methods combining multiple, imperfect and disparate sources of data to estimate quantities that are challenging to measure directly are becoming widespread (e.g., Spiegelhalter, Abrams and Myles, 2004; Ades and Sutton, 2006). Although little data may be available from which to directly estimate such quantities, there may be plenty of *indirect* information on related parameters that can be expressed as functions of the key parameters of interest. The synthesis of both direct and indirect data usually entails the formulation of complex probabilistic models, where the dependency of the data on the parameters is represented by a directed acyclic graph (DAG). Recent examples can be found in the fields of ecology (Clark et al., 2010), biochemical kinet-

Anne M. Presanis is Senior Investigator Statistician, MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 0SR, United Kingdom (e-mail: anne.presanis@mrc-bsu.cam.ac.uk). David Ohlssen is Senior Expert Methodologist, IIS Statistical Methodology, Novartis Pharmaceuticals Corporation, One Health Plaza, Building 419, room 2157, East Hanover, New Jersey, 07936-1080, USA. David Spiegelhalter is Winton Professor for the Public Understanding of Risk, Statistical Laboratory, University of Cambridge Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, United Kingdom. Daniela De Angelis is Programme Leader, MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 0SR, United Kingdom.

ics (Henderson, Boys and Wilkinson, 2010), environmental epidemiology (Jackson, Richardson and Best, 2008), health technology assessment (Welton et al., 2012), mixed treatment comparisons (Lu and Ades, 2006) and infectious disease epidemiology (Birrell et al., 2011).

With modern software it has become reasonably straightforward to draw inferences and make probabilistic predictions from such complex models. However, with complexity also comes a vital requirement that the conclusions of the model can be explained and justified, both for the "owners" of the model and any audience they wish to convince. There are two main issues: First, to identify the essential *drivers* of inference, assessing sensitivity to data or assumptions. Second, to judge whether the data are *consistent* with each other or with model assumptions. This assessment is crucial in syntheses of multiple sources of evidence, where these sources are dependent on shared parameters of interest and hence have the potential to conflict with each other (Lu and Ades, 2006; Presanis et al., 2008). The evidence arising from (i) prior distributions, (ii) the assumed form of the likelihood and (iii) other structural/functional model assumptions also has the potential to conflict with the different sources of data or with each other. The existence of such inconsistency and/or sensitivity to model assumptions would naturally lead to careful reexamination of the model and data sources, and a further iteration of the inference and model-criticism cycle recommended by Box (1980). O'Hagan (2003) reviews the connection between model checking and conflict detection in the context of complex stochastic systems.

This paper focusses on the issue of detecting and measuring conflict, particularly on diagnostics that are effective in the type of complex DAG-based models that evidence synthesis requires for substantive problems. Bayesian predictive p-values, in various guises, have been widely employed as a natural measure to assess the consistency of the components driving inference [e.g., Box, 1980; Gelman, Meng and Stern, 1996; Bayarri and Castellanos, 2007]. Marshall and Spiegelhalter (2007) proposed the idea of "node-splitting" to compare prior and likelihood in a hierarchical DAG-based model. A general framework that unifies these various approaches has been proposed (Dahl, Gåsemyr and Natvig, 2007; Gåsemyr and Natvig, 2009), but exploration of how to apply these ideas in real-world complex problems remains limited (Dias et al., 2010; Scheel, Green and Rougier, 2011). We review in Section 2 the literature on Bayesian model checking, with a focus on conflict detection. In Section 3 we describe a generalisation of Marshall and Spiegelhalter (2007) to any node in a DAG, with the aim of demonstrating how, in practice, such a diagnostic may be usefully employed to detect and measure different types of inconsistency in substantive problems. We give recommendations for strategies to construct the diagnostic in different contexts and to treat nuisance parameters. In Section 4 we then consider three detailed examples: an evidence synthesis to estimate HIV prevalence from many indirect data sources (4.1); an evidence synthesis to estimate influenza severity (4.2); and a bivariate normal random effects model for rat weights, illustrating multivariate conflict assessment (4.3). We end with a discussion in Section 5.

2. BACKGROUND

2.1 DAGs

It is now a standard procedure to use directed acyclic graphs to represent the qualitative conditional independence assumptions of a complex stochastic model; see, for example, Lauritzen (1996) for a full description. The crucial idea is that each node in the graph represents a stochastic quantity, related to other nodes through a series of "parent-child" relationships to form a DAG. Using an intuitive language in terms of familial relationships, the basic assumption represented is that any node is conditionally independent of its nondescendants given its parents. "Founder" nodes, that is, nodes with no parents, are assigned a prior distribution. Given the directed graphical structure, the joint distribution over all nodes is given by the product of all the conditional distributions of each child given its direct parents. Inference on DAGs is conducted when some nodes are observed as data and the resulting joint and marginal posterior distributions of the remaining nodes are needed. Substantial research has led to efficient exact and simulation-based algorithms implemented in various software (see Cowell et al., 1999, for a review), such as Markov chain Monte Carlo [MCMC Gamerman and Lopes, 2006].

Figure 1(a) shows a simple example of a DAG representing a model H for data $\mathbf{y}$ with parameters $\boldsymbol{\theta}$. The double circles represent the founder node assigned a prior $p(\boldsymbol{\theta}|H)$; the square represents observations $\mathbf{y}$, a child node of $\boldsymbol{\theta}$; and the solid arrow represents a distributional assumption $\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\theta}, H)$. Figure 1(b) shows a slightly more complex hierarchical model, where within units $i \in 1:k$, the n_i observations y_{ij}, $j \in$

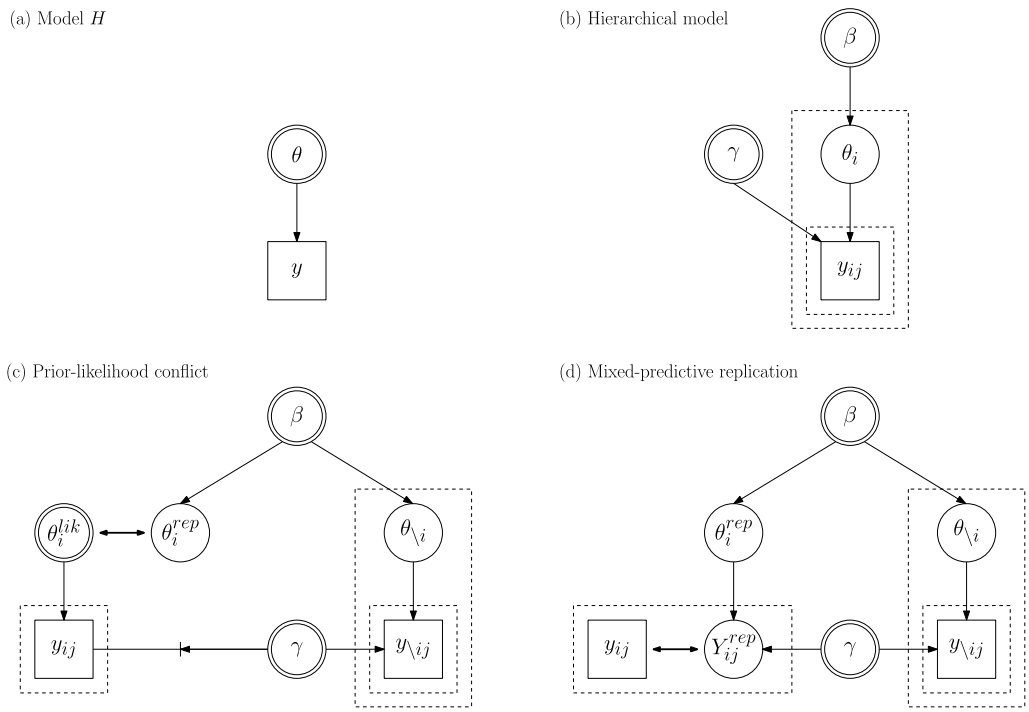

FIG. 1. *Examples of DAGS showing* (a) *a simple model* H; (b) *a hierarchical model*; (c) *prior-likelihood comparison via a node-split in model* (b); *and* (d) *cross-validatory mixed-predictive replication in model* (b).

$1:n_i$ are assumed drawn from a distribution with unit-specific parameters θ_i and global parameters γ (e.g., variances). At the next level of hierarchy, the θ_i are drawn from distributions with hyper-parameters β:

$$\begin{aligned} y_{ij} &\sim p(y_{ij}|\theta_i,\gamma), \quad i \in 1:k,\ j \in 1:n_i, \\ \theta_i &\sim p(\theta_i|\beta), \\ \beta,\gamma &\sim p(\beta,\gamma). \end{aligned}$$

Repetition over and within units is represented by the dashed rectangles, and the hyperparameters β and γ are the founder nodes. Continuing the analogy of a family, γ and θ_i are co-parents of the data $\mathbf{y}_i$ and within groups i, the data $\mathbf{y}_i$ are siblings.

2.2 Node-Splitting

Marshall and Spiegelhalter (2007) propose separating out the contributions of prior and likelihood to a unit-level parameter θ_i in a hierarchical model such as Figure 1(b), to compare their consistency. They do so by splitting the DAG into two independent partitions [Figure 1(c)]. The partition representing the likelihood of the ith unit's data is formed by drawing a replicate $\theta_i^{\text{lik}}|\mathbf{y}_i$ from a uniform reference prior, updated with the observations $\mathbf{y}_i$, that is, from the posterior distribution generated by $\mathbf{y}_i$ alone. This distribution is in effect the "data-translated" likelihood (Box and Tiao, 1992). For the partition representing the prior contribution to θ_i, a replicate $\theta_i^{\text{rep}}|\mathbf{y}_{\setminus i}$ is drawn from the "predictive-prior" $p(\theta_i|\mathbf{y}_{\setminus i})$ where $\mathbf{y}_{\setminus i}$ denotes the remaining data aside from unit i. The authors propose comparison of the posterior distributions [represented by the double-headed arrow in Figure 1(c)] of the two replicates θ_i^{lik} and θ_i^{rep} by defining a difference function, $\delta_i = \theta_i^{\text{rep}} - \theta_i^{\text{lik}}$, then calculating the "conflict p-value"

$$c_{MS1,i} = \Pr(\delta_i \le 0|\mathbf{y})$$

in the case of a one-sided hypothesis test of departures towards smaller θ_i than suggested by $\mathbf{y}_i$, or

$$c_{MS2,i} = 2\min\bigl(\Pr(\delta_i \le 0|\mathbf{y}), 1 - \Pr(\delta_i \le 0|\mathbf{y})\bigr)$$

if testing the null hypothesis $\delta_i = 0$ against $\delta_i \neq 0$. In the first case, values of $c_{MS1,i}$ close to either 0 or 1 indicate conflict. In the more general second case, a small value represents a high level of conflict, so the "conflict p-value" is possibly a misnomer: $c_{MS2,i}$ actually measures *consistency*. However, since the term has already been introduced in the literature, we continue to refer throughout to the conflict p-value.

Finally, note also that this example is one where in order to completely split the DAG into two independent partitions would require splitting a vector of

nodes, $\{\theta_i, \gamma\}$. However, if, for example, γ is a variance parameter in a normal hierarchical model, we may not actually be directly interested in examining conflict around γ, particularly if it is not strongly identifiable from a unit i alone. Marshall and Spiegelhalter (2007) therefore propose treating such parameters as nuisance parameters, and "cutting" feedback from unit i to γ to prevent the data $\mathbf{y}_i$ from influencing γ [e.g., using the "cut" function in the OpenBUGS software (Lunn et al., 2009)]. A cut in a DAG stops information flow in one direction, as opposed to a node-split which prevents information flow in both directions. This cut is represented by the diode shape between y_{ij} and γ in Figure 1(c). The two replicates θ_i^{rep} and θ_i^{lik} are therefore not entirely independent, since γ may still influence θ_i^{lik}; however, the authors believe any such influence will be small.

2.3 Bayesian Predictive Diagnostics

Bayesian predictive diagnostics to assess consistency are based on comparison of a discrepancy statistic with a reference distribution. The general setup is that of Figure 1(a), assuming a model H for $\mathbf{y}$ with parameters $\boldsymbol{\theta}$:

$$\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\theta}, H),$$
$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|H).$$

To assess whether the observed data could have been generated by the assumed model H, we compare the observed data $\mathbf{y}$, via a test statistic $T(\mathbf{y})$, to a reference distribution $p_T\{T(\mathbf{Y}^{\text{rep}})|H\}$ of the test statistic for hypothetical (replicated) data $\mathbf{Y}^{\text{rep}}$ under the assumed (null) model H. The p-value defined by where the observed value of the test statistic is located in the reference distribution measures the compatibility of the model with the data. The reference distribution depends on the way we assume the replicated data are generated from the null model and therefore on the exact components of inference we wish to compare. Various proposals have been made, broadly categorised into prior-, posterior- and mixed-predictive approaches. Note that to be able to interpret p-values in a meaningful way, the distribution of the p-values under the null prior model H is required.

The prior-predictive distribution (Box, 1980), in which the parameters are integrated out with respect to the prior, is a natural choice for the reference, as the p-values $\Pr\{T(\mathbf{Y}^{\text{rep}}) \geq T(\mathbf{y})|H\}$ are uniformly distributed under the null prior model H. The approach assesses prior-data conflict, most usefully in the case of informative priors. However, in the case of improper priors, prior-predictive replication is not defined. In practice, many analysts use very diffuse but proper priors to express noninformativeness, in which case prior-data comparison is not particularly useful, since almost any data will be plausible under such priors. Other related approaches to assessing prior-data conflict include: the adaptation by Evans and Moshonov (2006, 2007) of Box (1980) to minimally sufficient statistics; the use of logarithmic scoring rules (Dawid, 1984; Spiegelhalter et al. 1993, 1994) to assess conflict; and, more recently, the use of ratios of prior-to-posterior distances under different priors, using Kullback-Leibler measures (Bousquet, 2008).

The posterior-predictive distribution [e.g., Rubin, 1984; Gelman, Meng and Stern, 1996] was proposed as an alternative to the prior-predictive distribution for use when improper priors are employed. It results from integrating the parameters out with respect to the posterior rather than prior distribution, thereby assessing model-data rather than prior-data compatibility. Posterior-predictive checks have become widespread (Gelman et al., 2003). However, the p-values may not be uniformly distributed, since the data are used twice, in obtaining both the posterior and the posterior-predictive distribution (Bayarri and Berger 1999, 2000; Robins, van der Vaart and Ventura, 2000). Suggestions have therefore been made to avoid the conservatism of posterior-predictive p-values, including alternative p-values that are closer to uniform, but often difficult to compute, such as the conditional and partial posterior-predictive p-values (Bayarri and Berger, 1999, 2000; Robins, van der Vaart and Ventura, 2000; Bayarri and Castellanos, 2007). The predictive distributions in these approaches are defined by integrating out the unknown parameters $\boldsymbol{\theta}$ with respect to posterior distributions that are, respectively, (i) conditional on a sufficient statistic for $\boldsymbol{\theta}$; and (ii) constructed from the prior and from a likelihood defined to be conditional on the observed value of a test statistic, $T(\mathbf{y})$, so that the information in $T(\mathbf{y})$ contributing to the posterior is "removed" before integrating out $\boldsymbol{\theta}$. Alternative "double simulation" approaches, post-processing posterior-predictive p-values such that their distribution is uniform (Hjort, Dahl and Steinbakk, 2006; Johnson, 2007; Steinbakk and Storvik, 2009), require proper priors and are computationally demanding.

The mixed-predictive distribution (Gelman, Meng and Stern, 1996; Marshall and Spiegelhalter, 2007), in the context of hierarchical models such as Figure 1(b) with hyperparameters β, integrates out the parameters

$\boldsymbol{\theta} = \{\theta_i, i \in 1:k\}$ with respect to what Marshall and Spiegelhalter (2007) term the "predictive-prior" distribution, namely, $p^M(\boldsymbol{\theta}|\mathbf{y}, H) = \int p(\boldsymbol{\theta}|\beta, H) \cdot p(\beta|\mathbf{y}, H)\, d\beta$. This distribution is not the marginal posterior distribution of $\boldsymbol{\theta}$, but the distribution obtained by drawing replicates $\boldsymbol{\theta}^{\text{rep}}$ from the marginal posterior distribution of β. The parameters $\boldsymbol{\theta}$ are then integrated out, resulting in a mixed-predictive p-value that is still conservative, but less so than the posterior-predictive p-value.

Posterior- and mixed-predictive approaches are often carried out in a cross-validatory framework to avoid the double use of the data. Marshall and Spiegelhalter (2007) showed that under certain conditions, their conflict p-value is equivalent to the cross-validatory mixed-predictive p-value when this exists [Figure 1(c,d)]. In the mixed-predictive approach, compatibility between the observations $\mathbf{y}$ and the predictive distribution arising from both likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta}|\beta, H)$ is measured. In the conflict approach, the prior is compared with the posterior arising from the likelihood alone. Although the tests are mathematically the same under the conditions described by Marshall and Spiegelhalter (2007), the mixed-predictive approach tests model/data compatibility, whereas the conflict approach tests prior/likelihood compatibility. Other cross-validatory approaches for model checking include the observed relative surprise (Evans, 1997) and a Bayesian influence statistic (Jackson, White and Carpenter, 2012).

2.4 Node-Level Conflict Measures

The predictive diagnostics of the previous section in general assess a particular aspect of the whole model in comparison to the data. The node-splitting idea of Marshall and Spiegelhalter (2007) is, by contrast, a method of assessing prior-likelihood conflict locally at a particular node in the DAG. Other node-based diagnostics have been proposed also, including that of O'Hagan (2003). He proposed contrasting two sets of information informing a single node θ, where each set is summarised by a unimodal density or likelihood with parameters $\boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}_b$ respectively. The author proposes first normalising both densities/likelihoods to have unit maximum height, then considering the height $z = p_a(x_z|\boldsymbol{\lambda}_a) = p_b(x_z|\boldsymbol{\lambda}_b)$ at the point of intersection x_z of the two curves between the two modes. Taking $c_{\text{OH}} = -2\log(z)$ as the measure of conflict, this will be high if the two densities/likelihoods have little overlap. Bayarri and Castellanos (2007) extend the partial posterior predictive approach of Bayarri and Berger (1999, 2000) to hierarchical models and, in doing so, compare their method to several others, including O'Hagan's (2003) conflict measure and Marshall and Spiegelhalter's (2007) conflict p-value. They conclude that only their method and the conflict p-value consistently detect conflict, noting that O'Hagan's measure may be sensitive to the prior used and conservative, due to double use of the data.

Dahl, Gåsemyr and Natvig (2007) raise the same objection as Bayarri and Castellanos (2007) to the double use of data in O'Hagan's measure, and therefore propose a variation on this measure, for the simple normal hierarchical analysis of variance. They propose both a data-splitting approach and a conflict measure at an internal node θ of the DAG of the model, based on means and variances of cross-validatory "integrated posterior distributions" (IPDs). These IPDs are constructed by taking the information contribution to θ from one partition λ_a of the DAG (either a prior or likelihood), normalizing it to a density, and integrating it with respect to the posterior distribution of the parameters λ_b in the other partition (analogous to posterior- or mixed-predictive distributions for θ, see next section for details). This is in contrast to O'Hagan, who normalizes to unit height instead of to a density. The authors derive the distributions of their conflict measure for various data splittings under fixed known variances. They perform an extensive simulation study for the case of unknown variances assigned prior distributions, comparing their approach to O'Hagan's for various data splittings and prior distributions.

2.5 Unifying Approaches to Node-Level Conflict Diagnostics

The models of Figure 1 can be seen to be special cases of the generic DAG in Figure 2(a). The figure shows a generic model H, with an internal "separator" node or set of nodes $\boldsymbol{\theta}$. The evidence informing $\boldsymbol{\theta}$ is partitioned into two groups, $\boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}_b$. The set $\boldsymbol{\lambda}_a$ contains $\boldsymbol{\theta}$'s parents $\mathbf{pa}$ and a subset of child nodes $\mathbf{ch}_a$, with corresponding co-parents $\mathbf{cp}_a$ and siblings $\mathbf{si}$. The partition $\boldsymbol{\lambda}_b$ contains the remaining child nodes $\mathbf{ch}_b$ and corresponding co-parents $\mathbf{cp}_b$. The observed data $\mathbf{y}$ are split into $\mathbf{y}_a$ and $\mathbf{y}_b$, that are contained within the respective vectors of child and/or sibling nodes $\{\mathbf{ch}_a, \mathbf{si}\}$ and $\mathbf{ch}_b$. Each of the predictive diagnostics of Section 2.3 are special cases of Figure 2(a). Figure 2(b) shows the same generic model as (a), but with the node(s) $\boldsymbol{\theta}$ split into two copies, so that the evidence that the two partitions $\boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}_b$ provide about $\boldsymbol{\theta}$ may be compared. This figure represents a generalisation of

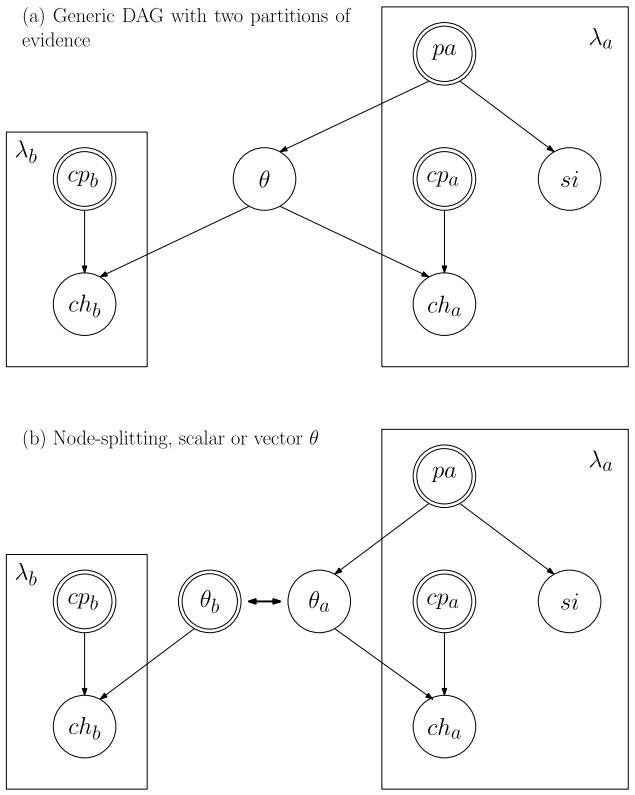

FIG. 2. *DAG* (a) *shows a generic model* H, *with the evidence informing an internal "separator" node or set of nodes* $\boldsymbol{\theta}$ *partitioned into two groups,* $\boldsymbol{\lambda}_a$ *and* $\boldsymbol{\lambda}_b$, *as in Gåsemyr and Natvig* (2009). *DAG* (b) *shows the same generic model as* (a), *but with the node(s)* $\boldsymbol{\theta}$ *split into two copies, so that the evidence that the two partitions* $\boldsymbol{\lambda}_a$ *and* $\boldsymbol{\lambda}_b$ *provide about* $\boldsymbol{\theta}$ *may be compared.*

Marshall and Spiegelhalter's (2007) node-splitting approach to any internal node(s) in a DAG.

The generic setup of Figure 2(a) is described by Gåsemyr and Natvig (2009), who generalise their earlier work in Dahl, Gåsemyr and Natvig (2007). They use the same cross-validatory IPDs as reference distributions, but consider p-values based on tail areas of the distributions, rather than a conflict measure based on means and variances. The authors consider first conflict at a data node, in which case the reference IPD is a posterior- or mixed-predictive distribution of one partition of data conditional on the remaining data, that is, in the cross-validatory setting of Figure 2(a). Gåsemyr and Natvig (2009) show that: for symmetric, unimodal IPDs, their tail area conflict measure is equivalent to the measure based on means and variances of Dahl, Gåsemyr and Natvig (2007); if the data are normally distributed, their measure is equivalent to the cross-validatory mixed-predictive p-value of Marshall and Spiegelhalter (2007); and if the IPDs are symmetric, unimodal and the data in the two partitions are conditionally independent, their measure is equivalent to the partial posterior predictive p-value of Bayarri and Berger (2000).

Gåsemyr and Natvig (2009) next consider conflict between two partitions $\boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}_b$ of a DAG at an internal node θ which is scalar. The IPDs are then predictive distributions for θ, conditional on the data in each partition:

$$p_a(\theta|\mathbf{y}_a) = \int f(\theta;\boldsymbol{\lambda}_a)\, p(\boldsymbol{\lambda}_a|\mathbf{y}_a)\, d\boldsymbol{\lambda}_a,$$

$$p_b(\theta|\mathbf{y}_b) = \int f(\theta;\boldsymbol{\lambda}_b)\, p(\boldsymbol{\lambda}_b|\mathbf{y}_b)\, d\boldsymbol{\lambda}_b,$$

where $f(\theta;\boldsymbol{\lambda}_a)$ and $f(\theta;\boldsymbol{\lambda}_b)$ are densities proportional to the likelihood factors informing θ in each of the two partitions, expressed as functions of θ. As in Dahl, Gåsemyr and Natvig (2007), Gåsemyr and Natvig (2009) propose normalising the likelihood terms to densities of θ conditional on the nodes in the partition, $\boldsymbol{\lambda}_a$ or $\boldsymbol{\lambda}_b$. The authors take a pair of independent samples (θ_a^*, θ_b^*) from the two predictive distributions $p_a(\theta|\mathbf{y}_a)$ and $p_b(\theta|\mathbf{y}_b)$, and define $\delta = \theta_a^* - \theta_b^*$. Their proposed conflict measures are tail area probabilities:

$$c_{GN3} = 1 - 2\min\{\Pr(\delta \le 0), 1 - \Pr(\delta \le 0)\},$$

$$c_{GN4} = \Pr\{p_\delta(\delta|\mathbf{y}_a,\mathbf{y}_b) \ge p_\delta(0|\mathbf{y}_a,\mathbf{y}_b)\},$$

where p_δ is the posterior density of δ. Gåsemyr and Natvig (2009) demonstrate that the data-data conflict tail areas they first considered are special cases of c_{GN3} and c_{GN4}. More generally, they also show that if the cumulative distribution functions of θ (corresponding to the predictive densities p_a and p_b, resp.) are normal, both c_{GN3} and c_{GN4} are uniform pre-experimentally and are equivalent to each other. The authors also extend their results to the general normal linear model when the covariance matrix is fixed and known. Note that c_{GN3} should be straightforward to obtain via simulation, using MCMC, for example, whereas c_{GN4} is much more computationally demanding, requiring, for example, a kernel estimate of p_δ. Gåsemyr and Natvig (2009) note that c_{GN3} is closely related to the conflict p-value of Marshall and Spiegelhalter (2007): by taking the function $f(\theta;\boldsymbol{\lambda}_b)$ to be proportional to the likelihood of the data $\mathbf{y}_b$, Gåsemyr and Natvig (2009) are implicitly assuming a uniform reference prior for θ, whereas Marshall and Spiegelhalter (2007) explicitly do so for the copy θ_b [Figure 2(b) versus (a)]. Finally, Gåsemyr and Natvig (2009) extend their framework to

multivariate node-splits $\boldsymbol{\theta}$, although their theoretical results are restricted to cases where the two predictive distributions are multivariate normal, and to general normal linear models with known covariances.

In a slightly different approach, that complements and is related to the conflict measures summarised here, Scheel, Green and Rougier (2011) propose a diagnostic plot to visualise conflict at any particular node θ in a DAG. The authors define a "local prior" for θ conditional on its parents and a "lifted likelihood" coming from θ's children, conditional on both θ and the co-parents of θ's children. The "local critique plot" then examines where the marginal posterior of θ lies, relative to both the local prior and the lifted likelihood.

3. EXTENDING THE CONFLICT p-VALUE TO ANY NODE

While the conflict measures summarised in the previous section are useful tools, the general framework introduced by Dahl, Gåsemyr and Natvig (2007) and Gåsemyr and Natvig (2009) uses idealised examples, such as normal models or general normal linear models with fixed covariances, to demonstrate uniformity of p-values. Furthermore, many of the other measures, such as post-processed p-values, are computationally expensive. In the context of complex evidence syntheses, a diagnostic for conflict is required that is both easy to compute and applicable in the more complex probabilistic models typical of evidence synthesis. We emphasise that in presenting a generalisation of the conflict p-value of Marshall and Spiegelhalter (2007) to any "separator" node(s) $\boldsymbol{\theta}$ in a DAG [Figure 2(b)], we are not aiming to prove uniformity of p-values in specific cases. Instead, we aim to present a framework for conflict detection in *practice*, demonstrating the utility of such methods in substantive, realistic examples.

In the context of Figure 2(b), the evidence informing $\boldsymbol{\theta}$ is comprised of the information coming from each of $\boldsymbol{\theta}$'s neighbours in the DAG. This information is generally in the form of either: (i) a (potentially predictive) prior distribution from $\boldsymbol{\theta}$'s parents and siblings, that may include likelihood terms from **si**; or (ii) likelihood contributions from $\boldsymbol{\theta}$'s children combined with priors for the co-parents. Note that Figure 2(b), although general, could be generalised even further if, for example, the co-parents have other children that are not children of $\boldsymbol{\theta}$, but also contain likelihood terms, indirectly informing $\boldsymbol{\theta}$.

The evidence surrounding $\boldsymbol{\theta}$ is split into two independent groups, and we compare the inferences about $\boldsymbol{\theta}$ resulting from each of the two partitions by comparing the two (independent) posterior distributions $p(\boldsymbol{\theta}_a|\mathbf{y}_a)$ and $p(\boldsymbol{\theta}_b|\mathbf{y}_b)$. We assess our null hypothesis that $\boldsymbol{\theta}_a = \boldsymbol{\theta}_b$. Our measure of conflict can be designed to reflect differences between data and a model, between different data sources, or between a prior and a likelihood. Simple examples of different types of node-splits are given in Figure 3, each of which is a special case of Figure 2(b). The examples include comparison between a likelihood and the remaining combined prior and likelihood, appropriate when questioning part of the data contributing to a node [Figure 3(b)] and a likelihood vs likelihood comparison [Figure 3(c)], directly

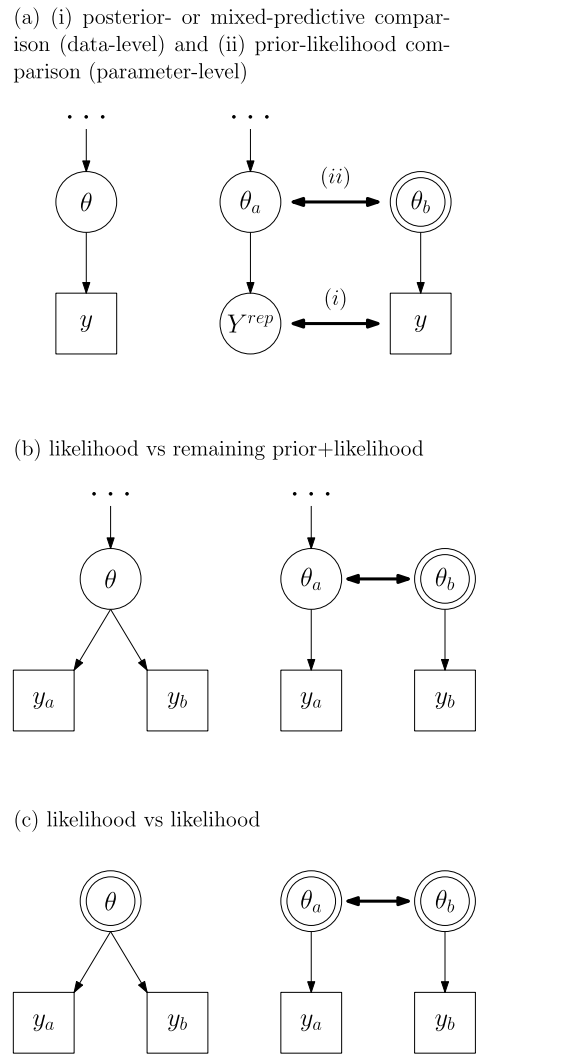

FIG. 3. *Examples of node-splits contrasting* (a) (i) *data and predictive distribution or* (ii) *prior and likelihood*; (b) *a likelihood term and the remaining combined prior and likelihood*; *and* (c) *two likelihood terms*, *assuming a reference prior for each* θ. *In each example*, *the original model is on the left and the node-split is shown on the right*.

contrasting sources of evidence without taking into account a prior.

3.1 Choice of a Reference Prior

The question of how to create the two posterior distributions to compare depends on the precise comparison we wish to make, and hence the specific node-splits we construct. In particular, the creation of posterior distributions depends partly on the choice of a reference prior to use for $\boldsymbol{\theta}_b$ or, indeed, for both $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$ in the case of node-splits of a form such as Figure 3(c). The aim is to choose a prior that turns what is effectively a likelihood term $[p(\mathbf{y}_b|\mathbf{cp}_b, \boldsymbol{\theta}_b)$, where $\mathbf{cp}_b$ is a vector of $\boldsymbol{\theta}_b$'s co-parents] into a posterior distribution, without the prior itself influencing the posterior. We therefore follow Box and Tiao (1992) and Kass (1990) in adopting uniform priors for a transformation $h(\boldsymbol{\theta}_b)$ of $\boldsymbol{\theta}_b$ to a scale such that the uniform prior is appropriate. The posterior distribution is then effectively the "data-translated likelihood" (Box and Tiao, 1992). As noted previously, Marshall and Spiegelhalter (2007) showed that under certain conditions, choosing a uniform prior results in the conflict p-value and the cross-validatory mixed-predictive p-value coinciding when the latter exists. They note also that in other situations, Box and Tiao (1992) showed that the Jeffreys' priors widely used as "noninformative" priors are equivalent to a uniform prior on an appropriate transformation of $\boldsymbol{\theta}_b$. We therefore recommend use of a Jeffreys' prior in general, although note that for some node-splits, choice of a "noninformative" prior may not be so straightforward. For some comparisons, we may not be able to assign a Jeffreys' prior or there may not be a natural choice of reference prior. We then rely on the likelihood in a specific partition dominating any prior chosen for founder nodes in that partition, though this assumption should be assessed, for example, through sensitivity analyses to the prior (see Section 5 for further discussion of this).

3.2 Comparison of Two Posterior Distributions

Considering for now scalar θ, to test the point-null hypothesis that $\theta_a = \theta_b$, Bayes factors could be employed. However, the approach is known to be difficult because of the high dependence on the precise form of the reference priors, as well as hard to evaluate using MCMC. Instead, we prefer to consider the plausibility of the hypothesis either directly if θ takes discrete values or using a p-value if the support of θ is continuous.

3.2.1 *Conflict at a discrete node.* If θ takes values in $0, 1, \ldots, K$, then we can directly evaluate

$$c = p(\theta_a = \theta_b|\mathbf{y}_a, \mathbf{y}_b) = \sum_{k=0}^{K} p(\theta_a = k|\mathbf{y}_a)\, p(\theta_b = k|\mathbf{y}_b).$$

As a simple example, consider a disease with known prevalence π and denote the event of having the disease by $\theta = 1$ and of no disease by $\theta = 0$. The prior probability of having the disease is therefore $p(\theta = 1) = \pi$. A diagnostic test Y has sensitivity s and specificity t, so that $p(Y = 1|\theta = 1) = s$ and $p(Y = 0|\theta = 0) = t$. If a positive test result $y = 1$ is observed, then the posterior probability of the disease is $p(\theta = 1|y = 1) = \pi s/\{\pi s + (1 - \pi)(1 - t)\}$.

If we wish to assess the conflict between the prior $\theta \sim \text{Bernoulli}(\pi)$ and the likelihood of observing a positive test result given θ, $p(y = 1|\theta)$, then, in this case, $K = 1$, θ_a is the copy representing the prior, θ_b is the copy representing the likelihood, $\mathbf{y}_a$ is the empty set and $\mathbf{y}_b = y$ [see also Figure 3(a)(ii)]. We assume a reference prior $p(\theta_b = 1) = 0.5$, and so obtain a reference posterior $p(\theta_b = 1|y = 1) = s/\{s + (1 - t)\}$. The conflict measure is then

$$c = \{\pi s + (1 - \pi)(1 - t)\}/\{s + (1 - t)\}.$$

For example, if a diagnostic test has sensitivity and specificity 0.9, then the conflict when observing a positive test result is $c = 0.1 + 0.8\pi$, which has a minimum of 0.1 for very rare diseases.

3.2.2 *Conflict at a continuous node.* For θ with continuous support, we can no longer calculate $p(\theta_a = \theta_b|\mathbf{y}_a, \mathbf{y}_b)$ directly. Instead, we consider the posterior probability of $\delta = h(\theta_a) - h(\theta_b)$, where h is the function (Section 3.1) we choose that turns θ into a parameter for which it is reasonable to assume a uniform prior. To judge whether the null hypothesis $\delta = 0$ is plausible, and as do Gåsemyr and Natvig (2009), we adapt Box's (1980) suggestion of calculating the probability that a predictive density is smaller than the predictive density at the observed data, to calculating the probability that the *posterior* density of δ is smaller than that at 0, that is,

$$c = \Pr\{p_\delta(\delta|\mathbf{y}_a, \mathbf{y}_b) < p_\delta(0|\mathbf{y}_a, \mathbf{y}_b)\}.$$

This can also be interpreted as adopting a log scoring rule (Spiegelhalter et al., 1993, 1994; Gneiting and Raftery, 2007) for δ:

$$c = \Pr\{-\log p_\delta(\delta|\mathbf{y}_a, \mathbf{y}_b) > -\log p_\delta(0|\mathbf{y}_a, \mathbf{y}_b)\},$$

which we can think of as the predictive probability of getting a higher penalty than if we believe the null hypothesis that $\delta = 0$.

A problem is the need for evaluation of the posterior density $p_\delta(\delta|\mathbf{y}_a, \mathbf{y}_b)$, which is not available analytically in any but the simplest of examples. However, the Jeffreys' transformation $h(\cdot)$ to a location parameter on the real line may also ensure that the posterior distribution of δ is symmetric and unimodal. In this case, the conflict measure is simply double the tail-area beyond 0:

$$c = 2 \times \min\{\Pr(\delta > 0|\mathbf{y}_a, \mathbf{y}_b), 1 - \Pr(\delta > 0|\mathbf{y}_a, \mathbf{y}_b)\},$$

which is easily implemented using MCMC by counting the number of times θ_a exceeds θ_b in the sample. As Gåsemyr and Natvig (2009) have shown, if $h(\theta_a), h(\theta_b)$—and hence δ if θ_a and θ_b are independent—are normally distributed, then the p-value is uniformly distributed a priori. A value of c close to 0 therefore indicates a low degree of consistency between the posterior distributions of θ_a and θ_b.

If the distribution is not symmetric, a one-sided tail area may be more appropriate. As we will see in some of the examples of Section 4, we may in any case be interested in one-tailed tests $\delta \geq 0$ or $\delta \leq 0$, in which case asymmetry is not a problem. Clearly, if the distribution is multi-modal, the tail area is not an appropriate measure of where 0 lies in the posterior distribution of δ (Evans and Jang, 2010). Then (Gåsemyr and Natvig, 2009) we may consider using a kernel density estimate of MCMC samples from the posterior distribution to empirically obtain $c = \Pr\{p_\delta(\delta|\mathbf{y}_a, \mathbf{y}_b) < p_\delta(0|\mathbf{y}_a, \mathbf{y}_b)\}$, though this will clearly be dependent on the choice of bandwidth and kernel. We defer further discussion of these issues to Section 5.

3.2.3 *Conflict at multiple continuous nodes*. As seen from the example in Marshall and Spiegelhalter (2007) and Figure 1(c) of this paper, to obtain completely independent partitions of evidence in a DAG, often (particularly in hierarchical models) a vector of nodes would need to be split. How the DAG should be split will be context-dependent: either we are interested in comparing what we can infer about all the separator nodes in the vector from the two partitions of evidence or some of the separator nodes are nuisance parameters [such as the variance γ in Figure 1(c)] in which we are not directly interested. If the latter is the case, we can impose a cut such as the one shown in Figure 1(c) to prevent information from one partition influencing the nuisance parameters. In the former case, we can split each node in the vector and examine the posterior distribution of $\boldsymbol{\delta} = \mathbf{h}_a - \mathbf{h}_b = \{h_1(\theta_{a1}), \ldots, h_k(\theta_{ak})\}^T - \{h_1(\theta_{b1}), \ldots, h_k(\theta_{bk})\}^T$, where k is the length of the vector and the functions $h_1, \ldots, h_k$ are the appropriate Jeffreys' transformations. The key question is then how to calculate a multivariate p-value to test $\boldsymbol{\delta} = \mathbf{0}$. We consider three options, dependent on the posterior distribution of $\boldsymbol{\delta}$:

(i) If we are willing to assume multivariate normality for the posterior $p_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\mathbf{y}_a, \mathbf{y}_b)$, and denoting the posterior expectation and covariance of $\boldsymbol{\delta}$ by $\mathbb{E}_p$ and Cov_p, respectively, then (Gåsemyr and Natvig, 2009) the standardised discrepancy measure

$$\Delta = \mathbb{E}_p(\boldsymbol{\delta})^T \mathrm{Cov}_p(\boldsymbol{\delta})^{-1} \mathbb{E}_p(\boldsymbol{\delta})$$

may be compared with a $\mathcal{X}^2$ distribution with k degrees of freedom to obtain a conflict measure $c = 1 - \Pr\{\mathcal{X}^2_k \leq \Delta\}$.

(ii) If we are not willing to assume multivariate normality, but the posterior density $p_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\mathbf{y}_a, \mathbf{y}_b)$ is still symmetric and uni-modal, we can sample points from the posterior (e.g., using MCMC) and for each sample $\boldsymbol{\delta}_i$, calculate its Mahalanobis distance from their mean

$$\Delta_i = \{\boldsymbol{\delta}_i - \mathbb{E}_p(\boldsymbol{\delta})\}^T \mathrm{Cov}_p(\boldsymbol{\delta})^{-1} \{\boldsymbol{\delta}_i - \mathbb{E}_p(\boldsymbol{\delta})\}.$$

Then a conflict measure is $c = \Pr\{\Delta_i > \Delta\}$, the proportion over the MCMC sample of points that are further away from the mean than is $\mathbf{0}$. This is a means of calculating the multivariate tail area probability, analogous to counting the number of times in the MCMC sample δ is greater than 0 in the univariate case.

(iii) Finally, if the posterior distribution is skew and/or multi-modal, we could, as in the univariate case, obtain a kernel density estimate of the posterior based on the MCMC samples and use the estimate to calculate the probability that the posterior density at $\boldsymbol{\delta}$ is less than (at a lower contour than) at $\mathbf{0}$:

$$c = \Pr\{p_{\boldsymbol{\delta}}(\boldsymbol{\delta}|\mathbf{y}_a, \mathbf{y}_b) < p_{\boldsymbol{\delta}}(\mathbf{0}|\mathbf{y}_a, \mathbf{y}_b)\},$$

that is, that $\mathbf{0}$ lies in the tail of the distribution.

Note that the third approach will again be dependent on the choice of bandwidth and kernel.

4. EXAMPLES

We now illustrate the use of conflict p-values to detect conflict in a series of three increasingly complex evidence syntheses. All analyses were carried out in OpenBUGS 3.2.2 (Lunn et al., 2009) and R 2.15.0 (R Development Core Team, 2005).

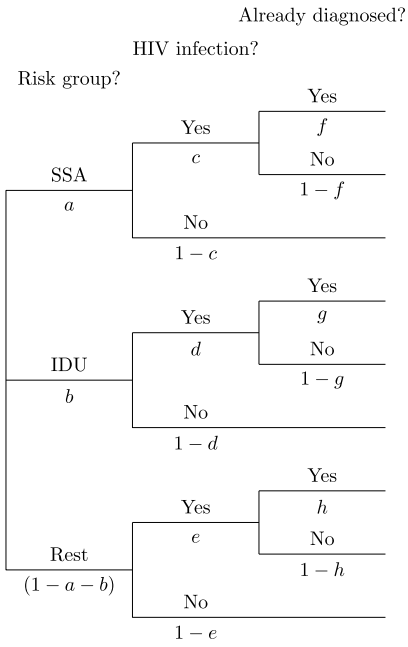

FIG. 4. *The HIV example: probability tree showing the epidemiological model from Ades and Cliffe (2002). "SSA" denotes women born in sub-Saharan Africa and "IDU" denotes injecting drug using women.*

4.1 HIV Example

Ades and Cliffe (2002) proposed a Bayesian synthesis of multiple sources of evidence to examine alternative strategies for screening HIV in prenatal clinics, with the specific aim of deciding whether to universally screen or to use targeted screening of high risk groups. Figure 4 shows the epidemiological part of their model, with the "basic parameters" a to h to be estimated. These letters represent the probabilities of women in prenatal clinics being in a particular risk group [a, b and $1 - a - b$, representing, respectively, women born in sub-Saharan Africa (SSA), women injecting drugs (IDU) and the remaining women]; being HIV infected (c, d and e, respectively, for the three risk groups); and being already diagnosed prior to clinic visit (f, g and h, resp., for HIV positive women in each of the three risk groups). Direct evidence is only available for a limited number of these parameters, but there is also indirect evidence informing functions ("functional parameters") of a to h as well as of an extra basic parameter w. Direct evidence is defined as a study with the aim of measuring a basic parameter. Indirect evidence is provided by the other studies through the logical functions that link the basic parameters. Table 1 shows the available data and the parameters, both basic and functional, that these inform, while Figure 5 shows a DAG of part of the model, demonstrating both the distributional and functional relationships. The basic parameters are the founder nodes to which prior distributions are assigned, whereas the functional parameters $p_1, \ldots, p_{12}$ are the probabilities informed directly by each of the 12 data sources.

The effect of the direct and indirect evidence on inference may be compared using two different types of node-splits: one at the level of the basic parameters and the second at the level of the probabilities $p_i, i \in 1, \ldots, 12$. The two types of node-splits are shown in Figure 6. The first node-split is carried out for each of the six basic parameters $\theta \in \{a, b, c, d, g, w\}$ for which direct data are available (studies 1 to 4, 10 and 11). The indirect evidence comprises all the remaining studies and the functional relationships assumed. The second type of node-split compares, for each $i \in 1, \ldots, 12$, the direct evidence on p_i provided by study i with the indirect evidence provided by the remaining studies, that is, a data-level cross-validation.

We adopt Jeffreys' prior for the nodes representing direct evidence in each case: $\theta_b \sim \text{Beta}(1/2, 1/2)$ for each $\theta \in \{a, b, c, d, g, w\}$ in the first set of node-splits; and $p_{ib} \sim \text{Beta}(1/2, 1/2)$ in partition b in the second set of node-splits [Figure 6(b)]. In both node-splits, the conflict p-values are two-tailed since we are testing for nonequality. Since the posterior densities of each difference function δ appear symmetric and uni-modal, the p-values are defined by taking twice the proportion of MCMC samples where $\delta \geq 0$ or $\delta < 0$, whichever is smaller. The results are based on two chains of 20,000 MCMC iterations each, after discarding a burn-in of 20,000 iterations.

Comparison of the direct and indirect evidence informing each $\theta \in \{a, b, c, d, g, w\}$ is shown in Figure 7, together with the conflict p-values. The plots and the p-values indicate that the direct evidence informing b and d, that is, studies 2 and 4, appear to be in conflict with the rest of the model. This is confirmed by the second type of node-split, as shown in Figure 8: the difference function $\delta_i = p_{ia} - p_{ib}$ is plotted, and the line $\delta_i = 0$ is shown together with the conflict p-value. Again, studies 2 and 4 are clearly in conflict with the rest of the model.

4.2 Influenza Example

In Presanis et al. (2011), the severity of the first two waves of the 2009 pandemic influenza outbreak experienced by England was estimated via a Bayesian

TABLE 1

HIV example: *data sources and the parameters they inform. "SSA" denotes sub-Saharan Africa and "IDU" denotes injecting drug using. "Seroprevalence" is the prevalence of HIV antibodies in blood samples from a "sero-survey"*

Source	Description of data	Parameter	Data y	n	y/n
1	Proportion of women born in SSA, 1999	$p_1 = a$	11,044	104,577	0.106
2	Proportion of women who are IDU in the last 5 years	$p_2 = b$	12	882	0.014
3	HIV prevalence in women born in SSA, 1997–1998	$p_3 = c$	252	15,428	0.016
4	HIV prevalence in IDU women, 1997–1999	$p_4 = d$	10	473	0.021
5	HIV prevalence in women not born in SSA, 1997–1998	$p_5 = \frac{db+e(1-a-b)}{1-a}$	74	136,139	0.001
6	HIV seroprevalence in pregnant women, 1999	$p_6 = ca + db + e(1 - a - b)$	254	102,287	0.002
7	Diagnosed HIV in SSA-born women as a proportion of all diagnosed HIV, 1999	$p_7 = \frac{fca}{fca+gdb+he(1-a-b)}$	43	60	0.717
8	Diagnosed HIV in IDU women as a proportion of diagnosed HIV in non-SSA-born women, 1999	$p_8 = \frac{gdb}{gdb+he(1-a-b)}$	4	17	0.235
9	Overall proportion of HIV diagnosed	$p_9 = \frac{fca+gdb+he(1-a-b)}{ca+db+e(1-a-b)}$	87	254	0.343
10	Proportion of infected IDU women diagnosed, 1999	$p_{10} = g$	12	15	0.800
11	Proportion of infected SSA-born women with serotype B, 1997–1998	$p_{11} = w$	14	118	0.119
12	Proportion of infected non-SSA-born women with serotype B, 1997–1998, assuming that 100% of infected IDU women have serotype B and that infected non-SSA-born non-IDU women have the same prevalence of serotype B as infected SSA-born women	$p_{12} = \frac{db+we(1-a-b)}{db+e(1-a-b)}$	5	31	0.161

evidence synthesis. Severity was measured by "case-severity ratios", the probability that an infection will lead to a more severe event such as hospitalisation (case-hospitalisation ratio CHR), intensive care (ICU) admission (case-ICU ratio CIR) or death (case-fatality ratio CFR). Restricting attention to the first wave only, Figure 9 is a DAG of the model, shown for a single age group (with the age index dropped for clarity). At the top, we have the three case-severity ratios, which can each be expressed as a product of component conditional probabilities $p_{i|j}$ of being a case at severity level i given severity level j. The severity levels are infection Inf, symptomatic infection S, hospitalisation H, ICU admissions I and death D, with Pop denoting the total (age-group specific) population of England. The case-severity ratios are functional parameters, whereas the probabilities $p_{i|j}$ are basic parameters assigned prior distributions. At the third layer of the DAG, the number of infections at each severity level $N_i, i \in \{\text{Inf}, S, H, I, D\}$ is a function of the conditional probabilities $p_{i|j}$ and the number at the less severe level j, N_j: $N_i = p_{i|j} N_j$.

The authors synthesised data, y_i, from surveillance systems monitoring influenza at different levels of severity, i (Figure 9), as well as serial sero-prevalence data indirectly informing the cumulative incidence of infection $p_{\text{Inf}|\text{Pop}}$. Some of the observed data were

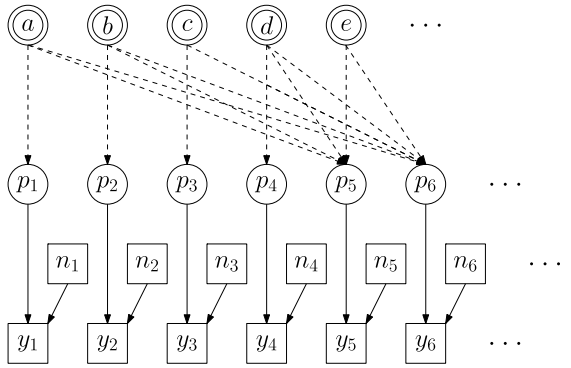

FIG. 5. *The HIV example*: *DAG showing part of the model. Note the functional relationships represented by the dashed arrows.*

recognised to under-ascertain the true number of infections at different severity levels, so were modelled as

$$y_i \sim \text{Binomial}(N_i, d_i),$$

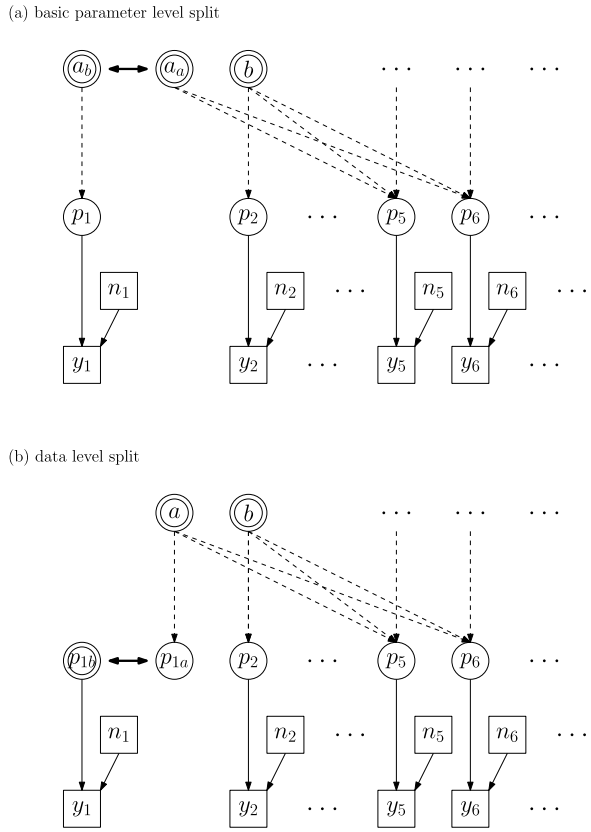

FIG. 6. *The HIV example*: *DAG showing the two types of node splits, one example of each. In* (a), *the node a is split to reflect direct* (a_b) *versus indirect* (a_a) *evidence. In* (b), *the node* p_1 *is split to reflect prior* (*all the indirect evidence on* p_{1a}) *versus likelihood* (*direct evidence on* p_{1b}) *in a cross-validation approach.*

where the probability parameters d_i represent the under-ascertainment or bias parameters. Estimates $\hat{N}_B$ of the number symptomatic were produced by the Health Protection Agency (HPA) from data on primary care consultations for influenza-like-illness and virological testing for the pandemic strain, adjusted for the proportion of symptomatic cases contacting primary care. Early on in the pandemic, these estimates were thought to be underestimates (B here stands for bias). The HPA estimates were therefore incorporated in the synthesis by modelling them as

$$\hat{N}_B \sim \text{Binomial}(N_S, d_S).$$

Informative Beta priors were given to the probabilities $p_{H|S}$ and $p_{S|\text{Inf}}$, representing estimates from a cure-rate model fitted to data on the first few thousand confirmed cases and estimates from the literature on seasonal influenza, respectively. The remaining probabilities $p_{i|j}$ were assigned Beta(1, 1) priors. An informative Beta prior representing estimates from a capture-recapture study was adopted for the ascertainment probability for death, d_D. The other ascertainment probabilities, d_H and d_S, were given Beta(1, 1) priors.

Given uncertainties over possible biases in both the sero-prevalence data and the HPA estimates $\hat{N}_B$, Presanis et al. (2011) carried out a number of sensitivity analyses to the inclusion of these "denominator" data, making different assumptions about which source, if any, is biased. An alternative approach is to split the DAG at the node N_S, having dropped from the model for now the bias parameter d_S, to assess the consistency of the different groups of evidence. Denote the full model of Figure 9 but with d_S removed by Model 1. We wish to compare the sero-prevalence data combined with the informative prior for the proportion symptomatic $p_{S|\text{Inf}}$ (the "parent" model 2) against the HPA estimates $\hat{N}_B$, combined with all the severe end data and priors (the "child" model 3). Figure 10 shows this node-split. In the child model, $\log(N_S^3)$ is assigned the Jeffreys' prior (uniform on the real line). The difference function we are interested in is $\delta = \log(N_S^2) - \log(N_S^3)$ and the tail probability we assess is the one-sided probability $c = \Pr\{\delta < 0\}$, since the HPA estimates are recognised underestimates. The p-values are calculated as the proportion of MCMC samples where $\delta < 0$.

Three chains of a million MCMC iterations each, with the first 200,000 discarded as burn-in, were thinned to every 10th iteration to obtain 240,000 posterior samples on which to base results. Figure 11 shows,

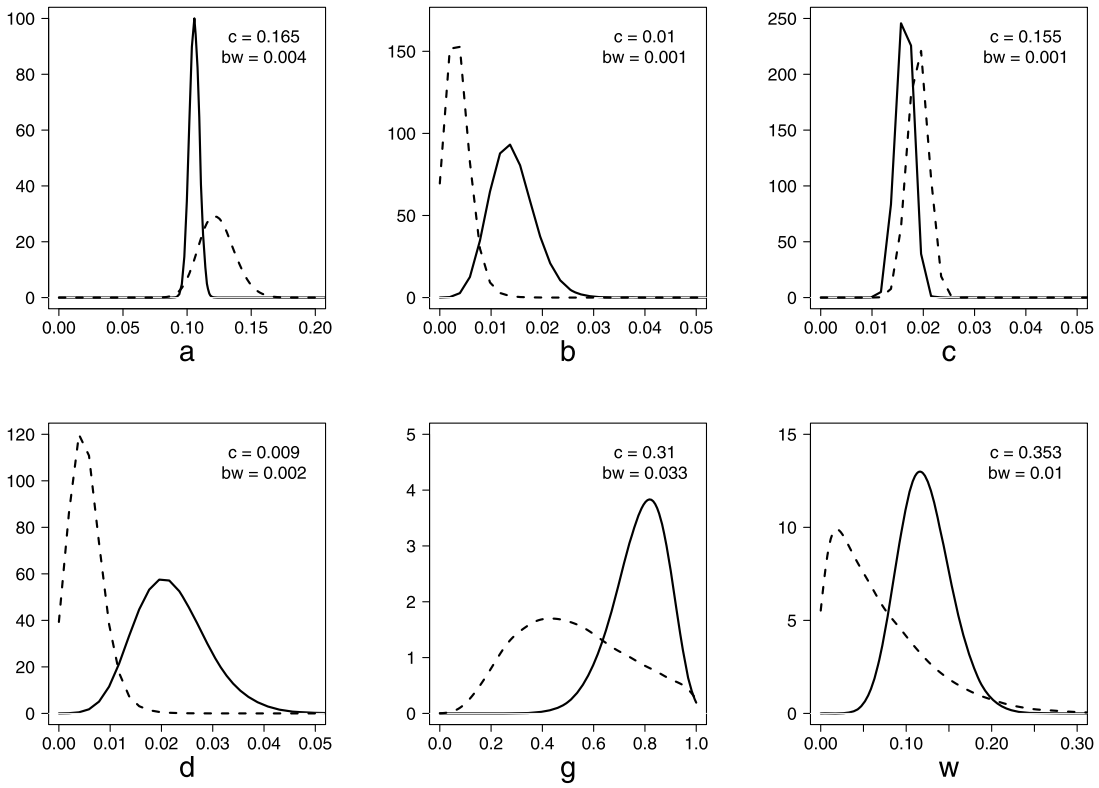

FIG. 7. *The HIV example*: *posterior distributions reflecting direct* (*solid lines*) *vs indirect* (*dashed lines*) *evidence at the* 6 *basic parameters with direct evidence available*, $a, \ldots, d, g, w$. *The conflict* p*-value* (c), *calculated as twice the proportion of MCMC samples where* $\delta \geq 0$ *or* $\delta < 0$, *whichever is smaller, is given in each plot. The bandwidth* (bw) *of the kernel density estimate used to plot the posterior distributions is also shown.*

for each of seven age groups, the posterior distributions of $\log(N_S)$ in each of the three models: the full model 1, the parent model 2 and the child model 3. Two facts are immediately apparent from these plots: first, that the estimates provided by the sero-prevalence data in the parent model are very uncertain, particularly in the adult age groups, where sample sizes were small; and second, the resulting influence of the more severe end data, including the HPA estimates, is seen in the closeness of the full model estimates to the child model estimates. What is less apparent from these plots is the extent of conflict between the two sets of information, except in the age group < 1 where the conflict is clear to see. The plots of the difference function δ in Figure 12, together with the conflict p-values shown, are required to assess the conflict in other age groups. From these, we see p-values less than 0.1 in the child age groups, providing evidence of conflict in these groups. By contrast, in the adult age groups, the uncertainty in the estimates of N_S^2 in the parent model 2 is so large that there is no conflict.

If we had not already suspected the HPA estimates were underestimates and wished to assess potential conflict using a two-sided test, then in this example calculation of the two-sided p-value would not be so straightforward. Particularly in the over-65 age group, the posterior difference function is skewed (Figure 12). We could therefore, as suggested in Section 3.2, use kernel density estimation to calculate the two-sided p-value. Figure 13 compares the one-sided to the resulting two-sided p-value for the 65+ group, using a bandwidth of 0.5.

4.3 Multivariate Example: Growth in Rats

A simpler example that nevertheless demonstrates the complexity of multivariate node-splitting is provided by data from a population growth experiment considered by Gelfand et al. (1990). The data comprise the weight y_{ij} of each of 30 rats ($i \in 1, \ldots, 30$) at ages 8, 15, 22, 29 and 36 days, indexed by $j \in 1, \ldots, 5$. The authors' null model H_0 assumes a normal error and random coefficient linear growth curves with time t_j measured in days centred on 22. The intercept and gra-

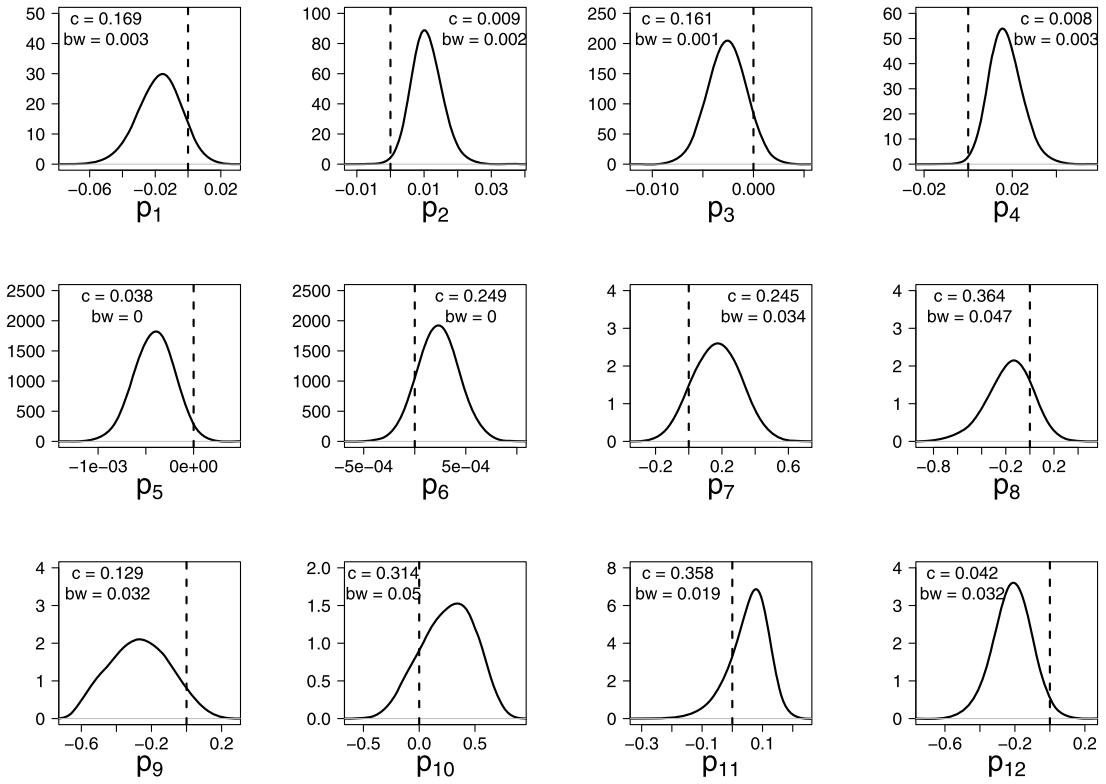

FIG. 8. *The HIV example*: *the posterior distribution of the difference function* δ (*solid lines*) *at each functional parameter* $p_i, i \in 1, \ldots, 12$ (*data-level cross-validation*). $\delta = 0$ *is shown by the dashed vertical line. The conflict* p*-value* (c), *calculated as twice the proportion of MCMC samples where* $\delta \geq 0$ *or* $\delta < 0$, *whichever is smaller, is given in each plot. The bandwidth* (bw) *of the kernel density estimate used to plot the posterior distributions is also shown.*

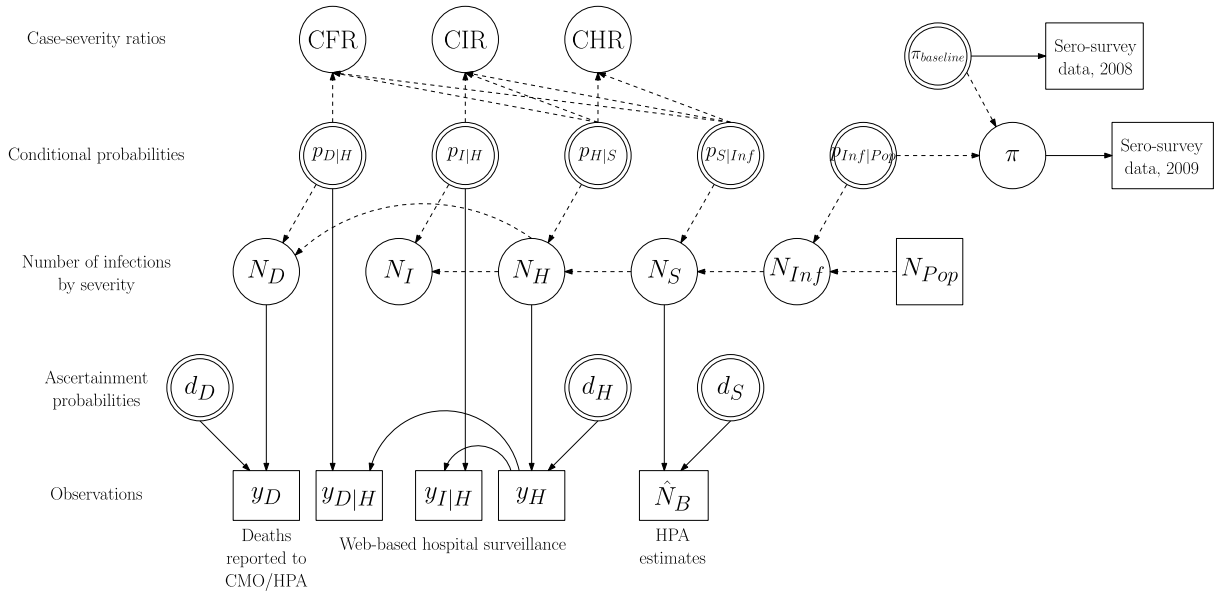

FIG. 9. *The influenza example*: *DAG for a single age-group* (*age index dropped for clarity*) *of the model reported in Presanis et al.* (2011).

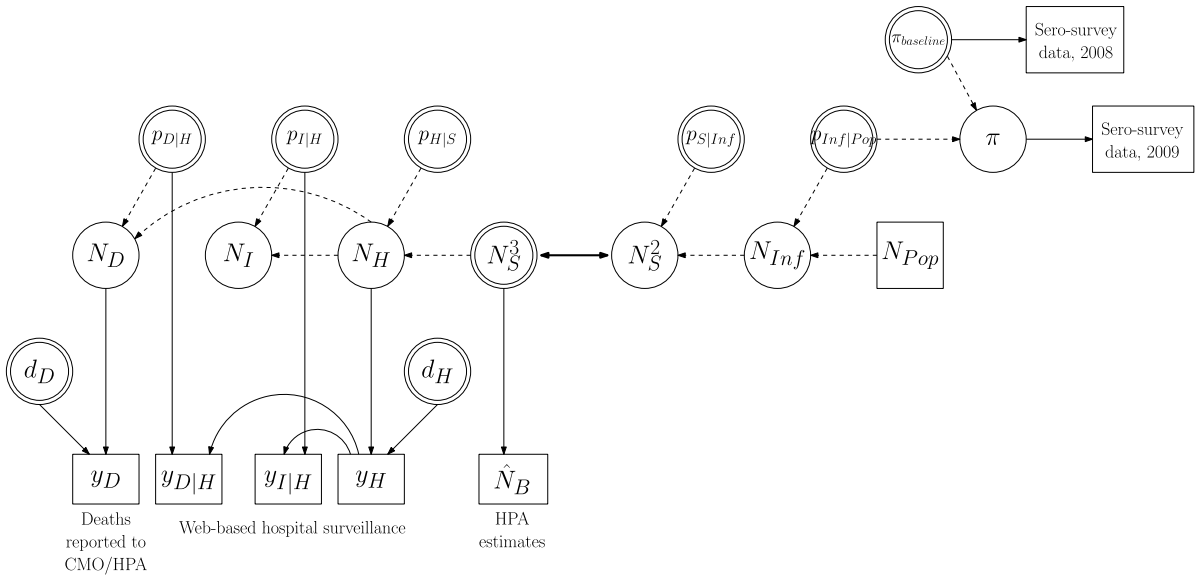

FIG. 10. *The influenza example: the node split at N_S. On the right is model 2, the "parent" model. On the left is model 3, the "child" model. The double-headed arrow represents the comparison between the two.*

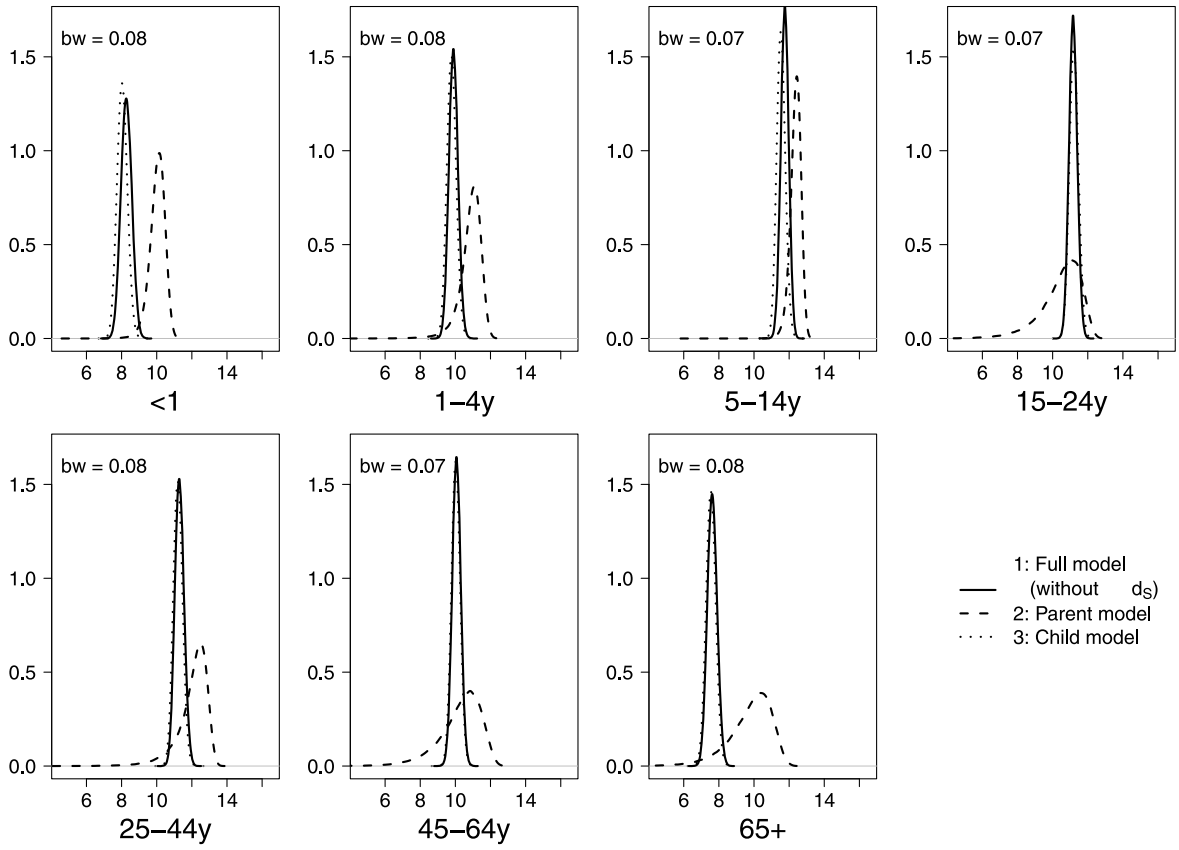

FIG. 11. *The influenza example: the posterior distribution of the number symptomatic N_S on the log-scale, by age group and model. The bandwidth (bw) of the kernel density estimate used to plot the posterior distributions is shown in the top left of each plot.*

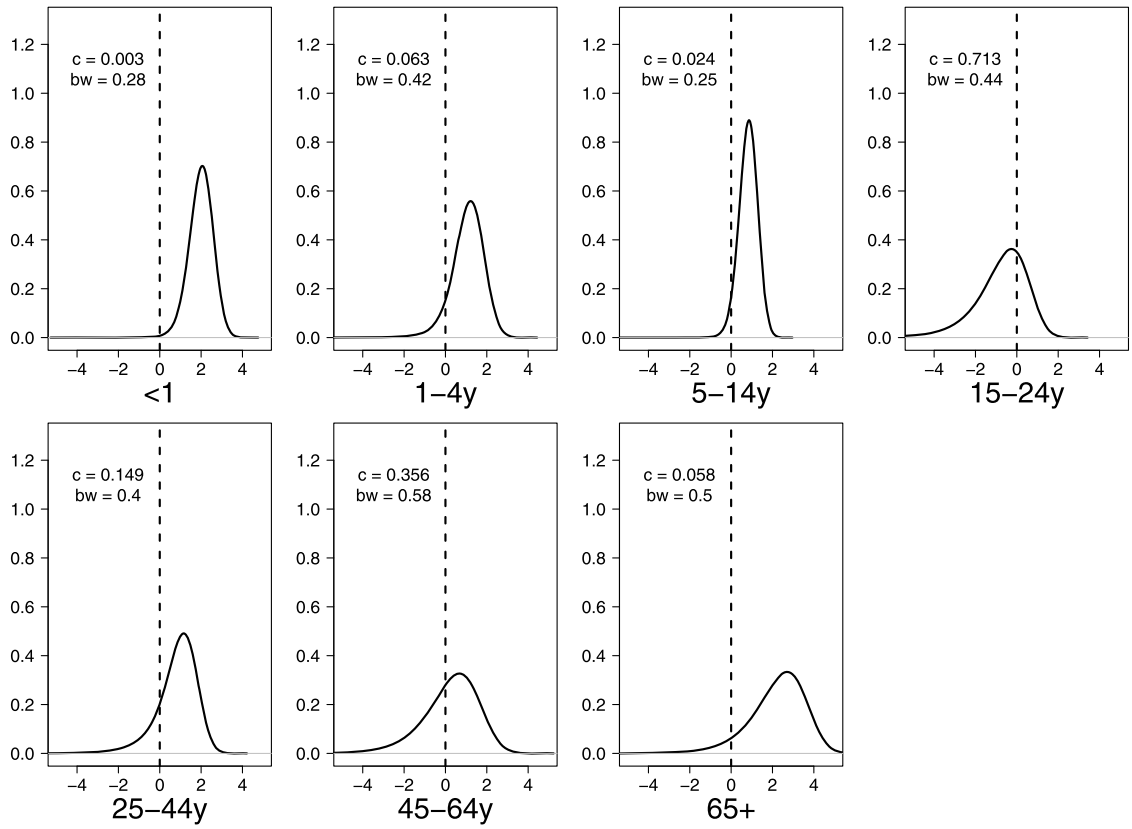

FIG. 12. *The influenza example*: *the posterior distribution of the difference function* $\delta = \log(N_S^2) - \log(N_S^3)$. *The vertical dashed line gives* $\delta = 0$. *The one-sided conflict p-value* (c), *calculated as the proportion of MCMC samples where* $\delta < 0$, *is given in each plot. The bandwidth* (bw) *of the kernel density estimate used to plot the posterior distributions is also shown.*

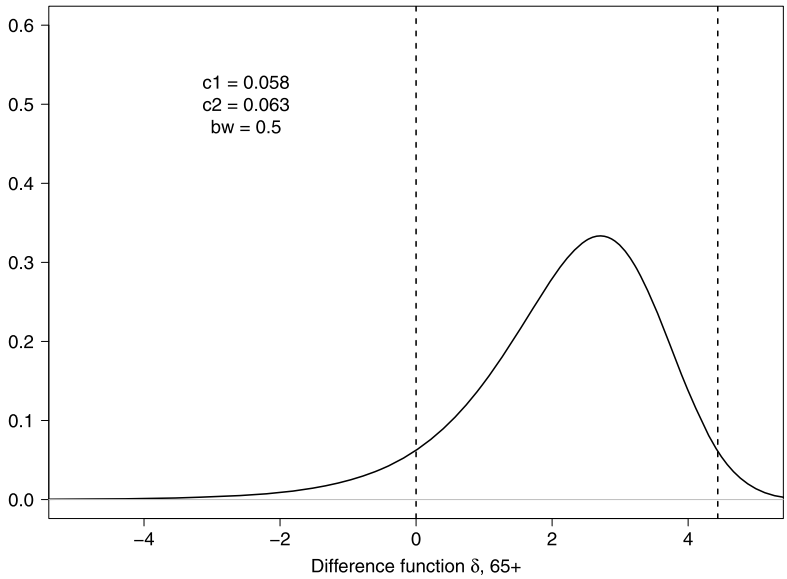

FIG. 13. *The influenza example*: *one-* (c_1) *and two-sided* (c_2) *p-values for the* 65+ *age group*, *calculated using kernel density estimation with bandwidth* bw. *The two vertical dashed lines show where* $\delta = 0$ *and the corresponding value* $\delta = k = 4.44$, *such that the density at* 0 *and at* k *is equal*, *lie in the posterior distribution of* δ.

dient are given a bivariate normal prior, so that

$$
\begin{aligned}
y_{ij} &\sim \mathrm{N}(\mu_{ij}, \sigma^2),\\
\mu_{ij} &= \phi_{i1} + \phi_{i2} t_j,\\
\boldsymbol{\phi}_i &\sim \mathrm{MVN}_2(\boldsymbol{\beta}, \boldsymbol{\Omega}),
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\phi}_i = (\phi_{i1}, \phi_{i2})^T$ and $\boldsymbol{\beta}, \boldsymbol{\Omega}, \sigma^2$ are given proper but very diffuse prior distributions:

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \mathrm{MVN}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}10^{-6} & 0\\ 0 & 10^{-6}\end{pmatrix}\right),\\
\mathbf{W} &\sim \mathrm{Wishart}\left(\begin{pmatrix}200 & 0\\ 0 & 0.2\end{pmatrix}, 2\right),\\
\boldsymbol{\Omega} &= \mathbf{W}^{-1},\\
\tau &\sim \Gamma(10^{-3}, 10^{-3}),\\
\sigma^2 &= \tau^{-1}.
\end{aligned}
$$

We wish to examine the suitability of the bivariate random effects distribution for each rat, which in this case requires assessing a multivariate node-split, at the vector of parameters $\boldsymbol{\phi}_i$ (Figure 14). For each rat i (i.e., each cross-validation), in one partition, we fit a fixed effects model to the data $\mathbf{y}_i$ to estimate the nodes denoted $\boldsymbol{\phi}_i^{\mathrm{lik}}$, which are assigned independent (improper) Jeffreys' priors. In the other partition, we predict the nodes $\boldsymbol{\phi}_i^{\mathrm{rep}}$ from the random effects model fitted to the data on the remaining rats, denoted by $\mathbf{y}_{\setminus i}$. As with the example from Marshall and Spiegelhalter (2007), to form a complete split in the DAG would require also splitting the variance parameter σ^2. Since our primary interest is in assessing the random effects distribution for $\boldsymbol{\phi}_i$ rather than for the rat-specific data $\mathbf{y}_i$, and as a rat-specific variance σ_i^2 may not be well identified from the data on one rat alone, we treat σ^2 as a nuisance parameter. We therefore place a "cut" in the DAG to prevent feedback from rat i's data to σ^2.

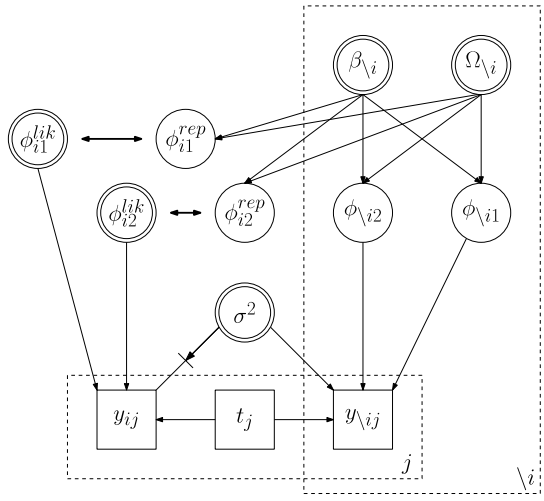

FIG. 14. *The rats example: DAG showing the comparison of the fixed effect model for rat i ($\boldsymbol{\phi}_i^{\mathrm{lik}}$) with the random effects prediction from the remaining rats ($\boldsymbol{\phi}_i^{\mathrm{rep}}$).*

A multivariate difference function was defined for each rat i: $\boldsymbol{\delta}_i = \boldsymbol{\phi}_i^{\mathrm{rep}} - \boldsymbol{\phi}_i^{\mathrm{lik}}$ and MCMC samples from the posterior distribution of $\boldsymbol{\delta}_i$ were obtained based on two chains of 20,000 iterations each, following a burn-in of 20,000 iterations. Plots for each rat of the samples from the joint posterior distribution of $\boldsymbol{\delta}_i$ suggest that at least uni-modality and symmetry hold approximately, and possibly bivariate normality also (see, e.g., rat 9 in Figure 15). We therefore calculate a conflict p-value based on each of the first two suggestions in Section 3.2.3, shown in Figure 16. Both methods of defining the p-value give similar results and suggest that rat 9 is discrepant, with p-values of 0.003 and 0.006 for the χ^2- and Mahalanobis-based methods, respectively. Figure 15 shows the joint posterior distribution of $\boldsymbol{\delta}_9$, with crosses denoting samples that are further away (in terms of Mahalanobis distance) from the posterior mean (the white star) than is the point $(0, 0)$.

A parallel literature exists on model diagnostics to identify unit-level outliers for classical multilevel models (Langford and Lewis, 1998). The basic idea of this diagnostic is to add extra fixed effects or dummy variables for the unit under consideration and compare this model and the null by comparing the change in deviance between the two, evaluated at the maximum likelihood estimate, to a χ_1^2 distribution. For example, in the rats study, an extra fixed effect could be added for the slope and intercept of rat i and this model (the alternative) could be compared with the null model (1). Ohlssen, Sharples and Spiegelhalter (2007) show that this method is equivalent to the Bayesian cross-validatory mixed-predictive p-value of Marshall and Spiegelhalter (2007) in balanced one-way random effects models with uniform priors.

5. DISCUSSION

We have described a generic simulation-based technique that can be used as a diagnostic for conflicting information at any node(s) $\boldsymbol{\theta}$ in a DAG, generalising the conflict p-value first proposed by Marshall and Spiegelhalter (2007). We have presented a framework focussed on a conflict diagnostic that is both useful and straightforward to construct for complex, typically nonstandard, evidence synthesis models. We have

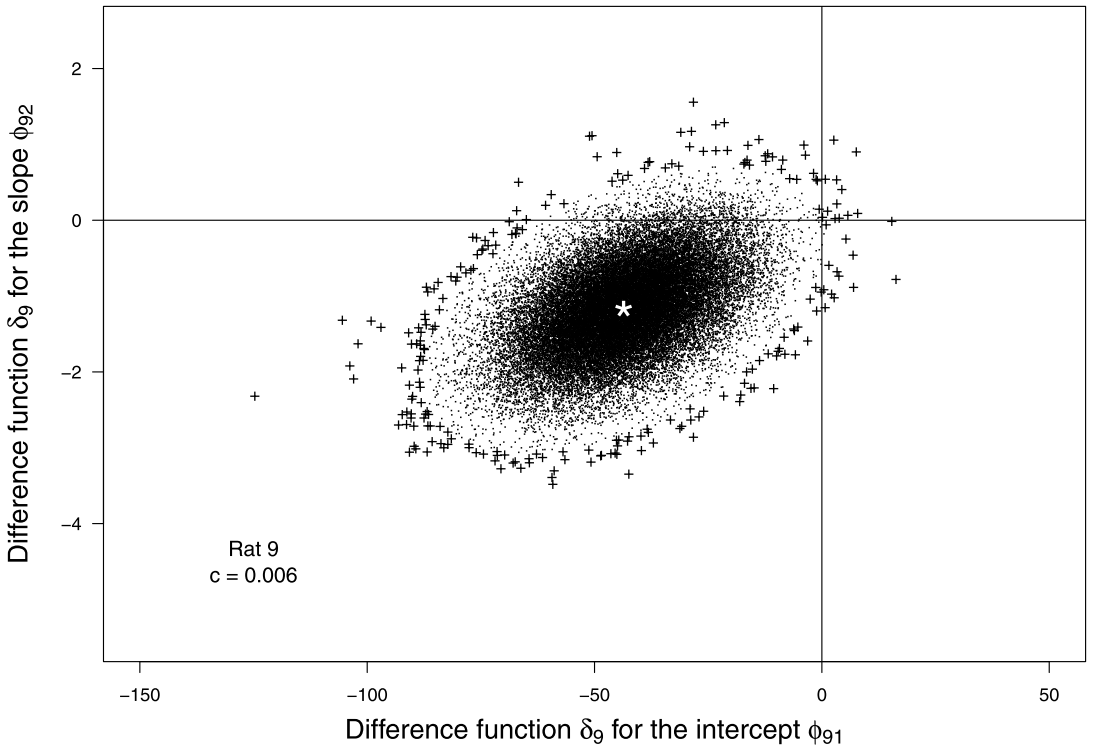

FIG. 15. *The rats example*: *joint posterior distribution of* $\boldsymbol{\delta}_9$. *Points more extreme than* $(0, 0)$ (*i.e.*, *further from the mean in terms of Mahalanobis distance and therefore lying in the tail of the distribution*) *are shown as crosses*. *The white star denotes the posterior mean* $\mathbb{E}_p(\boldsymbol{\delta}_9)$.

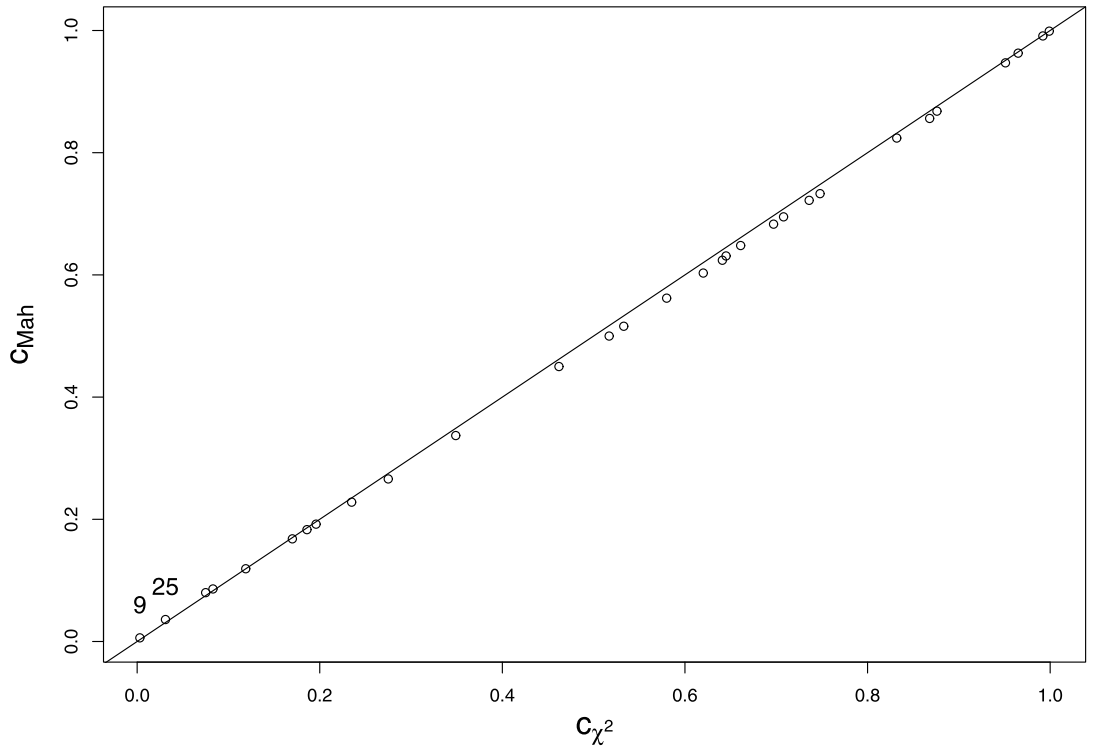

FIG. 16. *The rats example*: *on the x-axis*, c_{χ^2} *is the p-value using the* χ^2 *approach*; *on the y-axis*, c_{Mah} *is the p-value using the Mahalanobis approach*. *Rats* 9 *and* 25 *have p-values less than* 0.05.

given recommendations, via three substantive examples, for how to perform node-splits to assess different types of conflict. In particular, we have demonstrated different methods of handling multivariate node-splits, dependent on the context. If a "separator" parameter is not of primary interest, but a nuisance parameter, we suggest making use of the "cut" function in software such as OpenBUGS, to prevent information flow to the nuisance parameter from one partition. For multivariate "separator" nodes that are of importance, we recommend a hierarchy of options for defining a multivariate conflict p-value, dependent on the normality, uni-modality and symmetry or otherwise of the posterior distribution of the difference function. In focussing on nonstandard but realistic situations, our framework goes beyond that so far proposed in the literature [e.g., Bayarri and Castellanos, 2007; Gåsemyr and Natvig, 2009].

There are still practical and theoretical considerations raised by the analysis of conflict that require further investigation. For large complex models, a systematic examination of every potential conflict at every node in a DAG may induce a large computational cost, since—as with cross-validation—every node-split requires a new model run. Furthermore, such a systematic conflict assessment would result in the multiple comparison problem. To address these issues, different approaches may be taken. In practice, approximations to full cross-validation as suggested by Marshall and Spiegelhalter (2007) may be employed. Or it may be prudent to be selective about node-splits to examine, based on the context of the problem, as we were in the influenza example and to some extent in the HIV example. One strategy may be to employ a diagnostic such as comparison of posterior mean deviance to the number of data items (Dempster, 1997; Spiegelhalter et al., 2002) to detect lack of fit to particular items, which may then be an indicator of conflicting evidence (Ades and Cliffe, 2002; Presanis et al., 2008). Then the choice of node-splits to examine may be guided by the locations in the DAG of lack of fit. However, the posterior mean deviance has its own limitations as a diagnostic (Presanis et al., 2008) and, indeed, conflicting evidence may not necessarily manifest as lack of fit. This was the case with the influenza example (results not shown), where instead an informative prior for the proportion of infections which were symptomatic was shifted in the posterior to an implausibly low range. If systematic conflict assessment is indeed an aim of the analyst, then the problem of multiple testing may be addressed by combining node-splitting with methods for controlling the false discovery rate [FDR, Benjamini and Hochberg, 1995; Jones, Ohlssen and Spiegelhalter, 2008]. A further complication is the possibility of correlated hypothesis tests, since the different node-splits assessed may have nodes in common with each other. FDR methods do extend to correlated tests (Benjamini and Yekutieli, 2001), but such methods may have an impact on the power to detect conflict or identify the network of nodes in a DAG that are conflicting (Hothorn, Bretz and Westfall, 2008; Bretz, Hothorn and Westfall, 2011). A possibility for further investigation is to account for correlation through the general framework for multiple comparisons of Hothorn, Bretz and Westfall (2008), Bretz, Hothorn and Westfall (2011), although this requires asymptotic multivariate normality.

In models where flat improper priors are employed and the likelihood dominates the prior, the posterior will be approximately normal, so that the conditions of the framework of Gåsemyr and Natvig (2009) hold, and the conflict p-values will therefore be uniformly distributed under the null hypothesis that $\theta_a = \theta_b$. However, in practice, analysts modelling complex phenomena rarely use improper priors, perhaps instead using either (i) proper priors with very large variances; or (ii) informative priors to represent previous knowledge or to ensure identifiability. In case (i), approximate normality of the posterior difference function δ will again ensure the p-values are uniform under the null, though sensitivity analyses should be employed to check the prior is dominated by the likelihood. Furthermore, our recommendation of the use of Jeffreys' priors for appropriate transformations of θ_b should, we hope, result in a posterior difference function that is at least approximately (multivariate) normal. Sensitivity analysis may again be employed to assess our choice of reference prior (Jeffreys' or some other uniform prior) and transformation function $h(\cdot)$ for the split node. In case (ii), where informative priors are assigned to nodes other than θ_b, the posterior distribution of θ_a and hence the difference function δ are not guaranteed to be even approximately normal. This potential nonnormality may be exacerbated by small sample sizes, as was the case in our influenza example, where in some age groups the posterior difference function was somewhat skewed. Our suggestion is to then use kernel density estimation to obtain a conflict p-value, though clearly more work would be required to understand the distribution of such a p-value. Kernel density estimation also raises its own questions of how to choose a kernel

and a bandwidth with which to smooth, suggesting the need for sensitivity analyses to these choices. Kernel density estimation is also computationally demanding, particularly for multivariate surfaces, and may result in p-values that are not invariant to transformation. More generally, a conflict measure defined as the probability that a density function is smaller than a given threshold is preferable in cases where the distribution of δ is not symmetric and unimodal, since it would identify regions of surprise not only in the tails, but also in shallow anti-modes. However, the question of how to obtain such a p-value, that is also invariant to transformation, requires further investigation (Evans and Jang, 2010).

Our case studies in assessing conflict have demonstrated the great importance of visualisation, with plots of, for example, the posterior difference function or the posterior distributions corresponding to different partitions of a DAG adding to our understanding of what and where conflict is occurring. The influence of different partitions of evidence on estimation can also be visualised from such plots. It is important to note that node-splitting is a diagnostic, one step in the inference-criticism cycle and a pointer to where in a DAG further analysis is required. The next step is to understand, within the context of the specific problems under analysis, the reasons for the inconsistencies, and therefore to resolve the conflict. There are many possibilities for accommodating conflict, including: the exclusion of sources of evidence; the addition of extra variation to account for potential bias (e.g., Andrade and O'Hagan, 2006; Evans and Jang, 2011); and model elaboration to explicitly model biases or account for unexplained heterogeneity (e.g., DuMouchel and Harris, 1983; Lu and Ades, 2006; Greenland, 2009; Presanis et al., 2008; Welton et al., 2009; Turner et al., 2009; Higgins et al., 2012; White et al., 2012). Any such model development will then lead to the next iteration of inference and criticism, in the spirit of Box (1980) and O'Hagan (2003). Clearly, also, in any Bayesian analysis, sensitivity analyses, both to prior distributions, whether vague or informative, and in the case of evidence synthesis, to the sources of information included, are an important part of the model criticism process.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council [Unit Programme Numbers U105260566 and U105260557]. We thank the Health Protection Agency for providing the data for the influenza example and Richard Pebody, in particular, for useful discussions about the detected conflict. We are also grateful to the participants of the workshop "Explaining the results of a complex probabilistic modelling exercise: conflict, consistency and sensitivity analysis", sponsored by the Medical Research Council Population Health Sciences Research Network, in Cambridge in 2006, for many useful discussions. We acknowledge also the constructive comments of two referees, that led to an improved manuscript.

REFERENCES

ADES, A. E. and CLIFFE, S. (2002). Markov chain Monte Carlo estimation of a multiparameter decision model: Consistency of evidence and the accurate assessment of uncertainty. *Medical Decision Making* **22** 359–371.

ADES, A. E. and SUTTON, A. J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: Current approaches. *J. Roy. Statist. Soc. Ser. A* **169** 5–35. MR2222010

ANDRADE, J. A. A. and O'HAGAN, A. (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Anal.* **1** 169–188. MR2227369

BAYARRI, M. J. and BERGER, J. O. (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics*, 6 (*Alcoceber*, 1998) (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 53–82. Oxford Univ. Press, New York. MR1723493

BAYARRI, M. J. and BERGER, J. O. (2000). p values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142, 1157–1170. MR1804239

BAYARRI, M. J. and CASTELLANOS, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statist. Sci.* **22** 322–343. MR2416808

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BIRRELL, P. J., KETSETZIS, G., GAY, N. J., COOPER, B. S., PRESANIS, A. M., HARRIS, R. J., CHARLETT, A., ZHANG, X.-S., WHITE, P. J., PEBODY, R. G. and DE ANGELIS, D. (2011). Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. *Proc. Natl. Acad. Sci. USA* **108** 18238–18243.

BOUSQUET, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *J. Appl. Stat.* **35** 1011–1029. MR2522125

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

BOX, G. E. P. and TIAO, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley, New York.

BRETZ, F., HOTHORN, T. and WESTFALL, P. (2011). *Multiple Comparisons Using R*, 1st ed. Chapman & Hall/CRC, London.

CLARK, J. S., BELL, D., CHU, C., COURBAUD, B., DIETZE, M., HERSH, M., HILLERISLAMBERS, J., IBÁÑEZ, I., LADEAU, S., MCMAHON, S., METCALF, J., MOHAN, J., MORAN, E., PANGLE, L., PEARSON, S., SALK, C., SHEN, Z., VALLE, D. and WYCKOFF, P. (2010). High-dimensional coexistence based on individual variation: A synthesis of evidence. *Ecological Monographs* **80** 569–608.

COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York. MR1697175

DAHL, F. A., GÅSEMYR, J. and NATVIG, B. (2007). A robust conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.* **34** 816–828. MR2396940

DAWID, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292.

DEMPSTER, A. P. (1997). The direct use of likelihood for significance testing. *Statist. Comput.* **7** 247–252.

DIAS, S., WELTON, N. J., CALDWELL, D. M. and ADES, A. E. (2010). Checking consistency in mixed treatment comparison meta-analysis. *Stat. Med.* **29** 932–944. MR2752057

DUMOUCHEL, W. H. and HARRIS, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species. *J. Amer. Statist. Assoc.* **78** 293–315. MR0711105

EVANS, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Comm. Statist. Theory Methods* **26** 1125–1143. MR1450226

EVANS, M. and JANG, G. H. (2010). Invariant P-values for model checking. *Ann. Statist.* **38** 512–525. MR2589329

EVANS, M. and JANG, G. H. (2011). Weak informativity and the information in one prior relative to another. *Statist. Sci.* **26** 423–439. MR2917964

EVANS, M. and MOSHONOV, H. (2006). Checking for prior-data conflict. *Bayesian Anal.* **1** 893–914 (electronic). MR2282210

EVANS, M. and MOSHONOV, H. (2007). Checking for prior-data conflict with hierarchically specified priors. In *Bayesian Statistics and Its Applications* (A. K. Upadhyay, U. Singh and D. Dey, eds.) 145–159. Anamaya Publishers, New Delhi.

GÅSEMYR, J. and NATVIG, B. (2009). Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.* **36** 822–838. MR2573310

GAMERMAN, D. and LOPES, H. F. (2006). *Markov Chain Monte Carlo*: *Stochastic Simulation for Bayesian Inference*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2260716

GELFAND, A. E., HILLS, S. E., RACINE-POON, A. and SMITH, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85** 972–985.

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, London.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GREENLAND, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statist. Sci.* **24** 195–210. MR2655849

HENDERSON, D. A., BOYS, R. J. and WILKINSON, D. J. (2010). Bayesian calibration of a stochastic kinetic computer model using multiple data sources. *Biometrics* **66** 249–256. MR2756712

HIGGINS, J. P. T., JACKSON, D., BARRETT, J. K., LU, G., ADES, A. E. and WHITE, I. R. (2012). Consistency and inconsistency in network meta-analysis: Concepts and models for multi-arm studies. *Research Synthesis Methods* **3** 98–110.

HJORT, N. L., DAHL, F. A. and STEINBAKK, G. H. (2006). Post-processing posterior predictive p-values. *J. Amer. Statist. Assoc.* **101** 1157–1174. MR2324154

HOTHORN, T., BRETZ, F. and WESTFALL, P. (2008). Simultaneous inference in general parametric models. *Biom. J.* **50** 346–363. MR2521547

JACKSON, C., RICHARDSON, S. and BEST, N. (2008). Studying place effects on health by synthesising individual and area-level outcomes. *Social Science & Medicine* **67** 1995–2006.

JACKSON, D., WHITE, I. R. and CARPENTER, J. (2012). Identifying influential observations in Bayesian models by using Markov chain Monte Carlo. *Stat. Med.* **31** 1238–1248. MR2925692

JOHNSON, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Anal.* **2** 719–733. MR2361972

JONES, H. E., OHLSSEN, D. I. and SPIEGELHALTER, D. J. (2008). Use of the false discovery rate when comparing multiple health care providers. *J. Clin. Epidemiol.* **61** 232–240.

KASS, R. E. (1990). Data-translated likelihood and Jeffreys's rules. *Biometrika* **77** 107–114. MR1049412

LANGFORD, I. H. and LEWIS, T. (1998). Outliers in multilevel data. *J. Roy. Statist. Soc. Ser. A* **161** 121–160.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

LU, G. and ADES, A. E. (2006). Assessing evidence inconsistency in mixed treatment comparisons. *J. Amer. Statist. Assoc.* **101** 447–459. MR2256166

LUNN, D., SPIEGELHALTER, D., THOMAS, A. and BEST, N. (2009). The BUGS project: Evolution, critique and future directions. *Stat. Med.* **28** 3049–3067. MR2750401

MARSHALL, E. C. and SPIEGELHALTER, D. J. (2007). Identifying outliers in Bayesian hierarchical models: A simulation-based approach. *Bayesian Anal.* **2** 409–444. MR2312289

O'HAGAN, A. (2003). HSSS model criticism (with discussion). In *Highly Structured Stochastic Systems*, 1st ed. (P. J. Green, N. L. Hjort and S. Richardson, eds.). Oxford Univ. Press, New York.

OHLSSEN, D. I., SHARPLES, L. D. and SPIEGELHALTER, D. J. (2007). A hierarchical modelling framework for identifying unusual performance in health care providers. *J. Roy. Statist. Soc. Ser. A* **170** 865–890. MR2408982

PRESANIS, A. M., DE ANGELIS, D., SPIEGELHALTER, D. J., SEAMAN, S., GOUBAR, A. and ADES, A. E. (2008). Conflicting evidence in a Bayesian synthesis of surveillance data to estimate human immunodeficiency virus prevalence. *J. Roy. Statist. Soc. Ser. A* **171** 915–937. MR2530293

PRESANIS, A. M., PEBODY, R. G., PATERSON, B. J., TOM, B. D. M., BIRRELL, P. J., CHARLETT, A., LIPSITCH, M. and DE ANGELIS, D. (2011). Changes in severity of 2009 pandemic A/H1N1 influenza in England: A Bayesian evidence synthesis. *BMJ* **343** d5408+.

R DEVELOPMENT CORE TEAM. (2005). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1167, 1171–1172. MR1804240

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SCHEEL, I., GREEN, P. J. and ROUGIER, J. C. (2011). A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models. *Scand. J. Stat.* **38** 529–550. MR2833845

SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York.

SPIEGELHALTER, D. J., DAWID, A. P., LAURITZEN, S. L. and COWELL, R. G. (1993). Bayesian analysis in expert systems. *Statist. Sci.* **8** 219–283. MR1243594

SPIEGELHALTER, D. J., HARRIS, N. L., BULL, K. and FRANKLIN, R. C. G. (1994). Empirical evaluation of prior beliefs about frequencies: Methodology and a case study in congenital heart disease. *J. Amer. Statist. Assoc.* **89** 435–443.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639. MR1979380

STEINBAKK, G. H. and STORVIK, G. O. (2009). Posterior predictive p-values in Bayesian hierarchical models. *Scand. J. Stat.* **36** 320–336. MR2528987

TURNER, R. M., SPIEGELHALTER, D. J., SMITH, G. C. S. and THOMPSON, S. G. (2009). Bias modelling in evidence synthesis. *J. Roy. Statist. Soc. Ser. A* **172** 21–47. MR2655603

WELTON, N. J., ADES, A. E., CARLIN, J. B., ALTMAN, D. G. and STERNE, J. A. C. (2009). Models for potentially based evidence in meta-analysis using empirically based priors. *J. Roy. Statist. Soc. Ser. A* **172** 119–136. MR2655608

WELTON, N. J., SUTTON, A. J., COOPER, N. J., ABRAMS, K. R. and ADES, A. E. (2012). Evidence Synthesis in a Decision Modelling Framework. In *Evidence Synthesis for Decision Making in Healthcare* 138–150. Wiley, New York.

WHITE, I. R., BARRETT, J. K., JACKSON, D. and HIGGINS, J. P. T. (2012). Consistency and inconsistency in network meta-analysis: Model estimation using multivariate meta-regression. *Research Synthesis Methods* **3** 111–125.