Dias, S., Welton, N. J., Sutton, A. J., Caldwell, D. M., Lu, G., & Ades, A. E. (2011). NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials.(NICE DSU Technical Support Document in Evidence Synthesis; No. TSD4). National Institute for Health and Clinical Excellence.

NICE DSU TECHNICAL SUPPORT DOCUMENT 4: INCONSISTENCY IN NETWORKS OF EVIDENCE BASED ON RANDOMISED CONTROLLED TRIALS

REPORT BY THE DECISION SUPPORT UNIT

May 2011

(last updated April 2012)

Sofia Dias[1], Nicky J Welton[1], Alex J Sutton[2], Deborah M Caldwell[1], Guobing Lu[1], AE Ades[1]

[1] School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK
[2] Department of Health Sciences, University of Leicester, 2nd Floor Adrian Building, University Road, Leicester LE1 7RH, UK

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street Sheffield, S1 4DA;

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is a collaboration between the Universities of Sheffield, York and Leicester. We also have members at the University of Bristol, London School of Hygiene and Tropical Medicine and Brunel University. The DSU is commissioned by The National Institute for Health and Clinical Excellence (NICE) to provide a research and training resource to support the Institute's Technology Appraisal Programme. Please see our website for further information www.nicedsu.org.uk

ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

The NICE Guide to the Methods of Technology Appraisal[i] is a regularly updated document that provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The Methods Guide does not provide detailed advice on how to implement and apply the methods it describes. This DSU series of Technical Support Documents (TSDs) is intended to complement the Methods Guide by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in each topic area, and make clear recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE Technology Appraisals, whether manufacturers, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Dr Allan Wailoo

Director of DSU and TSD series editor.

[i] National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal, 2008 (updated June 2008), London.

Acknowledgements

The authors thank Ian White and Julian Higgins for drawing their attention to the difficulties in parameterising loop inconsistency models in the presence of multi-arm trials.

The DSU thanks Julian Higgins, Peti Juni, Eveline Nuesch, Steve Palmer, Georgia Salanti, Mike Spencer and the team at NICE, led by Zoe Garrett, for reviewing this document. The editor for the TSD series is Allan Wailoo.

The production of this document was funded by the National Institute for Health and Clinical Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the author only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

This report should be referenced as follows:

Dias, S., Welton, N.J., Sutton, A.J., Caldwell, D.M., Lu, G. & Ades, A.E. NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based on Randomised Controlled Trials. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

EXECUTIVE SUMMARY

In this document we describe methods to detect inconsistency in a network meta-analysis. Inconsistency can be thought of as a conflict between "direct" evidence on a comparison between treatments B and C, and "indirect" evidence gained from AC and AB trials. Like heterogeneity, inconsistency is caused by effect-modifiers, and specifically by an imbalance in the distribution of effect modifiers in the direct and indirect evidence. Checking for inconsistency therefore logically comes alongside a consideration of the extent of heterogeneity and its sources, and the possibility of adjustment by meta-regression or bias adjustment (see TSD3[1]). We emphasise that while tests for inconsistency must be carried out, they are inherently underpowered, and will often fail to detect it. Investigators must therefore also ask whether, if inconsistency is not detected, conclusions from combining direct and indirect evidence can be relied upon.

After an introduction outlining the document, Section 2 begins by defining inconsistency as a property of "loops" of evidence, and the Inconsistency Degrees of Freedom, which is approximately the number of independent loops of evidence. The relation between inconsistency and heterogeneity is explained, followed by a description of the difficulties created by multi-arm trials. In Section 3 we set out Bucher's original approach to assessing consistency in 3-treatment "triangular" networks, in larger "circuit" structures, and its extension to certain special structures where independent tests for inconsistencies can be created. Section 4 looks at detection of inconsistency in the general case, and we describe methods suitable for more complex networks. The first is the repeated use of the Bucher method to all the evidence loops in the network: this is a sound approach if it fails to detect inconsistencies, but may be difficult to interpret if inconsistencies are found. A second method is to compare the standard network consistency model to an "inconsistency", or unrelated mean effects, model. This is proposed as an efficient way of detecting inconsistency: sample WinBUGS code to implement fixed and random effects inconsistency models in a Bayesian framework is given in the Appendix, and results for two illustrative examples are provided. Section 4 closes with some comments on the relative merits of Bayesian and Frequentist methods with particular reference to sensitivity to prior distributions.

In Section 5 we review steps that can be taken to minimise the risk of drawing incorrect conclusions from indirect comparisons and network meta-analysis, which are the same steps that will minimise heterogeneity in pair-wise meta-analysis, and suggest some empirical

indicators that can provide reassurance. The question of how to respond to inconsistency is discussed in Section 6. Finally, the document ends with a set of brief summary statements and recommendations (Section 7).

CONTENTS

1. INTRODUCTION 8
2. NETWORK STRUCTURE: LOOPS, MULTI-ARM TRIALS, AND THE NUMBER OF INCONSISTENCIES. 10
2.1. EVIDENCE LOOPS 10
2.2. HETEROGENEITY VERSUS INCONSISTENCY 12
2.3. MULTI-ARM TRIALS 13
3. NETWORKS WITH INDEPENDENT TESTS FOR INCONSISTENCY 14
3.1. BUCHER METHOD FOR SINGLE LOOPS OF EVIDENCE 14
3.2. EXTENSION OF BUCHER METHOD TO NETWORKS WITH MULTIPLE LOOPS 16
4. METHODS FOR GENERAL NETWORKS 19
4.1. REPEAT APPLICATION OF THE BUCHER METHOD 19
4.2. INCONSISTENCY MODEL 20
4.2.1. Smoking Cessation Example 21
4.2.2. Thrombolytic Treatments Example 24
4.3. OTHER METHODS FOR DETECTING INCONSISTENCY 27
4.3.1. Variance measures of inconsistency 27
4.3.2. Node-Splitting 27
4.3.3. Bayesian and Frequentist approaches compared 28
5. MEASURES TO AVOID INCONSISTENCY 29
5.1. AVOIDING HETEROGENEITY 29
5.2. EMPIRICAL INDICATIONS OF HETEROGENEITY 31
6. RESPONSE TO INCONSISTENCY 31
7. SUMMARY AND CONCLUSIONS 33
7.1. CHOICE OF METHOD 33
7.2. PLACE OF INCONSISTENCY TESTING IN EVIDENCE SYNTHESIS. 33
7.3. RESPONSE TO INCONSISTENCY 34
8. REFERENCES 35
APPENDIX: WINBUGS CODE FOR ILLUSTRATIVE EXAMPLES 38
SMOKING CESSATION: RE MODEL, BINOMIAL LIKELIHOOD 38
THROMBOLYTIC TREATMENTS: FE MODEL, BINOMIAL LIKELIHOOD 39

TABLES, BOXES AND FIGURES

Table 1 Smoking example: Posterior summaries from Random Effects consistency and inconsistency models. Mean, standard deviation (sd), 95% Credible Interval (CrI) of relative treatment effects, and median of between-trial standard deviation (σ) on the log-odds scale; and posterior mean of the residual deviance (resdev), pD and DIC. 23
Table 2 Thrombolitics example: Posterior summaries, mean, standard deviation (sd) and 95% Credible Interval (CrI) on the log-odds ratio scale for treatments Y vs X for contrasts that are informed by direct evidence; and posterior mean of the residual deviance (resdev), pD and DIC, for the FE network meta-analysis and inconsistency models. 25

Box 1 Worked example: calculating Bucher's inconsistency estimate and approximate test for inconsistency. .16
Box 2 Worked example: chi-square test for inconsistency in the enuresis network.[16] 18

Figure 1 Possible treatment networks: treatments are represented by letters, lines connecting two treatments indicate that a comparison between these treatments has been made (in one or more RCTs) 11
Figure 2 Enuresis treatment network[16]: lines connecting two treatments indicate that a comparison between these treatments has been made (in one or more RCTs). 17
Figure 3 Plot of the individual data points' posterior mean deviance contributions for the consistency model (horizontal axis) and the inconsistency model (vertical axis) along with the line of equality. 23

Figure 4 Thrombolytics example network: lines connecting two treatments indicate that a comparison between these treatments (in one or more RCTs) has been made. The triangle highlighted in bold represents comparisons that have only been made in a three-arm trial. 24
Figure 5 Plot of the individual data points' posterior mean deviance contributions for the consistency model (horizontal axis) and the inconsistency model (vertical axis) along with the line of equality. Points which have a better fit in the inconsistency model have been marked with the trial number. 26

Abbreviations and definitions

DIC	Deviance information criterion
FE	Fixed effects
ICDF	Inconsistency degrees of freedom
MCMC	Markov chain Monte Carlo
RCT	Randomised controlled trial
RE	Random effects
TSD	Technical Support Document

1. INTRODUCTION

If pair-wise meta-analysis combines information from multiple trials comparing treatments A and B, network meta-analysis, also referred to as mixed treatment comparisons, or multiple treatment meta-analysis, combines information from randomised comparisons A vs B, A vs C, B vs C, A vs D and so on.[2-7] These methods all have the important property that they preserve randomisation.[8] Given a connected network of comparisons, network meta-analysis produces an internally coherent set of estimates of the efficacy of any treatment in the network relative to any other. A key assumption of network meta-analysis is that of evidence *consistency*. The requirement, in effect, is that in every trial *i* in the network, regardless of the actual treatments that were compared, the true effect δ_{iXY} of treatment Y relative to treatment X, is the same for every trial in a Fixed Effects (FE) model, i.e. $\delta_{iXY} = d_{XY}$, or exchangeable between-trials in a Random Effects (RE) model, i.e. $\delta_{iXY} \sim \text{Normal}(d_{XY}, \sigma^2)$. From this assumption the "consistency equations" can be deduced.[7,9,10] These assert, for example, that for any three treatments X, Y, Z, the fixed effects, or mean effects in a RE model, are related, as follows: $d_{YZ} = d_{XZ} - d_{XY}$.

Where doubts have been expressed about network meta-analysis, these have focussed on the consistency equations. This is because, unlike the exchangeability assumptions from which they are derived, which are notoriously difficult to verify, the consistency equations offer a clear prediction about relationships in the data that can be statistically tested. Note that consistency concerns the relation *between* the treatment contrasts, as distinct from heterogeneity, which concerns the variation between trials *within* each contrast (we use the term "contrast" to refer to a pair-wise comparison between two treatments).

Systematic empirical work on the validity of the consistency assumption has been limited, although it remains an active area. Song et al.[11] identified 44 datasets with triangle networks and used the Bucher method,[12] described in Section 3.1, to explore the existence of conflict between "direct" evidence on A vs B trials and "indirect" evidence inferred from AC and BC trials (see TSD2[9]). There were three cases where statistically significant inconsistencies (at $p<0.05$) were detected. However, the difference between the direct and indirect evidence was described by the authors as being clinically significant in only one of these cases. In this instance the doses used in the direct evidence were dissimilar to the doses used in the indirect evidence and the authors report that the inconsistency was resolved when the comparison was restricted to a similar range of doses. This example suggests a close relationship between between-trial heterogeneity and inconsistency between "direct" and "indirect" evidence. It is

also an example of a type of heterogeneity that might be removed by meta-regression (described in TSD3[1]). We return to these issues in later sections, but it should be understood that the logical place for an enquiry into consistency is alongside a consideration of heterogeneity and its causes, and where appropriate the reduction of heterogeneity through covariate adjustment (meta-regression) and bias adjustment (see TSD3[1]).

The first objective of this technical support document is to suggest robust methods for detection of inconsistency in evidence networks. However, it is important to note that failure to detect inconsistency does not imply consistency. As with other interaction effects, the evidence required to confidently rule out any but the most glaring inconsistency is seldom available, and in many cases, such as Indirect Comparisons, there is no way of testing the consistency assumptions at all. A second objective, therefore, is to clarify the measures that can be taken to minimise the risk of drawing incorrect conclusions from indirect comparisons and network meta-analysis, and to suggest some empirical indicators that might help assess what that risk might be.

The document takes the following form: firstly we discuss the effect of network structure on the number of potential inconsistencies, and we define Inconsistency Degrees of Freedom (ICDF) as the number of independent "loops" of evidence. We then discuss the relation between heterogeneity and inconsistency and the impact of multi-arm trials on the definitions of these terms. Section 3 then outlines Bucher's original approach to assessing consistency in 3-treatment, "triangular" networks of evidence[12] and how it can be extended to larger loops of evidence and to certain special structures. We then turn to the general case (Section 4), where it is not possible to carry out independent tests for each set of inconsistencies. First we consider the repeat application of the Bucher method to every triangle or closed loop in the network (Section 4.1). In Section 4.2 we propose a more general method, suitable for assessing consistency in any network, which compares the consistency model, on which a network meta-analysis suitable for coherent decision making must be based, to an "inconsistency model" (or unrelated mean effects model), in which the constraints forced by the consistency equations are removed. The latter is equivalent to having separate, unrelated, meta-analyses for every pair-wise contrast but with a common variance parameter in RE models. Sample code using the WinBUGS package[13] for fixed and random effects inconsistency models in a Bayesian Markov Chain Monte Carlo (MCMC) framework, is set out in the Appendix. In Section 4.3 we briefly review other methods for detecting inconsistency and the relative merits of Bayesian and Frequentist approaches.

In the remaining sections we suggest methods for avoiding inconsistency and note their relation to methods for avoiding heterogeneity (Section 5); we examine the question of how to respond to inconsistency if it is detected (Section 6); finally, we end with a set of brief summary comments and recommendations (Section 7).

The document should be seen as an adjunct to TSD2,[9] which sets out a generalised linear modelling framework for network meta-analysis, indirect comparisons and pair-wise meta-analysis. TSD2[9] explained how the same core model could be applied with different likelihoods and linking functions. It should be understood that this carries over entirely to the Bayesian models developed for inconsistency.

2. NETWORK STRUCTURE: LOOPS, MULTI-ARM TRIALS, AND THE NUMBER OF INCONSISTENCIES.

2.1. EVIDENCE LOOPS

We strongly recommend that the first step in checking for inconsistency should be to examine network diagrams carefully, as the structure can reveal particular features that may assist in the choice of analysis method. For example, it may be useful to use different line styles to highlight multi-arm trials, or to include information on the number of trials informing each comparison in the network diagram.

We begin by considering networks that consist only of two-arm trials, starting with a triangular network ABC (Figure 1(a)), where each edge represents the direct evidence comparing the treatments it connects. If we take treatment A as our reference treatment, a consistency model (TSD2[9]) has two basic parameters, say d_{AB} and d_{AC}, but we have data on three contrasts $d_{AB,}$ d_{AC} and d_{BC}. The latter, however is not an independent parameter, but is wholly determined by the two other parameters through the consistency equations. Setting aside the question of the number of trials informing each pair-wise contrast, we can see that there are two independent parameters to estimate and three sources of data. This generates one degree of freedom with which to detect inconsistency. More generally, we can define the "inconsistency degrees of freedom" (ICDF) as the number of pair-wise contrasts on which there is data, *N minus* the number of basic parameters, the latter being one less than the number of treatments, *nt*.[7] Thus if all trials are two-arm trials, the ICDF can be calculated from the number of treatments, *nt*, and the number of contrasts, *N*, on which there is

evidence as

$$\text{ICDF} = N - (nt - 1)$$

This accords with the common sense notion of "inconsistency", which has been at the heart of both previous methodological work[14] and empirical work,[11] which views it as a property of "loops" of evidence. Every additional *independent* loop in a network of two-arm trials represents one additional ICDF, and one further way in which potential inconsistency can be realised.

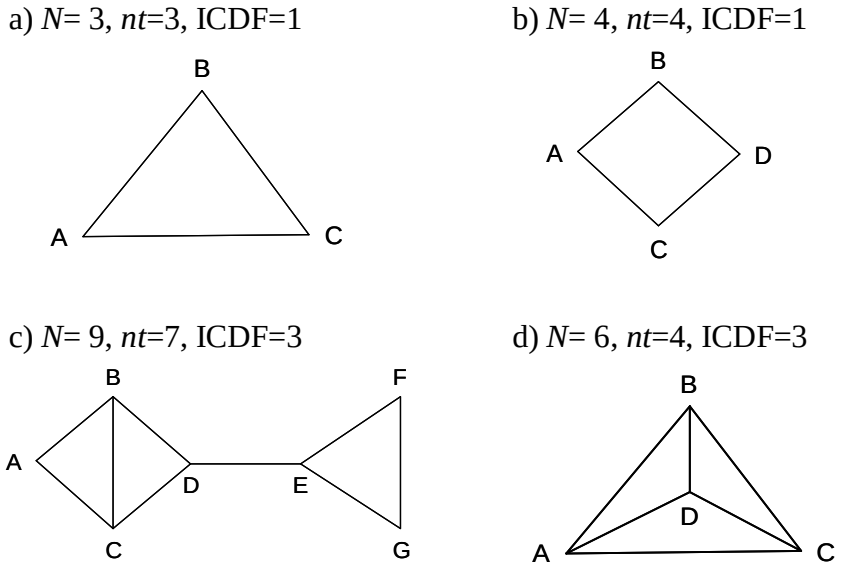

Figure 1 Possible treatment networks: treatments are represented by letters, lines connecting two treatments indicate that a comparison between these treatments has been made (in one or more RCTs).

The "square" network in Figure 1(b) consists of AB, AC, BD and CD trials. Here we have 1 inconsistency degree of freedom, as there are N=4 independent pieces of evidence, nt=4 treatments, and therefore nt-1=3 parameters in a consistency model, giving ICDF=4-(4-1)=1. In Figure 1(c) there are N=9 contrasts on which there is evidence, nt=7 treatments and therefore 6 parameters, giving ICDF=3. Note that the ICDF is equal to the number of *independent* loops. In Figure 1(c) there are two separate structures where inconsistency could be detected, first the triangle EFG, and second the square ABCD. In the square, one could count a total of 3 loops: ABC, BCD, and ABCD. However, there are only two *independent* loops in this second part of the structure, because if we know all the edges of any two loops, we would immediately know the edges of the third. Therefore there can be only two inconsistencies in the ABCD square.

Similarly, in Figure 1(d) one can count a total of 7 loops: four three-treatment loops (ACD, BCD, ABD, ABC), and three four-treatment loops (ABCD, ACDB, CABD). But there are only three independent loops, and one can confirm that with *N*=6, *nt*=4, and ICDF=3. It is not possible to specify *which* loops are independent, only how many there are.

2.2. HETEROGENEITY VERSUS INCONSISTENCY

Although we have characterised heterogeneity as between-trial variation *within* treatment contrasts, and inconsistency as variation *between* contrasts, the difference is subtle. Eventually, all heterogeneity in relative treatment effects is a reflection of an interaction between the treatment effect and a trial level variable (see TSD3[1]). To put it another way, heterogeneity reflects the presence of effect-modifiers. If we now consider the case of a triangular network (Figure 1(a)), if the effect-modifiers are present in the AB and AC trials, but not the BC trials, then we may observe "inconsistency". However, if the effect modifiers are more evenly balanced across the network, then it is more likely that we will not find inconsistency.

One might try to distinguish an imbalance occurring by chance from one due to an *inherent inconsistency*. Suppose, for example, patients in BC trials are inherently different because they cannot take treatment A, and this is in addition associated with a different treatment effect. These patients have perhaps already failed on A, or have had adverse side-effects, or they have markers that counter-indicate A. At the other extreme, it might be that the inconsistency was just the result of chance, in which the BC trials just happened to be those with an effect modifier. But between these extremes one might imagine that the BC trials are just *more likely* to concern patients who could not take A. In either case, heterogeneity and inconsistency are both reflections of a treatment effect modifier. Inconsistency is a special case of heterogeneity where there is an association between the effect modifier and the set of treatment contrasts. There is an immediate implication that inconsistency due to a chance imbalance in the distribution of effect modifiers, should become increasingly less likely to occur as the number of trials on each contrast increases.

Inconsistency checking is closely related to cross-validation for outlier detection (see TSD3[1]). However, in the presence of heterogeneity, cross-validation is based on the predictive distributions of effects, while the concept of inconsistency between "direct" and "indirect" evidence refers to inconsistency in expected (i.e. mean) effects and is therefore based on the posterior distributions of the mean effects. This will frequently result in a situation where, in

a triangular loop, in which one edge consists of a singleton trial, we may find inconsistency in the expected effects, while cross-validation fails to show that the singleton trial is an outlier. Of course, technically there is no reason why inconsistency checks cannot be made on the predictive distributions of the treatment effects, and this may be desirable if inference is to be based on the predictive treatment effects from a network meta-analysis. See TSD3[1] for further details.

2.3. MULTI-ARM TRIALS

When multi-arm trials are included in the network, that is trials with more than two arms, the definition of inconsistency becomes more complex. A 3-arm trial provides evidence on all three edges of an ABC triangle, and yet it cannot be inconsistent. In other words, although trial i estimates three parameters, $\delta_{i,AB}$, $\delta_{i,AC}$, $\delta_{i,BC}$, only two are independent because $\delta_{i,BC} = \delta_{i,\mathrm{AC}} - \delta_{i,\mathrm{AB}}$. There can therefore be no inconsistency within a 3-arm trial. Similarly, if *all* the evidence was from 3-arm trials on the same three treatments, there could be no inconsistency in the network, only between-trials heterogeneity.

The difficulty in defining inconsistency comes when we have both 2- and 3-arm trial evidence, for example AB, AC, BC and ABC trials. This raises two questions. The first question is: do we wish to consider evidence from an ABC trial as potentially inconsistent with evidence from an AB trial? Although one of the very first treatments of these data structures[3] *did* regard this as a form of inconsistency, it appears that in practice AB evidence from AB, ABC, and ABD studies has been synthesised without any special consideration being given to the presence of the further arms. In systematic reviews and meta-analyses of AB evidence, when multi-arm trials are available, only the AB arms are included and any further arms discarded. In all the published meta-analyses and systematic reviews, the issue of whether the presence of multiple arms might be associated with greater heterogeneity, i.e. inconsistency, has never been raised. For this reason, in everything that follows, inconsistency will be used only to refer to evidence loops, and loops of evidence that are potentially inconsistent can only arise from structures in which there are three distinct trials or sets of trials.

A second issue is parameterisation. AB, AC and BC evidence arises from 3 separate sources and can be inconsistent. But suppose there are *also* ABC trials? We know that these can contribute *independent* evidence on only two treatment effects, but it is not clear *which* two to

choose. One way to see what the implications are is to consider between-trial heterogeneity. If we are interested in the heterogeneity of the AB, AC and BC effects, we might begin by looking at each set of two-arm trials separately. But how should the 3-arm trials be used? They contribute further information on between-trials heterogeneity, but strictly speaking, they can only provide independent information on *two* of the three contrasts. Clearly, the choice will have an impact on both estimates of between-trial heterogeneity *and* the detection of inconsistency.

Thus, where there are mixtures of 2-arm and multi-arm trials, our definition of inconsistency as arising in loops creates inherent technical difficulties that cannot, as far as is known, be avoided. We return to the issue as we explain approaches to detecting inconsistency. The solutions we suggest are simple and practical, and, while still not entirely satisfactory, they are predicated on the assumption that the majority of trials are 2-arm trials and there is unlikely to be any material impact on detection of inconsistency. Conversely, if the proportion of multi-arm trials becomes higher, the distinction between heterogeneity and inconsistency, conceptualised as systematic differences between "direct" and "indirect" evidence, becomes harder to draw and less relevant.

3. NETWORKS WITH INDEPENDENT TESTS FOR INCONSISTENCY

A key consideration in consistency assessment in networks of evidence is whether independent tests for inconsistency can be constructed. Below we show how to construct independent tests, and explain the circumstances where this is possible. Section 4 sets out methods for the more general case which can be applied to any network. However, the methods in the following section should be used wherever possible as they provide the simplest, most complete, and easiest to interpret analyses of inconsistency possible.

3.1. BUCHER METHOD FOR SINGLE LOOPS OF EVIDENCE

The first and simplest method for testing consistency of evidence is due to Bucher et al.[12] It is essentially a "two-stage" method. The first stage is to separately synthesise the evidence in each pair-wise contrast; the second stage is a test of whether direct and indirect evidence are in conflict. A "direct" estimate of the C vs. B effect, $\hat{d}_{BC}^{Dir}$, is to be compared to an "indirect" estimate, $\hat{d}_{BC}^{Ind}$, formed from the AB and AC direct evidence

$$\hat{d}_{BC}^{Ind} = \hat{d}_{AC}^{Dir} - \hat{d}_{AB}^{Dir} \quad (1)$$

We assume that the direct estimates can either be estimates from individual trials, or they can be from pair-wise meta-analyses, whether fixed or random effects. Attaching to each direct estimate is a variance, for example Var($\hat{d}_{BC}^{Dir}$). As the direct estimates are statistically independent we have

$$Var(\hat{d}_{BC}^{Ind}) = Var(\hat{d}_{AC}^{Dir}) + Var(\hat{d}_{AB}^{Dir})$$

An estimate of the inconsistency, ω, can be formed by simply subtracting the direct and indirect estimates:

$$\hat{\omega}_{BC} = \hat{d}_{BC}^{Dir} - \hat{d}_{BC}^{Ind}$$
$$\text{Var}(\hat{\omega}_{BC}) = \text{Var}(\hat{d}_{BC}^{Dir}) + \text{Var}(\hat{d}_{BC}^{Ind}) = \text{Var}(\hat{d}_{BC}^{Dir}) + \text{Var}(\hat{d}_{AB}^{Dir}) + \text{Var}(\hat{d}_{AC}^{Dir})$$

An approximate test of the null hypothesis that there is no inconsistency can be obtained by referring $z_{BC} = \dfrac{\hat{\omega}_{BC}}{\sqrt{\text{Var}(\hat{\omega}_{BC})}}$ to the standard normal distribution. Box 1 provides a worked example.

It is easy to confirm that it makes no difference whether we compare the direct BC evidence to the indirect evidence formed through AB and AC, or compare the direct AB evidence to the indirect AC and BC, or the AC with the AB and BC. The absolute values of the inconsistency estimates will be identical, and will always have the same variance. This agrees with the intuition that, in a single loop, there can only be one inconsistency, as discussed above. Needless to say, the method can only be applied to 3 *independent* sources of data. Three-arm trials cannot be included: because they are internally consistent they will reduce the chances of detecting inconsistency (see Section 4.3 for further notes).

Box 1

Meta-analyses of (direct) randomised controlled trial evidence on three treatments for virologic suppression in patients with HIV[15] produced the following estimates on the log odds ratio (OR) scale, corresponding to the network in Figure 1(a):

	ln(OR)	standard error of ln(OR)
$\hat{d}_{AB}^{Dir}$	2.79	0.56
$\hat{d}_{AC}^{Dir}$	1.42	0.34
$\hat{d}_{BC}^{Dir}$	0.47	0.10

The indirect estimate for the relative effect of treatment C vs B was $\hat{d}_{BC}^{Ind} = 1.42 - 2.79 = -1.37$ with $Var(\hat{d}_{BC}^{Ind}) = 0.56^2 + 0.34^2 = 0.429$. Comparing this to the direct estimate $\hat{d}_{BC}^{Dir} = 0.47$, we have inconsistency estimate $\hat{\omega}_{BC} = 0.47 - (-1.37) = 1.84$ with $Var(\hat{\omega}_{BC}) = 0.10^2 + 0.429$. Then, $z_{BC} = 1.84 \Big/ \sqrt{0.439} = 2.78$ indicating there is evidence of inconsistency (p-value < 0.01).

Box 1 Worked example: calculating Bucher's inconsistency estimate and approximate test for inconsistency.

This method generalises naturally to the "square" network in Figure 1(b) which, like the triangle and any other simple "circuit" structure, has ICDF=1. An indirect estimate of any edge can be formed from the remaining edges, and the variance of the inconsistency term is the sum of the variances of all the comparisons. As the number of edges in the loop increases it becomes less and less likely that a real inconsistency will be detected due to the higher variance calculated for the inconsistency estimate.

3.2. EXTENSION OF BUCHER METHOD TO NETWORKS WITH MULTIPLE LOOPS

Figure 2 shows an example of a Cochrane Overview of Reviews[16,17] where there is data on 10 contrasts involving 8 treatments for childhood enuresis, and therefore ICDF=3. The special feature of this network is that all the inconsistencies can be seen as concerning estimates of the Alarm vs No treatment effect. In particular, there are four independent estimates of this parameter: one direct estimate and three indirect estimates via Psychological therapy, Imipramine and Dry bed training, respectively. In this situation an approximate chi-square (χ^2) test of inconsistency can be constructed on 3 degrees of freedom[16] which is an extension of the Bucher method – see Box 2.

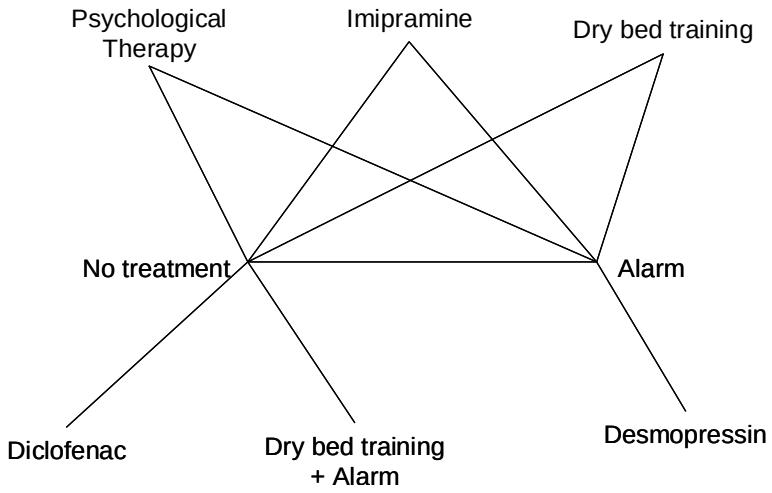

Figure 2 Enuresis treatment network[16]: lines connecting two treatments indicate that a comparison between these treatments has been made (in one or more RCTs).

Figure 1(c) represents a further pattern where the inconsistency analysis can be broken down into separate independent elements. Here there are a total of three independent loops and we anticipate ICDF=3. This is confirmed by observing that there are 9 contrasts on which there is evidence, 7 treatments, and therefore 6 parameters to estimate. In this case, one inconsistency relates to the loop EFG where there are two sources of evidence on any edge, while the other concerns the edge BC, on which there are 3 independent sources of evidence, one direct and two indirect. The most convenient way to analyse inconsistency in this structure is to break the problem down into the two separate and unrelated components. First, inconsistency in the EFG "triangle" can be examined using the simple Bucher approach (Section 3.1). Second, consistency between the three sources of evidence on the BC edge can be examined by calculating a statistic to refer to a χ^2_2 distribution, as described in Box 2. These two tests provide a complete analysis of the inconsistency in this network.

The Bucher method for triangle structures, and its extension to larger loops and to χ^2 tests, are all based on two-arm trials. Inclusion of multi-arm trials will lower the power of these tests to detect inconsistency. Our suggestion is that they are excluded entirely.

Box 2

To analyse the potential inconsistencies in the network in Figure 2, we consider the evidence on each of the relevant edges obtained from separate FE pair-wise meta-analyses, on the log relative risk (RR) scale:

			Direct estimates	
Control	**Treatment**	**Number of RCTs**	**ln(RR)**	**variance of ln(RR)**
No treatment	Alarm	14	-0.968	0.006
No treatment	Psychological Therapy	3	-0.371	0.012
Alarm	Psychological Therapy	3	0.386	0.038
No treatment	Imipramine	11	-0.261	0.001
Alarm	Imipramine	3	0.315	0.009
No treatment	Dry Bed Training	2	-0.198	0.012
Alarm	Dry Bed Training	3	-0.285	0.071

We need to compare the direct estimate of the relative effect of Alarm vs No treatment, $\hat{d}_1 = -0.968$ with the three possible indirect estimates, $\hat{d}_2, \hat{d}_3, \hat{d}_4$, given below:

	ln(RR)	**variance of ln(RR)**
Direct estimate of Alarm vs No treatment effect		
	-0.968	0.006
Indirect estimates of Alarm vs No treatment effect		
via Psychological Therapy	-0.757	0.050
via Imipramine	-0.576	0.010
via Dry Bed Training	0.087	0.083

Given 4 independent estimates $\hat{d}_1, \hat{d}_2, \hat{d}_3, \hat{d}_4$ of the relative treatment effect of Alarm vs No treatment, and their variances V_1, V_2, V_3 and V_4, an average treatment effect $\tilde{d}$ is estimated by inverse variance weighting as $\tilde{d} = \sum_{i=1}^{4} w_i \hat{d}_i \Big/ \sum_{i=1}^{4} w_i = -0.776$, where $w_i = 1/V_i$, i=1,2,3,4. An approximate χ^2 statistic, is given by $T = \sum_{i=1}^{4} w_i \left(\hat{d}_i - \tilde{d}\right)^2 = 18.8$. Referring T to a χ^2_3 distribution (the degrees of freedom are given by the number of independent estimates minus 1) suggests that there is evidence of inconsistency (p-value < 0.01). For more details see Caldwell et al.[16]

Box 2 Worked example: chi-square test for inconsistency in the enuresis network.[16]

4. METHODS FOR GENERAL NETWORKS

4.1. REPEAT APPLICATION OF THE BUCHER METHOD

Figure 1(d) shows a 4 treatment network in which there is data on every contrast and 3 possible inconsistencies. The difference between the networks in Figure 1(d) and Figure 1(c) is that in the former there are four three-treatment loops (ACD, BCD, ABD, ABC), and three four-treatment loops ABCD, ACDB, CABD, but we have already noted that the loops are not statistically independent. The further difficulty is that it is not possible to construct a set of independent tests to examine the 3 inconsistencies.

Another approach is to apply the Bucher method to each of the seven loops in the network in turn, which has the advantage of being simple to implement. However, when this is done, the number of loops, and hence the number of tests, will far exceed the number of inconsistencies that the network can actually have. For example in a network where N=14, nt=6 and ICDF=9,[18] the Bucher method was applied to every three-way, four-way and five-way loop leading to 20 tests of inconsistency, none of which suggested that inconsistency might be present. In another example[19] N=42, nt=12 and ICDF=31 is the maximum number of inconsistencies. However, repeated use of the Bucher method on each of the three-way loops in this network gives 70 estimates of inconsistency for the response outcome and 63 estimates for the acceptability outcome. In total six loops showed statistically significant inconsistency and the authors concluded that this was compatible with chance as 133 separate tests for inconsistency were performed. However, this conclusion could be questioned on the grounds that the 133 tests were not independent – there could not be more than 62 independent tests, and even this assumes that the two outcomes are unrelated.

Each application of the Bucher test is a valid test at its stated significance level. *If no inconsistencies are found,* at say $p<0.05$, when applying the test to all loops in the network, one can correctly claim to have failed to reject the null hypothesis of no inconsistency at this level, *or higher*. This is not necessarily very reassuring because of the inherently high variance of indirect evidence, especially in multi-sided loops. (The situation is analogous to concluding "no difference" from a small, under-powered trial). This is a problem common to all methods for detecting inconsistency, of course, not just the Bucher approach.

Difficulties in the interpretation of statistical tests arise if any of the loops show significant inconsistency, at say a $p<0.05$ level. One cannot immediately reject the null hypothesis at this level because a certain degree of multiple testing has taken place, and adjustment of

significance levels would need to be considered. However, because the tests are not independent, calculating the correct level of adjustment becomes a complex task. Further, in networks with multiple treatments, the total number of triangular, quadrilateral, and higher order loops may be extremely large.

The presence of multi-arm trials again causes complications. Our suggestion is that when a test on a loop ABC is being constructed, evidence from 3-arm ABC trials is excluded. However, ABC evidence on AB *should* be included when testing, for example, the ABD loop. Similarly ABCD trials would be excluded from tests on the ABCD loop, but included in studies of the ABCE, BCDE loops, and so on.

4.2. INCONSISTENCY MODEL

Instead of the repeat application of the Bucher method, in complex networks where independent tests cannot be constructed, we propose that the standard consistency model that was presented in TSD2[9] is compared with an *inconsistency* model. In the consistency model a network with *nt* treatments, A, B, C,… defines *nt*-1 "basic" parameters[20] d_{AB}, d_{AC},… which estimate the effects of all treatments relative to treatment A, chosen as the reference treatment. Prior distributions are placed on these parameters. All other contrasts are derived "functional" parameters, which can be defined as functions of the basic parameters by making the consistency assumption.

In the inconsistency model proposed here, each of the *N* contrasts for which evidence is available, represents a separate, unrelated, basic parameter to be estimated: no consistency is assumed. So, for a network such as that in Figure 1(b), the consistency model would estimate three relative treatment effect parameters, d_{AB}, d_{AC}, d_{AD}, from evidence on four contrasts (ICDF=1). The functional parameter $d_{CD} = d_{AD} - d_{AC}$ is defined from the basic parameters. The inconsistency model would estimate 4 relative treatment effect parameters, d_{AB}, d_{AC}, d_{BD}, d_{CD}, from the evidence on these 4 contrasts, without assuming any relationship between the parameters. Similarly for the network in Figure 1(d), the consistency model would estimate three relative treatment effect parameters, d_{AB}, d_{AC}, d_{AD}, from evidence on six contrasts (ICDF=3), while the inconsistency model would estimate 6 unrelated relative treatment effect parameters, d_{AB}, d_{AC}, d_{AD}, d_{BC}, d_{BD}, d_{CD}.

More formally, suppose we have a set of *M* trials comparing *nt*=4 treatments, A, B, C and D in any connected network. In a RE model the study-specific treatment effects for a study

comparing a treatment *X* to another treatment *Y*, $\delta_{i,XY}$, are assumed to follow a normal distribution

$$\delta_{i,XY} \sim N(d_{XY}, \sigma^2) \quad \text{for } i = 1,...,M \tag{2}$$

In a consistency model, *nt*-1=3 basic parameters are given vague priors: d_{AB}, d_{AC}, d_{AD} ~N(0,100^2), and the consistency equations define all other possible contrasts as:

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

In a RE inconsistency model, each of the mean treatment effects in equation (2) is treated as a separate (independent) parameter to be estimated, sharing a common variance σ^2. So, for the network in Figure 1(d), the six treatment effects are *all* given vague priors: d_{AB}, d_{AC}, d_{AD}, d_{BC}, d_{BD}, d_{CD} ~ N(0,100^2).

In a FE inconsistency model no shared variance parameter needs to be considered. The inconsistency model is then equivalent to performing completely separate pairwise meta-analysis of the data. However, fitting an inconsistency model to all the data has the advantage of easily accommodating multi-arm trials as well as providing a single global measure of model fit.

When multi-arm trials are included in the evidence, the inconsistency model can have different parameterisations depending on which of the multiple contrasts defined by a multi-arm trial are chosen (see Section 2.3). For example, a three-arm trial ABC can inform the AB and AC independent effects, or it can be chosen to inform the AB and BC effects (if B was the reference treatment), or the AC and BC effects (with C as reference). The code presented in the Appendix arbitrarily chooses the contrasts relative to the "first" treatment in the trial. Thus, ABC trials inform the AB and AC contrasts, BCD trials inform BC and BD etc. Choice of parameterisation will affect parameter estimates, and the tests of inconsistency. The presence of multi-arm trials also complicates calculation of the ICDF. Because ICDF corresponds to the number of independent loops, if a loop is formed from a multi-arm trial alone, it is not counted as an independent loop and must therefore be discounted from the total ICDF.[7]

4.2.1. Smoking Cessation Example

Twenty-four studies, including two three-arm trials, compared four smoking cessation counselling programs and recorded the number of individuals with successful smoking

cessation at 6-12 months. All possible contrasts were compared, forming the network in Figure 1(d), where A= no intervention, B= self-help, C= individual counselling and D= group counselling. This dataset has been previously analysed by Hasselblad[4] and Lu and Ades[7] among others. The consistency model, with either random or fixed effects, can be fitted using the code presented TSD2[9] (Programs 1(c) and 1(d) in the Appendix, respectively).

We now contrast the previous results with a RE inconsistency model estimating six independent mean treatment effects, as described in the previous section. The code is presented in the Appendix. Both the consistency and inconsistency models have a shared variance for the random effects distributions. Results for both models are presented in Table 1, along with residual deviance and DIC, measures of model fit[21] introduced in TSD2.[9] These are based on 100,000 iterations on three chains after a burn-in period of 20,000 for the consistency model and 100,000 iterations on three chains after a burn-in of 30,000 for the inconsistency model. The heterogeneity estimates, the posterior means of the residual deviance and the DICs are very similar for both models, although both are lower for the consistency model. Comparison between the deviance and DIC statistics of the consistency and inconsistency models provides an "omnibus" test of consistency.

Plotting the posterior mean deviance of the individual data points in the inconsistency model against their posterior mean deviance in the consistency model (Figure 3) provides information that can help identify the loops in which inconsistency is present. We expect each data point to have a posterior mean deviance contribution of about 1, with higher contributions suggesting a poorly fitting model.[21] In this example, the contributions to the deviance are very similar and close to 1, for both models. Two points have a higher than expected posterior mean deviance – these are the arms of two trials which have a zero cell – but the higher deviance is seen in both consistency and inconsistency models. In general, trial-arms with zero cells will have a high posterior mean of the residual deviance as the model will never predict a zero cell exactly. The parameter estimates are also similar for both models and there is considerable overlap in the 95% credible intervals. This suggests no evidence of inconsistency in the network.

Table 1 Smoking example: Posterior summaries from Random Effects consistency and inconsistency models. Mean, standard deviation (sd), 95% Credible Interval (CrI) of relative treatment effects, and median of between-trial standard deviation (σ) on the log-odds scale; and posterior mean of the residual deviance (resdev), pD and DIC.

	Network meta-analysis* (consistency model)			Inconsistency Model		
	Mean/Median	sd	CrI	Mean/Median	sd	CrI
d_{AB}	0.49	0.40	(-0.29, 1.31)	0.34	0.58	(-0.81, 1.50)
d_{AC}	0.84	0.24	(0.39, 1.34)	0.86	0.27	(0.34, 1.43)
d_{AD}	1.10	0.44	(0.26, 2.00)	1.43	0.88	(-0.21, 3.29)
d_{BC}	0.35	0.41	(-0.46, 1.18)	-0.05	0.74	(-1.53, 1.42)
d_{BD}	0.61	0.49	(-0.34, 1.59)	0.65	0.73	(-0.80, 2.12)
d_{CD}	0.26	0.41	(-0.55, 1.09)	0.20	0.78	(-1.37, 1.73)
σ	0.82	0.19	(0.55, 1.27)	0.89	0.22	(0.58, 1.45)
resdev†	54.0			53.4		
pD	45.0			46.1		
DIC	99.0			99.5		

* d_{BC}, d_{BD}, d_{CD} calculated using the consistency equations
† compare to 50 data points

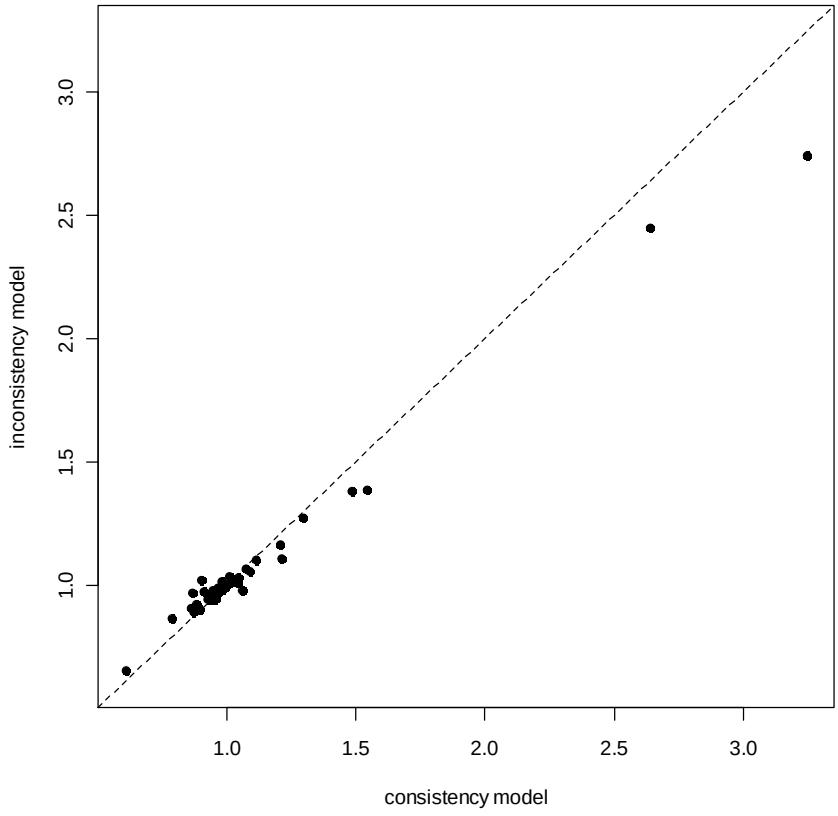

Figure 3 Plot of the individual data points' posterior mean deviance contributions for the consistency model (horizontal axis) and the inconsistency model (vertical axis) along with the line of equality.

4.2.2. Thrombolytic Treatments Example

The number of deaths in 30 or 35 days and the number of patients in each treatment arm from a dataset consisting of 50 trials comparing 8 thrombolytic drugs: streptokinase (SK, coded 1), alteplase (t-PA, 2), accelerated alteplase (Acc t-PA, 3), streptokinase plus alteplase (SK+t-PA, 4), reteplase (r-PA, 5), tenocteplase (TNK, 6), urokinase (UK, 8), anistreptilase (ASPAC, 9); and per-cutaneous transluminal coronary angioplasty (PTCA, 7), following acute myocardial infarction were included. This is a set of treatments defined in two comprehensive systematic reviews,[22,23] except that the trials involving either ASPAC or UK were excluded from the original analysis because these treatments were no longer available in the United Kingdom. Network meta-analyses of different subsets of this network have been previously studied[2,7] and consistency in the complete network has been previously assessed.[24]

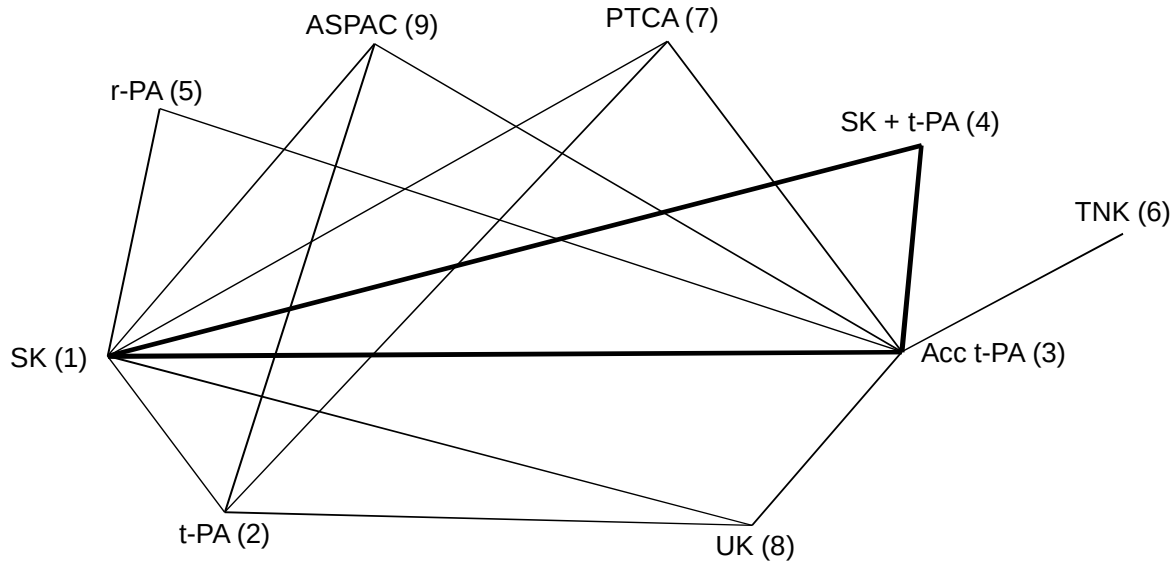

Figure 4 Thrombolytics example network: lines connecting two treatments indicate that a comparison between these treatments (in one or more RCTs) has been made. The triangle highlighted in bold represents comparisons that have only been made in a three-arm trial.

Figure 4 represents the treatment network, where each edge represents a direct comparison of the two treatments being connected. We can see that not all treatment contrasts have been compared in a trial, as there are treatment pairs which are not connected. There are 9 treatments in total and information on 16 pairwise comparisons, which would suggest an ICDF of eight. However, there is one loop, SK, Acc t-PA, SK+t-PA (highlighted in bold)

which is only informed by a three-arm and therefore cannot contribute to the number of possible inconsistencies. Discounting this loop gives ICDF=7.[7]

A FE network meta-analysis (consistency model) with a binomial likelihood and logit link (see TSD2[9]) was fitted to the data, taking SK as the reference treatment, i.e. the eight treatment effects relative to SK are the basic parameters and have been estimated, while the remaining relative effects were obtained from the consistency assumptions. A FE inconsistency model was also fitted which estimated 15 independent mean treatment effects (code presented in the Appendix). Results for the 15 contrasts on which there is information for both models are presented in Table 2, along with measures of model fit. These are based on 50,000 iterations on two chains after a burn-in period of 50,000 for the consistency model and 50,000 iterations on three chains after a burn-in of 20,000 for the inconsistency model.

Table 2 Thrombolitics example: Posterior summaries, mean, standard deviation (sd) and 95% Credible Interval (CrI) on the log-odds ratio scale for treatments Y vs X for contrasts that are informed by direct evidence; and posterior mean of the residual deviance (resdev), pD and DIC, for the FE network meta-analysis and inconsistency models.

treatments		Network meta-analysis* (consistency model)			Inconsistency Model		
X	Y	mean	sd	CrI	mean	sd	CrI
SK	t-PA	0.002	0.030	(-0.06,0.06)	-0.004	0.030	(-0.06,0.06)
SK	Acc t-PA	-0.177	0.043	(-0.26,-0.09)	-0.158	0.049	(-0.25,-0.06)
SK	SK + t-PA	-0.049	0.046	(-0.14,0.04)	-0.044	0.047	(-0.14,0.05)
SK	r-PA	-0.124	0.060	(-0.24,-0.01)	-0.060	0.089	(-0.23,0.11)
SK	PTCA	-0.173	0.077	(-0.32,-0.02)	-0.665	0.185	(-1.03,-0.31)
SK	UK	-0.476	0.101	(-0.67,-0.28)	-0.369	0.518	(-1.41,0.63)
SK	ASPAC	-0.203	0.221	(-0.64,0.23)	0.005	0.037	(-0.07,0.08)
t-PA	PTCA	0.016	0.037	(-0.06,0.09)	-0.544	0.417	(-1.38,0.25)
t-PA	UK	-0.180	0.052	(-0.28,-0.08)	-0.294	0.347	(-0.99,0.37)
t-PA	ASPAC	-0.052	0.055	(-0.16,0.06)	-0.290	0.361	(-1.01,0.41)
Acc t-PA	r-PA	-0.126	0.067	(-0.26,0.01)	0.019	0.066	(-0.11,0.15)
Acc t-PA	TNK	-0.175	0.082	(-0.34,-0.01)	0.006	0.064	(-0.12,0.13)
Acc t-PA	PTCA	-0.478	0.104	(-0.68,-0.27)	-0.216	0.119	(-0.45,0.02)
Acc t-PA	UK	-0.206	0.221	(-0.64,0.23)	0.146	0.358	(-0.54,0.86)
Acc t-PA	ASPAC	0.013	0.037	(-0.06,0.09)	1.405	0.417	(0.63,2.27)
resdev†		105.9			99.7		
pD		58			65		
DIC		163.9			164.7		

* All relative treatment effects not involving SK were calculated using the consistency equations

† compare to 102 data points

Although the inconsistency model has a lower posterior mean of the residual deviance and hence is a better fit to the data, the DICs are very similar for both models. This is because the inconsistency model has seven more parameters than the network meta-analysis model, as can be seen in the difference in the values of pD. A plot of the individual data points' posterior mean deviance contribution in each of the two models is presented in Figure 5. In this example, four data points show a much lower value of the posterior mean deviance in the inconsistency model, suggesting that a consistency model does not fit these points well. These four points corresponding to the two arms of trials 44 and 45, which were the only two trials comparing Acc t-PA to ASPAC. Comparing the posterior estimates of the treatment effects of ASPAC vs Acc t-PA in Table 2 we can see that these differ between the consistency and inconsistency models with no overlap in the 95% credible intervals. The fact that the two trials on this contrast give similar results to each other, which are in conflict with the remaining evidence, support the notion that there is a systematic inconsistency.

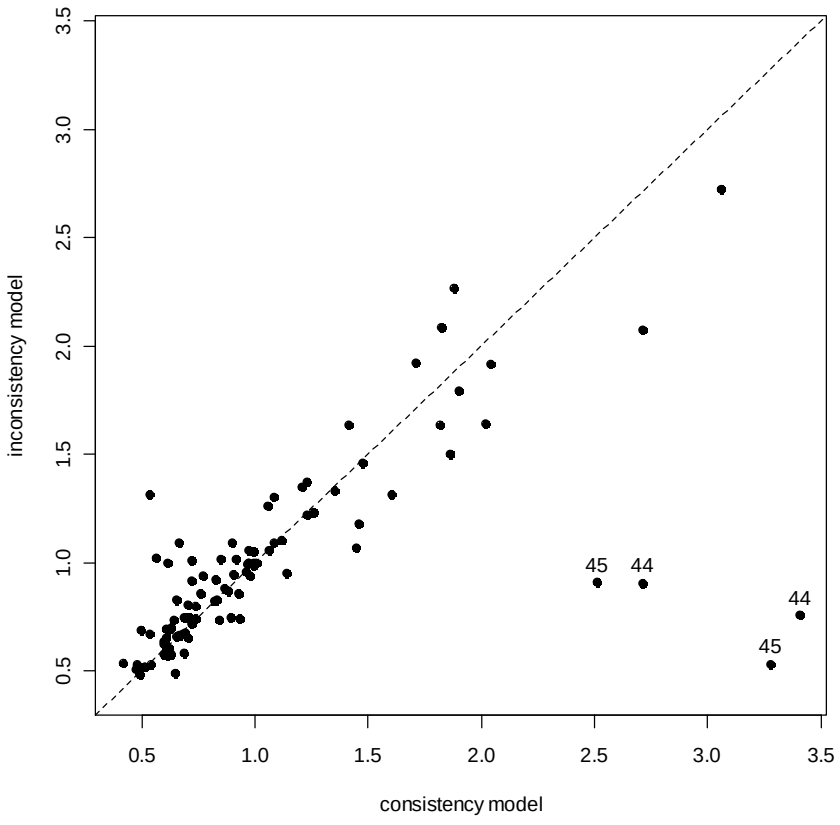

Figure 5 Plot of the individual data points' posterior mean deviance contributions for the consistency model (horizontal axis) and the inconsistency model (vertical axis) along with the line of equality. Points which have a better fit in the inconsistency model have been marked with the trial number.

4.3. OTHER METHODS FOR DETECTING INCONSISTENCY

4.3.1. Variance measures of inconsistency

In the inconsistency models described above a different basic parameter represents each contrast. For example, in the four treatment network in Figure 1(d) we have a 6-parameter model to estimate, rather than the 3 parameter consistency model. One can re-parameterise the 6 parameter inconsistency model so that instead of 6 treatment effect parameters (d_{AB}, d_{AC}, d_{AD}, d_{BC}, d_{BD}, d_{CD}) we have (d_{AB}, d_{AC}, d_{AD}, ω_{BC}, ω_{BD}, ω_{CD}) where:

$$\begin{cases} \omega_{BC} = d_{BC} - (d_{AC} - d_{AB}) \\ \omega_{BD} = d_{BD} - (d_{AD} - d_{AB}) \\ \omega_{CD} = d_{CD} - (d_{AD} - d_{AC}) \end{cases}$$

The $\omega_{BC}, \omega_{BD}, \omega_{CD}$ parameters are the "inconsistencies" between the "direct" and "indirect" evidence on these three edges. However, rather than considering the three inconsistency parameters as unrelated, we might assume that they all come from a random distribution, for example $\omega_{XY} \sim N(0, \sigma_\omega^2)$. This model has been proposed in other contexts by both Lumley,[14] who named the additional variance term "incoherence variance" and by Lu and Ades[7] who named it "inconsistency variance". Both authors suggest that this additional between-contrast variance can serve as a measure of inconsistency. We do not recommend this however, because measures of variance will have very wide credible intervals unless the ICDF is extremely high. Even then, large numbers of large trials on each contrast would be required to obtain a meaningful estimate. Furthermore, where there is a single loop (ICDF=1) it should be impossible to obtain any estimate of σ_ω^2. In spite of this a number of published applications of the Lumley[14] model have reported estimates of inconsistency variance in networks consisting of only a single loop,[25,26] and it seems likely that the model has not always been implemented in a way that takes account of the number of inconsistencies. See Salanti et al.[27] for further comments on this issue.

4.3.2. Node-Splitting

A more sophisticated approach, which needs to be implemented in a Bayesian MCMC framework, is "node splitting".[24] This technique allows the user to split the information contributing to estimates of a parameter (node), say, d_{XY} into two distinct components: the "direct" based on all the XY data (which may come from XY, XYZ, WXY trials) and the "indirect" based on all the remaining evidence. The process can be applied to any contrast

(node) in the network, and in networks of any complexity. Like the inconsistency model suggested above, which can be seen as a node-splitting approach in which a number of nodes are split at the same time, a shared variance term solves the difficulties created in a RE model when some contrasts are supported by only one or two trials. Node-splitting is a powerful and robust method that can be recommended as a further option for inconsistency analysis in complex networks. Node splitting can also generate intuitive graphics showing the difference between the "direct", "indirect" and the combined information. However, node-split models are not easy to parameterise, especially when multi-arm trials are present since, as with other inconsistency models, there may be more than one possible parameterisation. Furthermore, care should be taken to ensure that the split nodes refer to contrasts involved in generating potential inconsistencies.

4.3.3. Bayesian and Frequentist approaches compared

Compared to the Bayesian methods, applications of the Bucher approach is conceptually simpler and relatively easy to apply, although it requires two "stages". Bayesian approaches have the advantage of being "one-stage": there is no need to summarise the findings on each contrast first. The two-stage approach introduces a particular difficulty in sparse networks where the evidence on some contrasts may be limited to a small number of trials. This is that the decision as to whether to fit a Random Effects model must be taken for each contrast separately, and if there is only one study only a "Fixed Effect" analysis is available, even when there is clear evidence of heterogeneity on other contrasts. The likelihood of detecting an inconsistency, therefore, will be highly sensitive to the pattern of evidence. Caldwell et al.[16] present an example where the choice of Fixed or Random Effects summaries in the first stage determines whether inconsistency is detected in the second. Interestingly, the inconsistency model with its shared variance parameter offers a way "smoothing" the estimates of between-trial heterogeneity.

But sparse data also shows up drawbacks in the Bayesian methods, especially when a RE analysis is used in the underlying model. The difficulty is that the greater the degree of between-trials heterogeneity, the less likely it is for inconsistency to be detectable. The particular difficulty with Bayesian methods is that there is seldom enough data to estimate the between-trials variation. The practice of using vague prior distributions for the between-trials variation, combined with a lack of data, will generate posteriors which allow an unrealistically high variance. This, in turn, is likely to mask all but the most obvious signs of inconsistency.

In the smoking cessation example presented above, the posterior estimates show a high level of between-trials variation. Against this background only an exceptionally striking inconsistency could be detected. In the thrombolytic example, we used a FE analysis, but we have obtained very similar results with a RE model.[7,24] However, in this dataset, the extent of between-trials variation is unusually low, so that the inconsistency stands out (Figure 5). Our advice is to scrutinise the posterior distribution of the between-trials standard deviation from a consistency model, before embarking on an analysis of inconsistency based on Bayesian models. If the data has failed to rule out unrealistic values, consideration should be given to using informative priors, based on expert opinion or meta-epidemiological data (see TSD2,[9] Section 6.2 and TSD3[1]).

5. MEASURES TO AVOID INCONSISTENCY

While it is essential to carry out tests for inconsistency, the issue should not be considered in an overly mechanical way. Detection of inconsistency, like the detection of any statistical interaction, requires far more data than is needed to establish the presence of a treatment effect. Investigators will therefore nearly always fail to reject the null hypothesis of consistency. But this is not an indication that there is no inconsistency. Although a high level of heterogeneity increases the risk of inconsistency, it also lowers the chances that it will be detected. For this reason, even when inconsistency is not detected, and when, as with indirect comparisons, it *cannot* be detected because ICDF=0, the question that must always be asked is: "How reliable are conclusions based on indirect evidence or network meta-analysis?" A full consideration of this issue, which is still an area of active research interest, is beyond the scope of this document. However, below we outline measures that can help avoid inconsistency, and suggest some further empirical indicators that can provide some reassurance about the risk of inconsistency.

5.1. AVOIDING HETEROGENEITY

As discussed in Section 2.2, the mechanisms that potentially could create "bias" in indirect comparisons appear be to identical to those that cause heterogeneity in pair-wise meta-analysis. Thus, to ensure conclusions based on indirect evidence are sound, we must attend to the direct evidence on which they are based, as is clear from equation (1), repeated here: $\hat{d}_{BC}^{Ind} = \hat{d}_{AC}^{Dir} - \hat{d}_{AB}^{Dir}$. This states that if the direct estimates of the AB and AC effects are

unbiased estimates of the treatment effects *in the target population*, the indirect estimate of the BC effect must be unbiased as well. Conversely, any bias in the direct estimates, for example due to effect-modifying covariates arising from the patients not being drawn from the target population, will be passed on to the indirect estimates in equal measure. The term "bias" in this context must be seen broadly, comprising both internal and external threats to validity. This implies that if direct evidence on AB is based on trials conducted on a different patient population, and that a treatment effect modifier is present, what some may regard as an "incorrect generalisation" from the AB trials to draw inferences about the target population can be considered as (external) bias which will be inherited by any indirect estimates based on this data.

Thus, it seems that to a large extent the question "are conclusions based on indirect evidence reliable?" should be considered alongside the question "are conclusions based on pair-wise meta-analysis reliable?" Any steps that can be taken to avoid between-trial heterogeneity will be effective in reducing the risk of drawing incorrect conclusions from both pair-wise meta-analysis, indirect comparisons and network meta-analysis alike. Fortunately, the decision making context is likely to have already eliminated the great majority of potentially confounding factors. The most obvious sources of potential heterogeneity of effect, such as differences in dose or differences in co-therapies, will already have been eliminated in the scope, which is likely to restrict the set of trials to specific doses and co-therapies.

Clear cases where direct and indirect evidence are in conflict are rare in the literature.[11] Where inconsistency has been evident, it illustrates the danger introduced by heterogeneity, and in particular by the practice of trying to combine evidence on disparate treatment doses or treatment combinations within meta-analyses, often termed "lumping", as noted in the Introduction. The material used to illustrate the Bucher method for detecting inconsistency in Box 1[15] is a further example. Here, there were substantial, and independently recognised, differences in efficacy between the treatment combinations appearing in the direct evidence and those in the indirect evidence. It has been shown that if this is addressed, the difference between indirect and direct evidence is no longer statistically significant.[28]

In spite of all the limits on heterogeneity resulting from the narrow scope required to make a decision, there is still the potential for treatment effect modifiers to be present in trials, and unrecognised. Results may not have been broken down by confounding variables, and their distribution over the sample may not have been recorded. Among typical variables that frequently appear as effect modifiers are age, severity at baseline, and previous treatments, all

of which may be further confounded with each other. Investigators should make themselves aware of potential confounders, both within the network of evidence, and in previous literature, and consider the potential role of bias adjustment and meta-regression (see TSD3[1]), prior to synthesis and consistency checking.

5.2. EMPIRICAL INDICATIONS OF HETEROGENEITY

The above discussion suggests that the risk of inconsistency is greatly reduced if between-trial heterogeneity is low. Empirical assessment of heterogeneity can therefore provide some reassurance, or can alert investigators to the risk of inconsistency. Tests of homogeneity in the pair-wise comparisons, using I^2 [29] or χ^2 measures, can be used for this purpose (see TSD3[1]). Posterior summaries of the distribution of the between trials standard deviation, may be more useful because the extent of between-trial heterogeneity can be compared to the size of the mean treatment effects. A second useful indicator is the between-trials variation in the trial "baselines". If a large number of trials include comparisons with a reference treatment, perhaps placebo or a standard, and these arms all have similar proportions of events, hazards etc, then this suggests that the trial populations are relatively homogeneous and that there will be little heterogeneity in the treatment effects. If, on the other hand, the baselines are highly heterogeneous, while not meaning that the *relative* effects are also heterogeneous, it does at least constitute a warning that there is a potential risk of heterogeneity in the relative effects. This is an observation that has been made in the context of pair-wise synthesis. Heterogeneity in baselines can be examined via a Bayesian synthesis (see TSD5[30]).

6. RESPONSE TO INCONSISTENCY

There has been little work on how to respond to inconsistency when it is detected in a network. It would appear to be a reasonable principle that decisions should be based on models that are internally coherent, that is models in which $d_{YZ} = d_{XZ} - d_{XY}$, *and* that these models should fit the data. If the data cannot be fitted by a coherent model, then some kind of adjustment must be made. Any adjustment in response to inconsistency is *post hoc*, which emphasises the importance of identifying potential causes of heterogeneity of effect at the scoping stage, and potential internal biases in advance of synthesis.

One possible cause of inconsistency is a poor choice of scale of measurement, which can also lead to increased heterogeneity.[31] It is not always obvious whether to model treatment effects

on a risk difference, logit, or complementary log log scale. In TSD2[9] we emphasise that the choice of which scale was most appropriate is essentially an empirical one, although there is seldom enough evidence to decide on the basis of goodness of fit. Our experience, however, is that measurement scales that lead to a higher I^2 statistic also show more inconsistency on the measures described in Section 4.

Inconsistency in one part of the network does not necessarily imply that the entire body of evidence is to be considered suspect. However, inconsistency is a property of "loops", not individual contrasts. In a triangle structure, it is not possible to identify which contrast is "deviant". In the Enuresis example (Figure 2), we are essentially looking at four different estimates of a single contrast. If three agree and the fourth is different, it might be considered that three estimates have been "corroborated" and that the fourth is "deviant". In this particular instance, an examination of the different estimates (Box 2) does not suggest any such simple interpretation, and it would be necessary to review *all* these studies. It would be advisable, in fact, to *always* reconsider the entire network if inconsistency is located in any part of it, as the inconsistency throws doubt on the trial inclusion criteria and the potential presence of effect modifiers. Both are issues that might affect the entire network.

In the absence of corroboration, it is important to appreciate that once inconsistency is detected, there is little that statistical science can offer as remediation. For example it might be that the statistical analysis shows up one particular trial as having a bad fit within a consistency model. Very likely the poor statistical fit might disappear if this trial is removed, or if the observed treatment effect in the trial is adjusted. However, it is highly likely that a consistent network of evidence can also be obtained by removing or making adjustments to *other* trials. Worse, each different adjustment might be equally effective in reducing inconsistency, but each represents a very different interpretation of the evidence, and each produces very different estimates. There are clear examples of this in the literature on multi-parameter evidence synthesis in epidemiology applications.[32,33] The essential point is that inconsistency is not a property of individual studies, but of loops of evidence, and it may not always be possible to isolate which loop is "responsible" for the detected inconsistency, let alone which edge.[7] Where several alternative adjustments are available, a sensitivity analysis is essential.

The decision of how to address inconsistency cannot therefore be determined by statistical methods. A thorough review of the entire evidence base by clinical epidemiologists is required. This may result in the identification of one or more trials that are "different" in

some way, where, for example, a treatment-modifying covariate (e.g. dose) is present, or suspected. Investigators must then make a series of decisions: how might these factors relate to the target population for the decision? Are there specific trials that should not be included in the evidence base? Should the treatment effects observed in some trials be regarded as "biased" and adjusted in some way (see TSD3[1]) and if so, what data is available on which this adjustment can be based? A final option, of course, if there seems to be no explanation for an apparent inconsistency or heterogeneity, is to consider it a chance finding.

7. SUMMARY AND CONCLUSIONS

7.1. CHOICE OF METHOD

- Choice of method should be guided by the evidence structure.
- If it is possible to construct independent tests, then the Bucher test or its extensions to larger "circuit" structures and to chi-square tests represent the most simple and complete approach.
- In more complex networks, a repeated application of the Bucher method to all the possible loops produces interpretable results *as long as no "significant" inconsistencies are found.* If inconsistencies are found, correction for multiple testing is needed, but it is difficult to specify how this should be done.
- Within a Bayesian framework a consistency model can be compared to an "inconsistency" model. Analyses of residual deviance can provide an "omnibus" test of global inconsistency, and can also help locate it.
- Node splitting[24] is another effective method for comparing direct evidence to indirect evidence in complex networks.
- Measures of inconsistency variance[7] or incoherence variance[14] are not recommended as indicators of inconsistency

7.2. PLACE OF INCONSISTENCY TESTING IN EVIDENCE SYNTHESIS.

Logically, inconsistency testing should come after an examination of heterogeneity, and after adjustment for known causes of heterogeneity through meta-regression or bias adjustment (see TSD3[1]).

7.3. RESPONSE TO INCONSISTENCY

Decisions should be based on coherent models that fit the data. Careful examination of different sources of evidence may reveal that some estimates are “corroborated” and others not. If inconsistency is detected, the entire network of evidence should be reconsidered from a clinical epidemiology viewpoint with respect to the presence of potential effect modifiers.

8. REFERENCES

1. Dias, S., Sutton, A.J., Welton, N.J., Ades, A.E. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

2. Caldwell, D.M., Ades, A.E., Higgins, J.P.T. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ: British Medical Journal* 2005; 331(7521):897-900.

3. Gleser, L.J., Olkin, I. Stochastically dependent effect sizes. *The Handbook of Research Synthesis* 1994;339-355.

4. Hasselblad, V. Meta-analysis of multi-treatment studies. *Medical Decision Making* 1998; 18:37-43.

5. Higgins, J.P.T., Whitehead, A. Borrowing strength from external trials in a meta analysis. *Statistics in Medicine* 1996; 15(24):2733-2749.

6. Lu, G., Ades, A.E. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* 2004; 23(20):3105-3124.

7. Lu, G., Ades, A.E. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006; 101(474):447-459.

8. Glenny, A.M., Altman, D.G., Song, F., Sakarovitch, C., Deeks, J.J., D'amico, R. et al. Indirect comparisons of competing interventions. *Health Technology Assessment* 2005; 9(26):1-149.

9. Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

10. Lu, G., Ades, A. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* 2009; 10(4):792-805.

11. Song, F., Altman, D.G., Glenny, A.M., Deeks, J.J. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ: British Medical Journal* 2003; 326(7387):472-476.

12. Bucher, H.C., Guyatt, G.H., Griffith, L.E., Walter, S.D. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* 1997; 50(6):683-691.

13. Lunn, D.J., Thomas, A., Best, N.G., Spiegelhalter, D.J. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; 10:325-337.

14. Lumley, T. Network meta analysis for indirect treatment comparisons. *Statistics in Medicine* 2002; 21(16):2313-2324.

15. Chou, R., Fu, R., Hoyt Huffman, L., Korthuis, P.T. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. *The Lancet* 2006; 368(9546):1503-1515.

16. Caldwell, D.M., Welton, N.J., Ades, A.E. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. *Journal of Clinical Epidemiology* 2010; 63(8):875-882.

17. Russell, K., Kiddoo, D. The Cochrane Library and nocturnal enuresis; an umbrella review. *Evidence Based Child Health: A Cochrane Review Journal* 2006; 1(1):5-8.

18. Salanti, G., Marinho, V., Higgins, J. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *Journal of Clinical Epidemiology* 2009; 62(8):857-864.

19. Cipriani, A., Furukawa, T.A., Salanti, G., Geddes, J.R., Higgins, J.P.T., Churchill, R. et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The Lancet* 2009; 373(9665):746-758.

20. Eddy, D.M., Hasselblad, V., Shachter, R.D. Meta-analysis by the confidence profile method: The statistical synthesis of evidence. Academic Press, 1992.

21. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 2002; 64(4):583-639.

22. Boland, A., Dundar, Y., Bagust, A., Haycox, A., Hill, R., Mujica Mota, R. et al. Early thrombolysis for the treatment of acute myocardial infarction: a systematic review and economic evaluation. *Health Technology Assessment* 2003; 7(15):1-136.

23. Keeley, E.C., Boura, J.A., Grines, C.L. Primary angioplasty versus intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review of 23 randomised trials. *The Lancet* 2003; 361(9351):13-20.

24. Dias, S., Welton, N.J., Caldwell, D.M., Ades, A.E. Checking consistency in mixed treatment comparison meta analysis. *Statistics in Medicine* 2010; 29(7 8):932-944.

25. Elliott, W.J., Meyer, P.M. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *The Lancet* 2007; 369(9557):201-207.

26. Trikalinos, T.A., Olkin, I. A method for the meta analysis of mutually exclusive binary outcomes. *Statistics in Medicine* 2008; 27(21):4279-4300.

27. Salanti, G., Higgins, J.P.T., Ades, A.E., Ioannidis, J. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* 2008; 17(3):279-301.

28. Caldwell, D.M., Gibb, D.M., Ades, A.E. Validity of indirect comparisons in meta-analysis. *Lancet* 2007; 369(9558):270.

29. Higgins, J.P.T., Thompson, S.G. Quantifying heterogeneity in a meta analysis. *Statistics in Medicine* 2002; 21(11):1539-1558.

30. Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. NICE DSU Technical Support Document 5: Evidence synthesis in the baseline natural history model. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

31. Deeks, J.J. Issues in the selection of a summary statistic for meta analysis of clinical trials with binary outcomes. *Statistics in Medicine* 2002; 21(11):1575-1600.

32. Goubar, A., Ades, A.E., De Angelis, D., McGarrigle, C.A., Mercer, C.H., Tookey, P.A. et al. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; 171(3):541-580.

33. Presanis, A.M., De Angelis, D., Spiegelhalter, D.J., Seaman, S., Goubar, A., Ades, A.E. Conflicting evidence in a Bayesian synthesis of surveillance data to estimate human immunodeficiency virus prevalence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; 171(4):915-937.

APPENDIX: WINBUGS CODE FOR ILLUSTRATIVE EXAMPLES

Below we set out code to fit random and fixed effects inconsistency models to any network with a binomial likelihood and logit link function. The program codes are printed here, but are also available as WinBUGS system files from http://www.nicedsu.org.uk. Users are advised to download the WinBUGS files from the website instead of copying and pasting from this document. In TSD2[9] a generalised linear model framework was introduced, with explanations of how the code for the binomial/logit model could be adapted for other likelihoods and link functions, including Poisson/log, Normal/identity and other models. The inconsistency models below can be adapted in exactly the same way.

The code below is fully general and will work for any number of multi-arm trials with any number of arms. It is suitable for networks where there is information on all possible treatment contrasts (such as in the Smoking example presented in Section 4.2.1) or where there is information on just a subset of possible contrasts (such as in the Thrombolytic treatments example presented in Section 4.2.2). However, in the latter case, the WinBUGS output for contrasts that have no information will be redundant, i.e. the posterior distribution will be equal to the prior and no inferences can be made on these contrasts. We therefore recommend a careful consideration of the network structure before looking at the WinBUGS output from the code below.

SMOKING CESSATION: RE MODEL, BINOMIAL LIKELIHOOD

```
# Binomial likelihood, logit link, inconsistency model
# Random effects model
model{                                  # *** PROGRAM STARTS
for(i in 1:ns){                         # LOOP THROUGH STUDIES
   delta[i,1]<-0                        # treatment effect is zero in control arm
   mu[i] ~ dnorm(0,.0001)               # vague priors for trial baselines
   for (k in 1:na[i])  {                # LOOP THROUGH ARMS
      r[i,k] ~ dbin(p[i,k],n[i,k])      # binomial likelihood
      logit(p[i,k]) <- mu[i] + delta[i,k]   # model for linear predictor
      rhat[i,k] <- p[i,k] * n[i,k]      # expected value of the numerators
      dev[i,k] <- 2 * (r[i,k] * (log(r[i,k])-log(rhat[i,k]))  #Deviance contribution
        +  (n[i,k]-r[i,k]) * (log(n[i,k]-r[i,k]) - log(n[i,k]-rhat[i,k])))
    }
  resdev[i] <- sum(dev[i,1:na[i]])      # summed residual deviance contribution for this trial
   for (k in 2:na[i]) {  # LOOP THROUGH ARMS
      delta[i,k] ~ dnorm(d[t[i,1],t[i,k]] ,tau) # trial-specific LOR distributions
    }
 }
totresdev <- sum(resdev[])              # Total Residual Deviance
for (c in 1:(nt-1)) {                   # priors for all mean treatment effects
   for (k in (c+1):nt)  { d[c,k] ~ dnorm(0,.0001) }
 }
sd ~ dunif(0,5)                         # vague prior for between-trial standard deviation
var <- pow(sd,2)                        # between-trial variance
tau <- 1/var                            # between-trial precision
}                                       # *** PROGRAM ENDS
```

```
# Data (Smoking example)
# nt=no. treatments, ns=no. studies
list(nt=4,ns=24 )

r[,1]   n[,1]   r[,2]   n[,2]   r[,3]   n[,3]   t[,1]   t[,2]   t[,3]   na[]
9       140     23      140     10      138     1       3       4       3  # trial 1 ACD
11      78      12      85      29      170     2       3       4       3 # trial 2 BCD
75      731     363     714     NA      1       1       3       NA      2 # 3
2       106     9       205     NA      1       1       3       NA      2 # 4
58      549     237     1561    NA      1       1       3       NA      2 # 5
0       33      9       48      NA      1       1       3       NA      2 # 6
3       100     31      98      NA      1       1       3       NA      2 # 7
1       31      26      95      NA      1       1       3       NA      2 # 8
6       39      17      77      NA      1       1       3       NA      2 # 9
79      702     77      694     NA      1       1       2       NA      2 # 10
18      671     21      535     NA      1       1       2       NA      2 # 11
64      642     107     761     NA      1       1       3       NA      2 # 12
5       62      8       90      NA      1       1       3       NA      2 # 13
20      234     34      237     NA      1       1       3       NA      2 # 14
0       20      9       20      NA      1       1       4       NA      2 # 15
8       116     19      149     NA      1       1       2       NA      2 # 16
95      1107    143     1031    NA      1       1       3       NA      2 # 17
15      187     36      504     NA      1       1       3       NA      2 # 18
78      584     73      675     NA      1       1       3       NA      2 # 19
69      1177    54      888     NA      1       1       3       NA      2 # 20
20      49      16      43      NA      1       2       3       NA      2 # 21
7       66      32      127     NA      1       2       4       NA      2 # 22
12      76      20      74      NA      1       3       4       NA      2 # 23
9       55      3       26      NA      1       3       4       NA      2 # 24
END

#Initial values
# chain 1
list(sd=1,  mu=c(0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0),
d = structure(.Data = c(NA,0,0,0,     NA, NA,0,0,      NA,NA,NA,0), .Dim = c(3,4)))

# chain 2
list(sd=1.5,  mu=c(0,2,0,-1,0,  0,1,0,-1,0,  0,0,0,10,0,  0,10,0,0,0,  0,-2,0,0),
d = structure(.Data = c(NA,-2,0,5,     NA, NA,0,2,      NA,NA,NA,5), .Dim = c(3,4)))

# chain 3
list(sd=3,  mu=c(0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0),
d = structure(.Data = c(NA,-3,-3,-3,     NA, NA,-3,-3,      NA,NA,NA,-3), .Dim = c(3,4)))
```

THROMBOLYTIC TREATMENTS: FE MODEL, BINOMIAL LIKELIHOOD

```
# Binomial likelihood, logit link, inconsistency model
# Fixed effects model
model{                                  # *** PROGRAM STARTS
for(i in 1:ns){                         # LOOP THROUGH STUDIES
   mu[i] ~ dnorm(0,.0001)               # vague priors for trial baselines
   for (k in 1:na[i])  {                # LOOP THROUGH ARMS
      r[i,k] ~ dbin(p[i,k],n[i,k])      # binomial likelihood
      logit(p[i,k]) <- mu[i] + d[t[i,1],t[i,k]]   # model for linear predictor
      rhat[i,k] <- p[i,k] * n[i,k]      # expected value of the numerators
      dev[i,k] <- 2 * (r[i,k] * (log(r[i,k])-log(rhat[i,k]))  #Deviance contribution
        + (n[i,k]-r[i,k]) * (log(n[i,k]-r[i,k]) - log(n[i,k]-rhat[i,k])))
    }
  resdev[i] <- sum(dev[i,1:na[i]])      # summed residual deviance contribution for this trial
 }
totresdev <- sum(resdev[])              # Total Residual Deviance
for (k in 1:nt) { d[k,k] <- 0 }         # set effects of k vs k to zero
for (c in 1:(nt-1)) {                   # priors for all mean treatment effects
```

```
   for (k in (c+1):nt)  { d[c,k] ~ dnorm(0,.0001) }
  }
} # *** PROGRAM ENDS


# Data (Thrombolytic treatments example)
#nt=no. treatments, ns=no. studies;
list(nt=9,ns=50)

r[,1]   n[,1]   r[,2]   n[,2]   r[,3]   n[,3]   t[,1]   t[,2]   t[,3]   na[]   #   study ID
1472    20251   652     10396   723     10374   1       3       4       3      #   1
9       130     6       123     NA      NA      1       2       NA      2      #   2
5       63      2       59      NA      NA      1       2       NA      2      #   3
3       65      3       64      NA      NA      1       2       NA      2      #   4
887     10396   929     10372   NA      NA      1       2       NA      2      #   5
1455    13780   1418    13746   1448    13773   1       2       9       3      #   6
7       85      4       86      NA      NA      1       2       NA      2      #   7
12      159     7       157     NA      NA      1       2       NA      2      #   8
10      135     5       135     NA      NA      1       2       NA      2      #   9
4       107     6       109     NA      NA      1       4       NA      2      #   10
285     3004    270     3006    NA      NA      1       5       NA      2      #   11
11      149     2       152     NA      NA      1       7       NA      2      #   12
1       50      3       50      NA      NA      1       7       NA      2      #   13
8       58      5       54      NA      NA      1       7       NA      2      #   14
1       53      1       47      NA      NA      1       7       NA      2      #   15
4       45      0       42      NA      NA      1       7       NA      2      #   16
14      99      7       101     NA      NA      1       7       NA      2      #   17
9       41      3       46      NA      NA      1       7       NA      2      #   18
42      421     29      429     NA      NA      1       7       NA      2      #   19
2       44      3       46      NA      NA      2       7       NA      2      #   20
13      200     5       195     NA      NA      2       7       NA      2      #   21
2       56      2       47      NA      NA      2       7       NA      2      #   22
3       55      1       55      NA      NA      3       7       NA      2      #   23
10      94      3       95      NA      NA      3       7       NA      2      #   24
40      573     32      565     NA      NA      3       7       NA      2      #   25
2       61      3       62      NA      NA      3       7       NA      2      #   26
16      419     20      421     NA      NA      3       7       NA      2      #   27
5       69      3       71      NA      NA      3       7       NA      2      #   28
5       75      5       75      NA      NA      3       7       NA      2      #   29
59      782     52      790     NA      NA      3       7       NA      2      #   30
5       81      2       81      NA      NA      3       7       NA      2      #   31
16      226     12      225     NA      NA      3       7       NA      2      #   32
8       66      6       71      NA      NA      3       7       NA      2      #   33
522     8488    523     8461    NA      NA      3       6       NA      2      #   34
356     4921    757     10138   NA      NA      3       5       NA      2      #   35
13      155     7       169     NA      NA      3       5       NA      2      #   36
10      203     7       198     NA      NA      1       8       NA      2      #   37
3       58      2       52      NA      NA      1       9       NA      2      #   38
3       86      6       89      NA      NA      1       9       NA      2      #   39
3       58      2       58      NA      NA      1       9       NA      2      #   40
13      182     11      188     NA      NA      1       9       NA      2      #   41
2       26      7       54      NA      NA      3       8       NA      2      #   42
12      268     16      350     NA      NA      3       8       NA      2      #   43
5       210     17      211     NA      NA      3       9       NA      2      #   44
3       138     13      147     NA      NA      3       9       NA      2      #   45
8       132     4       66      NA      NA      2       8       NA      2      #   46
10      164     6       166     NA      NA      2       8       NA      2      #   47
6       124     5       121     NA      NA      2       8       NA      2      #   48
13      164     10      161     NA      NA      2       9       NA      2      #   49
7       93      5       90      NA      NA      2       9       NA      2      #   50
END


# Initial values
# chain 1
```

```
list(mu=c(0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,
        0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0),
d = structure(.Data = c(NA,0,0,0,0,0,0,0,0,  NA,NA,0,0,0,0,0,0,0,  NA,NA,NA,0,0,0,0,0,0,  NA,NA,NA,NA,0,0,0,0,0,
NA,NA,NA,NA,NA,0,0,0,0,  NA,NA,NA,NA,NA,NA,0,0,0,  NA,NA,NA,NA,NA,NA,NA,0,0,  NA,NA,NA,NA,NA,NA,NA,NA,0,
NA,NA,NA,NA,NA,NA,NA,NA,NA),  .Dim = c(9,9)) )

# chain 2
list(mu=c(0,0,10,0,-1,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,
        0,0,0,0,0,  3,0,0,-2,0,  0,0,-1,0,0,  0,-0.5,5,0.5,0.5,  0,0,0,2,0),
d = structure(.Data = c(NA,0,1,0,0,-2,0,0,0,  NA,NA,0,0,2,0,0,-2,0,  NA,NA,NA,0,0,0,0,0,0,  NA,NA,NA,NA,0,0,0,0,0,
NA,NA,NA,NA,NA,0,0,1,0,  NA,NA,NA,NA,NA,NA,0,0,-2,  NA,NA,NA,NA,NA,NA,NA,0,0,  NA,NA,NA,NA,NA,NA,NA,NA,0,
NA,NA,NA,NA,NA,NA,NA,NA,NA),  .Dim = c(9,9))  )

# chain 3
list(mu=c(0,0,10,0,5,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,  0,0,0,0,0,
        0,0,0,0,0,  3,0,0,-2,0,  0,0,-8,0,0,  0,-0.5,5,0.5,0.5,  0,0,0,2,0),
d = structure(.Data = c(NA,0,1,0,0,-5,0,5,0,  NA,NA,0,0,3,0,0,-2,0,  NA,NA,NA,0,0,0,0,0,0,  NA,NA,NA,NA,0,0,0,0,0,
NA,NA,NA,NA,NA,0,-5,1,0,  NA,NA,NA,NA,NA,NA,0,0,-4,  NA,NA,NA,NA,NA,NA,NA,0,0,  NA,NA,NA,NA,NA,NA,NA,NA,0,
NA,NA,NA,NA,NA,NA,NA,NA,NA),  .Dim = c(9,9))  )
```