Research Article

Statistics in Medicine

Received 8 May 2012, Accepted 28 January 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5764

Incorporating data from various trial designs into a mixed treatment comparison model

Susanne Schmitz,[a*†] Roisin Adams[b] and Cathal Walsh[a,b]

Estimates of relative efficacy between alternative treatments are crucial for decision making in health care. Bayesian mixed treatment comparison models provide a powerful methodology to obtain such estimates when head-to-head evidence is not available or insufficient. In recent years, this methodology has become widely accepted and applied in economic modelling of healthcare interventions. Most evaluations only consider evidence from randomized controlled trials, while information from other trial designs is ignored. In this paper, we propose three alternative methods of combining data from different trial designs in a mixed treatment comparison model. Naive pooling is the simplest approach and does not differentiate between-trial designs. Utilizing observational data as prior information allows adjusting for bias due to trial design. The most flexible technique is a three-level hierarchical model. Such a model allows for bias adjustment while also accounting for heterogeneity between-trial designs. These techniques are illustrated using an application in rheumatoid arthritis. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: mixed treatment comparison; network meta-analysis; multiple treatments meta-analysis; observational data; hierarchical modelling; rheumatoid arthritis

1. Introduction

1.1. Motivation

Synthesizing evidence is essential when evaluating healthcare interventions. Single study groups are often too small to make an informed decision, and when more than one trial investigates the effect in question, a conclusion should be drawn on the basis of all available evidence rather than one trial only.

Meta-analysis has developed to be a widely used tool to combine evidence from trials evaluating the same intervention. There is a growing demand, however, to estimate the relative efficacy between all alternative interventions used to treat the same medical condition. Such estimates are essential to inform choice of agent in a clinical setting and also form the basis for economic evaluations carried out by national organizations such as the National Institute for Health and Clinical Excellence in the UK and the National Centre for Pharmacoeconomics in Ireland to inform healthcare decision making. Because of ethical or financial reasons, direct evidence of alternative interventions is often sparse. Mixed treatment comparison (MTC) models are a generalization of meta-analysis that allow the estimation of relative efficacy between treatments where direct evidence is not available or is insufficient. MTC is also known as network meta-analysis or multiple treatment meta-analysis. In recent years, MTCs have become an accepted tool for evidence synthesis in the medical literature [1–3].

Combining evidence from different sources fits naturally in the Bayesian framework where the inclusion of all available evidence is anticipated [4]. A Bayesian approach to MTC modelling provides a powerful methodology that is flexible to deal with increasingly complex evidence structures [5]. Such models simultaneously estimate the efficacy between treatments that are not directly compared,

[a] *Trinity College Dublin, Dublin, Ireland*
[b] *National Centre for Pharmacoeconomics, Dublin, Ireland*
**Correspondence to: Susanne Schmitz, Department of Statistics, Trinity College Dublin, Ireland.*
[†] *E-mail: schmitzs@tcd.ie*

and borrowing strength across the network ensures an optimal use of the available evidence. Its value to evidence synthesis modelling has been recently highlighted [6].

One of the incentives for combining evidence from different sources is to make an informed decision on the basis of all available evidence. Nevertheless, MTCs are typically restricted to evidence based on randomized controlled trials (RCTs), while a wealth of evidence from different trial designs is ignored. RCTs are considered to be the gold standard of evidence, because its controlled setting minimizes potential bias. On the other hand, it has been argued that its experimental setting limits its external validity [7]. Although observational trials reflect reality well, there are other concerns such as overprecision and overestimation of the treatment effect [8]. An advantage of a Bayesian evidence synthesis is the opportunity to systematically include a wide range of data in the form of prior information. It is also the case that software is readily available within the paradigm with which to carry out the modelling of these complex structures. There is an argument that only including randomized evidence in a meta-analysis might be a suboptimal approach, because data from alternative study designs, while possibly less reliable, still contain additional information about the effect size [4, 9].

Being a generalization of meta-analysis, we argue that the same argument holds for MTCs. Three methods have been proposed for the inclusion of data from different trial designs into a standard meta-analysis: naive pooling, inclusion in the form of prior information and an extension of the model to a three-level hierarchical model [4, 10]. These methods differ in their flexibility to adjust for potential bias and to quantify consistency between-trial designs. Mak *et al.* [11] and Salpeter *et al.* [12] used observational evidence to inform prior information. Prevost *et al.* [10] have applied a hierarchical model to include nonrandomized trial data in a standard meta-analytic framework. In the paper presented here, we will generalize all three methods for the application in MTC modelling.

Using a case study in rheumatoid arthritis (RA), we will illustrate the effect of including evidence other than RCTs in an MTC and analyse the advantages and disadvantages of the three methods.

1.2. Potential bias in observational trials

Whereas RCTs may lack generalizability because of their strict inclusion criteria and clinical setting, randomization is likely to create a balanced distribution of known and unknown confounding factors in the treatment and the control groups [13]. This minimizes the occurrence of bias, and RCTs are therefore considered to be the gold standard of clinical evidence. Observational data on the other hand are collected in real-life environments, and although estimates are more generalizable to a wider patient population, such nonrandomized evidence can be prone to potential bias. When including observational data in evidence synthesis, it is hence important to acknowledge potential bias by adjusting the model to reflect these. There are two types of bias that may be present in observational estimates, nonsystematic and systematic bias, both of which can be modelled by adjusting the likelihood function [14, 15]. Let ξ be the bias, then a nonsystematic bias has the following form:

$$\xi \sim N(0, \tau_\xi) \tag{1}$$

Modelling bias in this way downweighs observational evidence by inflating the variance and therefore adjusts for overprecision. Modelling a systematic bias adjusts for overestimation or underestimation of the treatment effect by shifting the mean by a fixed amount $\xi = \mu_\xi$. A combination of both forms of bias has the following form:

$$\xi \sim N(\mu_\xi, \tau_\xi) \tag{2}$$

In practice, it is difficult to estimate the size of bias because of paucity of evidence. A sensitivity analysis in this respect is therefore essential. A range of applications of bias adjustments for observational evidence have been published [8, 16]. Both of these assume an overestimation of the treatment effect in observational studies; however, the data do not appear to consistently overestimate or underestimate the treatment effect [17, 18].

Two of the three methods described in this article allow for modelling of bias due to trial design.

1.3. The case study

Rheumatoid arthritis is a chronic, progressive and disabling autoimmune disease, causing swelling and damaging cartilage and bone around the joints. Any joint may be affected, but it is commonly the hands, feet and wrists. Common symptoms are joint swelling, pain, morning joint stiffness, poor sleep, fatigue and weight loss [19].

Over the past decade, enhanced understanding of the molecular pathogenesis has led to the development of biologic agents that target specific parts of the immune system. These innovative treatments have altered the path and face of RA and outcomes for patients and society. Tumour necrosis factor alpha antagonists (anti-tumour necrosis factor (anti-TNF)-α) are the first of the biologic treatment groups used in RA. There are currently five anti-TNF agents licensed for RA in Europe: adalimumab, certolizumab, etanercept, golimumab and infliximab. All of these agents have demonstrated considerable efficacy in placebo-controlled RCTs in patients who have had an inadequate response to conventional disease-modifying antirheumatic drugs (DMARDs) such as methotrexate (MTX) or sulfasalazine.

Although there is a wealth of RCT evidence available for these agents compared with either placebo or conventional DMARDs, there are currently very limited head-to-head RCTs of anti-TNF agents. Despite this, some estimate of relative efficacy to inform choice of agent is needed.

There exist a range of Bayesian MTC analyses estimating the relative efficacy of these treatments [3, 20–25]. None of these included evidence other than RCTs; however, there exists a large body of observational data collected via registries and open-label studies.

The objective of this paper is to facilitate the inclusion of evidence from different trial designs in an MTC framework. We will illustrate the advantages and disadvantages of three inclusion methods by using the case study in RA.

1.4. Efficacy measures

To estimate relative efficacy between treatments, one has to decide on a measure of disease activity and improvement. Commonly used measures in RA are the American College of Rheumatology outcome criteria, the Disease Activity Score and the Health Assessment Questionnaire (HAQ) score; details on these and other measures can be found elsewhere [26]. Dichotomized outcome measures have been shown to suffer from a loss of power to detect differences between treatments in MTC modelling [27]. A continuous outcome measure was therefore chosen for this analysis. The Disease Activity Score was reported in too few trials; for the purpose of this analysis, we therefore chose the HAQ score as measure of efficacy. The HAQ score represents the result of a self-report questionnaire in which patients rate their ability to perform daily life activities such as washing one's hair or getting in and out of a car. Values range from 0 to 3, where high values indicate a more severe disease status. The improvement in HAQ score is measured on a continuous scale, which enhances its sensitivity to change. It is a widely reported measure in clinical trials. Disease activity at baseline influences the effectiveness of the intervention [28]. We therefore model the improvement in HAQ score relative to baseline disease activity measured by the HAQ score at baseline.

2. Methods

2.1. Literature search and data extraction

2.1.1. Randomized controlled trials. A systematic literature review following the PRISMA method [29] was performed to identify trials meeting our inclusion criteria. The search included published studies up to and including October 2010 in PubMed, Embase and the Cochrane Database. The search was rerun in April 2012. Rheumatological inflammatory diseases other than RA were excluded from the search. The inclusion criteria were RCTs, patients with established RA who have had an inadequate response to MTX and who have been treated for at least 24 weeks (where 24-week data were not available, the data within 6 weeks either before or after 24 weeks were used). Both monotherapy and combination therapy were included with an explicit term in the statistical model allowing for the additional effect of MTX. A number of studies have published follow-up data. Kievit *et al.* [30] published data up to 12 months and found that treatment effect appears to be maintained for the anti-TNF treatment within this time scale. More details on the selection process are published elsewhere [3].

The literature review identified 16 RCTs meeting our inclusion criteria; 13 of which report the outcome measure of interest, improvement in HAQ score. The solid lines in Figure 1 show the network of available RCT evidence, and the extracted data are summarized in Table I.

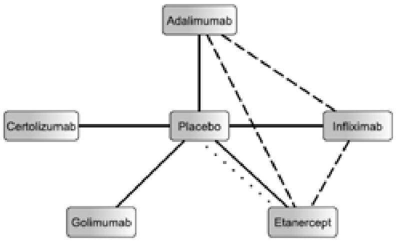

Figure 1. Combined network diagram: solid edges refer to randomized controlled trial evidence; dashed edges indicate observational evidence; dotted line refers to a matched observational trial.

2.1.2. Observational trials. Relaxing the inclusion criteria of the aforementioned literature search to include trials other than RCTs identified the observational data suitable for the inclusion in the analysis. The dashed lines in Figure 1 show the network of available observational evidence, and the data extracted from the observational trials are summarized in Table I.

Two of the trials include data collected from registries (Italy and the Netherlands). These registries were established to monitor the safety of these drugs and also to assess the effectiveness of these agents in the real-life setting. The Dutch Rheumatoid Arthritis Register was established to evaluate the effects of adalimumab, etanercept and infliximab on disease activity, functional ability, quality of life and medication costs in a real-life setting ($n = 707$). Bazzani et al. assessed the effectiveness of anti-TNF α agents by analysing the primary clinical outcomes in patients with active RA ($n = 1010$). Klareskog et al. carried out an open-label extension to assess the long-term safety and efficacy of etanercept in patients with RA ($n = 549$); it is a one-armed trial. Particular concern about one-armed trials has been expressed by Ades *et al.* [46]; they 'very strongly recommend that single arm studies, whether RCT or observational study, are excluded (from MTCs)'. We agree that unadjusted MTC methods do not provide reliable estimates. In this case, we must either exclude this large trial or find an appropriate placebo comparison. To do this, we match a control arm from a trial, whose baseline characteristics closely reflect those of Klareskog *et al.* [31]; we have used the MTX arm of Weinblatt *et al.* [32]. As Klareskog *et al.*, this trial examines combination therapy administering MTX in conjunction with etanercept in the treatment arm and MTX in the control arm. No great differences are found in important baseline characteristics mean age, number of previous DMARDs and baseline HAQ score. Because the construction of this link to placebo may introduce additional bias, this trial has been excluded from the analysis in a sensitivity analysis. Baseline demographics of all trials are also summarized in Table I.

2.2. Model development

In a previous analysis, we have fitted an MTC to the RCT evidence only [3]. The model calculates the difference in percentage improvement between the five agents by taking into account baseline HAQ score. The model allows for more than one treatment arm per study, and adjustments are made for the concurrent treatment with MTX. In the following, when indices refer to a biologic agent, 1 indicates placebo, 2 adalimumab, 3 infliximab, 4 etanercept, 5 golimumab and 6 certolizumab. The likelihood for the improvement in HAQ score in study i and study arm j, $\Delta[i,j]$, is normally distributed with mean $\mu_\Delta[i,j]$ and precision $\tau_\Delta[i,j]$:

$$\Delta[i,j] \sim N(\mu_\Delta[i,j], \tau_\Delta[i,j]) \tag{3}$$

Mean improvement is modelled relative to baseline HAQ score $\lambda[i,j]$; $\delta[i,j]$ measures the proportion of improvement in HAQ score taking values in [0, 1]:

$$\mu_\Delta[i,j] = \lambda[i,j] \cdot \delta[i,j] \tag{4}$$

The effect of MTX is assumed to be additive.

$$(\text{A} + \text{MTX}) \text{ vs A} \equiv (\text{B} + \text{MTX}) \text{ vs B} \tag{5}$$

Table I. Trial data.

Trial	Arm	N	ΔHAQ (SD)	HAQ_{base}	Age	DoD	DMARDs	MTX-dose
Weinblatt *et al.* [33]	P+	62	0.27 (0.6)	1.64	56	12	3	17
	Ada+	69	0.54 (0.6)	1.52				
	Ada+	67	0.62 (0.6)	1.55				
	Ada+	73	0.59 (0.5)	1.55				
Keystone *et al.* [34]	P+	200	0.24 (0.5)	1.45	56	11	2	17
	Ada+	207	0.56 (0.5)	1.44				
	Ada+	212	0.60 (0.5)	1.48				
Van de Putte *et al.* [35]	P	110	0.07 (0.5)	1.88	53	11	4	NU
	Ada	112	0.39 (0.6)	1.88				
	Ada	106	0.29 (0.6)	1.88				
	Ada	103	0.49 (0.5)	1.84				
	Ada	113	0.38 (0.6)	1.83				
Miyasaka [36]	P	87	−0.1 (0.6)	1.39	55	7	NU	NU
	Ada	87	0.2 (0.5)	1.57				
	Ada	91	0.2 (0.6)	1.64				
	Ada	87	0.4 (0.6)	1.77				
Kim *et al.* [37]	P+	63	0.2 (0.5)	1.3	49	7	NU	16
	Ada+	65	0.5 (0.6)	1.4				
Maini *et al.* [38]	P+	88	0.3 (0.5)	1.8	53	8	3	15
	Inf+	86	0.3 (0.5)	1.8				
	Inf+	86	0.5 (0.5)	1.8				
	Inf+	87	0.5 (0.6)	1.8				
	Inf+	81	0.4 (0.5)	1.5				
Zhang *et al.* [39]	P+	86	0.45 (–)	1.6	48	8	NR	NR
	Inf+	87	0.76 (–)	1.6				
Moreland *et al.* [40]	P	80	0.03 (–)	1.7	52	12	3	NU
	Eta	76	0.58 (–)	1.7				
	Eta	78	0.62 (–)	1.6				
Weinblatt *et al.* [32]	P+	30	0.4 (–)	1.5	50	13	3	19
	Eta+	59	0.7 (–)	1.5				
Keystone *et al.* [41]	P+	133	0.13 (0.4)	1.25	51	6	NU	15
	Gol	133	0.13 (0.7)	1.38				
	Gol+	89	0.38 (0.5)	1.38				
	Gol+	89	0.5 (0.5)	1.38				
Keystone *et al.* [42]	P+	199	0.18 (–)	1.7	52	6	1	14
	Cert+	393	0.60 (–)	1.7				
	Cert+	390	0.63 (–)	1.7				
Smolen *et al.* [43]	P+	127	0.14 (0.5)	1.6	52	6	1	13
	Cert+	246	0.5 (0.5)	1.6				
	Cert+	246	0.5 (0.5)	1.6				
Fleischmann *et al.* [44]	P	109	−0.07 (0.4)	1.6	54	10	2	NU
	Cert	111	0.39 (0.7)	1.4				
Bazzani *et al.* [45,]	Eta+	230	0.34 (–)	1.23	56	9	3	NR
	Ada+	283	0.34 (–)	1.20				
	Inf+	497	0.34 (–)	1.50				
Kievit *et al.* [30,]	Eta+	289	0.35 (0.5)	1.4	56	7	3	NR
	Ada+	267	0.42 (0.5)	1.3				
	Inf+	151	0.23 (0.5)	1.4				
Klareskog *et al.* [31]*◇	Eta+	549	0.80 (–)	1.8	53	7	3	NR
	P+	30	0.4 (–)	1.5				

N, number of patients; ΔHAQ, change in HAQ score; HAQ_{base}, HAQ score at baseline; Age, mean age (years); DoD, mean disease duration (years); DMARDS, mean number of previous disease-modifying antirheumatic drugs; MTX-dose, dose of methotrexate (mg); +, additional treatment with methotrexate; *, observational trials; NU, not used; NR, not reported; ◇, an open-label extension, the placebo arm is taken from Weinblatt *et al*. [32]; Ada, adalimumab; Inf, infliximab; Eta, etanercept; Gol, golimumab; Cert, certolizumab; P, placebo.

$$\begin{aligned}\delta[i,j] &= \nu[i] + \alpha[i,t[i,j]] + \beta \cdot I[i,j] \\ \alpha[i,b[i,j]] &= 0\end{aligned} \tag{6}$$

$t[i,j]$ refers to the treatment given in treatment arm j of study i and $b[i,j]$ to the corresponding baseline treatment. Equation (6) defines the difference of improvement $\alpha[i,t[i,j]]$ between treatment $t[i,j]$ and baseline treatment $b[i,j]$ while adjusting for the additional effect of MTX β. $I[i,j]$ indicates whether MTX was present or absent in treatment arm j of study i. Constant treatment effects are assumed when the same treatment is given in multiple treatment arms of the same study. Because of the lack of sufficient evidence, fixed effects are assumed for the MTX effect. For the biologic treatment effect, we assume a random effect model:

$$\alpha[i,t[i,j]] \sim N(\mu_\alpha[i,t[i,j]], \tau_\alpha) \tag{7}$$

Each $\mu_\alpha[i,t[i,j]]$ can be written as the difference between baseline parameters:

$$\mu_\alpha[i,t[i,j]] = a[t[i,j]] - a[b[i,j]] \tag{8}$$

where

$$a[1] = 0 \tag{9}$$

defines placebo to be the basic treatment and relative efficacy to placebo to be the basic parameters. Prior distributions need to be declared for baseline effects $\nu[i]$, basic parameters $a[k]$ and the MTX effect β and for the between-trial standard deviation parameter $\sigma_\alpha = \sqrt{\frac{1}{\tau_\alpha}}$.

We assume the following vague prior distributions. In line with WinBUGS coding, the normal distribution is parameterized by mean and precision parameter.

$$\begin{aligned}a[k] &\sim N(0, 0.0001) \\ b &\sim N(0, 0.0001) \\ \nu[i] &\sim N(0, 0.0001) \\ \sigma_\alpha &\sim \text{unif}(0, 0.25)\end{aligned} \tag{10}$$

A normal distribution with mean 0 and precision 0.0001 has a 95% confidence interval of $(-196, 196)$, which is sufficiently vague for measuring the relative treatment effects $a[k]$ and b as well as the baseline effects $\nu[i]$, which are measured as proportions and differences between proportions and therefore take values much smaller than this. For the between-trial standard deviation, a uniform prior in [0, 0.25] is sufficiently vague, because it allows the treatment effect to differ up to 100% between trials.

2.2.1. Naive pooling. Naive pooling treats all trial designs the same, one simply pools across all trials regardless their design. This is the simplest approach to combining different trial designs; it is not possible to downweigh designs of lesser quality or to adjust for potential bias.

Generally, we would not recommend this approach. However, in this particular application, pooling across all trials has one advantage: the resulting evidence network contains closed loops, Figure 1, which allows the estimation of consistency between direct and indirect evidence.

The MTC model for naive pooling remains the same. Because the observational trials contain more than two treatments per trial, a term allowing for the correlation between-trial arms has been included in the code. As proposed by the Decision Support Unit [47], inconsistency is first checked by fitting a model relaxing the consistency assumptions and comparing it to the original model. The deviance contributions of the data points can help to identify inconsistencies in the network. When trial level data are available, a node-split method has been proposed to estimate the consistency between direct and indirect evidence [48]. The evidence for each comparison that is part of a closed loop is split into evidence based on direct and evidence based on indirect sources. In this case, we can split the evidence for the following comparisons: (adalimumab versus placebo), (infliximab versus placebo), (etanercept versus placebo), (adalimumab versus infliximab), (adalimumab versus etanercept) and (infliximab versus etanercept). The code for this method provided by Dias *et al.* [48] can be easily adapted to fit this model.

2.2.2. As prior information. In the Bayesian framework, observational data can also be used to inform prior information. For this approach, observational trials are analysed separately, and results are then used as prior information for the RCT model. Only slight changes in the model are necessary. For the analysis of observational data alone, no adjustment for MTX needs to be made. MTX is given in each trial arm of all studies and hence cancels out because of the assumption made in Equation (5). Results from the observational analysis inform the basic parameters in the RCT model. The strongest link of the observational network should be used to make full use of the additional information. The basic parameters should therefore be adapted to correspond to the relative efficacy of each treatment versus etanercept, the only treatment that is directly connected to all other interventions in the network. The comparator of the basic parameters is defined in Equation (9). It thus becomes

$$a[4] = 0 \tag{11}$$

This approach allows adjustment for bias. The potential biases associated with observational data described earlier can be modelled by adjusting the prior distribution. To downweigh observational information, we can inflate the variance parameter, and to adjust for overestimation or underestimation of the treatment effect, we can shift the mean of the prior information. Because the actual bias is typically not known, bias adjustment is carried out as a sensitivity analysis.

2.2.3. Hierarchical modelling. A three-level Bayesian hierarchical model introduces a study type level between the study level and the overall level. A schematic illustration of such a model for the simplest case of an indirect comparison can be found in Figure 2. This is the most flexible approach to include different forms of evidence. In addition to being able to adjust for systematic bias and downweigh

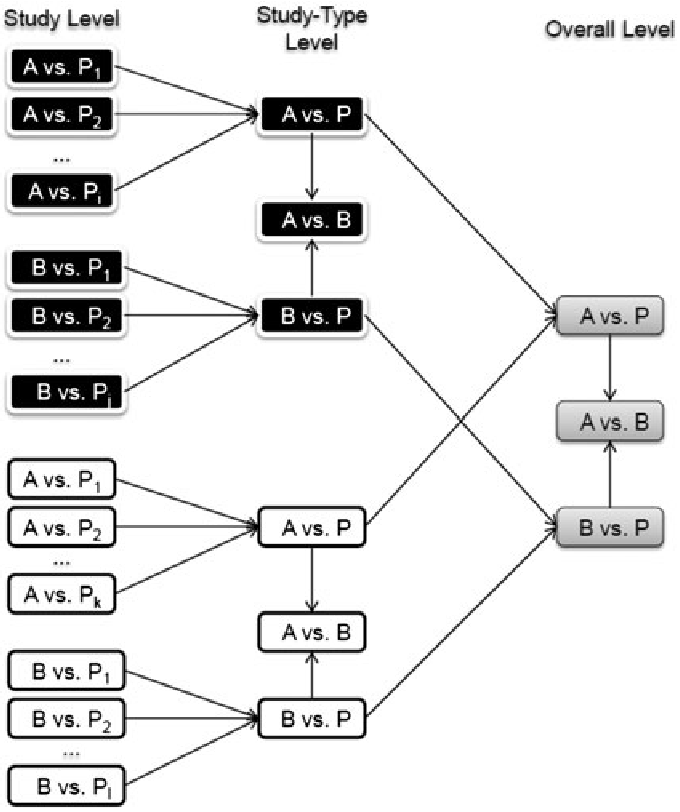

Figure 2. Three-level hierarchical model: schematic graph of three-level hierarchical model for simplest case indirect comparison, where trials comparing drug A and placebo (P) are combined with trials comparing drug B and P. Evidence from different trial designs is colour marked: black refers to randomized controlled trials and white to observational trials. Evidence from trials of the same design is combined to study type level estimates; study type estimates are in turn combined to obtain overall estimates (grey). Pairwise indirect estimates can be obtained on study type level as well as overall level.

different trial designs, this method allows the direct comparison of estimates on study type level to overall estimates and therefore gives an estimate of consistency between the designs. The model accounts for between design heterogeneity, and overall estimates become more conservative when study type estimates differ much.

Within the model, RCT data and observational data are each fit separately as described in the beginning of Section 2.2; that is, for each of the parameters introduced earlier, there exists a version for each trial design. Let $a_{\mathrm{RCT}}[k]$ and $a_{\mathrm{OBS}}[k]$ be the biologic treatment effects of drug k versus the baseline treatment based on RCT evidence and observational data, respectively. Assuming that RCT and OBS evidence are exchangeable, the trial design estimates are then combined to an overall measure of treatment effect using random effects:

$$\begin{aligned} a_{\mathrm{RCT}}[k] &\sim N(a[k], \tau) \\ a_{\mathrm{OBS}}[k] &\sim N(a[k], \tau) \end{aligned} \tag{12}$$

Prior distributions need to be specified for $a[k]$ and $\sigma = \sqrt{\frac{1}{\tau}}$; we assume the same vague prior distributions as before:

$$\begin{aligned} a[k] &\sim N(0, 0.0001) \\ \sigma &\sim \mathrm{unif}(0, 0.25) \end{aligned} \tag{13}$$

Relative efficacy between agents k and l can now be calculated from the baseline parameters as follows:

$$\mathrm{IC}[k,l] = a[k] - a[l] \tag{14}$$

The assumption of exchangeability between RCT evidence and observational evidence in Equation (12) can be relaxed by adjusting for potential bias in the various trial designs. We can adjust for overestimation using an additive factor to the mean or for overprecision using a multiplicative factor to the variance. Equation (12) then takes the following form:

$$\begin{aligned} a_{\mathrm{RCT}}[k] &\sim N(a[k], \tau) \\ a_{\mathrm{OBS}}[k] &\sim N(a[k] + \mu_{\xi}, \tau \cdot w_{\xi}) \end{aligned} \tag{15}$$

The code for the three-level hierarchical model can be found in supplementary material.[‡]

3. Results

3.1. Naive pooling

Table II summarizes the relative efficacy results from naive pooling. Results based on the RCT evidence are only listed for comparison. The results reveal how the additional information from observational data strengthens the evidence; the standard deviations of estimates including placebo, adalimumab, infliximab and etanercept become smaller when including observational data. Estimates not including these treatments do not change notably. Furthermore, we can see the impact of additional evidence on the mean effect estimates. The strongest effect is found in the relative efficacy of etanercept versus placebo. Only on the basis of RCT evidence, etanercept is estimated to improve the HAQ score by 31%; this effect shrinks to 21% when including observational trials. This suggests discrepancy between evidence from RCTs and evidence from observational trials. When including evidence other than RCTs, the difference of effect between the agents is less.

In the aforementioned model, all comparisons can be written in terms of the basic parameters. To detect inconsistencies in the network on the basis of RCT as well as observational data a model relaxing this consistency assumption is fit. The model estimates all relative treatment effects for which direct evidence is available without forcing the consistency in loops in the network. In the presence of inconsistency, this model should provide an improved fit to the data. Further details and the code for this model are provided elsewhere [47]. Figure 3 plots the posterior mean deviance of each data point based on the

[‡] *Supporting information may be found in the online version of this article.*

Table II. Naive pooling results.

Comparison (A vs B)	RCT only	Naive pooling	Inconsistency model
Ada vs P	0.21 (0.03)	0.22 (0.02)	0.21 (0.03)
Inf vs P	0.11 (0.04)	0.13 (0.03)	0.11 (0.04)
Eta vs P	0.31 (0.05)	0.21 (0.03)	0.26 (0.05)
Gol vs P	0.21 (0.05)	0.21 (0.06)	0.21 (0.06)
Cert vs P	0.26 (0.03)	0.26 (0.03)	0.26 (0.03)
Inf vs Ada	−0.10 (0.05)	−0.09 (0.03)	−0.10 (0.04)
Eta vs Ada	0.10 (0.06)	−0.01 (0.03)	−0.04 (0.04)
Gol vs Ada	0.01 (0.06)	−0.01 (0.06)	
Cert vs Ada	0.05 (0.04)	0.04 (0.04)	
Eta vs Inf	0.20 (0.07)	0.09 (0.03)	
Gol vs Inf	0.11 (0.07)	0.09 (0.07)	
Cert vs Inf	0.15 (0.05)	0.13 (0.04)	
Gol vs Eta	−0.10 (0.07)	0.00 (0.07)	
Cert vs Eta	−0.06 (0.06)	0.05 (0.05)	
Cert vs Gol	0.04 (0.06)	0.04 (0.07)	
DIC ($\overline{D}$, p_D)	−91.49 (−113.1, 21.7)	−115.2 (−143.8, 28.3)	−114.5 (−143.4, 28.9)

Difference in % Health Assessment Questionnaire improvement for each pairwise comparison (standard deviation). RCT, randomized controlled trial; DIC, deviance information criterion; $\overline{D}$, posterior mean residual deviance; p_D, effective number of parameters; Ada, adalimumab; Inf, infliximab; Eta, etanercept; Gol, golimumab; Cert, certolizumab; P, placebo.

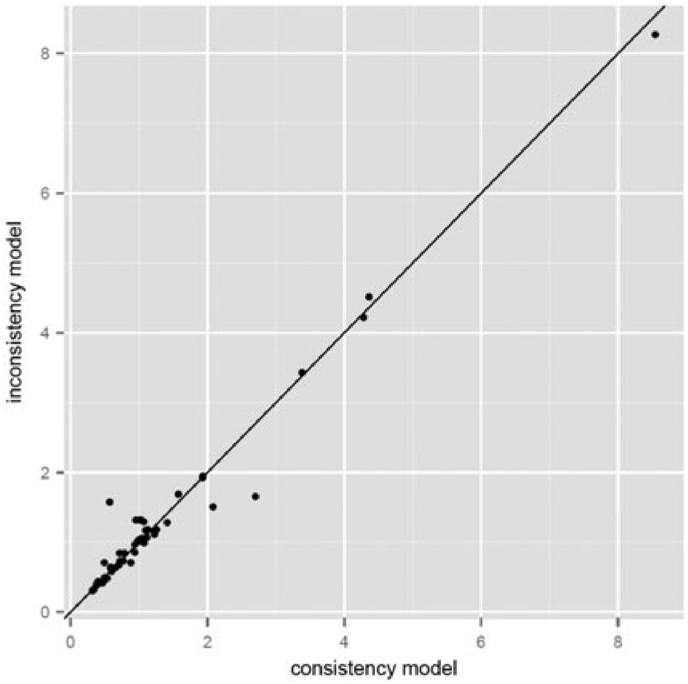

Figure 3. Plot of the individual data points' posterior mean deviance contribution for the original model (horizontal axis) and the model relaxing the consistency assumption (vertical axis).

model relaxing the consistency assumption against the original model. The contributions to the deviance are similar in both models. Arms 3 and 4 of Van de Putte *et al.* [35] as well as arm 4 of Miyasaka *et al.* [36] and arm 2 of Maini *et al.* [38] yield a deviance higher than expected. This can be attributed to the fixed effect assumption within trials. The efficacy estimates in these trial arms differ from those in the remaining arms of the same trial. The efficacy estimates based on the inconsistency model are also summarized in Table II. Most parameter estimates are very similar for both models, which suggests no evidence of inconsistency in the network. A slightly larger change is observed for the estimates including etanercept. For these parameters, the node-split method can be used to estimate the inconsistency.

Table III. Node split results.

	Eta vs P	Ada vs Eta	Inf vs Eta
DIC ($\overline{D}, p_D$)	−115.6 (−143.4, 27.8)	−115.2 (−142.9, 27.7)	−115.5 (−143.1, 27.6)
Direct	0.26 (0.04)	0.04 (0.04)	−0.06 (0.04)
Indirect	0.17 (0.04)	−0.05 (0.05)	−0.15 (0.05)

Mean difference in % Health Assessment Questionnaire improvement based on *direct* and *indirect* evidence (standard deviation). $\overline{D}$, posterior mean residual deviance; p_{D}, effective number of parameters; DIC, deviance information criterion; Ada, adalimumab; Inf, infliximab; Eta, etanercept; P, placebo.

The node-split method splits the evidence for the comparison of interest into direct and indirect sources. The differences between the estimates give a measure of the inconsistency. The deviance information criterion (DIC) is reported as a measure of model fit. In the presence of inconsistency, the node-split method provides an improvement in model fit. Results are summarized in Table III. While posterior distributions overlap, the results show a discrepancy between direct and indirect estimates for comparisons including etanercept. However, the decreases in DIC are relatively small, indicating a similar fit for all models. A change in DIC of more than 2 is said to indicate a significant improvement in model fit [49].

3.2. *As prior information*

Analysing the observational data on its own yields the following inputs as prior distributions for the basic parameters in the RCT model:

$$\begin{aligned} a[1] &\sim N(-0.18, 71) \\ a[2] &\sim N(0.04, 168) \\ a[3] &\sim N(-0.07, 168) \end{aligned} \tag{16}$$

As described earlier, the model has been adapted to define etanercept as the basic treatment; $a[1]$ now represents the efficacy of placebo versus etanercept, $a[2]$ the efficacy of adalimumab versus etanercept and $a[3]$ the efficacy of infliximab versus etanercept.

It is possible to reduce the weight given to observational data adjusting for a bias of the form given in Equation (1), which results in the inflation of the variance parameter:

$$a[i] \sim N(\mu_i, \tau_i \cdot w_\xi)$$

where w_ξ is the weight given to the observational evidence. Table IV shows the efficacy estimates including observational data as prior information for a range of different values of w_ξ. Estimates for golimumab and certolizumab are not affected because there is no additional information provided and are hence not included in the table. The results show how decreasing weight on observational information shifts the efficacy estimates towards the estimates based on RCTs only. This effect is most evident for the etanercept versus placebo estimate, where the largest difference between estimates is observed.

Table IV. Observational data as prior information results.

Comparison (A vs B)	No adj.	$w_\xi = 0.7$	$w_\xi = 0.5$	$w_\xi = 0.3$	$w_\xi = 0.1$	+30%	−30%
Ada vs P	0.22 (0.03)	0.22 (0.03)	0.22 (0.03)	0.21 (0.03)	0.21 (0.04)	0.22 (0.03)	0.21 (0.03)
Inf vs P	0.12 (0.04)	0.12 (0.04)	0.12 (0.04)	0.11 (0.04)	0.11 (0.04)	0.12 (0.04)	0.13 (0.04)
Eta vs P	0.24 (0.04)	0.25 (0.04)	0.25 (0.04)	0.27 (0.05)	0.27 (0.05)	0.25 (0.04)	0.23 (0.04)
Gol vs P	0.21 (0.06)						
Cert vs P	0.26 (0.03)						

Difference in % Health Assessment Questionnaire improvement for each biologic agent versus placebo (standard deviation). Results are shown for no bias adjustment (No adj.), for different values of w_ξ to account for overprecision and adjusting for a 30% overestimation and underestimation. Ada, adalimumab; Inf, infliximab; Eta, etanercept; Gol, golimumab; Cert, certolizumab; P, placebo.

To adjust for overestimation of the treatment effect, an additive factor μ_ξ onto the mean of the prior distribution is introduced.

$$a[i] \sim N(\mu_i + \mu_\xi, \tau)$$

The results assuming a 30% overestimation and 30% underestimation of the biologic treatment effect versus placebo are also summarized in Table IV. The results are not sensitive to bias due to overestimation or underestimation of the treatment effect. This can be explained by the structure of the observational evidence. Its strength lies in the relative efficacy of adalimumab, infliximab and etanercept, rather than in the comparisons with placebo. The relative efficacies between the biologic treatments are not affected greatly when assuming a 30% overestimation or underestimation of each agent when compared with placebo.

3.3. *Bayesian hierarchical model*

Fitting a three-level hierarchical model yields overall estimates of efficacy as well as estimates for each study type. Table V summarizes the biologic treatment effect estimates overall as well as study type level. The results show how RCT estimates and observational estimates are combined to overall estimates, which lie in between the trial design estimates. The overall mean effect estimates are the same as when including observational evidence as prior information. However, the standard deviation is slightly higher in the three-level model, because the model allows for uncertainty caused by combining evidence from different trial designs. Figure 4 shows shrinkage from data through each level to the estimate of the overall mean. Borrowing strength across the network results in the trial design level estimates being drawn towards the overall mean. This also explains the small changes in study type level estimates when adjusting for bias. The graph also shows that differences in effect between the treatments become smaller with the inclusion of observational evidence.

As before, it is possible to adjust for overprecision and overestimation. Table V summarizes the results when adjusting for overprecision and overestimation of the treatment effect in observational data. Adjusting for overprecision of observational data downweighs the impact of its information on the results. The results show how overall results shift towards the RCT only results when decreasing the weight on observational data; the same effect that has been found when using observational data as prior information. Study type level do not change and are therefore not recorded in the table. When adjusting for underestimation and overestimation, the effect on overall estimates is small. However, for the

Table V. Hierarchical model results.

Level	Comparison (A vs B)	No adj.	$w_\xi = 0.7$	$w_\xi = 0.5$	$w_\xi = 0.3$	$w_\xi = 0.1$	+30%	−30%
Overall	Ada vs P	0.22 (0.07)	0.22 (0.07)	0.22 (0.07)	0.22 (0.07)	0.21 (0.06)	0.23 (0.08)	0.21 (0.07)
	Inf vs P	0.12 (0.08)	0.12 (0.08)	0.12 (0.07)	0.12 (0.07)	0.12 (0.06)	0.12 (0.08)	0.12 (0.07)
	Eta vs P	0.24 (0.08)	0.25 (0.08)	0.25 (0.08)	0.26 (0.08)	0.28 (0.08)	0.27 (0.08)	0.22 (0.07)
	Gol vs P	0.21 (0.05)						
	Cert vs P	0.26 (0.03)						
RCT	Ada vs P	0.21 (0.02)					0.22 (0.03)	0.21 (0.02)
	Inf vs P	0.11 (0.04)					0.11 (0.04)	0.11 (0.04)
	Eta vs P	0.28 (0.06)					0.29 (0.05)	0.26 (0.06)
	Gol vs P	0.21 (0.05)						
	Cert vs P	0.26 (0.03)						
OBS	Ada vs P	0.23 (0.06)					0.18 (0.06)	0.27 (0.07)
	Inf vs P	0.13 (0.06)					0.09 (0.05)	0.17 (0.07)
	Eta vs P	0.21 (0.06)					0.17 (0.05)	0.25 (0.07)

Difference in % Health Assessment Questionnaire improvement for each biologic agent versus placebo (standard deviation) on study-type and overall levels. Results are shown for no bias adjustment (No adj.), for different values of w_ξ to account for overprecision and adjusting for a 30% overestimation and underestimation. RCT, randomized controlled trial; OBS, observational trial; Ada, adalimumab; Inf, infliximab; Eta, etanercept; Gol, golimumab; Cert, certolizumab; P, placebo.

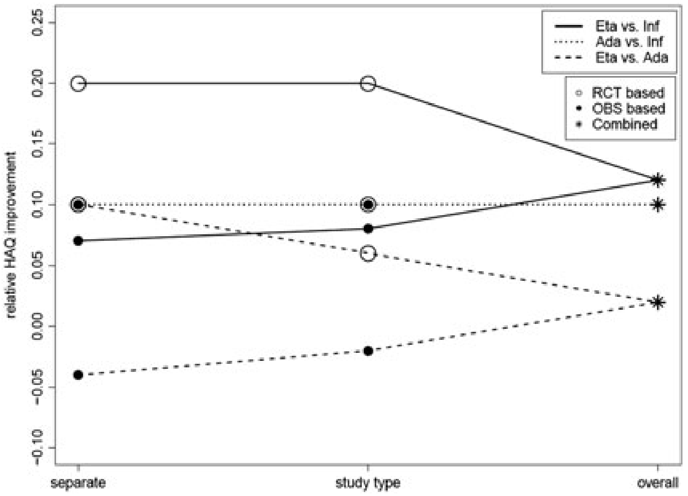

Figure 4. Borrowing strength across the network causes estimates to converge towards the overall mean when analysed within one model. This effect is called shrinkage. The figure plots relative efficacy estimates based on analysing randomized controlled trial data and observational data separately (separate), as well as based on the study type level when analysing data within one model (study type) and the combined efficacy estimates (overall).

etanercept effect, the results show an increase when adjusting for underestimation and a decrease when adjusting for overestimation of the treatment effect. On study type level, RCT estimates shift in line with the overall effects, whereas a larger effect is found for the observational estimates.

4. Discussion

The aforementioned techniques provide a framework for systematically including evidence from different trial designs in an MTC model. Naive pooling makes the strong assumptions of no differences between trial designs. The method does not allow for bias adjustments, and no additional uncertainty is taken into account. We therefore do not recommend naive pooling as a primary means of combining evidence from different trial designs. However, as demonstrated in this case study, in some situations, naive pooling may create the opportunity to estimate consistencies within the evidence network. In such situations, naive pooling may be worthwhile as the first step analysis. Summarizing observational evidence to inform the prior distribution in an MTC model allows the adjustment for potential bias because of overprecision or overestimation. This is a clear advantage over naive pooling; however, between trial design heterogeneity is not taken into account, and it is not possible to extend the model to include more than two different trial designs. The hierarchical modelling approach accounts for the uncertainty arising from combining information from different trial designs by using random effects. The hierarchy levels allow us to quantify the impact evidence from different designs have on the result while adjusting for potential bias. These advantages make hierarchical modelling the preferred method of including information from different designs.

Observational data are prone to two different biases, overestimation and overprecision of the treatment effect. Including observational data as prior information and in the form of a hierarchical model allows for bias adjustment. In the literature, assuming 30% overprecision has previously been used [8, 16]. Research has not found a consistent overestimation of the treatment effect among observational trials [17, 18]. In any case, the actual size of the bias is difficult to estimate, and it is important to vary this in a sensitivity analysis. Turner *et al.* [15] provided a detailed approach to modelling internal and external bias in evidence synthesis.

A challenge is given by open-label extensions. Such trials are typically one armed and cannot be included in the analysis without adjustment. In this case study, we took a matching approach, where baseline characteristics are compared across available trials to identify a suitable match [50]. However, such methods do not control for unobserved variables that may also affect the outcome. In a sensitivity analysis, we have excluded the one-armed trial [31] from the analysis; results based on the three methods are summarized in Table VI. Whereas the influence of observational data is slightly lower because only

Table VI. Results excluding Klareskog trial.

Comparison (A vs B)	NP	Prior	3-level
Ada vs P	0.22 (0.03)	0.22 (0.03)	0.23 (0.09)
Inf vs P	0.13 (0.03)	0.12 (0.04)	0.13 (0.09)
Eta vs P	0.22 (0.04)	0.25 (0.04)	0.26 (0.09)
Gol vs P	0.21 (0.06)	0.21 (0.06)	0.22 (0.05)
Cert vs P	0.26 (0.03)	0.26 (0.03)	0.26 (0.03)

Difference in % Health Assessment Questionnaire improvement for each biologic agent versus placebo (standard deviation). Results are shown for including observational evidence using naive pooling (NP), as prior information (Prior) and in a three-level hierarchical model (3-level). Ada, adalimumab; Inf, infliximab; Eta, etanercept; Gol, golimumab; Cert, certolizumab; P, placebo.

two additional trials are included in the analysis, the exclusion of Klareskog *et al.* [31] did not influence the results much. The main effect remains the drop in efficacy of the etanercept estimate. This is because the strength of the observational evidence lies in the relative efficacy between the anti-TNF agents. Adalimumab is doing better than etanercept in both three-armed observational trials [30, 45] (when considering baseline HAQ). The efficacy of etanercept changes more than the efficacy of adalimumab when including observational evidence, because the evidence for adalimumab in the RCT trials is stronger (five trials) than the evidence for etanercept (two trials). Welton *et al.* [51] have discussed the implications of including or excluding particular evidence sources in more detail.

For many disease areas, there are many observational data available providing additional information on treatment effectiveness. We think that it is important for an informed decision-making process to include all available evidence. Including observational data at base case analysis or in the form of a sensitivity analysis can greatly improve evidence synthesis as part of economic assessments, as well as choice of agent in a clinical setting. The methods described in this paper provide a flexible methodology to analyse the impact of such additional information.

References

1. Cooper N, Sutton A, Lu G, Khunti K. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. *Archives of Internal Medicine* 2006; **166**:1269 –1275.
2. Welton N, Cooper N, Ades A, Lu G, Sutton A. Mixed treatment comparison with multiple outcomes reported inconsistency across trials: evaluation of antivirals for treatment of influenza a and b. *Statistics in Medicine* 2008; **27**(27):5620–5639.
3. Schmitz S, Adams R, Walsh C, Barry M, FitzGerald O. A mixed treatment comparison of the efficacy of anti-TNF agents in rheumatoid arthritis for methotrexate non-responders demonstrates differences between treatments: a Bayesian approach. *Annals of Rheumatic Diseases* 2012; **71**(2):225–230.
4. Sutton A, Higgins J. Recent developments in meta-analysis. *Statistics in Medicine* 2008; **27**(5):625–650.
5. Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* 2004; **23**(20):3105–3124.
6. Cooper NJ, Peters J, Lai M, Juni P, Wandel S, Palmer S, Paulden M, Conti S, Welton N, Abrams K, Bujkiewicz S, Spiegelhalter D, Sutton A. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value in Health* 2011; **14**(2):371–380.
7. Rothwell P. External validity of randomised controlled trials: to whom do the results of this trial apply? *The Lancet* 2005; **365**(9453):82–93.
8. O'Rourke K, Walsh C, Hutchinson M. Outcome of beta-interferon treatment in relapsing-remitting multiple sclerosis: a Bayesian analysis. *Journal of Neurology* 2007; **254**(11):1547–54.
9. Ades A, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, Lu G. Bayesian methods for evidence synthesis in cost-effectiveness analysis. *Pharmacoeconomics* 2006; **24**(1):1–19.
10. Prevost T, Abrams K, Jones D. Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine* 2000; **19**(24):3359–3376.
11. Mak A, Cheung M, Ho R, AiCia Cheak A, Lau C. Bisphosphonates and atrial fibrillation: Bayesian meta-analyses of randomized controlled trials and observational studies. *BMC Musculoskeletal Disorders* 2009; **10**(113).
12. Salpeter S, Cheng J, Thabane L, Buckley N, Salpeter E. Bayesian meta-analysis of hormone therapy and mortality in younger postmenopausal women. *The American Journal of Medicine* 2009; **122**(11):1016–1022.
13. McCarron C, Pullenayegum E, Thabane L, Goeree R, Tarride J. The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms. *BMC Medical Research Methodology* 2010; **10**(64).
14. Spiegelhalter D, Abrams K, Myles J. *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. Wiley: NewYork, 2004.

15. Turner R, Spiegelhalter D, Smith G, Thompson S. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A* 2009; **172**(1):21–47.
16. Lilford R, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal* 1996; **313**:603–607.
17. Benson K, Hartz A. A comparison of observational studies and randomized, controlled trials. *The New England Journal of Medicine* 2000; **342**(25):1878–1886.
18. Concato J, Shah N, Horwitz R. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine* 2000; **342**:1887–1892.
19. Klareskog L, Catrina A, Paget S. Rheumatoid arthritis. *The Lancet* 2009; **979**:659–672.
20. Bergman G, Hochberg M, Boers M, Wintfeld N, Kielhorn A, Jansen J. Indirect comparison of tocilizumab and other biologic agents in patients with rheumatoid arthritis and inadequate response to disease-modifying antirheumatic drugs. *Rheumatoid Arthritis* 2010; **39**:425–441.
21. Nixon R, Bansback N, Brennan A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Statistics in Medicine* 2007; **26**(6):1237–1254.
22. Devine E, Alfonso-Cristancho R, Sullivan S. Effectiveness of biologic therapies for rheumatoid arthritis: an indirect comparison approach. *Pharmacotherapy* 2011; **31**(1):39–51.
23. Launois R, Avouac B, Berenbaum F, Blin O, Bru I, Fautrel B, Jouvert J, Sibilia J, Combe B. Comparison of certolizumab pegol with other anticytokine agents for treatmen of rheumatoid arthritis: a multiple-treatment Bayesian metaanalysis. *Journal of Rheumatology* 2011; **38**:1–11.
24. Turkstra W, NG S, Scuffham P. A mixed treatment comparison of the short-term efficacy of biologic disease modifying anti-rheumatic drugs in established rheumatoid arthritis. *Current Medical Research and Opinion* 2011; **27**(10):1885–1897.
25. Guyot P, Taylor P, Christensen R, Pericleous L, Poncet C, Lebmeier M, Drost P, Bergman G. Abatacept with methotrexate versus other biologic agents in treatment of patients with active rheumatoid arthritis despite methotrexate: a network meta-analysis. *Bayesian Analysis* 2011; **13**:1–18.
26. Saag K, Teng G, Patkar N, Anuntiyo J, Finney C, Curtis J, Paulus H, Mudano A, Piso M, Elkins-Melton M, Outman R, Allison J, Suarez Almazor M, Bridges A, Chatham W, Hochberg M, Maclea C, Mikuls T, Moreland L, O'Dell J, Turkiewicz A, Furst D. American College of Rheumatology 2008 recommendations for the use of nonbiologic and biologic disease-modifying antirheumatic drugs in rheumatoid arthritis. *Arthritis & Rheumatism* 2008; **59**(6):762–784.
27. Schmitz S, Adams R, Walsh C. The use of continuous data versus binary data in MTC models: a case study in rheumatoid arthritis. *BMC Medical Research Methodology* 2012; **12**(167). DOI: 10.1186/1471-2288-12-167.
28. Hyrich K, Watson K, Silman A, Symmons D, The BSR Biologics Register. Predictors of response to anti-TNF-α therapy among patients with rheumatoid arthritis: results from the british society for rheumatology biologics register. *Rheumatology* 2006; **45**(12):1558–1565.
29. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* 2009; **6**(7):e1000097. DOI: 10.1371/journal.pmed.1000097.
30. Kievit W, Adang E, Fransen J, Kuper H, van de Larr M, Jansen T, De Gendt C, De Rooij D, Brus H, Van Oijen P, Van Riel P. The effectiveness and medication costs of three anti-tumour necrosis factor alpha agents in the treatment of rheumatoid arthritis from prospective clinical practice data. *Annals of the Rheumatic Diseases* 2008; **67**:1229–1234.
31. Klareskog L, Gaubitz M, Rodriguez-Valverde V, Malaise M, Dougados M, Wajdula J. A long-term, open-label trial of the safety and efficacy of etanercept (enbrel) in patients with rheumatoid arthritis not treated with other disease-modifying antirheumtic drugs. *Annals of the Rheumatic Diseases* 2006; **65**:1578–1584.
32. Weinblatt M, Kremer J, Bankhurst A, Bulpitt K, Fleischmann R, Fox R, Jackson C, Lange M, Burge D. A trial of etanercept, a recombinant tumor necrosis factor receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. *New England Journal of Medicine* 1999; **340**(4):253–259.
33. Weinblatt M, Keystone E, Furst D, Moreland L, Weisman M, Birbara C, Teoh L, Fischkoff S, Chartash E. Adalimumab, a fully human anti-tumour necrosis factor alpha monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial. *Arthitis & Rheumatism* 2003; **48**(1):35–45.
34. Keystone E, Kavanaugh A, Sharp J, Tannenbaum H, Hua Y, Teoh L, Fischkoff S, Chartash E. Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy: a randomized, placebo-controlled, 52-week trial. *Arthritis & Rheumatism* 2004; **50**:1400–1411.
35. Van De Putte LBA, Atkins C, Malaise M, Sany J, Russell A, van Riel P, Settas L, Bijlsma J, Todesco A, Dougados M, Nash P, Emery P, Walter N, Kaul M, Fischkoff A, Kupper H. Efficacy and safety of adalimumab as monotherapy in patients with rheumatoid arthritis for whom previous disease modifying antirheumatic drug treatment has failed. *Annals of the Rheumatic Diseases* 2004; **63**:508–516.
36. Miyasaka N. Clinical investigation in highly disease-affected rheumatoid arthritis patients in Japan with adalimumab applying standard and general evaluation: the CHANGE study. *Modern Rheumatology* 2008; **18**(3):252–262.
37. Kim H, Lee K, Song Y, Dae-Hyun Y, Koh E, Yoo B, Luo A. A randomized, double-blind, placebo-controlled, phase III study of the human anti-tumor necrosis factor antibody adalimumab administered as subcutaneous injections in Korean rheumatoid arthritis patients treated with methotrexate. *APLAR Journal of Rheumatology* 2007; **10**(1):9–16.
38. Maini R, St Clair EW, Breedveld F, Furst D, Kalden J, Weisman M, Smolen J, Emery P, Harriman G, Feldmann M, Lipsky P. Infliximab (chimeric anti-tumour necrosis factor [alpha] monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a randomised phase III trial. *The Lancet* 1999; **353**(9194): 1932 –1939.
39. Zhang F, Hou Y, Huang F, Wu D, Bao C, Ni L, Yao C. Infliximab versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a preliminary study from China. *APLAR Journal of Rheumatology* 2006; **9**(2): 127–130.

40. Moreland L, Schiff M, Baumgartner A, Tindall E, Fleischmann R, Bulpitt K, Weaver A, Keystone E, Furst D, Mease P, Ruderman E, Horwitz D, Arkfeld D, Garrison L, Burge D, Blosch C, Lange M, McDonnell N, Weinblatt M. Etanercept therapy in rheumatoid arthritis: a randomized, controlled trial. *Annals of Internal Medicine* 1999; **130**(6):478–486.
41. Keystone EC, Genovese MC, Klareskog L, Hsia E, Hall A, Miranda P, Pazdur J, Bae S, Palmer W, Zrubek J, Wiekowski M, Visvanathan S, Wu Z, Rahman M. Golimumab, a human antibody to tumour necrosis factor (alpha) given by monthly subcutaneous injections, in active rheumatoid arthritis despite methotrexate therapy: the GO-FORWARD study. *Annals of the Rheumatic Diseases* 2009; **68**:789–769.
42. Keystone E, Van Der Heijde D, Mason D, Landewe R, van Vollenhoven R, Combe B, Emery P, Strand V, Mease P, Desai C, Pavelka K. Certolizumab pegol plus methotrexate is significantly more effective than placebo plus methotrexate in active rheumatoid arthritis: findings of a fifty-two-week, phase III, multicenter, randomized, double-blind, placebo-controlled, parallel-group study. *Arthritis & Rheumatism* 2008; **58**:3319–3329.
43. Smolen J, Landewe RB, Mease P, Brzezicki J, Mason D, Luijtens K, van Vollenhoven R, Kavanaugh A, Schiff M, Burmester G, Strand V, Vencovsky J, van der Heijde D. Efficacy and safety of certolizumab pegol plus methotrexate in active rheumatoid arthritis: the rapid 2 study. A randomised controlled trial. *Annals of the Rheumatic Diseases* 2008; **68**:797–804.
44. Fleischmann R, Vencovsky J, Van Vollenhoven R, Borenstein D, Box J, Coteur G, Goel N, Brezinschek H, Innes A, Strand V. Efficacy and safety of certolizumab pegol monotherapy every 4 weeks in patients with rheumatoid arthritis failing previous disease-modifying antirheumatic therapy: the fast4ward study. *Annals of the Rheumatic Diseases* 2009; **68**:805–811.
45. Bazzani C, Filippini M, Caporali R, Bobbio-Pallavicini F, Favalli E, Marchesoni A, Atzeni F, Sarzi-Puttini P, Gorla R. Anti-TNF alpha therapy in a cohort of rheumatoid arthritis patients: clinical outcomes. *Autoimmunity Reviews* 2009; **8**(3):260–265.
46. Ades A, Welton N, Lu G. *Introduction to Mixed Treatment Comparison*. University of Bristol: Bristol, UK, 2007. https://www.bris.ac.uk/cobm/research/mpes/mtc.html.
47. Dias S, Welton N, Sutton A, Caldwell D, Lu G, Ades A. NICE DSU technical support document 4: inconsistency in networks of evidence based on randomised controlled trials: report by the decision support unit. *Technical Report*, Decision Support Unit, ScHARR, University of Sheffield, 2012.
48. Dias S, Welton N, Caldwell D, Ades A. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* 2010; **29**:932–944.
49. Spiegelhalter D, Best N, Carlin B, van de Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* 2002; **64**(4):583–639.
50. d'Agostino R Jr., d'Agostino R Sr. Estimating treatment effects using observational data. *Journal of the American Medical Association* 2007; **297**(3):314–316.
51. Welton N, Ades A, Carlin B, Altman D, Sterne J. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A* 2009; **172**(1):119–136.