A review of methods for comparing treatments evaluated in studies which form disconnected networks of evidence

John W Stevens, Christine Fletcher, Gerald Downey, Anthea Sutton

Abstract

A network meta-analysis allows a simultaneous comparison between treatments evaluated in RCTs that share at least one treatment with at least one other study. Estimates of treatment effects may be required for treatments across disconnected networks of evidence, which requires a different statistical approach and modelling assumptions to account for imbalances in prognostic variable and treatment effect modifiers between studies. In this paper, we review and discuss methods for comparing treatments evaluated in studies which form disconnected networks of evidence. Several methods have been proposed but assessing which are appropriate often depends on the clinical context as well as the availability of data. Most methods account for sampling variation but do not always account for others sources of uncertainty. We suggest that further research is required to assess the properties of methods and the use of approaches that allow the incorporation of external information to reflect parameter and structural uncertainty.

1. Introduction

The National Institute for Health and Care Excellence (NICE) is responsible for making recommendations on the use of new treatments within the National Health Service in England. Amgen was recently invited to submit evidence to NICE in support of a Single Technology Appraisal (STA) (National Institute for Health and Care Excellence, 2015) of the clinical and cost-effectiveness of talimogene laherparepvec, an oncolytic viral immunotherapy derived from the herpes simplex virus type-1 (Kaufman et al., 2015), for the treatment of metastatic melanoma, a rare but serious form of skin cancer, within its European Union marketing authorisation i.e. adults with unresectable melanoma that is regionally or distantly metastatic (Stage IIIB, IIIC and IVM1a) with no bone, brain, lung or other visceral disease. The comparator treatments of interest were those representing the current standard of care in the UK: ipilimumab, vemurafenib and dabrafenib (for people with BRAF V600 mutation positive disease).

Talimogene laherparepvec has been evaluated against subcutaneous granulocyte-macrophage colony-stimulating factor (GM-CSF) in an open-label Phase 3 randomised controlled trial (RCT) known as OPTiM (Andtbacka et al., 2015). However, OPTiM did not include any of the comparator treatments of interest because ipilimumab, vemurafenib and dabrafenib were not available when the OPTiM study was designed or when the first subject was enrolled in April 2009. Therefore, a network meta-analysis (NMA) allowing an indirect comparison between talimogene laherparepvec and ipilimumab, vemurafenib and dabrafenib was required.

An NMA is an extension of a standard pairwise meta-analysis that coherently summarises all direct and indirect evidence about treatment effects and allows a simultaneous comparison to

be made between all pairs of treatments (Dias et al., 2013). The assumptions made in an NMA are: 1) the studies to be synthesised form a connected network of evidence such that there is a chain of pairwise comparisons that connects every treatment to every other treatment (a network that is connected provides an anchored indirect comparison with respect to a reference treatment); 2) randomisation is not broken so that treatment effects are estimated within studies before being combined across studies; 3) for every study included in the network, irrespective of the treatments that were actually compared, the true effect of Treatment B relative to Treatment A in Study i, δ_{iAB}, is the same in a fixed effect model (i.e. $\delta_{iAB} = d_{AB}$) or exchangeable between studies in a random effects model (i.e. $\delta_{iAB} \sim N(d_{AB}, \tau^2)$). An NMA makes use of the consistency equations which state that for any three treatments X, Y, Z, say, the population mean effects, d_{XY}, d_{ZY} and d_{ZX} are related such that:

$$d_{XY} = d_{ZY} - d_{ZX}.$$

It is assumed that the distribution of treatment effect modifiers is balanced between the Z, Y and Z, X studies, otherwise the indirect estimate of d_{XY} will be biased.

In situations when an anchored indirect comparison is not possible because studies do not share a common treatment, naïve or unadjusted indirect treatment comparison (ITC) could be performed by ignoring differences between studies in variables that affect response and effectively assuming that the data on each treatment arose from a single RCT (Song et al., 2003). When several studies evaluate a particular treatment, a naïve ITC would involve an arm-based synthesis of evidence across studies. Naïve ITCs and arm-based models have been criticised for potentially generating biased estimates of relative treatment effect by ignoring the randomisation within studies and are generally not recommended (Dias and Ades, 2016).

In the absence of a connected network of evidence, it is sometimes possible to form a connected network by adding one or more treatments to the comparator decision set to create a chain of pairwise comparisons that connects at least one treatment in each network to at least one treatment in another network, thereby forming an extended synthesis decision set (Ades et al., 2013). When this is not possible it will be necessary to use alternative methods of analysis and/or to make additional modelling assumptions to allow a valid ITC. Such modelling gives rise to an unanchored comparison in which there is no common reference treatment in each study.

The aim of this paper is to present the findings of a review of evidence synthesis methods to estimate the relative effect of treatments evaluated in studies forming disconnected networks of evidence. Although we mention methods for making indirect comparisons between treatments that can be applied when individual patient-level (IPD) are available for treatments from all studies (Faria et al., 2015), our focus is on methods that can be applied in situations where IPD is available on one or more studies but only aggregate data is available from studies of other comparator treatments. In addition, our interest was in methods for making indirect comparisons between treatments that have been, or can be applied, to data from studies of patients with advanced melanoma.

The paper is organised as follows: Section 2 describes the systematic review and presents the evidence network for talimogene laherparepvec and the comparator treatments; Section 3 describes the systematic review of evidence synthesis methods for comparing treatments across disconnected networks; Section 4 describes the methods that have been used to compare treatments across disconnected networks of evidence; Section 5 provides a discussion; Section 6 provides some concluding remarks.

2. Evaluation of the evidence network for talimogene laherparepvec

Amgen conducted a systematic literature review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Moher et al., 2009) to identify published RCTs (and non-RCTs) for the treatment of patients with metastatic malignant melanoma. MEDLINE, EMBASE and the Cochrane Central Register of Controlled Trials databases were searched from 1990 until 3 June 2015 using Ovid. Additional records were identified through other sources including clinical trial registries, previous regulatory and health technology assessment (HTA) reviews and conference abstracts. 3475 records were identified of which 97 records related to 59 RCTs. The inclusion criteria were restricted to Phase 3 RCTs published since 2010 to reflect the more recent and relevant melanoma treatments and their studies. Ten studies met the additional inclusion criterion. The systematic literature review found that the OPTiM study was isolated, having no treatment in common with the evidence base formed by the comparator treatments (Figure 1). Therefore, a conventional NMA of studies comprising the comparator decision set was infeasible.

[Figure 1 about here]

Evidence from studies comparing additional treatments was added to the comparator decision set in an attempt to form a connected network but was unsuccessful. Consequently, it was necessary to use alternative methods of analysis to generate indirect estimates of relative treatment effects.

3. Systematic review of evidence synthesis methods for comparing treatments across disconnected networks

A two-stranded approach was used to systematically review evidence synthesis methods in the scientific literature that allow comparisons to be made between treatments across disconnected networks of evidence (Figure 2). This entailed keyword searching of the MEDLINE database (via OvidSP) and pearl growing based on citation searching of 11 published journal articles dealing with novel approaches to making indirect comparisons between treatments forming disconnected networks (Ahn and French, 2010, Caro and Ishak, 2010, Gross et al., 2013, Ishak et al., 2015, Korn et al., 2008, Mandema, 2011, Mandema et al., 2011, Mercier et al., 2014, Signorovitch et al., 2010, Signorovitch et al., 2012b, Thom et al., 2015).

3.1 Keyword searching

Consideration was given to developing and implementing a keyword search strategy that was both sensitive and specific. The main objective was to identify methods that allow indirect

comparisons to be made between treatments because there is no direct evidence. However, using the terms "indirect comparisons" and "direct evidence" were thought to be insufficiently specific. A more specific phrase that encompassed the issue of there being no relevant or direct evidence in a target patient population was variants of "no head-to-head" and "absence of head-to-head" in combination with terms for "network meta-analysis", although we recognised that this did not specifically focus on methods for comparing treatments across disconnected networks. Therefore, we also included the term "disconnected network" in combination with terms for "meta-analysis". Details of the search strategy are provided in the Appendix.

MEDLINE via OvidSP (1946-Present including MEDLINE In-Process) was searched on 26 August 2015. 23 references were retrieved based on the keyword searching with 19 references remaining once duplicates were removed. One article was removed because it was one of the 11 originally known to the authors. None of the remaining 18 articles involved making comparisons between treatments across disconnected networks but were applications of conventional network meta-analyses.

3.2 Citation searching

Citation searches were conducted for each of the 11 published journal articles using the "cited reference" search feature of Web of Science. 343 cited references were retrieved with the majority of them (i.e. 258) referring to a single journal article (Korn et al., 2008). 328 unique references remained once duplicates were removed. The titles and, where necessary, abstracts of each article were reviewed to identify potentially relevant articles. 285 articles were excluded based on the titles and abstracts leaving 43 articles for consideration. Two articles were excluded because they were foreign language papers. Five articles were excluded because they were one of the 11 used in the citation searching process. Of the remaining 36 articles, full text articles were reviewed during which a further eight articles were excluded because they discussed the general issue of comparative effectiveness or were about specific clinical aspects, leaving 28 articles (Assawasuwannakit et al., 2015, Demin et al., 2012, Denney and Nucci, 2013, Dequen et al., 2012, Di Lorenzo et al., 2011, Di Lorenzo et al., 2012, Feng et al., 2013, Gibbs et al., 2012, Hoaglin, 2013, Kimko et al., 2012, Li et al., 2015a, Li et al., 2015b, Mandema et al., 2014, Mould, 2012a, Nie et al., 2013, Ravva et al., 2014, Reddy et al., 2013, Salinger et al., 2013, Signorovitch et al., 2011a, Signorovitch et al., 2011b, Signorovitch et al., 2012a, Signorovitch et al., 2015, Sikirica et al., 2013, Tiu and Kalaycio, 2012, Van Wart et al., 2013, Van Wart et al., 2014, Zhao et al., 2012, Zhou and Al-Huniti, 2013, Nixon et al., 2014).

4. Methods used to compare treatments in studies forming disconnected networks

The systematic review found that a variety of methods have been used or are available to estimate relative treatment effects when studies comprising the evidence base form a disconnected network. The following description of the methods used is based on the 28 articles identified through the citation searching, the 11 articles already known to the authors, and articles referred to by, or in response to, the reviewers.

4.1 Use of external controls

One approach that could be used to link disconnected networks is to make use of external evidence about the expected response to a control treatment in one or more studies. In the context of RCTs, the problems associated with the use of historical controls (such as a group of untreated patients at an earlier time) are well known and relate to the lack of randomisation and control for known and unknown baseline characteristics that might affect outcomes (International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2000). Studies making direct use of historical data make the constancy assumption that the expected response to a control treatment remains constant between the historical studies and the additional studies, which is unlikely, although some analysts have been known to use sample historical control data as if it had been generated as part of an RCT.

An alternative approach is to use the historical evidence to formulate a prior distribution for a parameter, μ_i, say, such as the mean for a continuous outcome or the log odds for a binary outcome for an external control treatment in study i in at least one study in each group of disconnected studies in order to facilitate a Bayesian random effects network meta-analysis (Schmidle et al., 2014, Viele et al., 2014). Spiegelhalter et al. (2004) (Spiegelhalter et al., 2004) discuss five different approaches to using historical control data, y_h, $h = 1, \dots, H$, that provide information about parameters θ_h from H studies:

- The simplest and most common approach that is used when analysing data arising from an RCT is to ignore the historical control data either because inferences are required to be made based on the sample data from the new RCT alone or because the historical data is thought not to provide any information about the parameter of interest.
- Alternatively, we could assume that the historical control groups are exchangeable such that $\theta_h \sim N(\theta, \tau^2)$. The prior distribution for μ_i is then the predictive distribution of the effect in a new study rather than the posterior distribution of the random effects mean, θ.
- We could assume that the θ_h are related to the target parameter such that $\theta_h = \mu_i + \delta_h$ where δ_h represents a bias that depends on study characteristics. The bias could be assumed to be known with some fixed value or unknown with probability distribution $\delta_h \sim N(\mu_b, \sigma_b^2)$.
- A slightly more arbitrary approach is to use a power prior that discounts the historical evidence such that $[lik(y_h|\theta_h)]^\alpha$, $0 < \alpha < 1$, where $\alpha = 0$ means that we ignore the historical evidence and $\alpha = 1$ means that we include the historical evidence without any discounting.
- Finally, we could assume that the parameter of interest is functionally dependent on the historical evidence, perhaps because of differences in variables that effect response, such that $\mu_i = f(\theta_1, \dots, \theta_H)$.

Korn et al. (2008) (Korn et al., 2008) considered the problem of creating an external control to use as a benchmark (or performance criteria) in future single-arm studies in patients with metastatic Stage IV melanoma. Their aim was to generate an estimate of the true response for

an untreated group corresponding to a sample of patients in the single-arm study. They developed prediction models (see Section 4.4) using data from 2100 patients in 42 randomised and single-arm Phase 2 studies involving 70 study arms of various treatments (assumed to be inactive) conducted between 1975 and 2005. Variables affecting overall survival (OS) and progression-free survival (PFS) were assessed using Cox proportional hazards models, and for OS and PFS events rates using logistic regression. Patient-level variables that were evaluated included sex, age, Eastern Cooperative Oncology Group performance status (ECOG PS), presence of visceral metastases and serum lactate dehydrogenase level; study-level variables that were evaluated included exclusion of patients with brain metastases, exclusion of patients with liver metastases, exclusion of patients with visceral metastases, previous treatment for metastatic disease and the year during which accrual was completed. Cox multiple regression analysis using complete cases suggested that gender, ECOG PS, visceral disease and brain metastases were predictive of OS and that pairwise interactions were not statistically significant. They generated an external control survivor function for an untreated group of patients as a weighted sum of n patients with baseline characteristics depending on the patients in the treated group such that $\bar{S}(t) = \frac{1}{n}\sum_{i=1}^{n} S_i(t)$, where $S_i(t) = [S_0(t)]^{HR}$ and $HR = \sum_{i=1}^{p} \beta_i x_i$, where the x_i are the covariates and the β_i are the coefficients corresponding to gender, ECOG PS, visceral disease and brain metastases. $S_0(t)$ represented the survivor function for female patients with ECOG PS 0, no visceral disease and without brain metastasis. However, the authors did not provide an assessment of the performance of the models. More importantly, the authors did not discuss uncertainty in the baseline survivor function or the impact of uncertainty and correlation associated with the estimated hazard ratios in the prediction model, and they assumed that there was no unexplained heterogeneity between studies. As proposed, the method does not produce a joint posterior distribution for parameters in a parametric survivor function corresponding to an untreated group. In addition, to implement the method it is necessary to have access to the prediction model responses for all patients in each study that did not include a concurrent untreated group. Although companies will have access to IPD from their own studies, it is unlikely that they will have access to IPD from other company's studies. It is more likely that relevant variables will be presented (if at all) as summary statistics in published articles. This is important in the case of non-linear models because the expectation of a function is not the same as the function evaluated as its expectation i.e. $E_X[f(X)] \neq f(E[\bar{X}])$. Consequently, the use of summary statistics about a comparator treatment to generate a posterior distribution from prediction models may generate biased estimates of parameters.

In the absence of any empirical evidence to inform parameters prior distributions could be generated using elicitation of experts' beliefs (O'Hagan et al., 2006).

Finally, the value of this approach in the context of a time-to-event outcome measure is in the ability to use more flexible modelling rather than assuming that the hazards for each treatment are proportional but it relies on the assumption that data can be reconstructed from published Kaplan-Meier survivor functions. The use of external controls was not considered as part of the submission to NICE of talimogene laherparepvec.

4.2 Treatment effect parameter

Abrams et al (2016) (Abrams et al., 2016) explored the impact of using various models to link disconnected networks using registry data. One approach was to use the registry data as if it came from a 'fictional study' that provided an estimate of the relative effect of treatments that were in separate networks of evidence formed by RCTs. The observational data was included without any discounting and also with discounting using a power prior. The modelling assumptions are the same as those for a network meta-analysis of RCTs, including that there are no internal biases.

In the absence of any empirical data, and as an alternative to generating a prior distribution for a study-specific response to treatment, we could generate a prior distribution for the population relative effect of two treatments across disconnected networks using elicitation of experts' beliefs (O'Hagan et al., 2006). For example, in Figure 1, we might generate a prior distribution for the population relative effect of GM-CSF compared to vemurafenib. The uncertainty represented by the prior distribution would affect comparisons between the networks but not the comparisons within networks (Dias et al., 2011; last updated April 2012). Goring et al, (2016) (Goring et al., 2016) discuss the use of prior information in a similar context but suggest that it should be suitably wide to reflect the overall uncertainty of this approach rather than reflecting genuine prior beliefs as would be the approach in a proper Bayesian analysis. We are not aware of any examples where elicitation of experts' beliefs has been used to generate prior distributions about relative treatment effects to facilitate indirect comparisons between treatments in disconnected networks. The elicitation of a relative treatment effect to connect networks was not considered as part of the submission to NICE of talimogene laherparepvec.

4.3 Random baseline models

Assuming a standard generalised linear model framework (Dias et al., 2013), the linear predictor can be written as:

$$\theta_{ik} = \mu_i + \delta_{i,bk} I_{\{k \neq 1\}}$$

where θ_{ik} is the population response in arm k of study i, μ_i is the study-specific baseline response in study i and $\delta_{i,bk}$ is the study-specific treatment effect of the treatment in arm k relative to the control treatment in arm b (b=1) in that study.

Random baseline models have been proposed in the context of conventional meta-analyses and assume that the μ_i are exchangeable (or conditionally exchangeable given prognostic variables) i.e. $\mu_i \sim N(\mu_{Base}, \tau_{Base})$ (Dias et al, 2013). In an NMA where not all studies may have included the reference treatment, it is necessary to ensure that the μ_i refer to the reference treatment. Random baseline models rely on the assumption that the baseline model is correct (Goring et al., 2016), and the main criticism against them is that they break the randomisation and assume that patients are randomised across studies as well as within studies; there is relatively little work comparing their properties to unconstrained baseline models.

Thom et al. (2015) (Thom et al., 2015) used random baseline models to form a connected network of RCTs by assuming that the placebo effects in each study were exchangeable across studies. The primary study was a placebo controlled adjunctive study stratified by one of four baseline treatments, which were treated as separate studies, but the authors also included data from single-arm, before-and-after observational studies. The indirect comparison of interest was between two treatments in two different strata that had no treatments in common with any other RCT. In addition, the distribution of baseline variables affecting response was different between the strata. The authors considered four separate models: 1) NMA of aggregate data from RCTs and observational studies; 2) NMA of IPD and aggregate data from RCTs and observational studies; 2) between-study and within-study covariate adjustments on the placebo effects; 4) within-study covariate adjustments on treatment effects. The authors also performed two separate sensitivity analyses, firstly by down-weighting the observational evidence using a power prior and secondly by constructing a prior distribution for a control arm for the observational studies but still including covariate adjustments. The authors acknowledged that their models have some limitations and involve several untestable modelling assumptions, including the use of random baseline models, but recommended this approach when networks are not connected.

4.4 Adjusted Treatment Response

Causal estimates of relative treatment effect across disconnected studies can be derived by modelling the probability of treatment assignment, generating a regression model for the outcome conditional on a set of covariates or a mixture of both (i.e. doubly robust estimation) (Faria et al., 2015). The aim of such adjusted treatment response methods is to generate adjusted responses for at least one treatment arm to account for differences between studies in prognostic variable and treatment effect modifiers. The methods make the strong assumption that there are no unobserved prognostic variables or treatment effect modifiers. Indirect estimates of relative treatment effects are then derived across studies after adjustment as if the treatments being compared had been included in the same study.

Stuart et al. (2011) (Stuart et al., 2011) and Hartman et al (2015) (Hartman et al., 2015) discuss the assumptions associated with methods for estimating the relative effect of treatments in a target patient population; these are summarised by Phillippo et al (2016) (Phillippo et al., 2016) for standard network meta-analyses, network meta-regression, and anchored and unanchored matching adjusted indirect comparisons and simulated treatment comparisons.

Methods for making indirect comparisons between treatments after adjusting treatment responses with a focus on unanchored comparisons are described below.

4.4.1 External Evidence-Based Adjustment

Differences between studies in the distribution of variables that effect response can be adjusted for based on external evidence. In the case of metastatic melanoma, Korn et al. (2008) (Korn et al., 2008) (see Section 4.1) showed that gender, ECOG PS, visceral disease and brain metastases were predictive of OS in patients with metastatic Stage IV melanoma

such that females, patients with an ECOG PS score of zero, patients with no visceral disease and patients with no brain metastases have better prognosis. Kotapati et al., (2011) (Kotapati et al., 2011) used the Korn model to adjust OS and compare treatments evaluated in studies which formed a disconnected network of evidence in an assessment of ipilimumab in the management of pre-treated patients with unresectable Stage III/IV melanoma. They reconstructed the OS probabilities over time from published Kaplan-Meier survivor functions and fitted Weibull distributions to the adjusted comparator treatment data as if the comparator treatment had been included in a target study. Specific details of the approach used were not provided in the conference abstract and presentation. However, the Korn et al. (2008) (Korn et al., 2008) model was developed in patients with mainly Stage IVM1b and Stage IVM1c melanoma and it may not be clinically relevant to patients with Stage IIIB, Stage IIIC and Stage IVM1a melanoma.

Bristol-Myers Squibb (BMS) Pharmaceuticals Ltd developed a modified Korn model that was used for the assessment of ipilimumab in patients with previously untreated unresectable Stage III or IV melanoma (National Institute for Health and Care Excellence, 2014) based on a different dataset to Korn et al. (2008) (Korn et al., 2008) and with the addition of the variable lactate dehydrogenase (LDH). The modified Korn model produced by BMS was:

$$\log(HR_t) = -0.154\overline{X}_{Gender=Female} - 0.400\overline{X}_{ECOGPS=0} - 0.285\overline{X}_{Visceral=No} - 0.306\overline{X}_{Brain=No} - 0.782\overline{X}_{LDH=Normal}$$

(Note: The original parameterisation of the Korn model (see Section 4.1) differed to the modified Korn model developed by BMS.)

The adjustment factor, HR_{Adj}, for a comparator treatment is given by the hazard ratio for the new treatment, HR_N, divided by the hazard ratio for the comparator treatment, HR_C i.e. $HR_{Adj} = {HR_N}/{HR_C}$. Adjusted survivor functions for the comparator treatment can then be generated as:

$$S_{Adj}(t) = S_C(t)^{HR_{Adj}}.$$

The adjustment can be made to Kaplan-Meier and parametric survivor functions. When comparator treatments are studied in more than one study, Kaplan-Meier survivor functions could be combined across studies using the Mantel-Haenszel method after adjustment for differences in studies. The process involves generating the number of patients at risk, the number of events and the number of censored observations from an adjusted survivor function in pre-defined time intervals and then pooling the data in each time interval using the Mantel-Haenszel method. This was the approach used to adjust OS for a comparator treatment in a study other than OPTiM as if the comparator treatment had been included in the OPTiM study (Quinn et al., 2016).

Some limitations with an external evidence-based adjustment approach, as generally applied, are that it assumes that differences between studies in all measured and unmeasured prognostic variables and treatment effect modifiers is captured by the prediction_model and,

as applied by Kotapati et al., (2011) (Kotapati et al., 2011) and Quinn et al., (2016) (Quinn et al., 2016), assumes that the regression coefficients are independent and estimated without uncertainty.

An alternative approach to estimating the adjustment factor would be to use a Bayesian approach, thereby quantifying uncertainty about the joint distribution between parameters and without having to assume asymptotic multivariate normality of the parameters. The adjustment factor could then be applied to a parametric survivor function. However, this would require access to the IPD for all studies and would involve identifying a suitable parametric survivor function to represent the observed data.

4.4.2 Iterative Proportional Fitting (IPF)

Kalton et al. (2003) (Kalton and Flores-Cervantes, 2003) described six methods for weighting sample estimates of response to match population values in the context of surveys where respondents are classified according to two or more variables each with two or more levels. In the context of a clinical study this corresponds to patients being classified according to two or more variables that affect response (e.g. gender and race) with two or more levels (i.e. males and females; white, black, other). The methods that they described were cell weighting, iterative proportional fitting (IPF) (also referred to as raking), linear weighting, generalised regression estimation (GREG) weighting, logistic regression weighting, and mixture of cell weighting and another method (Kalton and Flores-Cervantes, 2003).

Apart from IPF, we are not aware of any applications of the other five weighting methods for making indirect comparisons between treatments. IPF operates on the marginal distributions of the variables that affect response. The procedure is iterative in the sense that it starts by adjusting the sample row totals to correspond to the population (or target) row totals, and then adjusts the sample column totals to correspond to the population column totals, and continues until convergence is reached. The method assumes that there are no unobserved prognostic variables or treatment effect modifiers when making unanchored comparisons across studies. IPF has been used to make an indirect comparison between ponatinib and bosutinib in third line chronic phase chronic myloid leukemia (McGarry et al., 2016).

4.4.3 Propensity Score Matching Methods

A propensity score is the probability of treatment assignment conditional on observed variables that affect response and is estimated using logistic regression. There are four ways in which a propensity score can be applied: matching, with the most common approach being pair-matching in which pairs of patients treated with new and comparator treatments are found that have similar propensity scores, although other methods are available such as full matching (Stuart, 2010); inverse probability of treatment weighting (IPTW); stratification; and covariate adjustment.

Some limitations associated with propensity score matching methods are that estimates of treatment effect will be biased when there are unobserved prognostic variables and treatment effect modifiers (resulting in propensity score model misspecification) and when there is poor

overlap in the distribution of observed prognostic variables and treatment effect modifiers (resulting in extreme weights) (Austin and Stuart, 2015).

Conventional implementation of propensity score matching methods requires access to IPD on the new and comparator treatments, which (as in the case of talimogene laherparepvec) is generally not available. Section 4.4.5 discusses an approach to propensity score weighting when comparisons between treatments are required across studies in which there is IPD for one study and aggregate data for another study.

4.4.4 Entropy Balancing

Entropy balancing is an approach similar to propensity score methods for reweighting samples. As with propensity score methods, weights are estimated using a logistic regression but an assessment is then made whether the distributions of the covariates are similar subject to a set of predefined constraints on the moments of the covariate distributions (Hainmuller, 2012). Entropy balancing has been applied in the context of matching overall survival data in patients with non-small cell lung cancer to a population of patients defined by observational data (Happich et al., 2016).

Some limitations with entropy balancing as conventionally applied include that it requires access to IPD on the new and comparator treatments; it assumes that there are no unobserved prognostic variables or treatment effect modifiers; it is not possible to generate weights when the balancing constraints are inconsistent; the set of weights may include no positive weights in situations when there is limited data and extreme constraints; when there is limited overlap of the distributions of the covariates, the solution may involve extreme adjustments of the weights associated with some patients, which means that the final analysis may depend on a small set of highly weighted observations.

4.4.5 Matching-Adjusted Indirect Comparisons (MAIC)

Signorovitch et al., (2010) (Signorovitch et al., 2010), Signorovitch et al., (2012) (Signorovitch et al., 2012b) and Phillippo et al, (2016) (Phillippo et al., 2016) considered the problem of making indirect comparisons between treatments when there are differences between studies in variables that affect outcome. The method makes use of IPD from a study, P, of one of the treatments and weights the data using an approach similar to propensity score weighting so that their average covariate values matches those in a study, P', of the other treatments.

The estimator for treatment t in population represented by study P' is a weighted sum of the outcomes for patients in a population represented by study P:

$$\hat{\theta}_{tP'} = \frac{\sum_{i=1}^{N_{tP}} y_{itP} w_{it}}{\sum_{i=1}^{N_{tP}} w_{it}}$$

with weights, $w_{it} = \exp(\beta^T X_{it})$,corresponding to the odds of being included in the population represented by study P' versus study P, and X_{it} a vector of variables that affect

the outcome for patient i receiving treatment t. However, the weights cannot be estimated using conventional methods because it involves aggregate data from patients in study P'; Signorovitch et al. (2010) (Signorovitch et al., 2010) addressed this by proposing estimation based on the method of moments.

Signorovitch et al. (2010) (Signorovitch et al., 2010) claimed that the method can incorporate any number of continuous and categorical variables that affect response. Signorovitch et al. (2010) (Signorovitch et al., 2010) and Ishak et al. (2015) (Ishak et al., 2015) suggested that the method can be used to compare treatments across studies in which there is no common comparator, including single-arm studies. However, a limitation with this approach as usually applied is that it assumes that there are no unobserved prognostic variables or treatment effect modifiers, although Signorovitch et al. (2010) (Signorovitch et al., 2010) claimed that it is robust to model misspecification. Di Lorenzo et al. (2011) (Di Lorenzo et al., 2011) used this approach when making an adjusted indirect comparison between everolimus evaluated in a placebo controlled study and sorafenib evaluated in a single arm study.

Ishak et al. (2015) (Ishak et al., 2015) pointed out that to make the adjustment there needs to be overlap in the distributions of the covariates in each study. In the case of categorical outcomes, it would not be possible to adjust for a factor if a particular category is not represented in one of the studies e.g. gender might be an important prognostic factor but all patients were females in one study. In the case of a continuous outcome measure, it may not be possible to weight the values for which there is IPD so that they match the average baseline value in the comparator decision set. Extreme weights arise when there is poor overlap in the joint distribution of covariates between studies (Radice et al., 2012).

Belger at al (2015a) (Belger et al., 2015a) and Belger et al (2015b) (Belger et al., 2015b) considered the application of MAIC when there are multiple studies and treatments, and proposed a modification based on entropy balancing.

In the case of talimogene laherparepvec, the application of MAIC was inappropriate because it would produce adjusted responses as if talimogene laherparepvec had been evaluated in patient populations defined by the comparator treatments, which would be outside of its licensed indication.

4.4.6 Simulated Treatment Comparisons (STC)

STCs were introduced by Caro et al., (2010) (Caro and Ishak, 2010) and were described in further detail by Ishak (2015) (Ishak et al., 2015) and recently by Phillippo et al (2016) (Phillippo et al., 2016). STCs are similar to MAICs in that they generate adjusted responses for a treatment in a study for which there is IPD in order to match the sample characteristics of patients who received a comparator of interest in a separate study but differ in the way that the adjustments are made.

STCs use IPD from an index or reference study to generate a prediction model for the outcome measure of interest as a function of prognostic variables and treatment effect modifiers. The estimated coefficients are then applied to the average baseline characteristics

in the comparator study to generate predictions for treatments in the index study that reflect the sample of patients represented by the patients in the comparator study. The model for the data on the linear predictor scale is:

$$\theta_{tP}(\boldsymbol{X}) = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X} + (\beta_t + \boldsymbol{\beta}_2^T \boldsymbol{X}^{EM}) I_{\{t \neq b\}}$$

where β_0 is the response for the baseline treatment (b), β_t is the relative effect of treatment at $\boldsymbol{X} = \boldsymbol{0}$, $\boldsymbol{\beta}_1$ is a vector of coefficients corresponding to prognostic variables, $\boldsymbol{\beta}_2$ is a vector of coefficients corresponding to treatment effect modifiers $\boldsymbol{X}^{EM}$.

An estimate of the mean response for the new treatment as if it had been evaluated in study P' subject to the mean covariate values in study P'. Ishak et al., (2015) (Ishak et al., 2015) suggest that a benefit of using prediction models in the case of a continuous variable affecting response is in its ability to make predictions outside the range of values observed in the index study. However, the method produces biased estimates in the case of non-linear models because the expectation of a function is not the same as the function evaluated as its expectation i.e. $E_X[f(X)] \neq f(E[\bar{X}])$.

As with MAICs, Signorovitch et al. (2010) (Signorovitch et al., 2010) and Ishak et al. (2015) (Ishak et al., 2015) suggest that STCs can be used to compare treatments across studies in which there is no common comparator, including single-arm studies, the assumption being that there are no unobserved prognostic variables or treatment effect modifiers.

In the case of talimogene laherparepvec, the application of STCs was inappropriate because it would produce adjusted responses as if talimogene laherparepvec had been evaluated in patient populations defined by the comparator treatments, which would be outside of its licensed indication.

4.5 Model-Based Meta-Analysis (MBMA)

A model-based meta-analysis is an extension of conventional meta-analyses and is a relatively mature field in terms of applications; the first published examples appeared in the 1990s (Mould, 2012b). There is no single approach to, or guidance on, implementing a model-based meta-analysis, which may depend on the context as well as the analyst (Mould, 2012a) (Zhao et al., 2012) (Assawasuwannakit et al., 2015). The appropriateness and properties of a model-based meta-analysis is further complicated in models that combine a mixture of IPD data and aggregate data (Kimko et al., 2012) (Van Wart et al., 2013) (Ravva et al., 2014) (Van Wart et al., 2014).

Applications of MBMAs often involve fitting compartmental pharmacokinetic models but are increasingly being applied to other outcome measures collected longitudinally in studies with multiple doses and/or multiple treatments (Ahn and French, 2010) (Mandema et al., 2011) (Mandema et al., 2011) (Demin et al., 2012) (Gibbs et al., 2012) (Kimko et al., 2012) (Denney and Nucci, 2013) (Salinger et al., 2013) (Zhou and Al-Huniti, 2013) (Mandema et al., 2014) (Mercier et al., 2014) (Li et al., 2015b) (Li et al., 2015a). Some analysts adhere to the principle of concurrent controls and include fixed study effects while others assume

random study effects, which are generally not recommended in conventional meta-analyses. Mawdsley et al. (2016) (Mawdsley et al., 2016) showed how to implement a MBMA in the context of an NMA while respecting the randomisation. The rationale for the choice depends on previously published applications and the availability of data as well as the objectives of the analysis. In particular, the aim of a model-based meta-analysis is often seen to be broader than simply estimating a relative treatment effect, which is the main parameter estimated in a conventional meta-analysis. The evidence as a whole is typically used to describe the longitudinal placebo (or natural history) response in a particular disease in addition to describing any dose-response relationships or comparisons between multiple treatments by placing random effects on baseline responses as well as relative treatment effects in each study. Estimating treatment specific responses is also an objective in an HTA but the recommendation in this case is to fit separate treatment effect and baseline models and to combine the results from the separate models. The validity of MBMA models has typically been justified in terms of goodness-of-fit or their predictive ability without any discussion regarding the issue of respecting the randomisation.

Although we envisaged that a MBMA might be applicable in the context of time-to-event data, we did not find any specific examples of the analysis of OS and PFS, although Reddy et al. (2013) (Reddy et al., 2013) described a joint PK/PD and time-to-dropout model and Feng et al. (2013) (Feng et al., 2013) addressed the problem of assessing the exposure-response relationship of ipilimumab on overall survival using Cox proportional hazards regression adjusted for various covariates without accounting for study effects or heterogeneity between studies. The reason we did not find examples in the analysis of time-to-event data is most likely because such data are not strictly longitudinal within patient and the model for the data is a survivor function rather than a repeated measures model.

4.6 Multivariate Meta-Analysis

Disconnected networks can arise in the case of individual outcome measures within a study even though the studies as a whole might form a connected network. In this situation, it might be possible to borrow strength across outcome measures using a multivariate NMA (Achana and Cooper, 2014). Abrams et al (2016) (Abrams et al., 2016) used this approach to connect disconnected networks based on whether treatment was first or second-line in patients with rheumatoid arthritis. Multivariate NMA is a developing area of research that typically synthesises sample estimates of treatment effect (e.g. sample log hazard ratio) using a multivariate normal likelihood function. We are not aware of any published methodology on multivariate meta-analyses in the context of time-to-event outcome measures that model the underlying data generation process exactly and compare treatments in more flexible models that do not assume hazards are proportional for each treatment.

4.7 Class Effect Models

In a connected (i.e. anchored) network meta-analysis model it is sometimes possible to include a further hierarchy by assuming that treatment effects within classes are exchangeable such that:

$$\theta_{ik} = \mu_i + \delta^{c}_{i,bk} I_{\{k \neq 1\}}$$

$$\delta_{bk,c} \sim N\left(d_{Ak,c} - d_{Ab,c}, \tau^2\right)$$

$$d_{Ak,c} \sim N(\mu_c, \sigma_c^2)$$

where μ_c represents the pooled effect for the cth class of interventions, σ_c^2 represents the between-treatment variance within the cth class (which may assumed to be common to each class of treatments), and A is the reference treatment.

The model relies on the assumption that treatment can be classified into sensible classes. The advantages of this model, particularly when there is a limited amount of evidence about specific treatment effects within a class, are that it borrows strength about, and increases the precision of, individual estimates of treatment effect. However, this form of class effects model cannot be applied when the evidence base comprises disconnected networks.

Dequen et al. (2012) (Dequen et al., 2012) created a connected network at a class level where the treatments comprised a disconnected network by assuming that treatments were clinically equivalent within class. This approach meant that pairwise studies comparing treatments in the same drug class were excluded from the analysis and assumes that there is no treatment within drug class variability, including, for example, differences in effect according to dose. In addition, it raises the question whether decision-makers would be willing to approve treatments that might have relatively little evidence about their specific effect using evidence from other treatments, or whether companies would be prepared to accept this approach to sharing evidence.

Table 1 presents the treatment classes for the treatments in the assessment of talimogene laherparepvec. The OPTiM study is still isolated even after considering treatments as a class so that this approach was not feasible as part of the submission to NICE of talimogene laherparepvec.

5. Discussion

A network meta-analysis provides a basis for simultaneously comparing all treatments of interest even if they have not been compared directly in head-to-head studies but assumes that the studies form a connected network of studies and that the distribution of treatment effect modifiers is balanced across studies comparing different pairs of treatments. It might be possible to avoid disconnected networks by careful consideration of relevant treatments to include as comparators in RCTs at the design stage or by expanding the comparator decision set to include additional treatments to link disconnected networks at the analysis stage. However, including multiple comparator treatments in an RCT or repeating an RCT with different comparators may not be possible if the patient population is relatively small or if doing so is prohibitively expensive. It is inevitable that there will be situations when evidence about all relative treatment effects of interest will comprise studies forming disconnected networks of evidence for reasons including those associated with the assessment of talimogene laherparepvec in the OPTiM study; when there is no single standard of care

nationally or internationally; in single-arm studies of treatments in rare diseases (e.g. Waldenström's Macroglobulinaemia (National Institute for Health and Care Excellence); and in studies evaluating different treatment durations without control groups (e.g. ledipasvir–sofosbuvir for the treatment of chronic hepatitis C (National Institute for Health and Care Excellence, 2016a). Faced with such evidence, reimbursement agencies such as NICE must decide whether to recommend the new treatment based on an indirect estimate of the effect of the new treatment relative to comparators of interest.

In this paper, we have presented the results of a systematic review of methods and applications described in the scientific literature that address the problem of making indirect comparisons between treatments across disconnected networks that was motivated by an STA of talimogene laherparepvec for the treatment of advanced melanoma. Our work compliments that by Goring et al. (2016) (Goring et al., 2016), that appeared after we completed our systematic review, and also of that by Phillippo et al (2016) (Phillippo et al., 2016), which focused primarily on the validity of MAIC and STC and appeared after we our submitted our work for peer review.

The fundamental problem with making comparisons between treatments that have been evaluated in studies forming disconnected networks is that there may be differences between studies in the distribution of patient characteristics that are prognostic of response or are treatment effect modifiers. In this situation, a naïve, unadjusted indirect comparison produces a biased estimate of relative treatment effect and it is necessary to use alternative methods of analysis that are, by definition, not based on within-study estimates of treatment effect. In spite of the strong criticism that making comparisons between treatments evaluated in different studies, even after adjustment for observed variables that affect response, is a type of naïve indirect comparison and "its results are not worthy of consideration" (Hoaglin, 2013), an indirect estimate of relative effect must be generated to estimate the health benefits that might be achievable with the new treatment but that would be foregone by committing resources to the current treatment i.e. the opportunity cost.

Methods based on generating study-specific external controls and estimating relative treatment effects across networks using non-RCT evidence such as registry data or experts' beliefs preserve the ability to make simultaneous comparisons between treatments (assuming that there is not an imbalance in treatment effect modifiers in studies comparing different pairs of treatments). We are more receptive than Goring et al. (2016) (Goring et al., 2016) appear to be regarding the use of expert beliefs, although we acknowledge that elicitation must follow a justifiable, documented and transparent process, and can be resource intensive. Indeed, it is precisely in the context where there is no sample data with which to estimate parameters that prior distributions elicited from experts can be useful. The concern regarding random baseline models is well known but they provide a basis for incorporating sample data other than from RCTs and have been recommended in sparse disconnected networks (Thom et al., 2015). Matching adjusted indirect comparisons and simulated treatment comparisons may be useful in some contexts but it is important to appreciate that inferences depend on the population characterised by the sample of patients in the comparator study and that the population could potentially differ with each comparator of interest; these approaches were

not appropriate in the case of talimogene laherparepvec because inferences would be relative to the comparator treatment patient population rather than the talimogene laherparepvec patient population which would be outside its licensed indication. We are not aware of any research using MAICs or STCs that allow simultaneous inferences to be made across all treatment in the decision set in a specific population of interest.

Specification of the patient population for the decision problem is an important part of the decision-making process. Inferences following the application of adjusted treatment response methods will generally differ from those following a random effects NMA. In a random effects NMA it is assumed that the study-specific population treatment effects are exchangeable (i.e. related but different) and it is generally recommended that inferences are be based on predictive distributions of effects in new studies rather than on the mean of the random effect distribution (Higgins et al, 2009). Inferences based on adjusted treatment response effects will generally depend on the sample of patients in one of the studies and this may not be generalizable to the target population. In the case of talimogene laherparepvec, the aim was to generate an adjusted OS survivor function that would be expected for a comparator treatment in a study other than the OPTiM study as if the comparator treatment had been included in the OPTiM study.

In general, methods based on adjusted treatment responses have typically been proposed from a frequentist perspective which only account for sampling variation and do not allow for parameter uncertainty. Another source of uncertainty is structural uncertainty arising from model misspecification which produces biased estimates of relative effect, although it can be reduced using doubly robust estimation. Alternatively, it might be possible to incorporate external information to mitigate this or take a Bayesian perspective (Saarela et al, 2016). Generating joint posterior distributions about parameters should be seen as an important aim in health technology assessment in order to properly represent uncertainty about inputs to decision analytic models.

Finally, in the case of the talimogene laherparepvec submission to NICE, an indirect comparison with ipilimumab was made using an external evidence-based adjustment according to the modified Korn model and by presenting a naïve unadjusted indirect comparison. Sensitivity analyses were also performed by weighting each of the ipilimumab studies by line of therapy proportional to that observed in the OPTiM study. Although the NICE Evidence Review Group (ERG) acknowledged the effort made to generate an indirect estimate of relative treatment effect in the target patient population, the ERG considered the application of the modified Korn model inappropriate because it was developed using data from people with predominantly Stage IVM1b and Stage IVM1c disease, which have different disease trajectories to Stage III-IV1a disease. Nevertheless, the NICE Appraisal Committee concluded that talimogene laherparepvec is clinically and cost-effective in people for whom treatment with systemically administered immunotherapies is not suitable.

6. Conclusions

In conclusion, this review has identified various methods that have been proposed for dealing with the problem of estimating the relative effect of treatments across disconnected networks. We have described the main assumptions and limitations associated with each method. Assessing which method or methods are appropriate often depend on the clinical context as well as the availability of data. While data sharing initiatives should help to mitigate some of the limitations associated with studies that provide only aggregate responses, there is a need for further research on their use. In particular, the properties of frequentist methods and the robustness of results should be evaluated in simulation studies across a range of study sample sizes; using different models for prognostic variables and treatment effect modifiers that are unobserved at the analysis stage; and across different outcome measures such as time-to-event with non-proportional hazards. Furthermore, examples should be generated using a Bayesian approach that allows the incorporation of external information to reflect parameter uncertainty in addition to sampling variation.

Acknowledgements

We thank the two anonymous reviewers for their constructive comments.

Conflicts of interest

Christine Fletcher and Gerald Downey are employees and shareholders of Amgen.

This publication was funded by Amgen.

ScHARR provides consultancy advice to Amgen outside of the submitted work.

References

ABRAMS, K., BUJKIEWICZ, S., DEQUEN, P., JENKINS, D. & MARTINA, R. 2016. GetReal - Project No. 115546 WP1: Deliverable 1.5 (Case Study Review: Rheumatoid Arthritis) [Online]. Available: https://www.imi-getreal.eu/Portals/1/Documents/01%20WP1%20deliverables/Deliverable%20Report%20D1.5_Rheumatoid%20Arthritis_websiteversion.pdf [Accessed].

ACHANA, F. A. & COOPER, N. C. 2014. Network meta-analysis of multiple outcome measures accounting for borrowing of information across outcomes. BMC Medical Research Methodology.

ADES, A. E., CALDWELL, D. M., REKEN, S., WELTON, N. J., SUTTON, A. J. & DIAS, S. 2013. Evidence synthesis for decision making 7: A reviewer's checklist. Medical Decision Making, 33**,** 679-691.

AHN, J. E. & FRENCH, J. L. 2010. Longitudinal aggregate data model-based meta-analysis with NONMEM: approaches to handling within treatment arm correlation. Journal of Pharmacokinetics and Pharmacodynamics, 37**,** 179-201.

ANDTBACKA, R. H. I., KAUFMAN, H. L., COLLICHIO, F., AMATRUDA, T., SENZER, N., CHESNEY, J., DELMAN, K. A., SPITLER, L. E., PUZANOV, I., AGARWALA, S. S., MILHEM, M., CRANMER, L., CURTI, B., LEWIS, K., ROSS, M., GUTHRIE, T., LINETTE, G. P., DANIELS, G. A., HARRINGTON, K., MIDDLETON, M. R., MILLER JR, W. H., ZAGER, J. S., YE, Y., YAO, B., LI, A., DOLEMAN, S., VANDERWALDE, A., GANSERT, J. & COFFIN, R. S. 2015. Talimogene laherparepvec improves durable response rate in patients with advanced melanoma. Journal of Clinical Oncology, 33**,** 2780-2792.

ASSAWASUWANNAKIT, P., BRAUND, R. & DUFFULL, S. B. 2015. A model-based meta-analysis of the influence of factors that impact adherence to medications. Journal of Clinical Pharmacy and Therapeutics, 40**,** 24-31.

AUSTIN, P. C. & STUART, E. A. 2015. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Statistical Methods in Medical research**,** doi: 10.1177/0962280215584401.

BELGER, M., BRNABIC, A., KADZIOLA, Z., PETTO, H. & FARIES, D. Alternative weighting approaches for matching adjusted indirect comparisons (MAIC). (Abstract). . ISPOR 20th Annual International Meeting, 2015a Philadelphia, PA, USA.

BELGER, M., BRNABIC, A., KADZIOLA, Z., PETTO, H. & FARIES, D. Inclusion of multiple studies in matching adjusted indirect comparisons (MAIC). (Abstract). . ISPOR 20th Annual International Meeting, 2015b Philadelphia, PA, USA.

CARO, J. J. & ISHAK, K. J. 2010. No Head-to-Head Trial? Simulate the Missing Arms. PhamacoEconomics, 28**,** 957-967.

DEMIN, I., HAMREN, B., LUTTRINGER, O., PILLAI, G. & JUNG, T. 2012. Longitudinal Model-Based Meta-Analysis in Rheumatoid Arthritis: An Application Toward Model-Based Drug Development. Clinical Pharmacology & Therapeutics, 92**,** 352-359.

DENNEY, W. S. & NUCCI, G. 2013. Fasting Glucose Model-Based Meta-analysis: A Tool for Designing and Interpreting Early Diabetes Studies. Journal of Pharmacokinetics and Pharmacodynamics, 40**,** S141-S141.

DEQUEN, P., LORIGAN, P., JANSEN, J. P., VAN BAARDEWIJK, M., OUWENS, M. J. N. M. & KOTAPATI, S. 2012. Systematic Review and Network Meta-Analysis of Overall Survival Comparing 3 mg/kg Ipilimumab With Alternative Therapies in the Management of Pretreated Patients With Unresectable Stage III or IV Melanoma. Oncologist, 17**,** 1376-1385.

DI LORENZO, G., CASCIANO, R., MALANGONE, E., BUONERBA, C., SHERMAN, S., WILLET, J., WANG, X., LIU, Z. & DE PLACIDO, S. 2011. An adjusted indirect comparison of everolimus and sorafenib therapy in sunitinib-refractory metastatic renal cell carcinoma patients using repeated matched samples. Expert Opinion on Pharmacotherapy, 12**,** 1491-1497.

DI LORENZO, G., CASCIANO, R., MALANGONE, E., BUONERBA, C., SHERMAN, S., WILLET, J., WANG, X., LIU, Z. & DE PLACIDO, S. 2012. Authors reply: an adjusted

indirect comparison of everolimus and sorafenib therapy in sunitinib-refractory metastatic renal cell carcinoma patients using repeated matched samples. Expert Opinion on Pharmacotherapy, 13**,** 1079-1080.

DIAS, S. & ADES, A. E. 2016. Absolute or relative effects/ Arm-based synthesis of trial data. Research Synthesis Methods, 7**,** 23-28.

DIAS, S., SUTTON, A. J., ADES, A. E. & WELTON, N. J. 2013. Evidence synthesis for decision making 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials Medical Decision Making, 33**,** 607-617.

DIAS, S., WELTON, N. J., SUTTON, A. J. & ADES, A. E. 2011; last updated April 2012. NICE DSU technical Support Document 1: Introduction to evidence synthesis for decision making [Online]. Available: available from http://www.nicedsu.org.uk [Accessed 31 October 2016].

FARIA, R., HERNANDEZ-ALAVA, M., MANCA, A. & WAILOO, A. 2015. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal; Methods for comparative individual data [Online]. Available: http://scharr.dept.shef.ac.uk/nicedsu/wp-content/uploads/sites/7/2016/03/TSD17-DSU-Observational-data-FINAL.pdf [Accessed 22 May 2017].

FENG, Y., ROY, A., MASSON, E., CHEN, T.-T., HUMPHREY, R. & WEBER, J. S. 2013. Exposure-Response Relationships of the Efficacy and Safety of Ipilimumab in Patients with Advanced Melanoma. Clinical Cancer Research, 19**,** 3977-3986.

GIBBS, J. P., FREDRICKSON, J., BARBEE, T., CORREA, I., SMITH, B., LIN, S.-L. & GIBBS, M. A. 2012. Quantitative Model of the Relationship Between Dipeptidyl Peptidase-4 (DPP-4) Inhibition and Response: Meta-Analysis of Alogliptin, Saxagliptin, Sitagliptin, and Vildagliptin Efficacy Results. Journal of Clinical Pharmacology, 52**,** 1494-1505.

GORING, S. M., GUSTAFSON, P., LIU, Y., SAAB, S., CLINE, S. K. & PLATT, R. W. 2016. Disconnected by design: analytic approach in treatment networks having no common comparator. Research Synthesis Methods**,** doi: 10.1002/jrsm.1204.

GROSS, J. L., ROGERS, J., POLHAMUS, D., GILLESPIE, W., FRIEDRICH, C., GONG, Y., MONZ, B. U., PATEL, S., STAAB, A. & RETLICH, S. 2013. A novel model-based meta-analysis to indirectly estimate the comparative efficacy of two medications: an example using DPP-4 inhibitors, sitagliptin and linagliptin, in treatment of type 2 diabetes mellitus. Bmj Open, 3.

HAINMULLER, J. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20**,** 25-56.

HAPPICH, M., BRNABIC, A., FARIES, D., ABRAMS, K. R., WINFREE, K., GIRVAN, A., JONSSON, P., JOHNSTON, J. & BELGER, M. 2016. Reweighting RCT evidence to better reflect real life: A case study of the innovation medicines initiative [Online]. Available: https://www.imi-getreal.eu/Portals/1/Documents/Presentations/Happich%20et%20al%20-%20Reweighting%20RCT%20evidence%20to%20better%20reflect%20real%20life.pdf [Accessed 24 November 2016].

HARTMAN, E., GRIEVE, R., RAMSAHAI, R. & SEKHON, J. S. 2015. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. J R Stat Soc A, 178**,** 757-778.

HOAGLIN, D. C. 2013. An indirect comparison of everolimus versus sorafenib in metastatic renal cell carcinoma - a flawed analysis and a problematic response. Expert Opinion on Pharmacotherapy, 14**,** 1705-1706.

INTERNATIONAL CONFERENCE ON HARMONISATION OF TECHNICAL REQUIREMENTS FOR REGISTRATION OF PHARMACEUTICALS FOR HUMAN USE 2000. E10 Choice of control group and related issues in clinical trials.

ISHAK, K. J., PROSKOROVSKY, I. & BENEDICT, A. 2015. Simulation and Matching-Based Approaches for Indirect Comparison of Treatments. Pharmacoeconomics, 33**,** 537-549.

KALTON, G. & FLORES-CERVANTES, I. 2003. Weighting Methods. Journal of Official Statistics, 19**,** 81-97.

KAUFMAN, H. L., KOHLHAPP, F. J. & ZLOZA, A. 2015. Oncolytic viruses: a new class of immunotherapy drugs Nature Reviews Drug Discovery, 14**,** 642-662.

KIMKO, H., GIBIANSKY, E., GIBIANSKY, L., STARR, H. L., BERWAERTS, J., MASSARELLA, J. & WIEGAND, F. 2012. Population pharmacodynamic modeling of various extended-release formulations of methylphenidate in children with attention deficit hyperactivity disorder via meta-analysis. Journal of Pharmacokinetics and Pharmacodynamics, 39**,** 161-176.

KORN, E. L., LIU, P.-Y., LEE, S. J., CHAPMAN, J.-A. W., NIEDZWIECKI, D., SUMAN, V. J., MOON, J., SONDAK, V. K., ATKINS, M. B., EISENHAUER, E. A., PARULEKAR, W., MARKOVIC, S. N., SAXMAN, S. & KIRKWOOD, J. M. 2008. Meta-analysis of Phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future Phase II trials. Journal of Clinical Oncology, 26**,** 527-534.

KOTAPATI, S., DEQUEN, P. & OUWENS, M. 2011. Overall survival (OS) in the management of pretreated patients with unresectable stage III/IV melanoma: A systematic literature review and meta-analysis. Journal of Clinical Oncology, 29.

LI, H.-Q., XU, J.-Y., JIN, L. & XIN, J.-L. 2015a. The efficacy of placebo-adjusted taspoglutide on body weight reduction in clinical trials. Pharmazie, 70**,** 110-116.

LI, H. Q., XU, J. Y., JIN, L. & XIN, J. L. 2015b. Utilization of model-based meta-analysis to delineate the net efficacy of taspoglutide from the response of placebo in clinical trials. Saudi Pharmaceutical Journal, 23**,** 241-249.

MANDEMA, J. W. 2011. A Dose-Response Meta-Analysis for Quantifying Relative Efficacy of Biologics in Rheumatoid Arthriti. Clinical Pharmacology & Therapeutics, 90**,** 828-835.

MANDEMA, J. W., GIBBS, M., BOYD, R. A., WADA, D. R. & PFISTER, M. 2011. Model-Based Meta-Analysis for Comparative Efficacy and Safety: Application in Drug Development and Beyond. Clinical Pharmacology & Therapeutics, 90**,** 766-769.

MANDEMA, J. W., ZHENG, J., LIBANATI, C. & PEREZ RUIXO, J. J. 2014. Time Course of Bone Mineral Density Changes With Denosumab Compared With Other Drugs in Postmenopausal Osteoporosis: A Dose-Response-Based Meta-Analysis. Journal of Clinical Endocrinology & Metabolism, 99**,** 3746-3755.

MAWDSLEY, D., BENNETTS, M., DIAS, S., BOUCHER, M. & WELTON, N. J. 2016. Model-based network meta-analysis: A framework for evidence synthesis of clinical trial data. CPT Pharmacometrics Syst Pharmacol, 5**,** 393-401.

MCGARRY, L. J., YANG, M., CHIROLI, S., LUSTGARTEN, S. & DORER, D. J. 2016. Indirect comparison of efficacy of ponatinib versus bosutinib in 3rd-line chronic phase chronic myeloid leukemia using iterative proportional fitting [Online]. Available: https://www.ispor.org/research_pdfs/52/pdffiles/PCN21.pdf [Accessed 24 November 2016].

MERCIER, F., CLARET, L., PRINS, K. & BRUNO, R. 2014. A Model-Based Meta-analysis to Compare Efficacy and Tolerability of Tramadol and Tapentadol for the Treatment of Chronic Non-Malignant Pain. Pain Therapy, 3**,** 31-44.

MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D. G. & THE PRISMA GROUP. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med, 6**,** e1000097. doi:10.1371/journal.pmed.1000097.

MOULD, D. R. 2012a. Model-Based Meta-Analysis: An Important Tool for Making Quantitative Decisions During Drug Development. Clinical Pharmacology & Therapeutics, 92**,** 283-286.

MOULD, D. R. 2012b. Models for Disease Progression: New Approaches and Uses. Clinical Pharmacology & Therapeutics, 92**,** 125-131.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2014. NICE technology appraisal guidance 319: Ipilimumab for previously untreated advanced (unresectable or metastatic) melanoma [Online]. Available: https://www.nice.org.uk/guidance/TA319/documents/melanoma-previously-untreated-unresectable-stage-iii-or-iv-ipilimumab-id74-evaluation-report [Accessed 31 October 2016].

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE 2015. Single technology appraisal: User guide for company evidence submission template http://www.nice.org.uk/article/pmg24.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2016a. Ledipasivir-sofosbuvir for treating chronic hepatitis C [Online]. Available: https://www.nice.org.uk/guidance/ta363 [Accessed].

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. 2016b. Waldenstrom's macroglobulinaemia - ibrutinib [ID884] [Online]. [Accessed 24 November 2016].

NIE, L., ZHANG, Z., RUBIN, D. & CHU, J. 2013. Likelihood rewighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference. Annals of Applied Statistics, 7**,** 1796-1813.

NIXON, R., BERGVALL, N., TOMIC, D., SFIKAS, N., CUTTER, G. & GIOVANNONI, G. 2014. No Evidence of Disease Activity: Indirect Comparisons of Oral Therapies for the Treatment of Relapsing-Remitting Multiple Sclerosis. Advances in Therapy, 31**,** 1134-1154.

O'HAGAN, A., BUCK, C. E., DANESHKHAH, A., EISER, J. R., GARTHWAITE, P. H., JENKINSON, D. J., OAKLEY, J. E. & RAKOW, T. 2006. Uncertain Judgements: Eliciting Expert's Probabilities, Wiley.

PHILLIPPO, D. M., ADES, A. E., DIAS, S., PALMER, S., ABRAMS, K. R. & WELTON, N. J. 2016. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. [Online]. Available: http://scharr.dept.shef.ac.uk/nicedsu/technical-support-documents/population-adjusted-indirect-comparisons-maic-and-stc/ [Accessed 24 May 2017].

QUINN, C., MA, Q., KUDLAC, A., PALMER, S., BARBER, B. & ZHAO, Z. 2016. Indirect treatment comparison of Talomogene Laherparepvec compared with Ipilimumab and Vemurafenib for the treatment of patients with metastatic melanoma Adv Ther, 33**,** 643-657.

RADICE, R., RAMSAHAI, R., GRIEVE, R., KRIEIF, N., SADIQUE, Z. & SEKHON, J. S. 2012. Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. The International Journal of Biostatistics, 8.

RAVVA, P., KARLSSON, M. O. & FRENCH, J. L. 2014. A linearization approach for the model-based analysis of combined aggregate and individual patient data. Statistics in Medicine, 33**,** 1460-1476.

REDDY, V. P., KOZIELSKA, M., SULEIMAN, A. A., JOHNSON, M., VERMEULEN, A., LIU, J., DE GREEF, R., GROOTHUIS, G. M. M., DANHOF, M. & PROOST, J. H. 2013. Pharmacokinetic-pharmacodynamic modeling of antipsychotic drugs in patients with schizophrenia Part I: The use of PANSS total score and clinical utility. Schizophrenia Research, 146**,** 144-152.

SALINGER, D. H., MANDEMA, J. W., NEWMARK, R. D. & GIBBS, M. A. 2013. Model-Based Meta-Analysis Informs Phase 3 Head-to-Head Trial Simulations of Brodalumab and Competitors in Psoriasis. Journal of Pharmacokinetics and Pharmacodynamics, 40**,** S45-S46.

SCHMIDLE, H., GSTEIGER, S., ROYCHOUDHURY, S., O'HAGAN, A., SPIEGELHALTER, D. & NEUENSCHWANDER, B. 2014. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics, 70**,** 1023-1032.

SENN, S. 2004. Controversies concerning randomisation and additivity in clinical trials. . Statistics in Medicine, 23**,** 3729-3753.

SIGNOROVITCH, J., ERDER, M. H., XIE, J., SIKIRICA, V., LU, M., HODGKINS, P. S. & WU, E. Q. 2012a. Comparative effectiveness research using matching-adjusted indirect comparison: an application to treatment with guanfacine extended release or atomoxetine in children with attention-deficit/hyperactivity disorder and comorbid oppositional defiant disorder. Pharmacoepidemiology and Drug Safety, 21**,** 130-137.

SIGNOROVITCH, J. E., BETTS, K. A., REICHMANN, W. M., THOMASON, D., GALEBACH, P., WU, E. Q., CHEN, L. & DEANGELO, D. J. 2015. One-year and long-term molecular response to nilotinib and dasatinib for newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect comparison. Current Medical Research and Opinion, 31**,** 315-322.

SIGNOROVITCH, J. E., SIKIRICA, V., ERDER, M. H., XIE, J., LU, M., HODGKINS, P. S., BETTS, K. A. & WU, E. Q. 2012b. Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research. Value in Health, 15**,** 940-947.

SIGNOROVITCH, J. E., WU, E. Q., BETTS, K. A., PARIKH, K., KANTOR, E., GUO, A., BOLLU, V. K., WILLIAMS, D., WEI, L. J. & DEANGELO, D. J. 2011a. Comparative efficacy of

nilotinib and dasatinib in newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect comparison of randomized trials. Current Medical Research and Opinion, 27**,** 1263-1271.

SIGNOROVITCH, J. E., WU, E. Q., SWALLOW, E., KANTOR, E., FAN, L. & GRUENBERGER, J.-B. 2011b. Comparative Efficacy of Vildagliptin and Sitagliptin in Japanese Patients with Type 2 Diabetes Mellitus A Matching-Adjusted Indirect Comparison of Randomized Trials. Clinical Drug Investigation, 31**,** 665-674.

SIGNOROVITCH, J. E., WU, E. Q., YU, A. P., GERRITS, G. M., KANTOR, E., BAO, Y., GUPTA, S. R. & MULANI, P. M. 2010. Comparative Effectiveness Without Head-to-Head Trials A Method for Matching-Adjusted Indirect Comparisons Applied to Psoriasis Treatment with Adalimumab or Etanercept. Pharmacoeconomics, 28**,** 935-945.

SIKIRICA, V., FINDLING, R. L., SIGNOROVITCH, J., ERDER, M. H., DAMMERMAN, R., HODGKINS, P., LU, M., XIE, J. & WU, E. Q. 2013. Comparative Efficacy of Guanfacine Extended Release Versus Atomoxetine for the Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents: Applying Matching-Adjusted Indirect Comparison Methodology. Cns Drugs, 27**,** 943-953.

SONG, F., ALTMAN, D. G., GLENNY, A. & DEEKS, J. J. 2003. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ, 326.

SPIEGELHALTER, D. J., ABRAMS, K. R. & MYLES, J. P. 2004. Bayesian Approaches to Clinical Trials and Health-Care Evaluation, Wiley.

STUART, E. A. 2010. Matching methods for causal inference: A review and a look forward. Statistical Science, 25**,** 1-21.

STUART, E. J., COLE, S. R., BRADSHAW, C. P. & LEAF, P. J. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc A, 174**,** 369-386.

THOM, H. H., CAPKUN, G., CERULLI, A., NIXON, R. M. & HOWARD, L. S. 2015. Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension. BMC Medical Research Methodology, 15**,** 34.

TIU, R. & KALAYCIO, M. 2012. Targeted therapy for patients with chronic myeloid leukemia: clinical trial experience and challenges in inter-trial comparisons. Leukemia & Lymphoma, 53**,** 1263-1272.

VAN WART, S. A., SHOAF, S. E., MALLIKAARJUN, S. & MAGER, D. E. 2013. Population-based meta-analysis of hydrochlorothiazide pharmacokinetics. Biopharmaceutics & Drug Disposition, 34**,** 527-539.

VAN WART, S. A., SHOAF, S. E., MALLIKAARJUN, S. & MAGER, D. E. 2014. Population-based meta-analysis of furosemide pharmacokinetics. Biopharmaceutics & Drug Disposition, 35**,** 119-133.

VIELE, K., BERRY, S., NEUENSCHWANDER, B., AMZAL, B., CHEN, F., ENAS, N., HOBBS, B., IBRAHIM, J. G., KINNERSLEY, N., LINDBORG, S. & MICALLEF, S. 2014. Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical Statistics, 13**,** 41-54.

ZHAO, L., SHANG, E. Y. & SAHAJWALLA, C. G. 2012. Application of pharmacokinetics-pharmacodynamics/clinical response modeling and simulation for biologics drug development. Journal of Pharmaceutical Sciences, 101**,** 4367-4382.

ZHOU, D. & AL-HUNITI, N. 2013. Model Based Meta-analysis of Children's Depression Rating Scale: Revised (CDRS-R) in Children and Adolescents with Major Depressive Disorder. Journal of Pharmacokinetics and Pharmacodynamics, 40**,** S18-S18.

Table 1 Treatment classes involved in the assessment of talimogene laherparepvec

Intervention	**Class**
Talimogene laherparepvec	Oncolytic viral immunotherapy
Dacarbazine (DTIC)	Chemotherapy
Dabrafenib	BRAF inhibitor immunotherapy
Ipilimumab	CTLA-4 inhibitor immunotherapy
Vemurafenib	BRAF inhibitor immunotherapy
GM-CSF	Monomeric glycoprotein

Appendix

MEDLINE search strategy

1	("no head to head" or "no head-to-head").mp. (163)
2	(network meta-analys* or network meta analys* or network metaanalys*).mp. (755)
3	**1 and 2 (12) = Search 1**
4	disconnected network*.mp. (17)
5	(meta analys* or meta-analys* or metaanalys*).mp. (104932)
6	**4 and 5 (1) = Search 2**
7	("absence of head to head" or "absence of head-to-head").mp. (61)
8	**2 and 7 (10) = Search 3**

[mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]