



Causal inference and adjustment for reference-arm risk in indirect treatment comparison meta-analysis

Elyse Swallow^{*1}, Oscar Patterson-Lomba¹, Rajeev Ayyagari¹, Corey Pelletier², Rina Mehta² & James Signorovitch¹

¹Analysis Group Inc., Boston, MA 02199, USA

²US HEOR, Bristol-Myers Squibb Company, Princeton, NJ 08648, USA

*Author for correspondence: Tel.: +1 617 425 8483; Fax: +1 617 425 8000; Elyse.Swallow@analysisgroup.com

Aim: To illustrate that bias associated with indirect treatment comparison and network meta-analyses can be reduced by adjusting for outcomes on common reference arms. **Materials & methods:** Approaches to adjusting for reference-arm effects are presented within a causal inference framework. Bayesian and Frequentist approaches are applied to three real data examples. **Results:** Reference-arm adjustment can significantly impact estimated treatment differences, improve model fit and align indirectly estimated treatment effects with those observed in randomized trials. Reference-arm adjustment can possibly reverse the direction of estimated treatment effects. **Conclusion:** Accumulating theoretical and empirical evidence underscores the importance of adjusting for reference-arm outcomes in indirect treatment comparison and network meta-analyses to make full use of data and reduce the risk of bias in estimated treatments effects.

Lay abstract: Indirect treatment comparisons (ITCs) and network meta-analyses (NMAs) can help decision makers compare therapies that lack head-to-head randomized trials. However, these estimates are vulnerable to biases due to cross-trial differences in patient characteristics and other factors. In this study, we outline methods to reduce biases associated with ITC/NMA and apply them to three real-world examples (antiretroviral therapy for human immunodeficiency virus, treatments for Type 2 diabetes and biological treatments for psoriasis). Our results show that reference-arm adjustment can have a significant impact on indirectly estimated treatment effects and can improve consistency between indirect evidence and gold-standard evidence from randomized trials. ITC and NMA without reference-arm adjustment present an avoidable risk of misleading or biased treatment effects. We argue that reference-arm adjustment should always be considered and reported when feasible in ITC and NMA.

First draft submitted: 10 March 2020; Accepted for publication: 7 May 2020; Published online: 3 June 2020

Keywords: Bayesian approach • causal inference • Frequentist approach • indirect treatment comparison • network meta-analysis • reference-arm adjustment • treatment effect

Meta-analysis has long been used to summarize direct comparative evidence from multiple head-to-head randomized trials. Network meta-analysis (NMA), an extension of meta-analysis to indirect comparative evidence [1–3], has been increasingly used over the past decade to synthesize data from complex networks of clinical trials to allow inferences about the relative effects of treatments that have not been compared directly. Thus, NMAs have become important tools for comparative research and health technology assessments [4–6]. In these approaches, the estimated effect of drug A versus drug B may depend on direct evidence from head-to-head trials of these drugs, if available, as well as on indirect evidence from trials involving drug A but not drug B and vice versa. When direct evidence is unavailable, such as when head-to-head trials do not exist for new treatments, indirect treatment comparisons (ITCs) may provide the only comparative evidence to inform clinical or economic treatment preferences.

Causal inference for direct comparison meta-analyses is relatively uncomplicated because each head-to-head randomized trial, if well conducted and reported, estimates a causal effect. ITCs pose greater challenges for

causal inference since they involve cross-trial comparisons of nonrandomized treatment groups. Without adequate adjustment for cross-trial differences, ITCs may be biased [5,7].

The traditional adjusted ITC, introduced by Bucher *et al.* in 1997 [1], attempts to adjust for cross-trial differences by measuring treatment effects relative to a common reference arm (e.g., placebo). For example, given one set of trials providing a pooled odds ratio for response to drug A versus placebo ($OR_{A:P}$) and another set of trials providing a pooled odds ratio for response to drug B versus placebo ($OR_{B:P}$), the odds ratio for response to drug A versus drug B is estimated as $OR_{A:P}/OR_{B:P}$. As noted by Bucher *et al.* [1], this approach relies on the assumption that relative treatment effects are exchangeable across trials.

Detailed reviews and guidelines for ITCs have been developed for researchers and decision makers [5–9]. Some reviews have appropriately questioned whether traditional adjusted ITCs can rightly be called ‘adjusted,’ given that they do not provide adjustment for cross-trial differences in the same sense as regression, matching or propensity score-based approaches provide adjustment in observational studies [7,9,10]. The term ‘anchored’ ITC, as opposed to ‘adjusted’ ITC, has been proposed as being more accurate. Anchored ITCs can still be limited by imbalances in baseline characteristics of populations across clinical trials, which may in turn lead to bias. Moreover, although direct adjustments (e.g., via meta-regression) for observed baseline differences are possible, not all differences between populations may be observed. However, if a common comparator exists for a subset of the trials in the network of evidence, the comparator-arm outcome can integrate the effects of many other trial-level variables, some of which may not be observed. Thus, proper adjustment for the comparator-arm outcome can mitigate bias. It might at first seem that adjusting for outcomes would be problematic, since in most cases, adjustment should be based only on characteristics that are defined up to the time of treatment assignment. In fact, adjustment for outcomes is a well-known design error in statistical analyses that can introduce overadjustment bias [11]. In the present setting, however, it is important to note that the outcomes used for adjustment are occurring on the common reference arms, rather than on the treatment arms that are the target of inference. Because of randomization, outcomes on the reference arms can be treated as baseline characteristics for the treatment arms that are measured with error, meaning that the outcomes on the reference arms are telling of the treatment population at baseline, how it may differ across trials and when to account for cross-trial differences.

This paper considers indirect trial data under the potential outcomes framework for causal inference [12–14]. We focus on the elementary ITC in which no head-to-head trial exists for two treatments of interest, but each has been studied against a common reference arm (e.g., placebo). In this setting, we show that the assumption of exchangeable relative treatment effects across trials can lead to avoidable bias and can be formally tested without additional data. In particular, despite attempted adjustment for placebo arm responses, the traditional ITC based on $OR_{A:P}/OR_{B:P}$ may remain biased by cross-trial differences in placebo arm responses. We show that adjustment for placebo arm response, as described by Dias *et al.* [15] in a Bayesian setting, can avoid such biases and we provide a Frequentist version of this method. This approach to adjustment for the placebo arm response stems from models used to investigate the impact of baseline risk in direct comparison meta-analyses [16].

Both Frequentist and Bayesian approaches to ITC, with adjustment for the placebo-arm response, require weaker assumption for causal inference than the traditional anchored ITC model of Bucher *et al.* [1], which occurs as a special case. As we show below, analyses of real data examples indicate that significant departures from the traditional anchored ITC can be detected and resolved by adjustment for reference-arm response. For example, adjustment for the reference-arm response is shown to resolve a published discrepancy between direct and indirect meta-analyses of highly active antiretroviral therapy (HAART) in the treatment of HIV [17].

Materials & methods

ITC meta-analysis under the potential outcomes framework

Data from a collection of randomized trials can be described by a two-level model in which one level describes the distribution of underlying parameters across trials and the other level describes the distribution of observed outcomes within trials [3,18,19]. To identify causal effects of interest, this section focuses on the distribution of underlying parameters across trials and temporarily ignores within-trial sampling errors.

Let each trial be represented by (X, Y_0, Y_1, Z) , where X governs the population's response to placebo, Z indicates which treatment ($Z = 0$ or $Z = 1$) was trialed against placebo (Z can also be thought of as an indicator for trial assignment) and Y_0 and Y_1 govern responses to treatments 0 and 1, respectively. For example, X , Y_0 and Y_1 may represent the true mean values of a continuous outcome or the log odds of a binary response. The causal effect

of interest is the expected effect of treatment 1 versus treatment 0, $E[Y_1 - Y_0]$, under the same distribution that generated the observed trials. This causal effect is defined at the trial level, rather than the individual patient level.

The trial-level variables Y_0 and Y_1 are potential (or counterfactual) outcomes [12–14]. When no trial has directly compared treatment 0 versus 1, it is not possible to observe both Y_0 and Y_1 in the same trial. Under these circumstances, a placebo-controlled trial of treatment 0 or 1 can be represented by (X, Y, Z) , with $Y = ZY_1 + (1 - Z)Y_0$. Given data on (X, Y, Z) , the goal of an ITC is to estimate the causal effect of Z on Y , $E[Y_1 - Y_0]$. Because treatment assignment Z is not randomized, the distribution of X may differ between trials with $Z = 0$ and $Z = 1$. Furthermore, because X measures the same quantity as Y in a parallel treatment group, X is likely to be associated with Y . In other words, X is a potential confounder of the causal relationship between Z and Y . Adjustment for differences in X among trials is, therefore, important for causal inference in ITCs.

Estimation of $E[Y_1 - Y_0]$ given observations of (X, Y, Z) is a standard problem in causal inference [20,21] that is often addressed via regression adjustment (e.g., see Winship and Morgan [22]). A typical regression model for the effect of Z on Y , with adjustment for X , is:

$$Y = \alpha + \beta X + \theta Z + \varepsilon \quad (\text{Eq. 1})$$

where ε represents a random error with mean zero. This model is considered to ‘adjust’ for X in the sense that typical estimates of θ (e.g., least squares) will be consistent estimates of the causal effect $E[Y_1 - Y_0]$ provided that: the model for the conditional mean of Y given X and Z is correctly specified; and (Y_0, Y_1) is uncorrelated with Z conditional on X , which is often denoted as $(Y_0, Y_1) \perp\!\!\!\perp Z | X$ and referred to as ‘conditional exchangeability given X ’ or ‘absence of unmeasured confounding’ [23]. Formally, the effect of Z on Y adjusting for X should be considered a controlled direct effect (controlling for X) of Z on Y . The controlled effect of Z on Y is clearly the causal effect of interest. Indeed, interest in the controlled effect is the motivation for attempting to incorporate information from X into ITCs.

Causal inference via regression adjustment can be contrasted with traditional ITCs. Given data on (X, Y, Z) , the traditional anchored ITC of Bucher *et al.* [1] can be represented by the model

$$Y - X = \alpha + \theta Z + \varepsilon \quad (\text{Eq. 2})$$

where Y is the log odds of response on the treatment arm, X is the log odds of response on the placebo arm and ε represents a random error with mean zero. Note that Equation 2 is equivalent to fixing $\beta = 1$ in Equation 1, or

$$Y = \alpha + 1 \cdot X + \theta Z + \varepsilon \quad (\text{Eq. 3})$$

Because the effect of X is fixed at $\beta = 1$, rather than estimated from the data, the model in Equations 2 and 3 cannot adjust for X in the usual sense described for the model in Equation 1. Sufficient conditions for estimates of θ based on the model in Equation 2 to be unbiased for $E[Y_1 - Y_0]$ are: correct specification of the conditional mean of Y given X and Z ; and unconditional exchangeability of treatment contrasts: $(Y_1 - X, Y_0 - X) \perp\!\!\!\perp Z$. These assumptions have been stated previously [1].

Note that unconditional exchangeability is a stronger assumption than conditional exchangeability given X and is, therefore, more likely to be violated. In particular, estimates of θ based on Equation 2 can have substantial bias for $E[Y_1 - Y_0]$ if the placebo-arm response X is associated with the trialed treatment Z and with the treatment effect $Y - X$. In other words, despite attempted adjustment for the placebo arm responses, via and assumed and fixed coefficient $\beta = 1$, traditional anchored ITCs may remain biased by cross-trial differences in placebo-arm responses. Compared with Equation 2, the model in Equation 1 requires weaker assumptions for causal inference, contains Equation 1 as a special case, makes use of the data to estimate β rather than assuming a fixed value and it is more consistent with generally accepted approaches to causal inference based on regression adjustment. The model in Equation 1 could, therefore, improve causal inferences in ITCs.

Two features of this data setting warrant a close inspection of the relevant graphical models. The use of directed acyclic graphs to describe causal effects has been well described (e.g., see Greenland *et al.* [24] and Pearl *et al.* [25]). Briefly, Figure 1 depicts the causal models discussed above. Figure 1A gives a full model, in which outcomes on the active and placebo arms are all affected by trialed active treatment (Z) and unmeasured confounders (U). The arrow from Z to X may be justifiably excluded in some applications. However, knowing the active therapy under investigation could influence outcome assessments (e.g., ascertainment of suspected adverse events, time period associated with different background standards of care), even in the placebo arm. We have added this edge to the graph for completeness. While U is a cause of both X and Y , the controlled direct effect of Z on Y is not

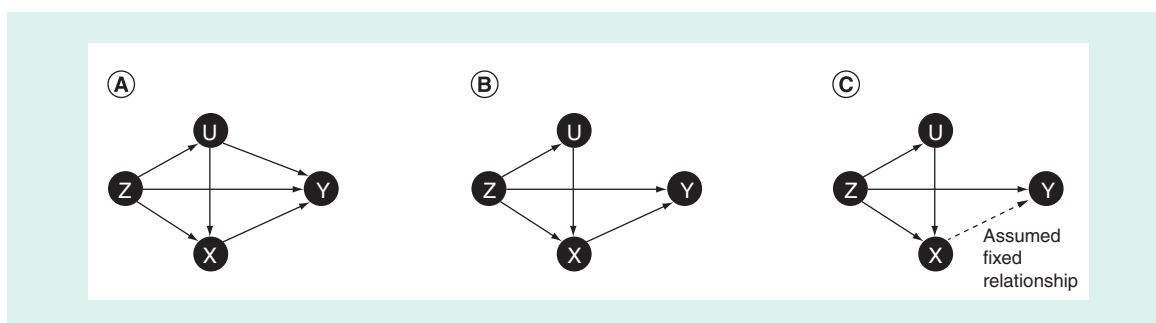


Figure 1. Directed acyclic graphs. (A) Directed acyclic graph for model (1). **(B)** Directed acyclic graph for model (2) assuming that ‘unobservables’ (U) have no direct effect on outcome (Y). **(C)** Directed acyclic graph for model (2), which assumes that a specific relationship between X and Y , for example, $(Y \cdot X)$, is conditionally independent of X and U given Z .

identifiable [26]. Arbitrary impacts of U on X and Y could explain any pattern of association regardless of the presence or absence of a true causal relationship. Note that, while U is arbitrary and represents ‘unobservables’ that we will never be able to condition on in practice, the causal model is unchanged in any practical sense by adding a double arrow between X and Y . It is well known that the controlled direct effect of Z on Y cannot be identified in this setting. To make progress, one can assume that U has no direct effect on Y , such that all effects of U on Y go through either X or Z . In this case it is possible to identify the controlled direct effect of Z on Y since all pathways between U and Y can be blocked by conditioning on Z and X (Figure 1B). The more stringent assumption of Bucher *et al.* [1], unconditional exchangeability, is represented by Figure 1C.

Frequentist approach: structural equation model

In this section, we extend the model of Thompson *et al.* [16] for relating treatment effect to underlying risk to the ITC of multiple treatments. We also introduce measurement error in the observations of X and Y . Building on the notation of Equation 1, indirect data in trial i can be described by the structural equation model

$$Y_i = \alpha_y + \theta_y Z_i + \beta X_i + \varepsilon_{yi} \quad (\text{Eq. 4})$$

$$X_i = \alpha_x + \theta_x Z_i + \varepsilon_{xi} \quad (\text{Eq. 5})$$

where ε_{xi} and ε_{yi} are Gaussian random variables with mean zero and with variances σ_x^2 and σ_y^2 , respectively. Independence of ε_{xi} and ε_{yi} is implied by conditional exchangeability of Z and (Y_0, Y_1) given X .

Let X_i^* and Y_i^* represent the observed data, which are assumed to be independent and to follow known distributions given X_i and Y_i . For example, in the case of a binary response, X_i and Y_i may represent the log odds of the response rates, such that $X_i^*|X_i$ is binomially distributed with mean $N_{xi}\exp(X_i)/(1 + \exp(X_i))$ and $Y_i^*|Y_i$ is binomially distributed with mean $N_{yi}\exp(Y_i)/(1 + \exp(Y_i))$, where N_{xi} and N_{yi} are known sample sizes in each arm of trial i . In the case of a continuous response, the variance of the response in each treatment arm of each trial is assumed to be known.

It is convenient to express the model in Equations 4 and 5 as a generalized linear mixed effects model [27,28] for the outcome in each arm of each trial. Define the treatment arm indicator $T_{ij} = j$ for $j = 0, 1$ (where 0 represents the reference treatment [e.g., placebo]) and let $W_{ij} = T_{ij}Y_i + (1 - T_{ij})X_i$ be the linear predictor on the j th arm in the i th trial. The two Equations 4 and 5 can then be combined as

$$W_{ij} = \alpha_x + \gamma T_{ij} + \theta_x Z_i + \eta T_{ij} Z_i + \varepsilon_{xi} + \xi_i T_{ij} \quad (\text{Eq. 6})$$

where $\gamma = \alpha_y + (\beta - 1)\alpha_x$, $\eta = \theta_y + (\beta - 1)\theta_x$ and $\xi_i = \varepsilon_{yi} + (\beta - 1)\varepsilon_{xi}$. The causal parameter of interest can be recovered as:

$$\theta_y = \eta - \theta_x \text{cov}(\varepsilon_{xi}, \xi_i) / \sigma_x^2$$

This model can be fit in SAS (SAS Institute Inc., Cary, NC, USA) using GLIMMIX or in R using lme4 with adaptive Gaussian quadrature to maximize the likelihood. A maximum likelihood estimate for θ_y can then be

obtained by plugging the maximum likelihood estimates for η , θ_x , $\text{cov}(\varepsilon_{xi}, \xi_i)$ and σ_x^2 into the above equation. Standard errors for the maximum likelihood estimate of θ_x are not readily available from model outputs from GLIMMIX or lme4, but statistical inference can be based on the profile likelihood. The traditional model of Bucher *et al.* [1] can be represented by setting $\beta = 1$ in Equation 1, which results in $\text{cov}(\varepsilon_{xi}, \xi_i) = \text{cov}(\varepsilon_{xi}, \varepsilon_{yi}) = 0$ and therefore $\eta = \theta_y$. The key modeling assumption of Bucher *et al.* (i.e., that $\beta = 1$), can therefore be tested via a one degree of freedom likelihood ratio test comparing models with versus without a free covariance parameter between the random effects ε_{xi} and ξ_i .

Bayesian approach

The Bayesian framework for ITCs in the context of a meta-analysis has been extensively described (e.g., Hoaglin *et al.* [7], Dias *et al.* [8]). In this section, we summarize this framework and extend its more basic formulation to adjust for reference-arm response, which has also been shown as part of the National Institute for Health and Care Excellence (NICE) guidance [15].

The Bayesian approach to meta-analysis combines the likelihood function of the data with a prior probability distribution about the parameters of interest to obtain a posterior probability distribution of such parameters. For the random effects network model of a binary outcome, the Bayesian formulation is as follows. The number of events r_{ij} is given by the binomial likelihood:

$$r_{ij} = \text{Binomial}(p_{ij}, n_{ij})$$

where p_{ij} and n_{ij} represent the probability of an event and the number of patients, respectively, in arm j of trial i . The parameter p_{ij} is modeled on the logit scale as:

$$\text{logit}(p_{ij}) = \mu_i + \delta_{i,1j} I_{\{k \neq 1\}} \quad (\text{Eq. 7})$$

where $I_{\{u\}} = 1$ if u is true and 0 otherwise, μ_i are trial-specific log odds of the outcome in the reference treatment (i.e., the treatment indexed 1) and $\delta_{i,1j}$ are the trial-specific log odds ratios of 'success' on the j th treatment group compared with reference. Moreover,

$$\delta_{i,1j} = N(d_{i,1j}, \sigma^2) = N(d_{i,j} - d_{i,1}, \sigma^2)$$

with σ^2 being between-trial heterogeneity ($\sigma^2 = 0$ for a fixed effect model, which assumes homogeneity of the underlying true treatment effects). In this framework, d and σ^2 are typically given independent noninformative priors:

$$d_{i,j} = N(0, \varphi^2), \quad \varphi^2 \gg 0$$

$$\sigma^2 = \text{Uniform}(0, 5)$$

Adjustment for reference-arm response is conducted in a similar way as a meta-regression that seeks to adjust for differences in a given baseline covariate (e.g., age). Specifically, the model in Equation 7 is modified to:

$$\text{logit}(p_{ij}) = \mu_i + \left(\delta_{i,1j} + (\beta_{1t_{ij}} - \beta_{1t_{i1}}) \left(\mu_i - \bar{\mu} \right) \right) I_{\{k \neq 1\}} \quad (\text{Eq. 8})$$

with t_{ij} representing the treatment in arm j of trial i and $\bar{\mu}$ is the mean of the log odds in the reference treatment arm ($j = 1$). Thus, the treatment effects are the estimated log odds ratios at the mean risk value.

We can then fit a model that assumes a common interaction effect for all treatments, such that $\beta_{1j} = b$ (for $j > 1$) and $\beta_{11} = 0$, which guarantees that the terms cancel out inactive versus active comparisons and no reference-arm risk adjustment is performed for trials that do not include the reference treatment. The b parameter is also given a noninformative prior: $b = N(0, \varphi^2)$.

To assess whether the estimated $\hat{b}$ coefficient is notably different from 0, we can calculate the 95% credible interval from the Markov chain Monte Carlo (MCMC) simulations. An interval not containing 0 would indicate

substantial evidence against the ‘unadjusted’ model. The deviance information criterion (DIC), a measure of model fit that penalizes model complexity, can also be compared between the adjusted and unadjusted models to assess the evidence in favor of adjusting for reference-arm response. Additionally, we can check if the treatment effect from the unadjusted model differs substantially from that obtained by the adjusted model; if so, this would indicate that lack of adjustment for the reference-arm response may result in significant confounding and bias, which can in turn have important implications for the comparative efficacy of the studied treatments.

An exploratory way to assess if the reference-arm response is a source of heterogeneity before conducting any adjusted meta-analysis is to plot the treatment effect (e.g., log odds ratio for a binary outcome) against the reference-arm rate (e.g., log odds of X) and fit a regression line (e.g., logistic regression of log odds of the outcome to the effects of trial, treatment, reference-arm log odds and the interaction between treatment and reference-arm log odds) [16]. Next, the plots and the regression results can be inspected to check if there are indications that the treatment effects change significantly with increasing reference-arm response.

Results

Applications

Highly active antiretroviral therapy

Chou *et al.* [17] described a discrepancy between direct and indirect meta-analyses of HAART with a protease inhibitor (PI-HAART) versus a non-nucleoside reverse transcriptase inhibitor (NNRTI-HAART). They identified 26 trials in total: 12 head-to-head trials directly comparing PI-HAART versus NNRTI-HAART, six trials comparing NNRTI-HAART versus two nucleoside reverse transcriptase inhibitors (NRTIs) and eight trials comparing PI-HAART versus two NRTIs. A direct comparison meta-analysis of the 12 head-to-head trials indicated a higher odds of virological suppression with NNRTI-HAART versus PI-HAART (odds ratio: 1.60; 95% CI: 1.31–1.96).

In contrast, an ITC based on the 14 trials, with two NRTIs as a common reference and using the method of Bucher *et al.* [1], suggested that NNRTI-HAART was associated with a lower odds of virological suppression than PI-HAART (odds ratio: 0.26; 95% CI: 0.07–0.91). The difference between these direct and indirect estimates was statistically significant and presented as a cautionary example. Chou *et al.* [17] hypothesized that rapid changes in management and outcomes of HIV infection could have led to cross-trial differences and biased the ITC.

Chou *et al.* [17] considered several sensitivity analyses, but none resolved the discrepancy between the direct comparison and the ITC. In our reanalysis of the Chou *et al.* data, we take as a starting point the sensitivity analysis, excluding from the indirect analyses three NNRTI trials that used delavirdine, which is not currently a recommended therapy because of poor potency. None of the 12 trials included in the direct comparison meta-analysis included delavirdine.

Differences in baseline characteristics among trials could result in discrepancies between the direct comparison and ITC if the aggregate baseline characteristics modify the effect of treatment on the odds ratio scale. Chou *et al.* [17] report baseline summary statistics for the proportion of females and CD4 count in each trial. However, adjustment for these trial-level baseline characteristics in meta-regression models could not explain the discrepancy between the direct comparison and the ITC (Figure 2 and Supplementary Table 1; see WinBugs code provided in the Custom code section in the Supplementary Material).

Cross-trial differences in reference-arm responses could also be hypothesized to bias the traditional ITC of Chou *et al.* [17]. The odds of virological suppression on the reference arms (treatment with two NRTIs) did not differ significantly between the trials of NNRTI-HAART versus those of PI-HAART in a logistic mixed effects model (odds ratio: 2.69; 95% CI: 0.76–9.50; $p = 0.12$). However, after adjusting for the fixed effects of the proportion of females and baseline CD4 count, trials versus NNRTI-HAART were associated with a significantly greater odds of response to two NRTIs than trials versus PI-HAART (odds ratio: 3.89; 95% CI: 1.17–12.90; $p = 0.03$). Adjustment for reference-arm response (NRTI), in addition to the proportion of females and baseline CD4 count, can eliminate the significant bias that Chou *et al.* [17] identified in their earlier analysis.

Therefore, we applied the Bayesian and Frequentist approaches described in the previous sections to the Chou *et al.* data to adjust for reference-arm risk in addition to the proportion of females and baseline CD4 counts. Both methods reversed the estimated treatment effect in the ITC, giving it the same direction and a similar magnitude to the pooled effect estimate based on head-to-head trials (Figure 2 and Supplementary Table 1). Note that, although adjustment for additional baseline factors increases the width of the CIs on the odds ratio scale, the widths on the log odds scale are nearly constant. Thus, the increased consistency between the indirect and direct analyses can be

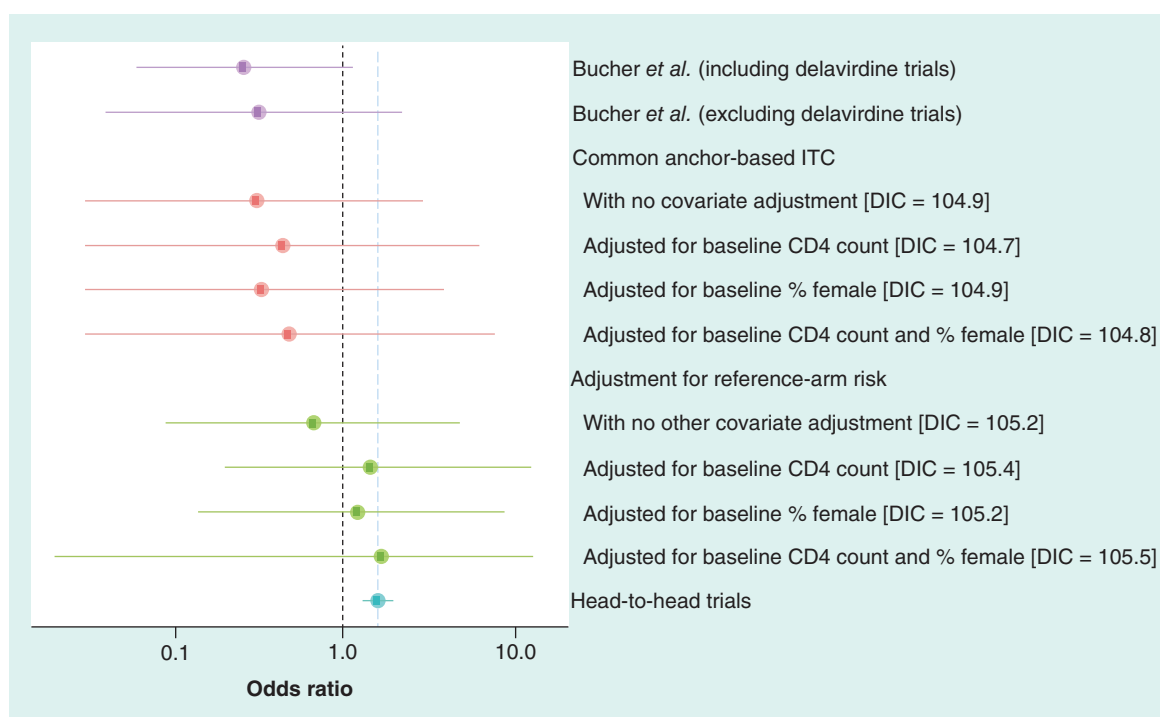


Figure 2. Bayesian approach: estimated odds ratios for virological suppression with non-nucleoside reverse transcriptase inhibitor-based versus protease inhibitor-based highly active antiretroviral therapy under different models for indirect comparison and the results from the head-to-head trials.

CD4: Cluster of differentiation 4; DIC: Deviance information criterion; ITC: Indirect treatment comparison.

attributed to moving the point estimate, rather than just increasing uncertainty. Note also that in the Frequentist analysis, a likelihood ratio test would reject the anchored ITC without adjustment for reference-arm risk or baseline characteristics (log likelihood: -35.1) in favor of the full model adjusting for these factors (log likelihood: -29.7) (likelihood ratio test: 10.8 on 3 degrees of freedom; $p = 0.012$).

In this ITC of NNRTI- and PI-based HAART, adjustment for multiple baseline differences, including reference-arm risk, was important to providing an indirect effect estimate consistent with direct randomized experiments.

Dipeptidyl peptidase 4 inhibitors in diabetes

Oral dipeptidyl peptidase 4 inhibitors are an established drug class in the treatment of Type 2 diabetes mellitus. The efficacy of two oral dipeptidyl peptidase 4 inhibitors, sitagliptin and vildagliptin, has been directly compared with placebo as therapy for the management of Type 2 diabetes mellitus in randomized trials. However, no head-to-head clinical trials have directly compared these two treatments in terms of their efficacy in reducing the percentage of hemoglobin A1c (%HbA1c) from baseline over 12 weeks.

A prior study assessed the relative efficacy of sitagliptin versus vildagliptin in %HbA1c reduction, examining the impact of adjustment for placebo arm effects [29]. A systematic literature review identified 11 randomized trials that included sitagliptin 100 mg once daily (seven trials) or vildagliptin 50 mg twice daily (four trials) as monotherapies for Type 2 diabetes. Random effects and fixed effects Bayesian NMAs were used to compare %HbA1c change from baseline to week 12. Models were fit with and without adjustment for comparator arm effect and the differences in %HbA1c reduction between sitagliptin and vildagliptin were assessed using the posterior mean difference. The results, provided in Table 1, indicate that adjustment for placebo-arm effects significantly impacted the ITC of week-12 outcomes. Without adjustment, the difference in efficacy was modest and insignificant, whereas after adjustment, the difference became larger and significant (95% credible intervals did not cross 0). Moreover, the regression coefficient associated with the placebo arm adjustment was significantly different from 0 and the DIC suggested better model fit after adjustment.

Table 1. Results of Bayesian network meta-analysis comparing the percentage of hemoglobin A1c reduction between sitagliptin and vildagliptin

Effects		Without placebo adjustment	With placebo adjustment
		Mean Δ %HbA1c (95% CrI)	Mean Δ %HbA1c (95% CrI)
Random effects	Sitagliptin 100 mg qd	-0.81 (-1.01, -0.61)	-0.70 (-0.80, -0.60)
	Vildagliptin 50 mg bid	-0.86 (-1.13, -0.58)	-1.02 (-1.18, -0.87)
	Beta	NA	-0.95 (-1.35, -0.54)
	Difference (vilda-sita)	-0.05 (-0.39, 0.30)	-0.32 (-0.52, -0.12)
	DIC	-27.2	-28.8
Fixed effects	Sitagliptin 100 mg qd	-0.82 (-0.90, -0.74)	-0.7 (-0.76, -0.64)
	Vildagliptin 50 mg bid	-0.9 (-1.03, -0.77)	-1.02 (-1.11, -0.92)
	Beta	NA	-0.98 (-1.29, -0.66)
	Difference (vilda-sita)	-0.08 (-0.23, 0.07)	-0.32 (-0.44, -0.19)
	DIC	-11.7	-26.4

Δ : Difference; bid: Twice daily; CrI: Credible interval; DIC: Deviance information criterion; HbA1c: Hemoglobin A1c; NA: Not applicable; qd: Once daily.

Biological treatments for psoriasis

The advent of biological therapies has improved treatment outcomes in moderate-to-severe psoriasis. Head-to-head randomized trials are currently unavailable for comparisons between most biological treatments given the number of available treatments. Moreover, substantial cross-trial variation in placebo arm response rates was identified in a recent systematic review [30] of trials of biological treatments for psoriasis. To assess the extent to which this variation in placebo arm response rates was a source of significant bias in cross-trial comparisons of biological treatment outcomes, Signorovitch *et al.* [31] conducted a Bayesian NMA that included an adjustment for reference-arm response rates, while revisiting previous NMAs based on the same data but that did not adjust for variation in the reference-arm response.

A total of 15 randomized trials of biological treatments for moderate-to-severe psoriasis were identified [32–46], 14 of which were placebo controlled. The primary efficacy outcome for this study was based on the Psoriasis Area and Severity Index (PASI) score, with reductions of 50, 75 and 90% from baseline PASI score were defined as PASI 50, 75 and 90, respectively. The relative efficacy of the biological treatments was evaluated using an ordinal model with a probit link for PASI 50, 75 and 90.

The analyses showed that the placebo-adjusted model fit the data significantly better than the previous unadjusted models and that the reference-arm response was an important confounder of the relative biological treatment effects. Specifically, the reference-arm adjustment coefficient was significantly different from zero, indicating that the adjustment for the reference-arm response reduced unexplained heterogeneity and improved model fit. Moreover, the placebo-adjusted model provided a considerable reduction in between-study heterogeneity compared with the unadjusted model. Finally, the DIC for the adjusted model was lower than the DIC for the unadjusted model, further indicating that the adjusted model was more parsimonious and fit the data better. Subsequent to this publication, reference-arm adjustment has become the standard approach applied in economic evaluations and network meta-analyses in psoriasis [47–49].

Discussion

Adjustment for baseline risk has been well-studied for pairwise meta-analyses, where it is rightly recognized as an important step for identifying effect modification and assessing generalizability [18,50]. Previous publications and guidance have also explained the value of reference-arm adjustment in ITCs and NMA [15]. However, despite this background, adjustment for baseline risk has not been widely used in ITCs. This paper has summarized theoretical principles and several real data applications that are supportive of broader use of reference-arm adjustment. We have also added a Frequentist option to the already available Bayesian approach. There is a need for further methodological research on reference-arm adjustment. However, based on the methods and examples available to

date, we argue that ITC and NMAs involving multiple trials with a common reference arm should either attempt to adjust for reference-arm response and report on the results of that attempt or provide a rationale for not adjusting.

The arguments in favor of further prioritizing conduct and reporting of reference-arm adjustment ITC and NMA are as follows:

- Adjustment makes more use of the available data & avoids unnecessary, testable assumptions

The reference-arm adjusted model includes the unadjusted model as a special case. Therefore, if the unadjusted model provided the best fit to the data, the adjusted and unadjusted models would return essentially equivalent results. It is also possible to statistically test the key assumption of the unadjusted model, as illustrated in the examples above – in other words, test whether there is a fixed relationship between reference and active arms, such as an odds ratio, around which variation or random effects are independent of the reference-arm outcomes. If sufficient trial data are available, this assumption should at least be tested rather than simply accepted. If sufficient data are not available to test this strong assumption, this is worth reporting as a limitation of the available data.

- Consistency with long-accepted approaches to causal inference outside of ITCs & NMAs

NICE Decision Support Unit (DSU) guidance indicates that adjustment for reference-arm responses in NMAs is in accordance with best practices in nonrandomized studies when a pretreatment characteristic varies substantially across treatment groups and these characteristics are likely to be associated with treatment effects [15].

Taking this a step further, we suggest that lack of adjustment for reference-arm response may often be poor practice. As described above, typical approaches to causal inference allow the observed data to inform adjustment for potential confounding factors, rather than assuming fixed relationships. For example, in multivariable regression analyses to estimate a treatment effect between nonrandomized groups, it would be unheard of to plug in an assumed coefficient value for a major suspected confounding variable, rather than using the data to estimate the value of that coefficient; however, as illustrated in [Equations 2 and 3](#) above, this is akin to how the anchor-based ITC/NMA model works without reference-arm adjustment.

- Supported by the same rationale for preferring anchor-based versus unanchored ITC

Preference for anchor-based ITC/NMA over unanchored naive comparisons is based on the belief that reference-arm outcomes reflect important variation across trials that should be accounted for when estimating treatment effects that incorporate indirect evidence. This is sound reasoning because reference-arm outcomes are likely to reflect the integrated effects of multiple observed and unobserved trial-level factors that are also likely to impact treatment arm outcomes [16]. However, following this reasoning further, we should aim to adjust for variation in reference-arm outcomes in the best possible way, making full use of the data, rather than assuming a fixed relationship.

- Statistical methods, both Bayesian & Frequentist, are available

Methodological guidance from the NICE DSU explains Bayesian approaches to adjust for reference-arm responses in ITC/NMA [15]. The present paper lays out a Frequentist approach, which may be preferred by some decision-makers.

- Multiple real data examples have already shown substantial impacts & clear improvements in estimated treatment effects owing to reference-arm adjustment

The examples summarized here show that adjustment for reference-arm outcomes can lead to significantly better model fit to the observed data and substantial changes in the estimated treatment effects. In particular, in the re-analysis of the HIV data from Chou *et al.*, the direction of the indirectly estimated treatment difference is reversed after reference-arm adjustment and brought into alignment with results from head-to-head randomized trials. These findings highlight the potential importance of reference-arm adjustment to reduction in bias. In addition, other sensitivity analyses that are typically conducted for ITC/NMA, such as use of fixed versus random effects, typically have less impact on estimated effects and do not as directly address the potential for confounding

as reference-arm adjustment. These considerations support greater priority on reference-arm adjustment, at least as a sensitivity analysis, in ITC/NMA.

Bayesian & Frequentist approaches

The purpose of this study was not to compare Bayesian and Frequentist approaches. However, a clear advantage of the Bayesian approach, compared with the Frequentist approach introduced here, is that the Bayesian calculations can be readily extended to settings with multiple comparator treatments and to networks in which not all trials share a single common reference arm (such as the third example related to psoriasis therapies). In the present example based on HIV data, the Bayesian and Frequentist point estimates were similar ([Supplementary Table 1](#)); however, the Bayesian approach had wider credible intervals. This is potentially a result of the use of vague priors for the trial-specific baseline risks. Hierarchical priors could be considered, especially when the trials have small sample sizes.

Effect heterogeneity, effect modification & confounding

Adjustment for reference-arm response in ITC/NMA is related to the issues of effect heterogeneity and modification in the underlying trials. In model (2) above (anchored ITC without reference-arm adjustment), effect heterogeneity is captured by ε . That is, the model allows some fluctuation of the treatment effect $Y-X$ across trials, centered around $\alpha + \theta Z$. However, effect heterogeneity that is associated with X constituted effect modification and can lead to bias that cannot be adjusted away under the model in (2). Moreover, typical approaches to detecting effect heterogeneity (e.g., Cochran's Q [Cochran 51]) and I^2 [Higgins *et al.* 50]) are global tests and are not powered to detect effect heterogeneity associated with X , which can confound traditional adjusted ITCs. Likewise, accounting for effect heterogeneity (e.g., using DerSimonian and Laird [18]) in the separate direct comparison meta-analyses for $Z = 0$ and $Z = 1$ will provide no adjustment for effect heterogeneity associated with X and will not remove bias from the indirect estimation of causal effects.

For these reasons, effect modification and confounding are not separable for ITC/NMA in the way that they are for pairwise meta-analyses of randomized controlled trials. That is, for pairwise meta-analyses, we always obtain an unconfounded pooled effect estimate across trials due to within-trial randomization. Effect modification can impact the generalizability of that estimate to populations that differ from the aggregate of the study populations, but it does not introduce confounding bias. However, in an ITC effect modification will result in confounding bias if the effect modifiers are not well balanced across studies. As shown in the examples reported in this paper, this bias can be substantial. Therefore, accounting for effect modification seems necessary and central to ITC/NMA, in contrast to pairwise meta-analyses where effect modification is often positioned as optional or secondary.

Good practices & appropriate interpretation

Several issues arise in the application and interpretation of anchor-based adjustment in ITC/NMA. First, since reference-arm outcomes are an important potential confounding factor, it is helpful to report these outcomes across trials and assess their heterogeneity, in addition to the usual reporting of relative treatment effects and assessment of heterogeneity in baseline characteristics. Second, when reference-arm adjustment is attempted, it will be necessary to decide which model should be preferred: the one with versus without adjustment. While DIC or other criteria are helpful for model selection, they should not be seen as determining criteria for whether reference-arm adjustments are needed. Rather, changes in the effect estimate, along with the clinical rationale for adjustment, should be the primary indicators of confounding bias and need for adjustment as is standard in other approaches to causal inference [52].

Limitations

There are a number of important limitations to reference-arm adjustment in ITC/NMA. First, even after adjusting for reference-arm response, along with any other observed factors, there may still be confounding bias owing to unobserved factors. Second, it is important to note that any attempt to infer causal effects at the individual level using only trial-level data would be subject to ecological bias [53,54]. Finally, it is critical to acknowledge that reference-arm adjustment is not always feasible in ITC/NMA. In general, if each treatment has only one trial available, it will not be possible to estimate the reference-arm effect. Rather, multiple trials per treatment are necessary to resolve the reference-arm effect. Similarly, it will often be necessary to assume that the reference-arm effects are consistent across treatments (i.e., there are not interactions between reference-arm effects and treatment

type) because of limited numbers of trials for some or all treatments. These limitations have been well described previously [15].

Areas for future research

Reference-arm adjustment raises a number of choices for model specification and future research would be helpful to inform these choices. In many cases, a single type of reference arm and outcome is widely available across trials and presents itself as an obvious choice for adjustment. However, in other cases, an evidence network might include multiple treatment arms and outcomes that could be used for adjustment. Research on how to select the treatment arm and outcome(s) to use for adjustment would be valuable. Second, if not all trials have a common reference arm, then reliance on the prior will increase and use of vague priors could become problematic (i.e., it may be better to use a hierarchical prior to help fill in the missing data in trials without the reference arm). There is a need for better understanding of approaches to address this situation. Finally, the sensitivity of anchor-based ITC/NMA to choice of effect measure (e.g., odds ratio vs risk difference) has been noted [55]. A hypothesis, which could be investigated in the future, is that reference-arm adjustment should reduce sensitivity to effect measure choice, since this sensitivity is exacerbated by differences in reference arm outcomes.

Conclusion

This paper has provided theoretical arguments and empirical examples that underscore the importance of adjusting for reference-arm effects in ITC/NMA. Based on these considerations, we advocate for higher priority to reference-arm adjustment in ITC/NMA design, reporting and interpretation. In particular, we recommend either attempting to adjust for reference-arm response and reporting the results of the attempt, reporting that adjustment is not feasible because of a limited number of trials and/or providing clinical arguments as to why reference-arm adjustment is not appropriate.

Summary points

- Indirect treatment comparison (ITC) approaches, such as network meta-analysis (NMAs), can provide valuable estimates of relative effects of treatments that have not been compared directly.
- Despite the value of ITC/NMA, estimates can be biased by imbalances in both observed and unobserved baseline characteristics and other factors that differ across clinical trials.
- Many cross-trial differences are manifested in differences in outcomes observed in reference arms (e.g., placebo arms) across trials. Methods to adjust for reference-arm outcomes have been proposed and included in guidance documents and have demonstrated value across multiple applications, but are not widely used in ITC/NMA.
- In this study, we first present strategies to adjust for reference-arm effects within a causal inference framework and lay out existing Bayesian and new Frequentist approaches that are then applied to three distinct examples of real world data related to antiretroviral therapy for human immunodeficiency virus, treatments for Type 2 diabetes and biological treatments for moderate-to-severe psoriasis.
- Results showed that reference-arm adjustment has a meaningful impact on estimated treatment effects, significantly improving model fit to observed data and yielding indirect estimates that are more consistent with the findings of randomized trials.
- Both the theoretical arguments and empirical examples highlighted in this work underscore the importance of adjusting for reference-arm effects in ITC/NMA to avoid potential bias.
- Reference-arm adjustment, while not always feasible, should be more consistently considered and reported in ITC.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/cer-2020-0042

Author contributions

J Signorovitch, R Ayyagari, E Swallow and O Patterson-Lomba were involved in the conception and design of the study and the analysis and interpretation of the data. C Pelletier and R Mehta were involved in the analysis and interpretation of the data. All authors contributed to the drafting of the paper and offered critical revisions for intellectual content. All authors agree to be accountable for all aspects of the work and approved of the final version for publication.

Financial & competing interests disclosure

J Signorovitch, R Ayyagari, E Swallow and O Patterson-Lomba are employees of Analysis Group Inc., which is a paid consultant to Celgene, now a wholly-owned subsidiary of Bristol-Myers Squibb. C Pelletier is an employee of Bristol-Myers Squibb, Summit, NJ, USA; R Mehta was an employee of Celgene at the time the study was conducted. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Editorial support was received from Peloton Advantage, LLC, an OPEN Health company, Parsippany, NJ, USA, sponsored by Bristol-Myers Squibb.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

Papers of special note have been highlighted as: • of interest

1. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J. Clin. Epidemiol.* 50(6), 683–691 (1997).
2. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat. Med.* 21(16), 2313–2324 (2002).
3. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat. Med.* 23(20), 3105–3124 (2004).
4. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ* 326(7387), 472 (2003).
5. Glenny AM, Altman DG, Song F *et al.* Indirect comparisons of competing interventions. *Health Technol. Assess.* 9(26), 1–134 iii–iv (2005).
6. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* 26(9), 753–767 (2008).
7. Hoaglin DC, Hawkins N, Jansen JP *et al.* Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. *Value Health* 14(4), 429–437 (2011).
8. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. National Institute for Health and Care Excellence (NICE), London, UK, (2020). www.ncbi.nlm.nih.gov/books/NBK310366/pdf/Bookshelf_NBK310366.pdf
- **A technical support document from the National Institute for Health and Clinical Excellence (NICE) that lays out the methodological foundations of network meta-analyses.**
9. Jansen JP, Fleurence R, Devine B *et al.* Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. *Value Health* 14(4), 417–428 (2011).
- **A report from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) that lays out the methodological foundations of network meta-analyses.**
10. Australian Government Department of Health. Pharmaceutical Benefits Advisory Committee (PBAC) technical working group reports: Indirect Comparisons Working Group (ICWG) report. Australian Government Department of Health, Canberra, Australia, (2020). www.pbs.gov.au/info/industry/useful-resources/pbac-feedback
11. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 20(4), 488–495 (2009).
12. Neyman J. [Sur les applications de la theorie des probabilités aux expériences agricoles: essai des principes]. *Roczniki Nauk Rolniczych* 10, 1–51 (1923).
13. Rubin DB. Estimating casual effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66(5), 688–701 (1974).
- **Introduces the causal inference framework.**
14. Holland PW. Statistics and causal inference. *J. Am. Stat. Assoc.* 81(396), 945–960 (1986).
15. Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU technical support document 3: heterogeneity: subgroups, meta-regression, bias and bias-adjustment. National Institute for Health and Care Excellence (NICE), London, UK, (2020). <http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD3-Heterogeneity.final-report.08.05.12.pdf>
- **A guidance document from the National Institute for Health and Care Excellence with methods for reference-arm adjustment.**
16. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat. Med.* 16(23), 2741–2758 (1997).

17. Chou R, Fu R, Huffman LH, Korthuis PT. Initial highly-active antiretroviral therapy with a protease inhibitor versus a non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. *Lancet* 368(9546), 1503–1515 (2006).
- **A direct meta-analysis of antiretroviral therapy for virological suppression of the HIV.**
18. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control. Clin. Trials* 7(3), 177–188 (1986).
19. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat. Med.* 14(24), 2685–2699 (1995).
20. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6(1), 34–58 (1978).
21. Rosenbaum PR, Rubin DB. Comment: estimating the effects caused by treatments. *J. Am. Stat. Assoc.* 79(385), 26–28 (1984).
22. Winship C, Morgan SL. The estimation of casual effects from observational data. *Annu. Rev. Sociol.* 25, 659–706 (1999).
23. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for casual effects. *Biometrika* 70(1), 41–55 (1983).
24. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 10(1), 37–48 (1999).
25. Pearl J, Glymour M, Jewell NP. *Causal Inference in Statistics: A Primer*. John Wiley and Sons, Inc, NJ, USA (2016).
26. Pearl J. Direct and indirect effects [technical report]. Presented at: *17th Conference on Uncertainty in Artificial Intelligence*. WA, USA (2001).
27. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 38(4), 963–974 (1982).
28. McCulloch CE, Searle SR. *Generalized, linear and mixed models*. John Wiley and Sons, Inc, NJ, USA, Hoboken (2000).
29. Ayyagari R. The importance of adjustment for placebo-arm effects in indirect comparisons of glycemic control in Type 2 diabetes mellitus. Presented at: *1st American Diabetes Association Middle East Congress*. Dubai, United Arab Emirates (2012).
- **An indirect comparison of hemoglobin A1C outcomes in Type 2 diabetes.**
30. Lamel SA, Myer KA, Younes N, Zhou JA, Maibach H, Maibach HI. Placebo response in relation to clinical trial design: a systematic review and meta-analysis of randomized controlled trials for determining biologic efficacy in psoriasis treatment. *Arch. Dermatol. Res.* 304(9), 707–717 (2012).
31. Signorovitch JE, Betts KA, Yan YS *et al.* Comparative efficacy of biological treatments for moderate-to-severe psoriasis: a network meta-analysis adjusting for cross-trial differences in reference arm response. *Br. J. Dermatol.* 172(2), 504–512 (2015).
- **A network meta-analysis that assesses the short-term efficacy of biological treatments for moderate-to-severe psoriasis.**
32. Gordon KB, Langley RG, Leonardi C *et al.* Clinical response to adalimumab treatment in patients with moderate to severe psoriasis: double-blind, randomized controlled trial and open-label extension study. *J. Am. Acad. Dermatol.* 55(4), 598–606 (2006).
33. Menter A, Feldman SR, Weinstein GD *et al.* A randomized comparison of continuous vs. intermittent infliximab maintenance regimens over 1 year in the treatment of moderate-to-severe plaque psoriasis. *J. Am. Acad. Dermatol.* 56(1), 31.e1–15 (2007).
34. Saurat JH, Stingl G, Dubertret L *et al.* Efficacy and safety results from the randomized controlled comparative study of adalimumab vs. methotrexate vs. placebo in patients with psoriasis (CHAMPION). *Br. J. Dermatol.* 158(3), 558–566 (2008).
35. Gottlieb AB, Matheson RT, Lowe N *et al.* A randomized trial of etanercept as monotherapy for psoriasis. *Arch. Dermatol.* 139(12), 1627–1632
36. Gottlieb AB, Evans R, Li S *et al.* Infliximab induction therapy for patients with severe plaque-type psoriasis: a randomized, double-blind, placebo-controlled trial. *J. Am. Acad. Dermatol.* 51(4), 534–542 (2004).
37. Leonardi CL, Powers JL, Matheson RT *et al.* Etanercept as monotherapy in patients with psoriasis. *N. Engl. J. Med.* 349(21), 2014–2022 (2003).
38. Leonardi CL, Kimball AB, Papp KA *et al.* Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 76-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 1). *Lancet* 371(9625), 1665–1674 (2008).
39. Papp KA, Tying S, Lahfa M *et al.* A global phase III randomized controlled trial of etanercept in psoriasis: safety, efficacy and effect of dose reduction. *Br. J. Dermatol.* 152(6), 1304–1312 (2005).
40. Papp KA, Langley RG, Lebwohl M *et al.* Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 52-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 2). *Lancet* 371(9625), 1675–1684 (2008).
41. Tying S, Gordon KB, Poulin Y *et al.* Long-term safety and efficacy of 50 mg of etanercept twice weekly in patients with psoriasis. *Arch. Dermatol.* 143(6), 719–726 (2007).
42. van de Kerkhof PC, Segaert S, Lahfa M *et al.* Once weekly administration of etanercept 50 mg is efficacious and well tolerated in patients with moderate-to-severe plaque psoriasis: a randomized controlled trial with open-label extension. *Br. J. Dermatol.* 159(5), 1177–1185 (2008).
43. Chaudhari U, Romano P, Mulcahy LD, Dooley LT, Baker DG, Gottlieb AB. Efficacy and safety of infliximab monotherapy for plaque-type psoriasis: a randomised trial. *Lancet* 357(9271), 1842–1847 (2001).

44. Reich K, Nestle FO, Papp K *et al.* Infliximab induction and maintenance therapy for moderate-to-severe psoriasis: a Phase III, multicentre, double-blind trial. *Lancet* 366(9494), 1367–1374 (2005).
45. Griffiths CE, Strober BE, van de Kerkhof P *et al.* Comparison of ustekinumab and etanercept for moderate-to-severe psoriasis. *N. Engl. J. Med.* 362(2), 118–128 (2010).
46. Menter A, Tying SK, Gordon K *et al.* Adalimumab therapy for moderate to severe psoriasis: a randomized, controlled Phase III trial. *J. Am. Acad. Dermatol.* 58(1), 106–115 (2008).
47. Armstrong AW, Puig L, Joshi A *et al.* Comparison of biologics and oral treatments for plaque psoriasis: a meta-analysis. *JAMA Dermatol.* 156(3), 258–269 (2020).
48. Armstrong AW, Betts KA, Signorovitch JE *et al.* Number needed to treat and costs per responder among biologic treatments for moderate-to-severe psoriasis: a network meta-analysis. *Curr. Med. Res. Opin.* 34(7), 1325–1333 (2018).
49. Institute for Clinical and Economic Review. Targeted immunomodulators for the treatment of moderate-to-severe plaque psoriasis: effectiveness and value. Condition update [final evidence report] (2020). https://icer-review.org/wp-content/uploads/2017/11/ICER_Psoriasis_Update_Draft_Report_04272018.pdf
50. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 327(7414), 557–560 (2003).
51. Cochran WG. The combination of estimates from different experiments. *Biometrics* 10(1), 101–129 (1954).
52. VanderWeele TJ. Principles of confounder selection [essay]. *Eur. J. Epidemiol.* 34(3), 211–219 (2019).
53. Piantadosi S, Byar DP, Green SB. The ecological fallacy. *Am. J. Epidemiol.* 127(5), 893–904 (1988).
54. Greenland S, Robins J. Invited commentary: ecologic studies—biases, misconceptions and counterexamples. *Am. J. Epidemiol.* 139(8), 747–760 (1994).
55. Signorovitch JE, Sikirica V, Erder MH *et al.* Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value Health* 15(6), 940–947 (2012).