REVIEW

The number needed to treat and relevant between-trial comparisons of competing interventions

Jeroen P Jansen[1]
Javaria Mona Khalid[2]
Michael D Smyth[3]
Haridarshan Patel[4]

[1]Evidence Synthesis and Decision Modeling, Precision Xtract, Oakland, CA, USA; [2]Evidence and Value Generation, Takeda International – UK Branch, London, UK; [3]Global Medical Affairs, Takeda Development Centre Europe Ltd, London, UK; [4]Evidence and Value Generation, Takeda International, Deerfield, IL, USA

Abstract: The number needed to treat (NNT) is considered an intuitive as well as popular effect measure. The aims of this review were to 1) explain why we cannot compare trial-specific NNT estimates for the competing treatments evaluated in different randomized controlled trials (RCTs) and 2) outline the principles of how relative treatment effects of different trials can be compared and results can be presented as NNT, without violating the principles of valid between-trial comparisons. Our premise is that ratio measures for relative treatment effects of response outcomes are less prone to effect modification than absolute difference measures of response outcomes. Accordingly, any between-trial comparisons of the efficacy of competing interventions using the study-specific ORs are less likely to be invalid or biased than comparisons based on the study-specific NNT estimates. However, treatment-specific ORs obtained from a meta-analysis or taken directly from an individual study can be transformed into consistent treatment-specific NNT estimates that allow for credible comparisons of treatments when these ratio measures are applied to the same reference response estimate. The theoretical discussion is illustrated with a relevant indirect comparison of biologics for the treatment of ulcerative colitis. Between-trial comparisons directly based on the NNT of individual trials may result in erroneous conclusions and should be avoided. Treatment-specific NNT estimates need to be based on the same probability of response with the common reference treatment against which the interventions are compared.

Keywords: biologics, indirect treatment comparison, network meta-analysis, treatment outcomes, ulcerative colitis

Introduction

Randomized controlled trials (RCTs) are considered important for informing clinical practice. Sound clinical decision making requires comparisons of all clinically appropriate interventions for a particular disease state. However, RCTs that simultaneously compare all interventions of interest are rarely available. As an alternative, a between-trial or indirect comparison of two or more treatments can provide useful estimates for the difference in treatment effects if each has been compared against the same comparator.[1,2]

Relative treatment effects can be presented with several statistical measures. For dichotomous outcome measures, such as treatment response, the OR, relative risk (RR), RR reduction, or risk difference (RD) is frequently used. Another measure suitable to express efficacy of a medical intervention is the number needed to treat (NNT). The NNT represents the expected number of patients who need to be treated with an

Correspondence: Haridarshan Patel
Evidence and Value Generation, Takeda International, 1 Takeda Pkwy, Deerfield, IL 60015, USA
Tel +1 847 809 4832
Email hari.patel@takeda.com

intervention to observe one patient successfully achieving the desired outcome in a specific patient population. NNT is popular for its ease of interpretation.[3]

In this paper, we discuss why we cannot simply compare trial-specific NNT estimates for the competing treatments evaluated in different RCTs and conclude that the treatment with the lowest NNT is the most efficacious. Next, we outline how the efficacy of competing interventions studied in different trials can be validly compared while still presenting results as NNT, which facilitates communication and interpretation. This is illustrated with a relevant comparison of biologics used to treat patients with moderately to severely active ulcerative colitis (UC).

Why we cannot compare the NNT of different clinical trials

The "gold standard" approach to assess the efficacy of a particular intervention is the RCT. These trials aim to quantify the independent effect of an intervention on an outcome of interest by randomly allocating patients to intervention and control groups, thus having a similar distribution of (un)known and (un)measured patient-related factors across these groups within a trial. Assuming that there are no other systematic differences between groups, the difference in outcomes between the treatment and control groups is attributable to the difference regarding interventions between these groups: the treatment effect.

Subgroups of patients distinguished by differences in their characteristics may respond to treatments differently, thereby demonstrating different relative treatment effects. This heterogeneity in relative treatment effects is due to patient characteristics that are effect modifiers. The extent of variation in relative treatment effects is also influenced by the statistical effect measure used. It has been suggested that, for response outcomes, the effect measures on multiplicative scales, such as the OR or RR, are less heterogeneous than the effect measures on additive scales, such as the RD.[4–6] Reviews of meta-analysis studies found a *P* value indicating between-trial heterogeneity to be less frequent in studies using ORs or RRs compared with those using RDs.[5,6] It remains unclear whether the differences observed are due to the sensitivity of the various effect measures, the impact of effect modification, or the influence of the statistical power of the measures to detect between-trial heterogeneity.[7] Nonetheless, identifying heterogeneity in relative treatment effects on a multiplicative scale seems to be of greater clinical relevance than identifying heterogeneity on the additive scale, and it has been argued that this suggests something about the underlying biology as to how effects operate.[7] For the remainder of this paper, we

build upon the premise of less heterogeneity with a multiplicative effect measure.

In a situation wherein we have a number of placebo-controlled trials that are all evaluating the same treatment, we assume that the extent of the differences in study design or patient characteristics between studies is not influencing how well the treatment works. Accordingly, we do not expect any heterogeneity in the relative treatment effects on the multiplicative scale, eg, all observed ORs are similar. However, if some of the patient characteristics are prognostic or risk factors for the outcome of interest independent of treatment, then we do see differences in the placebo response across studies. If we express the difference in response between the treatment and placebo groups in each trial on the additive scale with the RD, we do see between-study variation (despite similar ORs). This heterogeneity in relative treatment effects using the RD has nothing to do with the impact of patient characteristics on the treatment efficacy and is purely a result of differences in the placebo response. The NNT is the inverse of the RD (NNT =1/RD). As a result, the NNT is equally as sensitive to the baseline risk as the RD.

With indirect or between-trial comparisons, we like to capitalize on the within-trial randomization and use the trial-specific relative treatment effects as the measures to compare. It is well known that between-trial comparisons of relative treatment effects of interventions compared with the same control are biased if there are between-trial differences regarding study design and patient characteristics associated with the relative treatment effects.[2,8] Given the premise that effect measures on a multiplicative scale are less prone to effect modification by patient characteristics than effect measures on the additive scale, indirect or between-trial comparisons are less likely to be biased using the OR than the RD or NNT. Between-trial comparisons of the relative treatment effects on the multiplicative scale are only biased if there are between-trial differences in effect modifiers on this multiplicative scale. Between-trial differences in patient characteristics that are only prognostic factors, ie, factors having an impact on the outcomes independent of treatment, are not biasing the indirect comparison. However, with relative treatment effects expressed as RD or NNT, not only do the differences in patient characteristics that are effect modifiers on the multiplicative scale (and therefore likely effect modifiers on the additive scale) bias the between-trial comparison but also between-trial differences in patient characteristics impacting baseline risk (which now become effect modifiers for the NNT) will bias the indirect comparison. In other words, if there are two trials, one comparing B with placebo and the other C with placebo, simply concluding that with

a more favorable NNT one treatment is more efficacious than the other will be incorrect if the baseline risk differs between the trials.

How to derive NNT in the context of between-trial comparisons

Given the properties of the NNT, between-trial or indirect comparisons based on the NNT of individual trials should be avoided when there are differences in baseline risk, as outlined earlier. However, this does not mean that we need to avoid the NNT altogether. There are ways in which we can suitably make comparisons. First, we need to define the reference treatment against which all interventions are being compared (eg, placebo). Next, we need to estimate the OR for each of the competing interventions of interest relative to this reference treatment. We can take these estimates directly from the trials if, for each intervention, only one study is available and all use the same reference treatment of choice. Alternatively, we need to use (network) meta-analysis techniques to obtain these treatment-specific ORs. Third, the OR estimates can be transformed into treatment-specific NNT estimates by applying the ORs to the probability of response with the reference treatment (eg, placebo) corresponding to the population that we are interested in. Specifically, when the baseline probability of response to the reference treatment is expressed as odds, it can be multiplied with the ORs to obtain the odds of response for each of the interventions k and subsequently transformed into a probability. In other words,

$$\text{odds}_{\text{treatment } k} = \text{OR}_{\text{treatment } k \text{ vs. reference tx}} \cdot \text{odds}_{\text{reference tx}}, \text{ and}$$

$$p_{\text{treatment } k} = \frac{\text{odds}_{\text{treatment } k}}{1 + \text{odds}_{\text{treatment } k}}.$$

The probability of response to each treatment minus the probability of response to the overall reference treatment (or placebo) is the RD, that is,

$$\text{RD}_{\text{treatment } k \text{ vs. reference tx}} = p_{\text{treatment } k} - p_{\text{reference tx}}, \text{ which in}$$

turn can be transformed into NNT estimates:

$$\text{NNT}_{\text{treatment } k} = \frac{1}{\text{RD}_{\text{treatment } k \text{ vs. reference tx}}}.$$

These treatment-specific NNTs can now be interpreted and compared with the same confidence as the OR estimates.

Relevant between-trial comparisons

Indirect or between-trial comparisons of competing interventions need to be clinically relevant for decision making. More specifically, before any comparison, it is essential to define the target population of interest and the appropriate interventions for this population. In the absence of this step, we run the risk

of performing between-trial comparisons driven primarily by the availability of data, which may result in evaluations that are not informative or may even lead to incorrect decisions. For example, the RCTs evaluating biologics for inflammatory bowel disease may assess their efficacy as induction therapy (ie, inducing a response among patients with active disease), the efficacy of maintenance therapy (ie, maintaining a response or remission as a result of induction treatment), or both. Although one can perform an indirect comparison of RCTs that evaluated the maintenance of response among randomized patients having experienced an induction response, it has little relevance from a decision-making perspective. In routine practice, it is typically the case that a patient will continue with the same biologic for maintenance therapy as that with which induction response has been achieved; the choice for starting on a particular biologic will be made during active disease. A relevant comparison of the efficacy of biologics should focus on the induction phase of treatment followed by the maintenance phase. Furthermore, patients with inflammatory bowel disease can be stratified by known effect modifiers. In this case, stratification is crucially important because prior treatment history with anti-TNF agents is a strong effect modifier for treatment effects of biologics. Hence, comparative efficacy estimates may be biased if the results are combined with an overall population (ie, a mixed population of anti-TNF-naive and anti-TNF-experienced patients).

Illustrative example
Background

UC is a chronic disease characterized by inflammation and ulceration of the mucosa and submucosa of the large intestine and rectum. UC has a significant negative impact on patients' quality of life.[9,10] The aim of therapy is to induce and maintain disease remission.[11,12] Patients with moderately to severely active UC and an inadequate response or intolerance to conventional therapies and immunomodulators can be treated with biologics, including anti-TNFs such as golimumab, adalimumab, infliximab, and the anti-integrin agent vedolizumab.[13–18] The available evidence for the efficacy of these biologics is limited to a small number of RCTs, none of which include direct head-to-head comparisons of all available treatments. A between-trial comparison of relative treatment effects with these biologics is required to understand which treatment is most efficacious to guide treatment choice.

Based on the reported results for adalimumab trials, we first illustrated the impact of baseline risk when using NNT instead of a ratio measure to compare relative treatment

effects from different trials and highlight that standardization of baseline risk ensures consistent results between the OR and NNT. Second, based on the published studies identified with a systematic review, we performed an analysis to estimate NNT for response and remission with competing biologics for patients with moderately to severely active UC with and without a history of treatment with anti-TNFs. With this analysis, we took the approach as outlined in the previous sections to allow for a valid comparison of NNT estimates.

Evidence base

Relevant evidence used for the example analyses was identified using a systematic literature review.[19] In short, Medline, Embase, and Cochrane Library searches were conducted in December 2014 without time limits to identify published RCTs evaluating biologics for the induction and/or maintenance therapy of moderate-to-severe UC. Relevant RCTs had to meet the following predefined inclusion criteria: a patient population of moderate-to-severe UC that was either anti-TNF naive or experienced (either previously exposed to and/or failed); biologics of interest were infliximab, golimumab, adalimumab, and vedolizumab; and reporting clinical response and/or remission as outcome measures.

An overview of the included studies that formed the evidence base for the analyses in this example is provided in Tables S1 and S2.[13–18] For the anti-TNF-naive population, the available trials evaluated all the interventions of interest. For the anti-TNF-experienced population, the evidence base was limited to trials evaluating adalimumab and vedolizumab. Clinical outcomes of interest were clinical response and remission at the end of induction therapy as well as sustained response and remission at the end of maintenance therapy among patients starting induction therapy. Response and remission were defined according to the Mayo score reported in the individual trials. Adalimumab and infliximab clinical trials provided response and remission estimates at the end of induction therapy as well as sustained response and remission with maintenance therapy at 1 year of follow-up out of all patients starting induction therapy. Golimumab and vedolizumab trials had re-randomization of induction responders to maintenance therapy. To overcome these between-study differences in design, patient selection, and assessment of efficacy with maintenance therapy, the analysis was performed under the assumption that only patients who showed response with induction therapy would continue with the biologic as maintenance therapy (Figure S1). Recalculation of golimumab and vedolizumab maintenance outcomes was conducted in line as shown in Figure S1 to ensure appropriate

comparisons with adalimumab and infliximab trials. More specifically, the probability of induction response among all patients starting induction treatment was multiplied by the probability of sustained response at 1 year among patients who showed an induction response with the same treatment. The study-specific response and remission data used for the analyses are provided in Tables S3 and S4.[13–18]

Inconsistency between OR and NNT due to baseline RD

As summarized in Table 1,[14,15] the two RCTs that evaluated the efficacy of adalimumab as induction therapy for UC are presented. Based on the reported probabilities of remission for each treatment group, the study-specific OR, RD, and NNT for adalimumab relative to placebo were calculated. The point estimate for the OR in ULTRA-1[14] is 2.23 and 2.19 in ULTRA-2.[15] Although very similar, based on these estimates one would also expect the point estimates for the RD and NNT to be a little more favorable for ULTRA-1. However, that is not the case; ULTRA-1 showed an NNT of 10.8 and ULTRA-2 showed an NNT of 9.7. This can be explained by the difference in remission probabilities in the placebo arms of these two studies. It is easy to infer that, given the relative treatment effects in terms of ORs, NNT estimates would show greater heterogeneity across studies. When we assume the same probability of response for both ULTRA-1 and ULTRA-2 and apply the study-specific ORs to this estimate, then we get NNT estimates that are consistent with the OR estimates for both studies. The NNT is now more favorable for ULTRA-1 than for ULTRA-2 (NNT estimates standardized for placebo response are reported in Table 1). The results reported in Table 1 also illustrate that, without changing the OR, different NNTs are obtained when applied to different baseline risk estimates. As such, this example shows that baseline risk estimates need to be used that apply to the target population of interest to obtain meaningful NNT estimates for this population.

Using NNT to compare the efficacy of different biologics among patients with UC

The probabilities of induction response and remission, as well as the probabilities of sustained response and remission with each biologic relative to placebo among patients starting induction therapy, were estimated using Bayesian meta-analysis models for ordered categorical data.[1] Separate analyses were performed for anti-TNF-naive and anti-TNF-experienced populations. Clinical response and remission as

Table 1 Comparison of OR, RD, and NNT for adalimumab relative to placebo

Study-specific data									
Clinical trials	Group	Sample size	Patients in remission	Probability of induction remission	Odds of induction remission	OR	RD	NNT	Comment
ULTRA-1[14]	Placebo	130	12	0.092	0.10				OR and NNT estimates are inconsistent in terms of which study shows more favorable results
	Adalimumab	130	24	0.185	0.23	2.23	0.092	10.8	
ULTRA-2[15]	Placebo	145	16	0.110	0.12				
	Adalimumab	150	32	0.213	0.27	2.19	0.103	9.7	
Standardized to ULTRA-1 placebo response									
ULTRA-1[14]	Placebo			0.092	0.10				OR and NNT estimate are consistent
	Adalimumab			0.185	0.23	2.23	0.092	10.8	
ULTRA-2[15]	Placebo			0.092	0.10				
	Adalimumab			0.182[a]	0.22[a]	2.19	0.090[a]	11.2[a]	
Standardized to ULTRA-2 placebo response									
ULTRA-1[14]	Placebo			0.110	0.12				OR and NNT estimates are consistent
	Adalimumab			0.216[b]	0.28[b]	2.23	0.106[b]	9.4[b]	
ULTRA-2[15]	Placebo			0.110	0.12				
	Adalimumab			0.213	0.27	2.19	0.103	9.7	

Notes: [a]Obtained by multiplying ULTRA-2-specific OR for remission with ULTRA-1-specific odds of remission with placebo. [b]Obtained by multiplying ULTRA-1-specific OR for remission with ULTRA-2-specific odds of remission with placebo.
Abbreviations: NNT, number needed to treat; RD, risk difference.

defined in the clinical trials, wherein remission is a subset of the broader category of response, can be considered ordered categorical data with three mutually exclusive states: no response, response but no remission, and remission. Data of this nature can be analyzed with a multinomial probit model.[1] The multinomial model assumes that the treatment effects are the same for the three categories. The advantage of this approach over an analysis model that considers the clinical response and remission separately is that all possible outcomes (ie, nonresponse, response, and remission) are analyzed simultaneously, and uncertainty in the estimates is captured accurately. In the face of few RCTs, heterogeneity estimation was unreliable, and we relied on a fixed-effects model instead of a random-effects model. To not influence the observed results by prior belief, noninformative prior distributions were used for the model parameters to be estimated.

The relative treatment effect of each biologic vs placebo estimated on the probit scale with the fixed-effects multinomial probit model was transformed into probabilities of response and remission by combining it with the average placebo response across studies as the reference. Based on the probabilities of sustained response and remission among patients starting induction therapy, the NNT was then calculated for each biologic. Analyses were performed with the Markov Chain Monte Carlo method as implemented in the OpenBUGS software package, which avoided the challenges with calculating 95% CIs for the NNT.[20] The posterior distributions of these parameters of interest as obtained with

the Bayesian analysis were summarized by their median as a reflection of the "point estimate" and 95% credible intervals, which were constructed from the 2.5th and 97.5th percentiles.

Table 2 summarizes the estimated probabilities of clinical response and remission, as well as NNT estimates by treatment, for anti-TNF-naive patients. For the anti-TNF-naive subpopulation, infliximab showed the greatest efficacy in terms of response and remission at induction. At 1 year of treatment, however, vedolizumab demonstrated the most favorable NNT estimates for a sustained response and sustained remission among patients starting induction therapy. Table 3 summarizes the results of the analysis performed for the anti-TNF-experienced population. For this subpopulation, vedolizumab showed a higher probability of achieving response and remission at induction. In addition, vedolizumab demonstrated a higher probability of sustained response and sustained remission following the maintenance phase of treatment (at 1 year) and consequently has the more favorable NNT estimate.

Discussion

It is frequently assumed that ratio measures for relative treatment effects of response outcomes, such as the OR, are less prone to effect modification than the RD or its inverse, the NNT. Accordingly, any between-trial comparison of the efficacy of competing interventions using the study-specific ORs is expected to be less prone to differences in patient characteristics between studies than comparisons based on

Table 2 Estimated probability and NNT for response and remission with competing biologics among anti-TNF-naive patients with UC

Treatment	Estimate	95% credible interval	Estimate	95% credible interval	Estimate	95% credible interval
	Probability of induction response		Probability of sustained response at 1 year		NNT for sustained response at 1 year	
Placebo	0.34	(0.31; 0.37)	0.12	(0.09; 0.15)	Reference	
Infliximab[a]	0.69	(0.62; 0.76)	0.35	(0.24; 0.47)	4.4	(2.9; 8.6)
Adalimumab[b]	0.49	(0.42; 0.56)	0.22	(0.14; 0.32)	10.2	(5.0; 45.4)
Golimumab[c]	0.56	(0.47; 0.64)	0.31	(0.22; 0.40)	5.4	(3.5; 9.9)
Vedolizumab[d]	0.63	(0.51; 0.75)	0.40	(0.26; 0.55)	3.6	(2.2; 7.6)
	Probability of induction remission		Probability of sustained remission at 1 year		NNT for sustained remission at 1 year	
Placebo	0.09	(0.07; 0.11)	0.04	(0.03; 0.05)	Reference	
Infliximab[a]	0.34	(0.27; 0.41)	0.16	(0.09; 0.24)	8.4	(4.8; 18.4)
Adalimumab[b]	0.17	(0.13; 0.22)	0.08	(0.04; 0.14)	22.3	(9.7; 101.2)
Golimumab[c]	0.22	(0.16; 0.28)	0.13	(0.08; 0.20)	10.7	(6.3; 21.8)
Vedolizumab[d]	0.28	(0.18; 0.40)	0.19	(0.11; 0.32)	6.5	(3.5; 15.6)

Notes: [a]Infliximab 5 mg/kg induction therapy (weeks 0, 2, and 6) followed by every 8 weeks as maintenance therapy. [b]Adalimumab 160/80 mg induction therapy followed by 40 mg every other week as maintenance therapy. [c]Golimumab 200/100 mg induction therapy (weeks 0, 2, and 6) followed by 100 mg every 4 weeks as maintenance therapy. [d]Vedolizumab 300 mg induction therapy (weeks 0, 2, and 6) followed by every 8 weeks as maintenance therapy.
Abbreviations: NNT, number needed to treat; TNF, tumor necrosis factor; UC, ulcerative colitis.

Table 3 Estimated probability and NNT for response and remission with competing biologics among anti-TNF-experienced patients with UC

Treatment	Estimate	95% credible interval	Estimate	95% credible interval	Estimate	95% credible interval
	Probability of induction response		Probability of sustained response at 1 year		NNT for sustained response at 1 year	
Placebo	0.25	(0.18; 0.32)	0.04	(0.02; 0.08)	Reference	
Adalimumab[a]	0.32	(0.21; 0.44)	0.12	(0.04; 0.23)	13.5	(4.9; 80.9)
Vedolizumab[b]	0.44	(0.31; 0.58)	0.24	(0.12; 0.42)	5.1	(2.5; 15.8)
	Probability of induction remission		Probability of sustained remission at 1 year		NNT for sustained remission at 1 year	
Placebo	0.05	(0.03; 0.08)	0.01	(0.00; 0.03)	Reference	
Adalimumab[a]	0.07	(0.03; 0.13)	0.04	(0.01; 0.10)	39.2	(10.9; 283.1)
Vedolizumab[b]	0.13	(0.07; 0.22)	0.09	(0.03; 0.22)	12.0	(4.8; 44.5)

Notes: [a]Adalimumab 160/80 mg induction therapy followed by 40 mg every other week as maintenance therapy. [b]Vedolizumab 300 mg induction therapy (weeks 0, 2, and 6) followed by every 8 weeks as maintenance therapy.
Abbreviations: NNT, number needed to treat; TNF, tumor necrosis factor; UC, ulcerative colitis.

the study-specific NNT estimates and thereby less likely to be invalid or biased. However, the NNT is considered an intuitive measure to understand the efficacy of an intervention. To obtain consistent treatment-specific NNT estimates that allow for appropriate comparisons of competing interventions evaluated in different trials, the treatment-specific ORs derived from a meta-analysis or directly taken from an individual study need to be applied to one and the same reference response probability. In the absence of this approach, treatment choice informed by trial-specific NNT estimates may result in erroneous decisions.

Data sharing statement

All data were from publicly available publications of pivotal clinical trial results.

Author contributions

All authors contributed to the study concept and design, or data analysis or interpretation, drafting and revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Disclosure

Jeroen P Jansen is an employee of Precision Xtract, which received funding from Takeda. Haridarshan Patel is an employee of Takeda International, Deerfield, IL, USA. Javaria Mona Khalid is an employee of Takeda International, UK branch. Michael D Smyth was an employee of Takeda at the time of manuscript development. The authors report no other conflicts of interest in this work.

References

1. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33(5):607–617.
2. 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*. 2011;14(4):417–428.
3. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310(6977):452–454.
4. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med*. 2000;19(13):1707–1728.
5. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol*. 2001;54(10):1046–1055.
6. Deeks JJ, Althman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG, editors. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ Publishing Group; 2003:313–335.
7. VanderWeele TJ, Knol MJ. A tutorial on interaction. *Epidemiol Method*. 2014;3(1):33–72.
8. Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Med*. 2013;11:159.
9. Cohen RD, Yu AP, Wu EQ, Xie J, Mulani PM, Chao J. Systematic review: the costs of ulcerative colitis in Western countries. *Aliment Pharmacol Ther*. 2010;31(7):693–707.
10. Rubin DT, Dubinsky MC, Panaccione R, et al. The impact of ulcerative colitis on patients' lives compared to other chronic diseases: a patient survey. *Dig Dis Sci*. 2010;55(4):1044–1052.
11. Reinisch W, Sandborn WJ, Bala M, et al. Response and remission are associated with improved quality of life, employment and disability status, hours worked, and productivity of patients with ulcerative colitis. *Inflamm Bowel Dis*. 2007;13(9):1135–1140.
12. Vogelaar L, Spijker AV, van der Woude CJ. The impact of biologics on health-related quality of life in patients with inflammatory bowel disease. *Clin Exp Gastroenterol*. 2009;2:101–109.
13. Rutgeerts P, Sandborn WJ, Feagan BG, et al. Infliximab for induction and maintenance therapy for ulcerative colitis. *N Engl J Med*. 2005;353(23):2462–2476.
14. Reinisch W, Sandborn WJ, Hommes DW, et al. Adalimumab for induction of clinical remission in moderately to severely active ulcerative colitis: results of a randomised controlled trial. *Gut*. 2011;60(6): 780–787.
15. Sandborn WJ, van Assche G, Reinisch W, et al. Adalimumab induces and maintains clinical remission in patients with moderate-to-severe ulcerative colitis. *Gastroenterology*. 2012;142(2):257–265.e1-3.
16. Sandborn WJ, Feagan BG, Marano C, et al. Subcutaneous golimumab induces clinical response and remission in patients with moderate-to-severe ulcerative colitis. *Gastroenterology*. 2014;146(1):85–95; quiz e14-5.
17. Sandborn WJ, Feagan BG, Marano C, et al. Subcutaneous golimumab maintains clinical response in patients with moderate-to-severe ulcerative colitis. *Gastroenterology*. 2014;146(1):e101:96–109.e1.
18. Feagan BG, Rutgeerts P, Sands BE, et al. Vedolizumab as induction and maintenance therapy for ulcerative colitis. *N Engl J Med*. 2013;369(8):699–710.
19. Vickers AD, Ainsworth C, Mody R, et al. Systematic review with network meta-analysis: comparative efficacy of biologics in the treatment of moderately to severely active ulcerative colitis. *PLoS One*. 2016;11(10):e0165435.
20. Bender R. Calculating confidence intervals for the number needed to treat. *Control Clin Trials*. 2001;22(2):102–110.