Neurol Ther (2017) 6:1–9
DOI 10.1007/s40120-017-0063-y

REVIEW

Number Needed to Treat in Multiple Sclerosis Clinical Trials

Macaulay Okwuokenye · Annie Zhang · Amy Pace · Karl E. Peace

Received: November 23, 2016 / Published online: February 7, 2017

ABSTRACT

Clinicians are expected to select a therapy based on their appraisal of evidence on benefit-to-risk profiles of therapies. In the management of relapsing-remitting multiple sclerosis (RRMS), evidence is typically expressed in terms of risk (proportion) of event, risk reduction, relative and hazard rate reduction, or relative reduction in the mean number of magnetic resonance imaging lesions. Interpreting treatment effect using these measures from a RRMS clinical trial is fairly reliable; however, this might not be the case when treatment effect is expressed in terms of the number needed to treat (NNT). The objective of this review is to discuss the utility of NNT in RRMS trials. This article presents an overview of the methodological definition and characteristics of NNT as well as the relative merit of NNT use in RRMS controlled clinical trials, where endpoints are typically time-to-event and frequency of recurrent events. The authors caution against using NNT in multiple sclerosis, a clinically heterogeneous disease that can change course and severity unpredictably. The authors also caution against the use of NNT to interpret results in comparative trials where the absolute risk difference is not statistically significant, computing NNT using the time--to-event endpoint at intermediate time points, computing NNT using the annualized relapse rate, and comparing NNT across trials.

Keywords: Absolute risk difference; Annualized relapse rate; Controlled clinical trials; Number needed to treat; Relapsing-remitting multiple sclerosis; Statistical inference

INTRODUCTION

The number needed to treat (NNT) was originally used to provide greater clinical relevance to a small risk difference (RD) [1–3], particularly when that RD was converted to a percentage of control. For example, when the control (C) rate is 1% and the treatment (T) rate is 0.67% for an event of medical interest (e.g., myocardial infarction [MI]), the estimated absolute risk difference (ARD) is 0.0067–0.01 = 0.0033 or 0.33%. This becomes 33% (0.0033/0.01) when expressed as a percentage of C.

Enhanced content To view enhanced content for this article go to http://www.medengine.com/Redeem/9D77F0601D12B409.

M. Okwuokenye (✉) · A. Zhang · A. Pace
Biogen, Cambridge, MA, USA
e-mail: macaulay.okwuokenye@biogen.com

K. E. Peace
Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA

An interpretation of the ARD using the example above is that, on average, among 1000 patients treated with T and 1000 treated with C, one could expect 3.3 fewer patients to have the event in the T group than in the C group. Alternatively, one may wish to determine how many patients need to be treated to have one fewer event in the T group compared with the C group. The answer is provided by the NNT, which in this case would be $1/0.0033 = 303.03$ patients. In general, $\text{NNT} = 1/\text{ARD}$. NNT is usually rounded up to the nearest larger integer, which in this example is 304.

There is much debate in the literature [4, 5] about whether NNT is a useful measure to summarize treatment effect from a clinical trial with binary efficacy endpoints. The proponents of the summary measure find it appealing because they believe it is easy to interpret. However, some argue [6–8] that the basic definition of NNT is flawed and that NNT lacks statistical content.

Initially, it was coronary heart disease risk trials that spawned the use of NNT. These trials were large due to the relatively rare occurrence of the event in question and because of the large power needed to detect a clinically meaningful RD between the T and C groups. More recently, NNT has been used to interpret outcome data from clinical trials in many other disease areas, including multiple sclerosis (MS) [9–13], probably because of the purported ease of interpretation [14].

There is a need to clarify both the definition and use of NNT in summarizing outcome data from comparative relapsing-remitting MS (RRMS) trials so that reported NNTs can be critically appraised. This article provides an overview of the definition and methodological characteristics of NNT, and discusses the relative merits of its use in MS clinical trials, where endpoints are typically time-to-event or frequency of recurrent events.

Compliance with Ethics Guidelines

This article is based on previously conducted studies, and does not involve any new studies of human or animal subjects performed by any of the authors.

NNT FOR DICHOTOMOUS DATA

Historically, the use of NNT was based on dichotomous efficacy data and reported in trials in which T was better than C. In a clinical trial of MI (see example in the "Introduction"), an NNT of 304 is interpreted as indicating that 304 patients would need to be treated, on average, with the drug to prevent one MI. Adverse event (AE) data are collected in such trials and, by definition, reflect potential harm to the patient, particularly when the AEs are serious. Therefore, it is of clinical interest to compute the number needed to harm (NNH), which implies that the incidence in the T group is greater than that in the C group (if the event represents a potential harm). Of note, if the event represents a benefit, but its incidence is less in the T group than in the C group, it would be appropriate to use NNH instead of NNT. This is another complexity of interpreting NNT.

NNH is computed in the same way as NNT but represents a different measure. Consider the trial in the previous example, in which the NNT for MI was 304 patients. If the rates of proteinuria were 0.5% in the T group and 0.4% in the C group, then $\text{ARD} = 0.001$ and $\text{NNH} = 1000$. Thus, if MI and proteinuria were assessed in the same trial, 304 patients would need to be treated, on average, to prevent one fewer patient experiencing an MI, and 1000 patients would need to be treated to observe one more patient with proteinuria.

Statistical Properties of ARD and NNT

Population parameters (e.g., true ARD and NNT) are infrequently known. Data for a random sample from the population are used to estimate the parameters. This gives rise to point estimates. Different samples may contain data on different individuals. Thus, it is expected that estimates will vary from one sample to another.

Point estimates and their variability are used to compute a confidence interval (CI) with

lower and upper limits (L, U) on the population parameter. For any given data, the endpoints L and U are real numbers. To say that (L, U) is a 95% CI on the true ARD does not mean that the true ARD will fall within this interval, but rather that 95% of the intervals derived would contain the true mean ARD if the experiment was repeated multiple times.

The statistical methodology for computing CIs on the true ARD is well established. However, this is not the case for computing CIs on the true NNT. There are several reasons for this: (a) NNT is not defined in the interval (–1, +1), (b) the variance of NNT is difficult to estimate accurately, and (c) the sampling distribution appears to be triangular [15].

ARD is a summary statistic that synthesizes event information on patients in the T and C groups. Therefore, its variance consists of both within- and between-group variability, and it can be written in a closed form. This variance is a primary driver of the length of the CI on ARD. On the other hand, the NNT is a transformation of ARD, producing a single point estimate without quantifiable variability. A point estimate without its associated variability provides no information for one to judge the proximity of the sample estimate of treatment effect to the unknown population parameter (the true treatment effect) about which one wishes to make a statistical inference. Therefore, without an established method for quantifying the variability of a point estimate, it is impossible to know whether the estimated treatment effect is indicative of a real effect (see Hotelling [16]).

Several authors (e.g., Cook and Sackett [2], Schulzer and Mancini [17], Altman [18], and Bender [19]) have proposed different methods for computing the CI on NNT. For instance, Cook and Sackett [2] proposed computing the CI on ARD and then inverting and reversing the order of the endpoint of the interval to get an approximate CI on NNT.

There are instances, however, where the method of Cook and Sackett [2] and Altman [18] for computing CIs on NNT can produce unreasonable results in which the CI does not contain the point estimate of NNT. One such situation is when the ARD is not statistically significant, in which case the T group may be descriptively favored over the C group, or vice versa. When there is no statistically significant difference between treatment groups, there is no finite single value corresponding to instances where NNT is meaningful [8]. For example, when the estimate of ARD in a clinical trial comparing T to C was 0.05 and the 95% CI on the true ARD was (–0.10, +0.20), the point estimate of NNT is 20 patients, and by inversion the approximate 95% CI on NNT has a lower endpoint of +5 and an upper endpoint of –10. Not only is the point estimate of NNT not in the CI, but the CI is not valid (i.e., a positive lower limit and negative upper limit).

When there is no meaningful difference between treatment groups, NNT should not be computed. This is because when the ARD is not statistically significant, it is doubtful that the NNT has a meaningful clinical utility as there is little or no evidence that the observed ARD reflects the true state of nature. In general, if the (L, U) of a CI on the true ARD are negative and positive, respectively (ARD is not statistically significant), it is not advisable to generate a CI on the NNT by inverting the limits and reversing their order.

NNT IN MS CLINICAL TRIALS

Common efficacy outcome data from RRMS trials include the percentages of patients with events (estimated as cumulative time-to-event percentages for disability progression or first relapse), the frequency of recurrent events (annualized relapse rate [ARR]), and the number of magnetic resonance imaging lesions. Therapies showing higher efficacy could also have serious harmful side effects. Thus, if there is a desire to compute the NNT for MS therapies, it would be appropriate to report both the NNT and NNH to aid the evaluation of the benefit-to-risk profiles of the treatments. The following discussion is intended to help with the appraisal of NNT when computed in MS trials.

Time-to-Event Efficacy Endpoints for NNT

A time-to-event endpoint measures “if” an event occurred and “when” the event occurred.

Thus, for time to relapse or disability worsening, one can obtain the median (or other percentile) time to relapse or disability worsening as well as the proportion of patients with these events. When there is no censoring, the estimated cumulative percentage of patients with the event is the same as the binomial proportion (number relapsed/number at risk). For a time-to-event endpoint, the definition of NNT may be applied to cumulative time-to-event percentages. If a protocol specifies a fixed period of treatment, e.g., 2 years, and there is an interest in estimating the cumulative percentage of patients with events at 2 years, then time-to-event methods would have to be used given that the event is censored in some patients. If the 2-year cumulative relapse proportions are, for example, 30% for the T group and 40% for the C group, then ARD = 0.10 and NNT = 10; this is interpreted as indicating that 10 patients would need to be treated for up to 2 years in order to have, on average, one fewer patient experiencing a relapse.

NNT could be time dependent; hence, it should not be computed using time-to-event methods at intermediate points during the specified treatment period, as it may not lead to consistent meaningful results. This is because NNT would be difficult to interpret for MS trials where the ARD was smaller early and larger later in the treatment period, as the implication is that a greater number of patients would need to be treated for a shorter period of time and fewer would need to be treated for a longer period of time. In the DEFINE study [20], the Kaplan–Meier estimates of the proportion of patients that relapsed at week 24 are 0.167 and 0.110 for placebo and delayed-release dimethyl fumarate (DMF; also known as gastro-resistant DMF) twice daily (BID), respectively. This gives an NNT of 18, suggesting that 18 patients would need to be treated for up to 24 weeks in order to have, on average, one fewer patient experiencing a relapse. However, the Kaplan–Meier estimates of the proportion of patients that relapsed at week 96 in the same study are 0.461 and 0.270 for placebo and DMF BID, respectively. This gives a NNT of 6; i.e., the same therapy (DMF) at the same dose and with the same effect on biology yields a different NNT than that obtained in the same study when computed at a different time point. Additionally, NNT does not provide information on the average or median time before the event occurs. Therefore, the time ("when") aspect of the time-to-event endpoint is often not reflected by NNT. Furthermore, NNT could vary between the first time-to-event and the end of treatment. Besides, the course and severity of MS can change unpredictably [21]. Thus, the computation of NNT at intermediate time points during a treatment period is of doubtful utility.

ARR for NNT

For recurrent events such as MS relapse, measures similar to NNT (referred to as NNT-like measures, NNT-L) have been proposed based on reducing the occurrence of events [22]. ARR is computed as the total number of relapses in a given period divided by the total number of person-years in that period. Because a patient may experience more than one relapse over the period of a study, relapses within each patient may be correlated. Additionally, the use of person-years assumes that the probability of relapse is constant over time. For example, 50 relapses in 100 patients over 10 years of exposure would give the same relapse rate as 50 relapses in 1000 patients for 1 year of exposure. The idea is that person-years of exposure is the sum of person-years exposed for each patient summed over patients. However, because the change in ARR over time has been documented in MS, computing NNT using ARR may confound well-known changes in ARR over time [23–25]. We point out that some authors compute NNT-L measures using the definition of NNT and interpret them as one interprets NNT. We argue that such an interpretation is wrong.

The event-based NNT-L measure may be heavily influenced by small numbers of patients with multiple events (e.g., highly active patients) and may give a fractional value that is <1. When such a fractional value is rounded up to 1, it will suggest that one patient needs to be treated to prevent one relapse [26].

Table 1 presents hypothetical relapse data from 20 patients treated over 2 years, showing

that when some patients have multiple relapses, the event-based NNT does not equal the traditional NNT. Thus, it is easy to misinterpret the event-based NNT in Table 1 as indicating that five patients need to be treated, on average, to prevent one relapse. However, this is half the number of patients that need to be treated to prevent one relapse when NNT interpretation is based on the traditional definition of NNT. Hence, a distinction should be made between NNT and NNT-L measures. Of note, NNT is based on ARD: (T%–C%)/100, i.e., the difference in the percentages of patients with a critical event of interest by a specified time of exposure. Therefore, ARR should not be used to compute NNT.

The original definition of NNT provides an answer in terms of number of patients. NNT-L (such as rate per year) does not have this property. At best, the computation of such event-based measures results in the number of patient-years of treatment, not the number of patients. For a chronic disease such as MS, in which treatment may be lifelong, describing efficacy in terms of a lower number of patient-years of treatment may not be relevant for clinical decisions. Again, the fact that MS disease course and severity can change unpredictably invalidates any attempt to interpret treatment effect in terms of number of patient-years of treatment.

Table 2 shows data from two clinical trials (Copolymer 1 [NCT00004814] and CONFIRM [NCT00451451]) [27] where glatiramer acetate (GA) was administered at a dose of 20 mg/day for 2 years. The first study [28] is a pivotal phase

Table 1 Example of the difference between event-based NNT and traditional NNT

	Placebo patient number	**Number of placebo relapses over 2 years (*n*)**	**Treatment patient number**	**Number of treatment relapses over 2 years (*n*)**
	1	3	1	2
	2	2	2	1
	3	2	3	1
	4	1	4	0
	5	0	5	0
	6	0	6	0
	7	0	7	0
	8	0	8	0
	9	0	9	0
	10	0	10	0
Proportion of patients who relapsed		0.40		0.30
Total relapses (*n*)		8		4
ARR		8/20 PY = 0.40		4/20 PY = 0.20
Traditional NNT			1/(0.40–0.30) = 10	
Event-based NNT			1/(0.40–0.20) = 5	

ARR annualized relapse rate, *NNT* number needed to treat, *PY* patient-years

Table 2 Example of the same RRR but different NNT values from ARR and ARD

Results for GA	Copolymer 1	CONFIRM
Placebo group ARR	0.84 ($n = 126$)	0.40 ($n = 363$)
Treatment group ARR	0.59 ($n = 125$)	0.29 ($n = 350$)
Absolute ARR difference	0.25	0.11
Event-based NNT	4	10
RRR vs. placebo (%)	30[a,b]	29[a,b]
Proportion of placebo group who were relapse-free	0.27	0.59
Proportion of treatment group who were relapse-free	0.34	0.68
ARD	0.07	0.09
Traditional NNT	15	12

Data from the Copolymer 1 and CONFIRM MS clinical trials [25, 26]
ARD absolute risk difference, *ARR* annualized relapse rate, *GA* glatiramer acetate, *MS* multiple sclerosis, *NNT* number needed to treat, *RRR* relative rate reduction
[a] $p < 0.05$
[b] Same RRR

III comparative trial of patients with RRMS, and the second is a pivotal comparative trial of oral DMF, in which GA was included as a reference comparator for assay sensitivity [27]. The same relative rate reduction (RRR) vs. placebo on ARR was seen for Copolymer 1 and CONFIRM: 30% and 29%, respectively (Table 2). This is despite obtaining different NNTs from ARR and ARD. The fallacy in the event-based NNT from the two studies suggests that four and ten patients, respectively, would need to be treated with GA for one patient to benefit from treatment. However, when the traditional definition of NNT is applied to ARD, the NNT is 15 and 12, respectively.

Two clinical trials (DEFINE [NCT00420212] and CONFIRM [20, 27]) assessed DMF 240 mg BID over a period of 2 years. Table 3 details different RRRs on ARR for the two studies, identical NNTs computed from ARR and different NNTs computed from ARD. The RRR vs. placebo was 53% and 44%, respectively, for DEFINE and CONFIRM. However, the NNTs computed from ARR were the same for both studies (6), and the NNTs computed from ARD were different from the NNTs computed from ARR. Because the NNTs computed using ARR from the DEFINE and CONFIRM studies are identical, the known changes in ARR over time are obscured [23–25].

Direct Comparison of NNT Across Trials

Comparison of NNTs across MS trials warrants caution because of the possible difference in the underlying baseline risk between studies. Even though the baseline risk may be the same in the overall study populations, the baseline risk may differ within the levels of disease severity. Specifically, in MS trials, the computed NNT values in two studies may be the same and may suggest a treatment benefit in the overall study population, whereas the benefit might be limited to patient groups with low or intermediate risk, or who are newly diagnosed or treatment naïve. Thus, patient groups with a treatment benefit may differ across studies. Because the design, baseline characteristics, and percentages of these risk groups are likely to differ across studies, it should not be surprising that the NNT is different when compared across trials for the same therapy. As reported in Table 2, not only were the estimated NNTs for the comparison of GA vs. placebo different in the Copolymer 1 and

Table 3 Example of different RRRs but the same NNT calculated from ARR. Data from the DEFINE and CONFIRM MS clinical trials [18, 25]

	DMF[a] 240 mg twice daily	
	DEFINE	CONFIRM
Placebo group ARR	0.36	0.40
Treatment group ARR	0.17	0.22
Absolute ARR difference	0.19	0.18
Event-based NNT	6	6
RRR vs. placebo (%)	53	44
p value vs. placebo	<0.0001	<0.001
Proportion of placebo group who were relapse-free	0.54	0.59
Proportion of treatment group who were relapse-free	0.73	0.71
ARD	0.19	0.12
Traditional NNT	6	9

ARD absolute risk difference, *ARR* annualized relapse rate, *MS* multiple sclerosis, *NNT* number needed to treat, *RRR* relative rate reduction

[a] Delayed-release dimethyl fumarate (DMF; also known as gastro-resistant DMF)

CONFIRM trials, but the event-based and traditional NNTs produced conflicting results. The event-based NNT from Copolymer 1 was lower than that in the CONFIRM study; however, the converse was true with the traditional NNT. Additionally, ARDs, and hence NNTs, are greatly influenced by placebo risks and rates; therefore, they are study specific and cannot be generalized to all MS populations or compared across studies. As indicated by the GA example above, when comparing across studies, it is possible to arrive at a different conclusion using different or the same endpoints. It is only appropriate to compare NNTs directly when outcomes, study durations, and reference populations are similar, as results are directly related to the control or comparison group.

Furthermore, it is not advisable to perform a meta-analysis of individual NNT estimates from a series of clinical trials directly comparing treatments to get an overall NNT across the collection of trials. There are many reasons for this: (a) incompatibility between the average NNT from the meta-analysis as contrasted with the NNT produced by reciprocating the inverse of the average ARD from a meta-analysis; (b) the variance of individual NNTs is approximate, so the variance of the overall estimate of NNT across trials would compound the approximation; and (c) the variability of the control rate across trials may be large even when the ARD and NNT are consistent across trials, which impacts the interpretation of the combined estimate of NNT across trials (e.g., 100% T vs. 99% C and 5% T vs. 4% C both lead to an NNT of 100). Smeeth et al. [29] questioned the viability of conducting meta-analyses based on NNT. They acknowledge that, although the results are sometimes informative, they are usually misleading. In conclusion, if clinical trials are sufficiently similar to substantiate a scientifically meaningful meta-analysis of the RD or ARD, then meta-analysis [30] of RD or ARD across the trials should be performed to get an overall RD or ARD, after which the definition of NNT can be applied to the overall ARD.

CONCLUSION

This article provides a review of the definition and methodological characteristics of NNT; it also discusses the relative merits of its use in

comparative clinical RRMS trials. The authors caution against: (a) the use of NNT in comparative trials where the ARD is not statistically significant, (b) computing NNT using a time--to-event endpoint at intermediate time points, (c) computing NNT using ARR, and (d) making comparisons or performing inferential analyses of NNT across trials.

ACKNOWLEDGEMENTS

Biogen provided funding for editorial support in the development of this paper, and for the article processing charges; Ana Antaloae, PhD, from Excel Scientific Solutions incorporated feedback from authors, and Kristen DeYoung from Excel Scientific Solutions copyedited and styled the manuscript per journal requirements. Biogen reviewed and provided feedback on the paper to the authors. The authors had full editorial control of the paper, and provided their final approval of all content.

All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this manuscript, take responsibility for the integrity of the work as a whole, and have given final approval to the version to be published.

Disclosures. Macaulay Okwuokenye is an employee of and holds stock/stock options in Biogen. Annie Zhang was an employee of Biogen during manuscript development. Amy Pace was an employee of Biogen during manuscript development and holds stock in Biogen; she is currently an employee of Alexion Pharmaceuticals. Karl E. Peace is a consultant to Biogen.

Compliance with Ethics Guidelines. This article is based on previously conducted studies and does not involve any new studies of human or animal subjects performed by any of the authors.

Data Availability. Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

REFERENCES

1. Laupacis A, Sackett DL, Roberts RS. An asssessment of clinically useful measures of the consequences of treatment. N Engl J Med. 1988;318:1728–33.
2. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ. 1995;310:452–4.
3. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ. 1999;319:1492–5.
4. Moore RA, Gavaghan DJ, Edwards JE, Wiffen P, McQuay HJ. Pooling data for number needed to treat: no problems for apples. BMC Med Res Methodol. 2002;2:2.
5. Cates CJ. Simpson's paradox and calculation of number needed to treat from meta-analysis. BMC Med Res Methodol. 2002;2:1.
6. Grieve AP. The number needed to treat: a useful clinical measure or a case of the Emperor's new clothes. Pharm Stat. 2003;2:87–102.
7. Hutton JL. Number needed to treat: properties and problems. J R Stat Soc Ser A. 2000;163:403–15.
8. Hutton JL. Number needed to treat and number needed to harm are not the best way to report and assess the results of randomised clinical trials. Br J Haematol. 2009;146:27–30.
9. Klawiter EC, Cross AH, Naismith RT. The present efficacy of multiple sclerosis therapeutics: is the new 66% just the old 33%? Neurology. 2009;73:984–90.
10. Zakaria M. Smoke and mirrors: limited value of relative risk reductions for assessing the benefits of disease-modifying therapies for multiple sclerosis. Mult Scler Relat Disord. 2015;4:187–91.

11. Leist TP, Freedman MS, Miller AE, Dive-Pouletty C, Montalban X. Assessing comparative outcomes from teriflunomide and dimethyl fumarate studies in relapsing MS: use of "number needed to treat" analysis. In: 67th Annual Meeting of the American Academy of Neurology, 18–25 Apr 2015, Washington, DC, USA.

12. Leist TP, Miller AE, Thangavelu K, Hass S, Freedman MS. Number needed to treat analysis comparing teriflunomide and injectable disease-modifying therapies. In: 68th Annual Meeting of the American Academy of Neurology, 15–21 Apr 2016, Vancouver, BC, Canada.

13. Freedman MS, Montalban X, Miller AE, et al. Comparing outcomes from clinical studies of oral disease-modifying therapies (dimethyl fumarate, fingolimod, and teriflunomide) in relapsing MS: assessing absolute differences using a number needed to treat analysis. Mult Scler Relat Disord. 2016;10:204–12.

14. Mendes D, Alves C, Batel-Marques F. Benefit-risk of therapies for relapsing-remitting multiple sclerosis: testing the number needed to treat to benefit (NNTB), number needed to treat to harm (NNTH) and the likelihood to be helped or harmed (LHH): a systematic review and meta-analysis. CNS Drugs. 2016;30:909–29.

15. Okwuokenye M, Peace EK. The use of number needed to treat in clinical trials. 2015. **(Unpublished work)**

16. Hotelling H. Recent improvements in statistical inference. J Am Stat Assoc. 1931;26:79–87.

17. Schultzer M, Mancini GBJ. 'Unqualified success' and 'unmitigated failure'; number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment induced adverse effects. Int J Epidemiol. 1996; 25:704–12.

18. Altman DG. Confidence intervals for the number needed to treat. BMJ. 1998;317:1309–12.

19. Bender R. Calculating confidence intervals for the number needed to treat. Control Clin Trials. 2001;22:102–10.

20. Gold R, Kappos L, Arnold DL, et al. DEFINE Study Investigators. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. N Engl J Med. 2012;367:1098–107.

21. Lublin FD, Reingold SC, Cohen JA, et al. Defining the clinical course of multiple sclerosis: the 2013 revisions. Neurology. 2014;83:278–86.

22. Cook RJ. Number needed to treat for recurrent events. J Biom Biostat. 2013;4:167.

23. Tremlett H, Zhao Y, Joseph J, Devonshire V. Relapses in multiple sclerosis are age- and time-dependent. J Neurol Neurosurg Psychiatry. 2008; 79:1368–74.

24. Nicholas R, Straube S, Schmidli H, Pfeiffer S, Friede T. Time-patterns of annualized relapse rates in randomized placebo-controlled clinical trials in relapsing multiple sclerosis: a systematic review and meta-analysis. Mult Scler. 2012;18:1290–6.

25. Nicholas R, Straube S, Schmidli H, Schneider S, Friede T. Trends in annualized relapse rates in relapsing-remitting multiple sclerosis and consequences for clinical trial design. Mult Scler. 2011;17:1211–7.

26. Aaron SD, Fergusson DA. Exaggeration of treatment benefits using the "event-based" number needed to treat. CMAJ. 2008;179:669–71.

27. Fox RJ, Miller DH, Phillips JT, et al. CONFIRM Study Investigators. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. N Engl J Med. 2012;367:1087–97.

28. Johnson KP, Brooks BR, Cohen JA, Copolymer 1 Multiple Sclerosis Study Group, et al. Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: results of a phase III multicenter, double-blind placebo-controlled trial. Neurology. 1995;45:1268–76.

29. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading. BMJ. 1999; 318:1548–51.

30. Chen D-G, Peace KE. Applied meta-analysis using R. London: Taylor & Francis; 2013.